

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. All Rights Reserved. © The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2022

This work is subject to copyright. All rights are solely and exclusively licensed by the Publisher, whether the whole or part of the material is concerned, specifically the rights of reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

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
Beijing, China
Haikou, China
Tianjin, China

Xiaoli Chu, Ph.D.
Yue Huang, Ph.D.
Yong-Huan Yun, Ph.D.
Xihui Bian, Ph.D.

Contents

1	Chemometric Methods in Analytical Spectroscopy Technology	1
1.1	Introduction	1
1.1.1	Overview of Chemometrics	2
1.1.2	Analysis of Spectroscopy Combined with Chemometrics	15
1.1.3	Beginning of Modern Spectroscopy Technology—The Contribution of Karl Norris	20
	References	27
2	Modern Spectral Analysis Techniques	31
2.1	Introduction	31
2.2	Near-Infrared Spectroscopy	34
2.2.1	Micro Near-Infrared Spectral Analysis Technology	36
2.2.2	Online Near-Infrared Spectral Analysis Technology	37
2.2.3	Standard Methods for Near-Infrared Spectroscopy	39
2.3	Mid-Infrared Spectroscopy	47
2.3.1	Portable Mid-Infrared Spectral Analysis Technology	48
2.3.2	Online Mid-Infrared Spectral Analysis Technology	49
2.4	Raman Spectroscopy	50
2.4.1	Fourier Transform Raman Spectroscopy	51
2.4.2	Surface Enhanced Raman Scattering Spectroscopy	51
2.4.3	Confocal Raman Spectroscopy	53
2.4.4	Spatial Offset Raman Spectroscopy	55
2.4.5	Transmitted Raman Spectroscopy	57
2.4.6	Portable Raman Spectral Analysis Technology	59

2.4.7	Fiber Raman Spectral Analysis Technology	60
2.5	Ultraviolet-Visible Spectroscopy	61
2.6	Molecular Fluorescence Spectroscopy	64
2.6.1	Three-Dimensional Fluorescence Spectroscopy	65
2.6.2	Laser-Induced Fluorescence Spectroscopy	67
2.7	Low-Field NMR Spectroscopy	67
2.8	Terahertz Spectroscopy	70
2.9	Laser-Induced Breakdown Spectroscopy	72
2.10	Spectral Imaging	74
	References	80
3	Basis of Matrices and Mathematical Statistics	89
3.1	Basis of Matrix	89
3.2	Matrix Representation of Lambert-Beer's Law	92
3.3	Variance and Normal Distribution	93
3.4	Significance Test	97
3.5	Correlation Coefficient	99
3.6	Covariance and Covariance Matrix	100
3.7	Multivariable Graph Representation	102
3.7.1	Spatial Representation of Samples	102
3.7.2	Box Plot	104
3.7.3	Radar Chart	105
	References	108
4	Spectral Preprocessing Methods	111
4.1	Mean Centering	111
4.2	Auto-scaling	113
4.3	Normalization	114
4.4	Smoothing	114
4.4.1	Moving Average Smoothing	115
4.4.2	Savitzky-Golay Convolution Smoothing	116
4.4.3	Fourier Transform and Wavelet Transform	117
4.5	Continuum Removed	119
4.6	Adaptive Iteratively Reweighted Penalized Least Squares	120
4.7	Derivative	122
4.7.1	Norris Method	122
4.7.2	Savitzky-Golay Convolution for Derivative Calculation	123
4.7.3	Wavelet Transform for Derivative Calculation	125
4.7.4	Fractional Derivative	128
4.8	Standard Normal Variate and De-Trending	129
4.9	Multiplicative Scatter Correction	132
4.10	Vector Angle Conversion	134
4.11	Fourier Transform	135
4.12	Wavelet Transform	137
4.13	Image Moment Methods	144

4.14	External Parameter Orthogonalization	147
4.15	Generalized Least Squares Weighting	148
4.16	Loading Space Standardization	149
4.17	Oblique Projection	150
4.18	Orthogonal Signal Correction	151
4.18.1	Wold Algorithm	152
4.18.2	Fearn Algorithm	152
4.18.3	Direct Orthogonal Signal Correction Algorithm	154
4.18.4	Direct Orthogonal Algorithm	155
4.18.5	Application of Orthogonal Signal Correction Algorithm	156
4.19	Net Analyte Signal	157
4.20	Optical Path-Length Estimation and Correction	158
4.21	Two-Dimensional Correlation Spectroscopy	160
	References	162
5	Wavelength Selection Methods	169
5.1	Correlation Coefficient and Analysis of Variance Method	170
5.2	Simple-To-Use Interactive Self-modeling Mixture Analysis Method	173
5.3	Successive Projections Algorithm	174
5.4	Variable Importance in Projection	175
5.5	Interval Partial Least Squares Method	176
5.6	Moving Window PLS	176
5.7	Recursive Weighted PLS	178
5.8	Elimination of Uninformative Variables	178
5.9	Global Optimization Methods	181
5.9.1	Genetic Algorithm	181
5.9.2	Simulated Annealing Algorithm	184
5.9.3	Particle Swarm Optimization	185
5.9.4	Ant Colony Algorithm	187
5.10	Model Population Analysis-Based Methods	189
5.10.1	Competitive Adaptive Reweighted Sampling	190
5.10.2	Iteratively Retaining Informative Variables	192
5.10.3	Variable Combination Population Analysis	195
5.10.4	Other Methods	197
5.10.5	Wavelength Selection Method Based on Hybrid Strategy	197
5.11	The Selection of Spectral Preprocessing and Wavelength Selection Methods	200
	References	202

6	Spectral Dimensionality Reduction Methods	209
6.1	The Multicollinearity Problem	209
6.2	Principal Component Analysis	213
6.2.1	Theory of Principal Component Analysis	213
6.2.2	Determination of Principal Component Number	215
6.2.3	Algorithm of Principal Component Analysis	216
6.2.4	Application of Principal Component Analysis	217
6.2.5	Multivariate Resolution Alternating Least Squares	218
6.2.6	Band Target Entropy Minimization	219
6.2.7	Multilevel Simultaneous Component Analysis	221
6.3	Non-negative Matrix Factorization	222
6.4	Independent Component Analysis	224
6.5	Multi-dimensional Scaling Transformation	225
6.6	Isometric Mapping	226
6.7	Local Linear Embedding	229
6.8	T-Distributed Stochastic Neighborhood Embedding	230
6.9	Other Algorithms	233
	References	233
7	Linear Calibration Methods	237
7.1	Univariate Linear Regression	237
7.2	Multiple Linear Regression	238
7.3	Concentration Residual Augmented Classical Least Squares	239
7.4	Stepwise Linear Regression	240
7.5	Ridge Regression	241
7.6	Lasso Regression	241
7.7	Least Angle Regression	242
7.8	Elastic Net	243
7.9	Principal Component Regression	244
7.9.1	Theory	244
7.9.2	Method for Selecting the Optimal PCs	245
7.9.3	Partial Least Squares Regression	249
	References	252
8	Nonlinear Calibration Methods	255
8.1	Artificial Neural Network	255
8.1.1	Introduction	255
8.1.2	Back Propagation-Artificial Neural Network	260
8.1.3	Design of BP-ANN	264
8.1.4	Other Types of Neural Networks	267
8.1.5	Optimization of Neural Network Parameters	270
8.2	Support Vector Machine	271
8.2.1	Introduction	271
8.2.2	Support Vector Regression	277
8.2.3	Least Squares Support Vector Regression	280

8.2.4	Optimization of Support Vector Regression Parameters	281
8.3	Relevance Vector Machine	283
8.4	Kernel Partial Least Squares	285
8.5	Extreme Learning Machine	287
8.6	Gaussian Process Regression	289
	References	293
9	Method of Selecting Calibration Samples	297
9.1	Introduction	297
9.2	Kennard-Stone Method	302
9.3	Sample Set Partitioning Based on Joint X–Y Distances (SPXY) Method	303
9.4	Optimizable K-dissimilarity Selection Method	303
9.5	Other Methods	304
	References	307
10	Detection Methods for Outlier Samples	309
10.1	Detection of Outlier Samples During Calibration Process	309
10.2	Detection of Outlier Samples During the Prediction Process	310
10.3	Other Detection Methods	313
	References	314
11	Maintenance and Update of Calibration Model	317
11.1	Necessity	317
11.2	Recursive Exponentially Weighted PLS	321
11.3	Block-Wise Recursive PLS	323
11.4	Just-In-Time Learning and Active Learning	325
	References	325
12	Pattern Recognition Methods	329
12.1	Introduction	329
12.2	Unsupervised Pattern Recognition Methods	331
12.2.1	Similarity Coefficients and Distances	331
12.2.2	Hierarchical Cluster Analysis	333
12.2.3	K-Means Clustering	335
12.2.4	Fuzzy K-Means Clustering	337
12.2.5	Gaussian Mixture Model	339
12.2.6	Self-organizing Neural Network	340
12.3	Supervised Pattern Recognition Methods	343
12.3.1	Minimum Distance Discriminant Method	343
12.3.2	Canonical Variate Analysis	344
12.3.3	K-Nearest Neighbor	348
12.3.4	Soft Independent Modeling of Class Analogy	349
12.3.5	Logistic Regression	352
12.3.6	Soft-Max Classifier	354
12.3.7	Random Forest	356

12.3.8	Application of Regression Methods for Discriminant Analysis	359
12.4	Spectral Searching Methods	360
12.4.1	Introduction	360
12.4.2	Spectral Searching Algorithms	363
12.4.3	Improvements of Spectral Searching Algorithms	366
12.4.4	Spectral Searching Strategies and Applications	370
	References	374
13	Model Evaluation	381
13.1	Evaluation of Quantitative Calibration Model	381
13.1.1	Evaluation Parameters	381
13.1.2	Model Evaluation	385
13.2	Evaluation of Performance of Pattern Recognition Model	392
	References	397
14	Methods for Improving Prediction Ability of Model	399
14.1	Modeling Strategies for Improving the Robustness	399
14.2	Modeling Strategies Based on Local Samples	400
14.3	Ensemble Modeling Strategies	402
14.3.1	Bagging Ensemble Strategy	403
14.3.2	Boosting Ensemble Strategy	404
14.3.3	Stacked Ensemble Strategy	407
14.3.4	Stacked Generalization Strategy	409
14.4	Virtual Sample Modeling Strategy	411
14.5	Semi-supervised Learning Methods	413
14.6	Multi-target Regression Strategy	416
	References	417
15	Multi-spectral Fusion Technology	423
15.1	Fusion Strategies and Methods	423
15.2	Multi-block Partial Least Squares Method	428
15.3	Sequential and Orthogonal Partial Least Squares Method	430
15.4	Research on Application of Multi-Spectral Fusion	431
15.5	Future Prospect	436
	References	436
16	Multi-way Resolution and Calibration Methods	439
16.1	Introduction	439
16.2	Parallel Factor Analysis	441
16.3	Alternating Trilinear Decomposition	444
16.4	Multi-way Partial Least Squares	445
	References	449

17 Calibration Transfer Methods	451
17.1 Introduction	451
17.2 Traditional Algorithms	453
17.2.1 Spectral Subtraction Correction	453
17.2.2 Shenk's Algorithm	453
17.2.3 Direct Standardization	454
17.2.4 Piecewise Direct Standardization	454
17.2.5 Procrustes Analysis	456
17.2.6 Target Transformation Factor Analysis	456
17.2.7 Maximum Likelihood Principal Component Analysis	457
17.2.8 Slope/Bias Correction	457
17.3 Improvement of Traditional Algorithms	458
17.4 New Algorithms	462
17.4.1 Canonical Correlation Analysis	462
17.4.2 Spectral Space Transformation	463
17.4.3 Alternating Trilinear Decomposition	464
17.4.4 Multi-task Learning	465
17.4.5 Generalized Least Squares	466
17.4.6 Other Algorithms	467
17.5 Global Calibration, Robust Calibration, and Model Update	471
17.6 Progress of Applications	476
17.6.1 SBC Method	476
17.6.2 SSC Method	476
17.6.3 Shenk's Method	477
17.6.4 DS Method	478
17.6.5 PDS Method	479
17.6.6 CCA Method	482
17.6.7 Establishment of Global Model	482
17.6.8 Other Methods	484
References	484
18 Deep Learning Methods	503
18.1 Stacked Auto-encoder	504
18.2 Convolution Neural Network	507
18.2.1 Basic Structure of CNN	507
18.2.2 Optimistic Algorithm	513
18.2.3 Loss Function	514
18.2.4 Activation Function	515
18.2.5 Methods to Avoid Over-Fitting	519
18.2.6 Classical Convolution Neural Network Architecture	521
18.2.7 Popular Deep Learning Software Framework	527
18.2.8 Design of Convolution Neural Networks	529
18.2.9 Training of Convolution Neural Networks	532

18.2.10	Advantages and Disadvantages of Convolution Neural Network	535
18.2.11	Applications of Convolution Neural Network	535
18.3	Deep Belief Network	543
18.4	Transfer Learning	546
	References	550
19	Chemometrics Software and Toolkits	555
19.1	Introduction	555
19.2	Basic Structure and Functions of Software	555
19.3	Common Software and Toolkits	558
	References	560
20	Discussion of Some Issues	563
20.1	Comparison of Different Spectroscopic Analysis	563
20.2	Selection of Chemometric Methods	566
20.2.1	Selection of Multivariate Calibration Methods	567
20.2.2	Selection of Pattern Recognition Methods	568
20.2.3	Selection of Spectral Preprocessing Methods and Spectral Variables	571
20.3	Influencing Factors of Model Prediction Ability	572
20.3.1	Effect of Calibration Samples	573
20.3.2	Effect of Reference Data	575
20.3.3	Effect of Spectral Measurement Methods	579
20.3.4	Effect of Spectral Acquisition Conditions	581
20.3.5	Effect of Instrument Performance	587
20.4	Outlook	587
	References	591

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

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

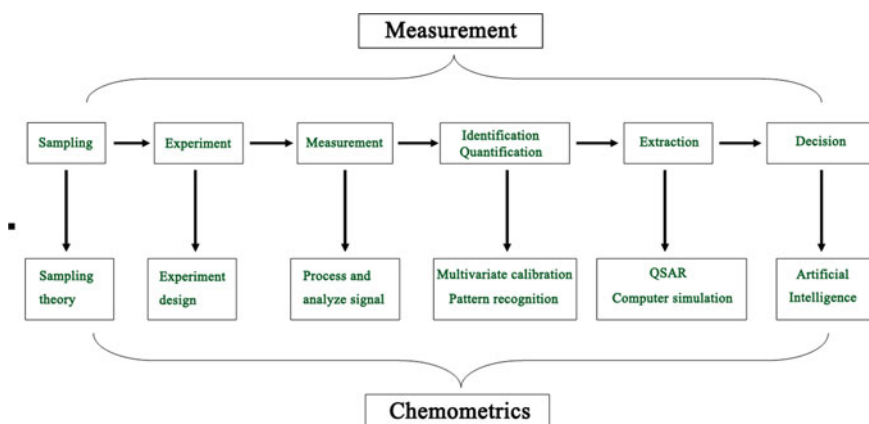


Fig. 1.1 Correspondence between the chemometrics and chemical measurement

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

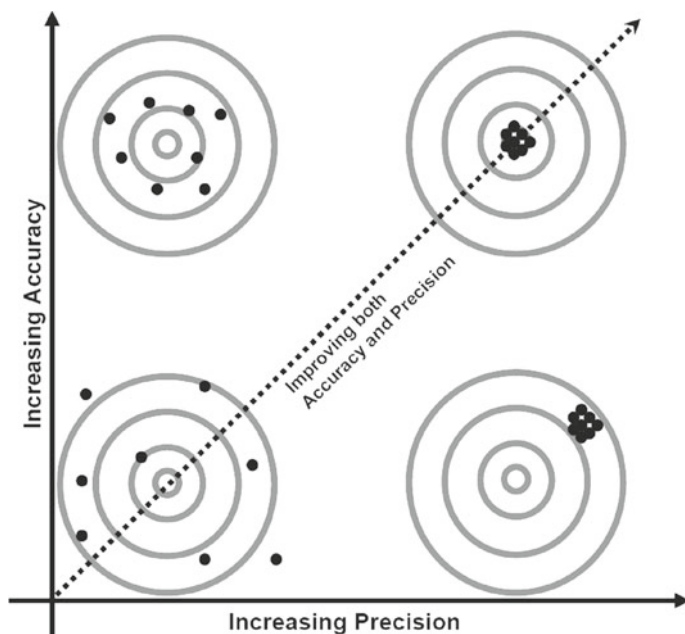


Fig. 1.2 Scheme to improve the accuracy and repeatability of analytical testing

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].

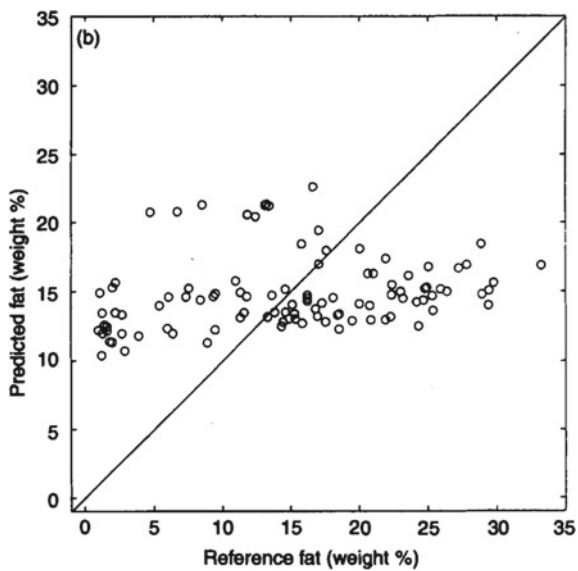


Fig. 1.3 Unary linear regression results of absorbance at 940 nm

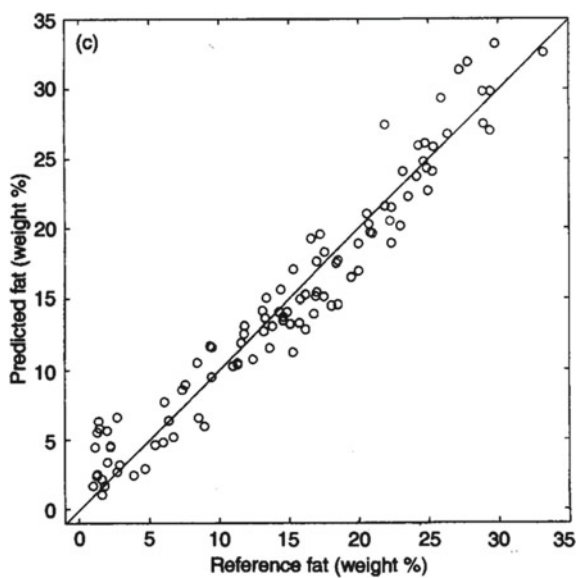


Fig. 1.4 The calibration result of the shortwave NIR full spectrum-PLS method

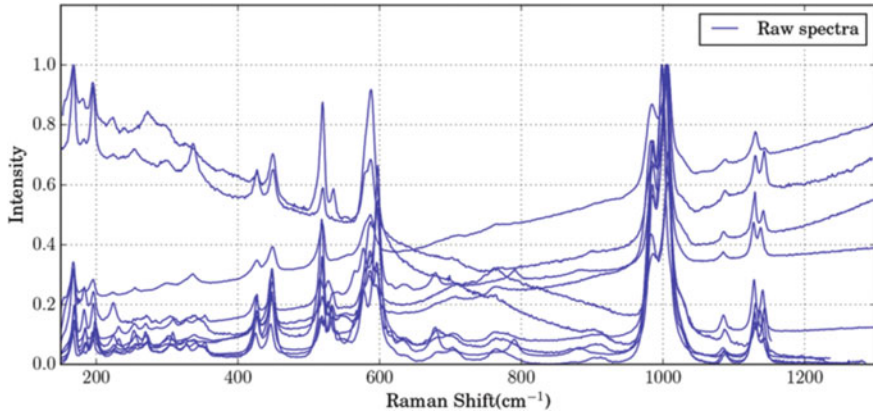


Fig. 1.5 Ten original Raman spectra of the same mineral from different origins in the RRUFF mineral database

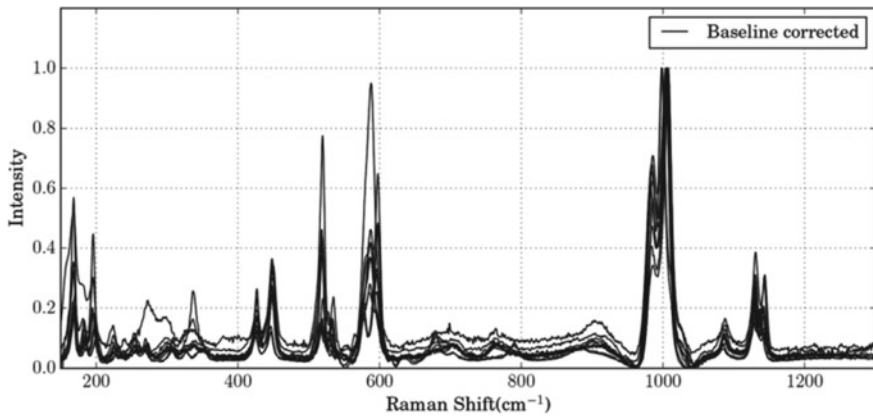


Fig. 1.6 Spectra of Fig. 1.5 after baseline correction

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].

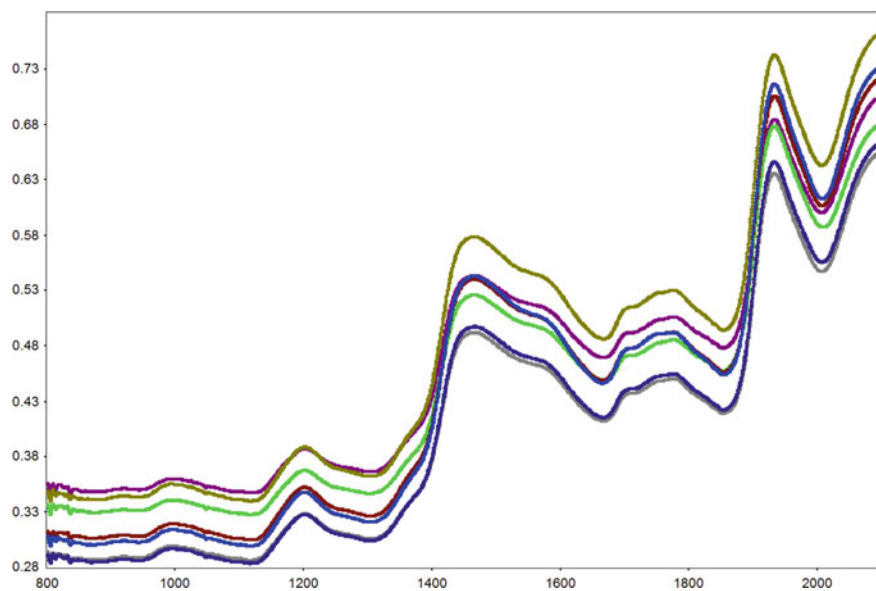


Fig. 1.7 Diffuse reflectance NIRS of different flour samples

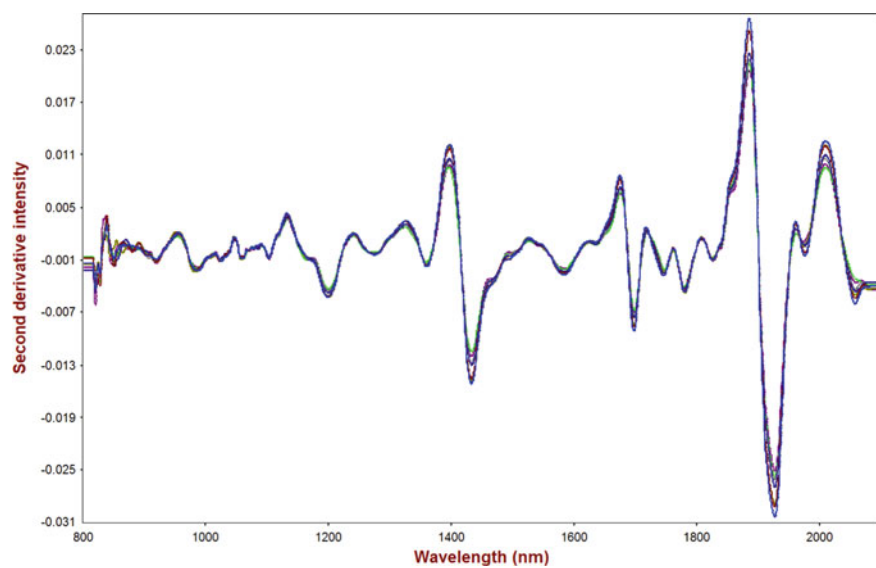


Fig. 1.8 Spectra of Fig. 1.7 after the second derivative processing

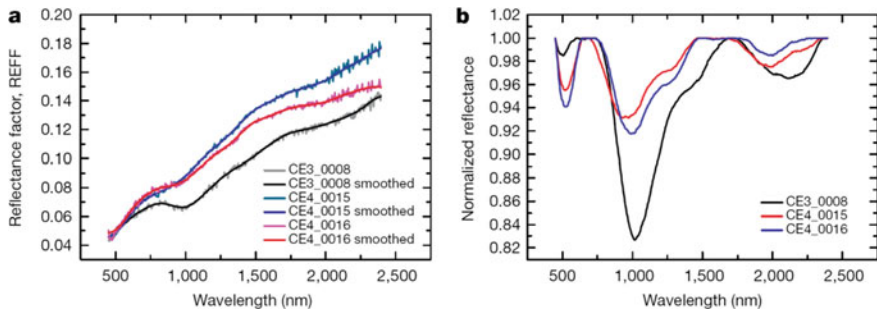


Fig. 1.9 a Diffuse reflectance near-infrared spectra of minerals on the lunar surface; Fig. 1.9 b Spectra processed by continuum removal method

- (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.

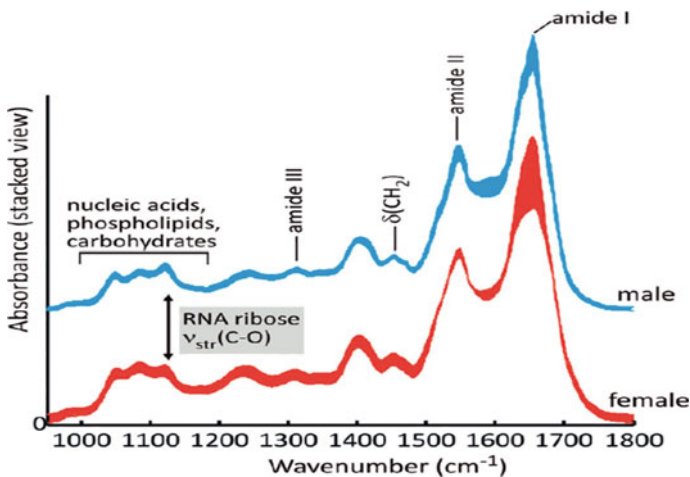


Fig. 1.10 Mid-infrared spectra of the root-tip material of different gender bird feathers

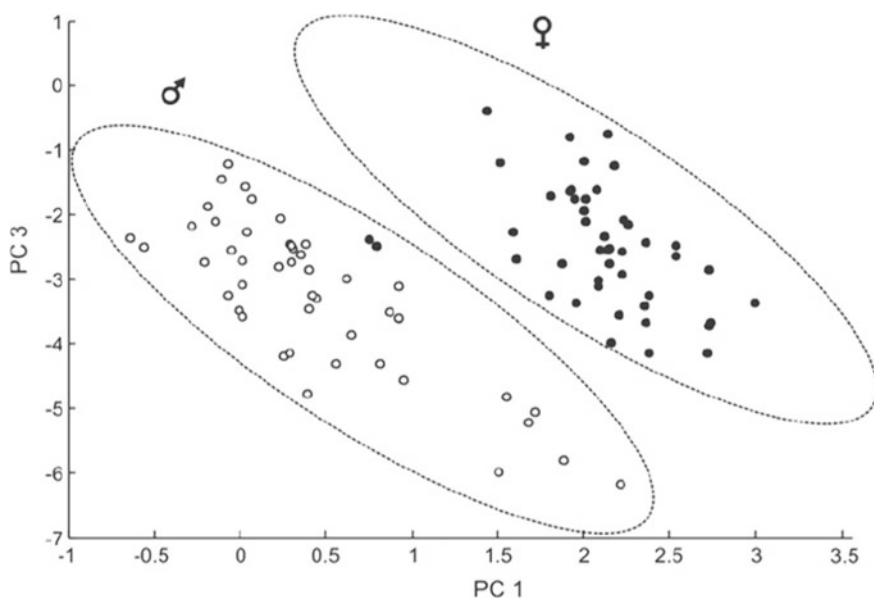


Fig. 1.11 The first and third principal component diagrams after principal component analysis

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

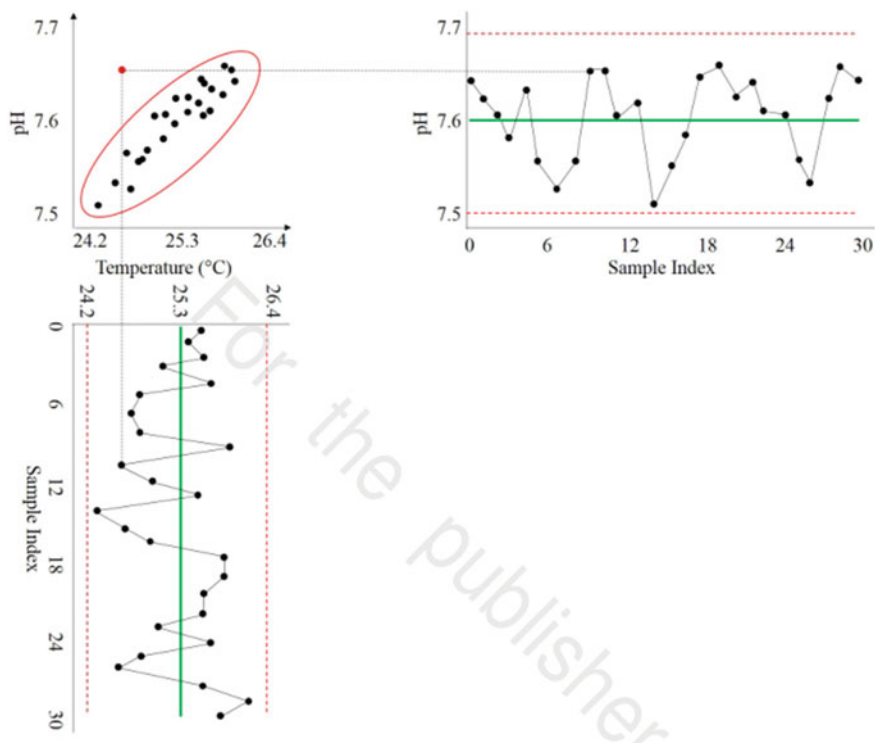


Fig. 1.12 Single and multi-variable control chart for judging abnormal points in the production process

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.

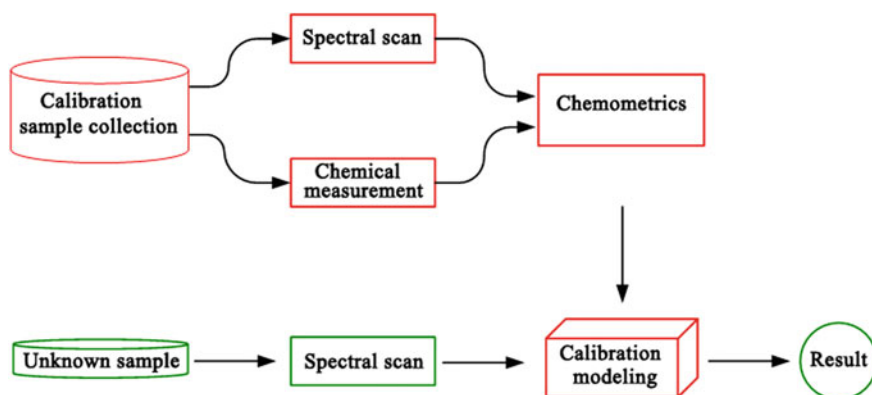


Fig. 1.13 Process of establishment of calibration model and prediction of unknown sample

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

from multiple measurements, use the same instrument and skilled operators to ensure the calibration as much as possible. Finally, the sample used for reference measurement must be the same as that used for spectral collection, and reference data and spectrum of sample should be tested as soon as possible after sampling, so as not to affect the accuracy of the calibration model due to changes in sample composition.

(5) Establishment of calibration

The sequence of calibration is roughly as follows: ① Formation a calibration matrix by using spectra and corresponding reference values; ② Mathematical transformation of the spectral data (i.e., preprocessing), such as derivative, wavelet transform, multiplicative scatter correction, mean-centering, etc.; ③ Selection of spectral variables (intervals), such as correlation coefficients, genetic algorithms, etc.; ④ Obtaining a quantitative calibration model by performing regression to preprocessing spectral and property values by PCR, PLS, or ANN. In the process, parameters such as the number of derivative points and the number of optimal PCs need to be determined; ⑤ Removal of outliers. An outlier refers to a sample whose predicted value obtained by the interactive validation is significantly different from its actual value; ⑥ Re-establish the model. After removing the outliers from the calibration set, the same calibration parameters are used to perform the regression again, and then repeat until a satisfactory quantitative model is obtained.

(6) Validation of model

After the model being built, a set of known samples (validation set) need to be used to validate the accuracy, stability, robustness, and transferability of the model. Validation set samples should contain all the components contained in the sample to be predicted, concentration range of which should cover at least 95% of that in the calibration set, with the uniform distribution. Plus, samples in the validation set should be enough for statistical testing, usually no less than 28 samples are required.

The robustness of a model refers to its performance against external interference factors. These influencing factors mainly include the replacement of the same type of test devices (such as cuvettes, optical fiber probes, integrating spheres, etc.), changes in the degree of fiber bending, replacement of light sources, replacement of reference materials (such as ceramic chips or barium sulfate powder, etc.), changes in sample loading conditions, changes in temperature (ambient and sample temperature), and changes in the physical state of particles (such as grain moisture content, changes in polymer particle size, residual solvents), etc.

The transferability of model mainly depends on the hardware differences between the instrument systems, and its essence is the replaceability of the spectrometer and its key components (optical systems such as interferometers). Transferability of model directly affects the generalization performance of the analytical method for the user. If the spectrometer of the same manufacturer does not have the transferability of the model, it is very difficult to share abundant model resources. Usually, different instruments have significant system deviations, and the spectra need to be corrected, that is, the calibration transfer, in order to get consistent results. However, there are also manufacturers that can achieve consistency between instruments, and their

calibration models can be directly used for the same type of spectrometer without any modification. That is, the model data is directly copied and transmitted (Calibration transport).

(7) Applicability criterion

Since it is impossible to establish a calibration model that covers all unknown samples, it is particularly important and necessary to establish the applicability criterion for the model. Before performing prediction on unknown samples, the validity and accuracy of results can be guaranteed only if the sample to be tested is within the range covered by the model.

Generally, there are three criteria can be used to ensure the applicability of the model. One is the Mahalanobis distance. If the Mahalanobis distance of tested sample is farer than the maximum distance of the calibration set, it indicates that the concentration of certain components in the tested exceeds the range of that of the calibration set. The second is the spectral residual. If the spectral residual of the tested is over the specified threshold, it means the tested contains components that are not in the calibration set sample. The third is the nearest neighbor distance. If the minimum value of the distance between the tested and all calibration samples (the nearest neighbor distance) is over the specified threshold, it is implied that the tested falls into a place where the distribution of the calibration set is relatively sparse, and the accuracy of the prediction result will be suspected.

Establishment of a robust, reliable, and highly accurate calibration model is the key to the success of the analytical method. The various links involved in the modeling process will affect the accuracy of the analysis. The main influencing factors include:

- (1) Influence of the calibration sample. It includes the representativeness, quantity, range, and distribution; storage; uniformity (such as the particle size, sprout rate, water content, color, and impurities of agro-products); preprocessing (such as crushing, slicing and extraction, etc.); accuracy of the reference data, etc.
- (2) Influence of acquisition condition. It includes spectral range, resolution, acquisition method (such as diffuse reflectance accessory is integrating sphere or fiber optic probe, choice of background material, choice of optical path in transmission method, etc.), temperature, uniformity, and consistency of sampling and loading, etc. Each type of sample (clear liquid, turbid suspension, milled powder, or coarse particles) has its most suitable measuring accessories.
- (3) Influence of chemometrics. It includes spectral pretreatments and their parameters, selection of wavelength variables, calibration methods, and parameters (linear/nonlinear methods, under/over-fitting judgments, and removal of outliers, etc.).
- (4) Influence of instrument (repeatability and long-term stability). It includes effective wavelength range, resolution, S/N ratio, baseline stability, wavelength accuracy and repeatability, absorbance accuracy and repeatability, temperature application range, and resistance to voltage fluctuations, etc.

1.1.2.2 Routine Analysis

After validation, the tested samples can be routinely analyzed. Spectra of the predicted samples should be collected in accordance with the measurement of calibration sample, such as resolution, background, sample and ambient temperature, loading method, and pretreatment method (milling), etc. Applicability of model should be judged before performing routine analysis on the tested sample. If the applicability criterion exceeds the threshold, the model is not suitable for the quantitative analysis of sample.

Models and instruments need to be tested regularly during routine analysis. It is also called the assurance or quality control of analysis. Routine analysis can be conducted in the following ways. ① Actual samples are used for regular verification, such as 2–3 times a week, and compared with the reference method. The absolute deviation should not exceed the range of reproducibility. ② If tested sample can be sealed, about 3–5 representative samples are selected and stored in a sealed enclosure. Routine analysis is done as once every 2 days, and evaluate by the quality control chart. ③ If the composition of tested sample is simple, the accuracy can be verified regularly by preparing standard samples.

If there were inconsistent results in the test, the spectra should be re-collected multiple times, and predictive analysis should be performed to ensure that the spectra are collected correctly, and then the accuracy of the reference data should be checked. If there are still significant differences, the hardware of the spectrometer needs to be fully tested and checked until the cause of the error is located.

1.1.2.3 Features of the Method

Compared with traditional methods, analytical method combining spectroscopy with chemometrics has the following significant advantages.

- (1) It can perform non-destructive analysis of complex mixtures of various forms, usually without sample processing. It directly acquires the spectra without chemical reagents, being as an environmentally friendly analysis technology.
- (2) Analytical speed is pretty fast and the efficiency is satisfactory. A spectrum can be used to determine the multiple composition and property data of the sample within a few seconds.
- (3) The repeatability and reproducibility of analysis are generally better than conventional analytical methods.
- (4) Easy to realize on-site rapid analysis.
- (5) The instrument has few wearing parts and consumables, and low maintenance.
- (6) Most online analyzers can use optical fiber transmission technology, which is suitable for harsh environments.

In fact, any analytical method has its advantages and also limitations, which is very helpful for users to decide whether to adopt or how to know this technology. The limitations are mainly as follows:

- (1) Quantitative and qualitative analyses almost completely depend on the calibration model that often needs to be established separately for different sample types with a lot of resources. Therefore, this method is not suitable for the small amount samples, and not suitable for analysis that can be completed easily by conventional methods.
- (2) Establishment of the calibration model is not done once and for all. In practice, calibration model needs to be continuously expanded and maintained according to the composition change of the tested sample.
- (3) Calibration model requires long-term stability of the spectrometer without the significant change of optical components in the instrument.

The above characteristics make this technology suitable for the following occasions:

- (1) Non-destructive analysis of natural complex system, such as simultaneous analysis of multiple components of petroleum products, and agro-products.
- (2) Fast analysis with highly frequent repeated measurements. Composition of the analyte has relatively strong stability, consistency, and repeatability, such as the laboratory of an oil refinery, food factory, or pharmaceutical factory. Calibration model sharing of branch companies can be realized by networked management.
- (3) Online real-time process analysis of large industrial plants such as petrochemicals and pharmaceuticals, in which the combination of process control and optimization systems can bring considerable economic benefits. Figure 1.14 shows the comparison of traditional offline analysis and modern online process analysis. It can be seen that the online analysis can more accurately reflect the change of the material concentration due to the real-time analysis.

Compared with other analysis, this type of method has the characteristics of the integration of hardware, software and modeling. Its accuracy is closely related to the quality of model established. Thus, the user should have sufficient knowledge in the analytical objects and fields, conventional analytical methods, spectroscopy, chemometric methods, and modeling strategies, to maximize the potential advantages of this technique.

1.1.3 Beginning of Modern Spectroscopy Technology—The Contribution of Karl Norris

The modern spectroscopy technology began with the research and application of near-infrared spectroscopy. Most of its original innovative work was done by a team led by Dr. Karl Norris, an engineer from the United States Department of Agriculture [18].

Near-infrared spectrum is the first non-visible region discovered by the British physicist F. W. Herschel (1739–1822). Until the 1960s, NIRS had not been well applied, mainly because the absorption is very weak, and the spectral bands are

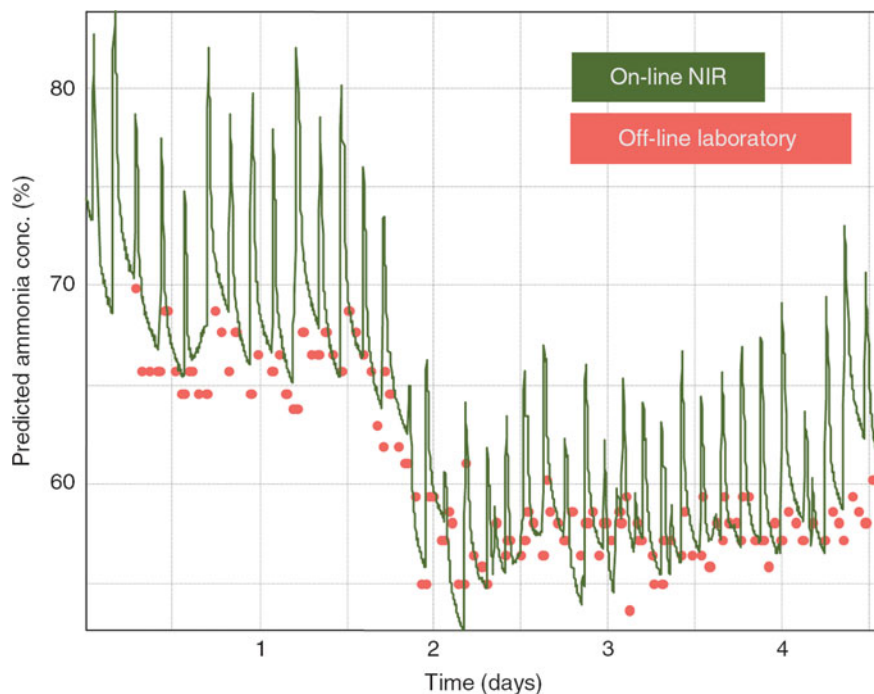


Fig. 1.14 Scheme of comparison between online process analysis and traditional offline analysis results

seriously overlapped. It is difficult to apply it with traditional spectroscopic quantitative (Lambert-Beer law) and qualitative analysis (characteristic absorption of functional groups). It was once called the “Garbage bin of spectroscopy” in spectroscopy. Instead, the epitaxial regions at both ends of the NIRS (ultraviolet-visible and mid-infrared) have been developed rapidly during this period.

In the 1940–1950s, there were also reports on the use of NIRS for quantitative analysis of epoxy compound functionality, polymer and phenolic plastic unsaturation, compound hydroxyl groups, and drug moisture [19–21]. Willis of the British Chemical Industry Company (ICI) used NIRS to characterize the structure of polymers, and to measure the thickness of polymer films [22]. But these researches and applications had been following the traditional mid-infrared spectroscopy and Lambert-Beer law of qualitative and quantitative analysis.

Modern near-infrared spectroscopy technology started from the work of Dr. Karl Norris [23–25], who was an engineer at the USDA Research Center (Batesville, Maryland). In 1949, he used his modified Beckman DU ultraviolet spectrometer to study the freshness of eggs through transmission measurement. He found that the absorption peak at 750 nm was the overtone absorption of the O-H group in water [26–28]. Unfortunately, due to the limitations of technology at that time, the relationship between spectrum and the quality of eggs was not established. An automatic egg

screening equipment was developed only based on the color of the egg shell. This work attracted the attention of the then U.S. President D. D. Eisenhower (Fig. 1.15). Karl Norris also discovered fruits and vegetables have obvious absorption bands at 700–800 nm, which laid the groundwork for Karl Norris' subsequent development of NIR non-destructive fruit quality analyzers (water core disease of apples, etc.) (Fig. 1.16) [29, 30].

Norris really started the study of NIRS in 1960 from the determination of moisture in seeds with his early ideas also based on the Lambert-Beer law. He found the one-variable quadratic polynomial quantitative relationship between the absorbance difference between the two wavelengths (1.94 and 2.08 μm) in the transmission spectrum and the water content, and obtained satisfactory results [31, 32]. The impression of this differentiated spectrum has a deep impact on Norris, and the effects of filter instrument wavelength screening and derivative spectroscopy to eliminate particles are all originated from this concept. Due to the toxicity of the carbon tetrachloride solvent, Norris began to experiment with the reflection method by bringing in the best spectrometer of Cary 14 at that time. But the performance of this instrument did

Fig. 1.15 President Eisenhower visited the automatic egg screening equipment invented by Karl Norris in 1953



Fig. 1.16 Near-infrared internal quality analyzer developed by Karl Norris and Neotec



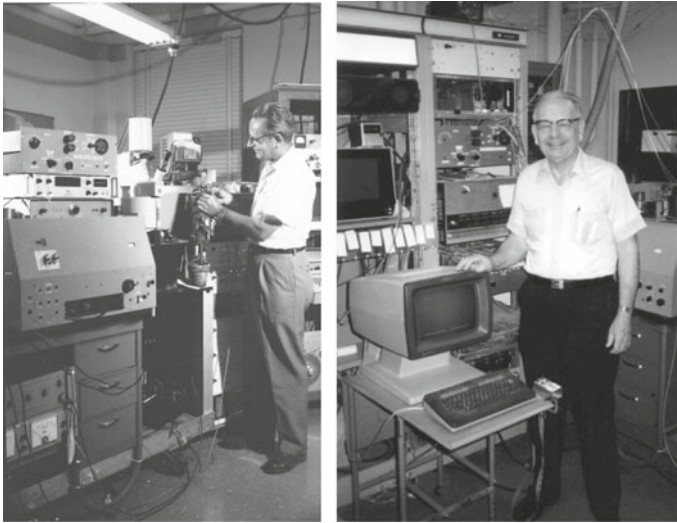


Fig. 1.17 Karl Norris and his modified Cary 14 spectrometer (In 1957 and 1988)

not meet their needs because the scanning speed was slow and there was no suitable reflection accessory. In the following years, with the development of electronic technology, Norris and his colleagues continued to transform it (Fig. 1.17), by updating sample chambers, optical path systems (changing dual optical paths to single optical paths), electronic devices, A/D conversion board, detector and computer, etc. It is right on this spectrometer called “The Norris Machine” that Norris opened the golden door to modern near-infrared spectroscopy technology [33–35].

First of all, Norris creatively replaced the absorbance ($A = \log 1/T$) in traditional spectral analysis with $A = \log 1/R$. This idea obviously did not conform to the Lambert-Beer law, basing on no theoretical basis, and was unanimously opposed by most spectroscopist at the time. But Norris himself is an agricultural engineer other than a spectroscopist, whose research orientation is to solve practical problems. In fact, his results were very positive cause there was indeed a strong correlation between $\log 1/R$ and moisture [36]. With the further research, his team found that the two-wavelength measurement of grain moisture would be interfered by other components in the sample, such as protein in wheat, oil in soybeans, etc. Later, Norris realized that NIRS can also measure the content of these interferences. By Norris’s work, six important wavelengths (1680, 1940, 2100, 2180, 2230, 2310 nm) have been screened out, laying a solid foundation for the later development of commercial filter instruments (Fig. 1.18). Meanwhile, in order to reduce the influence of particle size on the diffuse reflectance spectrum, Norris used the derivative method to process the spectra and proposed the “Karl Norris Derivative” method [37].

The work done by Karl Norris has the remarkable characteristics of modern spectroscopy technology: non-destructive analysis of whole grains, fast analysis speed, simultaneous analysis of multiple parameters based on spectral preprocessing and

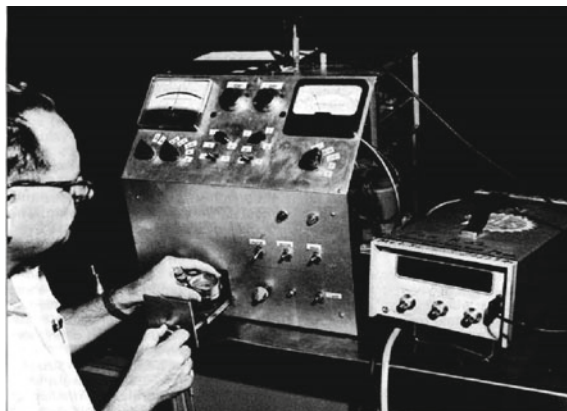


Fig. 1.18 In 1968, Karl Norris operated the first four-filter soy near-infrared analyzer prototype (originally based on the transmission measurement method of mixing crushed soybeans and carbon tetrachloride into a slurry, and later changed to the diffuse reflection measurement method)

multivariate calibration [38–41], etc. It is worth noting that compared with traditional analytical techniques, NIRS has two distinctive features from its inception. (1) It is recommended not to preprocess the samples, and solve the measurement problems of different forms of samples in the form of accessories. (2) It is recommended to bring the instrument to the place of sample instead of bring sample to the instrument (that is, on-site analysis and online analysis). These two characteristics have a profound influence on the development of modern spectroscopic technology.

Inspired by Norris, two companies, Dickey-John and Neotec, in the early 1970s, developed the first commercialized NIRS grain analyzer based on filter technology, which was a great milestone in the development process of NIRS technology [42–45]. Afterward, these instruments selected filters of different wavelengths, increased the number of filters, controlled temperature, and sealed optical system to adapt to harsh environments, according to different applications (such as grass and tobacco) etc.

These instruments had played a very important role in practical applications and greatly promoted the development of NIRS technology. For example, in Canada, Phil Williams used this near-infrared grain analyzer to quickly meet the demand for protein in the wheat export area [46, 47]. Because traders were always willing to pay more for wheat with high protein content, hundreds of such instruments entered large grain elevators and export areas. At the same time, some flour mills, soybean plants and food factories also began to use near-infrared analyzers.

In the late 1970s, grating scan near-infrared spectrometers began to appear with the key technologies all developed by “Norris Machine” as the prototype, such as Neotec Model 6100 and Technicon InfraAlyzer 500, etc. Figure 1.19 shows the NIR instrument companies that have evolved from two original manufacturers, DICKEY-john and Neotec.

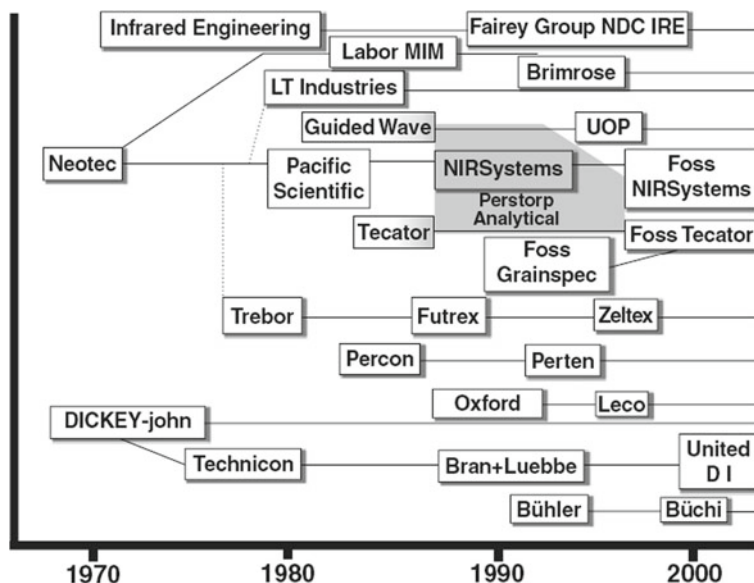


Fig. 1.19 Instrument companies evolved from DICKEY-john and Neotec

In 1975, the Canadian Grain Commission (CGC) designated the near-infrared method as the official method for protein detection. In 1980, the United States Department of Agriculture Federal Grain Inspection Service adopted NIRS method as the official standard method for determining wheat protein. In 1982, the American Association of Cereal Chemists (AACC) officially approved the method (AACC No. 39–00). Till now, Phil Williams estimates that over 90% of wheat world-wide is sold on the basis of protein testing by whole-grain NIRS instruments. After Australia adopted NIRS technology, the yield of rice increased by about 0.6 tons per hectare, the yield of wheat increased by about 1.1 tons, and the protein content of wheat increased by about 1% [48–51].

The work of Karl Norris, especially the “Norris Machine”, has gained wide attention in the agricultural field. Norris unreservedly imparted his research results to each visiting scholar with a selfless, generous, and open spirit of scientists, and conducted cooperation with them [52–58]. Undoubtedly, the laboratory of Karl Norris has become the cradle for cultivating masters of modern near-infrared spectroscopy. During that period, scholars who visited the laboratory of Karl Norris included John Shenk from Pennsylvania, USA, Fred McClure from North Carolina, USA, Phil Williams from Canada, Mutsuo Iwamoto from Japan, Karoly Kaffka from Hungary, and so on. These scholars later became outstanding practitioners and powerful promoters of NIRS technology. John Shenk established the first NIRS forage analysis network in the United States and developed the famous chemometrics software DOSISI and WinISI. After Mutsuo Iwamoto returned to Japan, under his leadership and influence, NIRS technology began to be widely used in Japan. Japan

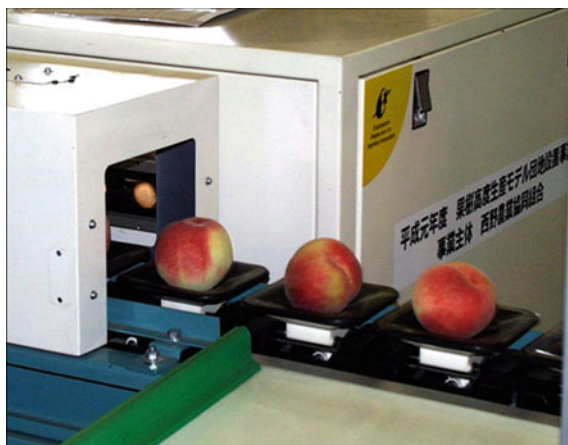


Fig. 1.20 Fruit near-infrared online sorting device developed by Mitsui Company

developed an automatic sorting device for fruits based on NIRS in the late 1980s. In the 1990s, Karl Norris visited the fruit NIR online sorting device developed by Mitsui in Shizuoka, Japan (Fig. 1.20). He said that “My dream has come true in Japan”. It can be concluded that Karl Norris’s contribution in cultivating international near-infrared masters is undoubtedly huge.

Karl Norris did a lot of work to promote the development of NIRS technology, and to obtain the support of some scientists at the time [59]. During this period, spectroscopists who began to support NIRS included Tomas Hirschfeld, Peter Griffiths and Bill Fateley, etc. The participation of these spectroscopists was very important in the formation of the theoretical system of NIRS technology. In 1984, under the advocacy of Tomas Hirschfeld, the American Society for Testing and Materials (ASTM) established a NIRS working group (E13.03.03) to study the standard method of NIRS technology.

In 1974, the Swedish chemist Wold and Professor Kowalski from the University of Washington created the discipline of chemometrics. Chemometrics is a branch of chemistry formed by combining mathematics, statistics, computer science, and chemistry. Its foundation is the rapid development of computer technology and the modernization of analytical instruments [60]. Unfortunately, the early stage of chemometrics was not combined with the application of near-infrared spectroscopy in agriculture. It was Karl Norris’ unremitting efforts that made chemometrics scientists gradually pay attention to this technology, which contributed to the rise of near-infrared spectroscopy technology [61, 62]. Some chemometric methods based on principal component analysis were beginning to be adopted by researchers, such as principal component regression and partial least squares, etc., which significantly improved the accuracy and reliability of the results of NIRS. In the mid-1990s, artificial neural

network methods had appeared in chemometrics commercial software for NIRS analysis. Since then, near-infrared spectroscopy and chemometrics have been developing and improving in interdependence, interaction, and mutual promotion [63, 64].

References

1. Yu LQ. Introduction to chemometrics. Changsha: Hunan Education Press; 1991.
2. Liang YZ, Yu LQ. Handbook of analytical chemistry (chapter 10-chemometrics). Beijing: Chemical Industry Press; 2000.
3. Brereton RG. Chemometrics: data driven extraction for science, 2nd ed. USA Roseland: Wiley; 2018.
4. Bakeev KA. Process analytical technology: spectroscopic tools and implementation strategies for the chemical and pharmaceutical industries. Oxford UK: Blackwell Publishing; 2005.
5. McClure FW. Near-infrared spectroscopy the giant is running strong. *Anal Chem.* 1994;66(1):42a–53a.
6. Williams P, Antoniszyn J, Manley M. Near infrared technology: getting the best out of light. Stellenbosch: Sun Press; 2019.
7. Chu XL, Li SH, Zhang T. New development of modern process analysis technology. Beijing: Chemical Industry Press; 2021.
8. Lavine BK, Brown SD, Booksh KS. 40 Years of chemometrics—from Bruce Kowalski to the future. Oxford: Oxford University Press; 2015.
9. Brown SD, Tauler R, Walczak B. Comprehensive chemometrics, 2nd ed. Elsevier; 2020.
10. Otto M. Chemometrics: statistics and computer application in analytical chemistry, 3rd ed. Verlag: Wiley; 2017.
11. Saeyns W, Trong NND, Beers VR, et al. Multivariate calibration of spectroscopic sensors for postharvest quality evaluation: a review. *Postharvest Biol Technol.* 2019;158:110981.
12. Mark H, Workman J. Statistics in spectroscopy, 2nd ed. Academic Press; 2003.
13. Lindon J. Encyclopedia of spectroscopy and spectrometry, 2nd ed. Academic Press; 2020.
14. Næs T, Isaksson T, Fearn T, et al. A user-friendly guide to multivariate calibration and classification. London: NIR Publications; 2002.
15. Liu JC, Osadchy M, Ashton L, et al. Deep convolutional neural networks for Raman spectrum recognition: a unified solution. *Analyst.* 2017;142:4067–74.
16. Li CL, Liu DW, Liu B, et al. Chang'E-4 initial spectroscopic identification of lunar far-side mantle-derived materials. *Nature.* 2019;569:378–82.
17. Steiner G, Bartels T, Stelling A, et al. Bird sexing by infrared spectroscopy. *Spectrosc Eur.* 2011;23(1):16–9.
18. Chen YH, Lu F, Yin LH. Local straight-line screening method: Research and development. *Scientia Sinica (Chimica).* 2010; 40(8):1142–48.
19. Meeker RL, Critchfield FE, Bishop ET. Water Determination by near infrared spectrophotometry. *Anal Chem.* 1962;34(11):1510–1.
20. O'Connor RT. Near-infrared absorption spectroscopy—a new tool for lipid analysis. *J Am Oil Chem Soc.* 1961;38(11):641–8.
21. Patterson WA. Non-dispersive types of infrared analyzers for process control. *Appl Spectrosc.* 1952;6(5):17–23.
22. Miller R. Professor Harry Willis and the history of NIR spectroscopy. *NIR News.* 1991;2(4):12–3.
23. Burns DA, Ciurczak EW. Handbook of near-infrared analysis. 3rd ed. New York: Marcel Dekker; 2007.
24. Davies T. Happy 90th birthday to Karl Norris. *Father NIR Technol, NIR News.* 2011;22(4):3–16.
25. Davies T. The history of near infrared spectroscopic analysis: past, present and future—"From Sleeping Technique to the Morning Star of Spectroscopy". *Analysis.* 1998;26(4):17–9.

26. Norris KH. Early history of near infrared for agricultural applications. *NIR News*. 1992;3(1):12–3.
27. Norris KH. History of NIR. *J Near Infrared Spectrosc*. 1996;4(1):31–7.
28. Davies AMC. The history of near infrared spectroscopy 1. *First NIR Spectr, NIR News*. 1991;2(2):12.
29. Rosenthal RD, Webster DR. On-line system sorts fruit on basis of internal quality. *Food Technol*. 1973; 27(1):52–6, 60.
30. Kawano S. Past, present and future near infrared spectroscopy applications for fruit and vegetables. *NIR News*. 2016;27(1):7–9.
31. Hart JR, Golumbic C, Norris KH. Determination of moisture content of seeds by near-infrared spectrophotometry of their methanol extracts. *Cereal Chem*. 1962;39(2):94–9.
32. Whetsel KB. Near-infrared spectrophotometry. *Appl Spectrosc Rev*. 1968;2(1):1–67.
33. Barton I FE. Progress in near infrared spectroscopy: the people, the instrumentation, the applications. *NIR News*. 2003;14(2):10–8.
34. Reeves Iii J, Delwiche SR. Near infrared research at the Beltsville agricultural research center (part 1): instrumentation and sensing laboratory. *NIR News*. 2005;16(6):9–12.
35. Reeves IJ. Near infrared research at the Beltsville agricultural research center (part 2). *NIR News*. 2005;16(8):12–3.
36. Norris KH. When diffuse reflectance became the choice for compositional analysis. *NIR News*. 1993;4(5):10–1.
37. Hopkins DW. What is a norris derivative? *NIR News*. 2001;12(3):3–5.
38. Workman JJ. A review of process near infrared spectroscopy: 1980–1994. *J Near Infrared Spectrosc*. 1993;1(4):221–45.
39. Wetzel DL. Near-infrared reflectance analysis sleeper among spectroscopic techniques. *Anal Chem*. 1983;55(12):1165a–a1176.
40. Williams P. Twenty-five years of near infrared technology-what were the milestones? *NIR News*. 1997;8(1):5–6.
41. McClure WF. Breakthroughs in NIR spectroscopy: celebrating the milestones to a viable analytical technology. *NIR News*. 2006;17(2):10–1.
42. Barton IE. Near infrared equipment through the ages and into the future. *NIR News*. 2016;27(1):41–4.
43. Davies T. NIR instrumentation companies: the story so far. *NIR News*. 1999;10(6):14–5.
44. Whetsel KB. The first fifty years of near-infrared spectroscopy in America. *NIR News*. 1991;2(3):4–5.
45. Whetsel KB. American developments in near infrared spectroscopy (1952–70). *NIR News*. 1991;2(5):12–3.
46. Williams P. Near infrared technology in Canada. *NIR News*. 1995;6(4):12–3.
47. Williams P. The Phil William’s episode. *NIR News*. 1992;3(2):3–4.
48. Batten G. An appreciation of the contribution of NIR to agriculture. *J Near Infrared Spectrosc*. 1998;6(1):105–14.
49. Bosco GL, James I. Waters symposium 2009 on near-infrared spectroscopy. *Trends Anal Chem*. 2010;29(3):197–208.
50. Paula C, Montesb JM, Williams P. Near infrared spectroscopy on agricultural harvesters: the background to commercial developments. *NIR News*. 2008;19(8):8–11.
51. Battena GD, Blakeneyb AB, Ciavarellaca S, et al. NIR helps raise crop yields and grain quality. *NIR News*. 2000;11(6):7–9.
52. Kaffka KJ. Near infrared technology in hungary and the influence of Karl H. Norris on our success. *J Infrared Spectrosc*. 1996; 4(1):63–7.
53. Iwamoto M, Kawano S, Ozaki Y. An overview of research and development of near infrared spectroscopy in Japan. *J Near Infrared Spectrosc*. 1995;3(4):179–89.
54. Miskelly D, Ronalds J, Miskelly DM, et al. Twenty-one years of NIR in Australia: a retrospective account with emphasis on cereals. *NIR News*. 1994;5(2):10–2.
55. Osborne B. Twenty years of NIR research at Chorleywood 1974–1993. *NIR News*. 1993;4(2):10–1.

56. Hildrum KI, Isaksson T. Research on near infrared spectroscopy at Matforsk 1979–1992. *NIR News*. 1992;3(3):14.
57. Davies T. Karl's London marathon. *NIR News*. 2002;13(3):3.
58. Gonczy JL. Developments in Hungary 1970–1990. *NIR News*. 1993;4(3):3–4.
59. Donaldson PEK. In Herschel's footsteps. *NIR News*. 2000;11(3):7–8.
60. Geladi P, Esbensen K. The start and early history of chemometrics: selected interviews. *J Chemom*. 1990;4:337–54.
61. Ritchie GE. Investigating NIR transmittance measurements through the use of the norris regression (Nr) algorithm: part 1: how do we come to "Norris Regression"? *NIR News*. 2002;13(1):4–6.
62. Norris KH, Williams PC. optimization of mathematical treatments of raw near-infrared signal in the measurement of protein in hard red spring wheat. I. Influence of particle size. *Cereal Chem*. 1984;61(2):158–65.
63. Geladi P, Dabakk E. An overview of chemometrics applications in near infrared spectrometry. *J Near Infrared Spectrosc*. 1995;3(3):119–32.
64. Fearn T. Chemometrics for NIR spectroscopy: past present and future. *NIR News*. 2001;12(2):10–2.

Chapter 2

Modern Spectral Analysis Techniques



2.1 Introduction

Light is an electromagnetic wave that moves in two orthogonal planes of electric and magnetism. The distance between two crests or troughs is seen as the wavelength, denoted by λ . Electromagnetic radiation is a stream of photons propagating through space at high speed, which has the property of both wave and particle. According to quantum theory, the emission or absorption of radiant energy is not continuous, but quantized. The smallest unit of this energy is “photon”, and the relationship between the energy E of each photon and its frequency ν and wavelength λ is shown in Eq. (2.1):

$$E = h\nu = hc/\lambda = hc\bar{\nu} \quad (2.1)$$

where E is the energy of photon and the unit is electron volt (eV) or joule (J), $1 \text{ eV} = 1.602 \times 10^{-19} \text{ J}$; h is Planck's constant, $h = 6.626 \times 10^{-34} \text{ J} \cdot \text{s}$; ν is the frequency, and the unit is Hertz (Hz) or second^{-1} (s^{-1}), representing the frequency of electromagnetic wave vibration per second; c is the speed of light, $c = 2.998 \times 10^{10} \text{ cm} \cdot \text{s}^{-1}$; λ is the wavelength, and the unit is meter (m), centimeter (cm), micron (μm) or nanometer (nm). $1 \text{ m} = 10^2 \text{ cm} = 10^6 \mu\text{m} = 10^9 \text{ nm}$; $\bar{\nu}$ is the wavenumber, and the unit is centimeter^{-1} (cm^{-1}), representing the number of vibration in the unit distance (cm) of the electromagnetic wave, and wavenumber and wavelength is reciprocal.

All kinds of electromagnetic radiation sorted based on the wavelength or frequency of the size of the order is called the electromagnetic spectrum. Table 2.1 listed the parameters related to electromagnetic waves used for spectral analysis. γ -rays have the shortest wavelength and the highest energy. The radio wave region has the longest wavelength and the lowest energy. If the wavelength or frequency is known, the energy required to produce different types of transitions in respective electromagnetic region can be calculated and vice versa. For example, the energy

Table 2.1 Relative parameters of electromagnetic wave

E/eV	V/Hz	λ	Electromagnetic wave	Transition type
$>2.5 \times 10^5$	$>6.0 \times 10^{19}$	<0.005 nm	γ ray region	Nuclear level
$2.5 \times 10^5 \sim 1.2 \times 10^2$	$6.0 \times 10^{19} \sim 3.0 \times 10^{16}$	$0.005 \sim 10$ nm	X ray region	K, L electronic energy levels
$1.2 \times 10^2 \sim 6.2$	$3.0 \times 10^{16} \sim 1.5 \times 10^{15}$	$10 \sim 200$ nm	vacuum ultraviolet region	
$6.2 \sim 3.1$	$1.5 \times 10^{15} \sim 7.5 \times 10^{14}$	$200 \sim 400$ nm	near ultraviolet region	Outer electron energy level
$3.1 \sim 1.6$	$7.5 \times 10^{14} \sim 3.8 \times 10^{14}$	$400 \sim 800$ nm	visible light region	
$1.6 \sim 0.50$	$3.8 \times 10^{14} \sim 1.2 \times 10^{14}$	$0.8 \sim 2.5$ μ m	near infrared region	Molecular vibrational energy level
$0.50 \sim 2.5 \times 10^{-2}$	$1.2 \times 10^{14} \sim 6.0 \times 10^{12}$	$2.5 \sim 50$ μ m	mid infrared region	
$2.5 \times 10^{-2} \sim 1.2 \times 10^{-3}$	$6.0 \times 10^{12} \sim 3.0 \times 10^{11}$	$50 \sim 1000$ μ m	far infrared region	Molecular rotational energy level
$1.2 \times 10^{-3} \sim 4.1 \times 10^{-6}$	$3.0 \times 10^{11} \sim 1.0 \times 10^9$	$1 \sim 300$ mm	microwave region	
$<4.1 \times 10^{-6}$	$<1.0 \times 10^9$	>300 mm	radio wave region	Spins of electrons and nuclei

required to motivate the valence electrons of a molecule or atom is 1~20 eV, and the corresponding wavelength of the electromagnetic wave within this energy range can be calculated as 1240~62 nm in Eqs. (2.2) and (2.3).

$$\lambda = \frac{hc}{E} = \frac{6.626 \times 10^{-34} \times 3.0 \times 10^{10}}{1 \times 1.602 \times 10^{-19}} \times 10^7 \text{ nm} = 1240 \text{ nm} \quad (2.2)$$

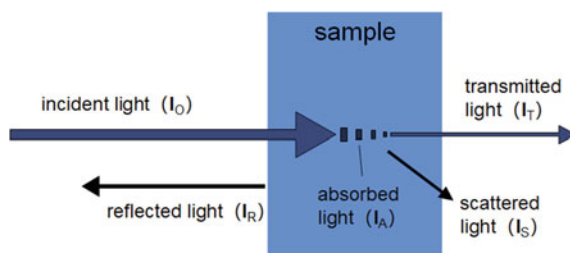
$$\lambda = \frac{hc}{E} = \frac{6.626 \times 10^{-34} \times 3.0 \times 10^{10}}{20 \times 1.602 \times 10^{-19}} \times 10^7 \text{ nm} = 62 \text{ nm} \quad (2.3)$$

For the electromagnetic spectrum with shorter wavelength (less than 10 nm) and greater energy (more than 10^2 eV), it is called the energy spectrum which has more obvious particle properties, and thus the analysis method was called the energy spectral analysis. The electromagnetic spectrum whose wavelength is greater than 1 mm and energy is less than 10^{-3} eV (such as microwave and radio waves) has obvious fluctuation, called the wave spectrum. The analysis method based on this spectrum is called the spectral analysis method. The electromagnetic spectrum whose wavelength and energy is between energy spectrum and wave spectrum is usually obtained by optical instruments, called optical spectrum. The analysis method built is thus called optical spectral analysis method, which is also called spectral analysis.

Spectral analysis is an analytical method measuring the wavelength and intensity of the emission, absorption, or scattering radiation generated by the transition between energy levels of the material internal quantum when the matter interacted with radiation energy. Spectroscopy can be divided into atomic spectroscopy and molecular spectroscopy.

Atomic spectroscopy is produced by the change of the outer or inner electron energy level of atoms, which has no superposition of molecular vibration and rotation energy level transition, and emits or absorbs some discontinuous radiation frequency (or wavelength). It is shown by line-spectrum methods, such as atomic emission spectrometry, atomic absorption spectrometry, atomic fluorescence spectrometry, and X-ray fluorescence spectrometry, etc. Molecular spectroscopy is produced by

Fig. 2.1 Schematic diagram of the interaction of electromagnetic radiation with matter



the change of electron energy level, vibrational energy level, and rotational energy level in a molecule, which is represented as band spectrum. These analytical methods include UV-Vis spectrophotometry, near-infrared (NIR) spectroscopy, infrared (IR) spectroscopy, molecular fluorescence spectroscopy, and molecular phosphorescence spectroscopy, etc.

Usually, the light emitted by matter contains a variety of frequency components, which is known as compound light. In spectral analysis, the light containing only one frequency component (i.e., monochromatic light) is often obtained by a certain method as an analytical method. In fact, the monochromatic light obtained by common analytical methods often contains more than one frequency component. The monochromaticity of monochromatic light is usually expressed by the width (or half width) of the spectral line. The narrower the width of the spectral line, the narrower the range of frequencies (or wavelengths) that the spectral line contains, and the better the monochromaticity of the light.

The optical analysis method commonly used in analytical chemistry is the spectroscopic method, which is an instrumental analysis method to extract useful information from the spectra of substances and to further determine the composition, content and structure of substances. As shown in Fig. 2.1, electromagnetic radiation interacts with matter to produce three types of spectra, including emission, absorption, and scattering.

(1) Emission spectroscopy

Materials obtain energy through excitation processes such as electroinduced excitation, thermally induced excitation or photoinduced excitation, and turn into excited atoms or molecules. When they make transition from excited state to low energy state or ground state, emission spectrum is generated and excess energy is emitted in the form of light as Eq. (2.4) shows



The method of qualitative and quantitative analysis by measuring the wavelength and intensity of the emission spectrum of a substance is called emission spectroscopy. According to the spectral region where the emission spectrum is located and the different excitation methods, emission spectrometry is divided into γ -ray spectrometry, X-ray fluorescence analysis, atomic emission spectrometry,

atomic fluorescence spectrometry, molecular fluorescence spectrometry, molecular phosphorescence spectrometry, chemiluminescence, etc.

(2) Absorption spectroscopy

When the electromagnetic radiation energy absorbed by the material and the energy of atom or molecule of the material meets the relationship of $\Delta E = h\nu$ required for the transition between the two or more energy levels of the nucleus, the absorption spectrum will be generated as Eq. (2.5) shows



Absorption spectroscopy includes Mossbauer spectroscopy, atomic absorption spectroscopy, UV-Vis spectroscopy, near-infrared spectroscopy, infrared spectroscopy, etc.

(3) Raman scattering spectroscopy

Monochromatic light with frequency of ν_0 shines on the transparent material, and the material molecules will scatter. If the scattering is the energy exchange between the photon and the material molecule, that is, not only the motion direction of the photon changes, but also its energy changes, then it is called Raman scattering. The frequency of this scattered light is different from that of the incident light, which is called the Raman shift. The magnitude of Raman shift is related to the vibrational and rotational energy levels of molecules, and the method of studying the structure and composition of substances using Raman shift is called Raman spectroscopy [1].

The optical analysis method based on the above spectrum is called spectral analysis method. Spectral analysis technology combining spectrum with chemometrics is called modern spectral analysis technology. Handheld or portable field rapid analysis, online analysis of industrial processes, spectral imaging analysis, and other practical applications are the most attractive parts of modern spectral analysis technology [2–5], and have also become the core content of modern process analysis technology (Fig. 2.2) [6–9]. These aspects are closely related to chemometrics method [10].

2.2 Near-Infrared Spectroscopy

NIR light is an electromagnetic wave between UV-Vis (UV-Vis) and mid-infrared light (MIR). Its wavelength range is 700–2500 nm (14,286–4000 cm^{-1}), and it could be further divided into two regions: short-wave (700–1100 nm) and long-wave (1100–2500 nm) NIR spectra. The short-wave region is also called the Herschel region in honor of Herschel's discovery of the infrared region (actually the NIR region) in 1800. Instruments extending from the UV-Vis spectrum often take the wavelength (nm or μm) as the horizontal coordinate unit, while instruments extending from the infrared spectrum, especially Fourier-type instruments, take the wavenumber (cm^{-1}) as the horizontal coordinate unit.

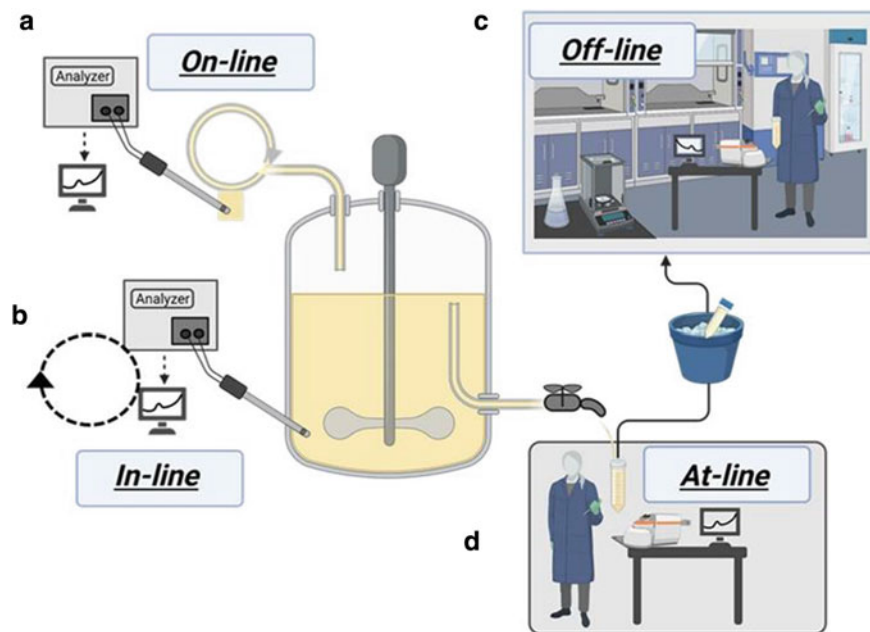


Fig. 2.2 Schematic diagram of various methods of implementing process analysis technology tools into bioprocess monitoring [9]

NIR spectra are mainly generated when molecular vibration transitions from ground state to higher energy level due to the non-resonance of molecular vibration, which mainly reflects the frequency doubling and frequency absorption of hydrogen containing groups X-H (such as C-H, N-H, O-H, etc.) vibration. NIR absorption wavelength and intensity of different groups (such as methyl, methylene, benzene ring, etc.) or the same group in different chemical environments are obviously different. NIR spectra have rich structure and composition information, which is very suitable for the measurement of physicochemical parameters of hydrogen-containing organic substances such as agricultural products, petrochemical products, and drugs.

Another feature of NIR spectra is the weak absorption strength. Compared with infrared spectrum (fundamental frequency), the probability of producing NIR spectrum is 1–3 orders of magnitude lower. On the one hand, NIR instruments are required to have high signal-to-noise ratio; on the other hand, it is very convenient for measurement. For example, it can measure the NIR spectrum of the liquid with the millimeter-scale colorimetric dish. Because the absorption coefficient of the material in the NIR region is small, its detection limit is usually 100 ppm, which is not suitable for trace analysis. In order to overcome its limitations, sample pretreatment (such as solid phase microextraction enrichment methods) can be used to improve the sensitivity. However in this case, NIR spectroscopy as a detection technology may not be the best choice.

NIR spectroscopy analysis technology also has certain limitations. NIR spectroscopy analysis is almost an indirect method of building models based on chemometrics. To establish a robust and reliable model, it requires a certain amount of labor, financial resources, and time investment. It is economical and fast for regular quality control, but not suitable for occasional analysis.

2.2.1 Micro Near-Infrared Spectral Analysis Technology

Because the NIR spectral region is between UV-Vis spectrum and MIR spectrum, the spectrometer has many ways of splitting, which brings great convenience to the miniaturization and microminiaturization of NIR spectral instruments. It takes less than 10 years for NIR spectrometers to evolve from benchtop, portable, handheld, to pocket-sized and miniature [11]. In recent years, some companies have been working on the development of miniature NIR spectrometer chips. For example, some companies have developed miniature NIR spectrometer chips with external size of 18 mm × 18 mm, thickness of 4 mm, weight of less than 10 g, and wavelength range of 1100–2500 nm, which is small enough to be integrated into smart phones and wearable devices. And future spectrometers will get smaller and smaller. Yang et al. used a special nanowire with a gradient band gap to replace the spectroscopic and detecting elements in the traditional spectrometer, and fabricated a light detector array on the nanowire for reducing the size of the traditional optical device to the nanoscale [12].

In recent years, research on the application of portable and miniature spectroscopic instruments in people's daily life has begun [13–15], and many concept products have appeared on the market, such as intelligent washing machine, intelligent red wine identification scanner, intelligent dehydration monitoring bracelet, clothing material identification instrument, and so on. Samsung Electronics filed a patent and exhibits a smartphone on its website that features a NIR spectrometer. The device's rear camera system provides a series of light sources at the top of the camera. When it shines light on an object, the camera receives reflected signals to generate spectral data. The smartphone is expected to measure the freshness and taste of fresh products as well as their nutritional value, such as fat, protein and carbohydrate content. It could also be used to measure the water-oil balance of skin, the quantity of sugar in a drink and, hopefully, even direct diagnostics in the medical area.

Miniature NIR spectrometer chips are increasingly integrated with robots and unmanned aerial vehicles. For example, there are already commercially available plastic sorting devices that combine robotic arms with spectrometers to quickly identify the types of waste plastic. The combination of NIR micro instruments and robots can even achieve a completely unmanned intelligent analysis laboratory: the process from sampling to data reporting is completely operated by the robot, and can work all day long, significantly improving the efficiency of analysis.

2.2.2 *Online Near-Infrared Spectral Analysis Technology*

The NIR light is longer than the UV light and shorter than the MIR light. The optical material used is quartz or glass, and the price of the instrument and measuring accessories is lower. NIR light can also be transmitted via a relatively inexpensive low-hydroxyl quartz fiber, which is suitable for remote online analysis of toxic materials or harsh environments. It also makes the design of spectrometers and measuring accessories more flexible and smaller. For example, there are a wide variety of commercially available optical fiber probes that can determine a wide variety of forms of samples. Using multi-channel optical switching technology could achieve that a NIR spectrometer can be used to measure multi-channel materials (3–15 channels) with the advantages of fast analysis speed and high measurement efficiency [16].

NIR spectral analysis technology has been applied in many fields, such as agriculture, petrochemical, pharmaceutical, food, etc., in the way of industrial chain. It can quickly and efficiently determine the chemical composition and physicochemical properties of samples. In the past decade, with the rise of process analysis technology in pharmaceutical and other fields, the application of NIR spectroscopy technology, especially online analysis, has been significantly improved.

At present, the process industry is in the transition period from the traditional production mode to the precise digital and intelligent modern production mode. “Self-perception” of information depth, “self-decision” of intelligent optimization, and “self-execution” of precise control are the three key characteristics of an intelligent factory, among which “self-perception” of information depth is the foundation of an intelligent refinery. The analysis data of molecular composition and physical properties of raw materials, intermediate materials and products is an important part of information perception. The modern process analysis technology with NIR spectroscopy as one of the core parts provides a very effective means for chemical information perception.

In petrochemical enterprises, taking gasoline pipeline automatic blending technology as an example, online NIR spectroscopy analyzer has become the standard equipment for this technology at present [17]. After more than 10 years of accumulation, China has established a relatively perfect gasoline NIR spectral database, which can predict a number of key physical properties (research octane number, knock resistance index, olefin, aromatic hydrocarbon, benzene, MTBE content, vapor pressure, etc.) of nearly ten components of gasoline and finished gasoline within 10 min. The blending optimization control system makes use of the blending effect between various gasoline components to calculate the relative proportion of blending components, namely, blending formula, in real time, to ensure that the blended gasoline products meet the quality specifications, and reduce the blending cost and excess quality to a minimum. This technology can bring economic benefits with tens of millions RMB to oil refining enterprises every year.

In feed production enterprises, with the increasingly fierce market competition, low cost of raw materials input, stable product quality, low processing consumption have become the key to stable survival in the market. Using NIR spectral analysis

technology can be real-time online detection of the quality parameters of raw materials, process products and finished products (such as moisture, protein, crude fiber, oil content, ash content, color, etc.). By optimizing control system, fine closed-loop adjustment during the product production process could be adjusted based on real-time quality of the product and objective product, which is able to ensure the quality of finished feed stability and realize the product yield and quality optimization. In addition, by virtue of the characteristics of scale production, it could bring more economic benefits for enterprises.

In the field of food industry, during the process of wheat milling, online NIR spectroscopy analyzer can be real-time determination of ash content of flour. By timely adjustment of the milling process, in the premise of ensuring the quality of flour, a higher powder yield could be got as far as possible. When mixing powder, user's requirements could be met by blending conform to the requirements of the quality of high value-added special powder according to the result of rapid analysis of NIR spectra. It can ensure no unqualified products or quality (protein) surplus phenomenon and make flour product quality long-term stability. Moreover, combined with the feedback control system, the fluctuation of protein content in flour (standard deviation) can be reduced to 0.1% by adjusting the amount of gluten. In some large meat production plants, online NIR spectroscopy is used to accurately determine the content of the main components in raw meat, which made operators can adjust the production process in time, optimize the ratio of raw material (such as the ratio of fat and lean meat), reduce the production cost, and increase the profit of enterprises. In dairy production enterprises, online NIR has been used to monitor the humidity and granularity of milk powder in the atomization dryer, and then optimize the drying process, such as temperature, feed speed, and airflow speed.

As shown in Fig. 2.3, in the process of traditional Chinese medicine extraction and production, online NIR spectroscopy can detect the changes of target components in the extraction in real time, and then ensure the extraction time and extraction end point (Fig. 2.4) [18]. In the purification process, the online NIR spectroscopy can detect the concentration change of the target component in the effluent in real time, to control the switch between the mobile phase and the eluent and determine the termination point of the elution process. This can collect the target component in the largest amount and reduce the number of impurities in the product. It can not only ensure product quality but also can avoid energy waste and reduce production costs. The concentration process can be controlled by detecting the concentration of water (solvent) or target component, and the end point of the concentration can also be determined instantly.

The application of online NIR spectroscopy technology in the fields of food, pharmaceutical, and chemical industry has just begun in China, which is in line with the general trend of fine management and intelligent processing, and will bring changes to the process industry [19, 20]. For a long time to come, the stable and hopeful basis on the application of online NIR spectroscopy in the process industry will not change. In addition, in the field of online screening of waste plastics, textiles and fruit products, the application of online NIR spectroscopy will be more and more extensive.

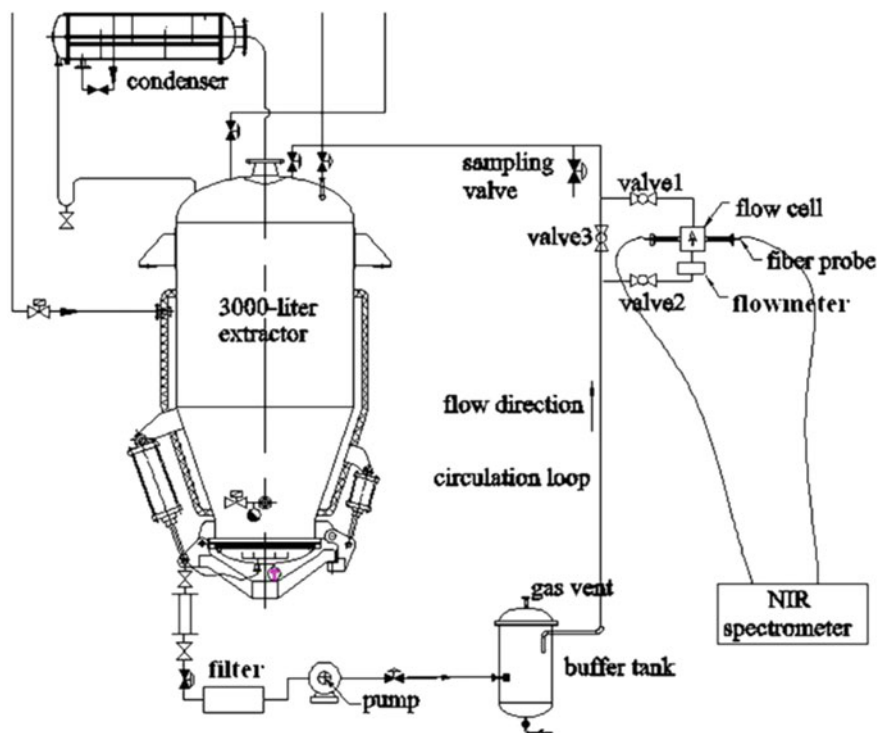


Fig. 2.3 Schematic diagram of online near-infrared spectroscopy for monitoring the extraction process of traditional Chinese medicine [18]

The implementation of online NIR spectroscopy technology is a multidisciplinary system engineering [21, 22], which requires the cooperation of multiple departments and a professional team implementing the subsequent operation and maintenance. In terms of the popularization of this technology, local customized design, manufacturing, implementation, operation, and maintenance have strong advantages.

2.2.3 Standard Methods for Near-Infrared Spectroscopy

NIR spectroscopy has achieved fruitful results in practical application, which is recognized and accepted by more and more applied enterprises. It plays an important role in industrial and agricultural production process as well as commerce. Up to now, above 100 standard methods of NIR spectroscopy have been promulgated all over the world, which will accelerate the popularization of NIR spectral analysis technology to a certain extent. The relevant NIR standard methods from international organizations and some countries are as follows:

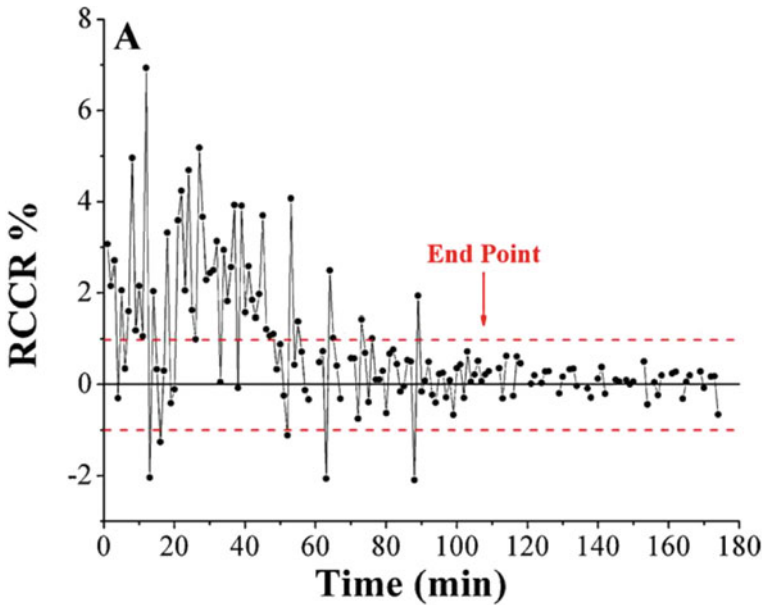


Fig. 2.4 The end point of extraction process was determined based on relative concentration changing rate (RCCR) [18]

1. ASTM E1655 Standard Practices for Infrared Multivariate Quantitative Analysis
2. ASTM E1790 Standard Practice for Near-Infrared Qualitative Analysis
3. ASTM D6122 Standard Practice for Validation of the Performance of Multivariate Online, At-Line, and Laboratory Infrared Spectrophotometer-Based Analyzer Systems
4. ASTM D3764 Practice for Validation of the Performance of Process Stream Analyzer Systems
5. ASTM D6342 Standard Practice for Polyurethane Raw Materials Determining Hydroxyl Number of Polyols by NIR Spectroscopy
6. ASTM D5845 Standard Test Method for Determination of MTBE, ETBE, TAME, DIPE, Methanol, Ethanol, and tert-Butanol in Gasoline by Infrared Spectroscopy
7. ASTM D6277 Standard Test Method for Determination of Benzene in Spark-Ignition Engine Fuels Using Mid-Infrared Spectroscopy
8. ASTM D6299 Practice for Applying Statistical Quality Assurance and Control Charting Techniques to Evaluate Analytical Measurement System Performance
9. ASTM D 7371 Determination of Biodiesel (Fatty Acid Methyl Esters) Content in Diesel Fuel Oil Using Mid-Infrared Spectroscopy (FTIR-ATR-PLS Method)
10. ASTM E2617 Standard Practice for Validation of Empirically Derived Multivariate Calibrations

11. ASTM E2891 Standard Guide for Multivariate Data Analysis in Pharmaceutical Development and Manufacturing Applications
12. ASTM D8321 Standard Practice for Development and Validation of Multivariate Analyses for Use in Predicting Properties of Petroleum Products, Liquid Fuels, and Lubricants based on Spectroscopic Measurements
13. ASTM E2898 Standard Guide for Risk-Based Validation of Analytical Methods for PAT Applications
14. ASTM E2056 Standard Practice for Qualifying Spectrometers and Spectrophotometers for Use in Multivariate Analyses, Calibrated Using Surrogate Mixtures
15. ISO 15063 Plastics-Polyols for use in the production of polyurethanes determination of hydroxyl number by NIR spectroscopy
16. ISO 21543 Milk products. Guidelines for the application of near-infrared spectrometry
17. ISO 12099 Animal feeding stuffs, cereals, and milled cereal products. Guidelines for the application of near-infrared spectrometry
18. ISO 17184-2014 Soil quality—Determination of carbon and nitrogen by near-infrared spectrometry (NIRS)
19. AACC 39-00 Near-Infrared Methods: Guidelines for Model Development and Maintenance
20. AACC 39-10 Near-infrared reflectance method for protein determination
21. AACC 39-11 Near-infrared reflectance method for protein—wheat flour
22. AACC 39-20 Near-infrared reflectance method for protein and oil determination—soybeans
23. AACC 39-21 Near-infrared method for whole-grain analysis
24. AACC 39-25 Near-infrared method for protein content in whole-grain wheat
25. AACC 39-70 Wheat hardness as determined by near-infrared reflectance
26. AACC 08-21 Prediction of Ash Content in Wheat Flour—Near-Infrared Method
27. AOAC 2007.04 Fat, Moisture, and Protein in Meat and Meat Products
28. AOAC 989.03 Fiber (acid detergent) and protein (crude) in forages: Near-infrared reflectance spectroscopic method
29. AOAC 991.01 Moisture in forage, near-infrared reflectance spectroscopy
30. AOAC 997.06. Protein (crude) in wheat. Whole grain analysis, Near-infrared spectroscopic method.
31. ICC 159 Determination of Protein by Near-Infrared Reflectance (NIR) Spectroscopy
32. ICC 202 Procedure for Near-Infrared (NIR) Reflectance Analysis of Ground Wheat and Milled Wheat Products
33. RACI 11.01 Determination of protein and moisture in whole wheat and barley by NIR
34. USP 856 Near-Infrared Spectroscopy
35. USP 1856 Near-Infrared Spectroscopy—Theory and Practice
36. USP 1039 Chemometrics
37. EP 2.2.40 Near-Infrared Spectroscopy

38. PSAG Guidelines for the development and validation of near-infrared (NIR) spectroscopy methods
39. CPMP&CVMP Note for guidance on the use of near-infrared spectroscopy by the pharmaceutical industry and the data requirements for new submissions and variations
40. RIVM Verification of the identity of pharmaceutical substances with near-infrared spectroscopy
41. EMA Guideline on the use of near-infrared spectroscopy by the pharmaceutical industry and the data requirements for new submissions and variations
42. FDA Development and submission of near-infrared analytical procedures, Guidance for industry, Draft guidance
43. AOCS Cd 1e Determination of Iodine Value by Pre-calibrated FT-NIR with Disposable Vials
44. AOCS Am 1a-09 Near-Infrared Spectroscopy Instrument Management and Prediction Model Development
45. JIS K0134
46. GOST 33,441 Vegetable oils. Determination of quality and safety by near-infrared spectrometry
47. GOST 32,041 Compound feeds, feed raw materials. Method for determination of crude ash, calcium and phosphorus content by means of NIR spectroscopy
48. GOST 31,795 Fish, marine products and products of them. Method of determining the fraction of total mass of protein, fat, water, phosphorus, calcium, and ash by the near-infrared spectrometry
49. GOST 32,040 Fodder, mixed, and animal feed raw stuff. Spectroscopy in near-infrared region method for determination of crude protein, crude fibre, crude fat and moisture
50. GOST R 51,038 Fodder and mixed fodder. Spectroscopia in near-infrared region method for determination of metabolizable energy
51. GOST 30,131 Oil-cake and ground oil-cake. Determination of moisture, oil and protein by infrared reflectance.

***Note**

ASTM	American Society for Testing and Materials
ISO	International Organization for Standardization
AACC	American Association of Cereal Chemists
AOAC	Association of Official Analytical Chemists
ICC	International Association for Cereal Science and Technology
AOCS	American Oil Chemists Society
RACI	Royal Australian Chemical Society
USP	U.S. Pharmacopeia,
EP	European Pharmacopoeia
PASG	Pharmaceutical Analytical Sciences Group

CPMP&CVMP	Committee for Proprietary Medicinal Products & Committee for Medicinal Products for Veterinary Use
RIVM	lang-nllRijksinstituut voor Volksgezondheid en Milieu
EMA	European Medicines Agency
FDA	Food and Drug Administration
JIS	Japanese Industrial Standards
GOST	Gosudarstvennyy standard (Russian Standard).

In recent two decades, China has promulgated 80 national, industrial, and local standards, involving chemical, food, agriculture, textile, and other fields. The relevant NIR standard methods from different levels in China are as follows:

1. GB/T 18,868-2002 Method for determination of moisture, crude protein, crude fat, crude fiber, lysine, and methionine in feeds—Near-infrared reflectance spectroscopy method
2. GB/T 12,008.3-2009 Plastics-Polyether polyols—Part 3: Determination of hydroxyl number
3. GB/T 24,895-2010 Inspection of grain and oils—General regulations for model authentication of near-infrared analysis and administration and maintenance of network
4. GB/T 25,219-2010 Inspection of grain and oils—Determination of starch content in maize—Near-infrared spectroscopy method
5. GB/T 24,900-2010 Inspection of grain and oils—Determination of moisture content in maize—Near-infrared spectroscopy method
6. GB/T 24,902-2010 Inspection of grain and oils—Determination of crude fat content in maize—Near-infrared spectroscopy method
7. GB/T 24,896-2010 Inspection of grain and oils—Determination of moisture content in paddy—Near-infrared spectroscopy method
8. GB/T 24,897-2010 Inspection of grain and oils—Crude protein determination in rice—Near-infrared spectroscopy method
9. GB/T 24,898-2010 Inspection of grain and oils—Determination of moisture content in wheat—Near-infrared spectroscopy method
10. GB/T 24,899-2010 Inspection of grain and oils—Determination of crude protein in wheat—Near-infrared spectroscopy method
11. GB/T 24,871-2010 Inspection of grain and oils—Crude protein determination in wheat flour—Near-infrared spectroscopy method
12. GB/T 24,872-2010 Inspection of grain and oils—Determination of ash content in wheat flour—Near-infrared spectroscopy method
13. GB/T 24,870-2010 Inspection of grain and oils—Crude protein and crude fat determination in soybean—Near-infrared spectroscopy method
14. GB/T 29,858-2013 Standard guidelines for molecular spectroscopy multivariate calibration quantitative analysis
15. GB/T 34,406-2017 Identification of pearl powder—Near-infrared spectroscopy method
16. GB/T 36,691-2018 Methyl vinyl silicone rubber—Determination of vinyl content—Near-infrared spectroscopy

17. GB/T 37,969-2019 Standard guidelines for near-infrared qualitative analysis
18. GB/T 7383-2020 Non-ionic surface active agents—Determination of hydroxyl value
19. GB/T 13,892-2020 Surface active agents—Determination of iodine value
20. ChP 2015 The Pharmacopoeia of the People's Republic of China (2015) 9104 Guidelines for Near-Infrared Spectrophotometry
21. NY/T 1423-2007 Method for Quick Discrimination of Meat and Bone Meal in Fishmeal and Ruminant Concentrate Supplement—Near-infrared reflectance spectroscopy method
22. NY/T 1841-2010 Non-destructive determination of soluble solid and titratable acidity in apple fruit by near-infrared spectroscopy method
23. NY/T 2797-2015 Non-destructive determination of fat in meat by near-infrared spectroscopy method
24. NY/T 2794-2015 Determination of amino acids content in peanut—Near-infrared spectroscopy method
25. NY/T 3105-2017 Determination of oil content in vegetable oilseeds—Near-infrared spectroscopy method
26. NY/T 3299-2018 Determination of oleic acid and linoleic acid in vegetable oilseeds—Near-infrared spectroscopy method
27. NY/T 3298-2018 Determination of crude protein content in vegetable oilseeds—Near-infrared spectroscopy method
28. NY/T 3297-2018 Determination of total phenolic compounds and tocopherols in rapeseed seeds—Near-infrared spectroscopy method
29. NY/T 3295-2018 Determination of erucic acid and glucosinolate in rapeseed—Near-infrared spectroscopy method
30. NY/T 3512-2019 Non-destructive determination of protein in meat—Near-infrared spectroscopy method
31. NY/T 3679-2020 Technical code of practice for screening high oleic acid peanut—Near-infrared spectroscopy method
32. SN/T 3896.1-2014 Quantitative analysis of fiber in textiles for import and export—Near-infrared spectroscopy method—Part 1: Mixture of polyester fiber and cotton fiber
33. SN/T 3896.2-2015 Quantitative analysis of fiber in textiles for import and export—Near-infrared spectroscopy method—Part 2: Mixture of polyester fiber and polyurethane fiber
34. SN/T 3896.3-2015 Textiles for import and export—Quantitative analysis of fiber—Near-Infrared spectroscopy method—Part 3: Mixture of polyamide fiber and polyurethane fiber
35. SN/T 3896.4-2015 Textiles for import and export—Quantitative analysis of fiber—Near-Infrared spectroscopy method—Part 4: Mixture of cotton fiber and polyurethane fiber
36. SN/T 3896.5-2015 Textiles for import and export—Quantitative analysis of fiber—Near-infrared spectroscopy method—Part 5: Mixture of polyester fiber and rayon fiber

37. SN/T 3896.6-2017 Textiles for import and export—Quantitative analysis of fiber—Near-infrared spectroscopy method—Part 6: Mixture of polyester fiber and wool fiber
38. SN/T 3896.7-2020 Quantitative analysis of fiber in textiles for import and export—Near-infrared spectroscopy method—Part 7: Mixture of polyester fiber and polyamide fiber
39. SN/T 3896.8-2020 Fiber quantitative analysis of textile for import and export—Near-infrared spectroscopy —Part 8: Mixture of cotton and polyamide fiber
40. SN/T 5233-2020 Import and export textile material test of moisture regain of raw cotton—Near-infrared spectroscopy method
41. SB/T 11,149-2015 Technical specifications of waste plastics collection and sorting
42. FZ/T 01,144-2018 Textiles—Quantitative analysis of fiber—Near-infrared spectroscopy method
43. FZ/T 01,150-2019 Textile -Test method for identification of bamboo fibre and viscose from bamboo—Near-infrared spectroscopy method
44. LY/T 2151-2013 Method for determination of holocellulose and acid-insoluble lignin in wood—Near-infrared spectroscopy method
45. LY/T 2053-2012 Standard method for near-infrared qualitative analysis of wood
46. GH/T 1260-2019 Method for Moisture, Total Polyphenols, and Caffeine in Instant Tea in Solid Form—Near-infrared reflectance spectroscopy method
47. GH/T 1259-2019 Method for moisture, total polyphenols, and caffeine in tea polyphenol products—Near-infrared reflectance spectroscopy method
48. QB/T 2812-2006 Paper-online determination of weight and moisture (The near-infrared spectroscopy method)
49. HG/T 3505-2020 Surface active agents—Determination of saponification value
50. DB12/T 347-2007 Rapid method for determination of crude protein in wheat and corn by near-infrared spectroscopy
51. DB22/T 1605-2012 Rapid and nondestructive detection of ash content, moisture, water-insoluble solids, water-saturated butanol extract in the Ginseng—Near-infrared spectroscopy method
52. DB32/T 2269-2012 Determination of Cottonseed Oil Content by Near-Infrared Spectroscopy
53. DB21/T 2048-2012 Determination of crude protein, crude fat, crude fibre, moisture, calcium, Sphosphorus, crude ash, water-soluble chlorides, amino in feeds Near-infrared reflectance spectroscopy
54. DB22/T 1812-2013 Rapid and nondestructive detection of polysaccharides in the Ginseng-Near-infrared spectroscopy method
55. DB53/T 497-2013 Guidelines for the establishment and validation of near-infrared calibration models for the main chemical constituents of tobacco and tobacco products
56. DB53/T 498-2013 Determination of main chemical components in tobacco and tobacco products—Near-infrared diffuse reflectance spectroscopy method

57. DB53/T 512-2013 Determination of the uniformity of blending of double-cut microwave expanded stalks by near-infrared spectroscopy method
58. DB34/T 2561-2015 Rapid analysis of conventional indicators in solid-state fermented grains—Near-infrared spectroscopy method
59. DB43/T 1065-2015 Determination of amino acids in feeds—Near-infrared reflectance spectroscopy method
60. DB34/T 3054-2017 The method of rapid determination of the main flavor components in strong flavor. Chinese spirits near-infrared spectroscopy
61. DB15/T 1229-2017 Test method for pure cashmere content of cashmere—Near-infrared reflectance spectroscopy method
62. DB34/T 2890-2017 Method for determination of the major components in tea—Near-infrared reflectance spectroscopy method
63. DB64/T 1554-2018 Method for fiber determination of cotton and polyester fiber-blended products—Near-infrared spectroscopy method
64. DB37/T 3635-2019 Technical specification for rapid screening of motor vehicle gasoline
65. DB37/T 3636-2019 Rapid detection method of motor vehicle gasoline near-infrared spectroscopy method
66. DB37/T 3637-2019 Technical Specification for Rapid Screening of Automobile Diesel Fuels
67. DB37/T 3638-2019 Rapid Detection Method of Automobile Diesel Fuels—Near-Infrared spectroscopy method
68. DB37/T 3639-2019 Technical specification for rapid screening of motor vehicle ethanol gasoline (E10)
69. DB37/T 3640-2019 Rapid detection method of motor vehicle ethanol gasoline (E10)—Near-infrared spectroscopy method
70. DB37/T 4118-2020 Rapid detection method of diesel engines NO_x reduction agent—Aqueous urea solution (AUS 32)—Near-infrared spectroscopy method
71. DB36/T 1127-2019 Method for determination of crude ash, calcium, phosphorus, and chlorides in feeds—Near-infrared reflectance spectroscopy method
72. DB34/T 3561-2019 The method for determining conventional indicators of brewing raw materials—Near-infrared spectroscopy method
73. DB12/T 955-2020 Determination of nitrogen and phosphorus in the slurry of dairy farm—Near-infrared diffused reflection spectroscopy
74. DB32/T 3881-2020 Intelligent factory of Chinese medicine—Quality control of the extraction processes by water-extraction and alcohol-precipitation
75. T/AHFIA 008-2018 Rapid determination method of physical and chemical indicators for brewing Daqu—Near-infrared spectroscopy method
76. T/GZTPA 0001-2020 Determination of the main chemical constituents in Guizhou green tea by near-infrared diffuse reflectance spectroscopy
77. GH/T 1337-2021 Rapid determination of impurity content of seed cotton—Near-infrared spectroscopy method
78. T/CIS 11001-2020 On-line detection of powder blending uniformity in the production of traditional Chinese medicine—Near-infrared spectroscopy method

79. T/CBJ 004-2018 The general analysis method of solid-state fermented grains
80. GB/T 40,467-2021 Livestock and poultry meat quality testing—Guideline for near-infrared spectroscopy method.

*Note

GB/T	China National Standards (recommendation)
NY/T	Agricultural Industry Standard of the People's Republic of China (recommendation)
SN/T	Industrial Standard of Import and Export Commodity Inspection of the People's Republic of China (recommendation)
SB/T	Commercial industry standard of the People's Republic of China (recommendation)
FZ/T	Textile Industry Standard of the People's Republic of China (recommendation)
LY/T	Forestry Industry Standards of the People's Republic of China (recommendation)
GH/T	Industry Standards for Supply and Marketing Cooperation of the People's Republic of China (recommendation)
QB/T	Standard for light industry of the People's Republic of China (recommendation)
DB/T	Local standards of the People's Republic of China (recommendation)
HG/T	People's Republic of China Chemical Industry Standard (recommended)
ChP	The Pharmacopoeia of the People's Republic of China
T/AHFIA	Food Industry Association Group Standards of Anhui province of China
T/GZTPA	Green Tea Brand Development Promotion Association Group Standards of Guizhou province of China
T/CIS	Group Standards of China Instrument and Control Society
T/CBJ	Group Standards of China Liquor Industry Association.

2.3 Mid-Infrared Spectroscopy

The MIR spectroscopy is commonly called IR spectroscopy with a spectral range of 400–4000 cm^{-1} , reflecting the spectral information of the vibration and rotation of material molecules, and the fundamental frequency absorption bands of the vast majority of organic compounds and inorganic ions appear in this region. Compared with the NIR spectral region, the MIR spectral region has strong absorption, relatively rich information, and strong group resolution ability, which can distinguish substances with very similar structure. This is also the reason why MIR spectroscopy has been mainly used for the analysis of molecular structure of substances for a long

time. In recent years, with the development of instrument manufacturing technology, chemometrics methods, and computers, MIR spectroscopy has been increasingly used in the field of emergency analysis and online process analysis.

2.3.1 Portable Mid-Infrared Spectral Analysis Technology

With the improvement of instrument manufacturing level and the new demand brought by social development, portable MIR spectrometer has been applied in more and more fields, such as product quality detection, environmental monitoring, and hazardous material leakage emergency monitoring. From the point of view of instrument type, most portable instruments still use Fourier transform type, but there are also other types, such as array detector type. From the point of view of measurement objects, there are special portable instruments suitable for a variety of amorphous samples (such as gas, viscous liquid, and solid powder). These portable instruments have been structurally modified to adapt to very stringent field environments. For example, some instruments can operate at temperatures ranging from 0 to 100% humidity and -10 to 50 °C.

In the determination of light oil products (gasoline and diesel), there are a number of special portable infrared analyzers, which mostly used transmission measurement way, automatic injection and cleaning, built-in a variety of chemometrics calibration models. They can quickly determine gasoline, diesel and jet fuel, and other conventional physical and chemical properties and chemical composition data. It is used in the field of intermediate control analysis and quality inspection in the circulation process.

There are also a number of portable MIR instruments dedicated to the determination of lubricating oil or biodiesel, which mainly used ATR measurement method and slightly transmission mode. They were employed to conduct quality monitoring of lubricating oil in the process of use, control analysis of biodiesel in the process of production, determination of mixing proportion of biofuel in the process of flow. In the aspect of quality monitoring of lubricating oil, mid-NIR spectroscopy combined with chemometrics can determine many physical and chemical indexes such as acid value, alkali value, and water content of lubricating oil. In terms of biodiesel analysis, MIR spectroscopy can be used to determine the composition of biodiesel and its feedstock, such as methyl ester and glycerol, as well as the mixing ratio of petrochemical diesel and biodiesel. With a few modifications, these instruments can be used for other analysis, such as ethanol in gasoline, ethanol in beverages, and heavy water (D₂O) in water.

Portable MIR spectrometer can be used for on-site identification and analysis of liquid, viscous, and gel materials. Such instruments are usually equipped with standard spectral libraries of thousands of substances including common laboratory chemicals, toxic industrial chemicals, chemical warfare agents, explosives, criminal investigation drugs, controlled drugs and precursors, common white powder spectra. By spectral retrieval, the material composition of the sample can be quickly identified.

It can be applied to military, fire, customs import and export, environmental protection, and health law enforcement departments, such as identification of unknown powder found in business district, identification of unknown liquid leaked in traffic accidents, on-site inspection of chemicals in the process of transportation [23].

Portable MIR spectrometer can be used for on-site gas analysis, and the inner wall of the gas pool is coated with precious metals such as gold or rhodium, which is highly corrosive. The optical path ranges from a few centimeters to a few meters, depending on the concentration of the gas to be measured. In order to detect trace concentrations of gas, some instruments are equipped with gas enrichment devices. These instruments are equipped with thousands of standard MIR spectra of gases, which can be used for emergency monitoring and analysis of environmental pollution accidents, identification, and monitoring of chemical weapons agents in anti-terrorist activities, on-site monitoring of labor health.

2.3.2 Online Mid-Infrared Spectral Analysis Technology

From the spectral theory and analysis principle, the project analyzed by NIR spectroscopy can also be analyzed by MIR spectroscopy if equipped with appropriate measurement accessories. Moreover, the sensitivity of MIR spectroscopy is an order of magnitude higher than that of NIR, and it generally can measure the content of more than 0.01% of the components. However, due to the high price of online MIR spectrometer and the limitation of measurement accessories, its application in the field of process industry (such as petrochemical and pharmaceutical) is far less extensive than that of online NIR spectrometer. Most applications of online MIR spectroscopy are also focused on experimental reaction processes, such as organic synthesis, polymerization, and biochemical reactions.

For liquid such as wine and milk with good fluidity, MIR spectral measurement (optical path is 20–200 μm) can be carried out by transmission mode, because transmission spectroscopy provides stronger structural information than ATR spectroscopy, and has advantages in the determination of low content substances [24]. In the detection of raw milk, MIR spectroscopy has been widely recognized as a rapid detection method in the international dairy industry. Through the establishment of the chemometrics calibration model, the items that can be analyzed include milk fat, milk protein, lactose, non-fat solids, total dry matter, density, and water mixing rate. In dairy production enterprises, MIR spectroscopy is used for the control and supervision of raw milk quality, as well as the standardized monitoring of the production process, so as to keep the balance of the physical and chemical components of each batch of products and ensure the continuity of the final product quality and the consistency of flavor (taste). The contents of low concentration urea (0.01–0.08%), acetone (0.00–0.02%), and microorganisms in milk can also be measured by MIR spectroscopy.

In recent years, online gas measurement technology combining MIR full-spectrum measurement and chemometrics is emerging quietly [25]. For example, Fourier transform MIR gas analyzers have been installed at the outlet of the desulfurization tower (sprayer) on the waste incinerator line. The waste incineration gas mainly contains gaseous pollutants (such as SO₂, NO, NO₂, CO, CO₂, HCl, HF, NH₃, etc.) and H₂O, among which the concentration of some gas components is sometimes very high, such as the concentration of H₂O up to 40 v%, the concentration of CO₂ up to 20 v%, etc. However, the concentration of HCl and HF is generally only 10~30 mg/m³, with a range ratio of more than 10⁴. Other analytical methods do not have such a wide dynamic measurement range as the quantitative analysis of MIR spectroscopy, so it is difficult to meet such requirements. The data monitored by the Fourier transform MIR gas analyzer can be used to adjust and control the operation of the desulfurization tower (for example, to control the amount of lime slurry in the spray tower), and can also be used as the basis for environmental assessment.

2.4 Raman Spectroscopy

Raman spectroscopy and MIR spectroscopy are both molecular vibration spectroscopy, but their generating principles are very different. MIR spectroscopy is the absorption spectrum, while Raman spectroscopy is the scattering spectroscopy. In 1928, when studying the light scattering of benzene, Indian physicist Raman found that in the scattered light, in addition to the scattered light with the same frequency as the incident light (namely, Rayleigh scattering), there are scattered light with different frequency with the incident light, namely, Raman scattering. The intensity of Raman scattering is extremely weak, with only 10⁻³~10⁻⁶ of Rayleigh scattering intensity.

Raman scattering is the result of inelastic collision between light and material molecules. The difference between the frequency of scattered light and incident light reflects the frequency of the photon corresponding to the difference of energy level of molecular vibration, which is called Raman shift. It has nothing to do with the frequency of incident light, and the wavenumber range is about 0~4000 cm⁻¹. For functional groups with weak MIR absorption, such as non-polar groups C=C, C-C, and S-S, strong absorption bands can be obtained in Raman spectroscopy. The chemical functional groups of various substances have the Raman vibration band with sharp and strong characteristics, which makes it easy to distinguish different substances. Moreover, the vibration band is also sensitive to the physical and chemical environment, so the position and strength of the band can also sensitively reflect the information of the structure and conformational change process of the relevant substances. The Raman spectral signal of water molecules is very weak, so it is easy to obtain the Raman spectra of water samples. In addition, Raman spectroscopy does not require the sample to own good light transmittance, so it is easy to obtain Raman spectra of turbid samples.

2.4.1 *Fourier Transform Raman Spectroscopy*

Raman spectrometers can be divided into dispersion Raman and Fourier transform Raman (FT-Raman) spectrometers according to different principles. Dispersive Raman is the principle of grating dispersion to obtain the spectrum. The lasers from UV, visible to NIR wavelength range can be used as the excitation source. FT-Raman uses Michelson interferometer to obtain Raman spectra by the way of Fourier transform. Most of the 1064 nm semiconductor lasers are used as the excitation source. Compared with the dispersive Raman spectrometer, FT-Raman has the advantages of fast scanning speed, good spectral reproducibility, high-frequency accuracy, wide measurement frequency range, high signal-to-noise ratio, small thermal effect, possibly overcoming fluorescence interference and directly passing through the biological tissue with NIR light to obtain useful information of the molecules in the tissue.

FT-Raman has many applications in drug analysis and food. Okumura et al. used FT-Raman spectroscopy combined with partial least squares (PLS) to establish a method for the rapid determination of indomethacin microcrystalline content, which could accurately predict the drug content in indomethacin tablets [26]. Szostak et al. established a PLS model based on FT-Raman spectroscopy to predict the content of active ingredients acetaminophen and diclofenac sodium in commercial suppositories [27]. This method can be promoted for rapid quantitative analysis of suppositories. FT-Raman can be used for the quality identification of unsaturated vegetable oils, such as unsaturation, iodine value, free fatty acids, oxidation stability, and adulteration identification [28].

2.4.2 *Surface Enhanced Raman Scattering Spectroscopy*

When some molecules are adsorbed to the surface of some rough metals, such as gold, silver, or copper, the intensity of their Raman signal is increased by 10^4 – 10^6 times, and the band position is not very different from the normal Raman spectrum. This unusual Raman scattering enhancement phenomenon is called surface enhanced Raman scattering (SERS) effect. In recent years, benefiting from the rapid development of laser technology and nanotechnology, SERS has been widely applied in the fields of interface and surface science, material analysis, biology, medicine, environment, and security [29]. For molecules with surface enhanced resonance Raman scattering (SERRS) effect, the intensity can be increased by 2–3 orders of magnitude if the excitation wavelength is adjusted to the absorption wavelength of the adsorbed molecules, and the detection limit can be as low as 10^{-9} mol/L.

SERS spectroscopy overcomes the disadvantage of low sensitivity of conventional Raman spectroscopy, and can obtain more material structure information. It has a broad application prospect in the field of on-site rapid screening, detection, and identification of pesticide residues, veterinary drug residues, and detection

of restricted or banned additives. Mamian-Lopez et al. took Klarite as the Raman enhanced substrate, and established a SERS analysis method for moxifloxacin, a fluoroquinolone antibiotic, with the detection limit of 0.085 mg/L and the limit of quantitation was 0.6 mg/L [30] by eliminating matrix effect through standard addition method and multiple curve resolution alternating least square method (MCR-ALS). Zhang et al. adopted commercial Klarite and Q-SERS substrate combined with principal component analysis (PCA) and partial least squares (PLS) method to establish SERS analysis method for enrofloxacin, furazolidone, and malachite green in fish, which could detect 1.0 $\mu\text{g/g}$ furazolidone and 200 ng/g malachite green in tilapia fillet [31]. Huang et al. used surface enhanced Raman spectroscopy combined with PLS to establish a quantitative model for malathion residue in Chinese cabbage, and the detection concentration of malathion in Chinese cabbage reached 1.08 mg/L [32]. Liu et al. realized the nondestructive detection of thiophosamine pesticide residues in navel orange with the enhanced base of floccule silver gel combined with chemometrics, and the detection limit reached 4.13 mg/L [33]. Based on silver nanorod array, Nie et al. built a quantitative model for predicting trace biformidine in honey through SERS and PLS. Compared with the traditional single variable quantitative model based on SERS single peak intensity, this multi-predicative model integrated all characteristic peaks of biformidine, improving the detection accuracy and anti-interference ability [34].

SERS spectroscopy has unique advantages such as non-invasive, high sensitivity, good selectivity, and small water interference, which makes it have good application prospects in life science, clinical laboratory, and it has become a very potential biological detection technology [35]. In combination with PCA and independent data t-test and other statistical methods to analyze Raman spectra, Liu et al. used SERS spectroscopy of human serum based on silver nanofilm solid device to carry out non-labeled and non-invasive detection of liver cancer at the molecular level. The diagnostic sensitivity was about 95.0% and the specificity was about 97.6% [36]. This non-labeling and non-invasive test have great potential for detecting cancer clinically.

Using SERS to classify and identify bacteria has become one of the hot spots in the field of microbial detection [37]. The application of SERS in bacterial classification was mainly to distinguish different species of bacteria and different types of the same species of bacteria. Raman spectroscopy has a relatively high information content, which is derived from the vibrational and rotational frequencies of the molecules in the sample. The molecular vibration frequencies of nucleic acids, proteins, lipids, and carbohydrates in bacteria are different, which are shown as their unique spectral peaks on Raman spectrum and can generate “whole biological fingerprint” that can be distinguished by pattern recognition methods. For example, urinary tract infections are a common condition, and the current gold standard for detecting infections is the traditional culture method, which costs a long time. Jarvis et al. used SERS combined with PCA and discriminant function analysis to study pathogenic bacteria of urinary tract infection, which successfully identified the main pathogenic bacteria groups of five different species [38].

In response to and disposal of public security emergencies involving chemical terrorism substances (such as chemical warfare agents, biological toxins, and other

highly toxic chemical substances), it is very important to carry out real-time, rapid, accurate, reliable, and highly sensitive on-site inspection. Due to its sensitivity, rapidity and portability, SERS spectroscopy has gradually attracted attention in the field of detection and security of chemical terrorist substances and is expected to be widely used in the fields of national defense security, public security, and on-site detection of chemical emergencies [39]. Surface enhanced Raman spectroscopy can be used for on-site detection and real-time and rapid analysis of trace and even ultra-trace drugs by portable Raman spectrometers, which has broad application prospects [40]. Dong et al. adopted dynamic SERS substrate and used PCA to reduce the dimension of the spectrum. Then, a discriminant model was established by support vector machine (SVM) and the accuracy rate of identifying the real urine of methamphetamine users reached 90% [41].

In addition to SERS, resonance Raman spectroscopy (RRS), coherent anti-Stokes Raman spectroscopy (CARS), and stimulated Raman spectroscopy (SRS) and the combination of these technologies with SERS (such as CARS-SERS) were used to enhance Raman signals [42–44].

2.4.3 Confocal Raman Spectroscopy

Confocal micro-Raman spectroscopy, also called micro-Raman, is a technique that combines Raman spectroscopy with micro analysis. In the essence of the spectrum, there is no difference between the micro-Raman and the ordinary laser Raman. The confocal microscope is introduced into only the optical path of the laser Raman so that the stray light from the defocused region of the sample can be eliminated, and the spatial filtering can be formed to ensure that the detector can capture the sample to be measured. By adjusting the position of the focal point, the laser can be focused to different depth of the sample, so as to realize the situ and nondestructive analysis of trace samples. Confocal micro-Raman spectroscopy has many advantages in microanalysis and determination, such as good separation effect, high sensitivity, simple equipment, and easy operation. Therefore, micro-Raman spectroscopy has been widely used in tumor detection, cultural relic archaeology, public security law, and other fields.

Microscopic Raman spectroscopy is a technology that can provide the spatial resolution of 0.5–1.0 μm for the study of the chemical structure of individual microorganism cells. In recent years, microscopic Raman spectroscopy has been used more and more in the study of single microorganism cells, which can distinguish the chemical composition of single microorganism cells in space. Due to the differences in the basic components such as protein, DNA, RNA, lipid, and carbohydrate in microbial cells, the Raman spectra of different species will be different to some extent. Therefore, these small spectral changes can be extracted and studied, which combined with chemometrics techniques to distinguish the species of microorganisms.

Laser optical tweezers Raman spectroscopy is a technology that combines laser optical tweezers with confocal Raman spectroscopy. This technology could capture,

manipulate, and measure single active cells in suspension under physiological conditions, which is used for the study of biological analysis. Laser Tweezers technology is a physical tool based on the mechanical effects of the laser, which utilizes the optical potential well formed by the interaction between strong converging light field and particles to capture particles. Optical tweezers have become a useful tool for trapping and manipulating biological particles, including cells, bacteria, viruses, and dielectric particles. The combination of optical tweezers and Raman spectroscopy can characterize molecules contained in individual organic droplets or microcapsules. The significant advantage of optical tweezers is the ability to confine Brownian particles in an aqueous solution to a small area, allowing for long periods of time to observe the properties of individual particles. Raman single-celled precise separation technology is a non-invasive and unmarked without damage of single cell technology and also a kind of quick and effective analysis tool for the identification of intracellular molecular composition, which can sort out the single-celled creatures without tags and intact, effectively identify the biological chemical composition and reflect the most real reaction cell in situ state of activity and function [45].

Kusic et al. employed single-cell Raman combined with SVM to classify and identify *Legionella* species associated with human diseases and other common aquatic pathogens, and established a Raman spectral database of 22 species of the genus as well as *Escherichia coli*, *Klebsiella pneumoniae*, and *Pseudomonas aeruginosa*. The study indicated that, Raman micro-spectroscopy can be used as a fast and reliable method to identify human pathogen *Legionella* species [46]. Klob et al., for urinary tract infections, detected the patient samples by the confocal Raman microscopy combined with SVM. The experimental results show that Raman technology can accurately detect urine samples of patient's body in the case of no medium in 2 h, and can determine the main bacterial infection. The accuracy can reach more than 92% [47]. Yogesh et al. also used micro-Raman spectroscopy combined with SVM to identify five common pathogenic bacteria of urinary tract infection, and the recognition accuracy was close to 90% [48]. Stockel et al. applied confocal Raman spectroscopy to detect 26 species of mycobacteria, including *Mycobacterium tuberculosis*, *Mycobacterium abscess*, and *Mycobacterium avium*, with a total of 8845 strains, and established a Raman spectral database of mycobacteria. Through SVM, unknown mycobacteria could be identified to the species level with an accuracy of 94.3% [49].

Li et al. used confocal Raman to detect in vitro tissues of nasopharyngeal carcinoma and established a pattern recognition model using PLS-DA, with diagnostic sensitivity and specificity of 85% and 88%, respectively [50]. Lee et al. employed the Raman spectra of extracellular vesicles collected by confocal microscopic Raman spectrometer and combine convolutional neural network (CNN) to diagnose prostate cancer with an accuracy rate of more than 93% [51]. Pablo et al. obtained the chemical fingerprints of colorectal cancer by Raman spectroscopy of living single cells, and classified cells by PCA and linear discriminant analysis (LDA) with an accuracy of 98.7%. Raman spectra can reveal the tumor cell sugar, phosphate, nucleic acid content, and protein α helix, folded β or $\alpha + \beta$ secondary structure, so as to distinguish between different cell types and different kinds of colorectal cancer cell lines,

and can further distinguish the different stages of the disease. This can be used as cell phenotype analysis in the important tool of clinical diagnosis [52]. Pilat et al. developed a microfluidic chip, which combined with optical tweezers technology to isolate single *E. coli*. By comparing the change of resonance Raman spectra of single-celled *E. coli* under the antibiotics pressure or not, more obvious changes of the peak were found. The results of PCA also show that statistically significant differences exist between them and research on drug resistance of individual bacteria can better understand the problem of heterogeneous drug resistance [53].

2.4.4 Spatial Offset Raman Spectroscopy

Confocal method can only determine the Raman spectra of solid samples within a depth of several hundred microns, while spatially offset Raman spectroscopy (SORS) can determine the Raman spectra of samples at a deeper depth. As shown in Fig. 2.5, the principle is that the incident focus of the laser source and the focus of the collecting lens in the spectral system are offset by a certain distance on the surface space of the sample measured. SORS can clearly distinguish the Raman spectra of the material and the container and realize the simultaneous identification of the material and the container, so as to analyze the chemical information inside the opaque sample. Container types include transparent plastic bags, opaque or colored high-density polyethylene plastic containers, colored or transparent glass containers, jute bags, and multi-layer paper bags. The SORS can effectively eliminate the fluorescence from the surface layer and truly realize the non-invasive and non-destructive fast detection [54].

SORS measurement method can be used to detect the authenticity of drugs through bottles or plastic blister packaging, as well as powder or liquid explosives in non-metallic containers [56]. As shown in Fig. 2.6, the SORS method can obtain the Raman characteristic spectral information of 30% hydrogen peroxide across the

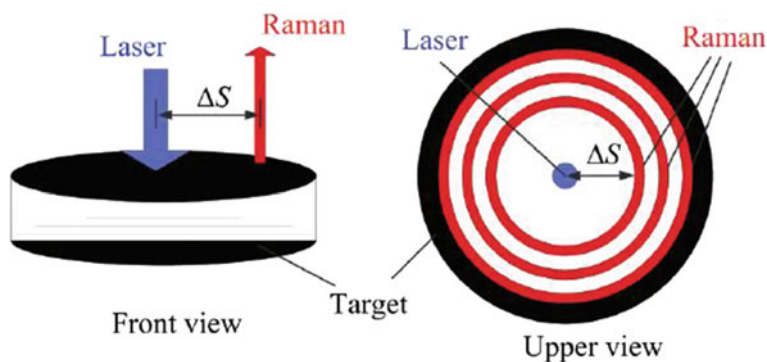


Fig. 2.5 SORS optical schematic [55]

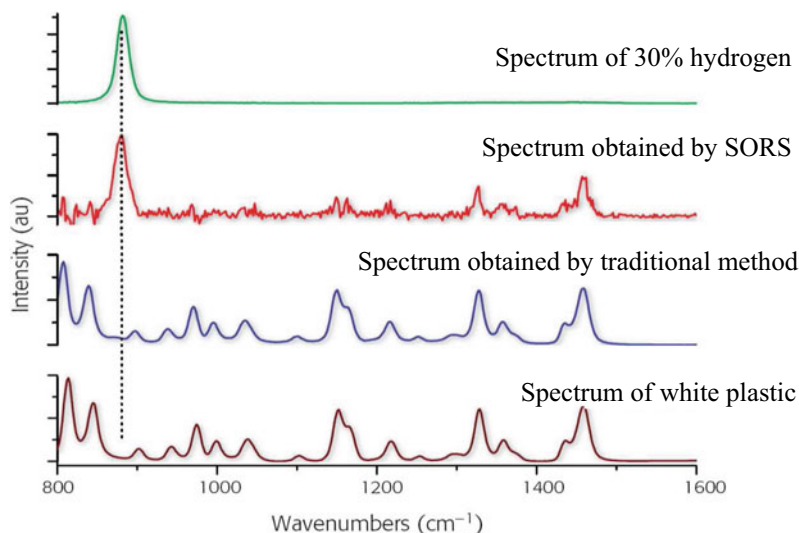


Fig. 2.6 Raman spectra of 30% hydrogen peroxide in a 1.5 mm white plastic bottle were measured using SORS method [56]

1.5 mm white plastic bottle, while the traditional measurement method can only obtain the spectral characteristics of the plastic bottle. Cobalt light systems company has developed the commercialized portable Raman spectrometer based on SORS, which is capable of acquiring the characteristic Raman spectra of raw materials through opaque packaging or containers and completing direct non-destructive identification of raw materials within 10 s. The product is in line with current GMP (CGMP) manufacturing practices. The company has also developed the commercially available insight portable Raman spectrometer, which has been approved by the European Civil Aviation Safety Supervisory Commission and is already used in some European airports to detect powder or liquid explosives through colored, opaque, or transparent plastic, glass, and paper packaging.

SORS is also applied in the medical field, such as the non-invasive diagnosis of subcutaneous skeletal disease and cancer [57]. Ding et al. used the method of SORS to analyze the healing of the thigh fracture in rats at 2 weeks and 4 weeks after fracture, and found that collagen mineralization and mineral carbonation increased significantly at 4 weeks after fracture than at 2 weeks after fracture. The test results of SORS were consistent with the radiological and material tests, indicating that SORS have the potential to evaluate the healing of fractures in vivo [58].

In addition to SORS, as shown in Fig. 2.7, there are also reverse SORS and slanted SORS [55].

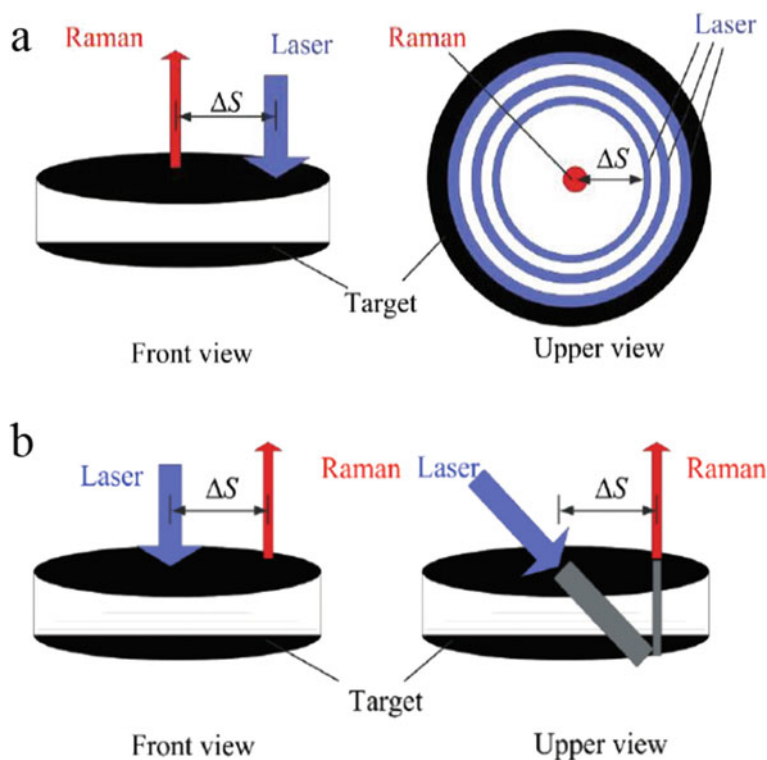


Fig. 2.7 Schematic diagram of SORS. **a** reverse SORS; **b** sloping SORS [55]

2.4.5 Transmitted Raman Spectroscopy

Traditional Raman spectra are measured by backscattering, but the transmission measurement method can obtain the information of the overall sample and effectively eliminate the fluorescence interference generated on the sample surface [59]. Figure 2.8 shows the positive and negative Raman spectra of 3.9 mm paracetamol tablets measured by the traditional backscattering method (a) and transmission method (b). One side of the tablet is covered by 2 mm trans-1, 2-stilbene. It can be seen that the reverse and inverse Raman spectra of tablets measured by the traditional backscattering method are significantly different. However, Raman spectra with the same reverse and inverse sides can be obtained by adopting transmission mode [60]. In terms of drug composition determination, many application examples show that the transmission Raman measurement method combined with the multivariate calibration method can give better quantitative results than the traditional backscattering measurement method [61, 62].

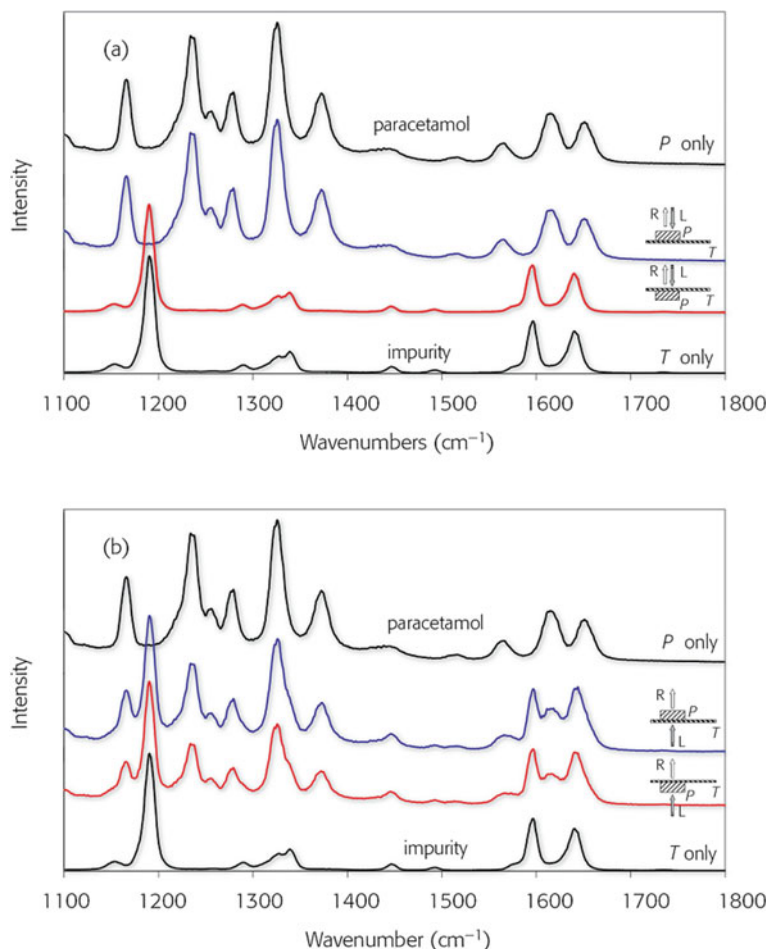


Fig. 2.8 Conventional backscatter (a) and transmission (b) measurements of the positive and negative Raman spectra of a paracetamol tablet contaminated with trans-1, 2-stilbene (L: incident laser; R: Raman scattered light; P: Paracetamol; T: trans-1, 2-stilbene) [60]

Cobalt light systems have developed a commercially available TRS100 transmission Raman analysis system that measures the content of multiple active ingredients in a complete tablet or capsule. The QTRam Raman spectrometric instrument developed by B&WTEK Company also uses the transmission method to collect Raman information through solid drug formulations, which can be used for rapid and nondestructive testing of drug composition uniformity in pharmaceutical companies.

2.4.6 *Portable Raman Spectral Analysis Technology*

In the actual application process, many occasions do not need the ultra-high resolution and sensitivity of confocal, portable Raman can complete most of the related applications. At present, portable/handheld Raman spectroscopy combined with chemometrics has been widely used in many fields, such as food safety, drugs, drug screening, and packaging material testing [63–65].

Sensory properties are an important indicator of food quality and portable Raman spectroscopy has been successfully applied to predict food sensory properties and quality grading based on sensory properties [66]. For example, Wang et al. successfully established the portable Raman spectroscopy quantitative model through PLS to predict the three sensory properties of pork loin juiciness, tenderness, and chewability, with an accuracy of over 80%. SVM was used to divide the sensory evaluation and corresponding Raman spectral data of pork loin into three levels according to tenderness and chewability. The accuracy of predicting pork with good grade reached 100% [67].

In recent years, portable Raman spectrometers have been used in archeological field research in the fields of precious artworks, manuscripts, pigments, ancient ceramics, and frescoes, providing a lot of convenience for in situ nondestructive testing of many large archaeological samples. There have been many reports on Raman spectroscopic studies of ancient cultural relics such as paleontology, ancient ceramics, glass, gems, ancient manuscripts, murals, textiles, and mummies. These results provide scientific basis for the identification of age and attributions of cultural relics, as well as the preservation and restoration of cultural relics [68].

The portable Raman spectrometer usually adopts a single wavelength laser of 785 nm, which has a signal-to-noise ratio 10~70 times higher than that of the laser of 1064 nm when detecting the sample with small fluorescence background. However, it will be seriously interfered when detecting the sample with strong fluorescence background. For this reason, Christesen et al. developed a 785 nm/1064 nm dual-wavelength handheld Raman spectrometer. The former is mainly used to detect samples with no fluorescence background and those with fluorescence background but whose spectrum is less affected by it, while the latter is mainly used to detect samples with strong fluorescence background [69].

Common Raman spectrometers have a spot diameter range of 50~500 μm on the sample, and it is difficult to guarantee the uniformity of the spectral sampling area for mixed non-uniform solid samples (such as tablets). Nowadays, commercially available portable Raman spectrometers use variable dynamic point sampling (VDPS) technique to obtain representative spectra. This measurement method keeps the sample stationary while the laser beam scans the sample at a high frequency of tens of Hertz in accordance with a preset grid trajectory. The spatially averaged Raman spectrum can be obtained in a very short time, thus obtaining a representative spectrum from the inhomogeneous sample. Fluotium interference is a key problem for both the user and the machine manufacturer. There is a commercial

and portable Raman spectrometer using shifted excitation Raman difference spectroscopy (SERDS) technology to eliminate the fluorescence interference [70, 71]. It uses two excitation sources with similar wavelength to excite the sample, respectively, to obtain two Raman spectra, and makes difference between the two spectra so that the effect of fluorescence can be effectively eliminated.

2.4.7 Fiber Raman Spectral Analysis Technology

Because the excitation light source and Raman scattering of the Raman spectrometer are both in the visible or NIR region, quartz fiber can be used to transmit the excitation light and collect and transmit the scattered light, while the spectrometer is placed far away from the harsh environment of the analysis site. The Raman fiber probe used for online analysis generally uses a backscattered 180° optical structure. In order to efficiently collect Raman scattered light and eliminate other interference-scattered light, there are a variety of commercial fiber probe forms. Because Raman spectrometers collect the scattering spectra of the measured material, no special sampling device is usually required. The objects can be various states of substances such as liquid, solid, and gas, which is especially suitable for online monitoring of multiphase polymer polymerization reactions.

Online Raman spectroscopy can track the polymerization process, measure the content of reactants, intermediates, and final products in real time, and be used for the study of reaction kinetics and the control analysis of the production process. For example, it was utilized to judge the appropriate reaction end point [72]. In addition to the insertion optical fiber probe, the non-contact optical fiber probe can also be used to monitor the whole reaction process through the optical window on the reactor wall. Many application examples show that Raman spectroscopy is very suitable for online analysis of emulsion polymerization processes containing high concentration of solid suspension, such as emulsion copolymerization of butyl, acrylate/methyl, and methacrylate. Raman spectroscopy was also used to monitor the emulsion polymerization process of droplets in a microfluidic device in real time.

In the field of petrochemical, Raman spectroscopy is very suitable for determination of the content of aromatic family of compounds, such as aromatics extraction of BTEX (benzene, toluene, ethylbenzene, and xylene) content, and C8 aromatics isomers, namely, paraxylene, xylene between, o-xylene, and ethyl benzene content of each isomer in the process of the separation [73]. In the simulated moving bed adsorption separation process and crystallization separation process of p-xylene, Raman spectroscopy has been used to measure the content of each component in real time, so as to adjust the process parameters in time, realize optimal control operation, and improve the stability of the production process and the purity of the product [74].

The effective combination of Raman spectroscopy and optical fiber probe makes it a useful tool for the diagnosis of living diseases [75, 76]. Yan et al. used the optical fiber Raman spectrometer combined with CNN to identify tongue squamous cancer

cells, and the sensitivity and specificity of the recognition result were 99.07% and 95.37%, respectively [77]. Raman spectroscopy has been used for the early diagnosis of osteoarthritis, osteoporosis, and other diseases as well as the assessment of fracture risk due to its ability to identify subtle molecular changes in bone tissue [78]. Raman spectroscopy can be used to evaluate bone composition parameters related to bone quality, such as mineral to matrix ratio, carbonate to phosphate ratio, mineral crystallinity, and collagen maturity. Buckley et al. used three different multivariate resolution methods of target band entropy minimization (BTEM), multiple curve resolution (MCR), and parallel factor analysis (PARAFAC) to process fiber optic Raman spectral data of bones. The results showed that all three methods could accurately reconstruct the ratio of phosphate to carbonate, and the error of each analysis was less than 2%. The results of PARAFAC are closest to the measured mineral to collagen ratio and are accurate enough to detect differences in components associated with osteoarthritis, osteoporosis, and osteogenesis insufficiency [79].

2.5 Ultraviolet-Visible Spectroscopy

UV-Vis absorption spectroscopy, also known as molecular electron transition spectroscopy, is produced by the electron transition in the outer layer of molecules after the absorption of UV or visible light. Its spectral range is 190–800 nm. Because the electron energy level in the molecule is greater than the vibrational and rotational energy levels, when the molecule absorbs light to realize the electron transition, the vibrational and rotational spectra of the molecule must be accompanied, and they overlap each other. Therefore, compared with the MIR spectrum, the absorption band of the UV spectrum is relatively wide. UV spectrum is only related to chromophore and auxochrome in molecules, mainly involving the part of electronic structure related to π electrons. In structural analysis, the role of UV spectrum is mainly to provide the size of the conjugated system of organic matter and the skeleton information related to the conjugated system. The number of absorption bands in UV-Vis spectrum is not large, and many compounds have very different structures. But as long as they have the same chromophore and auxochrome, their UV-Vis absorption spectra will be very similar.

There are three main types of UV-Vis spectrometers according to the different methods: filter, scanning grating, dispersion and fixed optical path array detector (CCD and PDA). Because the traditional scanning grating dispersive spectrometer has a rotating grating and many optical components, it is seldom used in online analysis. The filter type instrument is durable, cheap, and low resolution, and thus it is often used to form a relatively simple measurement system. The application of modern UV-Vis spectroscopy online analysis technology benefits from the development of optical fiber, array detector, and chemometrics. Under the irradiation of deuterium lamp or xenon lamp, the quartz fiber used in ordinary spectroscopy will be damaged due to the formation of “color center” when absorbing the deep UV light in the band of 214 nm, and its optical transmission performance will decay rapidly in

a short time. Therefore, special fiber resistant to UV exposure must be selected for the transmission of deep UV light. Optical fiber enables the spectrometer to conduct in situ measurement far away from dangerous measurement points. The emergence of PDA and CCD array detector makes the manufacturing of fast and continuous full-band online UV spectrometer become a reality. Since this type of instrument has no removable optical components, it is ideal for online analysis. The application of chemometrics can be used to analyze the overlapping spectra of complex mixtures and directly determine the concentration of multiple components.

UV-Vis spectroscopy is a classical analysis method in the petrochemical field. Most inorganic substances such as sulfur compounds and organic compounds with conjugated double bonds such as aromatic compounds, have characteristic absorption in the UV-Vis region. However, the lack of absorption of saturated hydrocarbons and simple straight-chain alcohols in this spectral range limits its application. On the other hand, the high molar absorbance coefficient of UV-Vis spectroscopy means that the sample with high aromatics content must be diluted before measurement, which limits its application in the online analysis of oil products to a certain extent. The advantage of UV-Vis method is that it has high sensitivity and the general substance can be measured $10^{-3} \sim 10^{-6}$ mol/L, so it is more suitable for the online determination of trace components.

In petrochemical enterprises, there are mainly two methods for online detection of H_2S and SO_2 ratio in the tail gas of sulfur recovery unit: gas chromatography and UV spectroscopy. UV spectroscopy is widely used due to its advantages of simplicity, high efficiency, short response time (a few seconds), and low maintenance cost [80]. Traditional online UV spectrometers (such as AMETEK's 880-NSL online gas analyzer) mostly employ non-dispersive filter mode. Four UV filters are 232 nm, 280 nm, 254 nm, and 400 nm, which are, respectively, used to measure the concentration and reference of H_2S , SO_2 , and S vapor. The reference datum is mainly used to compensate and correct the influence of unclean quartz window, change of light intensity, and other interference on the measurement accuracy. The measurement of H_2S and SO_2 will be disturbed by the presence of gases such as COS and CS_2 in the tail gas. Most of the modern online UV gas analyzers use the dispersive spectrometers (such as the 942-TG analyzer by Galvanic Company of Canada) to measure the UV spectrum of the whole band with a higher resolution (less than 1 nm). Combined with chemometrics methods (such as PLS), the influence of these interfering gases can be eliminated to a large extent. The accuracy of sulfur ratio measurement was significantly improved. At present, the online sulfur ratio analyzer has been widely used in the Klaus method sulfur recovery device, which plays a positive role in improving the conversion and recovery rate of sulfur as well as in environmental protection, energy saving, and emission reduction.

With minor modifications, these online instruments can also be used to analyze H_2S concentrations in natural gas, HCl purity and residual chlorine concentrations in the chlor-alkali industry, and SO_2 and NO_x concentrations in flue gas continuous emission monitoring systems (CEMS). It can also be used to monitor trace amounts of explosive aromatic organic compounds, such as toluene gas content in the production workshop. In addition, in the field of oil refining, online UV spectroscopy is also

used for the online measurement of aromatics content in oil products, such as residual BTX in the process of aromatics extraction, and the online measurement of the color of oil products, such as lubricants and solvents.

In the field of online analysis of environmental water quality, for a specific river system, the composition of the substances contained in it generally does not change much. The relationship between water quality parameters such as COD and UV absorbance of water samples is used to establish a regression model, and the water quality parameters are calculated indirectly. This kind of instrument has the characteristics of simple structure, fast real-time response, no secondary pollution, small maintenance, and so on, and is gradually recognized and selected by people. The development of UV spectrometers for the determination of COD in water quality can be divided into three stages: single wavelength, multi-wavelength, and continuous full spectrum. The single-wavelength method uses the absorbance value and COD value of the water sample at 254 nm to establish a regression curve. Due to the diversity and complexity of water components, especially the different components of organic matter in different water systems, their all-band UV absorption spectra are significantly different, and not all the maximum absorption wavelengths appear at 254 nm. Therefore, the applicability of the single-wavelength method is poor, and satisfactory results can not be obtained. Multi-wavelength analysis usually adopts dual wavelength. In addition to 254 nm, another wavelength such as 550 nm or 546 nm is selected for turbidity compensation. At present, some commercial online water quality analyzers adopt dual wavelength measurement. The use of full spectrum (200–750 nm) combined with chemometrics (such as PLS or artificial neural network, ANN) can more comprehensively reflect the internal information of water COD, and establish a multivariate correction model for specific measured water system, which can more accurately estimate the COD value of water [81, 82].

In papermaking enterprises, UV spectroscopy has been applied to the online detection of major components in the causticizing process of katerite pulping using ATR probe [83]. The purpose of black liquor recovery is to causticize the green liquor with high concentration of Na_2CO_3 to turn it into white liquor with high concentration of NaOH and Na_2S . The concentration of Na_2S , Na_2CO_3 , and NaOH in caustic process can be quantitatively analyzed by means of chemometrics methods. In addition, UV combined with ATR method can also be used to determine the composition of lye, the total content of dissolved solids in the black liquor, and the content of dissolved solids in the process of evaporation and concentration of black liquor.

In pharmaceutical enterprises, the concentration of active components in the dissolution process of drugs can be measured in real time by means of CCD array detector UV spectrometer combined with inserted optical fiber probe and multivariate calibration methods such as principal component regression (PCR) or PLS, so as to conduct detailed and accurate monitoring of the dynamic dissolution process of drugs [84–86]. A single spectrometer can simultaneously monitor the dissolution process of multiple drugs, or the dissolution of the same drug under different conditions, by means of a fiberoptic multi-channel switching device or a planar CCD detector. Another important application of online UV spectroscopy is to monitor

the concentration of cleaning solvents during the cleaning of batch reaction vessels (pharmaceutical, food, and beverage) to judge whether the vessel has been completely cleaned.

In the process of drug synthesis, the reaction system is measured online by UV spectroscopy, and the whole chemical reaction process (such as reaction kinetics and reaction mechanism), physical dissolution, and adsorption process, etc., are studied by chemometrics (three-dimensional data discrimination methods are often used) to provide useful information that is difficult to obtain by other methods [87–89].

2.6 Molecular Fluorescence Spectroscopy

Fluorescence is the emission light, and it is a photoluminescence phenomenon. Molecules are excited after absorbing light and then emit the light at the same or longer wavelengths than the absorbance light, which is called the phenomenon of photoluminescence. When the substance is in the ground state, electrons can jump to the excited state after absorbing light, and then return to the ground state. The emitted light is called fluorescence. But when the excited electrons move into the metastable triplet state, stay there for a while, and return to the ground state, the light emitted is called phosphorescence. Fluorescence analysis is a qualitative or quantitative analysis method based on the emitted fluorescence which can reflect the properties of the compound after the compound itself is irradiated by excitation light.

Fluorescence usually occurs in molecules with rigid and planar π -electron conjugated systems. With the increase of π -electron conjugation degree and molecular flatness, the fluorescence intensity increases, resulting in the corresponding red shift of the spectrum. The shape and intensity of the fluorescence spectrum also change with the increase of the number of benzene rings of aromatic hydrocarbon. Therefore, fluorescence spectroscopy is one of the special measuring methods to provide the distribution and concentration of aromatic components.

Molecular fluorescence analysis is characterized by its high sensitivity, with the minimum detection limit between 1 and 100 ppb, and even up to 0.01 ppb for substances with high fluorescence efficiency. Fluorescence analysis is about two orders of magnitude more sensitive than photometric methods. For example, for the determination of 3, 4-benzopyrene, the detection limit fluorescence analysis is on the order of ppb, whereas UV-Vis spectrophotometry is only on the order of ppm.

Since the 1980s, with the introduction of electronics, microprocessor, laser, and optical fiber, the progress in theory and application of fluorescence analysis has been promoted, and new technologies and methods such as synchronous fluorescence, three-dimensional fluorescence, time-resolved fluorescence, laser-induced fluorescence, dynamic fluorescence, and fluorescence imaging have emerged. Fluorescence analysis continues to develop toward the direction of real-time, trace, high efficiency, micro, in situ, and automation, and its application scope covers many fields such as agriculture, industry, environment, material science, life science, public security, and food engineering [90]. It is particularly worth mentioning that the fluorescence

analysis method of high sensitivity and high selectivity is bound up with the life sciences. New demand of life science constantly pushes the continuous development and improvement of new fluorescence analysis method in the instrument, method, and data processing. And its application in life science is also more and more extensive and in-depth [91].

2.6.1 Three-Dimensional Fluorescence Spectroscopy

Traditional fluorescence analysis methods pay more attention to the quantitative and qualitative analyses of substances by two-dimensional fluorescence spectroscopy, but fluorescence intensity is a function of excitation wavelength and emission wavelength. Such scanning results only at a certain excitation or emission wavelength cannot completely describe the fluorescence characteristics of substances, and it is difficult to provide complete information. Therefore, more and more attention has been paid to the study of three-dimensional fluorescence spectroscopy in recent years. Three-dimensional fluorescence spectroscopy can obtain the fluorescence intensity information when the excitation wavelength and emission wavelength change simultaneously, and can obtain more complete spectral information than conventional fluorescence spectra. Using the spectral information combined with chemometrics can accomplish more complex quantitative and qualitative analysis tasks in multi-component mixture systems.

Petroleum is mainly composed of hydrocarbons (95 ~ 99%) generated by hydrocarbon synthesis and some non-hydrocarbon components. Among them, aromatic hydrocarbons, especially polycyclic aromatic hydrocarbons, have high fluorescence efficiency. Fluorescence, therefore, have long been used in the oil and gas exploration process. The earliest way of the traditional fluorescence logging was not used to the formation of fluorescence spectra and just stay on the stage of macroscopic observation the fluorescence intensity. Emitting light colors was used to determine oil and gas composition, and the oil and gas content was judged through the luminous intensity. Subsequently, through one-dimensional (both excitation and reception are single wavelength), two-dimensional (single-wavelength excitation and receiving wavelength change), and three-dimensional (both excitation and receiving wavelength can change) changes, three-dimensional fluorescence has been applied in logging field so far. Three-dimensional fluorescence spectroscopy can not only give the oil concentration, fluorescence contrast level, oil index, and other parameters of the sample but also can be used for accurate identification and analysis of drilling fluid additives and crude oil by means of chemometrics.

Molecular fluorescence spectroscopy also plays an important role in the on-site monitoring of oil spill at sea and on land. Combined with the pattern recognition method in chemometrics, the types of oil spill (including crude oil, diesel oil, fuel oil, lubricating oil, gasoline, and edible oil) can be quickly identified [92, 93].

In recent years, the analysis strategy of three-dimensional fluorescence spectroscopy combined with multidimensional correction method has been more and

more widely applied in the fields of medicine, medicine, food, and environment. It can conduct direct, rapid, and simultaneous quantitative analysis of multi-component targets in complex test objects [94–96].

In terms of medical application research, Gu et al. used excitation emission matrix fluorescence combined with PARAFAC and other algorithms to carry out quantitative analysis of metoprolol and its metabolite-hydroxymetoprolol in plasma. In the case of plasma background interference and overlapping spectra of the two target analytes, this method achieves simultaneous quantitative analysis of the two components. Moreover, it is a simple and rapid method that requires simple dilution rather than complex pretreatment of plasma [97]. Ouyang et al. used three-dimensional fluorescence spectroscopy combined with self-weighted alternating trilinear decomposition (SWATLD) algorithm to simultaneously determine sulphiride and amisulpride, two antipsychotics in human serum samples. The results show that this method can still obtain satisfactory quantitative prediction results even if there is serious spectral overlap between target analytes, between analytes and background, and between analytes and other unknown disturbances [98].

In terms of the application research of environmental analysis, Qing et al. used SWATLD algorithm and three-dimensional fluorescence spectroscopy to simultaneously determine plant growth regulators 2-naphthaloxyacetic acid and 1-naphthaloxyacetic acid methyl ester in soil and sewage [99]. Manuel et al. used three-dimensional fluorescence spectroscopy and multivariate curve resolution alternating least squares (MCR-ALS) second-order calibration method to realize trace detection of the toxic substance tributyltin in water [100].

In terms of drug application research, Wang et al. quantitatively analyzed the contents of umbellolactone and scopolamine in Chinese traditional medicine, radix angelicae pubescentis, and Tibetan medicine, Saussurea mongolicus based on ATLD and three-dimensional fluorescence spectroscopy, indicating that such analysis strategies can accurately quantitatively analyze the contents of active components in the complex system of Chinese traditional medicine [101].

In terms of food application research, Zhong et al. used three-dimensional fluorescence spectroscopy and ATLD to quantitatively analyze the residual contents of thiabendazole and fuberidazole in red wine [102]. As highly efficient and broad-spectrum fungicides, thiabendazole and fuberidazole are widely used in the production, storage, and preservation of vegetables, fruits, and other crops. If improperly used, they will remain in grapes and grape products and enter the human body through the food chain, causing certain harm to human health. The traditional methods for determination of pesticide residues in grapes and wine are high-performance liquid chromatography (HPLC) or HPLC-mass spectrometry (HPLC-MS), which is complicated in pretreatment and time-consuming in analysis process. Because of the convenience and quickness of three-dimensional fluorescence spectroscopy, it can be used for simultaneous screening of thiamendazim and maisuillin in a large number of red wine samples. Zhu et al. combined three-dimensional fluorescence spectroscopy with PARAFAC and BP neural network (BP-NN) to establish years' identification model for clear flavor liquor, and the average identification accuracy reached 95% [103].

2.6.2 *Laser-Induced Fluorescence Spectroscopy*

Different from ordinary fluorescence analysis methods, laser-induced fluorescence (LIF) uses laser as the excitation source. However, the LIF process is a wavelength absorption and conversion process but not a scattering process. Because of high laser brightness, good monochromatism, and no stray light, LIF technology has the advantages of low detection limit and high sensitivity. Laser light source and weak signal detection technology make LIF spectroscopy reach the sensitivity limit of spectral analysis, which makes it have important applications in life science, environmental science, and other fields.

Compared with ordinary fluorescence spectroscopy, LIF is currently the best choice for in situ online analysis based on fluorescence technology [104]. At present, laser fluorescence radar is one of the most promising methods for detecting oil spill in the sea. The SLEAF system developed by Environmental Technology Centre of Canada and the AOL system jointly developed by NASA and NOAA are mature systems for the detection of Marine oil spill, both of which are developed based on LIF spectroscopy technology. The identification and quantitative analysis of oil pollutants, polycyclic aromatic hydrocarbon pollutants and organic pesticide pollutants in water or soil can be realized through the determination of aromatic hydrocarbons and their derivatives and organophosphorus groups in pesticides by LIF, without sampling and sample separation [105, 106]. The LIF spectroscopy technology can be used to dynamically remote measure the pollution status of a large area of water in real time, such as the parameters of dissolved organic matter (DOM), turbidity, and chlorophyll a concentration in water. After fusion with the GPS positioning system information, the pollution status distribution map of water can be drawn directly. In the field of agriculture, the telemetry of field crops by LIF spectroscopy can judge the growth state and nutrient condition of crops, and then guide agricultural production.

Hu et al. combined LIF spectroscopy with CNN to quickly identify the source of mine water inrush. The rapid identification and classification of mine water inrush are of great significance for underground flood prevention and control work [107]. In addition, the combination of LIF spectroscopy and chemometrics has also been used in the classification of plastics, identification of edible oil types, identification of counterfeit wine, and diagnosis of diseases [108–112].

In recent years, the combined technology of LIF and laser-induced breakdown spectroscopy (LIBS) has attracted more and more attention. For example, the combined technology realizes the highly sensitive detection of trace content of lead elements in water environment [113, 114].

2.7 Low-Field NMR Spectroscopy

The object of nuclear magnetic resonance (NMR) research is the nucleus of magnetic moment not equal to zero. When this kind of nuclei in external magnetic field, it can produce energy level splitting. If we use a particular frequency of radio source for irradiation samples, which makes its energy equal to the energy level difference, the

nucleus can be the transition between energy levels, and this phenomenon is called NMR. According to quantum mechanics, when the mass number or charge number of the nucleus is odd, and its magnetic moment is not equal to zero, which indicates the existence of nuclear magnetic resonance phenomenon. Proton hydrogen ^1H is a common element in organic molecules. Its mass number and charge number are odd, and its abundance is very large in nature. Therefore, proton nuclear magnetic resonance (^1H -NMR) is the most studied, the most sensitive and the most widely used NMR spectroscopy. In addition, there are ^{13}C , ^{19}F , and ^{31}P NMR spectroscopy.

NMR instrument is mainly composed of magnet, radio source, probe, receiver, and other parts. The function of magnets is to provide a stable high-intensity magnetic field. Radio frequency sources are used to supply fixed frequency electromagnetic radiation. The sample probe allows the sample tube to be fixed at a defined position in the magnetic field, and the receiving coil and transfer coil are also mounted in the sample probe to ensure that the position of the sample with respect to these components remains unchanged. The new spectrometer produced since the 1970s basically used radio frequency pulse to measure nuclear magnetic resonance. The radio frequency pulse is equivalent to a multi-channel transmitter, which simultaneously transmits a variety of frequencies to make the nuclei on different groups resonate at the same time, and the free induction attenuation signal (FID) of the multiple spectral lines of the nucleus mixed is obtained. FID is a time domain function which is transformed by using Fourier transform into a frequency domain function, and its measurement speed, sensitivity, and signal-to-noise ratio are improved remarkably.

The industrial online NMR analyzer was a time domain (TD) NMR analyzer used in petrochemical polypropylene and polyethylene installations in the mid-1990s. This kind of time domain or low-resolution NMR instrument is relatively simple in structure. Operating frequency is only 20 MHz. It can only give the total proton strength, relaxation time, and their distribution, and can online analyze polymer powder (such as polypropylene) melt flow rate, ethylene content, isotactic, crystallinity, density, and other physical and chemical indicators. The initial amplitude of NMR signal is proportional to the number of measured nuclei (hydrogen nuclei) in the sample to be measured, and the attenuation rate of the signal is related to the relaxation time of the sample. That is, it is related to the group and environment of the measured nucleus in the sample. For example, in isotactic and interisotactic polypropylene, the signal decays rapidly (T_2 time is short), whereas in atactic polypropylene, the signal decays much more slowly (T_2 time is long). More than 100 of these analyzers are currently in operation in industrial installations around the world.

Oil field logging is another important application of this kind of instrument. At present, NMR technology is increasingly used in oil production logging, and multi-dimensional NMR and imaging NMR have also begun to enter the logging field [115]. Online NMR instrument can provide geological parameters related to reservoir physical properties and reservoir fluid properties, such as effective porosity, movable fluid content, and oil saturation, so as to create conditions for timely and effective evaluation of the reservoir during the drilling process, and realize on-site interpretation and evaluation in time.

Desktop low-field NMR instruments are also widely used in quality control and laboratory research and development [116]. The NMR instrument time domain

obtains NMR signals which decay over time, and its intensity is proportional to the content of hydrogen, due to different morphology of hydrogen atoms with different relaxation time. Therefore, you can adopt different pulse sequence according to the requirement to distinguish the different morphology of hydrogen in the signals and get different morphology of hydrogen content. Combined with the method of chemometrics, the data of physical and chemical properties can be obtained. For example, the oil and water content of oil seeds, the fat and water content of milk powder, the water content of water-injected pork, the soluble and isometric xylene in polypropylene, the quality of fruit, and the lean meat, fat, and fluid content of living mice were determined [117–119].

Combining chemometrics with online NMR for large process industries emerged in the mid-1990s. At that time, an Israeli company developed a permanent magnet technology capable of producing a uniform magnetic field of 1.4 T, and developed an NMR instrument suitable for industrial online analysis. The instrument can obtain 60 MHz ^1H NMR spectra. It can give information about the chemical shift of hydrogen in the sample. At present, this technology has been applied to some extent in the petrochemical field, most of which are for the purpose of feasibility testing.

Theoretically, NMR can also measure the physical and chemical properties of oil products that can be measured by NIR and other molecular spectrometers [120–122]. However, because the sample must enter within the probe in the magnetic field, NMR not like molecular spectroscopy can be used in a fiber optic instruments and measuring sample separation device. Therefore, for industrial large-scale device or laboratory reaction kettle, the sample was introduced into the probe only through the bypass, and thus it cannot be achieved in the true sense of in situ online analysis. With automatic sample switching and cleaning systems, online NMR can also be used to measure multiple logistics, but this can affect the measurement speed to some extent. Large temperature differences between logistics can cause magnetic field fluctuations, which can significantly reduce the reproducibility of the analyzed data. Solid molecules such as wax and bitumen in heavy oils and paramagnetic substances such as iron also significantly affect the measurement of NMR spectra. Although the resolution of NMR spectroscopy is higher than that of NIR spectroscopy, for the quantitative and qualitative analyses of complex mixtures, multivariate calibration methods and pattern recognition methods are still needed, and the modeling task is still heavy.

In addition, compared with NIR spectrometers, NMR spectrometers has considerable cost in price, operation, and maintenance, and requires higher technical level of users and maintenance personnel, which brings some difficulties to the practical application of NMR technology in industry. Therefore, the analytical object of online NMR technology is limited to light oils such as naphtha, gasoline, and diesel. Its meas conventional physicochemical properties such as octane number, cetane number, distillation range, and group composition.

2.8 Terahertz Spectroscopy

Terahertz ($1 \text{ THz} = 10^{12} \text{ Hz}$) refers to the electromagnetic wave frequency in the range of $0.1\sim 10 \text{ THz}$ ($3.3\sim 333 \text{ cm}^{-1}$), located between infrared and microwave, in the transition stage from macroscopic electronics to microscopic photonics. Early terahertz had different names in different fields. In optics, it was called far infrared, while in electronics, it was called submillimeter wave, ultra-microwave. In photonics, lasers can be classified as continuous, semicontinuous, or pulsed depending on how they emit energy. In electronics, according to the shape of the signal, it can be divided into continuous wave and pulse wave. In addition to sine wave and a number of sinusoidal components of the continuous wave, others collectively are known as pulse wave.

Terahertz wave is between microwave and infrared. The source of terahertz wave can be obtained from optical methods and electronic methods. It can also be divided into continuous terahertz wave and pulse terahertz wave by using photonics and electronics methods for reference. The current research on terahertz wave technology is mainly in the form of pulse, and the research on continuous terahertz source is relatively few. Until the mid-1980s, infrared and microwave technologies on both sides of the terahertz band were relatively mature, but the understanding of the terahertz band was still very limited, resulting in the so-called “Terahertz Gap” [123].

Terahertz time domain spectroscopy (THz-TDS) and terahertz imaging are important methods and means in practical application of this technology. THz-TDS technology uses femtosecond laser pulses to generate and detect time-resolved terahertz electric field, and obtains spectral information of the samples through Fourier transform. Since the power of terahertz radiation is on the pW scale, which is smaller than the power of thermal background radiation. The thermal strain in the sample can be ignored. Terahertz spectroscopy techniques mainly include transmission, specular reflectance, diffuse reflectance, attenuated-total reflectance, and photopump-terahertz.

The terahertz spectrum of matter contains rich physical and chemical information, such as gas rotation, phonon vibration of condensed matter, and low-frequency vibration and rotation of biological macromolecules, which all respond in the terahertz band. Each molecule has a specific vibrational and rotational energy level, and usually the intramolecular vibration of material is mainly in the MIR band. However, the weak interactions between molecules (such as hydrogen bonds), skeleton vibrations (configuration bending) of macromolecules, rotation and vibrational transitions of dipoles, and low-frequency vibration absorption frequencies of crystal lattice are located in terahertz band. The molecular structure and related environmental information reflected by these vibrations have different absorption peaks in terahertz band. These spectral characteristics of organic molecules make it possible to use THz-TDS to identify the effects of structure, configuration and environment on the state of organic molecules. The study of the spectral properties of substances in this band is of great scientific significance and practical application value for the exploration

and comprehensive understanding of the structure and properties of substances and the interaction between molecules.

In recent years, with the continuous implementation of process analysis technology in foreign pharmaceutical enterprises, the application of terahertz combined with chemometrics in biomedical field has received more and more attention [124, 125]. TH_Z-TDS and terahertz imaging technology have made a lot of achievements and progress in the detection of drug components, the differentiation of isomers, the identification of drug polymorphs and pseudo-polymorphs, as well as the qualitative and quantitative analysis of mixtures. Some achievements have also been made in drug interaction, reaction mechanism, and reaction kinetics. For example, TH_Z-TDS has a high sensitivity to the crystal shape of the compound, which can reflect the phonon vibration mode in the crystal and effectively reflect the long-range ordered structure information of the crystal. Compared with X-ray powder diffraction, TH_Z-TDS has no preferred orientation problem. Different from IR and Raman spectroscopy, TH_Z-TDS mainly reflects the low frequency vibration of the whole molecule, the phonon vibration of the crystal, and the weak interaction between molecules such as hydrogen bonds. Although the molecular structure of different crystal forms is the same, the interaction between molecules leads to the different local environment inside the crystal. This changes the strength and location of the terahertz absorption peak, which will play a role in the drug synthesis, production, and storage process.

The unique advantages of terahertz wave (strong water absorption, non-destructive, and fast) make it also widely used in food quality and agricultural product quality inspection [126, 127]. For example, terahertz measurement of water content in food is to use the strong water absorption characteristics of terahertz wave. The quality of meat products can be analyzed by using the different wave absorption characteristics of lean meat and fat to terahertz. In addition, the low interference, non-ionizing properties of Terahertz can be used for rapid analysis of agricultural products based on imaging technology, such as real-time detection of moisture content in plants and foods. The terahertz spectroscopy also has a place in the study of gas spectroscopy because all or part of the rotational spectrum of gas molecules is located in the far-infrared region. Therefore, the chemical composition and concentration of mixed gases can be determined by TH_Z-TDS pulse with wide band. In the field of petrochemicals, some people have combined terahertz wave with chemometrics for quantitative and qualitative analyses of oil products, and obtained certain results. However, compared with NIR, MIR, and Raman spectroscopies, the advantages are not obvious at present, and there is a lack of application examples.

Terahertz spectroscopy, from the initial exploratory research, has gradually become a means of process analysis and detection, and gradually began to industrial applications. However, as a new analytical technique, there are still some problems and limitations. The existing terahertz radiation sources, detectors, and related components are complicated in structure, large in volume, and expensive, so it is difficult to be popularized and applied. Therefore, the miniaturization, low cost, and practicability of the instrument are the urgent problems to be solved in the application of this technology in practice. Due to the strong absorption of terahertz radiation by

water, the application of this technique in drug analysis in aqueous solution systems is limited. The spectral analysis and theoretical interpretation of terahertz spectroscopy are still in the initial stage of exploration, which needs to be further developed in the establishment of relevant theoretical models and simulation calculation.

2.9 Laser-Induced Breakdown Spectroscopy

Laser-induced breakdown spectroscopy (LIBS) is an elemental analysis technique based on atomic emission spectroscopy, also known as laser-induced plasma spectroscopy (LIPS). LIBS technology has many advantages, such as fast, multi-element simultaneous analysis, remote analysis, online analysis, and its applicability to extreme environments. It is called “chemical analysis star”. So far, LIBS technology has been successfully applied in many fields such as industry, medicine, military, archaeology, materials, space exploration, and so on [128, 129].

LIBS technology uses ultrashort-pulse laser to focus the sample surface (or inside the sample) to form a plasma, and then analyzes the plasma emission spectrum to determine the material composition and content of the sample. Its basic principle is as follows: (1) Pulse laser produced by the laser is focused on the surface of sample; (2) High-energy laser makes the surface melt and produce a large number of plasmas; (3) In the cooling process of the plasma, its coverage will decrease with the temperature continuously expand. Then, in the excited state of particles including atoms and ions will transition to the stability of the low level or the ground state; and (4) emission lines of a specific frequency are then generated during a transition. Different elements have different characteristic emission lines, so the types of elements can be analyzed according to different frequencies of emission lines, and the content of elements can be analyzed according to the intensity of the spectral lines.

The traditional univariate calibration curve method only uses a single characteristic spectral line corresponding to the element for quantitative analysis. However, due to the changes of self-absorption effect, element mutual interference, plasma physical parameters, and complex sample matrix effect and other factors, the position and strength of the characteristic spectral line often deviate from the theoretical value. Therefore, the accuracy of the relationship between the concentration and the intensity of the characteristic line established by the univariate curve is significantly reduced. The combination of LIBS and chemometrics can effectively solve the problems of spectral matrix effect deduction, overlapping peak resolution, and self-absorption effect correction, so as to improve the repeatability and accuracy of qualitative and quantitative analysis of LIBS technology to a large extent [130, 131].

As an important part of the active ingredients of traditional Chinese medicine, element is an indispensable characteristic parameter for the quality control of traditional Chinese medicine. In recent years, LIBS technology has been increasingly applied to the field of traditional Chinese medicine. In addition to the determination of element content in traditional Chinese medicine, it has also been used to identify the origin, authenticity and variety of traditional Chinese medicine. Wang et al. [132]

used LIBS technology combined with PCA and ANN technology to identify three kinds of medicinal materials, *angelica pubescens*, *Codonopsis pilosula*, and *Ligusticum wallichii* from different geographic origins, and proved that LIBS technology was an effective tool for the identification of traditional Chinese medicine. Zhao et al. [133] used LIBS technology combined with characteristic band extraction and chemometrics to identify different degrees of sulfur-fumigated *fritillaria thunbergii*, providing a basis for the identification of sulfur-fumigated Chinese herbal medicine, and contributing to the establishment of quality detection and grading and evaluation system of Chinese herbal medicine.

In the field of agricultural products and food analysis, LIBS technology is widely used in the detection of trace elements in food, quality control of products in the production and processing links, and safety assessment of food [134, 135]. Based on the content differences of Zn, Mg, Ca, Na, and K elements in beef, pork, and chicken, Bilge et al. used LIBS technology combined with PCA to identify meat types, and employed PLS method to qualitatively identify adulterated pork and chicken in beef samples [136]. Wang et al. used LIBS combined with discriminant analysis to identify 6 kinds of tea, including Longjing green tea, Mengding yellow bud, white tea, Tieguanyin, Wuyi black tea, and Pu 'er tea. Mg, Mn, Ca, Al, Fe, and K were selected as analysis indexes, and the classification accuracy of the validation set was 95.3% [137].

LIBS technology is very suitable for online analysis of element content in substances and their related physicochemical parameters [138]. For the composition of molten metal, LIBS has achieved various online analysis in the industrial field, including continuous online analysis of C, S, P, Si, and Mn in the molten iron of blast furnace iron drain, online analysis of C, Si, Mn, Cr, and Ni in molten steel of converter or AOD furnace, online analysis of P, Mg, Fe, Al, and Si in raw ore, concentrate, and tailings in the flotation process of phosphate ore, online analysis of Al, Cu, Fe, Mg, Mn, Si, and Cr in the filtrate liquid aluminum and online sorting and recycling of scrap metal [71–73]. LIBS technology can conduct online analysis of organic elements C, H, S, O, and inorganic elements Si, Ca, Mg, Fe, Al, as well as parameters such as ash content, calorimetric value, and volatile content of coal. Online analysis of coal quality is crucial to the safe, efficient, and economic operation of large boilers such as thermal power plants [139]. LIBS technology can also be used for online analysis of industrial production processes such as cement and potash fertilizer, and the elements analyzed include Ca, K, P, Al, Si, Fe, Mg, etc. [140].

Fast and accurate diagnosis of cancer is an important task to clinical medicine. Because of its advantages of simple equipment without pretreatment, microdamage, and real-time online detection, as well as the concentration difference of trace elements in normal cells and cancerous cells, LIBS technology combined with chemometrics methods is expected to become a powerful tool in vivo cancer diagnosis and classification [141–144].

2.10 Spectral Imaging

The imaging information is mostly shown as the form of wave. Wave is divided into shear and longitudinal wave, which can be expressed uniformly by wavelength. As shown in Fig. 2.9, longitudinal wave includes ultrasonic wave, etc., which is commonly used in B-mode ultrasound, color ultrasound, and photoacoustic imaging. The shear wave has electromagnetic wave, material wave, and so on. The electromagnetic wave includes radio frequency, microwave, infrared ray, visible light (including fluorescence, phosphorescence, etc.), UV ray, X-ray, and so on. Matter waves, which include α rays, β rays, or electrons, are short and can be distinguished up to angstroms.

Spectroscopic techniques (such as NIR, IR, Raman, terahertz, fluorescence, and LIBS) measure the average spectrum of a certain point (or a small region) of the sample, so that the average result of sample composition or properties is obtained, which is very suitable for the analysis of homogeneous substances. In order to obtain the spatial and concentration distributions of different components in non-uniform mixed samples, spectral imaging techniques, such as infrared, NIR, Raman, fluorescence, terahertz, and LIBS spectral imaging, are required [145–147]. Spectral imaging technology combines traditional optical imaging and spectral methods to obtain the spectrum of each point in the sample space at the same time, so as to further obtain the composition and structure information of each point in the space (Fig. 2.10) [148].

Previously, spectral imaging technology was mostly applied in the field of remote sensing, which combined imaging with spectroscopy. For each space image unit, dozens to hundreds of dispersion wavelength bandwidth of pixel about 10 nm continuous spectrum was formed when the detection of object space characteristics. The spectral range is in UV-Vis~NIR area (0.4~2.5 μm), so as to achieve the purpose of identifying the earth surface material directly from space. According to the different

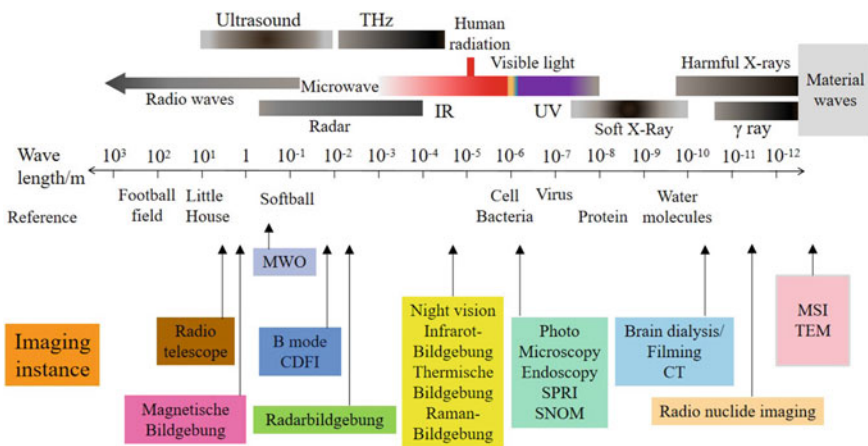


Fig. 2.9 Wave wavelength corresponding to the imaging

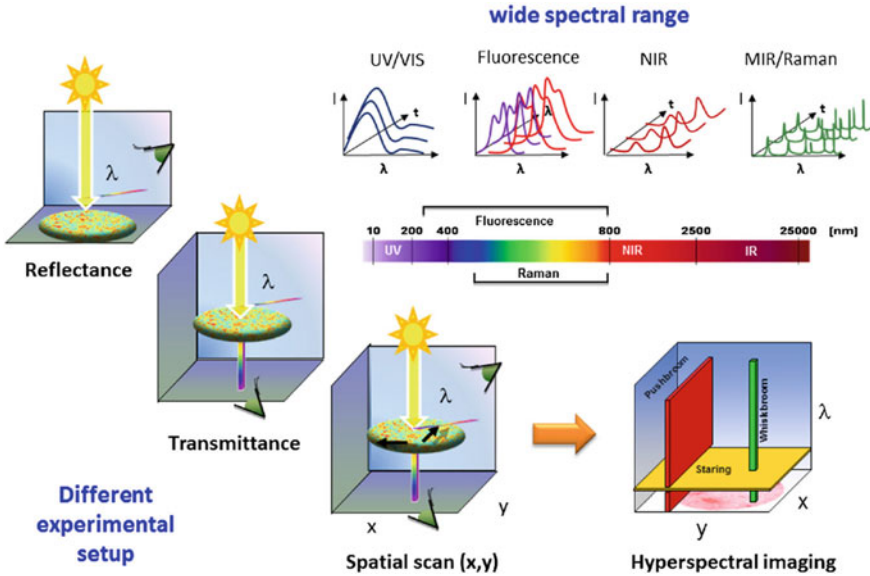


Fig. 2.10 Schematic diagram of spectral imaging techniques with different spectral ranges and different measurement methods [148]

spectral resolution, it can be divided into multi-spectral imaging and hyperspectral imaging [149].

Remote sensing spectral imaging is realized by the translational motion of the flight platform (such as aircraft and satellite) combined with the imaging spectrometer placed on the flight platform in a certain operating mode. The commonly used operating mode is whiskbroom and pushbroom. The whiskbroom imaging spectrometer uses the rotating scanning mirror of the motor and the flying platform in a forward motion to complete the two-dimensional space imaging, and the spectrum of each instantaneous field of view pixel is obtained by the line array detector. The pushbroom imaging spectrometer uses a planar array detector perpendicular to the direction of motion to complete two-dimensional space scanning in the forward motion of the flying platform, whose spatial scanning direction is the movement direction of the remote sensing platform. Spectral resolution and spatial resolution are two key technical indicators of remote sensing spectral imager. In order to obtain high-precision remote sensing monitoring results, vehicle-mounted spectral imaging system can be selected.

Remote sensing spectral imaging technology has been widely used in the fields of geology, agriculture, ocean, atmosphere, and military. It plays an increasingly important role in geological prospecting and mapping, atmospheric and environmental monitoring, agricultural and forest investigation, marine biological research, and other fields. In recent years, spectral imaging technology has gradually entered the laboratory and production site, and become a platform technology in analysis and detection. Spectral imaging has also been more and more replaced by the

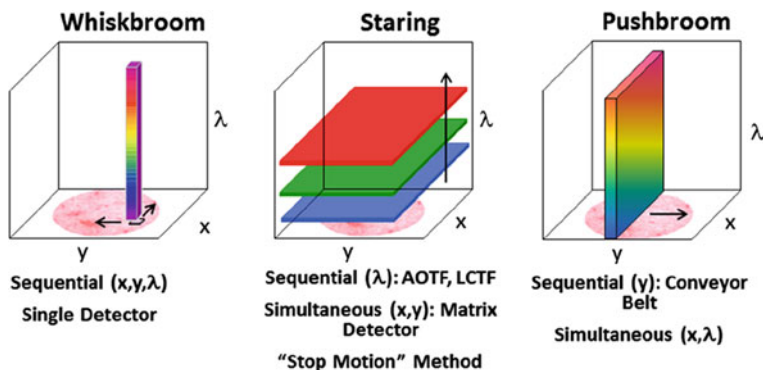


Fig. 2.11 Implementation of spectral imaging [148]

term of chemical imaging (CI). Spectral chemical imaging technology is currently becoming a complementary technology to traditional spectroscopy, and is gaining wide attention and practical application in the fields of pharmaceutical, agricultural, and food [150].

As shown in Fig. 2.11, there were three ways to implement spectral imaging: first, the sample is placed on the movable slide on the stage, and is moved along transverse and longitudinal directions to capture spectra point by point. The three-dimensional spectral image is composed. This type of imaging is called whisk broom imaging mode, mainly applied at spectral microscopic imaging. This way used a single point detector with a high spatial resolution, but the measurement time is long. Sometimes it takes several hours to test a sample. For the samples moving on the conveyor belt, the line array detector is often used, which is called push broom imaging mode. In recent years, with the emergence and application of liquid crystal tunable filter (LCTF) and acousto-optic tunable filter (AOTF) technology, and gradual transition of infrared focal plane array (FPA) detector from military to civilian, staring spectral imaging (STARING) is increasingly used in process analysis technology.

The data array obtained by spectral imaging is gained by scanning each space point of the sample at multiple discrete or continuous wavelengths. In fact, it is a three-dimensional data array composed of two-dimensional space and one-dimensional wavelength, which is called hypercube. As shown in Fig. 2.12, this hypercube matrix can be regarded as a composition of a series of spatially resolved spectra (called pixels) or a series of spectral resolved images (called image planes). Selecting an independent pixel will obtain the continuous spectrum of a specific spatial point of the sample. Similarly, selecting an image plane will obtain the intensity response (absorbance) of all spatial points of the sample at a specific wavelength, namely, the spectral image. Through spectral library retrieval or modern pattern recognition technology, the composition and distribution information of sample space can be identified, which can be expressed intuitively and clearly by color view, that is, chemical image.

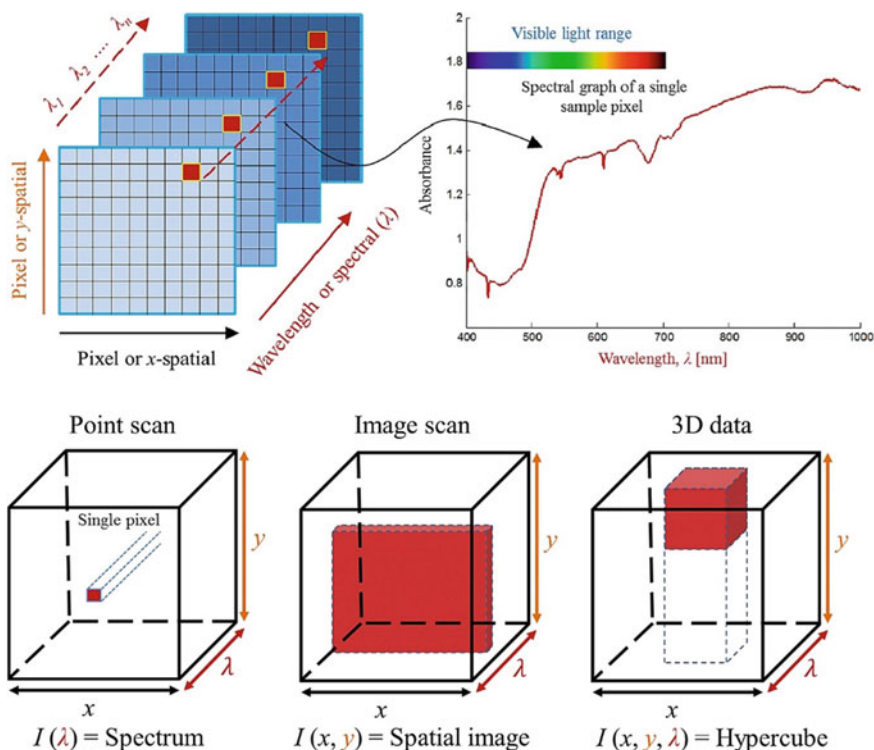


Fig. 2.12 Schematic diagram of the square array obtained by hyperspectral imaging (HSI). **a** HSI images as a function of wavelength, and **b** HSI images data structure [151]

Spectral imaging has a large amount of data. For example, the data array obtained by a 256×256 pixels array with 150 wavelength points contains 65,536 spectra, and each spectrum contains 150 wavelengths, and the spectral image of a sample has nearly one million data points in total. To mine useful information from such an information-intensive data array, that is, to transform spectral imaging into real chemical imaging, some modern chemometrics methods, such as data preprocessing and pattern recognition methods, are needed.

Spectral image data processing usually consists of the following three parts: (1) data preprocessing; (2) pattern recognition; and (3) chemical visualization and statistical analysis of data. Spectral image data preprocessing is similar to that of NIR spectroscopy, which aims to eliminate the influence of non-chemical information (such as scattering, noise, drift, etc.). The methods involved include smoothing, differential, standard normal variation (SNV), and multiple scattering correction (MSC). In general, spectra also need to be corrected by deducting the dark response, which is the detector response value after the light source is turned off and the lens is shielded.

The purpose of pattern recognition is to identify image regions with similar spectral characteristics. Common methods include unsupervised and supervised pattern

recognition methods [150]. Unsupervised methods, such as (PCA, K-means clustering, and fuzzy clustering, do not require training sets. Some monitoring methods require a set of training sets in advance, such as LDA, PLS-DA, ANN, and SVM. The training set data can be obtained from an image data array based on prior knowledge of the sample composition, from the imaging region of a component identified by an unsupervised method or using spectral imaging data from pure material samples. Before pattern recognition, the three-dimensional array needs to be first expanded into a two-dimensional spectral matrix according to the spectral direction, namely, each pixel corresponds to a spectrum. After the completion of pattern recognition, it can be restored to the original three-dimensional data array (Fig. 2.13). In addition, multidimensional analysis methods such as PARAFAC and multidimensional partial least squares (N-PLS) are also used to process spectral image data.

Chemical visualization and statistical analysis are to convert the above classification and recognition results into visual chemical composition distribution maps, which usually use gray or color maps with intensity scales to describe the comparison of chemical composition among image pixels. Histogram can be used to calculate the distribution number of pixels of chemical components in order to obtain quantitative information. Most of the commercialized spectral imaging instruments are equipped with image data processing software, which contains all the functions mentioned above and can be easily operated by the user.

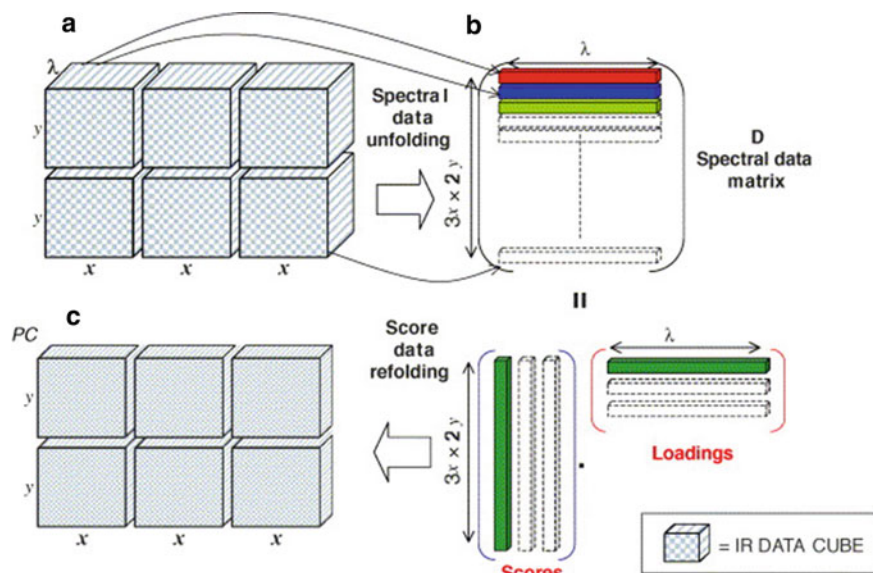


Fig. 2.13 Concatenation of IR images and principles of the unfold principal component analysis: **a** IR image concatenation; **b** principal component analysis; **c** creation of the scores image [152]

Spectral chemical imaging technology has been studied and applied in the fields of agriculture, food, medicine, and clinical medicine [153–155]. In the pharmaceutical field, NIR chemical imaging can be used to achieve high-throughput analysis of drugs. For example, NIR spectroscopy combined with chemometrics can be conveniently and intuitively used for the identification of counterfeit and substandard drugs, and can also be used for the identification and analysis of mixing uniformity, trace pollutants on drugs, and a small amount of degradable substances of active ingredients. In recent years, element imaging technology based on laser-induced breakdown spectroscopy (LIBS) has attracted much attention, which can realize spatial distribution imaging of elements in samples. It has a broad application prospect in biomedical, industrial production, and environmental detection, especially in drug metabolism and pathological analysis of biological tissue [156].

In the field of agriculture and food, spectral imaging combined with chemometrics can accurately measure the composition (such as water, protein, and starch) in single-grain grains, which can overcome the measurement error caused by the non-uniformity of samples by traditional spectral methods, and can also be used for the detection of insect pests inside grains [157]. In the application of fruit, it can detect the defects of fruit such as damage, bruise, and wormhole, and trace surface contaminants such as feces and organic residues. It can also be used for quality analysis, such as the hardness of peach and the total soluble solid content of strawberry. In addition, it can also be used for the analysis of various components in tobacco, compound feed, bacteria, and parasites in food, as well as the marbling grade of pork.

NMR imaging and terahertz imaging technology can not only obtain the surface characteristics of the sample but also can detect the internal structure, material composition, and its spatial distribution of the sample to achieve functional imaging. For example, NMR imaging can be used to detect the water distribution and proton mobility of fruits and vegetables in the storage process, observe the changes between the various organizational structures of fruits and vegetables, and judge the maturity of fruits and vegetables, as well as the degree of damage and deterioration, so as to provide a theoretical basis for the preservation of fruits and vegetables.

Terahertz imaging technology includes terahertz scanning imaging, terahertz real-time imaging, terahertz tomography, and terahertz near-field imaging, which can be used in biomedical, material quality inspection, and safety inspection. Tablets of coating membrane is one of the important factors affecting drug bioavailability. The features of coating membrane including thickness, structure, integrity, and consistency are very important to the quality of the drug. Using terahertz imaging of film-coated tablet of membrane, sugar pill, multi-layers, and gelatin soft capsule for three-dimensional imaging can obtain the thickness statistical distribution of the outer and inner membranes. In the field of biomedical science, due to the vigorous cell metabolism in the cancerous area, the water content in the biological tissue increases. Due to the strong absorption of terahertz by water, the cancerous area from the normal area can be distinguished by terahertz imaging.

In addition, terahertz has strong penetrability. Most non-polar materials do not absorb terahertz waves significantly. Terahertz waves can penetrate materials such as

ceramics, fats, fabrics and plastics with very little attenuation. Therefore, terahertz has a very good application prospect for safety inspection in public places [158].

In addition to the fields of medicine, agriculture, and food, spectral imaging technology has also been used in the identification of physical evidence, cultural relics, materials, geology, chemical synthesis, and biomedical fields. Some application examples include online identification of discarded plastic and paper, identification of soil composition of drilling cores, kinetic studies of chemical processes such as epoxy resin curing, clinical medical research (such as providing hemodynamic information related to functional brain activity), and disease diagnosis (such as cardiovascular disease and breast tumors) [159–163].

References

1. Pneg Y. *Nondestructive and rapid Raman spectral detection technology for edible agro-food quality*. Beijing: Science Press; 2019.
2. Chen Z, Lovett D, Morris J. Process analytical technologies and real time process control—a review of some spectroscopic issues and challenges. *J Process Control*. 2011;21:1467–82.
3. Rolinger L, Rüdtt M, Hubbuch J. A critical review of recent trends, and a future perspective of optical spectroscopy as PAT in biopharmaceutical downstream processing. *Anal Bioanal Chem*. 2020;412:2047–64.
4. Gendrin C, Roggo Y, Collet C. Pharmaceutical applications of vibrational chemical imaging and chemometrics: a review. *J Pharm Biomed Anal*. 2008;(48):533–53.
5. Rateni G, Dario P, Cavall F. Smartphone-based food diagnostic technologies: a review. *Sensors*. 2017;17:1453–553.
6. Ozaki Y, Huck C, Tsuchikawa S, et al. *Near-infrared spectroscopy: theory, spectral analysis, instrumentation, and applications*. Springer;2021.
7. Chu X, Zhang L, Yan Z. *Modern progress analytical technology: current development and future prospects*. Beijing: China Machine Press; 2016.
8. Chu X, Li S, Zhang T. *New development of modern process analytical technology*. Beijing: Chemical Industry Press; 2016.
9. Gerzon G, Sheng Y, Kirkitadze M. Process analytical technologies—advances in bioprocess integration and future perspectives. *J Pharm Biomed Anal*. 2022;(207):114379.
10. Wang Q, Shan P. *Molecular spectrum detection and data processing technology*. Beijing: Science Press; 2019.
11. Pasquini C. Near infrared spectroscopy: a mature analytical technique with new perspectives—a review. *Anal Chim Acta*. 2018;1026:8–36.
12. Yang Z, Albrow-Owen T, Cui H, et al. Single-nanowire spectrometers. *Science*. 2019;365:1017–20.
13. Tang Y, Jones E, Minasny B. Evaluating low-cost portable near infrared sensors for rapid analysis of soils from South Eastern Australia. *Geoderma Reg*. 2020;(20):e00240.
14. Kartakoullis A, Comaposada J, Cruz-Carrión A, et al. Feasibility study of smartphone-based Near Infrared Spectroscopy (NIRS) for salted minced meat composition diagnostics at different temperatures. *Food Chem*. 2019;278:314–21.
15. Jian X, Zhang LF, Yang H, et al. Spectral detection for quality and freshness index of main leaf vegetables based on smart cellphone. *Spectrosc Spectr Anal*. 2019;39:1524–9.
16. Lu W, Yuan H, Chu X. *Near infrared spectrometer*. Beijing: Chemical Industry Press; 2010.
17. da Silva NC, de Góes MARC, Domingos D, et al. NIR-based octane rating simulator for use in gasoline compounding processes. *Fuel*. 2019;243:381–9.

18. Wu Y, Jin Y, Li Y, et al. NIR spectroscopy as a process analytical technology (PAT) tool for on-line and real-time monitoring of an extraction process. *Vib Spectrosc.* 2012;58:109–18.
19. Pu Y-Y, O'Donnell C, Tobin JT, et al. Review of near-infrared spectroscopy as a process analytical technology for real-time product monitoring in dairy processing. *Int Dairy J.* 2020;(103):104623.
20. Grassi S, Alamprese C. Advances in NIR spectroscopy applied to process analytical technology in food industries. *Curr Opin Food Sci.* 2018;22:17–21.
21. Märk J, Karner M, Andre M, et al. Online process control of a pharmaceutical intermediate in a fluidized-bed drier environment using near-infrared spectroscopy. *Anal Chem.* 2010;(82):4209–15.
22. Ruangratanakorn J, Suwonsichon T, Kasemsumran S, et al. Installation design of on-line near infrared spectroscopy for the production of compound fertilizer. *Vib Spectrosc.* 2020;(106):103008.
23. Ryan JA, Compton SV, Brooks MA, et al. Rapid verification of identity and content of drug formulations using mid-infrared spectroscopy. *J Pharm Biomed Anal.* 1991;9:303–10.
24. Su W-H, Sun D-W. Mid-infrared (MIR) spectroscopy for quality analysis of liquid foods. *Food Eng Rev.* 2019;11:142–58.
25. Lu SL, Zhao HJ, Ren LB, et al. The online monitoring system of VOCs emitted by stationary pollution source based on FTIR. *Spectrosc Spectr Anal.* 2018;38:3106–11.
26. Okumura T, Otsuka M. Evaluation of the microcrystallinity of a drug substance, indomethacin, in a pharmaceutical model tablet by chemometric FT-Raman spectroscopy. *Pharm Res.* 2005;22:1350–7.
27. Szostak R, Mazurek S. Quantification of active ingredients in suppositories by FT-Raman spectroscopy. *Drug Test Anal.* 2013;5:126–9.
28. Dymińska L, Calik M, Albegar AMM, et al. Quantitative determination of the iodine values of unsaturated plant oils using infrared and Raman spectroscopy methods. *Int J Food Prop.* 2017;20:2003–15.
29. Lussier F, Thibault V, Charron B, et al. Deep learning and artificial intelligence methods for Raman and surface-enhanced Raman scattering. *TrAC Trends Anal Chem.* 2020;(124):115796.
30. Mamián-López MB, Poppi RJ. Quantification of moxifloxacin in urine using surface-enhanced Raman spectroscopy (SERS) and multivariate curve resolution on a nanostructured gold surface. *Anal Bioanal Chem.* 2013;405:7671–7.
31. Zhang Y, Huang Y, Zhai F, et al. Analyses of enrofloxacin, furazolidone and malachite green in fish products with surface-enhanced Raman spectroscopy. *Food Chem.* 2012;135:845–50.
32. Huang S, Wu Y, Hu J, et al. Rapid detection of malathion residues in Chinese cabbage by surface enhanced Raman spectroscopy. *Trans Chin Soc Agric Eng.* 2016;32:296–301.
33. Liu Y, Xie Q, Wang H, et al. Quantitative study on phosmet residues in navel oranges based on surface enhanced Raman spectra. *Laser Technol.* 2017;41:545–8.
34. Nie X-M, Wang J, Wang X, et al. Highly effective detection of amitraz in honey by using surface-enhanced Raman scattering spectroscopy coupled with chemometric methods. *Chin J Chem Phys.* 2019;(32):444–50.
35. Huo Y, Gao Z, Liu S, et al. Recent advances in surface-enhanced Raman spectroscopy for the detection of tumor markers. *Chin J.* 2020;65:1448–62.
36. Liu R, Xiong Y, Guo Y, et al. Label-free and non-invasive BS-SERS detection of liver cancer based on the solid device of silver nanofilm. *J Raman Spectrosc.* 2018;49:1426–34.
37. Liu S, Huo Y, Kang W, et al. Advances in bacterial detection based on raman spectroscopy. *Chin Sci Bull.* 2020;1448–62.
38. Jarvis RM, Goodacre R. Discrimination of bacteria using surface-enhanced Raman spectroscopy. *Anal Chem.* 2004;76:40–7.
39. Lu SH, Wang YS. Developments in detection of explosives based on surface enhanced Raman spectroscopy. *Spectrosc Spectr Anal.* 2018;38:1412–9.
40. Lu S, Wang z, Tian F. Application of illegal drugs detection based on surface enhanced raman spectroscopy. *Laser Optoelectron Prog.* 2018;(55):030004.

41. Dong R, Weng S, Yang L, et al. Detection and direct readout of drugs in human urine using dynamic surface-enhanced Raman spectroscopy and support vector machines. *Anal Chem.* 2015;87:2937–44.
42. Xu BB, Jin SZ, Jiang L, et al. A review of applications of resonance Raman spectroscopy. *Spectrosc Spectr Anal.* 2019;39:2119–27.
43. Man Y, Ang LI, Cao DC, et al. Stimulated Raman scattering microscopy and its application in biological sciences. *J Chin Electron Microsc Soc.* 2015;34:154–62.
44. Zhou Q, Yuan JH, Zhou W, et al. Coherent anti-stokes Raman scattering microscopy and its biomedical application. *J Chin Electron Microsc Soc.* 2015;34:261–71.
45. Wang D, He P, Wang Z, et al. Advances in single cell Raman spectroscopy technologies for biological and environmental applications. *Curr Opin Biotechnol.* 2020;64:218–29.
46. Kusić D, Kampe B, Rösch P, et al. Identification of water pathogens by Raman microspectroscopy. *Water Res.* 2014;48:179–89.
47. Kloß S, Kampe B, Sachse S, et al. Culture independent Raman spectroscopic identification of urinary tract infection pathogens: a proof of principle study. *Anal Chem.* 2013;85:9610–6.
48. Yogesha M, Chawla K, Bankapur A, et al. A micro-Raman and chemometric study of urinary tract infection-causing bacterial pathogens in mixed cultures. *Anal Bioanal Chem.* 2019;411:3165–77.
49. Stöckel S, Meisel S, Lorenz B, et al. Raman spectroscopic identification of mycobacterium tuberculosis. *J Biophotonics.* 2017;10:727–34.
50. Li Y, Huang W, Pan J, et al. Rapid detection of nasopharyngeal cancer using Raman spectroscopy and multivariate statistical analysis. *Mol Clin Oncol.* 2015;3:375–80.
51. Lee W, Lenferink AT, Otto C, et al. Classifying Raman spectra of extracellular vesicles based on convolutional neural networks for prostate cancer detection. *J Raman Spectrosc.* 2020;51:293–300.
52. Gala de Pablo J, Armistead FJ, Peyman SA, et al. Biochemical fingerprint of colorectal cancer cell lines using label-free live single-cell Raman spectroscopy. *J Raman Spectrosc.* 2018;(49):1323–32.
53. Pilát Z, Bernatová S, Ježek J, et al. Microfluidic cultivation and laser tweezers Raman spectroscopy of *E. coli* under antibiotic stress. *Sensors.* 2018;(18):1623.
54. Li YY, Ma JG, Li DC, et al. Research on spatial offset Raman spectroscopy and data processing method. *Spectrosc Spectr Anal.* 2020;40:71–4.
55. Zhu T, Liu Y, Wu J, et al. Development and application of spatially offset Raman spectroscopy. *Spectrosc Spectr Anal.* 2019;39:997–1004.
56. Eliasson C, Macleod N, Matousek P. Noninvasive detection of concealed liquid explosives using Raman spectroscopy. *Anal Chem.* 2007;79:8185–9.
57. Stone N, Baker R, Rogers K, et al. Subsurface probing of calcifications with spatially offset Raman spectroscopy (SORS): future possibilities for the diagnosis of breast cancer. *Analyst.* 2007;132:899–905.
58. Ding H, Lu G, West C, et al. Spatially offset raman spectroscopy for non-invasive assessment of fracture healing. In: *Photonic therapeutics and diagnostics XII: international society for optics and photonics*; 2016. p. 96894M.
59. Eliasson C, Macleod NA, Jayes LC, et al. Non-invasive quantitative assessment of the content of pharmaceutical capsules using transmission Raman spectroscopy. *J Pharm Biomed Anal.* 2008;47:221–9.
60. Matousek P, Parker A. Bulk Raman analysis of pharmaceutical tablets. *Appl Spectrosc.* 2006;60:1353–7.
61. Johansson J, Sparén A, Svensson O, et al. Quantitative transmission Raman spectroscopy of pharmaceutical tablets and capsules. *Appl Spectrosc.* 2007;61:1211–8.
62. Sparén A, Johansson J, Svensson O, et al. Transmission Raman spectroscopy for quantitative analysis of pharmaceutical solids. *Am Pharm Rev.* 2009;12:66–71.
63. Crocombe RA. Portable spectroscopy. *Appl Spectrosc.* 2018;72:1701–51.
64. Jehlička J, Culka A, Bersani D, et al. Comparison of seven portable Raman spectrometers: beryl as a case study. 2017;(48):1289–99.

65. Chandler L, Huang B, Mu TT. A smart handheld Raman spectrometer with cloud and AI deep learning algorithm for mixture analysis. In: Next-generation spectroscopic technologies XII: international society for optics and photonics;2019. p. 1098308.
66. Fowler SM, Schmidt H, van de Ven R, et al. Preliminary investigation of the use of Raman spectroscopy to predict meat and eating quality traits of beef loins. *Meat Sci.* 2018;138:53–8.
67. Wang Q, Lonergan SM, Yu C. Rapid determination of pork sensory quality using Raman spectroscopy. *Meat Sci.* 2012;91:232–9.
68. He Q-j, Wang L-Q. Research progress of raman spectroscopy on dyestuff identification of ancient relics and artifacts. In: *Guang pu xue yu Guang pu fen xi.* 2016;(36):401–07.
69. Christesen S, Guicheteau J, Curtiss J, et al. Handheld dual-wavelength Raman instrument for the detection of chemical agents and explosives. 2016(*Opt Eng*); 074103.
70. Wang H, Wang YZ, Zhao Y, et al. Latest methods of fluorescence suppression in Raman spectroscopy. *Spectrosc Spectr Anal.* 2017;37:2050–6.
71. Zou W, Cai Z, Wu J. Fluorescence rejection by shifted excitation Raman difference spectroscopy. *SPIE.* 2010.
72. Jin G, Huang X, Chen RH. Applications of real-time measurement technology with Raman spectroscopy for polymer synthesis and processing. *Spectrosc Spectr Anal.* 2016;36:2124–7.
73. Marteau P, Zanier-Szydłowski N, Aoufi A, et al. Remote Raman spectroscopy for process control. *Vib Spectrosc.* 1995;9:101–9.
74. Cansell F, Hotier G, Marteau P, et al. Method for regulating a process for the separation of isomers of aromatic hydrocarbons having from 8 to 10 carbon atoms. Google Patents. 1996.
75. Kong K, Kendall C, Stone N, et al. Raman spectroscopy for medical diagnostics—from in-vitro biofluid assays to in-vivo cancer detection. *Adv Drug Deliv Rev.* 2015;89:121–34.
76. Cordero E, Latka I, Matthäus C, et al. Raman spectroscopy: from basics to applications. *J Biomed Opt.* 2018;(23):071210.
77. Yan H, Yu M, Xia J, et al. Tongue squamous cell carcinoma discrimination with Raman spectroscopy and convolutional neural networks. *Vib Spectrosc.* 2019;(103):102938.
78. Ralbovsky NM, Lednev IK. Towards development of a novel universal medical diagnostic method: Raman spectroscopy and machine learning. *Chem Soc Rev.* 2020;49:7428–53.
79. Buckley K, Kerns JG, Parker AW, et al. Decomposition of in vivo spatially offset Raman spectroscopy data using multivariate analysis techniques. 2014;(45):188–92.
80. Tong X. On-line analysis and applicable control of H₂S/SO₂ ratio analyzer in sulfur recovery unit. *Process Autom Instrum.* 2009;30:23–30.
81. Langergraber G, Fleischmann N, Hofstaedter F, et al. Monitoring of a paper mill wastewater treatment plant using UV/VIS spectroscopy. *Water Sci Technol.* 2004;49:9–14.
82. van den Broeke J, Langergraber G, Weingartner A. On-line and in situ UV/vis spectroscopy for multi-parameter measurements: a brief review. *Spectrosc Eur.* 2006;18:S3–4.
83. Zhu H, Chai X, Wang S, et al. Attenuated total reflection UV/vis spectroscopic applications. *Prog Chem.* 2007;19(2–3):414–19.
84. Johansson J, Cauchi M, Sundgren M. Multiple fiber-optic dual-beam UV/Vis system with application to dissolution testing. *J Pharm Biomed Anal.* 2002;29:469–76.
85. Inman GW, Wethington E, Baughman K, et al. System optimization for in situ fiber-optic dissolution testing. *Pharm Technol.* 2001;25:92–100.
86. Florence AJ, Johnston A. Applications of ATR UV/vis spectroscopy in physical form characterisation of pharmaceuticals. *Spectrosc Eur.* 2004;4:24–7.
87. Levi MAB, Scarmínio IS, Poppi RJ, et al. Three-way chemometric method study and UV-Vis absorbance for the study of simultaneous degradation of anthocyanins in flowers of the *Hibiscus rosa-sinensis* species. *Talanta.* 2004;62:299–305.
88. Atole DM, Rajput HH. Ultraviolet spectroscopy and its pharmaceutical applications—a brief review. *Asian J Pharm Clin Res.* 2018;11:59–66.
89. Dai X, Song H, Liu W, et al. On-line UV-NIR spectroscopy as a process analytical technology (PAT) tool for on-line and real-time monitoring of the extraction process of *Coptis Rhizome*. *RSC Adv.* 2016;6:10078–85.

90. Warner IM, Callis JB, Davidson ER, et al. Fluorescence analysis: a new approach. *Anal Lett.* 1975;8:665–81.
91. Han R, Li Z, Fan Y, et al. Recent advances in super-resolution fluorescence imaging and its applications in biology. *J Genet Genomics.* 2013;40:583–95.
92. Pu C, Chu X, Tian S. The application of molecular fluorescence spectroscopy in analysis of crude oil. *Mod Sci Instrum.* 2012;1:129–33.
93. He X-f, Xiong A-b. Application and research progress of three-dimensional printing in the field of orthopaedics. *Chin J Tissue Eng Res.* 2017;21:428–32.
94. Liu W, Zhang L, Liu P, et al. FDOM conversion in karst watersheds expressed by three-dimensional fluorescence spectroscopy. *Water.* 2018;10:1427.
95. Peleato NM, Andrews RC. Comparison of three-dimensional fluorescence analysis methods for predicting formation of trihalomethanes and haloacetic acids. *J Environ Sci.* 2015;27:159–67.
96. Wang Z, Wu Z, Tang S. Characterization of dissolved organic matter in a submerged membrane bioreactor by using three-dimensional excitation and emission matrix fluorescence spectroscopy. *Water Res.* 2009;43:1533–40.
97. Gu H-W, Wu H-L, Liu Y-J, et al. Simultaneous determination of metoprolol and α -hydroxymetoprolol in human plasma using excitation–emission matrix fluorescence coupled with second-order calibration methods. 2012;(4):2781–93.
98. Nie JF, Wu HL, Xia AL, et al. Determination of sulphiride in human urine using excitation–emission matrix fluorescence coupled with second-order calibration. *Anal Sci Int J Jpn Soc Anal Chem.* 2007;23:1377.
99. Qing X-D, Wu H-L, Nie C-C, et al. Simultaneous determination of plant growth regulators in environmental samples using chemometrics-assisted excitation–emission matrix fluorescence: experimental study on the prediction quality of second-order calibration method. *Talanta.* 2013;103:86–94.
100. Bravo MM, Aguilar LF, Quiroz VW, et al. Determination of tributyltin at parts-per-trillion levels in natural waters by second-order multivariate calibration and fluorescence spectroscopy. *Microchem J.* 2013;106:95–101.
101. Wang L, Wu H-L, Yin X-L, et al. Simultaneous determination of umbelliferone and scopoletin in Tibetan medicine *Saussurea laniceps* and traditional Chinese medicine *Radix angelicae pubescentis* using excitation–emission matrix fluorescence coupled with second-order calibration method. *Spectrochim Acta Part A Mol Biomol Spectrosc.* 2017;170:104–10.
102. Zhong X, Liu Y, Yong L, et al. Three-dimensional fluorescence technique coupled with chemometric second-order calibration method for simultaneous detection of thiabendazole and fuberidazole in red wine. *Life Sci Instrum.* 2015;39:38–41.
103. Zhu ZW, Que LZ, Chen GQ, et al. Year discrimination of mild aroma chinese liquors using three-dimensional fluorescence spectroscopy combined with parallel factor and neural network. *Spectrosc Spectr Anal.* 2015;35:2573–7.
104. Han XS, Liu DP, Luan XN, et al. Discrimination of crude oil samples using laser-induced time-resolved fluorescence spectroscopy. *Spectrosc Spectr Anal.* 2016;36:445–8.
105. Wang X, Zhao NJ, Yu ZM, et al. Detection method progress and development trend of organic pollutants in soil using laser-induced fluorescence spectroscopy. *Spectrosc Spectr Anal.* 2018;38:857–63.
106. Huang Y, Zhao NJ, Meng DS, et al. Advance in the detection techniques of persistent organic pollutants by using fluorescence spectrometry. *Spectrosc Spectr Anal.* 2019;39:2107–13.
107. Hu F, Zhou M, Yan P, et al. Identification of mine water inrush using laser-induced fluorescence spectroscopy combined with one-dimensional convolutional neural network. *RSC Adv.* 2019;9:7673–9.
108. Wang X, Zhao NJ, Yin GF, et al. Classification and identification of plastic with laser-induced fluorescence spectroscopy based on back propagation neural network model. *Spectrosc Spectr Anal.* 2019;39:3136–41.
109. Fan Y, Wu RM, Ai SR, et al. Identification study of edible oil species with laser induced fluorescence technology based on liquid core optical fiber. *Spectrosc Spectr Anal.* 2016;36:3202–6.

110. Kapadia CR, Cutruzzola FW, O'Brien KM, et al. Laser-induced fluorescence spectroscopy of human colonic mucosa: detection of adenomatous transformation. *Gastroenterology*. 1990;99:150–7.
111. Mandrioli R, Morganti E, Mercolini L, et al. Fast analysis of amino acids in wine by capillary electrophoresis with laser-induced fluorescence detection. *Electrophoresis*. 2011;32:2809–15.
112. Stefan A-E, Jonas J, Katarina Svanberg MD, et al. Laser-induced fluorescence in medical diagnostics. *Proc SPIE*. 1990.
113. Li J, Xu M, Ma Q, et al. Sensitive determination of silicon contents in low-alloy steels using micro laser-induced breakdown spectroscopy assisted with laser-induced fluorescence. *Talanta*. 2019;194:697–702.
114. Lui SL, Godwal Y, Taschuk MT, et al. Detection of lead in water using laser-induced breakdown spectroscopy and laser-induced fluorescence. *Anal Chem*. 2008;80:1995–2000.
115. Freedman R. Advances in NMR logging. *J Petrol Technol*. 2006;58:60–6.
116. Mitchell J, Gladden LF, Chandrasekera TC, et al. Low-field permanent magnets for industrial process and quality control. *Prog Nucl Magn Reson Spectrosc*. 2014;76:1–60.
117. Zang X, Lin Z, Zhang T, et al. Non-destructive measurement of water and fat contents, water dynamics during drying and adulteration detection of intact small yellow croaker by low field NMR. *J Food Meas Charact*. 2017;11:1550–8.
118. Gai S, Zhang Z, Zou Y, et al. Rapid and non-destructive detection of water-injected pork using low-field nuclear magnetic resonance (LF-NMR) and magnetic resonance imaging (MRI). *Int J Food Eng*. 2019;15:1–9.
119. Feng L, Zhang M, Bhandari B, et al. Determination of postharvest quality of cucumbers using nuclear magnetic resonance and electronic nose combined with chemometric methods. *Food Bioprocess Technol*. 2018;11:2142–52.
120. Nordon A, McGill CA, Littlejohn D. Process NMR spectrometry. *Analyst*. 2001;126:260–72.
121. Bakeev KA. Process analytical technology: spectroscopic tools and implementation strategies for the chemical and pharmaceutical industries. Wiley & Sons;2010.
122. Edwards JC. A review of applications of NMR spectroscopy in the petroleum industry. *Spectrosc Anal Pet Prod Lubr*. 2011;16:423.
123. Song H-J, Nagatsuma T. Handbook of terahertz technologies: devices and applications. CRC Press;2015.
124. Li B, Zhao XT, Zhang YZ, et al. Progress on terahertz spectroscopic detection and analysis on antibiotics. *Spectrosc Spectr Anal*. 2019;39:3659–66.
125. Yang X, Zhao X, Yang K, et al. Biomedical applications of terahertz spectroscopy and imaging. *Trends Biotechnol*. 2016;34:810–24.
126. Afsah-Hejri L, Akbari E, Toudeshki A, et al. Terahertz spectroscopy and imaging: a review on agricultural applications. *Comput Electron Agric*. 2020;(177):105628.
127. Afsah-Hejri L, Hajeb P, Ara P, et al. A comprehensive review on food applications of terahertz spectroscopy and imaging. *Compr Rev Food Sci Food Saf*. 2019;18:1563–621.
128. Winefordner JD, Gornushkin IB, Correll T, et al. Comparing several atomic spectrometric methods to the super stars: special emphasis on laser induced breakdown spectroscopy, LIBS, a future super star. *J Anal At Spectrom*. 2004;19:1061–83.
129. Aragón C, Aguilera JA. Characterization of laser induced plasmas by optical emission spectroscopy: a review of experiments and methods. *Spectrochim Acta, Part B*. 2008;63:893–916.
130. Zhang T, Tang H, Li H. Chemometrics in laser-induced breakdown spectroscopy. 2018;(32):e2983.
131. Unnikrishnan V, Nayak R, Aithal K, et al. Analysis of trace elements in complex matrices (soil) by laser induced breakdown spectroscopy (LIBS). *Anal Methods*. 2013;5:1294–300.
132. Wang J, Liao X, Zheng P, et al. Classification of Chinese herbal medicine by laser-induced breakdown spectroscopy with principal component analysis and artificial neural network. *Anal Lett*. 2018;51:575–86.
133. Zhao YY, Zhu SS, He J, et al. Identification of *fritillaria thunbergii* treated by sulfur fumigation using laser-induced breakdown spectroscopy. *Spectrosc Spectr Anal*. 2018;38:3558–62.

134. Markiewicz-Keszycka M, Cama-Moncunill X, Casado-Gavalda MP, et al. Laser-induced breakdown spectroscopy (LIBS) for food analysis: a review. *Trends Food Sci Technol.* 2017;65:80–93.
135. Yu K, Ren J, Zhao Y. Principles, developments and applications of laser-induced breakdown spectroscopy in agriculture: a review. *Artif Intell Agric.* 2020;4:127–39.
136. Bilge G, Velioglu HM, Sezer B, et al. Identification of meat species by using laser-induced breakdown spectroscopy. *Meat Sci.* 2016;119:118–22.
137. Wang J, Zheng P, Liu H, et al. Classification of Chinese tea leaves using laser-induced breakdown spectroscopy combined with the discriminant analysis method. *Anal Methods.* 2016;8:3204–9.
138. Sun L, Yu H, Cong Z, et al. Applications of laser-induced breakdown spectroscopy in the aluminum electrolysis industry. *Spectrochim Acta, Part B.* 2018;142:29–36.
139. Lu Z, Mo J, Yao S, et al. Rapid determination of gross calorific value of coal using LIBS coupled with artificial neural networks (ANN) and genetic algorithm. *Energy Fuels.* 2017;31:3849–55.
140. Guo ZW, Sun LX, Zhang P, et al. On-line component analysis of cement powder using LIBS technology. *Spectrosc Spectr Anal.* 2019;39:278–85.
141. Wang Q, Xiangli W, Teng G, et al. A brief review of laser-induced breakdown spectroscopy for human and animal soft tissues: pathological diagnosis and physiological detection. *Appl Spectrosc Rev.* 2021;56:221–41.
142. Gaudiuso R, Melikechi N, Abdel-Salam ZA, et al. Laser-induced breakdown spectroscopy for human and animal health: a review. *Spectrochim Acta, Part B.* 2019;152:123–48.
143. Chen X, Li X, Yu X, et al. Diagnosis of human malignancies using laser-induced breakdown spectroscopy in combination with chemometric methods. *Spectrochim Acta, Part B.* 2018;139:63–9.
144. Wang J, Li L, Yang P, et al. Identification of cervical cancer using laser-induced breakdown spectroscopy coupled with principal component analysis and support vector machine. *Lasers Med Sci.* 2018;33:1381–6.
145. Gimenez Y, Busser B, Trichard F, et al. 3D imaging of nanoparticle distribution in biological tissue by laser-induced breakdown spectroscopy. *Sci Rep.* 2016;6:29936.
146. Qin J, Kim MS, Chao K, et al. Detection and quantification of adulterants in milk powder using a high-throughput Raman chemical imaging technique. *Food Addit Contam: Part A.* 2017;34:152–61.
147. Jolivet L, Leprince M, Moncayo S, et al. Review of the recent advances and applications of LIBS-based imaging. *Spectrochim Acta, Part B.* 2019;151:41–53.
148. Kessler RW. Perspectives in process analysis. *J Chemom.* 2013;27:369–78.
149. Maldonado AIL, Rodriguez-Fuentes H, Contreras JAV. *Hyperspectral imaging in agriculture, food and environment. BoD—Books on Demand;*2018.
150. Bai XB, Yu JS, Fu ZT, et al. Application of spectral imaging technology for detecting crop disease information: a review. *Spectrosc Spectr Anal.* 2020;40:350–5.
151. Munir MT, Wilson DI, Yu W, et al. An evaluation of hyperspectral imaging for characterising milk powders. *J Food Eng.* 2018;221:1–10.
152. Roggo Y, Edmond A, Chalus P, et al. Infrared hyperspectral imaging for qualitative analysis of pharmaceutical solid forms. *Anal Chim Acta.* 2005;535:79–87.
153. Boiret M, Rutledge DN, Gorretta N, et al. Application of independent component analysis on Raman images of a pharmaceutical drug product: pure spectra determination and spatial distribution of constituents. *J Pharm Biomed Anal.* 2014;90:78–84.
154. Yu H-D, Qing L-W, Yan D-T, et al. Hyperspectral imaging in combination with data fusion for rapid evaluation of tilapia fillet freshness. *Food Chem.* 2021;(348):129129.
155. Chandrasekaran I, Panigrahi SS, Ravikanth L, et al. Potential of near-infrared (NIR) spectroscopy and hyperspectral imaging for quality and safety assessment of fruits: an overview. *Food Anal Methods.* 2019;12:2438–58.
156. Alexandrino GL, Khorasani MR, Amigo JM, et al. Monitoring of multiple solid-state transformations at tablet surfaces using multi-series near-infrared hyperspectral imaging and multivariate curve resolution. *Eur J Pharm Biopharm.* 2015;93:224–30.

157. Johnson JB. An overview of near-infrared spectroscopy (NIRS) for the detection of insect pests in stored grains. *J Stored Prod Res.* 2020;(86):101558.
158. Feng C-H, Otani C. Terahertz spectroscopy technology as an innovative technique for food: current state-of-the-art research advances. *Crit Rev Food Sci Nutr.* 2021;61:2523–43.
159. Manley M. Near-infrared spectroscopy and hyperspectral imaging: non-destructive analysis of biological materials. *Chem Soc Rev.* 2014;43:8200–14.
160. Silvia S. Plastic waste monitoring and recycling by hyperspectral imaging technology. *Proc SPIE.* 2019.
161. Fei B. Chapter 3.6—Hyperspectral imaging in medical applications. In: Amigo JM, editor. *Data handling in science and technology.* Elsevier;2020. p. 523–65.
162. Halicek M, Fabelo H, Ortega S, et al. In-vivo and ex-vivo tissue analysis through hyperspectral imaging techniques: revealing the invisible features of cancer. *Cancers.* 2019;11:756.
163. Raeissi B, Bashir MA, Garrett JL, et al. Detection of different chemical binders in coatings using hyperspectral imaging. *J Coat Technol Res.* 2021; 1–16.

Chapter 3

Basis of Matrices and Mathematical Statistics



3.1 Basis of Matrix

The data used in analytical chemistry can be divided into scalars, vectors, matrices, and tensors. A scalar is a simple quantity, such as the volume of end point in titration analysis, by which the concentration of the sample under test can be calculated. The measurement data of instrument analysis is often not just a scalar, such as a spectrum (ultraviolet (UV) spectrum, infrared (IR) spectrum, etc.) measured by a spectrometer. Although it usually appears as a curve, in fact, this curve corresponds to a set of discrete values. In mathematics, a set of numbers can be represented by a vector. Thus, a vector can represent a spectrum (such as an UV, IR, chromatographic, or nuclear magnetic resonance (NMR) spectrum). If more than one spectrum is obtained from a measurement, a matrix of vectors can be used to represent the measurement result. A combination of instruments can often produce the data represented in this matrix. For example, the retention time of a sample is from 0 to 30 min, and the measurement wavelength is from 200 to 400 nm. If the sampling interval is 1 s, 1801 points will be measured. If the wavelength interval of the UV measurement is 2 nm, each spectrum will contain 101 points, resulting in a matrix of 1801×101 dimensions. This needs to be represented using a three-dimensional graph (as shown in Fig. 3.1). One row of this matrix corresponds to the spectrum of a chromatographic measurement point, while one column corresponds to the chromatogram measured at a wavelength. Using matrix to describe this kind of measurement data can completely express the measurement results, and it is very convenient for data processing. This kind of matrix data is widely used in chemometrics. Measurement of a sample with a combined instrument yield a matrix of data, and measurement of more than one sample yield several matrices that can form a tensor, which often contains more information [1, 2].

The same is true of spectral data processing. The independent variable (generally called the matrix \mathbf{X} , as shown in Fig. 3.2) with the dimension of $n \times m$ (n is the number of sample and m is the wavelength variable) a set of sample spectrum. Likewise, the dependent variable (generally called the matrix \mathbf{Y} , as shown in Fig. 3.2)

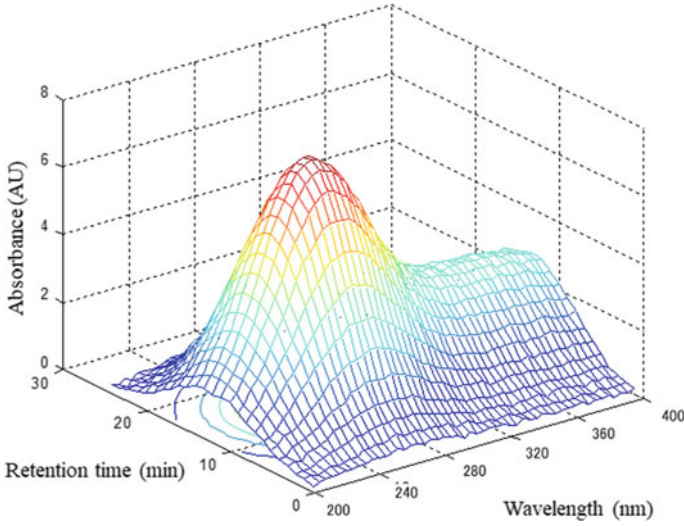


Fig. 3.1 Three-dimensional graphical representation of matrix data

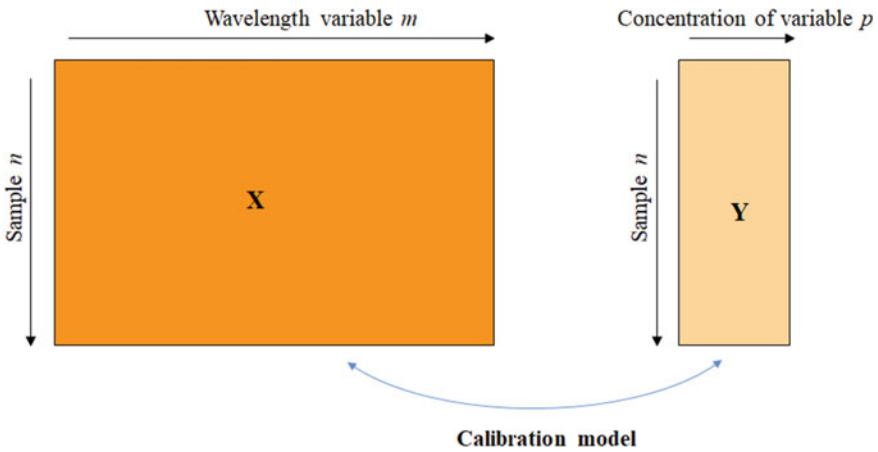


Fig. 3.2 Independent variable X matrix and dependent variable Y matrix

with the dimension of $n \times p$ (n is the number of sample and p is the number of concentrations) is composed of the corresponding concentrations of a set of properties (or components) data (e.g., wheat protein, starch and water, etc.). In fact, the establishment of correction model is to establish the relationship between the X matrix and the Y matrix. In addition to concentration, the Y matrix can also be an information matrix (e.g., category).

In chemometrics, scalars are often referred to as zero-dimensional data (or tensors of zero order) and are often represented with lowercase letters, such as $a = [0.05]$;

A vector is called a one-dimensional data (or a tensor of the first order) and is often represented by a boldface letter in lower case, such as $\mathbf{a} = [0.1 \ 0.5 \ 0.8 \ 0.7]$; Matrices are called two-dimensional data (or second-order tensors) and are usually

represented in bold letters in capital letters, such as $\mathbf{A} = \begin{bmatrix} 1 & 8 & 5 & 6 \\ 9 & 2 & 3 & 4 \\ 5 & 1 & 8 & 7 \\ 8 & 7 & 6 & 4 \end{bmatrix}$; tensors are

called three dimensional data (or tensors of third order). In addition, there are some relatively fixed expressions and operations on vectors and matrices, which are briefly introduced as follows [3, 4].

a_{ij} represents an element of the i th row and j th column of matrix \mathbf{A} , called the (i, j) element of matrix \mathbf{A} .

\mathbf{AB} represents the product of matrix \mathbf{A} and matrix \mathbf{B} , where the number of columns in matrix \mathbf{A} must equal to the number of rows in matrix \mathbf{B} . If \mathbf{A} is a matrix of $m \times s$ and \mathbf{B} is a matrix of $s \times n$, then the product of the matrices \mathbf{A} and \mathbf{B} is $\mathbf{AB} = \mathbf{C} = (c_{ij})$, and the product \mathbf{C} is an $m \times n$ matrix as follows:

$$c_{ij} = a_{i1}b_{1j} + a_{i2}b_{2j} + \dots + a_{is}b_{sj} = \sum_{k=1}^s a_{ik}b_{kj} \dots i = 1, 2, \dots, m; j = 1, 2, \dots, n \tag{3.1}$$

Matrix multiplication generally does not satisfy the commutative law, that is, $\mathbf{AB} \neq \mathbf{BA}$. But it satisfies the associative property, that is $(\mathbf{AB})\mathbf{C} = \mathbf{A}(\mathbf{BC})$.

\mathbf{A}^T represents the transpose of matrix \mathbf{A} , namely, the $m \times n$ matrix obtained by interchanging the rows and columns of the $n \times m$ matrix \mathbf{A} in the original order.

\mathbf{I} is the identity matrix, which is an $n \times n$ square matrix where every entry on the main diagonal is 1, and every other entry is 0.

$|\mathbf{A}|$ represents the determinant of a square matrix \mathbf{A} , and if $|\mathbf{A}| \neq 0$, \mathbf{A} is nonsingular matrix.

For a square matrix \mathbf{A} of order $n \times n$, if there is a number and a non-zero vector, satisfying $\mathbf{Ax} = \lambda\mathbf{x}$, λ is the eigenvalue of the square matrix \mathbf{A} , and \mathbf{x} is the eigenvector corresponding to the eigenvalue. QR (Orthogonal Trigonometry) decomposition method and Jacobian method are used to solve the eigenvalues of matrices.

$\text{tr}(\mathbf{A})$ represents the trace of square matrix \mathbf{A} , whose value is equal to the sum of the diagonal elements of square matrix \mathbf{A} . According to Veda's theorem, the sum of all eigenvalues of the matrix is the trace of the matrix. The sum of all eigenvalues of a matrix is the determinant of the matrix.

$\text{Rank}(\mathbf{A})$ represents the rank of matrix \mathbf{A} , which is the maximum linearly independent number of rows or columns in \mathbf{A} .

\mathbf{A}^{-1} represents the inverse of the matrix \mathbf{A} , which means $\mathbf{AA}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$. If \mathbf{A}^{-1} exists, \mathbf{A} is called a non-singular matrix, or \mathbf{A} is called a full rank matrix; otherwise, \mathbf{A} is called a singular matrix.

\mathbf{A} is said to be an orthogonal matrix if the square matrix \mathbf{A} of order $n \times n$ satisfies $\mathbf{A}^{-1} = \mathbf{A}^T$. The row and column vectors of an orthogonal matrix are orthogonal to each other, namely, the inner product of different rows or columns is zero.

\mathbf{A}^+ represents the generalized inverse of the matrix \mathbf{A} , which means $\mathbf{A}\mathbf{A}^+\mathbf{A} = \mathbf{A}$. $\|\mathbf{a}\|$ represents the mode or norm of vector $\mathbf{a} = [a_1 \ a_2 \ a_3, \dots, a_n]$ as follows:

$$\|\mathbf{a}\| = \|\mathbf{a}\| = \sqrt{a_1^2 + a_2^2 + \dots + a_n^2} \quad (3.2)$$

$$\|\mathbf{a}\| = \sqrt{(\mathbf{a}\mathbf{a}^T)} = \sqrt{\text{tr}(\mathbf{a}^T\mathbf{a})} \quad (3.3)$$

3.2 Matrix Representation of Lambert-Beer's Law

Lambert-Beer's law of multi-component systems can be expressed by matrix operations [5, 6]. If a mixture is composed of three components, the three components of the pure spectra with vector expressed as \mathbf{s}_1 , \mathbf{s}_2 , and \mathbf{s}_3 . If three kinds of components in a mixture of relative concentration of c_1 , c_2 , and c_3 , respectively, depending on the Lambert-Beer's law, the mixture of spectrum \mathbf{x} should be equal to the pure spectra of three components and corresponding to the sum of product of concentration. That is, $\mathbf{x} = c_1\mathbf{s}_1 + c_2\mathbf{s}_2 + c_3\mathbf{s}_3 + \mathbf{e}$, where \mathbf{e} is the measurement error of the instrument. Namely,

$$x_1 = c_1s_{11} + c_2s_{12} + c_3s_{13} + e_1 \quad (3.4)$$

$$x_2 = c_1s_{21} + c_2s_{22} + c_3s_{23} + e_2 \quad (3.5)$$

...

$$x_m = c_1s_{m1} + c_2s_{m2} + c_3s_{m3} + e_m \quad (3.6)$$

where m is the number of wavelength points, x_m represents the absorbance of the mixture at wavelength m , and s_{m3} represents the absorbance of the third pure component at wavelength m .

Let $\mathbf{x} = [x_1 \ x_2 \ \dots \ x_m]^T$, $\mathbf{S} = [\mathbf{s}_1^T \ \mathbf{s}_2^T \ \mathbf{s}_3^T]$, $\mathbf{c} = [c_1 \ c_2 \ c_3]^T$, $\mathbf{e} = [e_1 \ e_2 \ e_3]^T$, then the above equation can be expressed as a matrix product: $\mathbf{x} = \mathbf{cS} + \mathbf{e}$.

If there are n samples of such mixtures, according to matrix multiplication rules, they can be expressed as follows:

$$\begin{aligned}
 \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1m} \\ x_{21} & x_{22} & \dots & x_{2m} \\ \dots & \dots & \dots & \dots \\ x_{n1} & x_{n2} & \dots & x_{nm} \end{bmatrix} &= \begin{bmatrix} c_{11} & c_{12} & c_{13} \\ c_{21} & c_{22} & c_{23} \\ \dots & \dots & \dots \\ c_{n1} & c_{n2} & c_{n3} \end{bmatrix} \begin{bmatrix} s_{11} & s_{12} & \dots & s_{1m} \\ s_{21} & s_{22} & \dots & s_{2m} \\ s_{31} & s_{32} & \dots & s_{3m} \end{bmatrix} \\
 &+ \begin{bmatrix} e_{11} & e_{12} & \dots & e_{1m} \\ e_{21} & e_{22} & \dots & e_{2m} \\ \dots & \dots & \dots & \dots \\ e_{n1} & e_{n2} & \dots & e_{nm} \end{bmatrix} \quad (3.7)
 \end{aligned}$$

For the data of n samples with p components at m wavelengths, the matrix can be expressed as follows:

$$\mathbf{X}_{n \times m} = \mathbf{C}_{n \times p} \mathbf{S}_{p \times m} + \mathbf{E}_{n \times m} \quad (3.8)$$

3.3 Variance and Normal Distribution

In analytical chemistry, multiple measurements are often required to eliminate the impact of accidental errors. A set of n measurements (x_1, x_2, \dots, x_n) is usually described by two statistics, including mean and standard deviation [7, 8]:

$$\text{mean } \bar{x} : \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad (3.9)$$

$$\text{standard deviation } s : s = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}} \quad (3.10)$$

The standard deviation s and the square of the standard deviation s^2 (variance) are important statistics that describe how discrete a set of data is. For example, in NIR spectral analysis, for multiple spectral measurements of a sample, it is expected that the changes of repeated spectral measurements should be as small as possible. That is, the variance of the absorbance at each wavelength should be as small as possible. However, for the NIR spectra of a set of correction set samples, it is hoped to find the wavelength with a large corresponding absorbance variance, because the more information there is, the larger the variation of absorbance should be. Therefore, standard deviation or variance is often used to evaluate spectral repeatability in NIR spectral analysis, and it can also be used to select the wavelength range involved in the establishment of calibration model.

The standard deviation can be used to describe the degree to which the measured results are discrete from the mean, but it cannot be used to describe the distribution of

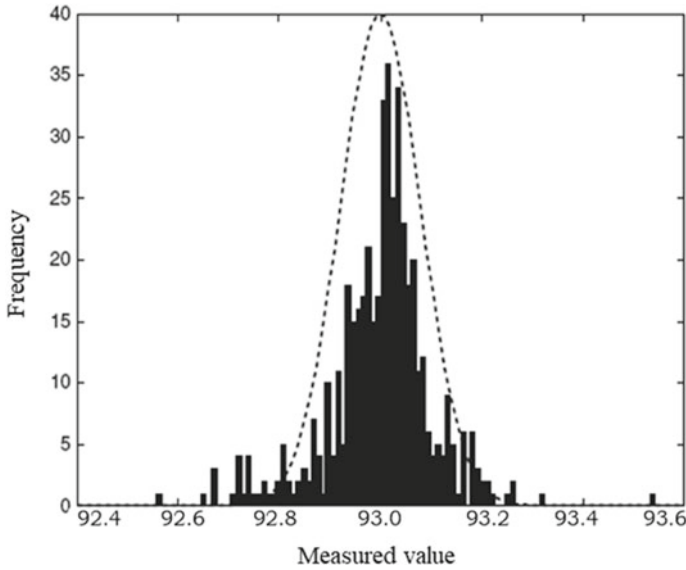


Fig. 3.3 Histogram of 200 measurements of the octane number of a gasoline sample

these data. The distribution of expression data needs to use a histogram (or frequency). For example, using NIR spectroscopy method determines the octane number of gasoline for a 200 times, where the mean value is 93.0, and among them 93.10 occurs for 37 times and 93.05 occurs for 35 times. Mapping the frequency of the occurrence of each measured value vs measured value is the histogram (or frequency chart), as shown in Fig. 3.3.

As can be seen from Fig. 3.3, the distribution is symmetric about the center, and the measurement results tend to gather toward the center value. The population mean can be written as μ , and the mean \bar{x} is actually an estimate of the population mean μ . Similarly, the population has a standard deviation, often expressed as σ , and the standard deviation of the sample s gives an estimate of the population standard deviation σ . In theory, the normal distribution curve, also known as the Gaussian distribution, is commonly used to study this kind of problem, and is described by the following formula:

$$f(x) = \left(\frac{1}{\sigma\sqrt{2\pi}} \right) e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{3.11}$$

The normal distribution curve is shown in Fig. 3.4, which is μ -symmetric. The larger the value of σ , the greater the degree of dispersion of the data, and the wider the curve. But the total area under the curve remains unchanged. As shown in Fig. 3.5, approximately 68% of the total measurements are in the $\pm 1\sigma$ range, and approximately 95% of the total measurements are in the $\pm 2\sigma$ range while approximately

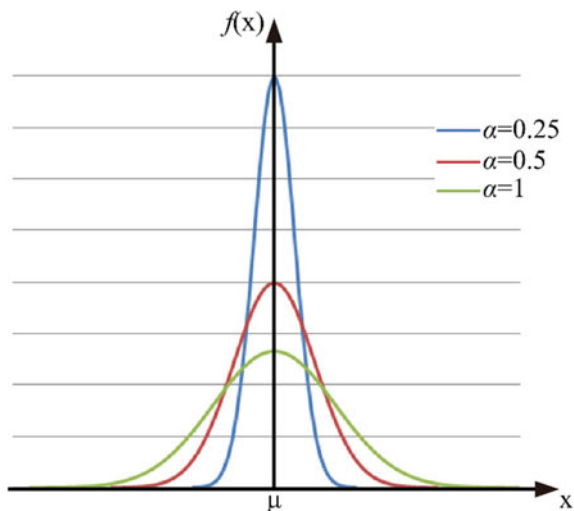


Fig. 3.4 Normal distribution with the same mean and different standard deviations

±1δ= 68.2% area under curve
 ±2δ= 95.4% area under curve
 ±3δ= 99.6% area under curve

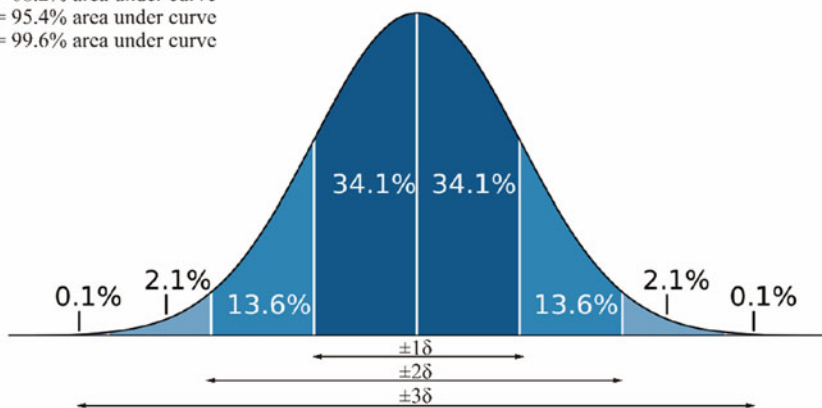


Fig. 3.5 Relationship between normal distribution area and standard deviation

99.7% of the total measurements are in the $\pm 3\sigma$ range. In analytical chemistry, the measurements obtained conform to a normal distribution in most cases.

The abscissa of the normal distribution curve is replaced by $u = (x-\mu)/\sigma$, which is called the standard normal distribution curve. It is expressed by $N(0,1)$, that is, the mean value is 0 and the variance is 1, which is used to represent the distribution of random errors as follows:

$$f(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}} \tag{3.12}$$

Table 3.1 Interval probability table of random error

Interval of random error u	Interval of the measured value (width of confidence interval)	Probability P (confidence) (%)
$-1\sigma \sim +1\sigma$	$\mu-1\sigma \sim \mu + 1\sigma$	68.3
$-1.96\sigma \sim +1.96\sigma$	$\mu-1.96\sigma \sim \mu + 1.96\sigma$	95.0
$-2\sigma \sim +2\sigma$	$\mu-2\sigma \sim \mu + 2\sigma$	95.5
$-2.58\sigma \sim +2.58\sigma$	$\mu-2.58\sigma \sim \mu + 2.58\sigma$	99.0
$-3\sigma \sim +3\sigma$	$\mu-3\sigma \sim \mu + 3\sigma$	99.7

The probability of the occurrence of measured values within the range (called confidence) can be expressed by the integral area of a certain interval as formula (3.13). Obviously, from $-\infty \sim +\infty$, the total probability P of the occurrence of all measured values is 1, that is, the area contained under the normal distribution curve is the sum of the occurrence probability of all measured data.

$$P = \int_{-\infty}^{+\infty} f(u)du = \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-u^2/2} du = 1 \tag{3.13}$$

The interval probability of random error calculated by integration is shown in Table 3.1.

For a finite number of measurements (n measurements), the statistic t is used to deal with and is defined as follows:

$$t = \frac{x - \mu}{s} \sqrt{n}. \tag{3.14}$$

t value is not only related to the confidence P but also to the degree of freedom f , expressed by $t_{\alpha,f}$, $\alpha = 1-P$ (α is called significance level), and $f = n-1$. For example, $t_{0.05,10}$ is the value of t when $P = 95\%$ and $f = 10$. As $f \rightarrow \infty$, the t -distribution becomes a normal distribution. T -distribution, also known as student distribution, can be obtained as Table 3.2 shows.

For a small amount of experimental data, the confidence interval of the mean value can be expressed as

$$\mu = \bar{x} \pm \frac{ts}{\sqrt{n}} \tag{3.15}$$

Table 3.2 *t*-values of different measurement times and different confidence degrees

Measurement times <i>n</i>	Confidence			
	90%	95%	99%	99.5%
2	6.314	12.706	63.657	127.32
3	2.920	4.303	9.925	14.089
4	2.353	3.182	5.841	7.453
5	2.132	2.776	4.604	5.598
6	2.015	2.571	4.032	4.773
7	1.943	2.447	3.707	4.317
8	1.895	2.365	3.500	4.029
9	1.860	2.306	3.355	3.832
10	1.833	2.262	3.250	3.690
11	1.812	2.228	3.169	3.581
21	1.725	2.086	2.845	3.153
∞	1.645	1.960	2.576	2.807

3.4 Significance Test

In the actual measurement, the mean value of the sample may not be equal to the true value. Such difference may be completely accidental error, or may contain systematic error. In order to distinguish the two cases, significance test should be introduced. Through significance test, if the analysis results are found to have significant differences, it can be judged that the analysis results have systematic errors. If there is no significant difference, it indicates that the difference of the analysis result is due to accidental error. In spectral analysis, the most commonly used test methods are *t*-test and *F*-test.

The *t*-test is used to judge whether there is significant difference between the mean value \bar{x} and the true value μ . The *t*-value is calculated as follows:

$$t = (\bar{x} - \mu) \frac{\sqrt{n}}{s} \quad (3.16)$$

In the spectral analysis, the most common test is the comparison test between the measured values of a group of samples with different component contents by using the spectral method and the reference method, which is called the paired *t*-test. Its essence is to judge whether the mean deviation (which should be close to zero) between the two methods is significantly different from the expected value (zero), that is, to judge whether there is a systematic error between the spectral method and the reference method. Paired *t*-test values are calculated as follows:

$$t = \bar{d} \frac{\sqrt{n}}{s} \quad (3.17)$$

where \bar{d} is the mean value of the corresponding difference between samples measured by the two methods, s is the standard deviation of the corresponding difference between samples measured by the two methods, and n is the number of samples.

Table 3.3 presents a set of comparative results of measuring olefin content in gasoline by fluorescence indicator method and near infrared (NIR) spectroscopy. Pairwise t -test is used to determine whether there is significant difference between the two methods. The results obtained were: $\bar{d} = -0.49$, $s = 1.56$, $t = 1.26$. Given the significance level $\alpha = 0.05$, the critical value $t_{(15, 0.05)} = 2.13$ was obtained. It can be seen that $|t|$ less than the critical value of 2.13 illustrates the two methods have no significant difference. In addition, t -test is also used to identify outliers of the calibration set in the spectral analysis.

F -test is mainly used to compare the variances of two sets of data. There are two situations: one is to use single-tail test to test whether method A is more precise than method B. The second is to determine whether there is a significant difference between the precision of test method A and B by double-tailed test. The expression of F -test is as follows:

Table 3.3 Comparison results of olefin content determination in gasoline by fluorescence indicator method and NIR spectroscopy

Sample	Olefin content / %		
	Fluorescence indicator method	NIR spectroscopy	Deviation between the two methods
Test01	33.08	30.90	-2.18
Test02	32.64	30.71	-1.93
Test03	28.99	31.49	2.50
Test04	28.06	29.75	1.69
Test05	26.95	27.87	0.92
Test06	26.59	23.66	-2.93
Test07	25.63	25.67	0.04
Test08	24.00	23.51	-0.49
Test09	23.70	22.60	-1.10
Test10	22.63	22.27	-0.36
Test11	27.24	26.52	-0.72
Test12	25.32	22.71	-2.61
Test13	23.71	22.97	-0.74
Test14	25.69	25.15	-0.54
Test15	21.23	23.06	1.83
Test16	25.00	23.81	-1.19
			$\bar{d} = -0.49$
			$s = 1.56$

$$F = \frac{s_1^2}{s_2^2} \quad (3.18)$$

In the above formula, the larger the numerator and the smaller the denominator, that is, to ensure that $F \geq 1$.

It should be noted that the significance level α should be the half of the double-tailed α if the single-tailed F distribution table is used for the double-tailed test.

3.5 Correlation Coefficient

In the spectral analysis, most of the mathematical problems encountered are the relationship between variables, such as the relationship between absorbance at different wavelengths, the relationship between absorbance and concentration, the relationship between properties and composition and the relationship between different properties. The relationship between variables can be mathematically divided into two categories: functional relationship and correlation relationship [9, 10]. Functional relations are deterministic relations, such as the relationship between circumference and radius. As long as the radius is determined, the circumference of a circle is determined. Correlation relationship is uncertain relationship, such as the relationship between height and weight, and the relationship between the teachers and graduation rates, which does exist a certain relation between variables, but not the one-to-one relationship. This relationship can be described by correlation (correlation coefficient R or determination coefficient R^2). Mathematical statistics methods such as regression analysis could be used to find out the inner relationship, so the relationship between these variables is also called a statistical relationship.

The linear correlation between the two variables is shown as the following three changes: (1) Positive correlation: when one variable increases or decreases, the other variable also increases or decreases accordingly; (2) Negative correlation: when one variable increases or decreases, the other variable decreases or increases; (3) No correlation: it means that the two variables are independent without linear correlation. In statistics, correlation coefficient R or determination coefficient R^2 is commonly used to describe the degree of linear correlation between two variables. The calculation formula of correlation coefficient R is as follows:

$$R = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} \quad (3.19)$$

where $x_i, y_i (i = 1, 2, \dots, n)$ is the sample value of two variables x and y ; \bar{x} and \bar{y} are the mean of the sample values of the two variables x and y , respectively; n is the number of samples of the two variables.

The values of R or R^2 are between -1 and 1. When the variables are fully correlated (the correlation coefficient is 1), it becomes a functional relationship. If there is no relationship between the variables, the correlation coefficient is close to or is zero.

In spectral analysis, the correlation coefficient is mainly used in two ways. One is to calculate the correlation between the predicted value (x_i) of the spectral method and the measured value (y_i) of the reference method for a group of samples. Another use is to calculate the correlation between the absorbance (x_i) of a wavelength and the concentration to be measured (y_i) of a set of samples.

3.6 Covariance and Covariance Matrix

The concept and operation of covariance and covariance matrix are often used in chemometrics [11]. The two variables x and y are measured for n times, and n groups of data (x_i, y_i) are obtained. Then the covariance of the two variables, $\text{cov}(x,y)$, is defined as follows:

$$\text{cov}(x, y) = \frac{1}{n - 1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) \tag{3.20}$$

If the correlation between variables x and y is poor, the absolute value of the covariance will be small, but the size of the covariance often depends on the ruler of the variable. The correlation coefficient is obtained by dividing the covariance by the product of the standard deviation of x and y .

$$R = \frac{\text{cov}(x, y)}{s_x s_y} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} \tag{3.21}$$

Assume that the data of m variables and n observations are shown in the Table 3.4. Variance of variable x_j can be calculated as follows:

Table 3.4 The data of m variables and n observations

The number of observations	Variables				
	x_1	x_2	x_3	...	x_m
1	x_{11}	x_{12}	x_{13}	...	x_{1m}
2	x_{21}	x_{22}	x_{23}	...	x_{2m}
...
n	x_{n1}	x_{n2}	x_{n3}	...	x_{nm}

$$s_j^2 = \frac{1}{n-1} (x_{ij} - \bar{x}_j)^2 \quad i = 1, 2, \dots, n; j = 1, 2, \dots, m \quad (3.22)$$

The covariance of variables x_j and x_k is calculated as follows:

$$\text{cov}(x_j, x_k) = \frac{1}{n-1} \sum_{i=1}^n (x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k) \quad i = 1, 2, \dots, n; j, k = 1, 2, \dots, m \quad (3.23)$$

The matrix composed of these variances and covariances is called variance-covariance matrix, or covariance matrix:

$$\mathbf{C} = \begin{bmatrix} s_1^2 & \text{cov}(1,2) & \cdots & \text{cov}(1,m) \\ \text{cov}(2,1) & s_2^2 & \cdots & \text{cov}(2,m) \\ \vdots & \vdots & \ddots & \vdots \\ \text{cov}(m,1) & \text{cov}(m,2) & \cdots & s_m^2 \end{bmatrix} \quad (3.24)$$

where $\text{cov}(x,x) = \mathbf{S}_x^2$, that is, the elements on the diagonal of the matrix are the variance of the variable. Since $\text{cov}(j,k) = \text{cov}(k,j)$, the covariance matrix is diagonal matrix.

The data in the above table is represented by a matrix:

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix} \quad (3.25)$$

The \mathbf{H} matrix is obtained by subtracting the mean value of each column from the element of each column of \mathbf{X} matrix:

$$\begin{bmatrix} x_{11} - \bar{x}_1 & x_{12} - \bar{x}_2 & \cdots & x_{1m} - \bar{x}_m \\ x_{21} - \bar{x}_1 & x_{22} - \bar{x}_2 & \cdots & x_{2m} - \bar{x}_m \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} - \bar{x}_1 & x_{n2} - \bar{x}_2 & \cdots & x_{nm} - \bar{x}_m \end{bmatrix} = \begin{bmatrix} h_{11} & h_{12} & \cdots & h_{1m} \\ h_{21} & h_{22} & \cdots & h_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ h_{n1} & h_{n2} & \cdots & h_{nm} \end{bmatrix} = \mathbf{H} \quad (3.26)$$

The mean of each of the columns in the \mathbf{H} matrix is zero. The covariance matrix \mathbf{C} can be obtained by mathematical processing of \mathbf{H} matrix:

$$\begin{aligned}
\text{cov}(\mathbf{X}) &= \frac{1}{n-1} \mathbf{H}^T \mathbf{H} = \frac{1}{n-1} \begin{bmatrix} h_{11} & h_{21} & \cdots & h_{n1} \\ h_{12} & h_{22} & \cdots & h_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ h_{1m} & h_{2m} & \cdots & h_{nm} \end{bmatrix} \begin{bmatrix} h_{11} & h_{12} & \cdots & h_{1m} \\ h_{21} & h_{22} & \cdots & h_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ h_{n1} & h_{n2} & \cdots & h_{nm} \end{bmatrix} \\
&+ \frac{1}{n-1} \begin{bmatrix} \sum_{i=1}^n h_{i1}^2 & \cdots & \cdots \\ \sum h_{i1} h_{i2} & \sum_{i=1}^n h_{i2}^2 & \vdots \\ \vdots & \vdots & \vdots \\ \cdots & \cdots & \sum_{i=1}^n h_{in}^2 \end{bmatrix} \\
&= \begin{bmatrix} s_1^2 & \text{cov}(1,2) & \cdots & \text{cov}(1,m) \\ \text{cov}(2,1) & s_2^2 & \cdots & \text{cov}(2,m) \\ \vdots & \vdots & \ddots & \vdots \\ \text{cov}(m,1) & \text{cov}(m,2) & \cdots & s_m^2 \end{bmatrix} = \mathbf{C} \quad (3.27)
\end{aligned}$$

Mahalanobis distance commonly used in spectral pattern recognition is a distance calculation method based on covariance, which is used to represent the similarity between unknown samples and certain types of samples [12]. Unlike Euclidean distance, it takes the interrelation of various characteristic variables (e.g., a piece of information about height will bring information about weight because the two variables are related) into account, and Mahalanobis distance is scale-invariant, namely, independent of the scale of measurement. In the spectral analysis, Mahalanobis distance is often used in the identification of outlier samples, cluster analysis, and discriminant analysis.

3.7 Multivariable Graph Representation

3.7.1 Spatial Representation of Samples

Spatial mapping of samples is helpful to study the relationship between samples. Usually, the spectral absorbance of a sample is used as the original characteristic variable to characterize the spatial distribution of a group of samples by two-dimensional or three-dimensional graphs. Since spectral variables are often with hundreds or even thousands of dimensions, it is necessary to select or compress and reduce the dimensionality of spectral variables (such as principal component analysis, PCA) before drawing. For example, Fig. 3.6 shows the two-dimensional spatial distribution of 210 cigarette samples of four different brands, which is characterized by the second-derivative absorbance at 5058 cm^{-1} and 4903 cm^{-1} [13].

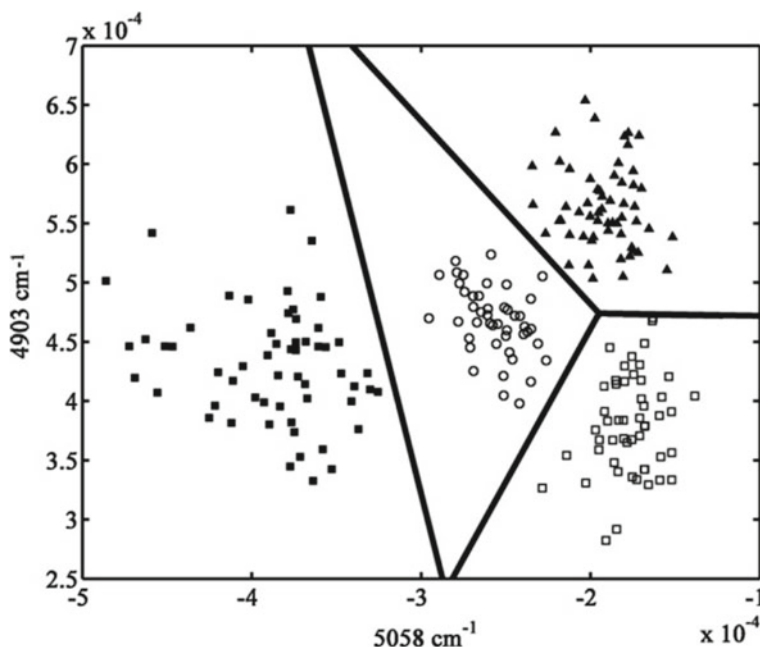


Fig. 3.6 Two-dimensional spatial distribution of 210 cigarette samples of four different brands at 5058 cm^{-1} and 4903 cm^{-1}

Figure 3.7 shows the three-dimensional spatial distribution of *Auricularia auricula* from four different producing areas by using PLS to reduce the dimensionality of NIR spectrum. It can be clearly seen from Fig. 3.7 that *Auricularia auricula* from different producing areas are grouped into one category, respectively

Fig. 3.7 Three-dimensional spatial distribution of the NIR spectra of four kinds of *Auricularia auricula* from different producing areas after dimensionality reduction by PLS

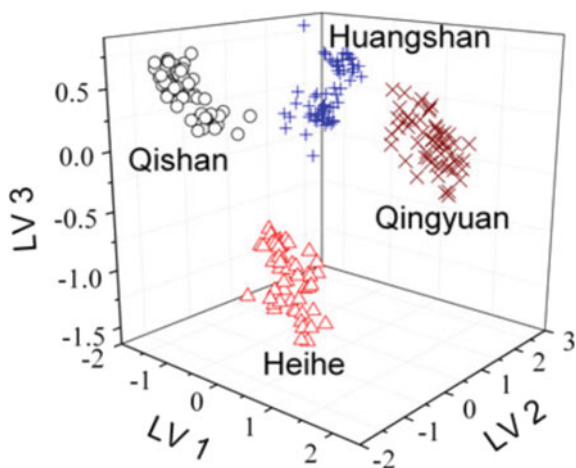
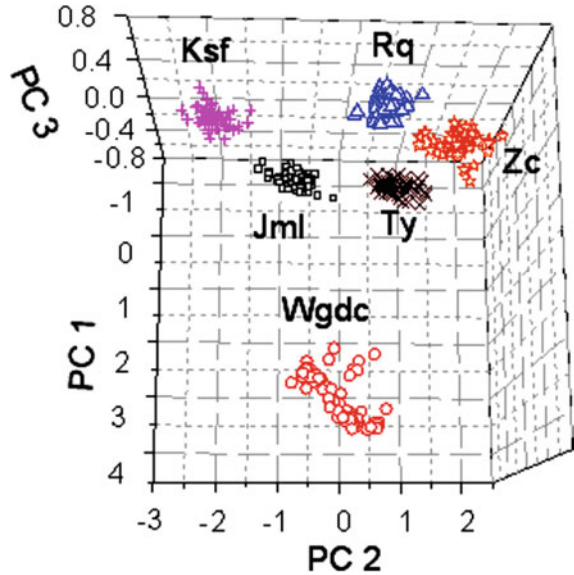


Fig. 3.8 Three-dimensional spatial distribution of the NIR spectra by PCA scores of different brands of instant noodles



[14]. Figure 3.8 shows a three-dimensional spatial distribution diagram of the classification of different brands of instant noodles by using NIR spectroscopy combined with PCA [17].

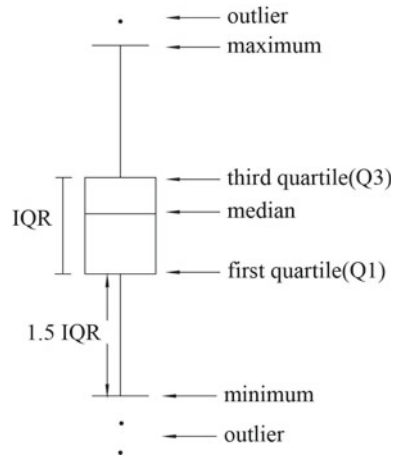
3.7.2 Box Plot

Box plot, also known as box plot, is used to reflect the central position and dispersion range of data distribution with the first quartile, median quartile, third quartile, and 1.5 times of the upper and the lower quartile range (IQR) in a set of data as shown in Fig. 3.9 [15]. Through the box-whisker diagram, we can roughly see whether the data has symmetry (skewness and tail weight), and judge abnormal samples. By drawing the box lines of multiple groups of data on the same coordinate, the distribution difference of each group of data can be clearly displayed [16].

Quartiles use three points to divide a set of data into four equal parts, each containing 25% of the data. What is commonly referred to as quartile data is the value at the 25% position and the value at the 75% position, which are, respectively, referred to as the lower quartile (Q_1) and the upper quartile (Q_3). When calculating the quartiles for ungrouped data, you first sort all the data, and then determine the position of the four quartiles. IQR is defined as follows:

$$\text{IQR} = Q_3 - Q_1 \quad (3.28)$$

Fig. 3.9 The elements of box plot



Two line segments similar to the median line are drawn at $Q_3 + 1.5IQR$ and $Q_1 - 1.5IQR$. These two line segments are the cut-off points of outliers, which are called the inner upper limit and inner lower limit. Points outside the inner limit represent outliers. Sometimes, two line segments are drawn at $Q_3 + 3IQR$ and $Q_1 - 3IQR$, and they are called outer upper limit and outer lower limit. The outliers between the inner limit and the outer limit are mild outliers, while those outside the outer limit are extreme outliers.

Figure 3.10 shows the deviation box plot of a set of samples of validation set predicted by PLS and Hierarchical Mixture of Linear Regressions (HMLR) models. From Fig. 3.10, it can be clearly seen that the HMLR algorithm is superior to the PLS algorithm [17]. Figure 3.11 is a turn-around time (TAT) box diagram of six analytical instruments in a laboratory, from which the operation of each instrument in different periods can be visually seen. It is helpful for the laboratory managers to grasp the use of the instruments, to take targeted measures to improve the efficiency of the instruments and enhance the visibility and controllability of device management [18].

3.7.3 Radar Chart

Radar chart is a widely used method for mapping multivariate data at present, and the relationship and rule among samples can be studied intuitively by using radar chart [19]. Suppose a sample data set have m variables, entirely standard drawing of radar chart is as follows: first, a circle is draw and then divided into m equal parts with the line from the center of circle to each part. The m lines will be seen as the coordinate axis. Make appropriate scale for each coordinate axis according to the value of each variable so that there is a scale on the corresponding coordinate axis for the value of each variable. For any sample, determine its coordinates on m axes,

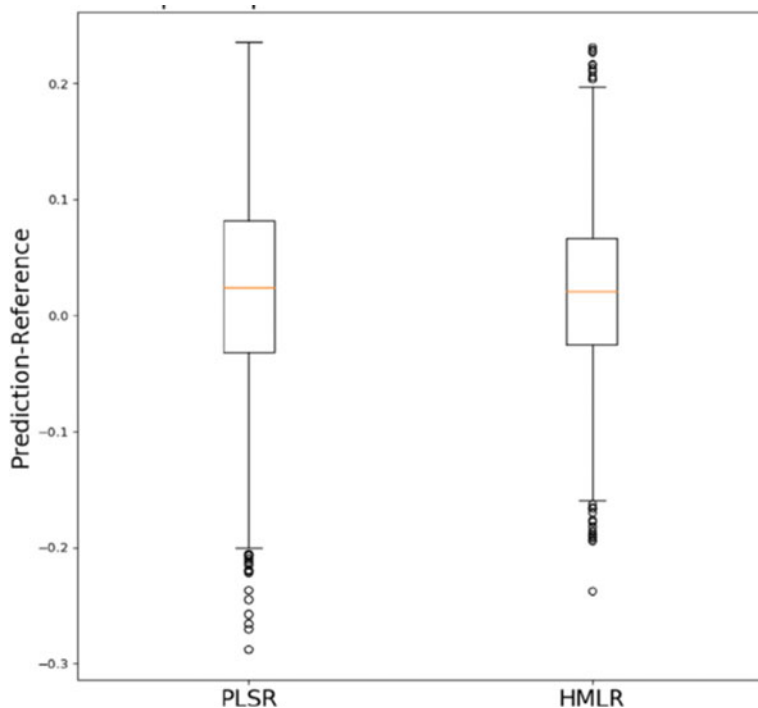


Fig. 3.10 Box plot of prediction errors of the two algorithms

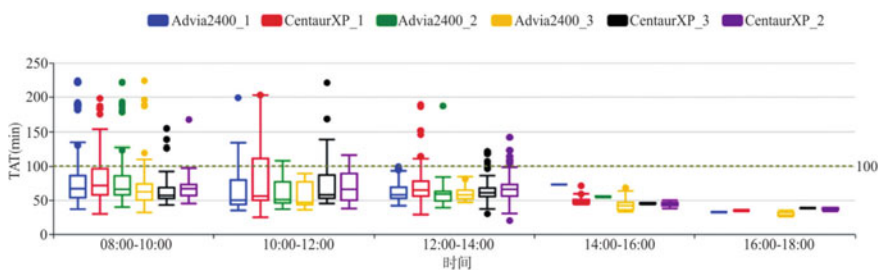


Fig. 3.11 Turn-around time (TAT) box plot of different analytical instruments in the laboratory over different periods [18]

respectively, point out its coordinates on each coordinate axis, and connect m points in turn to obtain an m -side shape. In this way, each sample can be represented by m -side shape. When observing the shape of each m -side shape, the similarity or inherent law between the samples can be analyzed. When the sample number is small, all the sample can be drawn in a circle; when the number of samples is large, a m -side shape can be drawn for each sample for analysis.

As shown in Fig. 3.12, Qin et al. used radar chart for the extraction of the NIR

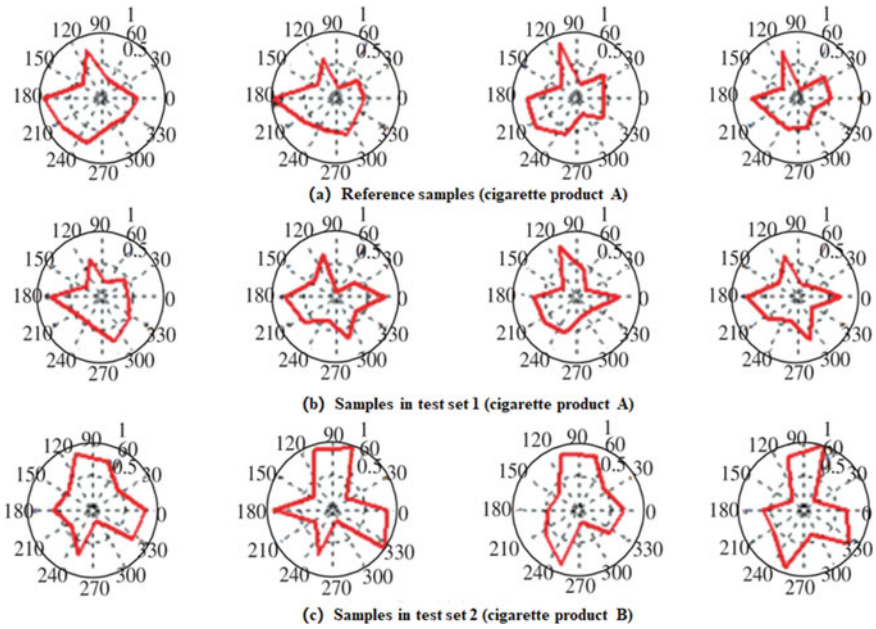


Fig. 3.12 PCA-radar chart generated from the spectra of different sample sets

spectra of tobacco leaves. The method firstly conducted PCA for the dimension reduction of the NIR spectral data, and the radar chart was then employed for the visual description of the quality stability trends of the products. Two characteristics of amplitude and angle of gravity vector were defined. Then the model of the quality stability and the abnormal type was established to guarantee the unification of integrity and fuzziness of the spectra data and achieve quality monitoring and counterfeit identification of the cigarette products [23].

As shown in Fig. 3.13, radar chart is widely used in aquaphotomics [20–24],

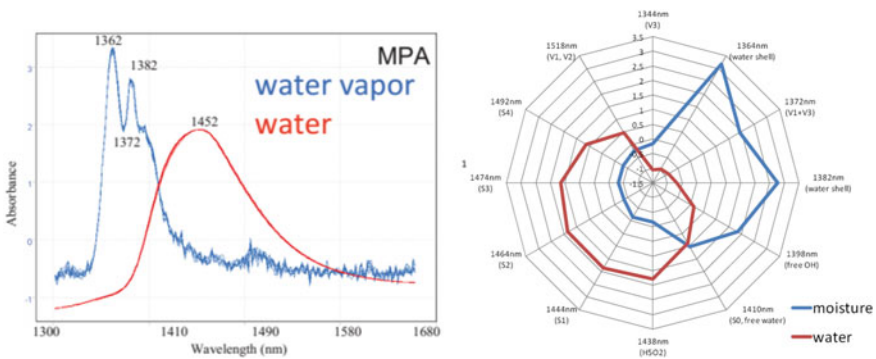


Fig. 3.13 NIR spectral radar chart of water and water vapor [30]

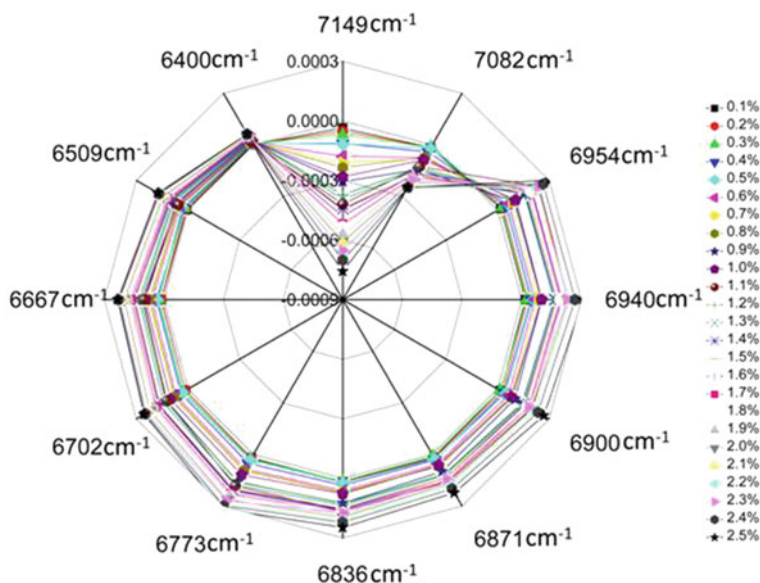


Fig. 3.14 NIR spectral radar chart of methanol-YPD medium solutions (with different methanol concentrations) [29]

which can obtain more information about changes in the chemical structure of water molecules [25–28]. As shown in Fig. 3.14, Li et al. made radar chart according to the absorbance of different low-content methanol-yeast extract peptone dextrose (YPD) medium solutions based on water matrix coordinates (WAMACS) of 12 water spectral peaks [28, 29]. As can be seen from Fig. 3.14, when methanol is added to YPD medium, the absorbance will still change even if the increase is 0.1%, indicating that the addition of methanol will disturb the covalent bond and hydrogen bond of water molecules in the solution, thus changing the spectrum of water. It can be observed from the radar chart that the band of 7149–6954 cm^{-1} makes a great contribution to distinguishing the solution of low concentration methanol-YPD medium. This may be due to the low methanol content and the presence of a large amount of water in the medium, which makes the weak hydrogen bond and symmetric O-H stretching vibration in the solution more abundant. Therefore, this band plays a dominant role in the water spectrum.

References

1. Liang YZ, Xu QS. Instrumental analysis of complex systems-white, gray and black analytical systems and their multivariate methods. Beijing: Chemical Industry Press; 2012.
2. Xu L. Chemometrics: principles and applications of some important methods. Beijing: Science Press; 2004.

3. Ni Y. Application of chemometrics in analytical chemistry. Beijing: Science Press; 2004.
4. Xu L, Shao X. Chemometric methods. 2nd ed. Beijing: Science Press; 2001.
5. Mark H, Workman J. Chapter 2-Elementary matrix algebra: Part 1. In: Mark H, Workman J (eds) Chemometrics in Spectroscopy, 2nd edn. Academic Press; 2018. p. 11–7.
6. Mark H, Workman J. Statistics in spectroscopy. 2nd ed. Amsterdam: Academic Press; 2003.
7. Miller J, Miller JC. Statistics and chemometrics for analytical chemistry. Pearson Education Limited, 2018.
8. Rutan SC. Chemometrics in analytical spectroscopy. Cambridge: Royal Society of Chemistry; 1996.
9. Gemperline P. Practical guide to chemometrics. 2nd ed. Boca Raton: CRC Press; 2006.
10. Varmuza K, Filzmoser P. Introduction to multivariate statistical analysis in chemometrics. 1st ed. Boca Raton: CRC Press; 2009.
11. Meyers RA, Mesilaakso M. Encyclopedia of analytical chemistry: applications, theory, and instrumentation. New Jersey: Wiley; 2009.
12. Mark HL, Tunnell D. Qualitative near-infrared reflectance analysis using Mahalanobis distances. *Anal Chem.* 1985;57:1449–56.
13. Moreira EDT, Pontes MJC, Galvão RKH, et al. Near infrared reflectance spectrometry classification of cigarettes using the successive projections algorithm for variable selection. *Talanta.* 2009;79:1260–4.
14. Liu F, He Y. Discrimination of producing areas of *Auricularia auricula* using visible/near infrared spectroscopy. *Food Bioprocess Technol.* 2011;4:387–94.
15. Otto M. Chemometrics: statistics and computer application in analytical chemistry. Wiley; 2016.
16. Dai B. A study on the application of PSO-SVM to the classification of wheat seed based on boxplot method. *J Hexi Univ.* 2018;34:19–25.
17. Cui C, Fearn T. Hierarchical mixture of linear regressions for multivariate spectroscopic calibration: an application for NIR calibration. *Chemom Intell Lab Syst.* 2018;174:1–14.
18. Lu L, Wen D, Zheng Y, et al. Development and application of quality index intelligent monitoring system on automatic assembly line. *Chin J Clin Lab Sci.* 2020;38:302–5.
19. Cui J, Gao H, Hong W. Research on identification of radix *pueraria* power based on radar graph feature extraction and near infrared spectra. *Chin High Technol Lett.* 2015;25:719–24.
20. Fan M, Zhao Y, Liu Y, et al. Aquaphotomics of near infrared spectroscopy. *Prog Chem.* 2015;27:242.
21. Tsenkova R. Aquaphotomics: dynamic spectroscopy of aqueous and biological systems describes peculiarities of water. *J Near Infrared Spectrosc.* 2009;17:303–13.
22. Tsenkova R, Munćan J, Pollner B, et al. Essentials of aquaphotomics and its chemometrics approaches. *Front Chem.* 2018;6:363.
23. Su T, Sun Y, Han L, et al. Revealing the interactions of water with cryoprotectant and protein by near-infrared spectroscopy. *Spectrochim Acta Part A Mol Biomol Spectrosc.* 2022(266), 120417.
24. Wang S, Wang M, Han L, et al. Insight into the stability of protein in confined environment through analyzing the structure of water by temperature-dependent near-infrared spectroscopy. *Spectrochim Acta Part A Mol Biomol Spectrosc.* 2022(267), 120581.
25. Kato Y, Munćan J, Tsenkova R, et al. Aquaphotomics reveals subtle differences between natural mineral, processed and aged water using temperature perturbation near-infrared spectroscopy. *Appl Sci.* 2021;11:9337.
26. Kaur H, Künemeyer R, McGlone A. Investigating aquaphotomics for temperature-independent prediction of soluble solids content of pure apple juice. *J Near Infrared Spectrosc.* 2020;28:103–12.
27. Muncan J, Tsenkova R. Aquaphotomics—From innovative knowledge to integrative platform in science and technology. *Molecules.* 2019;24:2742.
28. Li D, Li L, Quan S, et al. A feasibility study on quantitative analysis of low concentration methanol by FT-NIR spectroscopy and aquaphotomics. *J Mol Struct.* 2019;1182:197–203.

29. Li D. The determination of biomass, glycerol and methanol in fermentation process based on near infrared spectroscopy and aquaphotomics. Jinan: Shandong University; 2019.
30. Cattaneo T, Bazar G, Gowen A, et al. Water monitoring with hyperspectral techniques. *Transitional Waters Bull.* 2015;9:11–9.

Chapter 4

Spectral Preprocessing Methods



In addition to the chemical information of the sample itself, the spectrum also contains other irrelevant information, such as electrical noise, sample background, and stray light. Therefore, it is necessary to eliminate this irrelevant information by spectral preprocessing before modeling by chemometrics [1–3]. The commonly used spectral preprocessing methods include mean centering, auto-scaling, normalization, smoothing, derivatives, standard normal variate transformation, multiplicative scatter correction, Fourier transform, wavelet transform, orthogonal signal correction, and net analyte signal [4, 5].

4.1 Mean Centering

Mean centering (MC) is the average spectrum that subtracts the sample spectrum from the calibration set. The average column of the transformed spectral matrix of calibration set \mathbf{X} (number of samples $n \times$ number of wavelengths m) is zero. When spectral analysis models are built by multivariate calibration methods, this method correlates the change of the spectra rather than the absolute amount of the spectra with the change of properties or composition to be measured. Therefore, before establishing a spectral quantitative or qualitative model, mean centering is often used to increase the difference among the spectra of different samples. Thereby the robustness and predictive power of the model are improved [6]. While transforming spectral data in this way, the same processing for properties or compositional data is often performed. Mean centering is one of the most commonly used data preprocessing before establishing quantitative and qualitative models.

Average spectrum of the calibration set sample is first \bar{x} calculated.

$$\bar{x}_k = \frac{\sum_{i=1}^n x_{i,k}}{n} \quad (4.1)$$

where n is the number of samples in calibration set, $k = 1, 2, \dots, m$, and m is the number of wavelength points. For unknown sample spectrum \mathbf{x} ($1 \times m$), the centered spectrum $\mathbf{x}_{\text{centered}}$ by mean centering is given below:

$$\mathbf{x}_{\text{centered}} = \mathbf{x} - \bar{\mathbf{x}} \tag{4.2}$$

Figure 4.1 shows the original near-infrared (NIR) spectra of 80 corn samples. Figure 4.2 displays the spectra preprocessed by mean centering.

Fig. 4.1 Original NIR spectra of 80 corn samples

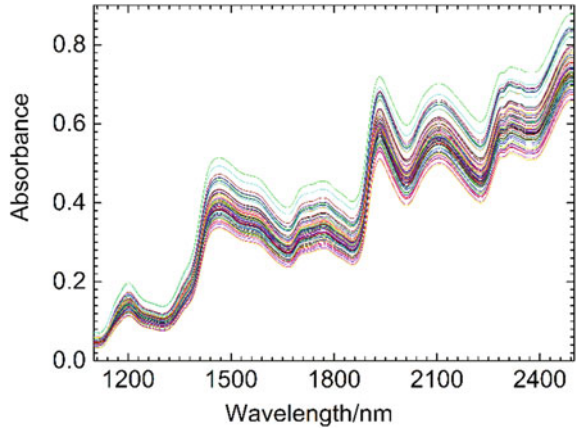
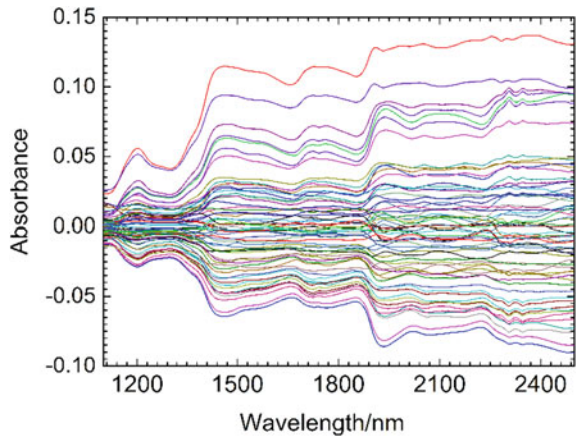


Fig. 4.2 NIR spectra of 80 corn samples preprocessed by mean centering



4.2 Auto-scaling

Auto-scaling is also known as mean variance, which is mean-centering divided by the standard deviations of spectral matrix in calibration set. Firstly, the average spectrum of the samples in calibration set $\bar{\mathbf{x}}$ is calculated. Then the standard deviation spectrum \mathbf{s} is also calculated for the samples in calibration set.

$$s_k = \sqrt{\frac{\sum_{i=1}^n (x_{i,k} - \bar{x}_k)^2}{n - 1}} \quad (4.3)$$

where n is the number of samples in calibration set, $k = 1, 2, \dots, m$, and m is the number of wavelength points.

The mean centralization is first performed for the unknown spectrum \mathbf{x} ($1 \times m$), and then the variables are scaled by dividing the standard deviation spectrum \mathbf{s} to obtain the auto-scaled spectrum.

$$\mathbf{x}_{\text{autoscaled}} = \frac{\mathbf{x} - \bar{\mathbf{x}}}{\mathbf{s}} \quad (4.4)$$

Auto-scaled spectra are with a column mean of zero and a variance of 1. Auto-scaling is particularly useful for modeling low-concentration components because it gives the same weights to all wavelength variables in the spectra. Thus, auto-scaling is a commonly used spectral data transformation. Figure 4.3 shows the auto-scaled NIR spectra of 80 corn samples.

Fig. 4.3 NIR spectra of 80 corn samples preprocessed by auto-scaling

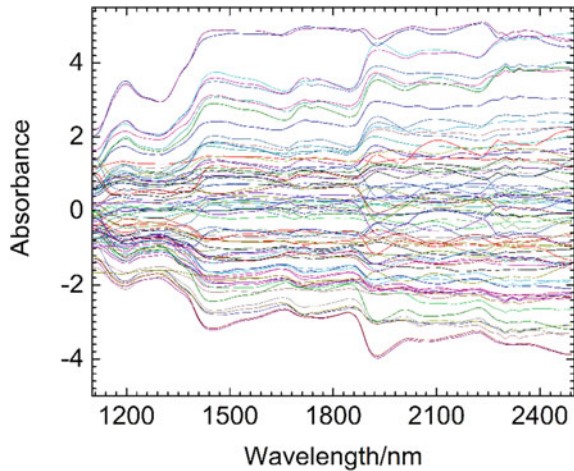
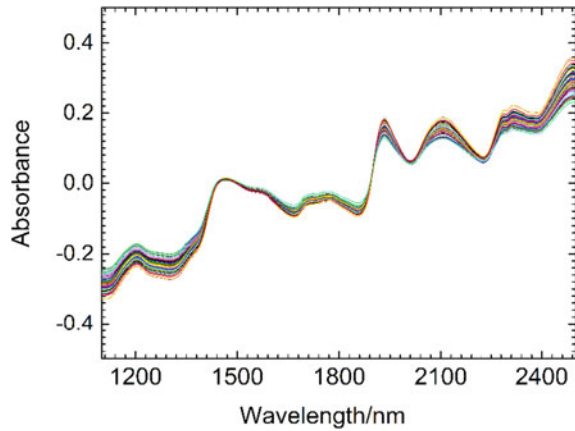


Fig. 4.4 NIR spectra of 80 corn samples preprocessed by normalization



4.3 Normalization

Normalization has many algorithms, such as area normalization, maximum normalization, and average normalization. In spectral analysis, the most commonly used normalization is vector normalization. For a spectrum \mathbf{x} ($1 \times m$), the vector normalization algorithm is as follows:

$$\mathbf{x}_{\text{normalized}} = \frac{\mathbf{x} - \bar{x}}{\sqrt{\sum_{k=1}^m x_k^2}} \quad (4.5)$$

where $\bar{x} = \frac{\sum_{k=1}^m x_k}{m}$, m is the number of wavelength points, $k = 1, 2, \dots, m$. This is often used to correct spectral changes caused by small light path differences. Figure 4.4 is NIR spectra of 80 corn samples preprocessed by normalization.

4.4 Smoothing

The spectral signals obtained by the spectrometer contain both useful information and random error, i.e., noise. Signal smoothing is one of the most commonly used de-noising methods. It mainly takes the average of multiple measurements to reduce the noise and improve the signal-to-noise ratio when the noise contained in the spectrum is zero mean random white noise. The commonly used signal smoothing methods are moving average smoothing and Savitzky-Golay convolution smoothing.

Moreover, Fourier transformation and wavelet transformation can also be used for spectral de-noising.

4.4.1 Moving Average Smoothing

The moving average smoothing is shown in Fig. 4.5, it selects a window of smoothing with a certain width ($2w+1$), with an odd number of wavelength points in each window. Then the measured value of k wavelength point is replaced by the central wavelength point k in the window and the mean measured \bar{x}_k at w before and after k . At last, k from left to right is moved to complete the smoothing of all points.

$$x_{k,smooth} = \bar{x}_k = \frac{1}{2w + 1} \sum_{i=-w}^{+w} x_{k+i} \tag{4.6}$$

When using the moving average smoothing method, the window of smoothing width is an important parameter. If the window width is too small, the de-noising effect is not perfect. If the window width is too large and its mean is calculated, the useful information is smoothed while smoothing the noise, resulting in the distortion of the spectral signal (Figs. 4.6 and 4.7).

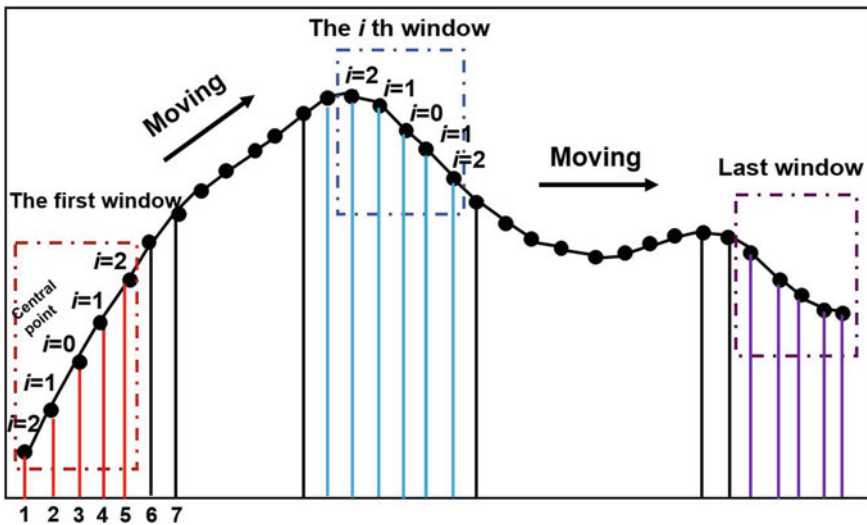


Fig. 4.5 Schematic diagram of window moving smoothing method

Fig. 4.6 Smoothing effects of moving average smoothing with different window size

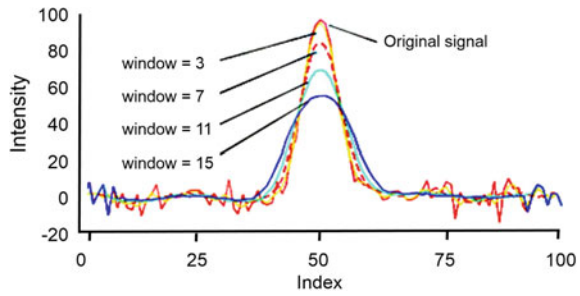
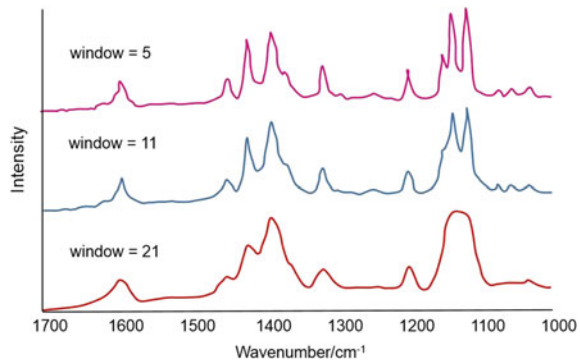


Fig. 4.7 Loss of spectral information caused by over-smoothing



4.4.2 Savitzky-Golay Convolution Smoothing

Savitzky-Golay (S-G) convolution smoothing [7] is also known as polynomial smoothing. The average after smoothing at wavelength k is as follows:

$$x_{k,smooth} = \bar{x}_k = \frac{1}{H} \sum_{i=-w}^{+w} x_{k+i} h_i \tag{4.7}$$

In Eq. 4.7, h_i and H are the smoothing factor and the normalization factor, respectively, where $H = \sum_{i=-w}^{+w} h_i$. The purpose of multiplying each measurement by the smoothing factor h_i is to reduce the effect of smoothing on useful information. h_k can be obtained by using polynomial fit based on the principle of least squares. Table 4.1 shows the cubic polynomial SG smoothing coefficients.

The basic idea of S-G convolution smoothing is similar to moving average smoothing. The difference between them is that S-G convolution smoothing is polynomials least square fitting of the data in the moving window, which is essentially a weighted average with more emphasis on the central role of the central point. S-G convolution smoothing is a widely used de-noising method. The effect of moving window width (often referred to number of smoothing points) is significantly lower than the moving average smoothing (Fig. 4.8). Figure 4.9 shows the NIR spectrum

Table 4.1 Savitzky-Golay smoothing coefficients for cubic polynomial with different moving window size

Points	25	23	21	19	17	15	13	11	9	7	5
-12	-253										
-11	-138	-42									
-10	-33	-21	-171								
-9	62	-2	-76	-136							
-8	147	15	9	-51	-21						
-7	222	30	84	24	-6	-78					
-6	287	43	149	89	7	-13	-11				
-5	343	54	204	144	18	42	0	-36			
-4	387	63	249	189	27	87	9	9	-21		
-3	422	70	284	224	34	122	16	44	14	-2	
-2	447	75	309	249	39	147	21	69	39	3	-3
-1	462	78	324	268	42	162	24	84	54	6	12
0	467	79	329	269	43	167	25	89	59	7	17
1	462	78	324	264	42	162	24	84	54	6	12
2	447	75	309	249	39	147	21	69	39	6	-3
3	422	70	284	224	34	122	16	44	14	-2	
4	387	63	249	189	27	87	9	9	-21		
5	343	54	204	144	18	42	0	-36			
6	287	43	149	89	7	-13	-11				
7	222	30	84	24	-6	-78					
8	147	15	9	-51	-21						
9	62	-2	-76	-136							
10	-33	-21	-171								
11	-138	-42									
12	-253										
	5175	805	3059	2261	323	1105	143	429	231	21	35

with noise by cubic polynomials with different moving window widths. It can be seen that the smoothing effect is improved significantly with the number of smoothing points increasing.

4.4.3 Fourier Transform and Wavelet Transform

The Fourier transform for the spectrum is to decompose the spectrum into the sum of sine waves with different frequencies. Compared with the useful signal, instrument noise has the characteristics of small amplitudes and high frequencies. Therefore, high frequencies are deleted. The original spectrum is reconstructed by inverse

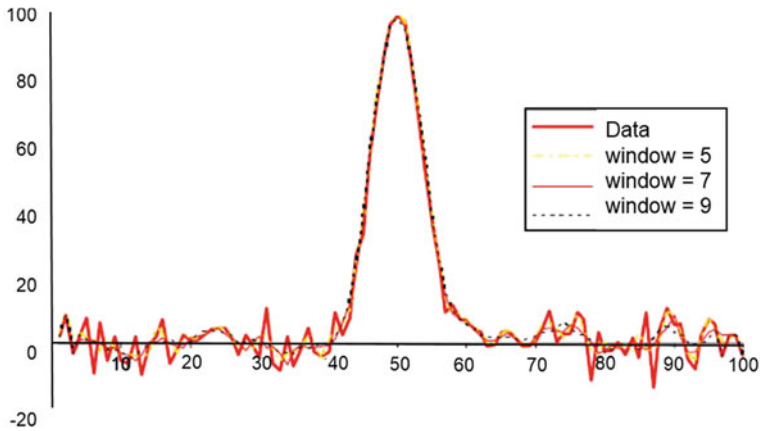


Fig. 4.8 Smoothing effects of quadratic polynomial S-G method with different window sizes

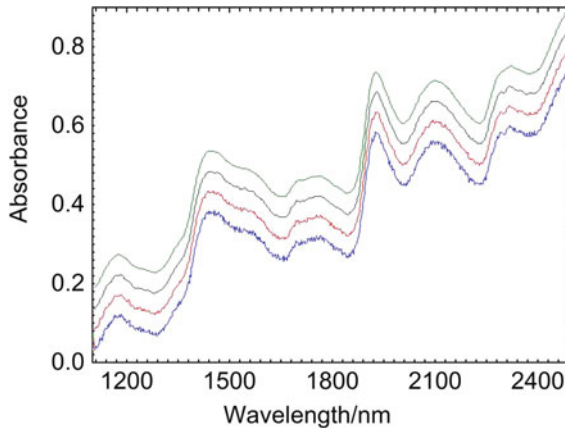


Fig. 4.9 Original and smoothing spectra by S-G smoothing method. From bottom to top are the noised spectrum, the smoothing spectra by S-G cubic polynomial with 5 point, 11 point, and 23 point, respectively. An artificial shift was added to absorbance for distinguishing different spectra

Fourier transform from low frequencies. Most spectral noise is eliminated and the spectrum is smoothed by this method.

The basic idea of wavelet transform for de-noising is similar to the Fourier transform. Wavelet coefficients corresponding to the high-frequency scales obtained by the wavelet transform are removed, and then the spectrum after de-noising is reconstructed.

The principles of spectral de-noising of Fourier transform and wavelet transform are detailed in Sects. 4.11 and 4.12 of this book.

4.5 Continuum Removed

Continuum removed, also known as envelope removal, is a spectral processing that effectively enhances the absorption characteristics of interest [8, 9]. It effectively highlights the absorption and reflection characteristics of the spectral curves and normalizes the reflectivity between 0 and 1, which facilitates the comparison of characteristic values with other spectral curves, thus the characteristic bands for quantitative and qualitative analysis are extracted. The curve connecting the protruding peak points on the spectral curve point-by-point linear is defined as the “Continuous” or “envelope”. Moreover, the outer angle of the line at the peak point is greater than 180° . The value on the original spectral curve is divided by the corresponding value on the envelope, namely, the spectral de-envelope. Intuitively, envelopes are equivalent to the “shell” of a spectral curve.

Continuum removed is widely used in the field of reflection spectra. For example, Han et al. [10] screened out the characteristic variables related to the soil organic matter content and established a calibration model to predict the soil organic matter content by the hyperspectral reflectance of soil samples to use continuum removed. Li et al. [11] improved the correlation between the raw canopy spectrum and the nitrogen content of the leaves by processing the spectrum of wheat leaves by continuum removed.

Continuum removed has many algorithms, of which Clark et al. are commonly used to propose shell coefficients [12–14]. The implementation steps are as follows:

- (1) All the local maxima on the spectral curve, that is, the protruding envelope “peak” values are calculated by derivation. Then the maximum is obtained by comparing all the local maxima.
- (2) Taking the maximum point as an endpoint of the envelope, the slopes of the lines connecting between this point and the maximal values in long wave direction (the direction of the increase in wavelength) are calculated. Then, the point of the maximum slope as the next endpoint of the envelope, this point is regarded as the starting point of the cycle until the last point.
- (3) The maximum point as an endpoint of the envelope, a similar calculation is done in the short-wave direction (the direction of the decrease in wavelength). Then, taking the point with the minimum slope as the next endpoint of the envelope, this point is regarded as the starting point for cycling until the last point;
- (4) The envelope is formed by connecting all endpoints along the direction of increasing wavelength. The reflectance of the corresponding band on the envelope is divided by the actual spectral reflectance to obtain the normalized value of envelope elimination.

Continuum removed can also be calculated by the following equations for the selected spectral interval:

$$R'_j = \frac{R_j}{R_{\text{start}} + k(\lambda_j - \lambda_{\text{start}})} \quad (4.8)$$

$$k = \frac{R_{\text{end}} - R_{\text{start}}}{\lambda_{\text{end}} - \lambda_{\text{start}}} \quad (4.9)$$

where R_j is the reflectance at wavelength j ; R_j' is the reflectance after R_j is processed; λ_j is the wavelength at j ; λ_{start} and λ_{end} are the starting and ending wavelengths of the selected spectral range, respectively; R_{start} and R_{end} are the reflectance at the starting and ending wavelengths of the selected spectral range, respectively; and k is the slope between the starting wavelength and the ending wavelength of the spectrum.

4.6 Adaptive Iteratively Reweighted Penalized Least Squares

Adaptive iteratively reweighted penalized least squares (airPLS) is a stepwise approximation background fitting algorithm [15]. It achieves background deduction by introducing parameters that adjust the smoothness and fidelity of the curve to obtain a spectrum with subtracted background. The algorithm consists of two main aspects: a penalized least squares algorithm for the smoothing of the signals and an adaptive iteration to convert the penalized process into a penalized least squares algorithm for the baseline estimation.

(1) Penalized least squares algorithm

If \mathbf{x} is the spectral analysis signal, \mathbf{z} is the fitted vector, and the number of wavelength points of them are both m . The accuracy of \mathbf{x} and \mathbf{z} can be expressed as the sum of the squared differences of their errors.

$$F = \sum_{i=1}^m (x_i - z_i)^2 \quad (4.10)$$

The roughness of a vector \mathbf{z} can be expressed as the sum of the squared differences of its two adjacent terms:

$$R = \sum_{i=2}^m (\mathbf{z}_i - \mathbf{z}_{i-1})^2 = \sum_{i=1}^{m-1} (\Delta \mathbf{z}_i)^2 \quad (4.11)$$

The balance between fidelity and roughness can be expressed in terms of fidelity plus a penalty for roughness as follows:

$$Q = F + \lambda R = \|\mathbf{x} - \mathbf{z}\|^2 + \lambda \|\mathbf{Dz}\|^2 \quad (4.12)$$

where λ is an adjustable parameter. The λ is larger, the fitted \mathbf{z} is smoother. D corresponds to the difference matrix, e.g., $D\mathbf{z} = \Delta\mathbf{z}$. By finding the partial derivative of the vector \mathbf{z} and making it equal to 0 ($\partial Q/\partial\mathbf{z} = 0$), an easily solvable equation for the linear system can be obtained.

$$(\mathbf{I} + \lambda D^T D)\mathbf{z} = \mathbf{x} \quad (4.13)$$

The above equation is a smoothing method by a penalized least squares algorithm. A weight vector w of fidelity is introduced to perform baseline calibration by the penalized least squares algorithm. Then w is placed at the corresponding position of 0 in the peaked segment, so that the fidelity of \mathbf{z} with respect to m becomes

$$F = \sum_{i=1}^m w_i (\mathbf{x}_i - \mathbf{z}_i)^2 = (\mathbf{x} - \mathbf{z})' W (\mathbf{x} - \mathbf{z}) \quad (4.14)$$

where W is the diagonal matrix of w_i on the diagonal. The above equation becomes

$$(W + \lambda D^T D)\mathbf{z} = W\mathbf{x} \quad (4.15)$$

The above linear equation system is solved to obtain the fitted vector \mathbf{z} .

$$\mathbf{z} = (W + \lambda D^T D)^{-1} W\mathbf{x} \quad (4.16)$$

(2) Adaptive iterative reweighting

The adaptive iterative reweighting is similar to weighted least squares and iterative penalized least squares. However, the weights are calculated by different methods. Moreover, the smoothing of the fitted baseline is controlled by adding a penalty. Each step of the adaptive iterative reweighting process involves solving the following weighted least squares problem:

$$Q^t = \sum_{i=1}^m w_i^t |\mathbf{x}_i - \mathbf{z}_i^t|^2 + \lambda \sum_{j=2}^m |\mathbf{z}_j^t - \mathbf{z}_{j-1}^t|^2 \quad (4.17)$$

where the weight vector w is obtained by the adaptive iterative method, and an initial value $w_0 = 1$ is given at the beginning. The w for each iteration step can be obtained as follows after assignment:

$$w_i^t = \begin{cases} 0 & \mathbf{x}_i \geq \mathbf{z}_i^{t-1} \\ e^{\frac{t(\mathbf{x}_i - \mathbf{z}_i^{t-1})}{|d^t|}} & \mathbf{x}_i < \mathbf{z}_i^{t-1} \end{cases} \quad (4.18)$$

where the vector d_t is the negative part of the difference between the vector \mathbf{x} and the last fitted background \mathbf{z}^{t-1} over the course of t iterations.

The fitted value \mathbf{z}^{t-1} is the baseline value from the previous iteration. The i th point is considered to be part of the peak when its value is greater than the selected baseline value. Therefore, its weight is set to zero to ignore its role in the next iteration of the fit. Iterations and reweights are performed continuously so as to eliminate the points within peak positions automatically and gradually and background points are retained in the weight vector w .

The termination conditions for the iterations are usually:

$$|d_t| < 0.001 \times |\mathbf{x}| \quad (4.19)$$

The airPLS algorithm has been widely used to eliminate baseline drift in Raman spectra caused by fluorescence and it has become a common baseline correction method. Excellent results have been obtained by this method for the baseline correction of NIR, LIBS and other spectra [16, 17].

Other methods are used for spectral baseline correction include polynomial fitting (ModPoly) [18], iterative polynomial smoothing (IPSA) [19], adaptive minimal-extreme baseline fitting (AdaptMinmax) [20], asymmetric weighted penalized least squares (AsLS) [21], asymmetric reweighted penalized least squares (ArPLS) [22], and locally symmetric reweighted penalized least squares (LSRPLS) [23]. The above baseline correction methods are single raw spectrum input and single corrected output (SISO). Another type of baseline correction method is multiple raw spectral inputs and single corrected output (MISO). Yao et al. [24] have proposed the BRACK method based on independent component analysis (ICA) and mixing entropy criterion.

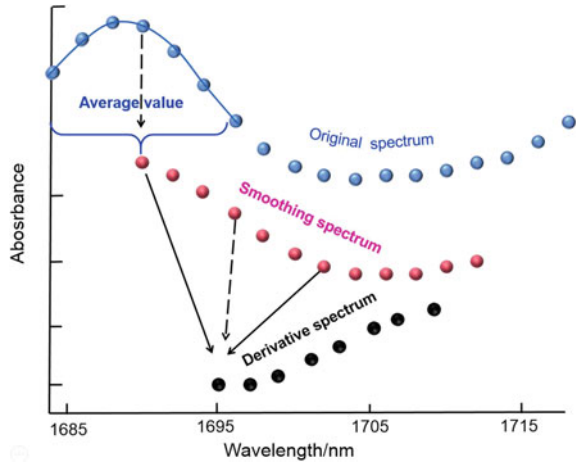
4.7 Derivative

First derivative (1st Der) and second derivative (2nd Der) are commonly used preprocessing methods for baseline correction and resolution enhancement in spectral analysis. There are generally two methods for spectral derivation: direct differences and S-G derivative methods.

4.7.1 Norris Method

The direct difference is the simplest derivative method for discrete spectra. For a discrete spectrum \mathbf{x}_k , first and second derivative spectra at wavelength k with gap size g are calculated as follows, respectively.

Fig. 4.10 Schematic diagram of Norris derivative method (first derivative with 7-point smoothing and 3-point difference width)



First derivative:

$$x_{k,1st} = \frac{x_{k+g} - x_{k-g}}{g} \tag{4.20}$$

Second derivative:

$$x_{k,2nd} = \frac{x_{k+g} - 2x_k + x_{k-g}}{g^2} \tag{4.21}$$

To eliminate the noise caused by the spectral transformation, the original spectrum is often smoothed before derivative. This method was first proposed by Norris et al., so it is often called the Norris derivative method [25]. As shown in Fig. 4.10, the Norris derivative of 7-point smoothing and 3-point gap size are performed on the spectrum. The spectrum is denoised by the moving average smoothing with a window width of 7 points, and then by the direct difference with a gap size of 3 points.

For spectra with high resolution and many wavelength sampling points, the derivative spectrum obtained by direct difference is not much different from the actual. However, the derivative obtained by this method has large errors for the spectra of sparse wavelength sampling points. Then the S-G convolution derivative method can be used.

4.7.2 Savitzky-Golay Convolution for Derivative Calculation

S-G convolution smoothing can also be used to obtain spectral derivatives. The derivative coefficients which are similar to smooth coefficients can be calculated by least squares. Tables 4.2 and 4.3 show first and second derivative coefficients

Table 4.2 S-G first derivative coefficients with cubic polynomials and different moving window sizes

Point	25	23	21	19	17	15	13	11	9	7	5
-12	30,866										
-11	8602	3938									
-10	-8525	815	84,075								
-9	-20,982	-1518	10,032	6936							
-8	-29,236	-3140	-43,284	68	748						
-7	-33,754	-4130	-78,176	-4648	-98	12,922					
-6	-35,003	-4567	-96,947	-7481	-643	-4121	1,133				
-5	-33,450	-4530	-101,900	-8700	-930	-14,150	-660	300			
-4	-29,562	-4098	-95,338	-8574	-1002	-18,334	-1578	-294	86		
-3	-23,806	-3350	-79,564	-7372	-902	-17,842	-1796	-532	-142	22	
-2	-16,649	-2365	-56,881	-5363	-673	-13,843	-1489	-503	-193	-67	1
-1	-8558	-1222	-29,592	-2816	-358	-7506	-832	-296	-126	-58	-8
0	0	0	0	0	0	0	0	0	0	0	0
1	8558	1222	29,592	2816	358	7506	832	296	126	58	8
2	16,649	2365	56,881	5363	673	13,843	1489	503	193	67	-1
3	23,806	3350	79,564	7372	902	17,842	1796	532	142	-22	
4	29,562	4098	95,338	8574	1002	18,334	1578	294	-86		
5	33,450	4530	101,900	8700	930	14,150	660	-300			
6	35,003	4567	96,947	7481	643	4121	-1133				
7	33,754	4130	78,176	4648	98	-12,922					
8	29,236	3140	43,284	-68	-748						
9	20,982	1518	-10,032	-6936							
10	8525	-815	-84,075								
11	-8602	-3938									
12	-30,866										
	1,776,060	197,340	3,634,092	255,816	23,256	334,152	24,024	5148	1188	252	12

obtained by S-G with cubic polynomials, respectively.

The derivative spectra can effectively eliminate the interference from baseline and other backgrounds to distinguish overlapping peaks, and improve resolution and sensitivity. However, it also introduces noise and reduces the signal-to-noise ratio. In the derivation, the selection of difference width (often known as derivative or differential points) is quite important. If the difference width is too small, the noise is large, which affects the prediction ability of the built analytical model. If the difference width is too large, the spectrum becomes excessive smoothing. Then a lot of details in the spectrum are lost. The optimal value can be selected by plotting the difference width with the root mean squared error of calibration (RMSEC) or root mean square error of prediction (RMSEP). Moreover, it is generally considered that the difference width should not exceed 1.5 times the half-peak width of the curve

Table 4.3 S-G second derivative coefficients with cubic polynomials and different moving window sizes

Points	25	23	21	19	17	15	13	11	9	7	5
-12	92										
-11	69	77									
-10	48	56	190								
-9	29	37	133	51							
-8	12	20	82	34	40						
-7	-3	5	37	19	25	91					
-6	-16	-8	-2	6	12	52	22				
-5	-27	-19	-35	-5	1	19	11	15			
-4	-36	-28	-62	-14	-8	-8	2	6	28		
-3	-43	-35	-83	-21	-15	-29	-5	-1	7	5	
-2	-48	-40	-98	-26	-20	-44	-10	-6	-8	0	2
-1	-51	-43	-107	-29	-23	-53	-13	-9	-17	-3	-1
0	-52	-44	-110	-30	-24	-56	-14	-10	-20	-4	-2
1	-51	-43	-107	-29	-23	-53	-13	-9	-17	-3	-1
2	-48	-40	-98	-26	-20	-44	-10	-6	-8	0	2
3	-43	-35	-83	-21	-15	-29	-5	-1	7	5	
4	-36	-28	-62	-14	-8	-8	2	6	28		
5	-27	-19	-35	-5	1	19	11	15			
6	-16	-8	-2	6	12	52	22				
7	-3	5	37	19	25	91					
8	12	20	82	34	40						
9	29	37	133	51							
10	48	56	190								
11	69	77									
	26,910	17,710	33,649	6783	3876	6188	1001	429	462	42	7

peak. Figures 4.11 and 4.12 are spectra of 80 corn NIR spectra preprocessed by the S-G first and second derivatives, respectively.

4.7.3 Wavelet Transform for Derivative Calculation

Wavelet transform is used in the calculation of the spectral derivatives by the special properties of the wavelet basis function [26, 27]. It realized mainly by continuous wavelet transform and discrete wavelet transform.

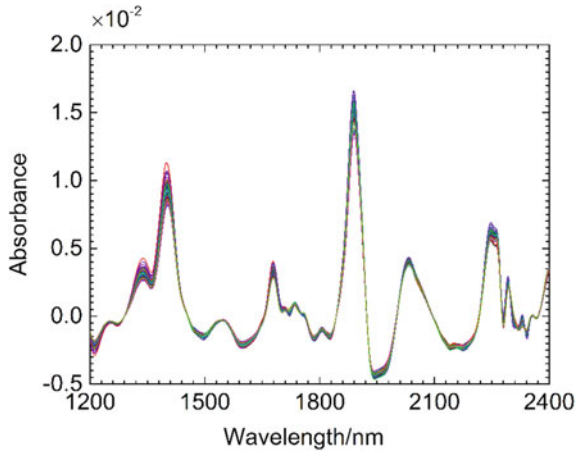


Fig. 4.11 S-G first derivative NIR spectra of 80 corn samples with 11-point cubic polynomial

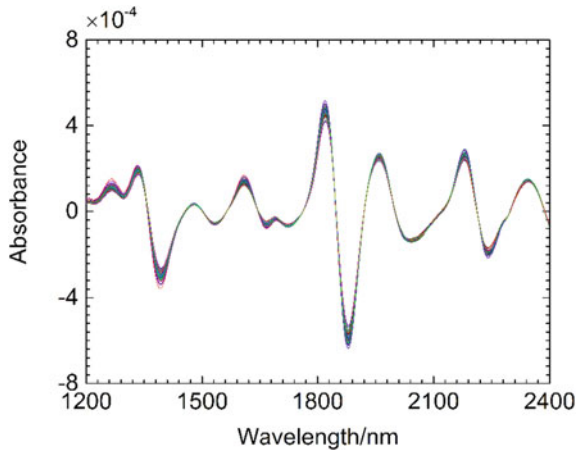


Fig. 4.12 S-G second derivative NIR spectra of 80 corn samples with 21-point cubic polynomial

(1) Continuous wavelet transform

The derivative of the spectrum can be approximated by continuous wavelet transform with specific wavelet functions. For example, Haar wavelet is a step function, which is convolution with the spectrum and becomes the first derivative. Then the second derivative can be obtained by continuing convolution. Since the wavelet transform has the function of smoothing and filtering noise, this method can solve the noise problem when the higher-order derivative is calculated. Figure 4.13 shows the second derivative spectral obtained by the sym2 wavelet function (scale 12) and the

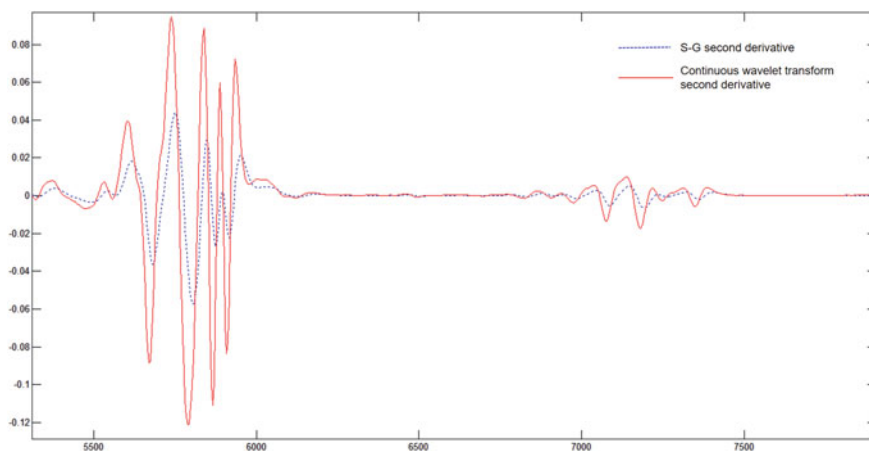


Fig. 4.13 Second derivative spectra by CWT with sym2 and 12 scale and SG with 13 points

commonly used S-G second derivative (13 points). It can be seen that the characteristics of the second derivative spectral obtained by the wavelet transform are more significant and the signal intensity is stronger. Figure 4.14 shows the NIR spectra of

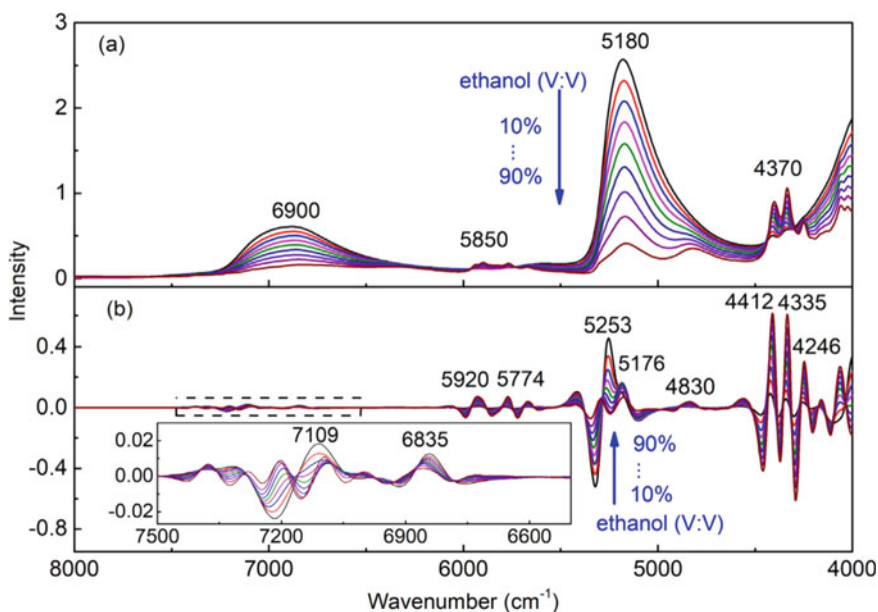


Fig. 4.14 NIR spectra of the ethanol-water mixtures and their fourth derivative spectra obtained by WT

the ethanol-water mixtures and their fourth derivative spectra obtained by wavelet transform [28].

(2) Discrete wavelet transform

Alexander et al. used the Daubechies group wavelet function (expressed by D_{2m}) to calculate the derivative of analytical signal [29]. The first derivative of the spectral vector \mathbf{x} can be expressed as

$$\mathbf{X}^{(1)} = \mathbf{C}_{1,D_{2m}} - \mathbf{C}_{1,D_{2\tilde{m}}} \cdots m \neq \tilde{m} \quad (4.22)$$

where m is a positive integer in the range of 1 to 10. $\mathbf{C}_{1,D_{2m}}$ and $\mathbf{C}_{1,D_{2\tilde{m}}}$ are the approximation signals of \mathbf{x} obtained by D_{2m} and $D_{2\tilde{m}}$ discrete wavelet transform, respectively.

The higher-order derivative of the spectrum can be calculated by taking the lower-order as an input to discrete wavelet transform. The principle of wavelet transform to decompose the spectrum is detailed in Sect. 4.12 of this book.

The wavelet basis functions used for first derivative calculation are Daubechies 1, bior 1.1, bior 1.3, Gaussian 1. The wavelet basis functions for second derivative calculation are Daubechies 1, Symlets 2, Coiflet 1, bior 2.2, bior 2.6, Gaussian 1, and Hat Mexican. The wavelet basis functions used for third derivative calculation have Daubechies 3, Symlets 3, bior 3.1, bior 3.5, and Gaussian 3. The wavelet basis functions used for fourth derivative calculation are Daubechies 4, Symlets 4, bior 4.4, Coiflet 2, and Gaussian 4.

In order to improve the signal-to-noise ratio, the derivative methods proposed by Li et al. based on singular perturbation and Taylor series can also be used for high-order derivative spectrum calculation [30].

4.7.4 Fractional Derivative

Traditional spectral derivatives use integer orders (commonly used are first and second). However, it has been reported that the optimal results of spectral derivatives are not all at integer order derivatives, but between zero and first derivatives or between first and second derivatives [31]. Compared with integer order derivatives, fractional order derivatives can more accurately reveal the changes of spectral details with the change of the order. Therefore, they can better characterize the details of the spectrum and balance the contradiction between spectral resolution and signal intensity.

Fractional derivative has a variety of algorithms, of which fractional order S-G derivative method (FOSGD) is more commonly used. The specific algorithm can be found in the relevant literature [32].

4.8 Standard Normal Variate and De-Trending

The standard normal variate transformation (SNV) is primarily used to eliminate the effects of solid particle size, surface scattering, and optical path changes on the NIR diffuse reflection spectra [33]. The influence of granularity size on the diffuse spectra is shown in Fig. 4.15 and particle size on the NIR spectra of wheat grains and wheat flour in Fig. 4.16. SNV has the same formula as the normalization algorithm. However, the normalizations algorithm processes spectral columns of the spectral matrix while SNV processes spectral rows of the spectral matrix.

SNV transforms the spectrum as follows:

$$x_{SNV} = \frac{x - \bar{x}}{\sqrt{\frac{\sum_{k=1}^m (x_k - \bar{x})^2}{(m-1)}}} \tag{4.23}$$

where $\bar{x} = \frac{\sum_{k=1}^m x_k}{m}$, m is the number of wavelength points, $k = 1, 2, \dots, m$.

In order to improve the correction effect of the SNV method, Bi et al. firstly segmented the spectrum and then performed local SNV for each interval, which has better result than SNV performed on the full spectrum [34]. Based on the idea of physical factors such as the particle size, Rabatel et al. proposed the weighted standard normal variate transformation, namely, variable sorting for normalization (VSN), giving different weights to different wavelength variables before performing SNV [35, 36]. In addition, other methods used to improve SNV correction effect include

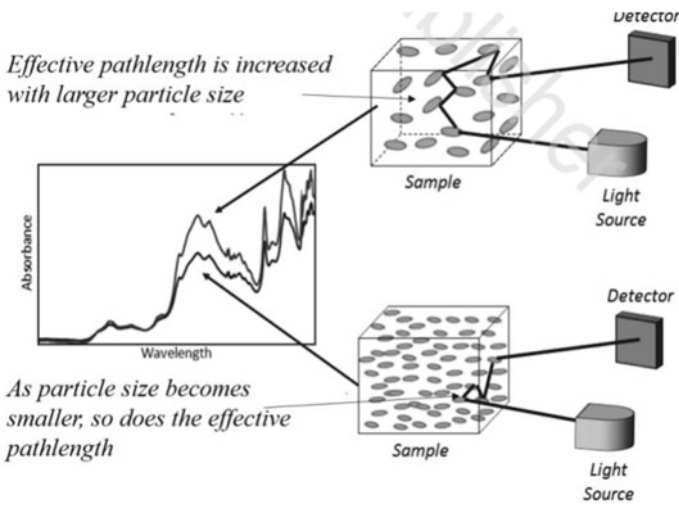


Fig. 4.15 Effect of the granularity size on the diffuse reflectance spectra

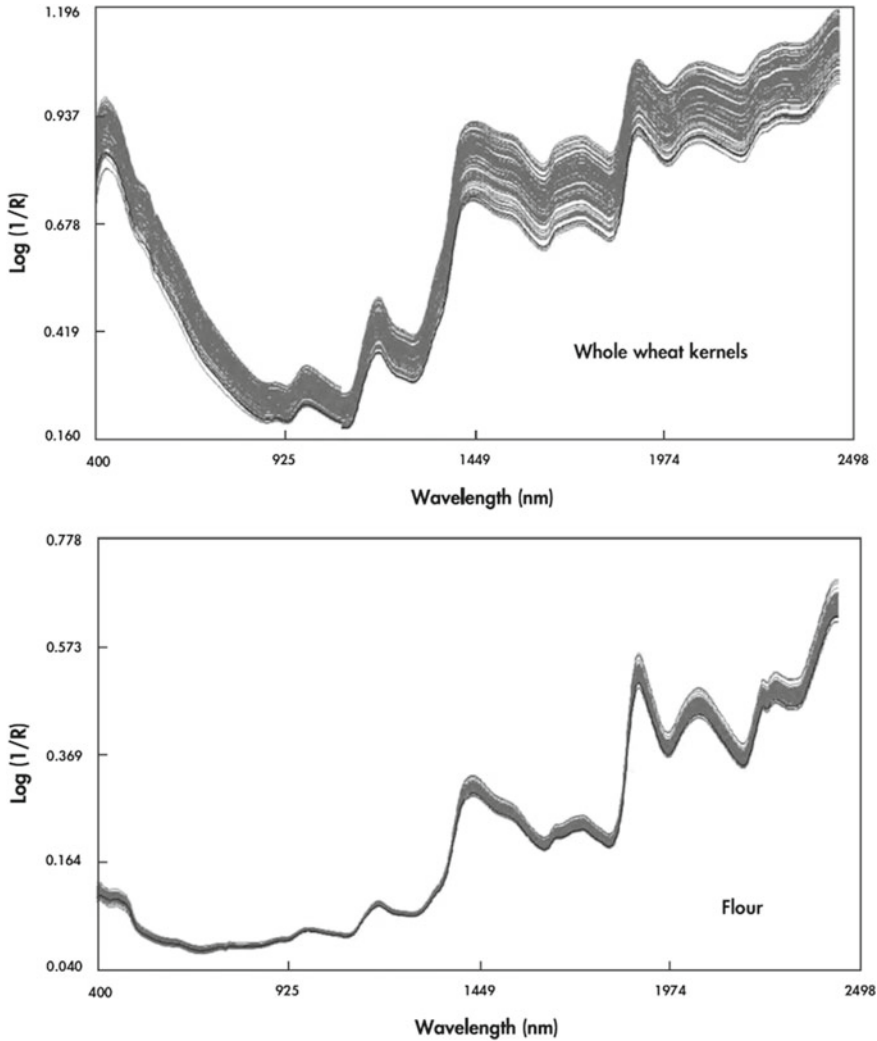


Fig. 4.16 Effects of granularity on NIR spectra for wheat grains and wheat flour

probabilistic quotient normalization (PQN), robust normal variate transformation (RNV), etc. [37].

De-trending is commonly used for spectra after SNV to eliminate baseline drift of diffuse reflectance spectra. The algorithm is very direct. Firstly, the spectrum x and wavelength λ are fitted by a polynomial to form a trend line d , then d is subtracted from x . The algorithm can be used in conjunction with SNV or alone. Reflection spectral units are usually converted into $\log 1/R$ before using SNV. Figures 4.17 and 4.18 show the NIR spectra of 80 corn samples preprocessed by SNV and SNV + de-trending, respectively.

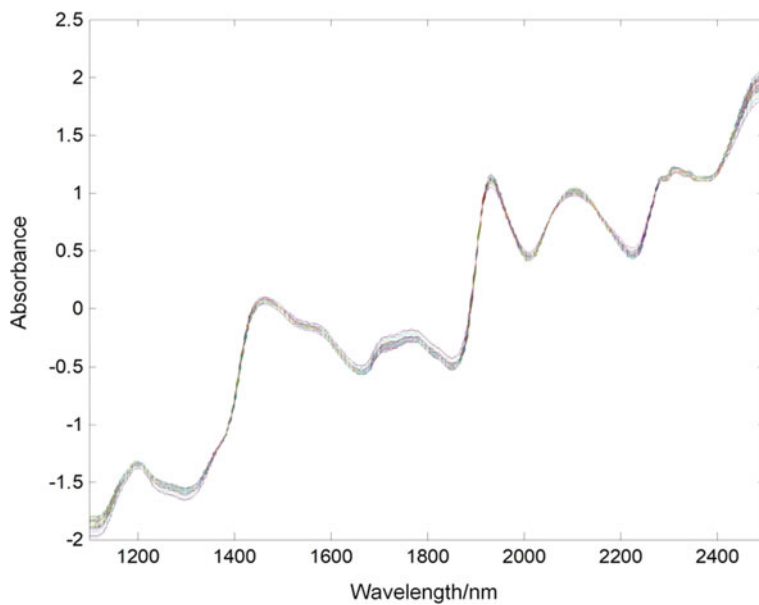


Fig. 4.17 NIR spectra of 80 corn samples preprocessed by SNV

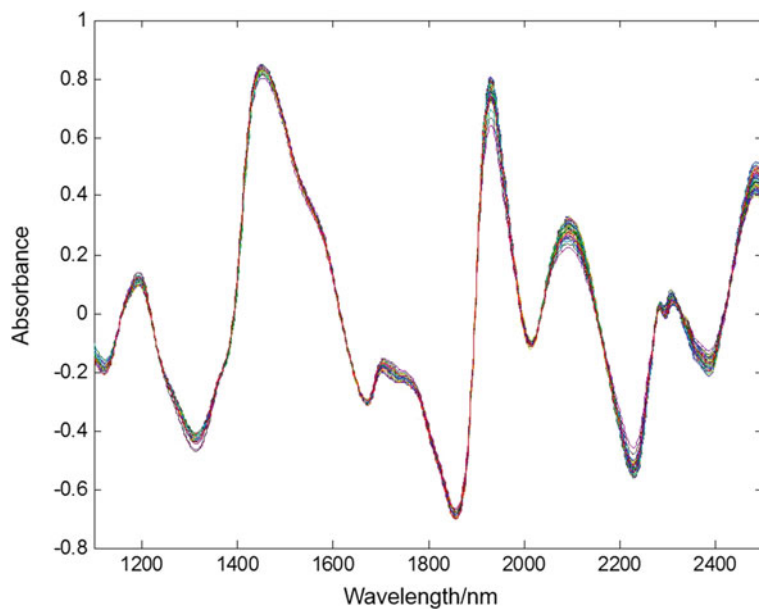


Fig. 4.18 NIR spectra of 80 corn samples preprocessed by SNV and quadratic polynomial de-trending algorithms

4.9 Multiplicative Scatter Correction

The purpose of multiplicative scatter correction (MSC) is basically the same as SNV, which is mainly to eliminate the effect of scattering from uneven particle distribution and particle size. It is widely used in solid diffuse reflection and slurry transmission (reflection) spectroscopy [38]. The MSC algorithm has the same properties as the standardization, which is based on the spectral matrix of a set of samples for calculation.

For a spectrum \mathbf{x} ($1 \times m$), the specific algorithm of MSC is as follows.

- (1) The average spectrum $\bar{\mathbf{x}}$ (i.e., the “ideal spectrum”) of the calibration set samples is calculated.
- (2) Linear regression between \mathbf{x} and $\bar{\mathbf{x}}$ is performed

$$\mathbf{x} = b_0 + \bar{\mathbf{x}}b \quad (4.24)$$

where b_0 and b are found by the least squares.

- (3) MSC transforms the spectrum as follows:

$$\mathbf{x}_{\text{MSC}} = (\mathbf{x} - b_0)/b \quad (4.25)$$

The average spectrum of sample in the calibration sets is required for MSC. That is, b_0 and b of the spectrum are first obtained, then MSC transformation is performed. The MSC algorithm assumes that the scattering is independent of the wavelength and the concentration variation of the sample. Thus, the effect may be bad when processing sample spectra with wide variation of component properties. It is proved that MSC is linearly correlated with SNV. Furthermore, the processing results of the two methods should also be similar [39]. Figure 4.19 shows the NIR spectra of 80 corns samples after MSC. It can be seen that the effect of MSC is similar to that of SNV.

Based on the MSC algorithm, many improved MSC methods have been presented [40–43], such as piecewise MSC (PMSC), loopy MSC (LMSC), extended MSC (EMSC), inverse signal correction (ISC), extended ISC (EISC), etc.

PMSC is a one-dimensional linear regression of \mathbf{x}_i against the mean spectrum $\bar{\mathbf{x}}_i$ in the wavelength range of the moving window with w . The slope b_i and the intercept a_i of each moving window segment are found in turn by the least squares method.

LMSC replaces the mean spectrum of original spectra by MSC transformation of the spectra in calibration set $\bar{\mathbf{x}}$, and then performs repeated MSC processing repeatedly.

EMSC is a polynomial regression of \mathbf{x} with and the average spectrum $\bar{\mathbf{x}}$, that is,

$$\mathbf{x} = b_0\bar{\mathbf{x}} + b_1 + \bar{\mathbf{x}}^2 b_2 \quad (4.26)$$

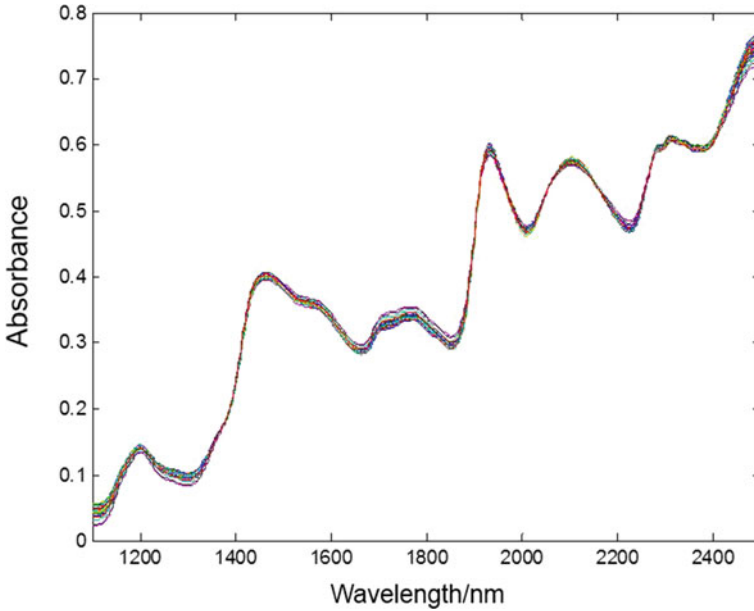


Fig. 4.19 NIR spectra of 80 corn samples preprocessed by MSC

Or a polynomial regression of x with the spectrum \bar{x} and the wavelength vector λ is performed, that is,

$$x = b_0 + \bar{x}b_1 + \lambda d_1 + \lambda^2 d_2 \tag{4.27}$$

$$x_{MSC} = (x - b_0 - \lambda d_1 - \lambda^2 d_2) / b_1 \tag{4.28}$$

ISC is the substitution of x and \bar{x} in the one-dimensional linear regression equation, that is, $\bar{x} = b_0 + \bar{x}b_1$. EISC is a polynomial regression between the \bar{x} and the mean spectrum x and the wavelength vector λ , that is,

$$\bar{x} = b_0 + \bar{x}b_1 + \lambda d_1 + \lambda^2 d_2 \tag{4.29}$$

Many improved algorithms based on EMSC are presented, such as spectral interference subtraction (SIS), eliminating internal repetitive differences by fusing orthogonal projections, etc. [44–46].

In order to eliminate the influence of out-of-bounds samples and high leverage point samples on the calibration results, Silalahi et al. proposed a robust and generalized MSC method [47].

4.10 Vector Angle Conversion

Vector angle conversion (VAC) is used to eliminate multiplicative interference in the spectrum due to scattering and refraction. A spectrum can be considered as a vector in the data space, where the vector mode (length) represents the measured intensity. The vector direction is determined by the composition of the system, which is expressed as the angle in space to the determined coordinates. The multiplicative factor b causes the intensity to change, which is the vector mode and is changed, the system composition is not changed, so the vector direction remains unchanged, that is, the vector angle does not change with the mode. Therefore, the multiplicative factor b can be eliminated by vector angle conversion [48].

As shown in Fig. 4.20, the spectrum S of the mixture consists of two components (spectrum a and spectrum b), that is, vector a and vector b together form vector S . When a and b are reduced in equal proportion, S and S' do not change in direction. However, it is only when the ratio of a and b changes that their combined vector changes direction. Furthermore, the vector angle and the system composition ratio have a functional relationship, which is independent of the vector mode.

A fixed vector a in space exists at an angle θ with a vector S that varies with composition and can be calculated by the dot product:

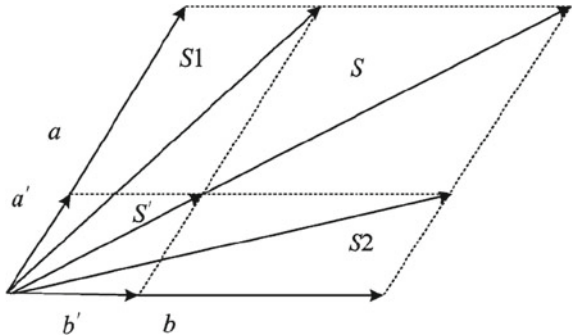
$$\cos \theta = \frac{a \cdot S}{|a| |S|} \quad (4.30)$$

If S is affected by multiplicative factor b and becomes S' , $S' = bS$, substituting into the formula for calculating the included angle can be obtained as

$$\frac{a \cdot S'}{|a| |S'|} = \frac{a \cdot bS}{|a| |bS|} = \frac{a \cdot S}{|a| |S|} = \cos \theta \quad (4.31)$$

The above formula illustrates that the vector angle is not affected by the multiplicative factor b . Moreover, it can be proved that there is a linear relationship between

Fig. 4.20 Schematic diagram of the relationship between the composition of the sample system and the vector direction



vector angle and component concentration. Therefore, a quantitative calibration model can be established by using the vector angle instead of the spectrum.

The basic steps of the VAC method [49] are shown as follows:

- (1) An appropriate reference vector \mathbf{a} is selected to calculate the angle between the direction measurement vector \mathbf{S} and it. The reference vector \mathbf{a} should be orthogonal to the background, but not orthogonal or similar to the measured component. To obtain the reference vector, the spectral matrix of calibration set is usually processed by singular value decomposition (SVD) or principal component analysis (PCA). Its loading for first principal component can approximately meet this requirement.
- (2) The measurement signal \mathbf{S} and the reference vector \mathbf{a} are divided into m intervals, respectively (or by using a moving window). The angle cosine between each interval vector pair is calculated to form the angle cosine vector $[\cos\theta_1 \cos\theta_2 \dots \cos\theta_m]$. The spectra of all samples in the calibration set are converted into the corresponding included angle cosine vectors, which constitutes the cosine vector matrix.

The quantitative model between the cosine vector matrix and the concentration vector is established by multiple linear regression or PLS, etc. For the spectrum \mathbf{x} of the sample to be measured, it is first divided into m intervals, then the angle cosine vector is calculated with the corresponding interval of the reference vector \mathbf{a} . Finally, the concentration value is predicted by the established calibration model.

In addition to quantitative analysis, the spectra multiplicatively corrected by VAC can also be used for discriminative analysis [50].

4.11 Fourier Transform

Fourier Transform (FT) plays an important role in signal processing technique, which enables the conversion between frequency domain functions and time domain functions as shown in Fig. 4.21 [51]. In a spectrometer using the Michael interference

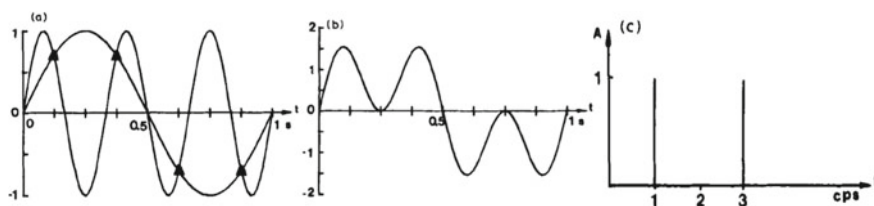


Fig. 4.21 Schematic diagram of FT converting time domain function to frequency domain function for two sine functions with periods of 1 s and 1/3 s in the time domain (a), the sum of the two sine functions in the time domain (b) and their frequency domain result obtained by FT (c)

principle, the interferogram (time domain spectrum) can be converted into a spectrum (frequency domain spectrum) by the Fourier transform.

The FT of a spectrum is the decomposition of the spectrum into a superposition sum of many sinusoidal waves of different frequencies. By this method, spectral de-noising, data compression, and information extraction can be realized.

For m discrete spectral data points x_0, x_1, \dots, x_{m-1} at equal wavelength intervals, the discrete Fourier transform (DFT) is

$$\mathbf{x}_{k.FT} = \frac{1}{m} \sum_{j=0}^{m-1} \mathbf{x}_j \exp\left(\frac{-2i\pi kj}{m}\right) \quad k = 0, 1, \dots, m-1 \quad i = \sqrt{-1} \quad (4.32)$$

The inverse Fourier transform (IFT) is as follows:

$$\mathbf{x}_j = \sum_{k=0}^{m-1} \mathbf{x}_k \exp\left(\frac{-2i\pi kj}{m}\right) \quad j = 0, 1, \dots, m-1 \quad i = \sqrt{-1} \quad (4.33)$$

The imaginary part of the original data \mathbf{x}_j is zero, and its Fourier transform frequency spectrum \mathbf{x}_k , FT is composed of real and imaginary parts \mathbf{x}_k , $FT = R_k + iL_k$, where

$$R_k = \frac{1}{m} \sum_{j=0}^{m-1} \mathbf{x}_j \cos\left(\frac{2\pi kj}{m}\right) \quad (4.34)$$

$$L_k = -\frac{1}{m} \sum_{j=1}^{m-1} \mathbf{x}_j \sin\left(\frac{2\pi kj}{m}\right) \quad (4.35)$$

The power spectrum (PS) of FT is

$$PS_k = R_k^2 + L_k^2 \quad (4.36)$$

Instrument noise is smaller in amplitude and higher in frequency compared to the useful information signal. Therefore, higher frequency signals are deleted to eliminate most spectral noise and make the signal smoother. By the low-frequency signals, the original spectral data is reconstructed by inverse Fourier transform to achieve noise removal (as shown in Fig. 4.22).

Derivative and convolution can also be performed on the raw spectral data based on FT to improve the resolution [53]. Furthermore, the Fourier coefficients or power spectra obtained by Fourier transform are directly involved in building quantitative calibration models or pattern recognition models as feature variables, which can greatly reduce the computing time without sacrificing accuracy.

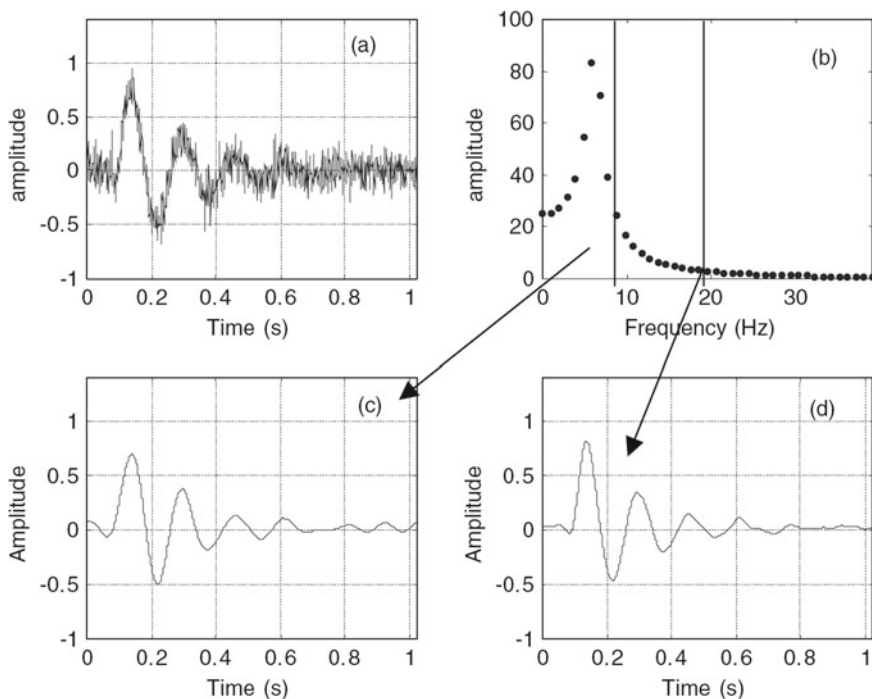


Fig. 4.22 Example of signal de-noising by FT for original noisy signal [52] (a), frequency signal by FT (b), reconstructed signal after intercepting frequency less than 10 Hz (c), reconstructed signal after intercepting frequency less than 20 Hz (d)

4.12 Wavelet Transform

The signals are decomposed into a series of accumulation of sine waves of different frequencies by FT. Since sine waves are not limited in time, FT can better delineate the frequency characteristics of the signal. However, it has no resolution in the space-time domain and cannot be used for local analysis. The basic idea of wavelet transform (WT) is similar to the FT, which is to decompose the signal into a superposition of a series of wavelet functions, all of which are obtained by translating and scaling a mother wavelet function. Wavelet analysis has positive localization properties in both the time and frequency domains. It can replace high-frequency components with gradually finer time or spatial domain substitution steps size, so that it can focus on arbitrary details of the object. Therefore, WT is known as a “mathematical microscope” for analyzing signals, and has a wide range of applications in signal processing of analytical chemistry.

The essence of wavelet transform is to project the signal $x(t)$ onto the wavelet $\Psi_{a,b}(t)$, which is the inner product of $x(t)$ and $\Psi_{a,b}(t)$, to obtain wavelet coefficients

that are easy to process. The wavelet coefficients are processed according to the need of analysis, then the processed signal is obtained by inverse transform of the processed wavelet coefficients.

Wavelets are function family that satisfy certain conditions generated by $\Psi(t)$ stretching and translation $\Psi_{a,b}(t)$:

$$\Psi_{a,b}(t) = \frac{1}{\sqrt{|a|}} \Psi \left[\frac{t-b}{a} \right], a, b \in R, a \neq 0 \quad (4.37)$$

where a is used to control the dilation, which called the scale parameter; b is used to control the position, called the translation parameter, which becomes the wavelet base or wavelet mother function; and $\Psi(t)$ must satisfy two conditions.

- (1) Small: $\Psi(t)$ rapidly converges to zero or rapidly decays to zero.
- (2) Wave:

$$\int_{-\infty}^{+\infty} \Psi(t) dt = 0 \quad (4.38)$$

In the wavelet transform of the analyzed signal, the discrete wavelet transform (DWT) is generally used.

Discrete wavelet definition: $a = a_0^m$ ($a_0 > 1, m \in Z$), $b = nb_0 a_0^m$ ($b_0 \in R, n \in Z$) Therefore,

$$\Psi_{m,n}(t) = a_0^{-\frac{m}{2}} \Psi(a_0^{-m} t - nb_0) \quad (4.39)$$

Generally $a_0=2, b_0=1$, which called dyadic wavelet. For k discrete spectral data points x_1, x_2, \dots, x_k at equal wavelength intervals, and its discrete dyadic wavelet transform is

$$WT_x(m, n) \leq x_i, 2^{-\frac{m}{2}} \Psi(2^{-m} t_i - n) \geq \sum_{i=1}^k 2^{-\frac{m}{2}} \Psi(2^{-m} t_i - n) x_i \quad (4.40)$$

The above formula illustrates that the wavelet transform is actually a projection of the discrete signal onto the wavelet basis function, with different m and n representing different resolutions (scales) and different time domains (translations). Different m and n can adjust wavelet function in different local time domains and different resolutions.

Compared with the basic functions used in FT (only trigonometric functions), the wavelet functions used in WT are not unique, that is, $\Psi(t)$ has diversity, the

same problem is analyzed with different wavelet functions sometimes the results are very different. Therefore, the selection of wavelet function is a difficulty in the practical application of WT. At present, the best wavelet function is usually selected by comparing the results of experience or continuous trial.

In the numerous wavelet basis function family, some wavelet functions have been proved effective in practice. The most commonly used in spectral analysis mainly include Haar, Daubechies (dbN), Coiflet, and Symlets.

$\Psi m, n(t)$ generally does not have an analytic expression. In order to realize finite DWT. Mallat proposed multi-resolution signal decomposition (MRSD) or Pyramid is often used in numerical calculation, which is also known as Mallat algorithm.

Discrete representation of $\Psi m, n(t)$ as a pair of low-pass filters $\mathbf{H} = \{h_p\}$ and high-pass filters $\mathbf{G} = \{g_p\}$, ($p \in \mathbb{Z}$) with $\{h_{p^*}\}$ and $\{g_{p^*}\}$ being the corresponding mirror filters. For k discrete spectral data points x_1, x_2, \dots, x_k at first-class wavelength intervals, denoted as $C(p)$, the orthogonal discrete dyadic wavelet decomposition can be written as

$$\mathbf{C}^j(i) = \sum_{p \in \mathbb{Z}} \mathbf{h}^*(p - 2i) \mathbf{C}^{j-1}(p) \tag{4.41}$$

$$\mathbf{D}^j(i) = \sum_{p \in \mathbb{Z}} \mathbf{g}^*(p - 2i) \mathbf{C}^{j-1}(p) \tag{4.42}$$

where $j = 0, 1, \dots, J, J$ is the highest decomposition order. Due to the orthogonality of decomposition, the original signal C^0 can be reconstructed by C^j and D^j :

$$\mathbf{C}^{j-1}(i) = \sum_{p \in \mathbb{Z}} \mathbf{h}(i - 2p) \mathbf{C}^j(p) + \sum_{p \in \mathbb{Z}} \mathbf{g}(i - 2p) \mathbf{C}^{j-1}(p) \tag{4.43}$$

The relationship between the scale parameter a and j is $a = 2^j$. The resolution is defined as $1/a$. As j increases, the scale binary expansion of the decomposition and the detail resolution decreases. C_j and D_j are called discrete approximation and discrete detail at 2^j resolution, respectively. That is, C_j denotes the low-frequency component with frequencies below 2^{-j} , while D_j denotes the high-frequency component with frequencies between 2^{-j} and 2^{-j+1} .

The low-pass filter $\mathbf{H} = \{h_p\}$ and the high-pass filter $\mathbf{G} = \{g_p\}$ have the following relationship:

$$\mathbf{g}_p = (-1)^p h_{p-1} \quad \text{and} \quad \sum_{p \in \mathbb{Z}} \mathbf{h}_p = \sqrt{2}, \quad \sum_{p \in \mathbb{Z}} \mathbf{g}_p = 0 \tag{4.44}$$

The wavelet basis (scale function and wavelet function) can be generated by the given filter coefficients. The approximation and detail coefficients of the wavelet can be derived directly from the filter coefficients. It is not necessary to know exactly the wavelet basis function, which greatly simplifies the calculation.

For a spectrum \mathbf{x} ($1 \times k$), the above DWT based on Mallat algorithm can be expressed in a matrix as

$$x_{WT} = \mathbf{W}\mathbf{x}^T \tag{4.45}$$

where x_{WT} is called the wavelet coefficient and \mathbf{W} is a matrix of order $k \times k$ containing the approximation and detail coefficients associated with the specified wavelet, that is, $\mathbf{W} = \begin{bmatrix} \mathbf{G} \\ \mathbf{H} \end{bmatrix}$. Its function is to perform two related convolution calculations on \mathbf{x} using a low-pass filter \mathbf{H} and a high-pass filter \mathbf{G} , respectively.

Figure 4.23 shows a schematic diagram of the cubic wavelet transform decomposition of a spectral vector using the Mallat algorithm. The dimension of the final wavelet coefficients obtained is the same as that of the original spectrum. The computational examples of wavelet decomposition and reconstruction are given in Figs. 4.24 and 4.25, respectively (Fig. 4.26).

Figure 4.26 is the diffused reflectance NIR spectrum of polypropylene powder. Figure 4.27 shows the high-frequency discrete details cd_2 (a), cd_4 (b) and low-frequency approximation ca_9 (c) obtained by using db4 mother wavelet function to decompose the NIR spectrum in Fig. 4.26 nine times. It can be clearly seen that the low-frequency approximation signal ca_9 contains mainly the strong background information of the spectrum. The detail signal cd_2 is mainly high-frequency noise.

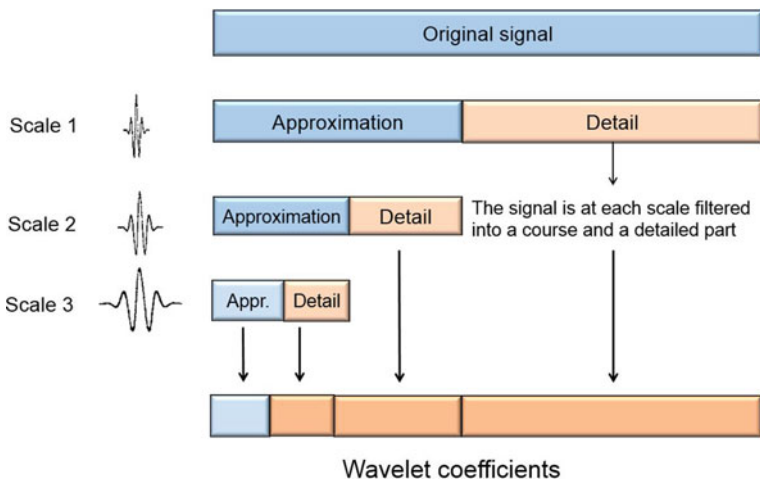


Fig. 4.23 Schematic diagram three-time decomposition of the original spectrum by MRSD algorithm

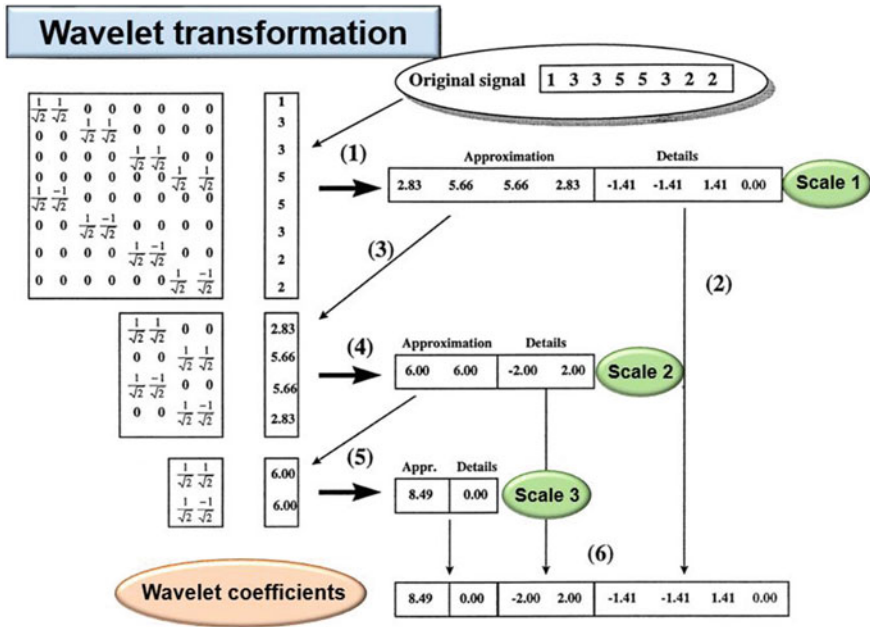


Fig. 4.24 Calculation example of wavelet decomposition of original signal using filter coefficient matrix [54]

The cd4 clearly distinguishes the effective feature information in the original spectrum. Similar to FT, WT can be used in spectral analysis for spectral de-noising, spectral data compression, extraction of feature information, etc. [55].

The general steps of WT for smoothing and noise filtering are as follows:

- (1) The original spectrum is decomposed by WT to obtain high-frequency and low-frequency wavelet coefficients.
- (2) The wavelet coefficients which are considered to represent noise are removed by threshold (called noise filtering), or the wavelet coefficients which are considered to high-frequency (low-scale) elements are removed by threshold (called smoothing).
- (3) The filtered spectral signal is obtained by inverse transformation of the processed one. The threshold usually has two forms: hard threshold, where all wavelet coefficients below the threshold are set to zero; Soft thresholding, where wavelet coefficients smaller than the threshold are set to zero and the threshold is subtracted from the absolute value of wavelet coefficients larger than the threshold. A number of reported estimation methods for thresholds are presented, such as simple soft and hard threshold, sure method, visu method, hybrid method and minmax method, etc.

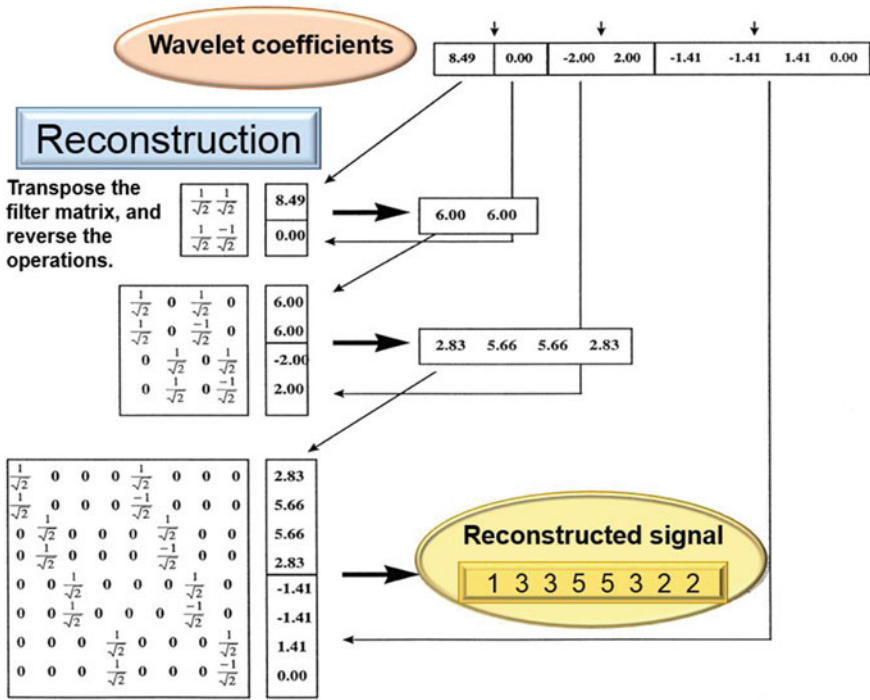


Fig. 4.25 Calculation example of reconstructing wavelet coefficients by filter coefficients transpose matrix [54]

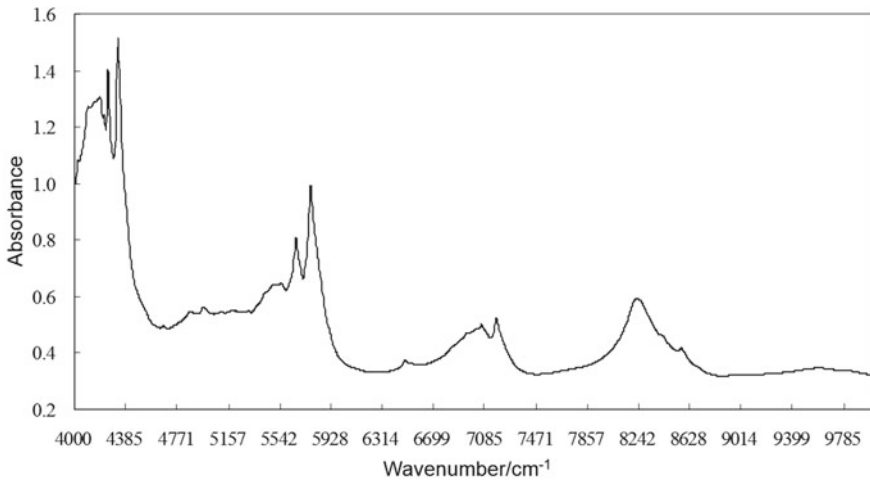


Fig. 4.26 Diffused reflectance NIR spectrum of polypropylene powder

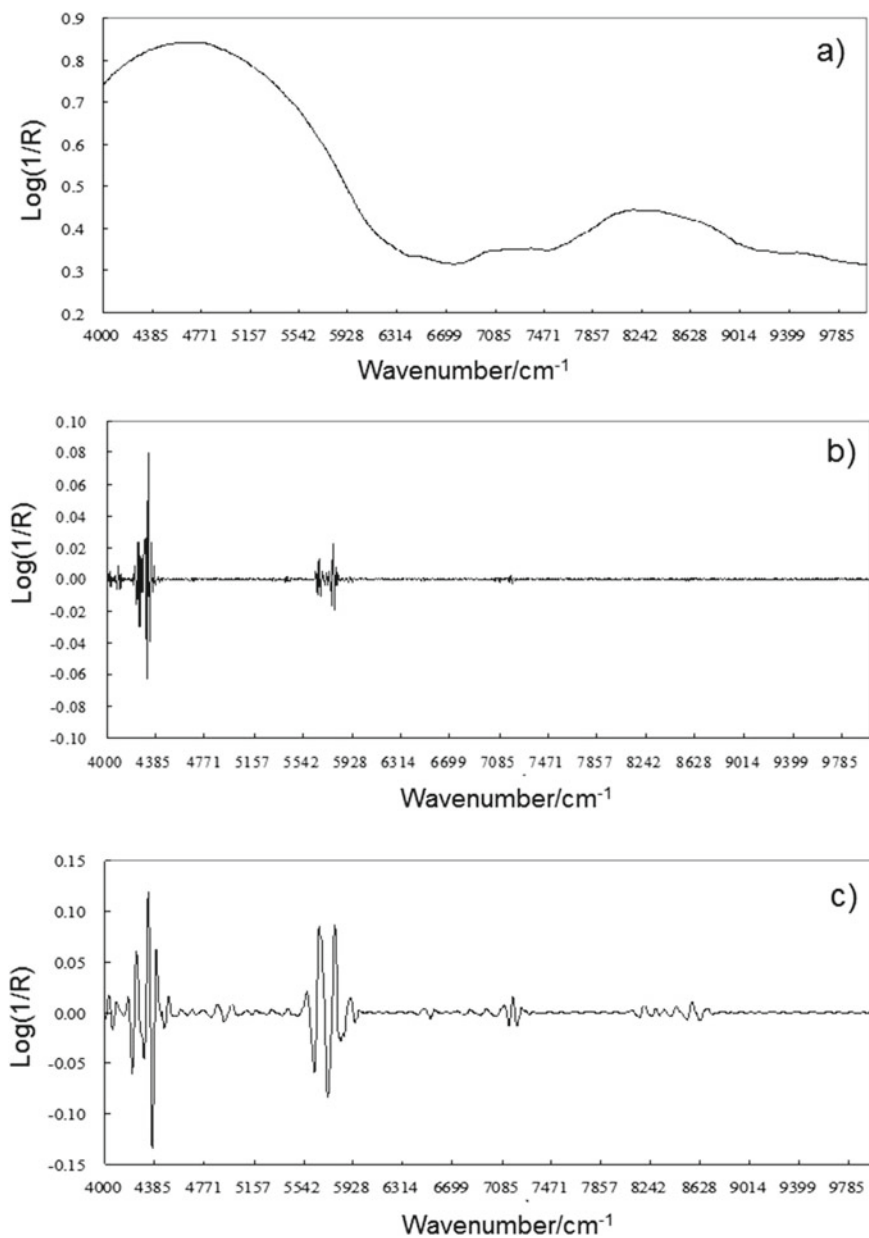


Fig. 4.27 Detail and approximation signals obtained by Db4 wavelet decomposition for NIR spectrum of polypropylene sample, approximation signal ca9 (a), high-frequency signal cd2 (b) and cd4 (c)

The basic principle of data compression by WT is similar to de-noising and generally takes the following steps:

- (1) WT is applied to the original data to obtain wavelet coefficients.
- (2) The threshold method is used to remove coefficients from the wavelet coefficients that are sufficiently small to be considered as not representing useful information and the processed coefficients are saved. When needed, the original data is obtained by inverting the transformation. The threshold is generally determined by empirical values or obtained by trial. For example, WT can compress the IR spectral database and reduce the storage space of the spectral library.

WT can also be used to extract feature information. Wavelet coefficients reflecting different information can be obtained after the decomposition of the original spectrum by WT. The wavelet coefficients related to the components to be measured can be determined by a priori knowledge or trial methods, which can be directly used to build multivariate quantitative or qualitative calibration models for the feature variables.

In addition, the WT can be used for the calculation of spectral derivatives, as shown in Sect. 4.7.3 of this book.

4.13 Image Moment Methods

Moment invariants are one of the important features of image description and are mainly used to represent geometric features of image. Because of its powerful multi-resolution capability and inherent invariance, even if the image undergoes changes such as rotation, scaling, and translation, the calculated moments remain basically unchanged. Tchebichef and Krawtchouk are two commonly used discrete orthogonal moments with good performance. Image moments have been widely used for image processing in fields such as computer vision. In recent years, image moments have been applied in quantitative or qualitative analysis of substances by spectroscopic techniques because of their advantages such as multi-resolution and inherent invariance [56–58].

Since Tchebichef moment is an orthogonal moment with discrete orthogonal polynomial as the basis function, it does not include any numerical approximation and does not require coordinate space transformation. It has no information redundancy in the representation of image information, thus has better performance. By Tchebichef image moment, the characteristic information of the target can be effectively extracted from spectra to build a model and satisfactory results can be obtained without any complicated preprocessing. Furthermore, the interference problems such as overlapping peaks, noise, and scattering can be solved by the multi-resolution capability of image moments [59, 60].

For a spectrum $f(x)$ with N number of wavelength points, the Tchebichef curve moment (T_n) at order n is defined as

$$T_n = \frac{1}{\rho(n, N)} \sum_{x=0}^{N-1} t_n(x) f(x) \quad (n = 0, 1, 2, \dots, N-1) \quad (4.46)$$

$$\rho(n, N) = \frac{N}{2n+1} (N^2 - 1)(N^2 - 2^2) \dots (N^2 - n^2) \quad (4.47)$$

where $t_n(x)$ is a discrete polynomial of Tchebichef curve moment at order n , $\rho(n, N)$ is the square of the parametric number, and $f(x)$ is the absorbance at wavelength point x .

The recurrence relationship is as follows:

$$\rho(n, N) = \frac{2n-1}{2n+1} (N^2 - n^2) \rho(n-1, N) \quad (4.48)$$

$$(n+1)t_{n+1}(x) - (2n+1)(2x - N + 1)t_n(x) + n(N^2 - n^2)t_{n-1}(x) = 0 \quad (4.49)$$

where $t_n(x)$ and $\rho(n, N)$ are normalized to make the results more stable and with less fluctuation range as follows:

$$\tilde{t}_n(x) = \frac{t_n(x)}{\beta^2(n, N)} \quad (4.50)$$

$$\tilde{\rho}(n, N) = \frac{\rho(n, N)}{\beta^2(n, N)} \quad (4.51)$$

where $\beta(n, N)$ is the scale factor and is a function of N , usually defined as $\beta(n, N) = N^2$.

The standardized Tchebichef curve moment (T_n) is defined as

$$T_n = \frac{1}{\tilde{\rho}(n, N)} \sum_{x=0}^{N-1} \tilde{t}_n(x) f(x) \quad (n = 0, 1, 2, \dots, N-1) \quad (4.52)$$

$$\tilde{\rho}(n, N) = \frac{N}{2n+1} \left(1 - \frac{1}{N^2}\right) \left(1 - \frac{2^2}{N^2}\right) \dots \left(1 - \frac{n^2}{N^2}\right) \quad (4.53)$$

The recurrence relation for the standardized Tchebichef curve moment is

$$\tilde{t}_0(x) = 1 \quad (4.54)$$

$$\tilde{t}_1(x) = (2x - N + 1)/N \quad (4.55)$$

$$n\tilde{t}_n(x) - (2n-1)\tilde{t}_1(x)\tilde{t}_{n-1}(x) + (n-1)\left(1 - \frac{(n-1)^2}{N^2}\right)\tilde{t}_{n-2}(x) = 0 \quad (4.56)$$

$$\tilde{\rho}(0, N) = N \quad (4.57)$$

$$\tilde{\rho}(n, N) = \frac{2n-1}{2n+1} \left(1 - \frac{n^2}{N^2} \right) \tilde{\rho}(n-1, N) \quad (4.58)$$

The reconstructed spectrum is calculated by the following equation:

$$\hat{f}(x) = \sum_{n=0}^{nN} T_n \tilde{t}_n(x) \quad (4.59)$$

where nN is the maximum order of the Tchebichef curve moment during the reconstruction, $nN \leq N$.

The reconstruction error ε can be selected by reconstructing the error nN defined as

$$\varepsilon = \sum_{x=0}^{N-1} \left| f(x) - \hat{f}(x) \right| \quad (4.60)$$

Liu et al. applied Tchebichef image moment to the quantitative analysis of mixtures by IR spectroscopy and solved the problem of inaccurate quantitative analysis results due to spectral overlap and shift in IR spectroscopy [56]. Pan Zhao et al. combined Krawtchouk image moment with fluorescence spectra and generalized regression neural networks to establish a quantitative model for predicting PAH content, which obtain accurate analytical results [57]. Xue et al. used Zernike moment to extract features from grayscale images of 3D fluorescence spectra and then established a quantitative model for humic acids, which obtained more reliable and accurate results compared with N-way partial least squares and alternating trilinear decomposition methods [58]. Yin et al. introduced a modeling method combining Tchebichef image moment and PLS in the quantitative analysis of terahertz spectra to predict the content of zinc oxide in rubber additive mixtures, which improved the accuracy and stability of the analysis [59]. Li et al. used Tchebichef image moment for UV-Vis spectra to establish a prediction of skin whitening agents in cosmetics. Zhu et al. used Tchebichef image moments to process NIR spectra of naphtha and established calibration models for predicting the composition of detailed families, whose prediction results were superior to conventional multivariate calibration methods [61].

4.14 External Parameter Orthogonalization

External parameter orthogonalization (EPO) is a spectral preprocessing method based on principal component analysis (PCA) [62]. Assuming that the external interference variables and concentration variables in the spectrum are independent, the purpose of EPO is to project the spectrum into a space orthogonal to the interference variables (e.g., sample temperature and water content in the sample, etc.) to achieve the role of interference filtering. The spectral matrix is decomposed as

$$\mathbf{X} = \mathbf{X}\mathbf{P} + \mathbf{X}\mathbf{Q} + \mathbf{E} \quad (4.61)$$

where \mathbf{P} and \mathbf{Q} are the projection operator matrices on the concentration and interference subspaces, respectively, and \mathbf{E} is the residual matrix.

The main steps of the EPO are described below by the moisture content in the sample as the interfering variable [63, 64].

\mathbf{X}_{dry} is spectral matrix ($n \times m$) without water in calibration set, and $\mathbf{X}_{M_1}, \mathbf{X}_{M_2}, \dots, \mathbf{X}_{M_k}$ are the spectral matrices of water-containing samples corresponding to the \mathbf{X}_{dry} calibration set samples with k different water contents. Each matrix size is also $n \times m$, where n is the number of samples and m is the number of wavelength variables.

- (1) The average spectral vectors of $\mathbf{X}_{\text{dry}}, \mathbf{X}_{M_1}, \mathbf{X}_{M_2}, \dots, \mathbf{X}_{M_k}$ matrices are calculated, respectively, $\bar{\mathbf{x}}_{\text{dry}}, \bar{\mathbf{x}}_{M_1}, \bar{\mathbf{x}}_{M_2}, \dots, \bar{\mathbf{x}}_{M_k}$.
- (2) The difference spectra of $\bar{\mathbf{x}}_{M_1}, \bar{\mathbf{x}}_{M_2}, \dots, \bar{\mathbf{x}}_{M_k}$ and $\bar{\mathbf{x}}_{\text{dry}}$ are calculated, respectively. The difference spectral matrix \mathbf{D} with dimension $k \times m$ is formed.
- (3) The singular value decomposition is performed on the covariance matrix of matrix \mathbf{D}

$$[\mathbf{U}, \mathbf{S}, \mathbf{V}] = \text{svd}(\mathbf{D}^T\mathbf{D}) \quad (4.62)$$

- (4) The dimension f of the EPO is set. The submatrix \mathbf{V}_f of the f factors in front of matrix \mathbf{V} is taken.
- (5) The \mathbf{Q} matrix is calculated as follows:

$$\mathbf{Q} = \mathbf{V}_f \mathbf{V}_f^T \quad (4.63)$$

- (6) The projection matrix \mathbf{P} is calculated with \mathbf{I} as the unit matrix

$$\mathbf{P} = \mathbf{I} - \mathbf{Q} \quad (4.64)$$

For a spectrum \mathbf{x}_M of an arbitrary water content sample, the corrected anhydrous spectrum \mathbf{x}_{EPO} is obtained as

$$\mathbf{x}_{\text{EPO}} = \mathbf{x}_M \mathbf{P} \quad (4.65)$$

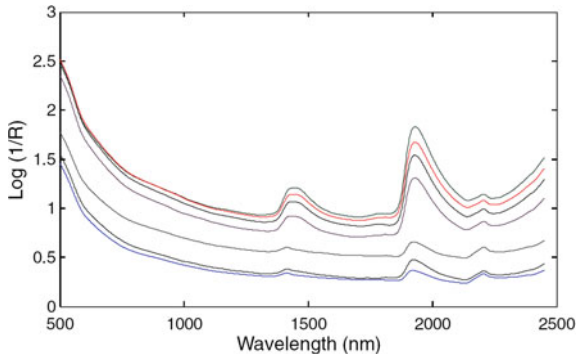


Fig. 4.28 Diffused reflectance NIR spectra of soil samples with different water contents

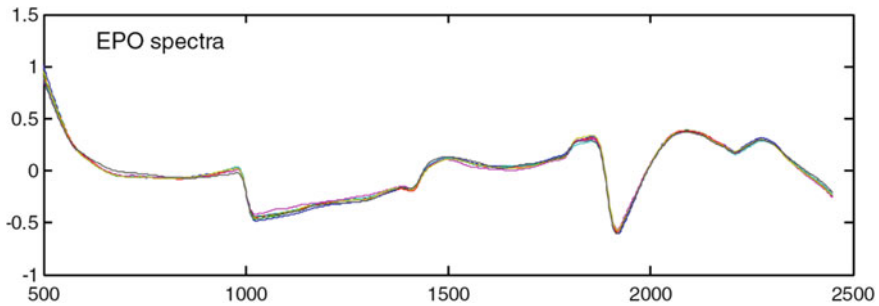


Fig. 4.29 Spectra preprocessed by EPO method

The EPO is mainly used to eliminate the influence of temperature and water content in the sample on the spectra [65, 66]. Figure 4.28 shows the diffuse reflectance NIR spectra of soils with different water contents. Figure 4.29 shows the results after the EPO. It can be seen that this method can eliminate the influence of moisture on the spectra very well.

4.15 Generalized Least Squares Weighting

The generalized least squares weighting (GLSW) is similar to the EPO, which focuses on removing the effect of external interference (e.g., temperature) on the spectrum by constructing a filter [67–69].

The main steps of the GLSW are described below with the sample temperature as the interference variable.

- (1) X_{T_1} and X_{T_2} are the spectral matrix of the calibration set samples at two temperatures, T_1 and T_2 , respectively. Firstly, mean centralization is performed

on \mathbf{X}_{T1} and \mathbf{X}_{T2} , respectively. Then the difference spectral matrix \mathbf{X}_d of the two matrices is calculated.

- (2) The covariance matrix \mathbf{C} is calculated as follows:

$$\mathbf{C} = \mathbf{X}_d^T \mathbf{X}_d \quad (4.66)$$

- (3) \mathbf{C} is decomposed to obtain the left eigenvector \mathbf{V} and the diagonal matrix \mathbf{S} of singular values:

$$\mathbf{C} = \mathbf{V} \mathbf{S}^2 \mathbf{V}^T \quad (4.67)$$

- (4) The matrix \mathbf{D} is calculated as

$$\mathbf{D} = \sqrt{\frac{\mathbf{S}^2}{\alpha} + \mathbf{I}} \quad (4.68)$$

where \mathbf{I} is the unit matrix and α is the weight parameter, which is generally between 0.0001 and 1. The smaller α is, the stronger the filtering ability is.

- (5) The filtering matrix \mathbf{P} is calculated as

$$\mathbf{P} = \mathbf{V} \mathbf{D}^{-1} \mathbf{V}^T \quad (4.69)$$

For the spectrum \mathbf{x}_T obtained at one temperature, the spectrum \mathbf{x}_{GLSW} at the reference temperature obtained after its correction is

$$\mathbf{x}_{\text{GLSW}} = \mathbf{x}_T \mathbf{P} \quad (4.70)$$

4.16 Loading Space Standardization

The loading space standardization (LSS) method focuses on eliminating the effect of sample temperature on the spectra [70, 71]. \mathbf{X}_{T1} , \mathbf{X}_{T2} , ..., \mathbf{X}_{TK} represent the spectral matrix obtained for n calibration set samples at temperatures T_1 , T_2 , ..., T_K , respectively. The main steps of the LSS are as follows:

- (1) \mathbf{X}_{T1} , \mathbf{X}_{T2} , ..., \mathbf{X}_{Tk} are combined to obtain the spectral matrix:

$$\mathbf{X}_{\text{comb}} = [\mathbf{X}_{T1} \mathbf{X}_{T2} \dots \mathbf{X}_{Tk}] \quad (4.71)$$

- (2) Singular value decomposition of \mathbf{X}_{comb} is performed:

$$[\mathbf{U}, \mathbf{S}, \mathbf{V}] = \text{svd}(\mathbf{X}_{\text{comb}})$$

- (3) The principal component f is set, and \mathbf{T} is calculated by taking the \mathbf{U}_f and \mathbf{S}_f of the first f factors:

$$\mathbf{T} = \mathbf{U}_f \mathbf{S}_f \quad (4.72)$$

- (4) The loading matrix is calculated for \mathbf{X}_{T_k} :

$$\mathbf{V}_{T_k}^T = \mathbf{T}^+ \mathbf{X}_{T_k}, k = 1, 2, \dots, K \quad (4.73)$$

- (5) Relational model between the corresponding elements of the loading matrix \mathbf{V}_{T_k} and temperature is established as follows:

$$v_{T_k,i,j} = a_{i,j} + b_{i,j} T_k + c_{i,j} T_k^2 \quad (4.74)$$

where $v_{T_k,i,j}$ is the (i, j) th element of \mathbf{V}_{T_k} .

After the model parameters $a_{i,j}$, $b_{i,j}$, and $c_{i,j}$ are obtained, the spectrum \mathbf{x}_{test} firstly obtained at temperature t , then \mathbf{x}_{test} can be normalized to the spectrum $\mathbf{x}_{\text{stand}}$ corresponding to any reference temperature t_{ref} according to the following equation:

$$v_{t,i,j} = a_{i,j} + b_{i,j} t + c_{i,j} t^2, v_{t_{\text{ref}},i,j} = a_{i,j} + b_{i,j} t_{\text{ref}} + c_{i,j} t_{\text{ref}}^2 \quad (4.75)$$

$$\mathbf{x}_{\text{stand}} = \mathbf{x}_{\text{test}} (\mathbf{V}_t^T) + (\mathbf{V}_{t_{\text{ref}}} - \mathbf{V}_t)^T + \mathbf{x}_{\text{test}} \quad (4.76)$$

where $v_{t,i,j}$ is the (i, j) th element of \mathbf{V}_t and $v_{t_{\text{ref}},i,j}$ is the (i, j) th element of $\mathbf{V}_{t_{\text{ref}}}$.

4.17 Oblique Projection

Oblique projection is a mathematical method for extracting the spectra of pure compounds from the spectra of complex mixtures [72, 73]. The oblique projection divides the spectral data space \mathbf{X} into two parts, one part is the vector subspace \mathbf{S} of the component to be measured and the other part is the adjacent subspace \mathbf{H} of the sample which is composed of other components besides the component to be measured. The oblique projection is to model the spectral signal of the pure substance, i.e., the oblique projection operator, which separates the spectral signal \mathbf{S} of the pure substance to be measured from the mixture spectra \mathbf{X} .

The separation model, i.e., the oblique projection operator, is established for the component vector \mathbf{S} and the background signal \mathbf{H} in the known modeling sample. The oblique projection operator \mathbf{E}_{SIH} is

$$\mathbf{E}_{\text{SIH}} = \mathbf{S} (\mathbf{S}^T \mathbf{P}_H \mathbf{S})^{-1} \mathbf{S}^T \mathbf{P}_H \quad (4.77)$$

$$\text{where } \mathbf{P}_H = \mathbf{I} - \mathbf{H}(\mathbf{H}^t\mathbf{H})^{-1}\mathbf{H}^t \quad (4.78)$$

The pure signal \mathbf{c} of the component to be measured in the mixed sample spectrum \mathbf{x} can be separated by the oblique projection:

$$\mathbf{c} = \mathbf{x} \mathbf{E}_{S|H} \quad (4.79)$$

The pure signal \mathbf{c} of the measured component is separated by the oblique projection. The ratio of this signal maximum c_{\max} to the pure spectral signal s_{\max} of the measured component is used as the pure signal intensity of the measured component \mathbf{I} . The intensity \mathbf{I} is proportional to the concentration of the component to be measured in the mixture which can establish a standard working curve [74, 75].

4.18 Orthogonal Signal Correction

The spectral preprocessing methods mentioned above only process the spectral data without considering the influence of the concentration matrix. Thus, some useful chemical information for establishment of the calibration model may be lost or noise removing may be incomplete, which affects the quality of the model during the preprocessing. Orthogonal signal correction (OSC) and net analysis signal (NAS) are both spectral preprocessing methods based on the participation of concentration matrix. The basic principle of these kinds of preprocessing methods is to remove the information in the spectral matrix irrelevant to the components by orthogonal projection before establishing the quantitative calibration model, in order to simplify the model and improve the prediction ability of the model.

OSC has three ways to realize: orthogonal signal correction (OSC) [76], direct orthogonal signal correction (DOSC) [77], and direct orthogonal (DO) [78, 79], where OSC has a variety of algorithms.

Generally, the first several principle components of PCR or PLS are usually not the matrix information related to concentration, but unrelated to concentration when the spectral matrix has a small relation to the concentration matrix or the background noise of the spectral matrix is too large. Therefore, before the quantitative calibration model is established, the unrelated spectral signals to the concentration matrix are filtered by orthogonal mathematical methods. It can reduce the number of principle components of the model and further improve the prediction ability. In addition, OSC can be used to solve problems such as model transfer in multivariate calibration, as well as outlier detection. The following mainly describes several common algorithms for OSC.

4.18.1 Wold Algorithm

Wold et al. [76] first proposed the idea of OSC, the specific algorithm is as follows:

- (1) The original spectral matrix $\mathbf{X}_{n \times k}$ and concentration vector $\mathbf{Y}_{n \times 1}$ in calibration set are mean centering centralized or standardization.
- (2) The first main component score vector t of \mathbf{X} is calculated that is spectral matrix.
- (3) Orthogonal treatment of t to \mathbf{Y} :

$$t_{new} = \left(I - Y(Y^T Y)^{-1} Y^T \right) t \quad (4.80)$$

- (4) The weight vector w is calculated, w is the regression coefficient obtained by PLS or PCR between \mathbf{X} and t_{new} .
- (5) The new t is calculated as

$$t = \mathbf{X}w \quad (4.81)$$

- (6) Determine whether there is $\|t - t_{old}\| / \|t\| < 10^{-6}$, if met on to the next step, otherwise return to the step (3).
- (7) The loading vector calculated as follows:

$$p^T = t^T X / t^T t_{new} \quad (4.82)$$

- (8) The orthogonal signal in \mathbf{X} is subtracted as given below:

$$X = X - t p^T \quad (4.83)$$

- (9) Return to step (2) until the cycle has completed the required number of main factors f (f is the number of main factors to be orthogonal).
- (10) For the predicted vector x_{new} , the corrected spectrum can be obtained from the weight w and loading p :

$$t = x_{new}^T - w \quad (4.84)$$

$$x_{OSC}^T = x_{new}^T - t p^T \quad (4.85)$$

4.18.2 Fearn Algorithm

Fearn proposed a simple and fast OSC algorithm based on the Wold algorithm [80].

- (1) The original calibration set spectral matrix \mathbf{X} ($n \times k$) and concentration of matrix \mathbf{Y} ($n \times 1$) are to be mean centering, centralization, or standardization.
- (2) \mathbf{M} is calculated (I-unit matrix) as follows:

$$\mathbf{M} = \mathbf{I} - \mathbf{X}^T \mathbf{Y} (\mathbf{Y}^T \mathbf{X} \mathbf{X}^T \mathbf{Y})^{-1} \mathbf{Y}^T \mathbf{X} \quad (4.86)$$

- (3) \mathbf{Z} is calculated as

$$\mathbf{Z} = \mathbf{X} \mathbf{M}$$

- (4) The singular value of \mathbf{Z} is decomposed as follows:

$$[U, S, V] = svd(\mathbf{Z}^T) \quad (4.87)$$

- (5) The previous number of f characteristic values to be orthogonal processing \mathbf{g} and the corresponding loading matrix \mathbf{C} are obtained as follows:

$$\mathbf{g} = \text{diag}(\mathbf{S}_f) \quad (4.88)$$

$$\mathbf{C} = \mathbf{V}_f \quad (4.89)$$

- (6) The weight vector \mathbf{w}_i is calculated as:

$$\mathbf{w}_i = \mathbf{M} \mathbf{X}^T \mathbf{C}_i / \mathbf{g}_i^T, i = 1, 2, \dots, f \quad (4.90)$$

- (7) The score vector is calculated as follows:

$$t_i = \mathbf{C}_i \mathbf{g}_i^T \quad (4.91)$$

- (8) The loading vector is calculated as follows:

$$p_i = \mathbf{X}^t t_i / t_i^T t_i \quad (4.92)$$

- (9) The orthogonal signal in \mathbf{X} is subtracted as given below:

$$X_{OSC} = \mathbf{X} - \sum_{i=1}^f t_i p_i^T \quad (4.93)$$

- (10) For the predicted vector \mathbf{x}_{new} , the corrected spectrum is determined by the weight \mathbf{w} and loading \mathbf{p} :

$$t = \mathbf{x}_{new}^T \mathbf{w} \quad (4.94)$$

$$x_{OSC}^T = x_{new}^T - tp^T \quad (4.95)$$

Based on the Fearn algorithm, the partial orthogonal signal correction (POSC) [81] was proposed to solve the local characteristics of orthogonal unrelated. The results of two sets of NIR spectral data show that their performance is slightly better than OSC, but at the same time it brings the problem of selecting window size. The comparison between the Wold algorithm and the Fearn algorithm shows that the Wold has a mathematical basis for \mathbf{t} and \mathbf{p} , but not for \mathbf{w} , while the Fearn has a theoretical basis for \mathbf{w} but not for \mathbf{t} and \mathbf{p} . Li combined the two methods and proposed a new OSC algorithm [82].

4.18.3 Direct Orthogonal Signal Correction Algorithm

Different from the Wold algorithm, direct orthogonal signal correction (DOSC) algorithm which is proposed by Westerhuis [77] firstly orthogonalizes the spectral matrix \mathbf{X} and concentration matrix \mathbf{Y} . Then principal components analysis (PCA) is performed on the orthogonalized \mathbf{X} to obtain \mathbf{T} and \mathbf{P} . The specific algorithm is as follows:

- (1) The original spectral matrix \mathbf{X} ($n \times k$) and concentration matrix \mathbf{Y} ($n \times 1$) in calibration set are mean centering, centralized, or standardized.
- (2) \mathbf{M} is calculated as follows:

$$M = X^T \left((X^T)^{-1} \right)^T Y \quad (4.96)$$

- (3) \mathbf{Z} is calculated as follows:

$$Z = X - MM^{-1}X \quad (4.97)$$

- (4) PCA is performed on \mathbf{ZZ}^t , and the first f principal component score matrix \mathbf{T}_f needed for orthogonal processing is selected.
- (5) The weight matrix \mathbf{W}_f is calculated and the broad inverse \mathbf{X}^{-1} is obtained by partial least squares regression (PLS).

$$\mathbf{W}_f = \mathbf{X}^{-1} \mathbf{T}_f \quad (4.98)$$

- (6) The new \mathbf{T}_f is calculated as follows:

$$T_f = XW_f \quad (4.99)$$

- (7) The loading matrix \mathbf{P}_f is calculated as follows:

$$P_f = X^T T_f / T_f^T T_f \quad (4.100)$$

(8) \mathbf{X}_{DOSC} is calculated as follows:

$$X_{DOSC} = X - T_f P_f^T \quad (4.101)$$

(9) For the predicted vector x_{new} , the corrected spectrum is determined by the weight \mathbf{W} and loading \mathbf{P} .

$$T = x_{new}^T W \quad (4.102)$$

$$x_{OSC}^T = x_{new}^T - T P^t \quad (4.103)$$

4.18.4 Direct Orthogonal Algorithm

The difference between the direct orthogonal (DO) algorithm and the Wold algorithm is that Wold algorithm uses inverse partial least squares regression to filter out the signals unrelated to the concentration matrix, while DO algorithm filters out the signals unrelated to the concentration matrix directly by orthogonal spectral matrix [78, 79]. Therefore, the DO algorithm is simpler and faster than the OSC algorithm. The two algorithms have some differences in the results of the actual preprocessing of the spectrum. The DO operation steps are as follows:

- (1) The original calibration set spectral matrix \mathbf{X} ($n \times k$) and concentration of matrix \mathbf{Y} ($n \times 1$) are to be mean centering, centralization, or standardization.
- (2) \mathbf{M} is calculated as follows:

$$M = X^T Y (Y^T Y)^{-1} \quad (4.104)$$

- (3) \mathbf{Z} is calculated as follows:

$$Z = X - Y M^T \quad (4.105)$$

- (4) PCA is carried out on Z , and the score matrix \mathbf{T}_f and loading matrix \mathbf{P}_f , which need orthogonal processing, are taken as the first f .
- (5) The new \mathbf{T}_f is calculated as

$$\mathbf{T}_f = X \mathbf{P} \quad (4.106)$$

- (6) X_{OD} is calculated as

$$X_{OD} = X - \mathbf{T}_f \mathbf{P}_f^T \quad (4.107)$$

- (7) For the predicted vector x_{new} , the corrected spectrum is determined by the loading \mathbf{P}

$$\mathbf{T} = x_{\text{new}}\mathbf{P} \quad (4.108)$$

$$X_{OD}^T = x_{\text{new}}^T - \mathbf{TP}^T \quad (4.109)$$

4.18.5 *Application of Orthogonal Signal Correction Algorithm*

When using the OSC algorithm to preprocess the spectrum, the following two problems should be paid attention to. (1) The selection of principal components (PCs) for spectral orthogonalization. 1–5 PCs are generally selected. However, the final determination of the number of PCs depends on the prediction results of unknown samples. Therefore, it can be selected by plotting the standard error of prediction (SEP) of the validation set with the number of PCs. (2) The influence of concentration matrix accuracy on spectral orthogonal results. The accuracy of the concentration reference matrix is of great importance to spectral orthogonal processing. The measurement results of the reference method are not accurate, some information related to the concentration matrix is filtered out when the spectrum is orthogonal processed with the data, part of the information related to the concentration matrix is filtered out, while the irrelevant signals are retained, thus making the prediction ability of the calibration model worse. Therefore, in the use of orthogonal spectral preprocessing method, it is important to ensure the accuracy of the concentration reference data.

OSC algorithm is proposed to solve the problem of calibration transfer for NIR analysis [83]. Subsequently, the OSC algorithm was used as a comparison method for almost all calibration transfer problems [84, 85]. Geladi et al. [86] compared several calibration transfer methods (FIR, WT, PDS, and S-G smoothing). The prediction of lake water pH analysis model is established by the NIR spectrum of sediments at the bottom of the lake to transfer between different instruments, to prove that the results of OSC preprocessing method are perfect. The OSC method used by Blanco et al. [87] effectively eliminated the differences between the two types of NIR spectral data (online and laboratory solid drugs) and obtained better calibration and prediction results than first derivative, SNC and MSC preprocessing methods.

The individual PLS method can eliminate nonlinear and other irrelevant variables to a certain extent in the calibration process. Therefore, in most cases, the OSC algorithm does not significantly improve the predictive ability of the model, nor does it substantially simplify the number of principal factors used by the model (the sum of the number of OSC orthogonal main factors and the number of OSC-PLS main factors are basically the same as the number of main factors used in the use of PLS alone) [88]. Bertran et al. [89] tried to use the OSC algorithm to improve the NIR spectral prediction ability for low-concentration components. Although the

result was not ideal, the OSC algorithm can explain the spectral features well and intuitively. Trygg et al. [90–92] also incorporated OSC into the PLS regression steps and proposed a number of new multivariate calibration methods.

Zhang et al. [93] research results showed that OSC for PLS modeling cannot effectively improve the model prediction ability due to the existence of overfitting. Therefore, they combined the modeling of OSC and MLR to obtain better results than the PLS model.

4.19 Net Analyte Signal

Net analyte signal (NAS) [94–96] is also a preprocessing algorithm involving concentration matrix, which was first proposed by Lorber [97]. Its basic idea is basically the same as OSC, which is to remove information irrelevant to the components in the spectral matrix by orthogonal projection. The specific algorithm of NAS is as follows:

- (1) The original calibration set spectral matrix \mathbf{X} ($n \times k$) and concentration matrix \mathbf{Y} ($n \times 1$) are to be mean centering, centralization, or standardization.
- (2) \mathbf{Z} that is the part of \mathbf{X} that is orthogonal with \mathbf{Y} is calculated as follows:

$$\mathbf{Z} = (\mathbf{I} - \mathbf{Y}\mathbf{Y}^T/(\mathbf{Y}^T\mathbf{Y}))\mathbf{X} \quad (4.110)$$

where \mathbf{I} is a $n \times n$ unit matrix.

- (3) PCA is carried out on \mathbf{Z} , and the first f loading matrices $\mathbf{P} = \mathbf{P}_f$ that need orthogonal processing are taken.
- (4) The orthogonal projection matrix \mathbf{R} is calculated (\mathbf{I} -unit matrix, $k \times k$) as follows:

$$\mathbf{R} = \mathbf{I} - \mathbf{P}_f\mathbf{P}_f^T \quad (4.111)$$

- (5) The processed \mathbf{X}_{nas} by NAS is calculated as follows:

$$\mathbf{X}_{\text{nas}} = \mathbf{X}\mathbf{R} \quad (4.112)$$

- (6) The predicted vector x_{nas} is calculated as follows:

$$x_{\text{nas}} = x_{\text{new}}\mathbf{R} \quad (4.113)$$

The spectral matrix of the calibration set after NAS processing spectral is generally established by CLS, PLS, or PCR. In addition, NAS is used to calculate figure of merit of multivariate calibration models, such as sensitivity, selectivity, detection limits, and confidence intervals, as well as to detect outlier and select wavelengths [98–101].

Boschetti et al. [102] used NAS/CLS to establish a calibration model for measuring the content of two additives in rubber by the NIR spectroscopy, with results comparable to PLS and PCR used separately. Hector et al. [103] compared the differences between NAS and OSC algorithms, and established models with OSC/CLS, OSC/PLS, NAS/CLS, and NAS/PLS, respectively; however, the prediction ability of NAS and OSC models was not significantly improved. Berger et al. [104–106] have proposed a new multivariate calibration method based on NAS-Mixed Linear Analysis (HLA). Xu et al. [107] also proposed a calibration method based on NAS that does not require the selection of the best number of factors. Faber et al. [108] used NAS to evaluate the effects of spectral preprocessing methods MSC, first-order derivatives, and second-order derivatives on the predictive ability of NIR calibration models.

In the spectral analysis field, in addition to OSC and NAS algorithms, interference elimination algorithms (IIR) [109] and orthogonal algorithms which proposed by Ferre et al. are all the preprocessing methods involving concentration matrix [110]. The IIR method is mainly used to solve the problem of measuring low-concentration substances by NIR or IR spectroscopy.

4.20 Optical Path-Length Estimation and Correction

Optical path-length estimation and correction (OPLEC) is an algorithm that combines spectral correction with regression [111–113]. The method first uses the calibration sample set spectral matrix and the corresponding concentration vector to estimate the light scattering multiplicative effect parameters caused by the difference physical properties of each sample in the calibration sample set. Then “double calibration strategy” was used to eliminate the spectral scattering multiplicative effect of the unknown sample to be measured. This method can effectively separate the multiplicative effect caused by the difference of physical properties of the sample from the spectral contribution caused by the change of chemical component content.

The influence of solid particle size or solid content in turbid liquids on the sample spectrum can be expressed by the following model:

$$\mathbf{x} = b \sum_{i=1}^g c_i \mathbf{s}_i + d + e \quad (4.114)$$

where \mathbf{x} is the spectrum; c_i and \mathbf{s}_i are the concentration of component and the pure spectrum of the i th component, respectively; d is the deviation of the model; e is the measurement error of the spectrum; and b is the change of light transmission of light in the sample due to change in the physical properties of the sample, resulting in a multiplicative effect. b varies from sample due to different physical properties of the sample. One of the main ideas of the OPLEC is to estimate the multiplicative effect b of each corrected sample based on the spectral data of the corrected sample set.

The OPLEC estimates that the multiplicative effect b is also based on the PCA. When the calibration set spectral matrix X ($n \times m$) and concentration vector y ($n \times 1$), n is the number of calibration set samples and m is the number of spectral wavelength variables, the main steps of the OPLEC method are as follows:

- (1) SVD decomposition of spectral matrix X :

$$[U, S, V] = svd(X) \quad (4.115)$$

- (2) The principal component number g was set (g is abstract active chemical group fraction in the sample), the previous number of g factor U_g is taken.
 (3) After deduction, it can be deduced that the multiplicative effect vector b of the calibration set samples can be obtained by solving the following constraint minimization problem:

$$\min_b f(b) = \frac{1}{2} b^T ((I - U_g U_g^T) + diag(y/w)(I - U_g U_g^T)diag(y/w))b \quad (4.116)$$

the constraint is $-b \leq -1$.

In Eq. 4.116, $diag(y/w)$ is y/w 's diagonal matrix and w is the weight parameter, which can be set to the maximum value in concentration vector y .

The multiplicative effect vector b in the upper formula can be solved by the quadratic programming.

- (4) The following two calibration models are established:

$$diag(b)y = \alpha_1 1 + X\beta_1 \quad (4.117)$$

$$b = \alpha_2 1 + X\beta_2 \quad (4.118)$$

where 1 is the vector of element 1, $diag(b)$ is the diagonal matrix, and its diagonal element is the corresponding element of vector b .

The parameters of the two calibration models, α_1 , β_1 , α_2 , and β_2 , are available by PLS.

For the spectral x_{un} of the sample to be tested, the multiplicative effect can be eliminated by the ratio of the predicted values of the two corrected models, thus predicting the concentration value of the sample to be tested y_{un} .

$$y_{un} = \frac{\alpha_1 1 + x\beta_1}{\alpha_2 1 + x\beta_2} \quad (4.119)$$

Surface-enhanced Raman scattering (SERS) has the characteristics of high sensitivity, strong spectral characteristics, and fast detection. However, the SERS signal intensity of complex system samples depends not only on the concentration of the

substance to be tested in the sample, but also is related to the physical properties of the SERS substrate, such as the shape, particle size, and aggregation of nanoparticles. The reproducibility and stability of common SERS substrates are poor, which made the accuracy of SERS quantitative analysis results not meet ideal requirements. Based on OPLEC and combined with the internal standard method, Hu and Jin et al. proposed a series of multiplicative effect models for SERS (MEMSERS) for surface-enhanced Raman spectroscopic quantitative analysis, which could effectively eliminate the effect of changes in the inhomogeneity physical properties of SERS substrates on the accuracy of quantitative analysis results [114–116].

4.21 Two-Dimensional Correlation Spectroscopy

Strictly speaking, two-dimensional correlation spectroscopy (2DCOS) is not a preprocessing method, but a combination of spectral experimental and data processing method.

In 1986, Noda [117] firstly proposed an experimental scheme to obtain 2DCOS, which dynamically altered the absorption spectrum of the sample by acting on the sample system in a certain form of perturbation (initially a low-frequency disturbance of sine waveforms). Then a mathematical correlation analysis is performed on time-varying spectrum to generate a two-dimensional correlation infrared spectrum. Subsequently, in 1993, Noda proposed the concept of a generalized 2DCOS. The external perturbation is extended from the fixed form of sinusoidal waveforms to any form that can cause changes in spectral signal, such as temperature, concentration, pressure, sample composition, reaction time, magnetic field, etc. Furthermore, 2DCOS is extended from IR spectroscopy to NIR, Raman, fluorescence, electron spin resonance spectrum, and other technical fields [118].

The change in the regional molecular environment induced by perturbation can be represented by the corresponding change of time in various spectra. The instantaneous fluctuation of this spectrum is often referred to as the dynamic spectrum of the system. In the IR spectrum, typical changes in the observed dynamic spectrum include changes in absorption intensity, displacement of absorption peaks changes in directional absorption (dichromatic effect), and so on. Different types of interference can cause different responses in the system, which is making spectral changes different. 2DCOS can be obtained by some simple mathematical processing of these dynamic spectra, mainly cross-correlation analysis.

Assuming that the external disturbance acts on the sample system to be studied, a series of dynamic spectra $x(\nu, t)$ is obtained between the maximum (\mathbf{T}_{\max}) and minimum (\mathbf{T}_{\min}) of the external disturbance variable, where ν is the spectral coordinate (e.g., wavenumber, wavelength, displacement, etc.), and the external disturbance variable t can be a variable such as temperature, pressure, or concentration.

First, the dynamic spectrum is transformed from time domain to frequency domain spectrum by FT.

$$\mathbf{X}_1(\omega) = \int_{-\infty}^{+\infty} \mathbf{x}(v_1, t) e^{-i\omega t} dt = \mathbf{X}_1^{\text{Re}}(\omega) + i\mathbf{X}_1^{\text{Im}}(\omega) \quad (4.120)$$

where $\mathbf{X}_1^{\text{Re}}(\omega)$ and $\mathbf{X}_1^{\text{Im}}(\omega)$ are the real and imaginary parts of $\mathbf{x}(v_1, t)$ after the transformation, respectively, and ω represents independent frequency components varying with time. Similarly, the conjugate function of the dynamic spectral Fourier transformation is

$$\mathbf{X}_2(\omega) = \int_{-\infty}^{+\infty} \mathbf{x}(v_2, t) e^{+i\omega t} dt = \mathbf{X}_2^{\text{Re}}(\omega) - i\mathbf{X}_2^{\text{Im}}(\omega) \quad (4.121)$$

Mathematical cross-correlation analysis of a pair of dynamic spectral signals transformed by Fourier measured at different spectral variables v_1 and v_2 results in two-dimensional correlation intensity.

$$\mathbf{X}(v_1, v_2) = \frac{1}{\pi(T_{\text{max}} - T_{\text{min}})} \int_0^{+\infty} X_1(\omega) X_2(\omega) d\omega = \varphi(v_1, v_2) + i\psi(v_1, v_2) \quad (4.122)$$

where $\varphi(v_1, v_2)$ and $\psi(v_1, v_2)$ are the real and imaginary parts, respectively. They correspond to the synchronization and asynchronous correlation spectral intensity of dynamic spectral changes.

In the actual calculation, the Hilbert transformation matrix method is used [118]. For the dynamic spectral matrix \mathbf{X} ($n \times m$) obtained by n experimental conditions, m is the number of wavelength point of the spectrum, and its synchronization correlation spectrum can be calculated as follows:

$$\varphi(i, j) = \frac{1}{n-1} \sum_{k=1}^n x_{k,i} x_{k,j} \quad (4.123)$$

where $x_{k,i}$ is the absorbance at the i th wavelength in the spectrum obtained by the k th experimental condition, $i, j = 1, \dots, m$. It can also be expressed as a matrix:

$$\varphi = \frac{1}{n-1} \mathbf{X}^T \mathbf{X} \quad (4.124)$$

The asynchronous correlation spectra are calculated as follows:

$$\psi = \frac{1}{n-1} \mathbf{X}^T \mathbf{H} \mathbf{X} \quad (4.125)$$

where H is the Hilbert transformation matrix ($n \times n$), the elements

$$h_{i,j} = \frac{1}{\pi(j-i)}, (i \neq j), \quad h_{i,j} = 0, (i = j), \quad i, j = 1, \dots, n \quad (4.126)$$

2DCOS can be visualized by three-dimensional or two-dimensional contour map, which is convenient for analyzing two-dimensional information intuitively. In a contour plot of a 2DCOS, the z-axis value is represented by contour in the x-y-plane. The 2DCOS is a flexible and effective spectral analysis technique, which emphasizes the subtle characteristics of spectral changes caused by external disturbances, improves the spectral resolution, and also can analyze the interaction between molecules.

Two-dimensional infrared (2D IR) correlation spectroscopy is a widely used analytical method, which has been successfully applied in polymer, protein, liquid crystal materials, biology, and other research fields. China has edited a book entitled "Atlas of Two-dimensional Correlation Infrared Spectroscopy for Traditional Chinese Medicine Identification", which can be used to distinguish different levels of complex traditional Chinese medicine by 2D IR correlation spectroscopy. 2DCOS has also made many research achievements in NIR spectroscopic analysis [119, 120]. For example, Wu et al. [121] studied the hydrogen bond interaction between amino groups in polyamide and polyurethane, revealing the existence of different hydrogen bond states in the sample. Ozaki et al. [122] demonstrated that two-dimensional correlation NIR spectroscopy has its unique advantages in studying the two-dimensional structure of proteins. Liu et al. [123] studied the molecular structure and stability of ascorbic acid through two-dimensional correlation NIR spectroscopy, and identified the authenticity of traditional Chinese medicines. Barton et al. [124] used 2DCOS to analyze the differences between NIR spectrometers. Sasic et al. [125] extended the traditional wavelength-wavelength 2DCOS to the sample-sample 2DCOS and analyzed the NIR spectrum of temperature-sensitive oleic acid and milk with different protein content, which obtained effective quantitative and qualitative information.

References

1. Rinnan A, Van Den BF, Engelsen SB. Review of the most common pre-processing techniques for near-infrared spectra. *Trends Anal Chem.* 2009;28:1201–22.
2. Engel J, Gerretzen J, Szymanska E, et al. Breaking with trends in pre-processing? *Trends Anal Chem.* 2013;50:96–106.
3. Rinnan A. Pre-processing in vibrational spectroscopy-When Why and How. *Anal Methods.* 2014;6:7124–9.
4. Lee LC, Liong CY, Jemain AA. A contemporary review on data preprocessing (DP) practice strategy in ATR-FTIR spectrum. *Chemom Intell Lab Syst.* 2017;163:64–75.
5. Wan XH, Li G, Zhang MQ, et al. A review on the strategies for reducing the non-linearity caused by scattering on spectrochemical quantitative analysis of complex solutions. *Appl Spectrosc Rev.* 2020;55(5):351–77.

6. Seasholtz MB, Kowalski BR. The effect of mean centering on prediction in multivariate calibration. *J Chemom.* 1992;6(2):103–11.
7. Savitzky A, Golay MJE. Smoothing and differentiation of data by simplified least squares procedures. *Anal Chem.* 1964;36:1627–39.
8. Dotto AC, Dalmolin RSD, Grunwald S, et al. Two preprocessing techniques to reduce model covariables in soil property predictions by vis-NIR spectroscopy. *Soil Tillage Res.* 2017;172:59–68.
9. Vasat R, Kodesova, Klement R, et al. Simple but Efficient signal pre-processing in soil organic carbon spectroscopic estimation. *Geofis Int.* 2017; 298:46–53.
10. Han ZY, Zhu XC, Liu Q, et al. Hyperspectral inversion models for soil organic matter content in the Yellow River Delta. *J Plant Nutr Fertil.* 2014;20(6):1545–52.
11. Li FL, Chang QR. Estimation of winter wheat leaf nitrogen content based on continuum removed spectra. *J Agric Mach.* 2017;48(7):174–9.
12. Clark RN, Roush TL. Reflectance spectroscopy: quantitative analysis techniques for remote sensing applications. *J Geophys Res.* 1984;89(B7):6329–40.
13. Xu YJ, Hu GD, Zhang ZF. Continuum removal and its application to the spectrum classification of field object. *Geogr Geo-Inf Sci.* 2005;21(6):11–4.
14. Cao WX, Cheng T, Zhu Y, et al. *Crop growth spectrum monitoring.* Beijing: Science Press; 2020.
15. Zhang ZM, Chen S, Liang YZ. Baseline correction using adaptive iteratively reweighted penalized least squares. *Analyst.* 2010;135(5):1138–46.
16. Chen ZG, Shen TT, Yao JD, et al. Signal enhancement of cadmium in lettuce using laser-induced breakdown spectroscopy combined with pyrolysis process. *Molecules.* 2019;24(13):2517.
17. Li YQ, Pan TH, Li HR, et al. Non-invasive quality analysis of thawed tuna using near infrared spectroscopy with baseline correction. *J Food Process Eng.* 2020;43(8):13445.
18. Lieber CA, Mahadevan-Jansen A. Automated method for subtraction of fluorescence from biological Raman spectra. *Appl Spectrosc.* 2003;57(11):1363–7.
19. Wang T, Dai LK. Background subtraction of Raman spectra based on iterative polynomial smoothing. *Appl Spectrosc.* 2017;71(6):1169–79.
20. Cao A, Pandya AK, Serhatkulu GK, et al. A robust method for automated background subtraction of tissue fluorescence. *J Raman Spectrosc.* 2007;38(9):1199–205.
21. Baek SJ, Park A, Ahn YJ, et al. Baseline correction using asymmetrically reweighted penalized least squares smoothing. *Analyst.* 2015;140(1):250–7.
22. He SX, Zhang W, Liu LJ, et al. Baseline correction for Raman spectra using an improved asymmetric least squares method. *Anal Methods.* 2014;6(12):4402–7.
23. Zhao H, Chen WX, Xu XD, et al. Baseline correction for Raman spectra based on locally symmetric reweighted penalized least squares. *Chin J Lasers.* 2018;45(12):274–85.
24. Yao J, Su H, Yao ZX. Blind source separation of coexisting background in Raman spectra. *Spectrochim Acta Part A Mol Biomol Spectrosc.* 2020; 238:118417.
25. Hopkins DW. What is a Norris derivative? *NIR News.* 2001;12(3):3–5.
26. Shao XG, Pang CY. Calculation of approximate derivative using continuous wavelet transform. *Comput Appl Chem.* 2000;17(3):57–60.
27. Elzanfaly ES, Hassan SA, Salem MY, et al. Continuous wavelet transform, a powerful alternative to derivative spectrophotometry in analysis of binary and ternary mixtures: a comparative study. *Spectrochim Acta Part A Mol Biomol Spectrosc.* 2015;151:945–55.
28. Shao XG, Cui XY, Wang M, et al. High order derivative to investigate the complexity of the near infrared spectra of aqueous solutions. *Spectrochim Acta Part A Mol Biomol Spectrosc.* 2019;213:83–9.
29. Alexander KML, Chau FT, Gao JB. Wavelet transform: a method for derivative calculation in analytical chemistry. *Anal Chem.* 1998;70(24):5222–9.
30. Li ZG, Wang QY, Lv JT, et al. Improved quantitative analysis of spectra using a new method of obtaining derivative spectra based on a singular perturbation technique. *Appl Spectrosc.* 2015;10(6):39–41.

31. Xu JG, Feng XL, Guan L, et al. Fractional differential application in reprocessing infrared spectral data. *Chem Autom Instrum.* 2012;39(3):347–51.
32. Zheng KY. Study on model optimization and model transfer algorithm for near infrared spectroscopy. Shanghai: East China University of Science and Technology; 2013.
33. Barnes RJ, Dhanoa MS. Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra. *Appl Spectrosc.* 1989;43(5):772–7.
34. Bi YM, Yuan KL, Xiao WQ, et al. A local pre-processing method for near-infrared spectra, combined with spectral segmentation and standard normal variate transformation. *Anal Chim Acta.* 2016;909:30–40.
35. Rabatel G, Marini F, Walczak B, et al. VSN: variable sorting for normalization. *J Chemom.* 2020; 34(2):e3164.
36. Sun XD, Subedi P, Walker R, et al. NIRS prediction of dry matter content of single olive fruit with consideration of variable sorting for normalisation pre-treatment. *Postharvest Biol Technol.* 2020; 163:111140.
37. Mishra P, Roger JM, Rutledge DN, et al. MBA-GUI: a chemometric graphical user interface for multi-block data visualisation, regression, classification, variable selection and automated pre-processing. *Chemom Intell Lab Syst.* 2020; 205:104139.
38. Isaksson T, Naes T. The effect of multiplicative scatter correction and linearity improvement in NIR spectroscopy. *Appl Spectrosc.* 1988;42(7):1273–84.
39. Dhanoa MS, Lister SJ, Sanderson R, et al. The link between multiplicative scatter correction (MSC) and standard normal variate (SNV) transformations of NIR spectra. *J Near Infrared Spectrosc.* 1994;2:43–7.
40. Afseth NK, Kohler A. Extended multiplicative signal correction in vibrational spectroscopy, a tutorial. *Chemom Intell Lab Syst.* 2012;117:92–9.
41. Windig W, Shaver J, Bro R. Loopy MSC: a simple way to improve multiplicative scatter correction. *Appl Spectrosc.* 2008;62(10):1153–9.
42. Pedersen DK, Martens H, Nielsen JP, et al. Near-infrared absorption and scattering separated by extended inverted signal correction (EISC): analysis of near-infrared transmittance spectra of single wheat seeds. *Appl Spectrosc.* 2002;56(9):1206–14.
43. Helland IS, Naes T, Isaksson T. Related versions of the multiplicative scatter correction method for preprocessing spectroscopic data. *Chemom Intell Lab Syst.* 1995;29:233–41.
44. Martens H, Stark E. Extended multiplicative signal correction and spectral interference subtraction: new preprocessing methods for near infrared spectroscopy. *J Pharm Biomed Anal.* 1991;9(8):625–35.
45. Liland KH, Kohler A, Afseth NK. Model-based pre-processing in Raman spectroscopy of biological samples. *J Raman Spectrosc.* 2016;47:643–50.
46. Kohler A, Böcker U, Warringer J, et al. Reducing inter-replicate variation in fourier transform infrared spectroscopy by extended multiplicative signal correction. *Appl Spectrosc.* 2009;63(3):296–305.
47. Silalahi DD, Midi H, Arasan J, et al. Robust generalized multiplicative scatter correction algorithm on pretreatment of near infrared spectral data. *Vib Spectrosc.* 2018;97:55–65.
48. Yao ZX, Sun ZQ, Yuan HF, et al. Correction multiplicative effects in Raman spectra through vector angle transformation. *Spectrosc Spectr Anal.* 2016;36(2):419–23.
49. Xie JC, Yuan HF, Song CF, et al. Online determination of chemical and physical properties of ploy (Ethylene Vinyl Acetate) pellets using a novel method of near-infrared spectroscopy combined with angle transform. *Anal Methods.* 2019;11:2435–42.
50. Zhu ZQ, Yuan HF, Song CF, et al. High-speed sex identification and sorting of living silkworm pupae using near-infrared spectroscopy combined with chemometrics. *Sens Actuators, B Chem.* 2018;268:299–309.
51. Deming SN, Michotte Y, Massart DL, et al. *Chemometrics: a textbook.* Elsevier Science; 1988.
52. Chau FT, Liang YZ, Gao JB, et al. *Chemometrics: from basics to wavelet transform chemometrics-from basics to wavelet transform.* Wiley-Interscience; 2004

53. Hassan SA, Abdel-Gawad SA. Application of wavelet and fourier transforms as powerful alternatives for derivative spectrophotometry in analysis of binary mixtures: a comparative study. *Spectrochim Acta Part A Mol Biomol Spectrosc.* 2018;191:365–71.
54. Trygg J, Wold S. PLS regression on wavelet compressed NIR spectra. *Chemom Intell Lab Syst.* 1998;42(1–2):209–20.
55. Tian GY, Chu XL, Yuan HF. Near infrared spectra analysis of diesel by wavelet transform combined with partial least square regression method. *Comput Appl Chem.* 2006;23(10):971–4.
56. Liu JJ, Li BQ, Wang X, et al. Applying Tchebichef image moments to the simultaneous quantitative analysis of the four components in corn based on raw NIR spectra. *Chemom Intell Lab Syst.* 2018;173:14–20.
57. Pan Z, Cui YY, Wu XJ, et al. Krawtchouk moment method for the quantitative analysis of polycyclic aromatic hydrocarbons based on fluorescence three-dimensional spectra. *Spectrosc Spectr Anal.* 2018;38(12):139–43.
58. Xue W, Bao Q, Li H, et al. An efficient approach to the quantitative analysis of humic acid in water. *Food Chem.* 2016;190:1033–9.
59. Yin XH, Guo C, Feng ML, et al. Quantitative study on terahertz spectra of zinc oxide based on Tchebichef image moments. *Laser Technol.* 2019;43(6):747–52.
60. Li SS, Yin B, Zhai HL, et al. An effective approach to the quantitative analysis of skin-whitening agents in cosmetics with different substrates based on conventional UV-Vis determination. *Anal Methods.* 2019;11(11):1500–7.
61. Zhu L, Lu SH, Zhang YH, et al. An effective and rapid approach to predict molecular composition of naphtha based on raw NIR spectra. *Vib Spectrosc.* 2020; 109:103071.
62. Roger JM, Chauchard F, Bellon-Maurel V. EPO-PLS external parameter orthogonalisation of PLS application to temperature-independent measurement of sugar content of intact fruits. *Chemom Intell Lab Syst.* 2003;66(2):191–204.
63. Minasny B, Mcbratney AB, Bellon-Maurel V. Removing the effect of soil moisture from NIR diffuse reflectance spectra for the prediction of soil organic carbon. *Geoderma.* 2011;167–168:118–24.
64. Sheng WN, Sun CY, Han TS, et al. External parameter orthogonalization based temperature calibration on near infrared diffuse spectra for glucose measurement. *Nanotechnol Precis Eng.* 2017;15(5):425–9.
65. Ge Q, Han TS, Liu R, et al. Temperature correction of NIR reflectance spectrum of noninvasive blood glucose measurement based on EPO. *Spectrosc Spectr Anal.* 2020;40(5):1483–8.
66. Yu L, Hong YS, Zhu YX, et al. Removing the effect of soil moisture content on hyperspectral reflectance for the estimation of soil organic matter content. *Spectrosc Spectr Anal.* 2017;37(7):2146–51.
67. Martens H, Høy M, Wise BM, et al. Pre-whitening of data by covariance-weighted pre-processing. *J Chemom.* 2003;17(3):153–65.
68. Fu QB, Suo H, He XP, et al. Transfer calibration for alcohol determination using temperature induced shortwave near infrared spectra. *Spectrosc Spectr Anal.* 2012;32(8):2080–4.
69. Sun CY, Han TS, Guo C, et al. The correction methods for near infrared spectrum of glucose aqueous solution to reduce the influence from temperature. *Spectrosc Spectr Anal.* 2017;37(11):3391–8.
70. Chen ZP, Morris J, Martin E. Correction of temperature-induced spectral variations by loading space standardization. *Anal Chem.* 2005;77:1376–84.
71. Wang SX, Li LM, Zhong LJ, et al. Recent developments on chemometric methods for the analysis of complex spectral measurements. *J Anal Sci.* 2011;27(6):104–9.
72. Yang XN, Yao ZX, Sun SH, et al. Rapid determination of Benzalkonium chloride in eye drops with ultraviolet spectrum based on oblique projection and space angle criterion. *J Instrum Anal.* 2016;35(3):337–41.
73. Hu AQ, Yuan HF, Yao ZX, et al. A new multivariate quantitative method of spectral analysis for multicomponent system. *Spectrosc Spectr Anal.* 2014;34(11):3040–4.

74. Zhu ZQ, Yuan HF, Hu AQ, et al. Study on the fast quantitative analysis of the content of DME adulterated in LPG. *Spectrosc Spectr Anal.* 2016;36(4):978–80.
75. Rong HT, Song CF, Yuan HF, et al. Rapid quantitative analysis of content of the additive in gasoline for motor vehicles by near-infrared spectroscopy. *Spectrosc Spectr Anal.* 2015;35(10):2757–60.
76. Wold S, Antti H, Lindgren F. Orthogonal signal correction of near-infrared spectra. *Chemom Intell Lab Syst.* 1998;44:175–85.
77. Westerhuis JA, Jong SD, Smilde AK. Direct orthogonal signal correction. *Chemom Intell Lab Syst.* 2001;56:13–25.
78. Andersson CA. Direct orthogonalization. *Chemom Intell Lab Syst.* 1999;47:51–63.
79. Pierna JAF, Massart DL, Ricoux P, et al. Direct orthogonalization: some case studies. *Chemom Intell Lab Syst.* 2001;55:101–8.
80. Fearn T. On orthogonal signal correction. *Chemom Intell Lab Syst.* 2000;50:47–52.
81. Feudale RN, Tan HW, Brown SD. Piecewise orthogonal signal correction. *Chemom Intell Lab Syst.* 2002;63:129–38.
82. Li BB, Morris AJ, Martin EB. Orthogonal signal correction: algorithmic aspects and properties. *J Chemom.* 2002;16(11):556–61.
83. Sjöblom J, Svensson O, Josefson M. An evaluation of orthogonal signal correction applied to calibration transfer of near infrared spectra. *Chemom Intell Lab Syst.* 1998;44:229–44.
84. Fearn T. Review: standardisation and calibration transfer for near infrared instruments: a review. *J Near Infrared Spectrosc.* 2001;9(1):229–44.
85. Tan A, Myles J, Brown SD, et al. Transfer of multivariate calibration models: a review. *Chemom Intell Lab Syst.* 2002;64:181–92.
86. Geladi P, Bärning H, Dåbakk E. Calibration transfers for predicting lake-water Ph from near infrared spectra of lake sediments. *J Near Infrared Spectrosc.* 1999;7(2):251–64.
87. Blanco M, Coello J, Montoliu I. Orthogonal signal correction in near infrared calibration. *Anal Chim Acta.* 2001;434(1):125–32.
88. Svensson O, Kourti T, Macgregor JF. An investigation of orthogonal signal correction algorithms and their characteristics. *J Chemom.* 2002;16:176–88.
89. Bertran E, Iturriaga H, Maspoch S, et al. Effect of orthogonal signal correction on the determination of compounds with very similar near infrared spectra. *Anal Chim Acta.* 2001;431:303–11.
90. Trygg J, Wold S. Orthogonal projections to latent structures (O-PLS). *J Chemom.* 2002;16(3):119–28.
91. Trygg J. O2-PLS for qualitative and quantitative analysis in multivariate calibration. *J Chemom.* 2002;16(6):283–93.
92. Trygg J, Wold S. O2-PLS, a two-block (X-Y) latent variable regression (LVR) method with an integral OSC filter. *J Chemom.* 2003;17(1):53–64.
93. Zhang X, Yuan HF, Guo Z, et al. Study on building MIR model using orthogonal signal correction. *Spectrosc Spectr Anal.* 2011;31(12):3228–31.
94. Lorber A. Net analyte signal calculation in multivariate calibration. *Anal Chem.* 1997;69(8):1620–6.
95. Faber NM. Efficient computation of net analyte signal vector in inverse multivariate calibration models. *Anal Chem.* 1998;70(23):5108–10.
96. Joan F, Brown SD, Rius FX. Improved calculation of the net analyte signal in inverse multivariate calibration. *J Chemom.* 2001;15:537–53.
97. Lorber A. Error propagation and figures of merit for quantification by solving matrix equations. *Anal Chem.* 1986;58(6):1167–72.
98. Faber NM. Characterizing the uncertainty in near-infrared spectroscopic prediction of mixed-oxygenate concentrations in gasoline: sample-specific prediction intervals. *Anal Chem.* 1998;70(14):2972–82.
99. Ferre J, Rius FX. Detection and correction of biased results of individual analytes in multicomponent spectroscopic analysis. *Anal Chem.* 1998;70(9):1999–2007.

100. Goicoechea HC, Olivieri AC. Wavelength selection by net analyte signals calculated with multivariate factor-based Hybrid Linear Analysis (HLA). A theoretical and experimental comparison with Partial Least-Squares (PLS). *Analyst*. 1999; 124(5):1999–2007.
101. Boque R, Rius FX. Multivariate detection limits estimators. *Chemom Intell Lab Syst*. 1996;32(1):11–23.
102. Boschettia CE, Olivieri AC. Net analyte preprocessing: a new and versatile multivariate calibration technique. Analysis of mixtures of rubber antioxidants by NIR spectroscopy. *J Near Infrared Spectrosc*. 2001; 9:245–54.
103. Goicoechea HC, Olivieri AC. A comparison of orthogonal signal correction and net analyte preprocessing methods. Theoretical and experimental study. *Chemom Intell Lab Syst*. 2002; 63:129–38.
104. Berger AJ, Koo TW, Itzkan I, et al. An enhanced algorithm for linear multivariate calibration. *Anal Chem*. 1998;70(3):623–7.
105. Qi YP, Wu YT, Li TH, et al. Theory of hybrid linear analysis and its application in the analysis of multicomponent system. *Chin J Anal Chem*. 2002;30(4):401–5.
106. Goicoechea HC, Olivieri AC. Enhanced synchronous spectrofluorometric determination of tetracycline in blood serum by chemometric analysis. Comparison of partial least-squares and hybrid linear analysis calibrations. *Anal Chem*. 1999; 71(19):4361–8.
107. Xu L, Schechter I. A calibration method free of optimum factor number selection for automated multivariate analysis. Experimental and theoretical study. *Anal Chem*. 1997;69(18):3722–30.
108. Faber NM. Multivariate sensitivity for the interpretation of the effect of spectral pretreatment methods on near-infrared calibration model predictions. *Anal Chem*. 1999;71(3):557–65.
109. Hansen PW. Pre-processing method minimizing the need for reference analyses. *J Chemom*. 2001;15:123–31.
110. Ferre J, Brown SD. Reduction of model complexity by orthogonalization with respect to non-relevant spectral changes. *Appl Spectrosc*. 2001;55(6):708–14.
111. Chen ZP, Morris J, Martin E. Extracting chemical information from spectral data with multiplicative light scattering effects by optical path-length estimation and correction. *Anal Chem*. 2006;78:7674–81.
112. Jin JW, Chen ZP, Li LM, et al. Quantitative spectroscopic analysis of heterogeneous mixtures: the correction of multiplicative effects caused by variations in physical properties of samples. *Anal Chem*. 2011;84(1):320–6.
113. Chen ZP, Lovett D, Morris J. Process analytical technologies and real time process control a review of some spectroscopic issues and challenges. *J Process Control*. 2011;21:1467–82.
114. Hu M, Chen ZP, Chen Y, et al. Quantification of methimazole in plasma and tablet samples by surface enhanced Raman spectroscopy in combination with multiplicative effects model. *Chin J Anal Chem*. 2015;43(5):759–64.
115. Jin JW. Novel chemometric models and methods for quantitative spectroscopic analysis of complex systems. Hunan University;2014.
116. Xia TH, Chen ZP, Chen Y, et al. Improving the quantitative accuracy of surface enhanced Raman spectroscopy by the combination of microfluidics with a multiplicative effects model. *Anal Methods*. 2014;6:2363–70.
117. Shen Y, Peng Y, Wu PY, et al. Two_Dimensional(2D)correlation spectroscopy. *Prog Chem*. 2005;17(3):499–513.
118. Noda I. Determination of two-dimensional correlation spectra using the hilbert transform. *Appl Spectrosc*. 2000;54(7):994–9.
119. Ozaki Y. Two-dimensional near infrared correlation spectroscopy: principle and its applications. *J Near Infrared Spectrosc*. 1998;6(1):19–31.
120. Lu J, Xiang BR, Liu H. Two-dimensional Near-infrared correlation spectroscopy: theory and application. *Prog Pharm Sci*. 2007;31(7):303–8.
121. Wu P, Yang Y, Siesler HW. Two-dimensional near-infrared correlation temperature studies of an amorphous polyamide. *Polymer*. 2001;42(26):10181–6.

122. Ozaki Y, Murayama K, Wang Y. Application of two-dimensional near-infrared correlation spectroscopy to protein research. *Vib Spectrosc.* 1999;20(2):127–32.
123. Liu H, Xiang BR, Qu LB. Structure analysis of ascorbic acid using near-infrared spectroscopy and generalized two-dimensional correlation spectroscopy. *J Mol Struct.* 2006;794(1–3):12–7.
124. Barton FE II, de Haseth JA, Himmelsbach DS. The use of two-dimensional correlation spectroscopy to characterise the differences in research grade instruments. *J Near Infrared Spectrosc.* 2006;14(6):357–62.
125. Sasic S, Ozaki Y. Wavelength-wavelength and sample-sample two-dimensional correlation analyses of short-wave near-infrared spectra of raw milk. *Appl Spectrosc.* 2001;55(2):163–72.

Chapter 5

Wavelength Selection Methods



In the method of multivariate calibration in combination with spectra, the traditional view is that the multivariate calibration method (such as partial least squares, PLS) has strong anti-interference ability and can participate in the establishment of multivariate calibration model using the full spectrum. With the further research and application of PLS and other methods, it is possible to obtain better quantitative calibration models by selecting characteristic wavelengths or wavelength intervals with specific methods. Wavelength (variable) selection can simplify the model, improve the operation efficiency of the model, and strengthen the interpretability of the model. More importantly, when uncorrelated or nonlinear variables are eliminated, the calibration model with high prediction ability and good robustness can be obtained [1–3]. Therefore, the selection of wavelengths has become one of the key steps in the process of establishing calibration model, and has also become a research hotspot in the field of chemometrics and spectral analysis [4, 5].

In 2012, Mehmood et al. reviewed the variable selection algorithm based on PLS, and classified the algorithm as shown in Fig. 5.1 [6]. According to the classification of variable selection methods in machine learning, they are divided into three categories: filter methods, wrapper methods, and embedded methods. In the filter method, the variables are evaluated independently without considering the dependence or synergy between variables. The commonly used methods include correlation coefficient method and analysis of variance (ANOVA) method. The wrapper method takes the correlation between variables into account and selects the combination with the best performance by evaluating the influence of the combination effect of variables on the model performance. The commonly used methods include interval PLS (iPLS) method and genetic algorithm (GA). The embedded method selects variables while establishing the model. The most commonly used strategy is to restrict the complexity of the model by adding regular terms, such as least absolute shrinkage and selection operator (Lasso) method, and the selection of variables in random forest (RF) also belongs to the embedded method. In addition, according to the continuous features of NIR spectroscopy, wavelength interval selection (WIS) and wavelength point selection (WPS) are often used to classify variable selection methods for NIR

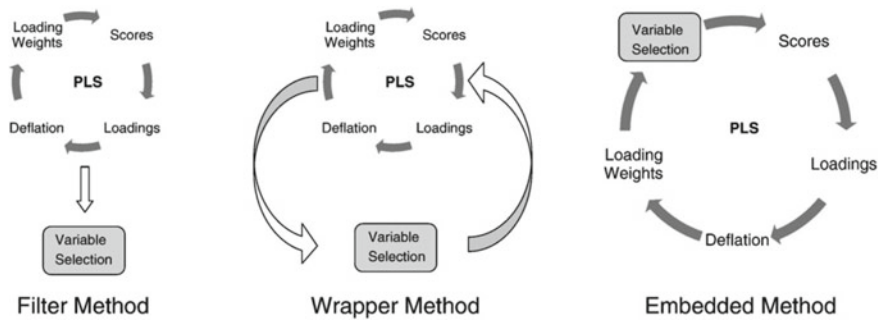


Fig. 5.1 Schematic diagram of PLS-based variable selection process, including filter methods, wrapper methods, and embedding methods [6]

spectral data [4]. WIS methods regard a number of continuous wavelengths or a wavelength band as a unit. Whereas WPS methods consider each wavelength point as a unit (i.e., a variable) when conducting variable selection, resulting in that the selected variables are discrete. Moreover, some classifications of variable selection methods are also based on selection process and final output [4].

In 2019, Yun et al. thoroughly reviewed the variable selection algorithms in analysis of NIR spectral data. Four factors, including initialization of variables, modeling method, evaluation metric, and selection strategy, were used to generalize variable selection methods (Fig. 5.2). Initialization of variables considers the number of the input of variables in the first step. Some methods take all variables into account for initialization, while sampling methods such as Monte Carlo (MC) sampling [7], Bootstrap sampling [8], and binary matrix sampling (BMS) [9] are often used to generate subsets of variables. Modeling method factor is to select a modeling method to build the relationship between the selected variables and the property of interest. The prediction performance of built model with selected variables is assessed based on an evaluation metric. Finally, a selection strategy is used to determine the optimal variable subset including filter-based, extreme value, sequential, exhaustive, intelligent optimization algorithm-based (IOA-based), and model population analysis-based (MPA-based) searches.

5.1 Correlation Coefficient and Analysis of Variance Method

Correlation coefficient method is to calculate the correlation between the corresponding absorbance vector \mathbf{x} of each wavelength in the spectral matrix of calibration set and the concentration vector \mathbf{y} of the component to be measured in the concentration matrix. Then a wavelength-correlation coefficient R diagram or the

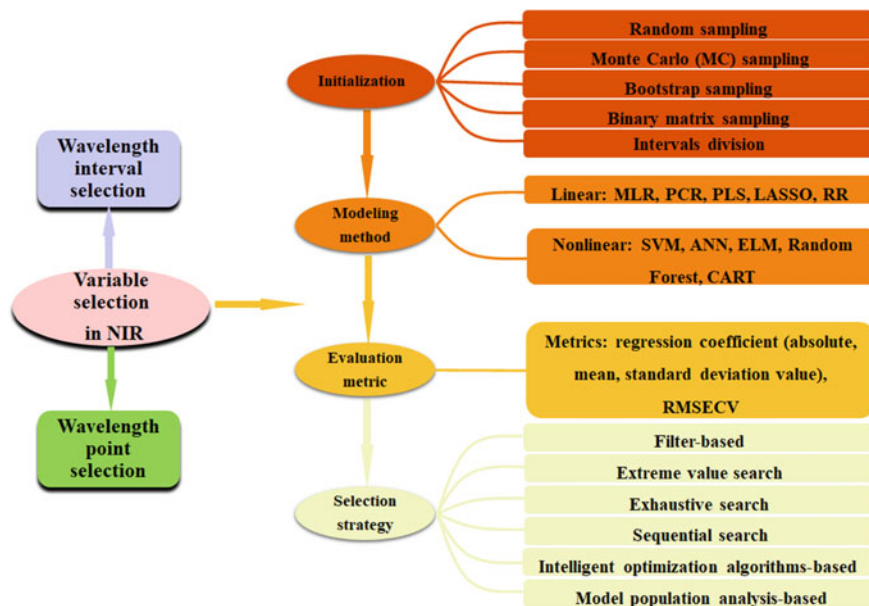


Fig. 5.2 Classification of variable selection methods based on four factors including initialization of variables, modeling method, evaluation metric, and selection strategy

determination coefficient R^2 diagram is drawn, from which the greater the corresponding absolute value of correlation coefficient (or determination coefficient), the more information the wavelength has. Therefore, the wavelengths whose correlation coefficients are greater than the threshold value can be selected to participate in building the model based on the given threshold value of known chemical knowledge. The correlation coefficient R is calculated by the following formula:

$$R = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}} \quad (5.1)$$

$$\bar{x} = \left(\sum_{i=1}^n x_i \right) / n \quad (5.2)$$

$$\bar{y} = \left(\sum_{i=1}^n y_i \right) / n \quad (5.3)$$

where n is the number of samples in the calibration set.

Since the correlation coefficient method is based on linear statistical method, the results selected by this method are often unreliable when the sample distribution of nonlinear correlation and calibration set is not uniform.

If the mathematical operation data between the wavelength (such as the difference between wavelengths or the ratio between wavelengths) has a linear relationship with the concentration value, then the wavelength can be selected through the two-dimensional graph of the correlation coefficient between the difference or the ratio of wavelengths and the concentration value (Fig. 5.3) [10, 11].

The method of ANOVA is to obtain the wavelengths-standard deviation graph through the ANOVA of the spectral matrix in the calibration set at various wavelengths. The larger the standard deviation is, the more significant the spectral change is. Similar to the correlation coefficient method, a threshold value is given to select the wavelength band. Since ANOVA is not used to optimize the selection of wavelength for the component to be measured, it is seldom used in quantitative models and more used in qualitative models.

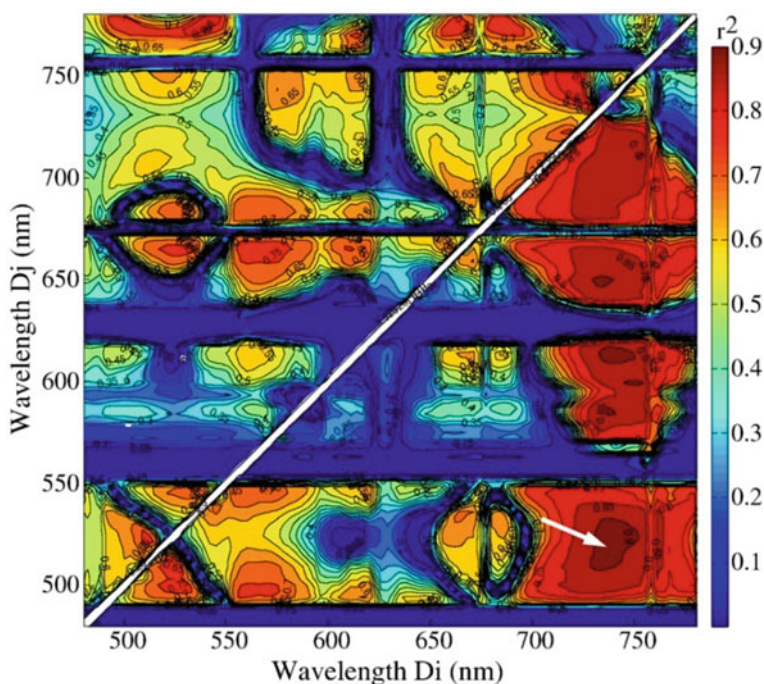


Fig. 5.3 Two-dimensional diagram of correlation coefficient between the difference or the ratio of wavelengths and the concentration value ($D_{740\text{nm}}$ and $D_{522\text{nm}}$ pointed by the arrow are selected variables) [10]

5.2 Simple-To-Use Interactive Self-modeling Mixture Analysis Method

Simple-to-use interactive self-modeling mixture analysis (SIMPLISMA) is used to distinguish the pure component spectrum from the mixture spectral array method. One of the key steps is to identify the variable (pure wavelength) of the pure component, and the principle of selection is to maximize the variance and be uncorrelated [12, 13].

Let the spectral matrix \mathbf{X} ($n \times m$), where n is the number of samples and m is the number of wavelength points. The steps of SIMPLISMA algorithm are as follows:

- (1) Select the first variable and calculate the purity value $p_{i,1}$:

$$p_{i,1} = \frac{\sigma_i}{(\mu_i + \alpha)} \quad (5.4)$$

where σ_i is the standard deviation of the i th wavelength, μ_i is the mean value of the i th wavelength, α is the compensation term used as the correction factor of the low absorption intensity (noise level) variable, and $i = 1, 2, \dots, m$.

- (2) Select the j th variable ($j \geq 2$). First, the correlation matrix \mathbf{C} is calculated as follows:

$$\mathbf{C} = \mathbf{X}_u \mathbf{X}_u^T / n \quad (5.5)$$

where \mathbf{X}_u is the spectral matrix after row area normalization. Correlation weight coefficient $\omega_{i,j}$ is calculated as follows:

$$\omega_{i,j} = \begin{vmatrix} c_{i,i} & c_{i,p_1} & \cdots & c_{i,p_{j-1}} \\ c_{p_1,i} & c_{p_1,p_1} & \cdots & c_{p_1,p_{j-1}} \\ \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots \\ c_{p_{j-1},i} & \cdots & \cdots & c_{p_{j-1},p_{j-1}} \end{vmatrix} \quad (5.6)$$

where j represents the j th variable to be selected, p_{j-1} represents the $(j-1)$ th variable has been selected, and p_1 represents the first variable has been selected. When the j th variable is highly correlated with the selected $(j-1)$ th variable, the value of $\omega_{i,j}$ is close to zero; otherwise, the value of $\omega_{i,j}$ is large. The purity value of the wavelength is generally calculated as follows:

$$p_{i,j} = \frac{\sigma_i}{(\mu_i + \alpha)} \omega_{i,j} \quad (5.7)$$

wherein $\omega_{i,1}$ of the selected first variable is calculated as follows:

$$\omega_{i,1} = \frac{\mu_i^2 + \sigma_i^2}{\mu_i^2 + (\sigma_i + \alpha)^2} \quad (5.8)$$

The wavelength with the largest purity value is successively selected as the j th variable until the number of variables required to complete the selection is completed.

The spectral matrix \mathbf{S} ($r \times m$) of the pure compound in the spectral matrix \mathbf{X} ($n \times m$) of the mixture can be identified by using the SIMPLISMA algorithm, where r is the pure group fraction in the mixture, and the matrix decomposition formula is as follows:

$$\mathbf{X} = \mathbf{C}\mathbf{S} \quad (5.9)$$

where \mathbf{C} is the concentration matrix of the pure compound ($n \times r$), the absorbance of the pure wavelength determined by the analysis is used as the value of the \mathbf{C} matrix, and the number of wavelengths selected is the pure group fraction r in the mixture (the value of r can be determined by **SVD** decomposition), then calculate as follows:

$$\mathbf{S} = (\mathbf{C}^T\mathbf{C})^{-1}\mathbf{C}^T\mathbf{X}^T \quad (5.10)$$

The actual concentration of the pure compound is calculated after the normalization of the \mathbf{S} matrix. Due to the strong characteristic peak information of Raman spectra, SIMPLISMA algorithm has a good application effect in Raman spectroscopy, especially surface-enhanced Raman scattering (SERS), for detection of pesticide residues in fruits and vegetables [14, 15]. Qin et al. used spatial migration Raman spectroscopy combined with SIMPLISMA algorithm to obtain the visual distribution map of lycopene in tomatoes with different maturity [16]. Khodabakhshian et al. used SIMPLISMA algorithm to extract the spectra of pure tannic acid components from pomegranate fruit samples, and then conducted spectral information divergence (SID) to distinguish four different maturity stages of pomegranate fruits, including immature stage (S1), semi-ripe stage (S2), semi-ripe stage (S3), and fully mature stage (S4) [17].

5.3 Successive Projections Algorithm

The successive projections algorithm (SPA) is a forward loop selection method that starts at one wavelength and calculates its projections upon unchosen wavelengths in each loop [18, 19]. The largest wavelength of the projected vector is introduced into the wavelength combination. Each new selected wavelength has the least linear relationship with the previous one. For the spectral matrix \mathbf{X} ($n \times m$) of the calibration set, given the number of wavelengths to be selected h , the approach of SPA algorithm is as follows:

- (1) Before the start of the first iteration ($p = 1$), select a column vector \mathbf{X}_j in the spectral matrix, denote as $\mathbf{X}_{k(0)}$, that is, $k(0) = j, j = 1, \dots, m$.
- (2) Denote the set of column vector positions that have not been selected as $\mathbf{s}, \mathbf{s} = \{j, 1 \leq j \leq m, j \notin \{k(0), \dots, k(p-1)\}\}$.
- (3) Calculate the projection of the remaining column vector $\mathbf{X}_j (j \in \mathbf{s})$ and the currently selected vector $\mathbf{X}_{k(p-1)}$, respectively:

$$\mathbf{P}\mathbf{X}_j = \mathbf{X}_j - (\mathbf{X}_j^T \mathbf{X}_{k(p-1)}) \mathbf{X}_{k(p-1)} (\mathbf{X}_{k(p-1)}^T \mathbf{X}_{k(p-1)})^{-1}, \quad j \in \mathbf{S} \quad (5.11)$$

- (4) Extract the sequence number of the wavelength that has the maximum projection value: $k(p) = \arg(\max(\|\mathbf{P}\mathbf{X}_j\|)), j \in \mathbf{s}$.
- (5) Let $\mathbf{X}_j = \mathbf{P}\mathbf{X}_j, j \in \mathbf{s}$.
- (6) $p = p + 1$. If $p < h$, return to Step (2) for a new loop.

Finally, the wavelength selected is $k(p), p = 0, \dots, h-1$. For each initial $k(0)$, cross-validation analysis of MLR or PLS is carried out after a loop, and the $k(p)$ corresponding to the minimum root mean square error of cross validation (RMSECV) was the final selection result.

SPA method has been applied in multivariate quantitative and qualitative analysis of a variety of spectra and achieved good results [20–22].

5.4 Variable Importance in Projection

For PLS regression, in addition to using the regression coefficient to select variables, important variables can also be screened through the weight vector \mathbf{w} , score vector \mathbf{t} , and load vector \mathbf{q} obtained in the PLS regression model. Variable importance in projection (VIP) is a wavelength selection method based on PLS regression model, which assesses the importance of the independent variable by its explanatory ability to the dependent variable [23, 24].

The VIP value of each wavelength is calculated by the following formula:

$$\text{VIP}_j = \sqrt{\frac{m \sum_{k=1}^h \left(q_k^2 t_k^T t_k \left(\frac{w_{jk}}{\|w_k\|} \right)^2 \right)}{\sum_{k=1}^h q_k^2 t_k^T t_k}} \quad (5.12)$$

where $j = 1, 2, \dots, m$, and m is the number of the whole wavelengths, while h is the optimal number of PLS principal component.

VIP comprehensively considers the contribution of spectrum to the construction of PLS score and the explanatory ability of PLS score to the concentration variable, which represents the importance of wavelength to model fitting. The interpretation ability of the wavelength to the concentration is transmitted through the PLS score. If the score has a strong interpretation ability to the concentration and the variable plays an important role in the construction of the score, then the VIP value of the variable will be large, indicating that this wavelength has a strong interpretation ability to the concentration [25]. The wavelength points of which VIP values are greater than 1 are usually selected as characteristic variables.

In addition to the VIP method, some methods also employed the generated parameters from PLS regression model, such as selectivity ratio (SR) by Kvalheim et al. [26] and significance multivariate correlation (sMC) by Tran et al. [27].

5.5 Interval Partial Least Squares Method

Interval partial least squares method (iPLS) is a wavelength interval selection method proposed by Nørgaard et al. [28, 29]. Its principle focuses on important spectral regions and removing interference from other regions to divide the whole spectrum into several sub-intervals with equal width. PLS regression is then conducted on each sub-interval to find the intervals whose prediction performance of model surpass the full-spectrum model based on RMSECV values. “The main advantage of using iPLS is the graphical output giving an overview of the spectra data and in displaying interesting spectral areas which could be selected” [30].

As shown in Fig. 5.4, in order to find a better interval combination with several different intervals corresponding to low RMSECV values, backward iPLS (BiPLS) [31] and forward iPLS (FiPLS) [31], synergy iPLS (SiPLS) based on greedy algorithm [29], and GA-iPLS based on genetic algorithm [30] were developed as expanding iPLS.

5.6 Moving Window PLS

iPLS divides the interval with a fixed and equal width, and the intervals are not overlapped with each other, which may lose some spectral information due to continuous features of NIR, resulting in making the optimization space of variable combination smaller. The basic idea of moving window PLS (MWPLS) is to continuously move a window along the spectral axis and build a model by cross validation for each moving wavelength interval. The overlapping intervals obtained by MWPLS can provide more spectral information than non-overlapping ones by iPLS. The sum of squares of residuals (PRESS or SSR) corresponding to a series of different windows (moving wavelength intervals) and the number of PLS factors can be obtained. Afterward, the spectral interval with high information content related to the component to

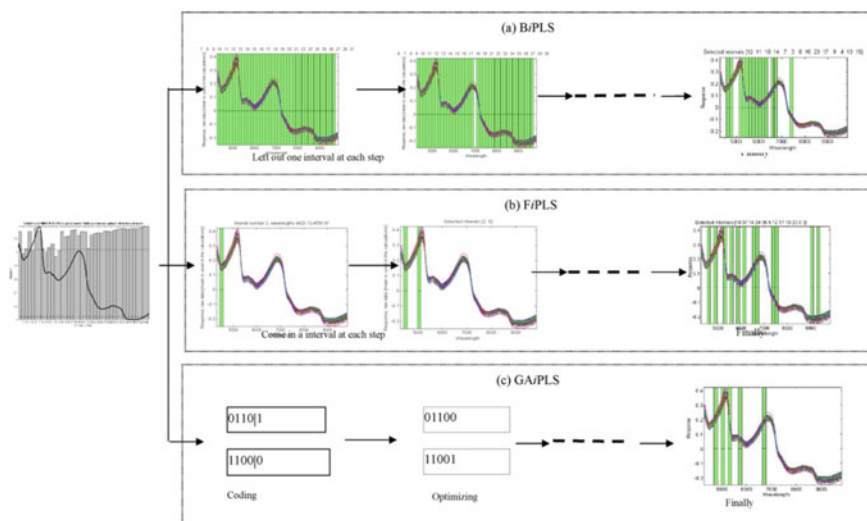


Fig. 5.4 Schematic diagram of wavelength interval selection by BiPLS, FiPLS, and GA-iPLS [30]

be measured can be selected by plotting it as shown in Fig. 5.5) [32]. As can be seen from Fig. 5.5, the information in the spectral range of $700\text{--}800\text{ cm}^{-1}$ is obviously better than that in the spectral range of $2400\text{--}3000\text{ cm}^{-1}$.

The width of window is a very important parameter in both the iPLS method and the MWPLS method. On the basis of MWPLS, some people proposed searching combination MWPLS (SMWPLS) and changeable size MWPLS (CSMWPLS) to better optimize the combination of wavelength interval, which allow for more precise wavelength interval selection [33, 34].

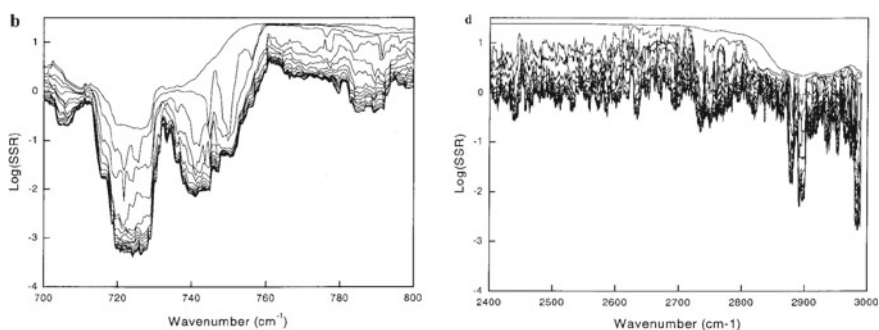


Fig. 5.5 Sum of squares of residuals in different spectral intervals obtained by MWPLS method [32]

5.7 Recursive Weighted PLS

Recursive weighted PLS (rPLS) is a variable selection method proposed by Rinnan et al. [35]. The basic idea is to take the regression coefficient of PLS regression model as the weight of corresponding variables, and recursively act on the original data matrix to increase the contribution degree of important variables and reduce the contribution degree of minor variables. Regression coefficients can reflect the importance of variables. A regression coefficient close to 0 indicates a minor variable, while a regression coefficient with a larger absolute value indicates an important variable.

The rPLS method recursively uses the estimated regression coefficient to reevaluate the independent variable, and then determines the reduced subset of the variable. PLS regression is carried out on this subset. The recursive relation is as follows:

$$\mathbf{X}_i = \mathbf{X}_i \text{diag}(\mathbf{b}_{i-1}) \quad (5.13)$$

where \mathbf{X}_i is the new variable with updated weight \mathbf{X} , \mathbf{X}_{i-1} is the variable with updated weight \mathbf{X} previously, and \mathbf{b}_{i-1} is the regression coefficient of the previous model. The algorithm first establishes the standard PLS model between \mathbf{X}_1 (initial independent variable) and \mathbf{y} and obtains the regression coefficient \mathbf{b}_1 . Then the PLS model is repeatedly established according to the above recursive relation until the regression coefficient no longer changes. That is, only elements 0 and 1 are included in the final regression coefficient vector.

This method is usually able to converge to a finite number of variables (usually equivalent to the number of factors in PLS models), which is very conducive to the interpretability of the model. In addition, rPLS method only needs to determine the number of factors, without setting parameters such as threshold or confidence interval. Thus, it has a high degree of automation.

5.8 Elimination of Uninformative Variables

The elimination of uninformative variables (UVE) [36] is a kind of wavelength selection method based on PLS regression coefficient \mathbf{b} , and the basic idea of this method is to take the regression coefficient as the measurement index of wavelength importance. The specific algorithm is as follows:

- (1) PLS regression was performed on the spectral matrix \mathbf{X} ($n \times m$) and concentration matrix \mathbf{y} ($n \times 1$) of the calibration set, and the optimal number of PLS factors f was selected.
- (2) A noise matrix \mathbf{R} ($n \times m$) is artificially generated, and \mathbf{X} and \mathbf{R} are combined to form a matrix \mathbf{XR} ($n \times 2m$), in which the first m is listed as \mathbf{X} and the second m is listed as \mathbf{R} .

- (3) Build PLS model with matrix \mathbf{XR} and \mathbf{y} by removing one sample at a time based on cross validation, then n PLS regression coefficients were obtained to form matrix \mathbf{B} ($n \times 2m$).
- (4) Calculate the standard deviation \mathbf{s} ($1 \times 2m$) and mean vector \mathbf{me} ($1 \times 2m$) of matrix \mathbf{B} ($n \times 2m$) in columns, and then calculate the stability as follows:

$$h_i = me_i/s_i, i = 1, 2, \dots, 2m \tag{5.14}$$

- (5) In the interval of $[m + 1, 2m]$, take the maximum absolute value of h as $h_{max} = \max(\text{abs}(\mathbf{h}))$.
- (6) In the interval of $[1, m]$, the variables in matrix \mathbf{X} of which h values are less than h_{max} are removed, and the remaining variables are formed into a new matrix \mathbf{X}_{UVE} selected by UVE method.

UVE method integrates noise and concentration information in the selection of wavelengths, which is also intuitive and practical (Fig. 5.6). Some literatures have shown that the UVE result is superior to the variable selection method based on correlation coefficients and other methods. Wavelength selection methods using PLS regression coefficient \mathbf{b} or weight \mathbf{w} also include interactive variable selection (IVS) [37] and ordered predictors selection (OPS) [38].

The combination of Monte Carlo (MC) sampling method and PLS regression coefficient \mathbf{b} for the screening of variables is a kind of method that has attracted more attention and wide application recently. In this method, MC strategy is introduced into UVE-PLS to replace the traditional leave one out cross validation (LOOCV) [7]. Each time, a certain proportion of samples were randomly selected from the sample set as the training samples, and the PLS regression model was established to obtain the regression coefficient \mathbf{b} . This step was then repeated for N times. It uses MC

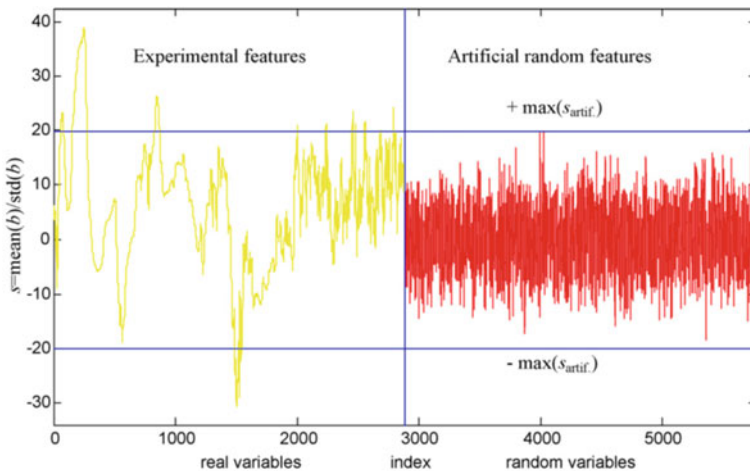


Fig. 5.6 Schematic diagram of selecting wavelength by UVE method [30]

sampling or random sampling to randomly select a part of samples from the calibration set for PLS modeling. The processes of sampling and modeling are repeated for hundreds of times, and then the wavelength corresponding to the significant regression coefficient \mathbf{b} is selected in accordance with certain rules. The MC-UVE method, which is composed of MC and UVE method, has attracted extensive attention [39]. The regression coefficient matrix \mathbf{B} ($N \times m$) composed of N regression coefficients of PLS models is obtained, then the i th variable ($i = 1, 2, \dots, m$) can be calculated by the following formula:

$$h_i = \text{mean}(\mathbf{b}_i) / \text{std}(\mathbf{b}_i), \quad i = 1, 2, \dots, m \quad (5.15)$$

where $\text{mean}(\mathbf{b}_i)$ and $\text{std}(\mathbf{b}_i)$ represent the mean and standard deviation of the i th variable's regression coefficient, respectively. The larger the absolute value of h_i is, the more important the corresponding variable is. Whether to remove the i th variable is decided according to the absolute value of h_i . As shown in Fig. 5.7 [40], since its variables are directly determined by the stability, it is more convenient than the UVE method to estimate the cutoff threshold by adding random noise variables into the original data matrix.

Han et al. adopted an integration strategy to improve the stability of the MC-UVE algorithm. The results showed that the selected cumulative frequencies of each wavelength varied from high to low after repeated running of MC-UVE, and the reliability and prediction ability of the MC-UVE algorithm were significantly improved after removing wavelengths with lower selected frequencies by setting thresholds [41].

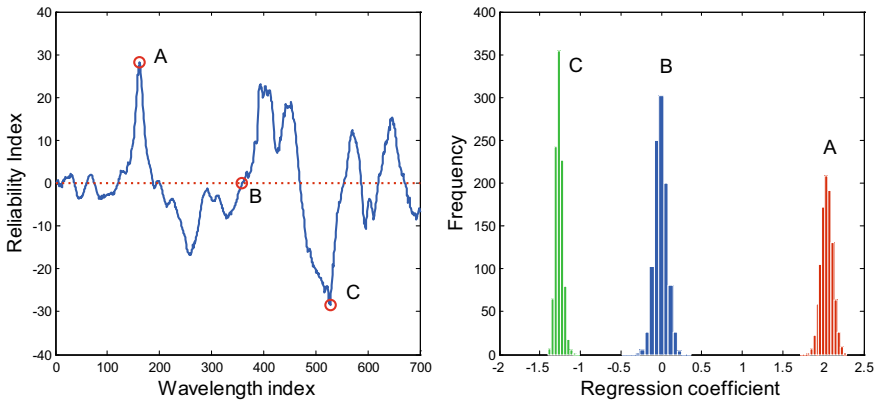


Fig. 5.7 Stability of variable (left panel) and regression coefficient frequency (right panel) obtained by MC-UVE method [40]

5.9 Global Optimization Methods

Random search and global optimization algorithms, also known as swarm intelligence or metaheuristics, such as genetic algorithm (GA), simulated annealing algorithm (SA), tabu search (TS), ant colony optimization (ACO), particle swarm optimization (PSO), cuckoo search (CS), firefly, bat algorithm (BA), gravitational search algorithm (GSA), random frog (RF), grey wolf optimizer (GWO), whale optimization algorithm (WOA), and cat swarm optimization (CSO), have shown strong search ability in solving real problems [42]. They can approach the optimal solution of the problem in a reasonable time. These algorithms involve artificial intelligence, statistical thermodynamics, biological evolution, and bionics, and most of them are based on certain natural phenomena, so they are also called intelligent optimization algorithms [43]. These methods are easy to introduce heuristic logic rules, and the algorithm principle is intuitive and easy to code and implement, and can find the global optimal solution with a large probability. One of the biggest characteristics of these methods is that they can retain the combination advantage among variables. These advantages have made the stochastic optimization algorithm successfully applied to many optimization problems, such as artificial neural network (ANN) or support vector machine (SVM) parameter optimization and wavelength selection in spectral data.

5.9.1 Genetic Algorithm

Genetic algorithm (GA) was originally proposed by Holland in 1975. It refers to the natural selection and genetic mechanism in the biological world and uses the operation of operators such as selection, crossover, and mutation to keep the variables with better objective function value and eliminate the ones with worse objective function value by means of continuous genetic iteration, finally achieving the optimal result. At present, genetic algorithm has been widely applied in the field of analytical chemistry, among which good results have been obtained in variable selection [44, 45].

The realization of GA mainly includes five basic elements: parameter coding, initialization of the population, design of fitness function, genetic operation design, convergence criterion, and selection of variables. It can be seen from Fig. 5.8 about the specific genetic algorithm implementation flow diagram.

(1) Parameter coding

Because GA is not suitable to deal with spatial data directly, it is necessary to express them as genotype string structure data of genetic space by encoding, which generally adopts binary string form based on 0/1 character. A problem with m parameters (such as wavelength) can be represented by a string of vectors (corresponding to chromosomes) containing $m \times p$ characters (corresponding to genes), where p represents the

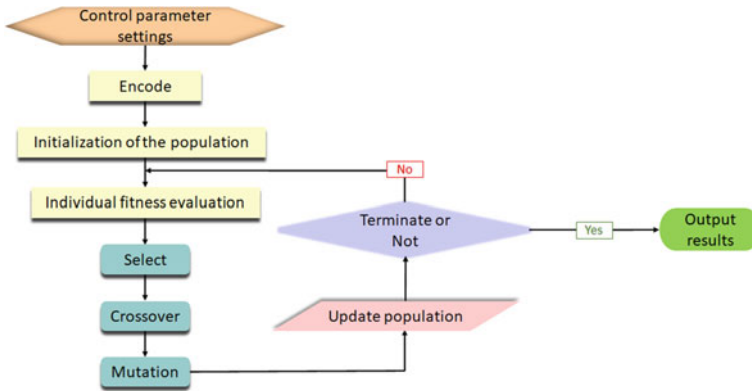


Fig. 5.8 Flow chart of GA algorithm implementation

number of genes required for each parameter. For wavelength selection, p usually selects 1, that is, each gene in a chromosome which corresponds to an actual parameter. If the gene is 1, it means the parameter it represents is selected. While the gene is 0, it is not selected.

(2) Initialization of the population

An initial population of a given size is generated randomly or according to certain restrictive conditions. The size of the population, i.e., the number of individuals (chromosomes), can be selected according to the number of parameters (genes), generally 30–100.

(3) Design of fitness function

The GA evaluates the individual according to the fitness function, which is used as the basis for future genetic operation. In the whole process of search evolution, only fitness function is related to the specific problem solved. Therefore, the determination of fitness function is very important. For wavelength selection, the fitness function can adopt the correlation coefficient (R), RMSECV, or RMSEP between the predicted value and the actual value of the dependent variable as parameters in the process of cross validation or prediction.

(4) Genetic manipulation design

Selection: Selection operator, also known as replication operator, directly inherits the individuals with high fitness to the next generation by selection or creates new individuals through crossover or mutation and then inherits to the next generation. The selection operation is based on the fitness assessment of the individuals in the population. The purpose of selection is to avoid genetic defects and improve global convergence and computational efficiency. Selection methods include fitness ratio, optimal preservation, certain sampling, and sorting selection, among which the most

commonly used selection method is fitness ratio method, also known as the wheel method where the selection probability of each individual is proportional to its fitness.

Crossover: Crossover is the process that when two paired chromosomes exchange parts of their genes in a way that creates two new individuals. It is the most important operator in the GA and also the main method of generating new individuals, by which the search process is achieved to a large extent. Therefore, it determines the global search ability of the GA. The crossover operators include random one-point crossover, two-point and multi-point crossover, and uniform crossover and arithmetic crossover, and the crossover probability is generally selected with 0.5–0.8. Before crossover operation, individuals in the group must be paired. At present, random pairing strategy is commonly used, that is, N individuals in the group randomly form $N/2$ pairs of paired individuals, and crossover operation is carried out between two individuals in these paired individuals.

Mutation: Mutation is the complement operation of some genes in the individual chromosome coding string, that is, 0 becomes 1 or 1 becomes 0. The purpose of introducing mutation operator is to maintain the diversity of population, prevent premature convergence phenomenon, and improve the local search ability of GA. The crossover operator and mutation operator are combined to complete the global search and local search of the search space, so that the GA can complete the optimization process with good search performance. The simplest mutation operator is the basic location mutation operator, namely, one or more genes are randomly selected from the individual to change with the mutation probability, which is in the range of 0.01 ~ 0.1. In addition, there are mutation operators such as uniform mutation, non-uniform mutation, boundary mutation, and Gaussian mutation.

(5) Convergence criterion

All conventional mathematical programming methods have strict convergence criteria in mathematics, but the convergence criteria of GA are basically heuristic. Therefore, GA has more criteria, such as calculation time, computer variables, or the quality of the solution to determine criteria. Selection of the number of genetic iterations is a common convergence termination condition, and its value range is generally 100–1000.

(6) Selection of variables

After the termination of genetic iteration, all the variables were rearranged according to the selection frequency, and then the optimal number of variables was selected by plotting the number of selected variables and the fitness function, and the selected variables were then obtained.

Because of its global optimization and easy realization, GA has become a more commonly used and effective wavelength selection method. The selection of wavelengths can not only optimize the model and improve its prediction ability, but also establish a robust model with little influence of external factors such as ambient temperature. In addition, the spectral region of the component to be measured can be better explained by the selected characteristic wavelength. GA can be used to select

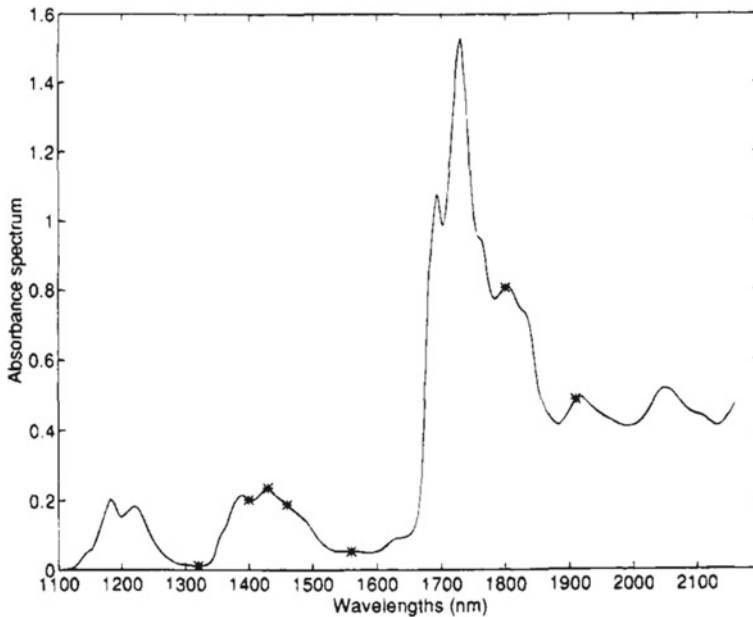


Fig. 5.9 The wavelength marked with asterisk is the variable selected by GA [44]

the variables of various quantitative calibration methods, as shown in Fig. 5.9. GA and MLR are combined to determine the hydroxyl number of polyether polyols by NIR spectroscopy, and MLR method based on the seven variables selected by GA is equivalent to the results of full spectrum by PLS.

However, the following problems should be paid attention to in practice: (1) Since the initial population of the GA is selected at random, selection, crossover, and mutation also have strong randomness. Thereafter, the consistency of each wavelength selection result cannot be guaranteed. (2) When using the GA, the ratio between the number of wavelengths and the number of samples in calibration set is generally less than 4 according to experience, otherwise the results obtained are not reliable. (3) The appropriate fitness function is particularly important for the GA, and the results obtained by different fitness functions will be quite different.

5.9.2 Simulated Annealing Algorithm

Simulated annealing (SA) algorithm was proposed by Kirkpatrick et al. in 1983 [46]. Its basic ideas originate from the principle of metal annealing. Annealing is the heating and cooling of a material at a specific rate to increase the size of the grain and reduce defects in the crystal lattice. The atoms in the material will stay at the position where the internal energy has a local minimum. Heating increases the

energy, and the atoms will leave the original position and move randomly in other positions. Annealing cools more slowly, making it more likely that the atoms will find their internal energies lower than they originally were. According to the Metropolis criterion, the probability that a particle tends to equilibrium at temperature T is $e^{-\Delta E/(kT)}$, where E is the internal energy at temperature T , ΔE is its change value, and k is Boltzmann's constant. The steps of SA to solve the combination optimization problem are as follows: the internal energy E is simulated as the objective function value F and the temperature T evolves into the control parameter T . Starting from the initial solution and the initial value of the control parameter T_0 , the iteration of "generating a new solution \rightarrow calculating the objective function \rightarrow judging whether to accept \rightarrow accepting or rejecting" is repeated for the current solution, and the T value is gradually attenuated. The current solution at the termination of the algorithm is the approximate optimal solution [47]. The steps for SA are as follows:

- (1) Set the termination temperature T_e , the initial temperature T_0 , the cooling coefficient β , and the total number of iterations L , and generate an initial solution \mathbf{x}_0 at random. Let \mathbf{x}_{best} be equal to \mathbf{x}_0 , and calculate the objective function value $E(\mathbf{x}_0)$.
- (2) Set the iteration number $i = 1$.
- (3) For the current optimal solution \mathbf{x}_{best} , a new solution \mathbf{x}_{new} is generated according to a neighborhood function. First calculate the new objective function value $E(\mathbf{x}_{\text{new}})$, and then calculate the increment of the objective function value $\Delta E = E(\mathbf{x}_{\text{new}}) - E(\mathbf{x}_{\text{best}})$. If $\Delta E < 0$, $\mathbf{x}_{\text{best}} = \mathbf{x}_{\text{new}}$; if $\Delta E > 0$, calculate the probability $p = \exp(-\Delta E / T_i)$ and generate a uniformly distributed random number s in the interval $[0, 1]$. if $p > s$, $\mathbf{x}_{\text{best}} = \mathbf{x}_{\text{new}}$, otherwise \mathbf{x}_{best} does not change.
- (4) $i = i + 1$. If i reaches the maximum number of iterations L , the iteration will be terminated; otherwise, step (3) will be returned.
- (5) The current objective function value is compared with the historical objective function value. If it is smaller, the historical value is updated with the parameters of the current state. Then, let $T_{k+1} = T_k \cdot \beta$ to cool the temperature.
- (6) If $T_{k+1} > T_e$, return to step (3) after initializing the number of iterations ($i = 1$); otherwise, the calculation ends and the wavelength index value corresponding to the minimum value of the objective function comes as output.

5.9.3 Particle Swarm Optimization

Particle swarm optimization (PSO) was first proposed by Eberhart and Kennedy in 1995 [48]. Its basic concept is derived from the study on the foraging behavior of birds, and it is an evolutionary computing technology derived from the study on the predation behavior of birds. It finds the optimal region in the complex search space through the interaction between particles [49, 50].

PSO is similar to GA, which is a global optimization technology based on swarm evolution. The system initializes a group of random solutions and finds the optimal

value through iteration. However, there is no crossover and mutation process like GA; instead, it follows the optimal particle in the solution space to search. Compared with GA, PSO is simple, easy to implement, and has no more parameters to adjust.

In the PSO algorithm, the solution of each optimization problem is a bird in spatial search, which is abstracted as a particle without mass and volume and extended to a multi-dimensional space. The position and flight speed of a particle in multi-dimensional space are, respectively, represented as a vector. All particles have an adaptive value determined by an evaluation function. In addition to knowing the best position they have found so far and current position, the particles also know the best position they have found so far for all the particles in the entire population. The particle is determined by its own experience and the best experience of its companions. The steps of PSO algorithm are as follows:

- (1) Initialize a group of random particles. The number of particles depends on the complexity of the problem. For general optimization problems, a good result can be obtained by taking 20–40 particles. Randomly initialize the position and velocity of each particle in the crowd and make them be scattering across the entire space. The i th particle is represented by an m -dimensional vector (spectrum) $\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{im})$. The “flying” velocity of the i th particle, that is, the rate of position change of the i th particle, is expressed as $\mathbf{v}_i = (v_{i1}, v_{i2}, \dots, v_{im})$.
- (2) Evaluate the fitness of each particle. Store the current position and fitness value of each particle in the pbest of each particle, and store the position and fitness value of all the individuals with the best fitness value in gbest. Remember the optimal position of the i th particle so far, that is, the individual optimal position is $\mathbf{p}_i = (p_{i1}, p_{i2}, \dots, p_{im})$. The optimal position of the whole particle swarm so far is the global optimal position $\mathbf{p} = (p_{g1}, p_{g2}, \dots, p_{gm})$, and the algorithm assumes that all particles move toward the individual and global optimal positions.
- (3) Update the velocity and position of the particle with the following formula:

$$v_{id}(\text{new}) = w \times v_{id}(\text{old}) + c_1 r_1 \times (p_{id} - x_{id}) + c_2 r_2 \times (p_{gd} - x_{id}) \quad (5.16)$$

$$x_{id}(\text{new}) = x_{id}(\text{old}) + \mu \times v_{id}(\text{new}) \quad (5.17)$$

where $d = 1, 2, \dots, m$ and w are non-negative constants, called inertia factors, which are used to balance global search and local search, and are between 0 and 1. The learning factors c_1 and c_2 are non-negative constants and usually take an integer value between 0 and 4. r_1 and r_2 are random numbers between 0 and 1. μ is called a constraint factor and is used to control the weight of speed.

- (4) For each particle, the fitness value is compared with the best position experienced. If it is better than the previous position, it can be seen as the current best position.

- (5) Compare all current values of p_{best} and g_{best} , and update g_{best} .
- (6) If the stop condition is met (the minimum error standard is specified or the iteration has reached the specified number of times), the search will stop and the results will be output; otherwise, the search will return to Step (3) and continue.

In order to avoid the convergence of the PSO algorithm to the local search, the ability of the algorithm to overcome the local search is enhanced to force a certain percentage of particles to fly randomly (for example, 10%) without following the two optimal values. PSO has developed into a variety of deformation and improvement algorithms, such as PSO with compression factor [51], PSO with changeable learning factor, variable dimension PSO [52], and second-order oscillating PSO [53].

5.9.4 Ant Colony Algorithm

Ant colony optimization (ACO) algorithm was proposed by Marco Dorigo in 1992 [54], which was inspired by the behavior of ants finding their way in search of food. Ants are social insects, and a group of ants working together can easily find the shortest path from the nest to the food source. Through a large number of studies, it has been found that individuals of ants transmit information through pheromones left on their paths, and ants can guide their forward direction according to the concentration of pheromones. Therefore, the more ants that travel along a particular path there are, the more probability it is that a latecomer will choose that path. This constitutes a positive feedback phenomenon of ant colony behavior.

In the ACO algorithm, human worker colony with a finite size can cooperate to search for a better solution to solve the optimization problem. Each ant builds a feasible solution from the selected initial state according to the criteria given by the problem. Each ant collects information about the characteristics of the problem and its own behavior, and uses this information to modify the presentation of the problem. Instead of direct communication, ants use pheromones to guide information exchange. Each ant can find a solution, but it is probably a bad one. High-quality solutions can be found through global cooperation among all the individuals in the group [55, 56]. The flow of the basic ACO is shown in Fig. 5.10.

There are many strategies for variable selection using ACO algorithm [57, 58], one of which is introduced as follows:

- (1) Construct pheromone vector τ , whose dimension is $1 \times m$, and m is the number of variables. Initialize it by assigning all elements to 1. The number of ants in the ant colony φ , the number of variables l to be selected, and the number of iterations are determined first.
- (2) Calculate probability vector p and cumulative probability vector cp as given in the following formula:

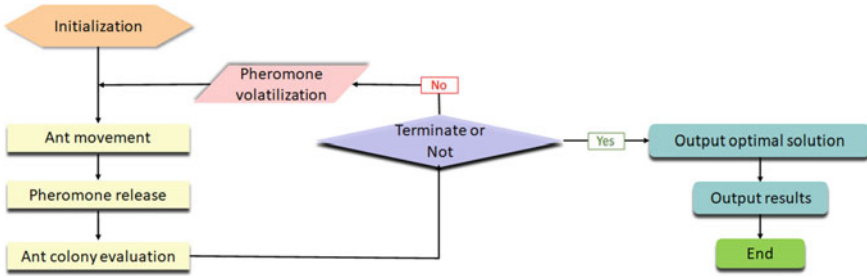


Fig. 5.10 Flow chart of the basic ACO algorithm

$$p_k = \frac{\tau_k}{\sum_{k=1}^m \tau_k}, k = 1, 2, 3, \dots, m \quad (5.18)$$

$$cp_k = \sum_{k=1}^k p_k, k = 1, 2, 3, \dots, m \quad (5.19)$$

- (3) The cumulative probability vector cp initializes the ant colony φ according to the random variable generated by uniform distribution, so that the variable with high pheromone concentration can be selected more frequently.
- (4) Calculate the fitness of each ant according to the following formula:

$$\text{Fitness function : } G_f = \frac{1}{PRESS_f \times l_f}, f = 1, 2, 3, \dots, \varphi \quad (5.20)$$

$$\text{Normalize } Gn_f : Gn_f = 0.8 \times \frac{G_f}{\sum_{f=1}^{\phi} G_f}, f = 1, 2, 3, \dots, \varphi \quad (5.21)$$

- (5) The first 50% ants with the highest fitness (φ_{best}) were used to update the pheromone vector τ according to the fitness. The higher the fitness, the more pheromones would be released as formula 5.22 shows.

$$\tau_k(t+1)\tau_k(t)[\tau_k(t) \times Gn_f], f = 1, 2, 3, \dots, \varphi_{\text{best}}, k \in \beta_f \quad (5.22)$$

where β_f is the variable selected by the f th ant.

- (6) Volatilization of pheromones, that is, pheromones will dissipate over time to prevent the infinite accumulation of pheromones as formula 5.23 shows.

$$\tau_k \tau_k \times \rho \quad (5.23)$$

where ρ is constant, $0 \leq \rho \leq 1$.

- (7) Determine whether the termination condition is met. Otherwise, return to step (2) to continue iteration.

ACO algorithm has been successfully used for the selection of efficient wavelengths in the scope of Vis-NIR [59], NIR [60–62], MIR, Raman [63, 64], and fluorescence [65] spectroscopy with various kinds of applications.

5.10 Model Population Analysis-Based Methods

Model population analysis (MPA) firstly proposed by Li et al. [5] is a general framework for developing a new type of chemometrics algorithm for modeling in various aspects of variable selection, outlier detection, model comparison, and applicability domain definition [66]. The core idea of MPA is to statistically extract and analyze useful information from output of the sub-models built with a large population of generated sub-dataset. The three key elements of MPA are as follows:

- (1) random sampling is used to randomly generate N sub-datasets (e.g., 500);
- (2) for each sub-dataset, a sub-model is built, and there are thus N sub-models; and
- (3) Statistically analyze an outcome of interest of all N sub-models.

As can be seen from Fig. 5.11, for variable selection methods based on MPA, a large population of sub-datasets are obtained with the aid of several sampling methods on both variable space and sample space, including MC sampling, Bootstrap sampling, binary matrix sampling, and permutation. For each sub-dataset, a sub-model is then built with the modeling method such as PLS, PCR, MLR, and SVM. Thereafter, the distributions of regression coefficients for uninformative variables and informative variables, the distribution of RMSECV for each variable subset, and the difference between the two distributions of RMSECV with and without the specific variable are used to assess the variable or variable subset based on some criteria, such as mean value, standard deviation (STD), 95% confidence interval, statistical test, and so on. During the last decade, a great many variable selection methods based on MPA strategy were proposed, such as competitive adaptive reweighted sampling (CARS) [67], variable combination population analysis (VCPA) [68], modified VCPA (mVCPA) [68], permutation combination population analysis (PCPA) [69], random frog (RF) [70], stability and variable permutation (SVP) [71], iteratively retains informative variables (IRIV) [9], iteratively variable subset optimization (IVSO) [72], variable permutation population analysis (VPPA) [73], bootstrapping soft shrinkage (BOSS) [8], variable iterative space shrinkage approach (VISSA) [74], randomization test (RT) [75], sampling error profile analysis-least absolute shrinkage and selection operator (SEPA-LASSO) [76], and weighted voting strategy-LASSO (WV-LASSO) [77].

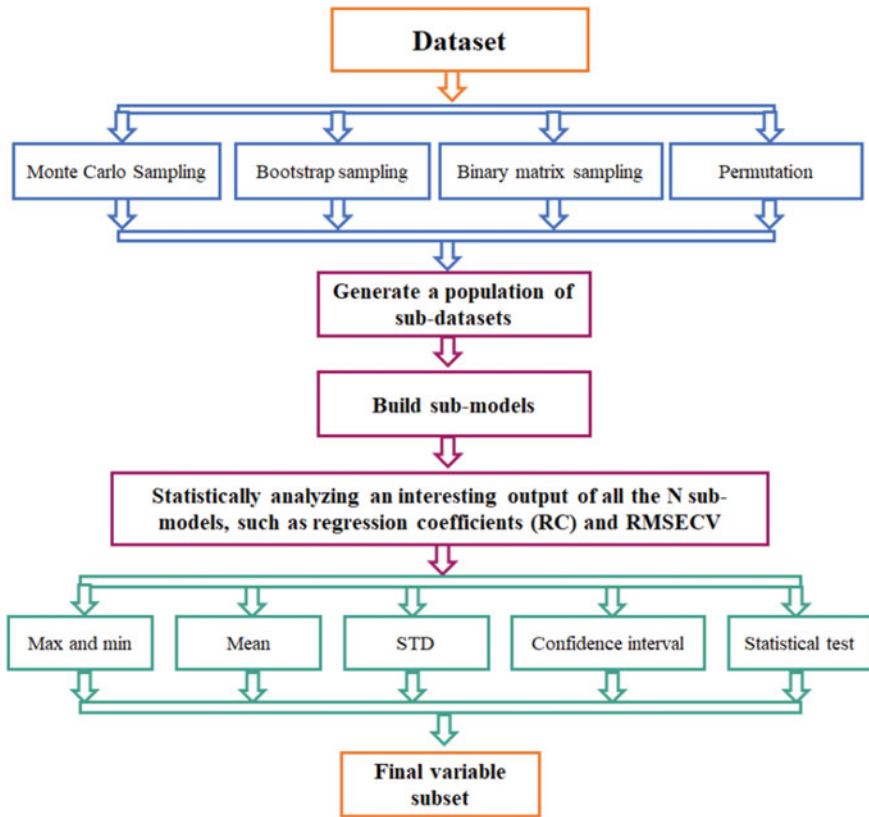


Fig. 5.11 The framework and three key elements of model population analysis (MPA) strategy in developing variable selection methods

5.10.1 Competitive Adaptive Reweighted Sampling

The competitive adaptive reweighted sampling (CARS) algorithm is very popular in analysis of spectral data [67]. In this method, each variable is regarded as an individual and the selection process of variables is in an iterative way. At the same time, the exponentially decreasing function (EDF) is introduced to control the rate of remaining number of variables, which has high computational efficiency, and can overcome the combination explosion problem in variable selection to some extent, and then screen out the optimal subset of variables. The implementation steps of the algorithm are as follows:

- (1) MC sampling method was used for sampling N times, and each time 80% samples were randomly selected from the samples set as the calibration set. The extracted spectral matrix \mathbf{X} ($n \times m$) and concentration matrix \mathbf{y} ($n \times 1$) were used to establish PLS regression models.

- (2) Eliminate wavelengths with relatively small absolute value of regression coefficient by enforce with the aid of EDF. During the i th sampling, the rate of remaining number of wavelengths r_i can be obtained according to the following EDF formula:

$$r_i = ae^{-ki} \tag{5.24}$$

where a and k are constants, and the calculation formula of a and k are as follows:

$$a = \left(\frac{m}{2}\right)^{\frac{1}{N-1}} \tag{5.25}$$

$$k = \frac{\ln\left(\frac{m}{2}\right)}{N-1} \tag{5.26}$$

It can be seen that in the first sampling, all m variables are used for modeling, r_1 is thus equal to 1. When the N th sample is run, only two wavelengths are used, so $r_N = 2/m$. Figure 5.12 vividly shows the EDF attenuation process with N of 50. In the first stage, the number of variables decreases quickly for fast selection of variables and, in the second stage, the number of variables decreases slowly for refined selection of variables. This way can not only improve the calculation speed, but also screen out important variables.

- (3) The variables with large absolute value of regression coefficient in PLS model were screened out through N times of sampling, and the PLS regression

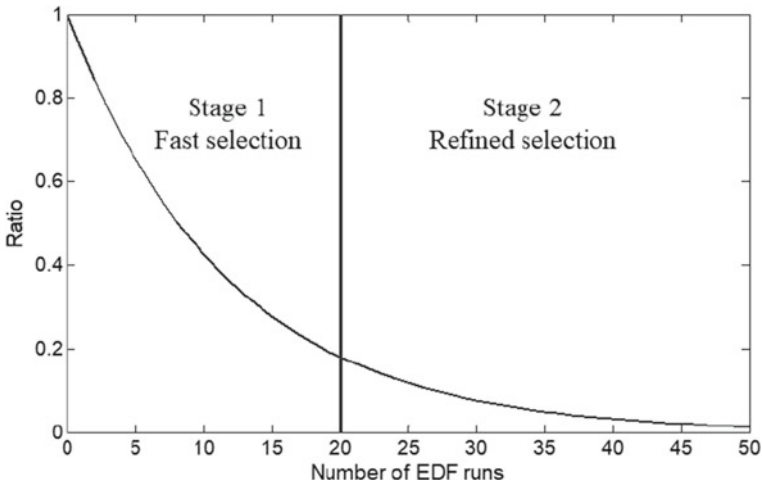


Fig. 5.12 Graphical illustration of exponentially decreasing function (EDF). In the first stage, the number of the variables is reduced rapidly. While in the second stage, it decreases in a mild way, namely, refined selection

model was established with the new variable subset generated each time. The RMSECV of each model was calculated, and the variable subset with the smallest RMSECV value was selected as the optimal variable subset.

Currently, CARS algorithm has been widely used in the selection of wavelength variables of NIR spectroscopy, UV-Vis spectroscopy, Raman spectroscopy and laser induced breakdown spectroscopy [78–81], and its effect is better than SPA and UVE methods in most cases. However, owing to random sampling, CARS presents the unstable result when implemented many times.

5.10.2 Iteratively Retaining Informative Variables

Iteratively retaining informative variables (IRIV) is a representative variable selection method based on MPA framework. Specially, IRIV first adopts BMS to get N data subset from a given set of samples. All variables can be divided into strongly informative variable, weakly informative variable, uninformative variable, and interfering variable based on statistical analysis. It iteratively removes uninformative and interfering variables which are useless for the model and retains the informative variables for the model. For the original spectral data of p -dimensional variables of m samples, IRIV selects variables through the following four steps:

- (1) Generate a binary matrix \mathbf{A}' with m rows and p columns containing only “1” and “0”. The number “1” represents the variables that are included for modeling, while “0” represents the variables that are not included. In each column of \mathbf{A}' , the number of ones and zeros in each column is the same. Figure 5.13 shows the process of generation of sub-datasets by BMS method. Each PLS model is then established according to the samples selected from each row of matrix \mathbf{A} . RMSECV value obtained from fivefold cross validation was taken as the evaluation metric, and the vector with the size of $m \times 1$ was denoted as \mathbf{RMSECV}_0 .
- (2) To assess each variable’s importance through its interaction with other variables, in the i th column of the matrix \mathbf{A} ($i = 1, 2, \dots, p$), “1” is replaced by “0”, and “0” is replaced by “1” to get a new matrix \mathbf{B} as shown in Fig. 5.13. Similarly, PLS model is established based on the samples selected from each row of matrix \mathbf{B} , and a vector of $m \times 1$ is obtained, which is denoted as \mathbf{RMSECV}_i . Define φ_0 and φ_i to evaluate the importance value of each variable:

$$\varphi_{0k} = \begin{cases} k^{th} \mathbf{RMSECV}_0 & \text{if } A_{ki} = 1 \\ k^{th} \mathbf{RMSECV}_i & \text{if } B_{ki} = 1 \end{cases}, \quad \Phi_{ik} = \begin{cases} k^{th} \mathbf{RMSECV}_0 & \text{if } A_{ki} = 0 \\ k^{th} \mathbf{RMSECV}_i & \text{if } B_{ki} = 0 \end{cases} \quad (5.27)$$

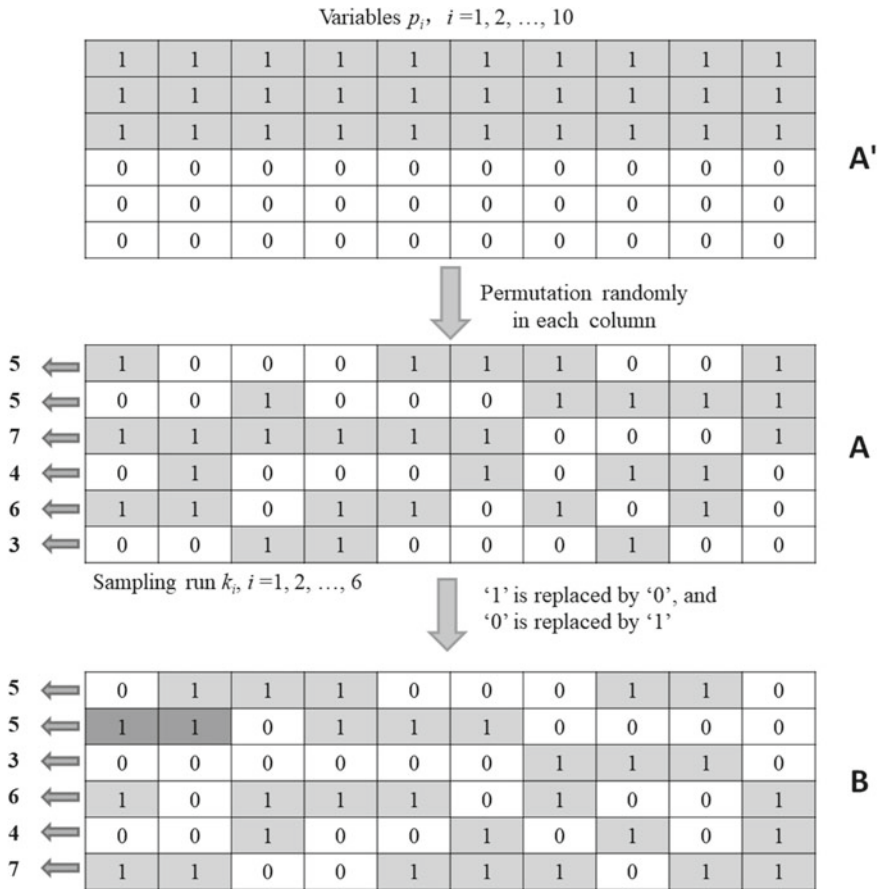


Fig. 5.13 Graphical illustration of binary matrix sampling (BMS)

where k th represents the k row in the vector, and k th \mathbf{RMSECV}_0 and k th \mathbf{RMSECV}_i represent the values of the k th row in the vectors \mathbf{RMSECV}_0 and \mathbf{RMSECV}_i , respectively. The mean values of φ_0 and φ_i are denoted as $M_{i,in}$ and $M_{i,out}$, respectively. DM_i is obtained by subtracting the two mean values of $M_{i,in}$ and $M_{i,out}$. $P = 0.05$ is defined as the threshold for Mann-Whitney U test, and the variables are finally divided into four categories:

if $DM_i < 0$ and $P_i < 0.05$, it is a strongly informative variable;

if $DM_i < 0$ and $P_i > 0.05$, it is a weakly informative variable;

if $DM_i > 0$ and $P_i > 0.05$, it is uninformative variable;

if $DM_i > 0$ and $P_i < 0.05$, it is the interfering variable.

- (3) In each iteration, strongly and weakly informative variables are retained, and uninformative variables and interfering variables are eliminated. Return to Step (1) and proceed to the next iteration until only the strongly and weakly informative variables are left.
- (4) Backward elimination is used to further optimize the variable subset with informative variables. Firstly, the PLS model of all left informative variables, denoted as t , was established to obtain $RMSECV_t$. Then, by eliminating the j th variable ($j = 1, 2, \dots, t$), PLS model was established for $t-1$ variables to obtain $RMSECV_{-j}$. If $RMSECV_{-j}$ is less than $RMSECV_t$, the j th variable would be excluded, otherwise it would be retained. In this process, the remaining variables are the final characteristic variables.

Figure 5.14 shows four types of variables screened by the IRIV method for the diesel fuels data set. 1236 nm is a strongly informative variable. RMSECV significantly increased after this variable was removed from the model. 1050 nm is a weakly informative variable. RMSECV slightly increased after this variable was removed from the model. 1468 nm is an uninformative variable, and RMSECV decreased slightly after this variable was removed from the model. 1502 nm is an interfering variable, and RMSECV decreased significantly when this variable was removed from the model.

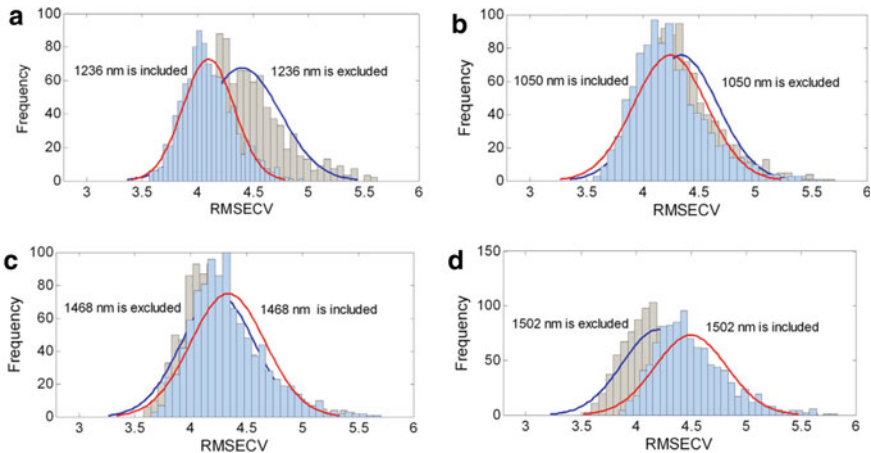


Fig. 5.14 Four types of variables screened by IRIV method (**a** strongly informative variable, **b** weakly informative variable, **c** uninformative variable, **d** interfering variable) [9]

5.10.3 Variable Combination Population Analysis

Variable combination population analysis (VCPA) is another widely used method based on the framework of MPA [68]. Like IRIV, VCPA also uses BMS to generate a large group of random variable combinations. When there are enough random variable combinations, the model prediction error of these combinations should be a normal distribution, as shown in Fig. 5.15. The left and right ends of the distribution are the enrichment areas of good models and bad models, respectively. The probability of better variable combination in the enrichment area of good model is large, while the probability of occurrence in the enrichment area of bad model is small. Based on this feature, VCPA only takes a certain proportion of good models in the distribution each time, and counts the frequency of each variable existing in these good models. Afterward, it assesses the variables according to the ranking of frequency from high to low, and then removes a certain proportion of variables forcibly at the end by using EDF because these variables make the least contribution to the good model. Next, it continues to use BMS to generate a large group of random variable combinations with the remaining variables, and iteratively retained variables according to the ratio of remaining variable by EDF. VCPA uses EDF to shrink the variable space continuously, and retained variables that contribute greatly to the model for each iteration. When the variable space becomes smaller, the mean RMSECV value of all variable combinations gradually decreases as shown in Fig. 5.16. Therefore, with a small and optimized variable space, VCPA can then select the variable subset that significantly improves the prediction performance of the model.

The detailed process of VCPA algorithm is as follows:

Step (1): Use the BMS method to generate a binary matrix \mathbf{M} with only “1” or “0”. The matrix \mathbf{M} ($k \times p$) contains k rows and p columns. The rows represent the number of random samples. The columns correspond to the columns of sample matrix \mathbf{X} .

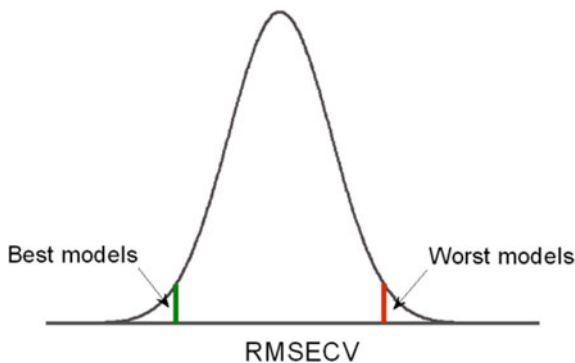


Fig. 5.15 Distribution of RMSECV values based on a population of built models. The lower the RMSECV, the better the model. The best models are located in left side of the distribution [68]

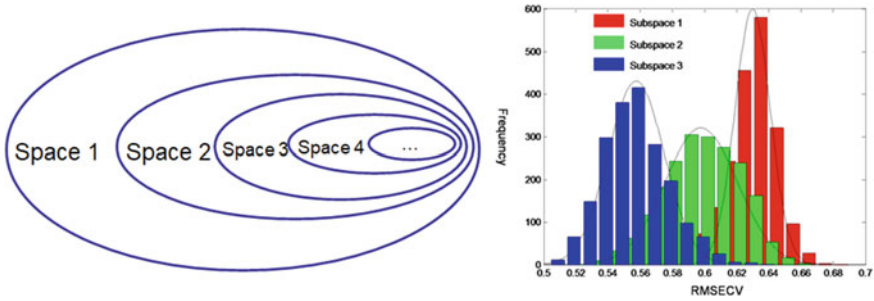


Fig. 5.16 The evolution of variable spaces after each EDF iteration and the RMSECV distribution with the corresponding variable space

Each column in \mathbf{M} is disrupted. After disruption, each row is a random combination of “0” and “1”, that is, each row is a combination of random variables, and the number of “1” in each column remains unchanged.

Step (2): Each row in \mathbf{M} is a random variable subset. Each row is modeled by PLS, and the RMSECV value is calculated to assess each variable subset. The smaller the RMSECV value, the better the model.

Step (3): Sort RMSECV value from small to large, and the ratio, σ , of all RMSECV values is regarded as good model, i.e., the top $k\sigma$ of RMSECV ranking corresponds to the combination of variables. Then count the number of occurrences of each variable in this $k\sigma$ combinations. The more variables appear, the more they contribute to a good model, and vice versa.

Step (4): For the variables with small contribution from the good model, the EDF is used to forcibly remove them through Eq. 5.28:

$$\theta = \frac{\ln(p/\omega)}{N} \tag{5.28}$$

where N is the number of EDF iterations and the proportion of remaining variables for the i th EDF and θ is a constant that controls the EDF. When i is 0, all variables are included for modeling, only ω variables are retained at the N th time. Thus, θ can be computed through Eq. 5.25 and EDF function (Eq. 5.21).

Step (5): After removing a certain proportion of variables according to EDF, return to step (1) with the number of remaining variables until the number of EDF iterations has conducted completely. The variable space is shrunk continuously based on EDF process as shown in Fig. 5.16, which means that it can better select the important variables in the small and optimized variable space after all EDF iterations have been conducted completely.

Step (6): Determine the optimal variable subset by investigating all possible combinations through greedy algorithm. The larger the ω , the more combinations there

are. The preset of ω should be based on the computation ability. Generally, ω is set to 14. When $\omega = 14$, the number of all combinations is $2^{14}-1 = 16,383$.

5.10.4 Other Methods

In addition to the wavelength selection methods commonly used in spectral analysis introduced above, there are many variable selection methods for NIR spectral data. They include iterative prediction weighting [82], iterative reweighted partial least squares [83], Boruta algorithm based on random forest [84], ridge regression method based on least square L2 regularization [85], Lasso method based on L1 regularization [86], least angle regression method [87], regularization of PLS using the elastic net method with L1 and L2 regularizations [88], regularized PLS [89], and sparse PLS (SPLS) [90]. In the regularized PLS method, both L1 and L2 norm penalty regularization terms are introduced to generate sparsity of the model. The sparse solution of the principal component load coefficient is solved by an alternating iterative algorithm to achieve spectral data reduction and the selection of key wavelengths.

Machine learning algorithms tend to have the problem of the samples with high dimensionality, which makes it difficult to select key variables from numerous variables in the regression or classification. But if the built model with these variables is a sparse model, indicating only a few variables have made contribution to the model; in other words, most variables made no contribution or tiny contribution. At this time, the variable with the coefficient is non-zero and can be only focused, which is the variable selection method through sparse model.

5.10.5 Wavelength Selection Method Based on Hybrid Strategy

With the continuous development of variable selection methods, the joint use of various different algorithms has been paid more and more attention. By taking the advantage of the complementarity among different algorithms, these hybrid methods firstly select the wavelength interval or wavelength point roughly, and then finely and optimally select fewer and more effective variables. The prediction ability of the models built on this basis is usually better than that of the single variable selection method. Yu et al. reviewed variable selection methods based on hybrid strategy and systemically classify them into two categories such as two-step and three-step hybrid strategy [91].

Two-step hybrid methods are formed by combining two different methods. Most of them employed the methods in sequence. The later method makes a further optimization on the variable subset optimized by the former method as the above method

mentioned. For example, Yu et al. compared a variety of variable screening methods and found that the CARS-SPA method screened 37 characteristic wavelengths from 2001 wavelengths in the hyperspectral full band, and the PLS model established for soil organic matter content had the best effect [92]. Liu et al. selected wavelengths of siPLS-IRIV hybrid method for the identification of olive oil quality by NIR spectroscopy, and the result was superior to that of siPLS method alone [93]. Liang et al. used CARS-IRIV algorithm to screen the hyperspectral characteristic variables and established the LS-SVM model for predicting the soluble solid content of Korla perfumed pear, which simplified the operation of the model and improved the prediction accuracy [94]. Cai et al. used MC-UVE-SPA algorithm to extract 27 effective variables from 4254 variables in the original NIR spectrum, and built an analytical model for predicting strawberry soluble solid content by combining color features [95]. Wang et al. combined UVE-CARS to screen wavelengths of hyperspectral data, and predicted and visualized total flavonoids content in *Cerasus Humilis* fruit during storage periods [96]. Three-step hybrid methods employed three different variable selection algorithms to thoroughly select characteristic variables. Yu et al. applied iPLS, modified VCPA (mVCPA), VIP, GA, and IRIV to construct four three-step hybrid methods as iPLS-VIP-GA, iPLS-VIP-IRIV, iPLS-mVCPA-GA, and iPLS-mVCPA-IRIV [91] as shown in Fig. 5.17. In the first step, iPLS was used to select several informative wavelength intervals in rough way. In the second step, VIP and mVCPA were applied to make fine selection to further filter some unimportant variables and shrink variable space. In the third step, based on the variables retained in the

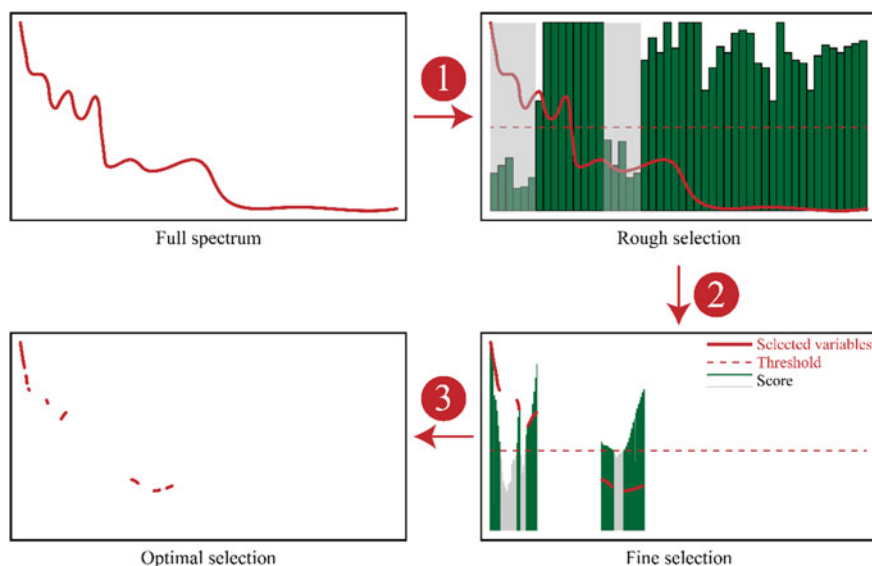


Fig. 5.17 Flow chart of the three-step hybrid strategy. Step 1: rough selection. Step 2: fine selection. Step 3: optimal selection [97]

first two steps, GA and IRIV were employed to make a further optimization to determine the optimal variable subset, which exploit the advantage of optimization ability of GA and IRIV with fewer variables. They have been applied successfully into two benchmark NIR data sets including beer and tobacco, and detecting the freshness of tilapia fillets. The results showed that the three-step hybrid methods present a better prediction performance than two-step hybrid methods and other single methods [97].

Besides, some hybrid methods were created by embedded the core idea of one method into another. For example, Yun et al. introduced the idea of MWPLS algorithm into random frog (RF) algorithm and proposed the interval random frog (iRF) method [98]. Yun et al. determined the structure of a proportion of chromosomes in the initial population of GA by PLS large regression coefficient (LRC) and proposed the LRC-GA-PLS method which can make the GA optimization better toward the optimal solution and efficiently screen key wavelengths [99].

In addition to the above two ways of hybrid method, another hybrid method first employs different variable selection methods to obtain different variable subsets, and the final optimal variable subset is determined based on different selected variable subsets using set operation in mathematics, such as intersection or union. Shen et al. also conducted intersection fusion of variables obtained from a variety of wavelength selection methods. As shown in Fig. 5.18, VIP method, Boruta algorithm, GA-RF algorithm, and GA-SVM algorithm were adopted to select characteristic wavelengths, respectively, and then the intersection of all selected wavelengths is

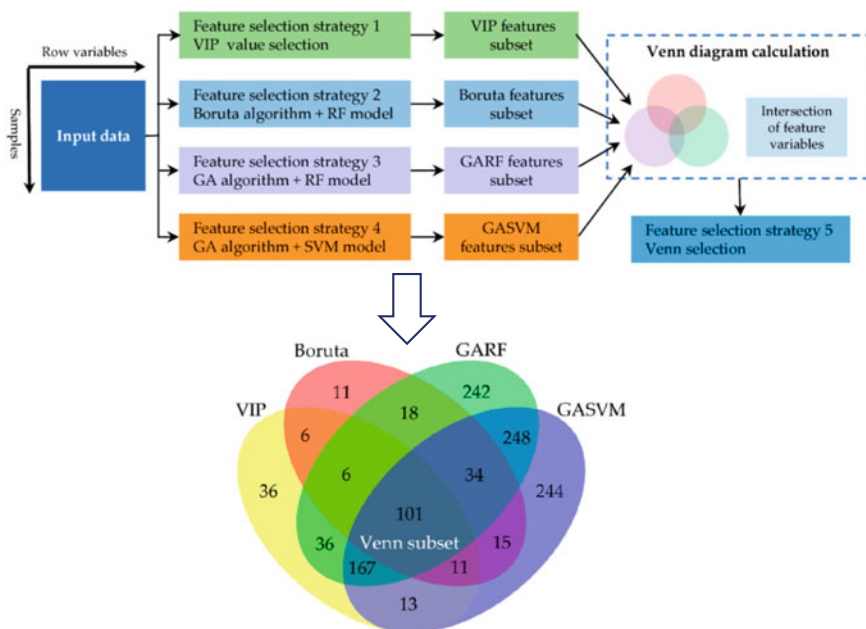


Fig. 5.18 Schematic diagram of wavelength selection by multiple algorithms combined with Venn diagram [100]

obtained through Venn graph [100]. Six variable selection methods including regression coefficient, Lasso, CARS, rPLS, sMC, and minimum redundancy maximum relevance (mRMR) were used by Song et al. to obtain six optimal variable subsets for LIBS data sets, respectively. The six variable subsets were sorted based on their RMSECV values, and the first three variable subsets with lower RMSECV value were fused by union operation to gain the final variable subset [109]. Although the number of this kind of hybrid methods is still relatively small, it deserves our attention as they can consider the good results of many methods to make a decision.

5.11 The Selection of Spectral Preprocessing and Wavelength Selection Methods

Spectral preprocessing and wavelength selection are the critical steps in the establishment of multivariate quantitative and qualitative models, directly determining the prediction ability and long-term reliability of the models. At present, there are dozens of spectral preprocessing and wavelength selection methods involved in literatures, and each method, such as wavelet transform, has different functions and parameters. Therefore, in the practical application, we will encounter the problem of how to select the optimal method and the optimal order of spectral preprocessing and wavelength selection methods. At the same time, we need to consider the influence of outlier samples and linear and nonlinear modeling methods.

Generally, the optimal spectral preprocessing method is not the same for different analytical systems and problems to be solved. But certain rules can be found. For example, derivative method is generally used for baseline correction, while multiplicative scatter correction (MSC), standard normal variate (SNV), and second derivative methods are used for diffuse reflection spectrum to eliminate light scattering caused by uneven particle distribution, and wavelet transform (WT) can effectively eliminate spectral background and improve the robustness of the model. If properly used, wavelength selection methods can always simplify the model and improve the ability of prediction. In the order of spectral preprocessing operation, baseline correction is usually carried out first, then noise is eliminated, and scattering correction and normalization are carried out in subsequence. However, in specific applications, some possible methods or their combinations still need to be compared to obtain the best results [101]. Diwu et al. also conducted the research on the influence of 120 combinations of ten preprocessing methods on the NIR spectral model, and the results showed that different sample sets had different optimal spectral preprocessing methods [102].

If the analytical system is relatively complex, only one spectral preprocessing method could not get better results. In this case, different spectral preprocessing and wavelength selection methods can be combined to obtain the expected results, but the combination of different spectral preprocessing methods and wavelength selection methods and their execution order still need to be optimized.

In the early stage, it is proposed in some literature that factor design method was adopted to solve the combination problem of preprocessing methods [103]. Gerretzen et al. also adopted design of experimental (DoE) method to integrate variable selection into the preprocessing selection approach to enhance the objective interpretation of built model [104, 105]. Zhao et al. used systematic tracking mapping to select the best combination of preprocessing methods, wavelength selection methods, and quantitative calibration methods at the same time (Fig. 5.19) [106]. Laxalde et al. used GA to optimize the selection of the combination of preprocessing method and wavelength selection and achieved good results (Fig. 5.20) [107]. Stefansson et al. also presented a fast method for performing GA-PLS to allow wavelength selection to be evolved containing variables from a mixture of different preprocessing techniques [108]. Based on the idea of system modeling, Gao et al. adopted D-optimal experimental design to optimize the modeling parameters such as spectral preprocessing method and wavelength selection method globally, thus improving the robustness and predictive ability of the model [109]. It is also an important development direction in the future to integrate the preprocessing and wavelength selection methods into the multivariate calibration step to form new calibration and preprocessing methods, rather than using them separately before calibration.

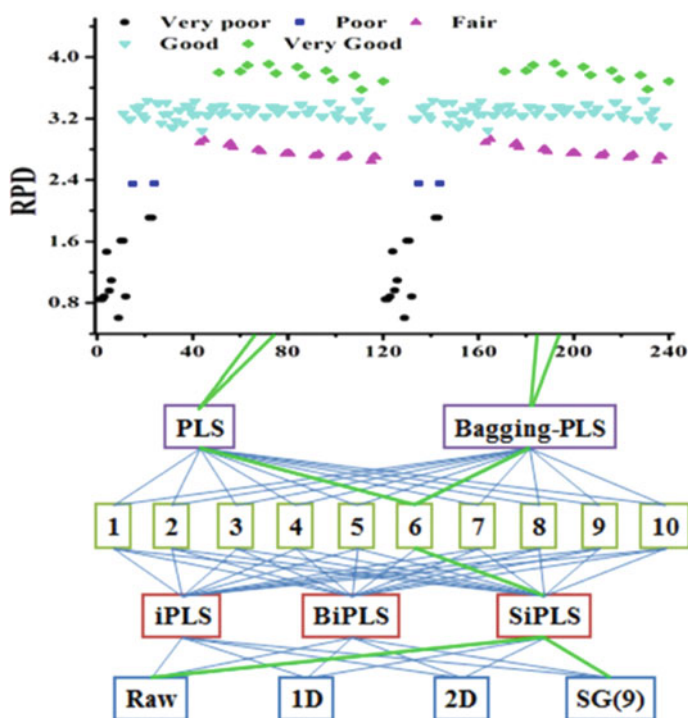


Fig. 5.19 Systematic tracking mapping was used to select the optimal model [106]

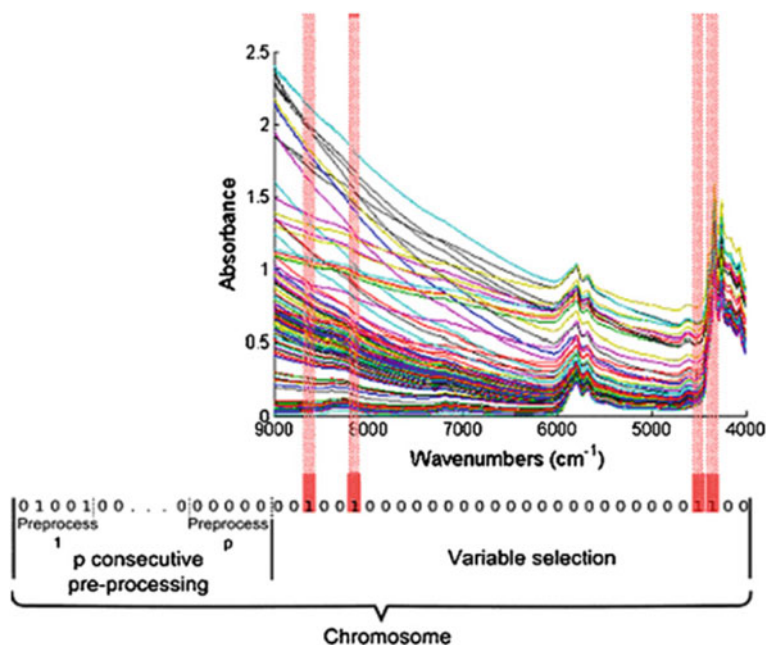


Fig. 5.20 Encoding the co-optimization problem by GA for spectral preprocessing and wavelength selection methods [107]

References

1. French AG, Jouan-Rimbaud D, Massart DL, et al. Wavelength selection method for multicomponent spectrophotometric determinations using partial least squares. *Analyst*. 1995;120:2787–92.
2. Song XZ, Tang G, Zhang LD, et al. Research advance of variable selection algorithms in near infrared spectroscopy analysis. *Spectrosc Spectr Anal*. 2017;37:1048–52.
3. Zhang J, Hu Y, Zhou L-X, et al. Progress of chemometric algorithms in near infrared spectroscopic analysis. *J Instrum Anal*. 2020;39:1196–203.
4. Yun Y-H, Li H-D, Deng B-C, et al. An overview of variable selection methods in multivariate analysis of near-infrared spectra. *TrAC Trends Anal Chem*. 2019;113:102–15.
5. Li H-D, Liang Y-Z, Cao D-S, et al. Model-population analysis and its applications in chemical and biological modeling. *TrAC Trends Anal Chem*. 2012;38:154–62.
6. Mehmood T, Liland KH, Snipen L, et al. A review of variable selection methods in Partial Least Squares Regression. *Chemom Intell Lab Syst*. 2012;118:62–9.
7. Cai W, Li Y, Shao X. A variable selection method based on uninformative variable elimination for multivariate calibration of near-infrared spectra. *Chemom Intell Lab Syst*. 2008;90:188–94.
8. Deng B-C, Yun Y-H, Cao D-S, et al. A bootstrapping soft shrinkage approach for variable selection in chemical modeling. *Anal Chim Acta*. 2016;908:63–74.
9. Yun Y-H, Wang W-T, Tan M-L, et al. A strategy that iteratively retains informative variables for selecting optimal variable subset in multivariate calibration. *Anal Chim Acta*. 2014;807:36–43.
10. Inoue Y, Sakaiya E, Zhu Y, et al. Diagnostic mapping of canopy nitrogen content in rice based on hyperspectral measurements. *Remote Sens Environ*. 2012;126:210–21.

11. Hong Y, Chen S, Zhang Y, et al. Rapid identification of soil organic matter level via visible and near-infrared spectroscopy: effects of two-dimensional correlation coefficient and extreme learning machine. *Sci Total Environ.* 2018;644:1232–43.
12. Windig W, Guilment J. Interactive self-modeling mixture analysis. *Anal Chem.* 1991;63:1425–32.
13. Moreira EDT, Pontes MJC, Galvão RKH, et al. Near infrared reflectance spectrometry classification of cigarettes using the successive projections algorithm for variable selection. *Talanta.* 2009;79:1260–4.
14. Hu B, Sun D-W, Pu H, et al. Rapid nondestructive detection of mixed pesticides residues on fruit surface using SERS combined with self-modeling mixture analysis method. *Talanta;* 2020(217), 120998.
15. Zhai C, Peng Y, Li Y, et al. Extraction and identification of mixed pesticides' Raman signal and establishment of their prediction models. *J Raman Spectrosc.* 2017;48:494–500.
16. Qin J, Chao K, Kim MS. Nondestructive evaluation of internal maturity of tomatoes using spatially offset Raman spectroscopy. *Postharvest Biol Technol.* 2012;71:21–31.
17. Khodabakhshian R. Feasibility of using Raman spectroscopy for detection of tannin changes in pomegranate fruits during maturity. *Sci Hortic.* 2019(257), 108670.
18. Araújo MCU, Saldanha TCB, Galvão RKH, et al. The successive projections algorithm for variable selection in spectroscopic multicomponent analysis. *Chemom Intell Lab Syst.* 2001;57:65–73.
19. Soares SFC, Gomes AA, Araujo MCU, et al. The successive projections algorithm. *TrAC Trends Anal Chem.* 2013;42:84–98.
20. Khanmohammadi M, Garmarudi AB, Ghasemi K, et al. Artificial neural network for quantitative determination of total protein in yogurt by infrared spectrometry. *Microchem J.* 2009;91:47–52.
21. Chen H, Tan C, Lin Z. Identification of ginseng according to geographical origin by near-infrared spectroscopy and pattern recognition. *Vib Spectrosc.* 2020(110), 103149.
22. Huang Y, Dong W, Sanaeifar A, et al. Development of simple identification models for four main catechins and caffeine in fresh green tea leaf based on visible and near-infrared spectroscopy. *Comput Electron Agric.* 2020(173), 105388.
23. Chong I-G, Jun C-H. Performance of some variable selection methods when multicollinearity is present. *Chemom Intell Lab Syst.* 2005;78:103–12.
24. He W-Q, Yan W-J, He G-Q, et al. Study on the wavelength selection based on VIP analysis in noninvasive measurement of blood components. *Spectrosc Spectr Anal.* 2016;36:1080–4.
25. Favilla S, Durante C, Vigni ML, et al. Assessing feature relevance in NPLS models by VIP. *Chemom Intell Lab Syst.* 2013;129:76–86.
26. Kvalheim OM. Interpretation of partial least squares regression models by means of target projection and selectivity ratio plots. *J Chemom.* 2010;24:496–504.
27. Tran TN, Afanador NL, Buydens LMC, et al. Interpretation of variable importance in Partial Least Squares with Significance Multivariate Correlation (sMC). *Chemom Intell Lab Syst.* 2014;138:153–60.
28. Nørgaard L, Saudland A, Wagner J, et al. Interval partial least-squares regression (iPLS): a comparative chemometric study with an example from near-infrared spectroscopy. *Appl Spectrosc.* 2000;54:413–9.
29. Leardi R, Nørgaard L. Sequential application of backward interval partial least squares and genetic algorithms for the selection of relevant spectral regions. *J Chemom.* 2004;18:486–97.
30. Xiaobo Z, Jiewen Z, Povey MJW, et al. Variables selection methods in near-infrared spectroscopy. *Anal Chim Acta.* 2010;667:14–32.
31. Zou X, Zhao J, Li Y. Selection of the efficient wavelength regions in FT-NIR spectroscopy for determination of SSC of 'Fuji' apple based on BiPLS and FiPLS models. *Vib Spectrosc.* 2007;44:220–7.
32. Jiang J-H, Berry RJ, Siesler HW, et al. Wavelength interval selection in multicomponent spectral analysis by moving window partial least-squares regression with applications to mid-infrared and near-infrared spectroscopic data. *Anal Chem.* 2002;74:3555–65.

33. Du YP, Liang YZ, Jiang JH, et al. Spectral regions selection to improve prediction ability of PLS models by changeable size moving window partial least squares and searching combination moving window partial least squares. *Anal Chim Acta*. 2004;501:183–91.
34. Kasemsumran S, Du YP, Maruo K, et al. Improvement of partial least squares models for in vitro and in vivo glucose quantifications by using near-infrared spectroscopy and searching combination moving window partial least squares. *Chemom Intell Lab Syst*. 2006;82:97–103.
35. Rinnan Å, Andersson M, Ridder C, et al. Recursive weighted partial least squares (rPLS): an efficient variable selection method using PLS. *J Chemom*. 2014;28:439–47.
36. Centner V, Massart D-L, de Noord OE, et al. Elimination of uninformative variables for multivariate calibration. *Anal Chem*. 1996;68:3851–8.
37. Lindgren F, Geladi P, Rännar S, et al. Interactive variable selection (IVS) for PLS. Part 1: theory and algorithms. *J Chemom*. 1994(8), 349–363.
38. Roque JV, Cardoso W, Peternelli LA, et al. Comprehensive new approaches for variable selection using ordered predictors selection. *Anal Chim Acta*. 2019;1075:57–70.
39. Niu X, Zhao Z, Jia K, et al. A feasibility study on quantitative analysis of glucose and fructose in lotus root powder by FT-NIR spectroscopy and chemometrics. *Food Chem*. 2012;133:592–7.
40. Deng B-C, Yun Y-H, Liang Y-Z. Model population analysis in chemometrics. *Chemom Intell Lab Syst*. 2015;149:166–76.
41. Han Q-J, Wu H-L, Cai C-B, et al. An ensemble of Monte Carlo uninformative variable elimination for wavelength selection. *Anal Chim Acta*. 2008;612:121–5.
42. Brezočnik L, Fister I, Podgorelec V. Swarm intelligence algorithms for feature selection: a review. *Appl Sci*. 2018;8:1521.
43. Bin J, Fan W, Zhou J-H, et al. Application of intelligent optimization algorithms to wavelength selection of near-infrared spectroscopy. *Spectrosc Spectr Anal*. 2017;37:95–102.
44. Jouan-Rimbaud D, Massart D-L, Leardi R, et al. Genetic algorithms as a tool for wavelength selection in multivariate calibration. *Anal Chem*. 1995;67:4295–301.
45. Leardi R. Application of genetic algorithm–PLS for feature selection in spectral data sets. *J Chemom*. 2000;14:643–55.
46. Kirkpatrick S, Gelatt CD, Vecchi MP. Optimization by simulated annealing. *Science*. 1983;220:671–80.
47. Shi J, Hu X, Zou X, et al. A heuristic and parallel simulated annealing algorithm for variable selection in near-infrared spectroscopy analysis. *J Chemom*. 2016;30:442–50.
48. Kennedy J, Eberhart R. Particle swarm optimization. In: *Proceedings of ICNN'95-international conference on neural networks*; 1995, p. 1942–8.
49. Tao Q-B, Shen Q, Zhang X-Y, et al. Simultaneous determination of multicomponent by wavelength selection using particle swarm optimization algorithm. *Chin J Anal Chem*. 2009;37:1197–200.
50. Cao H, Wang Y, Yang S, et al. A wavelength selection method based on random decision particle swarm optimization with attractor for near-infrared spectral quantitative analysis. *J Chemom*. 2015;29:289–99.
51. Wang J, Wang C, Zhu X, et al. Application of soft sensor in welding seam tracking prediction based on LSSVM and PSO with compression factor. In: *2013 25th Chinese control and decision conference (CCDC)*; 2013, p. 2441–6.
52. Zhang P, Xu Z, Wang Q, et al. A novel variable selection method based on combined moving window and intelligent optimization algorithm for variable selection in chemical modeling. *Spectrochim Acta Part A Mol Biomol Spectrosc*. 2021(246), 118986.
53. Ma Q, Lei X, Zhang Q. Mobile robot path planning with complex constraints based on the second-order oscillating particle swarm optimization algorithm. In: *2009 WRI world congress on computer science and information engineering*; 2009, p. 244–8.
54. Dorigo M, Birattari M, Stutzle T. Ant colony optimization. *IEEE Comput Intell Mag*. 2006;1:28–39.
55. Shamsipur M, Zare-Shahabadi V, Hemmateenejad B, et al. Ant colony optimisation: a powerful tool for wavelength selection. *J Chemom*. 2006;20:146–57.

56. Shen Q, Jiang J-H, Tao J-C, et al. Modified ant colony optimization algorithm for variable selection in QSAR modeling: QSAR studies of cyclooxygenase inhibitors. *J Chem Inf Model.* 2005(45), 1024–9.
57. Shamsipur M, Zare-Shahabadi V, Hemmateenejad B, et al. An efficient variable selection method based on the use of external memory in ant colony optimization. Application to QSAR/QSPR studies. *Anal Chim Acta.* 2009(646), 39–46
58. Goodarzi M, Freitas MP, Jensen R. Ant colony optimization as a feature selection method in the QSAR modeling of anti-HIV-1 activities of 3-(3,5-dimethylbenzyl)uracil derivatives using MLR, PLS and SVM regressions. *Chemom Intell Lab Syst.* 2009;98:123–9.
59. Hu L, Yin C, Ma S, et al. Rapid detection of three quality parameters and classification of wine based on Vis-NIR spectroscopy with wavelength selection by ACO and CARS algorithms. *Spectrochim Acta Part A Mol Biomol Spectrosc.* 2018;205:574–81.
60. Allegrini F, Olivieri AC. A new and efficient variable selection algorithm based on ant colony optimization. Applications to near infrared spectroscopy/partial least-squares analysis. *Anal Chim Acta.* 2011(699), 18–25.
61. Xiaowei H, Xiaobo Z, Jiewen Z, et al. Measurement of total anthocyanins content in flowering tea using near infrared spectroscopy combined with ant colony optimization models. *Food Chem.* 2014;164:536–43.
62. Zhang Y, Li M, Zheng L, et al. Spectral features extraction for estimation of soil total nitrogen content based on modified ant colony optimization algorithm. *Geoderma.* 2019;333:23–34.
63. Fallahzadeh O, Dehghani-Bidgoli Z, Assarian M. Raman spectral feature selection using ant colony optimization for breast cancer diagnosis. *Lasers Med Sci.* 2018;33:1799–806.
64. Guo Z, Wang M, Wu J, et al. Quantitative assessment of zearalenone in maize using multivariate algorithms coupled to Raman spectroscopy. *Food Chem.* 2019;286:282–8.
65. Ranzan L, Trierweiler LF, Trierweiler JO. Prediction of sulfur content in diesel fuel using fluorescence spectroscopy and a hybrid ant colony-Tabu Search algorithm with polynomial bases expansion. *Chemom Intell Lab Syst.* 2020(206), 104161.
66. Yun Y-H, Wu D-M, Li G-Y, et al. A strategy on the definition of applicability domain of model based on population analysis. *Chemom Intell Lab Syst.* 2017;170:77–83.
67. Li H, Liang Y, Xu Q, et al. Key wavelengths screening using competitive adaptive reweighted sampling method for multivariate calibration. *Anal Chim Acta.* 2009;648:77–84.
68. Yun Y-H, Wang W-T, Deng B-C, et al. Using variable combination population analysis for variable selection in multivariate calibration. *Anal Chim Acta.* 2015;862:14–23.
69. Geng J, Yang C, Luo Q, et al. iPCPA: interval permutation combination population analysis for spectral wavelength selection. *Anal Chim Acta.* 2021(1171), 338635.
70. Li H-D, Xu Q-S, Liang Y-Z. Random frog: an efficient reversible jump Markov Chain Monte Carlo-like approach for variable selection with applications to gene selection and disease classification. *Anal Chim Acta.* 2012;740:20–6.
71. Chen J, Yang C, Zhu H, et al. A novel variable selection method based on stability and variable permutation for multivariate calibration. *Chemom Intell Lab Syst.* 2018;182:188–201.
72. Wang W-T, Yun Y-H, Deng B-C, et al. Iteratively variable subset optimization for multivariate calibration. *RSC Adv.* 2015;5:95771–80.
73. Bin J, Ai F, Fan W, et al. An efficient variable selection method based on variable permutation and model population analysis for multivariate calibration of NIR spectra. *Chemom Intell Lab Syst.* 2016;158:1–13.
74. Deng B-C, Yun Y-H, Liang Y-Z, et al. A novel variable selection approach that iteratively optimizes variable space using weighted binary matrix sampling. *Analyst.* 2014;139:4836–45.
75. Xu H, Liu Z, Cai W, et al. A wavelength selection method based on randomization test for near-infrared spectral analysis. *Chemom Intell Lab Syst.* 2009;97:189–93.
76. Zhang R, Zhang F, Chen W, et al. A new strategy of least absolute shrinkage and selection operator coupled with sampling error profile analysis for wavelength selection. *Chemom Intell Lab Syst.* 2018;175:47–54.
77. Zhang R, Zhang F, Chen W, et al. A variable informative criterion based on weighted voting strategy combined with LASSO for variable selection in multivariate calibration. *Chemom Intell Lab Syst.* 2019;184:132–41.

78. Chen H, Tan C, Lin Z, et al. Quantifying several adulterants of notoginseng powder by near-infrared spectroscopy and multivariate calibration. *Spectrochim Acta Part A Mol Biomol Spectrosc.* 2019;211:280–6.
79. Jiang H, Xu W, Ding Y, et al. Quantitative analysis of yeast fermentation process using Raman spectroscopy: Comparison of CARS and VCPA for variable selection. *Spectrochim Acta Part A Mol Biomol Spectrosc.* 2020(228), 117781.
80. Wu D, Meng L, Yang L, et al. Feasibility of laser-induced breakdown spectroscopy and hyperspectral imaging for rapid detection of thiophanate-methyl residue on mulberry fruit. *Int J Mol Sci.* 2019;20:1–14.
81. Xu D, Fan W, Lv H, et al. Simultaneous determination of traces amounts of cadmium, zinc, and cobalt based on UV–Vis spectrometry combined with wavelength selection and partial least squares regression. *Spectrochim Acta Part A Mol Biomol Spectrosc.* 2014;123:430–5.
82. Forina M, Casolino C, Pizarro MC. Iterative predictor weighting (IPW) PLS: a technique for the elimination of useless predictors in regression problems. *J Chemom.* 1999;13:165–84.
83. Cummins DJ, Andrews CW. Iteratively reweighted partial least squares: a performance analysis by monte carlo simulation. *J Chemom.* 1995;9:489–507.
84. Jaiswal JK, Samikannu R. Application of random forest algorithm on feature subset selection and classification and regression. In: 2017 world congress on computing and communication technologies (WCCCT); 2017, p. 65–8.
85. Zhang M, Liu X-H, He X-K, et al. Study on the application of ridge regression to near-infrared spectroscopy quantitative analysis and optimum wavelength selection. *Spectrosc Spectr Anal.* 2010;30:1214–7.
86. Mei C-L, Chen Y, Yin L, et al. Wavelength selection by siPLS-LASSO for NIR spectroscopy and its application. *Spectrosc Spectr Anal.* 2018;38:436–40.
87. Yan S-K, Yang H-H, Hu B-C, et al. Variable selection method of NIR spectroscopy based on least angle regression and GA-PLS. *Spectrosc Spectr Anal.* 2017;37:1733–8.
88. Huang X, Luo Y-P, Xu Q-S, et al. Elastic net wavelength interval selection based on iterative rank PLS regression coefficient screening. *Anal Methods.* 2017;9:672–9.
89. Allen GI, Peterson C, Vannucci M, et al. Regularized partial least squares with an application to NMR spectroscopy. *Stat Anal Data Mining ASA Data Sci J.* 2013;6:302–14.
90. Chun H, Keleş S. Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *J R Stat Soc Ser B (Stat Methodol).* 2010;72:3–25.
91. Yu H-D, Yun Y-H, Zhang W, et al. Three-step hybrid strategy towards efficiently selecting variables in multivariate calibration of near-infrared spectra. *Spectrochim Acta Part A Mol Biomol Spectrosc.* 2020(224), 117376.
92. Yu L, Yongsheng H, Zhou Y, et al. Wavelength variable selection methods for estimation of soil organic matter content using hyperspectral technique. *Trans Chin Soc Agric Eng.* 2016;13:95–102.
93. Liu G-H, Han W-Q, Jiang H. Study on quality identification of olive oil based on near infrared spectra. *Spectrosc Spectr Anal.* 2016;36:2798–801.
94. Liang K, Liu Q, Pan L, et al. Detection of soluble solids content in “Korla fragrant pear” based on hyperspectral imaging and CARS-IRIV algorithm. *J Nanjing Agric Univ.* 2018;41:760–6.
95. Cai D, Tang C, Liang Y, et al. Establishment of quantitative analysis model for detecting the soluble solids content in strawberry by merging near infrared spectroscopy and color parameters. *Food Ferment Ind.* 2020;46:218–24.
96. Wang B, He J, Zhang S, et al. Nondestructive prediction and visualization of total flavonoids content in *Cerasus Humilis* fruit during storage periods based on hyperspectral imaging technique. *J Food Process Eng.* 2021(44), e13807.
97. Yu H-D, Zuo S-M, Xia G, et al. Rapid and nondestructive freshness determination of tilapia fillets by a portable near-infrared spectrometer combined with chemometrics methods. *Food Anal Methods.* 2020;13:1918–28.
98. Yun Y-H, Li H-D, E. Wood L R, et al. An efficient method of wavelength interval selection based on random frog for multivariate spectral calibration. *Spectrochim Acta Part A Mol Biomol Spectrosc.* 2013(111), 31–6.

99. Yun Y-H, Cao D-S, Tan M-L, et al. A simple idea on applying large regression coefficient to improve the genetic algorithm-PLS for variable selection in multivariate calibration. *Chemom Intell Lab Syst.* 2014;130:76–83.
100. Shen T, Yu H, Wang Y-Z. Discrimination of *Gentiana* and its related species using IR spectroscopy combined with feature selection and stacked generalization. *Molecules.* 2020;25:1442.
101. Lee LC, Liong C-Y, Jemain AA. A contemporary review on Data Preprocessing (DP) practice strategy in ATR-FTIR spectrum. *Chemom Intell Lab Syst.* 2017;163:64–75.
102. Diwu P-Y, Bian X-H, Wang Z-F, et al. Study on the selection of spectral preprocessing methods. *Spectrosc Spectr Anal.* 2019;39:2800–6.
103. Olsson RJO. Optimizing data-pretreatment by a factorial design approach. *Near InfraRed Spectrosc.* 1992, 103–7.
104. Gerretzen J, Szymańska E, Jansen JJ, et al. Simple and effective way for data preprocessing selection based on design of experiments. *Anal Chem.* 2015;87:12096–103.
105. Gerretzen J, Szymańska E, Bart J, et al. Boosting model performance and interpretation by entangling preprocessing selection and variable selection. *Anal Chim Acta.* 2016;938:44–52.
106. Zhao N, Ma L, Huang X, et al. Pharmaceutical analysis model robustness from bagging-PLS and PLS using systematic tracking mapping. *Front Chem.* 2018;6:1–7.
107. Laxalde J, Ruckebusch C, Devos O, et al. Characterisation of heavy oils using near-infrared spectroscopy: Optimisation of pre-processing methods and variable selection. *Anal Chim Acta.* 2011;705:227–34.
108. Stefansson P, Liland KH, Thiis T, et al. Fast method for GA-PLS with simultaneous feature selection and identification of optimal preprocessing technique for datasets with many observations. *J Chemom.* 2020(34), e3195.
109. Gao R-L, Yang P-S, Xu G, et al. Study on establishment of near-infrared quantitative model for salivianolic acid b in naoxintong capsule based on the system modeling idea. *Spectrosc Spectr Anal.* 2020;40:3573–8.

Chapter 6

Spectral Dimensionality Reduction Methods



6.1 The Multicollinearity Problem

The premise of multiple linear regression (MLR) is that the independent variables must be independent from each other. However, there is often a certain degree of correlation between the spectral variables, resulting in statistical multicollinearity. Multicollinearity means the high correlation among the independent variables in the linear regression model. The value of the regression coefficient obtained $\hat{\mathbf{b}}$ is unstable and difficult to interpret due to the existence of highly correlated relationships among the independent variables. The regression coefficients may become very sensitive to small changes in the sample data, making the values of the regression coefficients difficult to estimate precisely. There may be even a phenomenon that the positive and negative signs of the regression coefficients are opposite to those of theoretical research or experience [1].

For example, if $\mathbf{X} = \begin{bmatrix} 1 & 2 \\ 1 & 2.00001 \end{bmatrix}$, $\mathbf{y} = \begin{bmatrix} 3 \\ 3.00001 \end{bmatrix}$, then $\hat{\mathbf{b}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$. But if there are test errors, $\mathbf{y} = \begin{bmatrix} 3.00011 \\ 2.99990 \end{bmatrix}$, $\hat{\mathbf{b}} = \begin{bmatrix} 44.9985 \\ -20.9992 \end{bmatrix}$.

It can be seen that when \mathbf{y} value changes slightly, the regression coefficient changes greatly, even the positive and negative sign changes. The reasons for the above results are that matrix \mathbf{X} has serious collinearity, i.e., \mathbf{X} is an ill-conditioned matrix, and serious errors will be occurred when the inverse of $\mathbf{X}^T \mathbf{X}$ is obtained.

However, if $\mathbf{X} = \begin{bmatrix} 1 & 2 \\ 3 & 0.00001 \end{bmatrix}$, \mathbf{y} is equal to the $\begin{bmatrix} 3 \\ 3.00001 \end{bmatrix}$ and $\begin{bmatrix} 3.00011 \\ 2.99990 \end{bmatrix}$, respectively. The value of the regression coefficient $\hat{\mathbf{b}}$, $\begin{bmatrix} 1 \\ 1 \end{bmatrix}$ and $\begin{bmatrix} 0.99996 \\ 1.00007 \end{bmatrix}$ can be obtained, respectively. It can be seen that if the variables of the \mathbf{X} matrix are independent from each other, the regression coefficient calculated by MLR is robust.

At present, there are several commonly used multicollinearity diagnostic methods.

- (1) Correlation coefficient diagnostic method of independent variables: calculate the pairwise correlation coefficient of variables. If the value of correlation coefficient between independent variables is very large, it indicates that there is a strong linear relationship between the corresponding two independent variables. However, it is limited to the linear correlation between two variables, and it is invalid for the collinearity among multiple variables.
- (2) Variance inflation factor (VIF) can be diagnostics by Eq. 6.1.

$$VIF_i = \frac{1}{1-R_i^2}, \quad (i = 1, 2, 3, \dots k) \quad (6.1)$$

where k is the number of variables, R_i^2 is the coefficient of determination obtained by taking the i th independent variables as the dependent variables and using the remaining $(k-1)$ variables as the multiple linear regression. The closer R_i^2 is to 1, the larger the VIF_i is, which indicates that the collinearity between the i th variable and other independent variables are stronger. It can be used to diagnose the extent to which variable is affected by multicollinearity.

- (3) Conditional number diagnostic method: Singular value decomposition is performed on the \mathbf{X} matrix, and the ratio of the maximum and minimum singular values is calculated, namely, the conditional value. The range of the conditional value is $1 \sim \infty$, and the larger the conditional value is, the greater the possibility of the existence of collinearity is. For example, $\mathbf{X} = \begin{bmatrix} 1 & 2 \\ 1 & 2.00001 \end{bmatrix}$

the conditional value is 1×10^6 , $\mathbf{X} = \begin{bmatrix} 1 & 2 \\ 3 & 0.00001 \end{bmatrix}$, the conditional value is 1.77.

Taking pattern recognition classification as an example, Hughes et al. gave the relationship among the complexity of measurement data, average recognition accuracy, and the number of calibration samples [2]. The measurement data complexity here refers to the degree of detail of the data acquired by the measuring device, namely, the dimensions of the feature data (the number of wavelength points in the spectra). As shown in Fig. 6.1, with the constant increase of the dimension of feature data, if the number of calibration samples are small and cannot meet the requirement of the dimension increase of feature space, the higher-dimensional features will cause the classification accuracy to increase first and then decrease, which is called the Hughes phenomenon. Therefore, for finite samples in practical applications, there is an optimal dimension of feature data to achieve the optimal classification accuracy. Therefore, dimensionality reduction of spectral data is also an effective method to reduce the Hughes phenomenon.

As shown in Fig. 6.2, the methods for dimensionality reduction of spectral data mainly include feature selection and feature extraction. As shown in Fig. 6.3, feature selection is to select a feature subset from the feature set. Feature selection does not change the properties of the original feature space, just selects some important

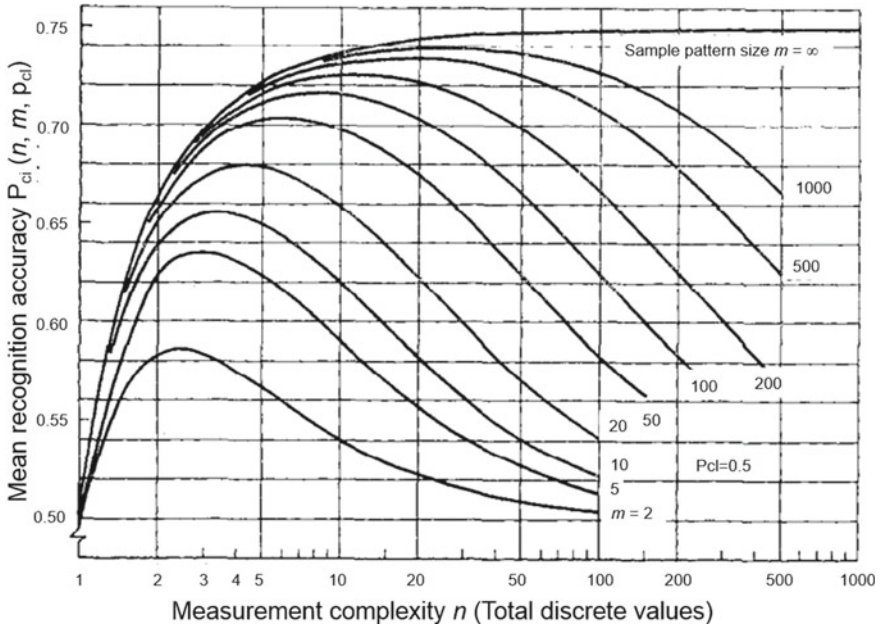


Fig. 6.1 The accuracy of classification for finite data set [2]

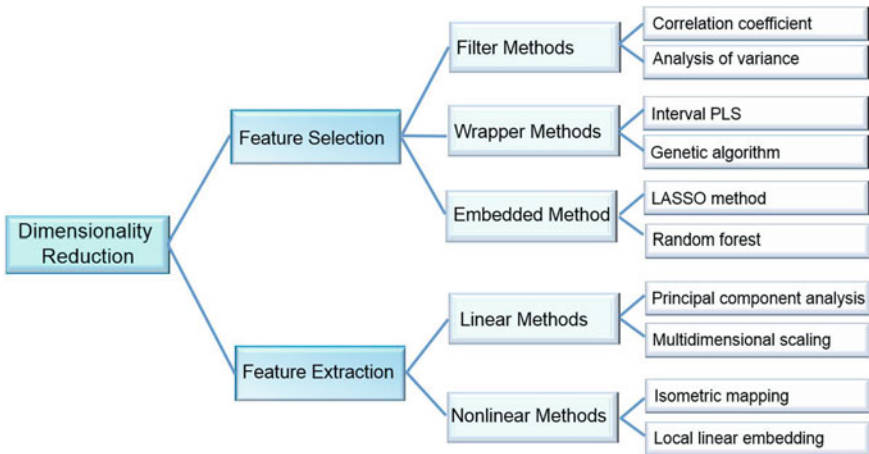


Fig. 6.2 Classification diagram of the realization methods of spectral data dimensionality reduction

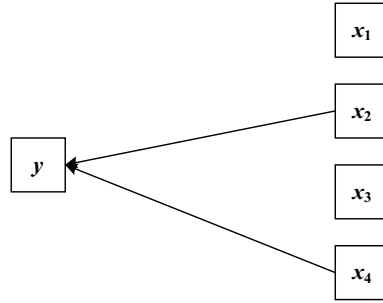


Fig. 6.3 Schematic diagram of feature selection

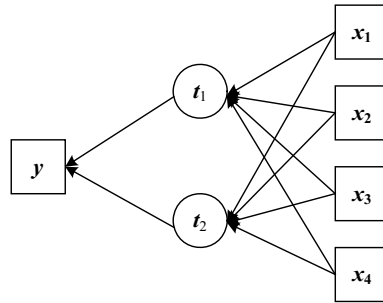


Fig. 6.4 Schematic diagram of feature extraction

features from the original space to form a new low-dimensional space. The commonly used feature selection methods (wavelength variable selection methods) are introduced in Chap. 5. As shown in Fig. 6.4, feature extraction (feature transformation) refers to the transformation of the original feature space to generate a new feature space with lower dimension and each dimension is independent of each other. This chapter mainly introduces the method of spectral feature extraction.

Feature extraction is divided into linear and nonlinear methods. Linear methods include principal component analysis (PCA), independent component analysis (ICA) and multi-dimensional scaling (MDS). The nonlinear methods include isometric mapping (ISOMAP), local linear embedding (LLE), and t-distributed stochastic neighborhood embedding (T-SNE). At present, most of these nonlinear methods are proposed based on the strategy of manifold learning, which is a hot spot in pattern recognition and machine learning research. It can reduce the dimension of high-dimensional data space nonlinearly, reveal its manifold distribution, and find out the specific low-dimensional structure hidden in the high-dimensional spectral data. Manifold learning has been widely used for dimensionality reduction and feature extraction of spectral data in recent years.

6.2 Principal Component Analysis

6.2.1 Theory of Principal Component Analysis

Principal component analysis (PCA) plays an important role in chemometrics. In fact, PCA is a very traditional technique of multivariate statistical analysis, first proposed by Hotelling in 1933.

The key purpose of PCA is to reduce the dimension of data and transform the original variables, so that a few new variables are linear combinations of the original variables. At the same time, these variables should express the data characteristics of the original variables as much as possible without losing of information [3, 4]. PCA transforms the data into a new coordinate system (see Fig. 6.5) such that the maximum variance of any data projection is at the first principal component (PC1), the second maximum variance in the second coordinate (PC2), and so on. The new variables obtained by the transformation are orthogonal to each other and unrelated to each other, which eliminate the overlapping parts among many coexisting information, that is, the possible multicollinearity among variables is eliminated.

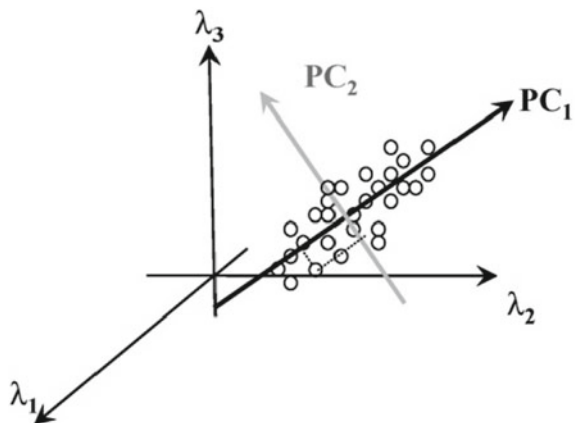
PCA decomposes the spectral matrix \mathbf{X} ($n \times m$) into the sum of the cross products of m vectors, that is:

$$\mathbf{X} = \mathbf{t}_1 \mathbf{p}_1^T + \mathbf{t}_2 \mathbf{p}_2^T + \mathbf{t}_3 \mathbf{p}_3^T + \dots + \mathbf{t}_m \mathbf{p}_m^T \quad (6.2)$$

where \mathbf{t} is called the score vector and \mathbf{P} is called the loading vector, or principal component (PC). It can also be written in the following matrix form: $\mathbf{X} = \mathbf{T} \mathbf{P}^T$, where $\mathbf{T} = [\mathbf{t}_1 \mathbf{t}_2 \dots \mathbf{t}_n]$ is called the score matrix and $\mathbf{P} = [\mathbf{p}_1 \mathbf{p}_2 \dots \mathbf{p}_m]$ is called the loading matrix, as shown in Fig. 6.6.

Each score vector is orthogonal to each other, that is, for any i and j , when $i \neq j$, $\mathbf{t}_i^T \mathbf{t}_j = 0$. Each loading vector is also orthogonal, and the length of each loading vector

Fig. 6.5 Schematic diagram of principal component analysis



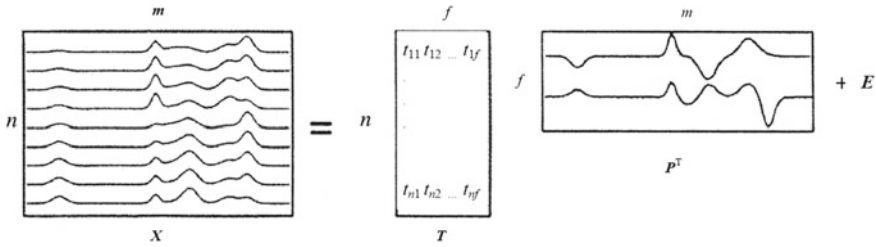


Fig. 6.6 Schematic diagram of matrix decomposition by PCA

is 1, that is, $\mathbf{p}_i^T \mathbf{p}_j = 0, i \neq j; \mathbf{p}_i^T \mathbf{p}_i = 1, i = j$. From the above vector properties, it is not difficult to get: $\mathbf{t}_i = \mathbf{X} \mathbf{p}_i$. This illustrates the mathematical significance of PCA, namely, each score vector is actually a projection of the matrix \mathbf{X} in the direction of its corresponding loading vector \mathbf{p} . The length of vector \mathbf{t}_i reflects the coverage degree of matrix \mathbf{X} in the direction of \mathbf{p}_i and reflects the relationship among samples. The greater its length is, the greater the coverage or variation range of \mathbf{X} in the direction of \mathbf{p}_i is.

As shown in Fig. 6.5, loading vector \mathbf{p}_1 represents the direction in which matrix \mathbf{X} has the greatest change (variance), \mathbf{p}_2 is perpendicular to \mathbf{p}_1 , and represents the second largest change direction in \mathbf{X} , and \mathbf{p}_m represents the smallest change direction in \mathbf{X} . From the point of probability statistics, the greater the variance of a random variable, the more information it contains; if the variance of a variable is zero, the variable is a constant and does not contain any information. When there is a certain degree of linear correlation among the variables in the matrix \mathbf{X} , the change of \mathbf{X} will be mainly reflected in the directions of the first few loading vectors, and the projection of \mathbf{X} on the last few loading vectors is very small, so it can be considered that they are mainly caused by measurement noise.

In this way, the PCA decomposition of matrix \mathbf{X} can be written as follows:

$$\mathbf{X} = \mathbf{t}_1 \mathbf{p}_1^T + \mathbf{t}_2 \mathbf{p}_2^T + \mathbf{t}_3 \mathbf{p}_3^T + \dots + \mathbf{t}_f \mathbf{p}_f^T + \mathbf{E} \tag{6.3}$$

where \mathbf{E} is the error matrix and represents the change of \mathbf{X} in the direction of the loading vector from \mathbf{p}_f to \mathbf{p}_m . Since the error matrix \mathbf{E} is mainly caused by measurement noise, ignoring \mathbf{E} will not cause significant loss of large amount of information in the data, and will also play the effect of removing noise. In practical applications, the number of principal components (PCs) f is often much smaller than m , so as to serve the purpose of data compression and feature extraction.

It can be proved that the PCA of \mathbf{X} is actually equivalent to the eigenvector analysis of covariance matrix $\mathbf{X}^T \mathbf{X}$ of \mathbf{X} . The loading vectors of matrix \mathbf{X} are actually the eigenvectors of matrix $\mathbf{X}^T \mathbf{X}$. If the eigenvalues of the $\mathbf{X}^T \mathbf{X}$ are arranged as follows: $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_m$, then the eigenvectors $\mathbf{p}_1, \mathbf{p}_2, \dots, \mathbf{p}_m$ are the loading vectors of the matrix \mathbf{X} .

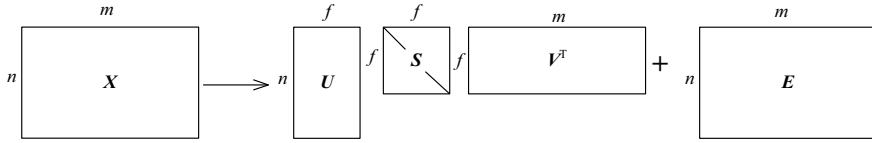


Fig. 6.7 Schematic diagram of matrix decomposition by SVD

PCA and singular value decomposition (SVD) are closely related to each other. The SVD can decompose the real number matrix of any order into the product of three matrices (Fig. 6.7), i.e.,

$$X = USV^T \tag{6.4}$$

where S is a diagonal matrix and collects the singular values of matrix X . In fact, it is the square root of the eigenvalues of covariance matrix $X^T X$. U and V^T are standard column orthogonality and standard orthogonal matrix, respectively. Column and row eigenvectors corresponding to these eigenvalues are collected. In fact, the product of matrix U and matrix S is equal to the score matrix T in the PCA and matrix V is equal to the loading matrix P .

The PCA of the spectral matrix X can be interpreted as follows, the loading vector P can be understood as the normalized spectra of the “pure component” extracted from the spectra of the mixture system, and the corresponding score vector t can be understood as the weight of the “pure component” in different samples, namely, the concentration. That is to say, the original spectra of the samples can be reconstructed by multiplying these “pure components” by their corresponding weights and summing them up, which is consistent with the Lambert-Beer law and the principle of additive property in spectral analysis.

6.2.2 Determination of Principal Component Number

The sum of the first f eigenvalues of the covariance matrix of $\sum_{i=1}^f \lambda_i$ is divided by the sum of all its eigenvalues $\sum_{i=1}^{\min(n,m)} \lambda_i$, which is called the cumulative contribution rate of the first f PCs, and represents the proportion of the data changes explained by the first f PCs in the total data changes. The number of selected PCs depends on the cumulative variance contribution ratio of the PCs. Generally, the number of PCs required to make the cumulative variance contribution rate greater than 85–95% can represent most of the information provided by the original variable. The eigenvalues can be plotted against each PC to select the number of PCs, as shown in Fig. 6.8. The number of PCs f should be selected as 5.

The PC number of the X matrix can also be determined by the indicated function method (IND), which is defined as follows:

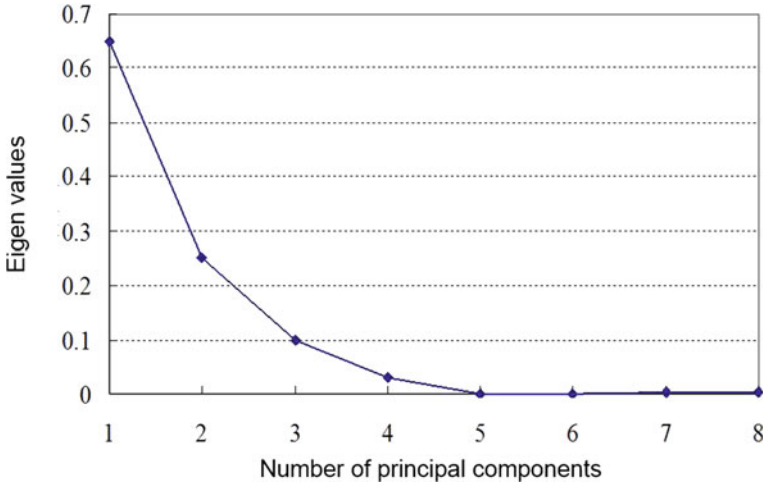


Fig. 6.8 Variation of eigenvalue with the number of principal components

$$IND = \sqrt{\frac{\sum_{i=f+1}^{\min(n,m)} \lambda_i}{\max(n, m)[\min(n, m) - f]^5}} \tag{6.5}$$

Starting from $f = 1$, calculate the IND value corresponding to different f . The IND value decreases gradually with the increase of f , and then increases again, so there is a minimum value. The f corresponding to the minimum value of the IND function is the number of PCs.

For spectral analysis, sometimes the loading matrix P can be used to help determine the PC. When the loading vector corresponding to a component number obviously shows a noise trend, it indicates that it is near the number of PCs of the spectral matrix.

6.2.3 Algorithm of Principal Component Analysis

In practice, the nonlinear iterative partial least squares (NIPALS) algorithm proposed by H Wold in 1966 is often used to calculate PCA. NIPALS algorithm is more suitable for microcomputer calculation, because it is an iterative algorithm with simple calculation steps and fast speed. In the calculation, PCs with large variance or large eigenvalue are given first, rather than all factors are calculated at once. The process of NIPALS algorithm is as follows.

- (1) Take a vector \mathbf{x} in \mathbf{X} as the starting value of \mathbf{t} : $\mathbf{t} = \mathbf{x}$.
- (2) Calculate \mathbf{p}^T

$$\mathbf{p}^T = \mathbf{t}^T \mathbf{X} / \mathbf{t}^T \mathbf{t} \quad (6.6)$$

(3) Normalize \mathbf{p}^T

$$\mathbf{p}^T = \mathbf{p}^T / \|\mathbf{p}\| \quad (6.7)$$

(4) Calculate \mathbf{t}

$$\mathbf{t} = \mathbf{X} \mathbf{p} / \mathbf{p}^T \mathbf{p} \quad (6.8)$$

- (5) Compare the new and the old \mathbf{t} , and see if the convergence condition is satisfied. If the convergence condition is met, proceed to step (6), otherwise jump back to step (2).
- (6) If the required PCs have been completed, the calculation will be stopped. Otherwise, calculate the residual matrix \mathbf{E}

$$\mathbf{E} = \mathbf{X} - \mathbf{t} \mathbf{p}^T \quad (6.9)$$

(7) Replace \mathbf{X} with \mathbf{E} , go back to step (1), and find the next PC.

After NIPALS calculation, \mathbf{X} is transformed into an orthogonal PC matrix \mathbf{T} . It can be proved that the eigenvector \mathbf{p} obtained by NIPALS algorithm is the eigenvector of matrix $\mathbf{X}^T \mathbf{X}$. The NIPALS algorithm has been widely used in chemometrics because of its high speed, simple steps, and easy application on computers.

For the spectrum of unknown sample \mathbf{x}_{un} , the score vector \mathbf{t}_{un} of the spectrum of the unknown sample can be calculated through the loading matrix \mathbf{P} obtained above

$$\mathbf{t}_{un} = \mathbf{x}_{un} \mathbf{P} \quad (6.10)$$

6.2.4 Application of Principal Component Analysis

The purpose of calculating PCs is to project the high-dimensional data into a lower-dimensional space. At the same time, these new variables can reflect the information of the original variables as much as possible and are independent of each other.

For PCA model,

$$\mathbf{X} = \mathbf{T} \mathbf{P}^T + \mathbf{E} \quad (6.11)$$

Score matrix \mathbf{T} can be used as characteristic variable for quantitative analysis, such as input variable of MLR, namely, principal component regression (PCR), input variable of artificial neural network (ANN), support vector regression (SVR), etc. [5]. The score matrix T is also often used in qualitative analysis, such as the characteristic

variable to calculate the Mahalanobis distance among samples to judge the outlier samples. In fact, the PC score vector can be directly plotted in two or three dimensions, and the classification of different samples can be realized through graphics display of computer screen. In addition, the spectral residual matrix \mathbf{E} can also be used for qualitative analysis (such as SIMCA method, the identification of spectral residual outlier samples, etc.).

The Hotelling T^2 statistics and Q statistics in the multivariate statistical process control (MSPC) are calculated based on the score matrix \mathbf{T} and the residual matrix \mathbf{E} , respectively [6, 7].

6.2.5 Multivariate Resolution Alternating Least Squares

Multivariate curve resolution-alternating least squares (MCR-ALS) is a bilinear-based spectral matrix decomposition method. It uses alternating method to iterate. The concentration and spectral distribution curves of the pure components in the complex system can be obtained [8–10].

The expression of decomposition of spectral matrix \mathbf{D} in MCR-ALS model is similar to that of PCA. In essence, the score matrix is further analyzed in PC space, so as to estimate the pure spectra. The model of MCR-ALS is as follows:

$$\mathbf{D} = \mathbf{C}\mathbf{S}^T + \mathbf{E} \quad (6.12)$$

where \mathbf{D} is the spectral matrix constituted by n samples of different concentrations, the dimension is $n \times m$, and m is the number of wavelength points of the spectra. There are k independent chemical components in the mixture; \mathbf{C} is the concentration matrix of pure components, and the dimension is $n \times k$. \mathbf{S} is a pure component spectral matrix with a dimension of $m \times k$. \mathbf{E} is the measurement error matrix.

As shown in Fig. 6.9, the calculation steps of MCR-ALS are as follows:

- (1) Determine the number of components of the spectral matrix. PCA (SVD decomposition) or prior knowledge is usually used to determine the number of components of the system.
- (2) Initialize pure component concentration matrix \mathbf{C} or pure component spectral matrix \mathbf{S} . Simple-to-use interactive self-modeling mixture analysis (SIMPLISMA) method is usually used for the initial pure component spectral matrix \mathbf{S} . The specific algorithm of SIMPLISMA is described in Sect. 5.2 of this book.
- (3) Repeat the iterative calculation of \mathbf{C} and \mathbf{S} through the following two expressions until convergence is achieved, and the iterative calculation ends:

$$\mathbf{C} = \mathbf{D}(\mathbf{S}^T)^+ \quad (6.13)$$

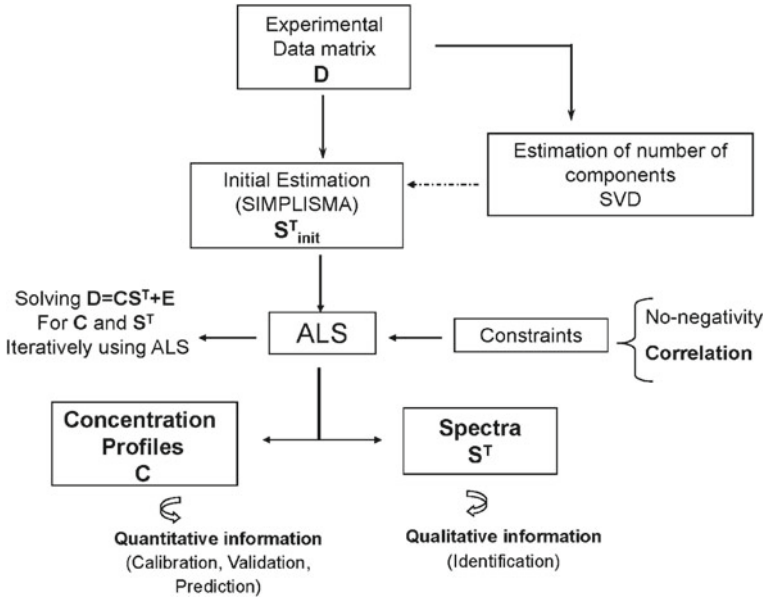


Fig. 6.9 Flowchart of resolution-alternating least squares algorithm (MCR-ALS)

$$S^T = C^+D \tag{6.14}$$

where $(S^T)^+$ and C^+ are the generalized inverse matrices of S^T and C , respectively. If S^T and C are full rank matrices, their generalized inverse matrices are $S(S^T S)^{-1}$ and $(C^T C)^{-1}C$, respectively.

In the calculation of MCR-ALS, there are permutation uncertainties, strength uncertainties, and rotation uncertainties, so it is necessary to introduce constraints to reduce or restrain the uncertainty of MCR-ALS. Commonly used constraints include concentration and spectral non-negative constraint, closure constraint, unimodality constraint, and correlation constraint [11–13].

6.2.6 Band Target Entropy Minimization

Before introducing band target entropy minimization (BTEM), a brief introduction to target transformation factor analysis (TTFA) is given. In this method, the spectral matrix of a mixture is decomposed to determine whether there is a target object in the mixture.

The expression of decomposition of the mixture spectral matrix D by TTFA is the same as that of PCA. Its essence is to rotate the loading vector so as to resolve the pure spectra. The model of TTFA is as follows:

$$\mathbf{D} = \mathbf{CS}^T + \mathbf{E} \quad (6.15)$$

where \mathbf{D} is the spectral matrix constituted by n mixture samples of different concentrations, and the dimension is $n \times m$, and m is the number of wavelength points of the spectra. There are k independent chemical components in the mixture; \mathbf{C} is the concentration matrix of pure components, and the dimension is $n \times k$. \mathbf{S} is a pure component spectral matrix with a dimension of $m \times k$. \mathbf{E} is the measurement error matrix.

The main steps of TTFA algorithm:

- (1) PCA is performed on \mathbf{D} at first.

$$\mathbf{D} = \mathbf{UV}^T + \mathbf{E} \quad (6.16)$$

where \mathbf{U} , \mathbf{V} , and \mathbf{E} are score, loading, and residual matrix, respectively. The dimensions of \mathbf{U} and \mathbf{V} are $n \times k$ and $m \times k$, respectively.

\mathbf{V} contains all the spectral information of \mathbf{S} , so the loading vector is also called the abstract spectrum. Therefore, any pure spectral \mathbf{s} vector in \mathbf{S} can have a linear representation of the loading vector \mathbf{V} .

$$\mathbf{s} = \mathbf{V}\mathbf{r} \quad (6.17)$$

where \mathbf{r} is a rotation vector and its dimension is $k \times 1$.

- (2) If the reference spectrum of the target object is \mathbf{s}^0 , then the rotation vector \mathbf{r} can be obtained by the least square method:

$$\mathbf{r} = (\mathbf{V}^T\mathbf{V})^{-1}\mathbf{V}^T\mathbf{s}^0 \quad (6.18)$$

- (3) Calculate the reconstructed spectrum \mathbf{s} by Eq. 6.17.
- (4) Check the similarity between the reference spectrum \mathbf{s}^0 and the reconstructed spectrum \mathbf{s} . If the consistency between the two is verified, the target is considered to exist in the mixture. Otherwise, the target is not contained in the mixture.

BTEM decomposes the spectral matrix of the mixture in the same way as TTFA, except that the reference spectrum \mathbf{s}^0 is not needed when calculating the rotation vector \mathbf{r} . Instead, the objective function is designed by combining the concept of target band and information entropy minimization in the spectrum. The optimal rotation vector \mathbf{r} was searched by simulated annealing algorithm to minimize the objective function. Finally, the pure component spectrum of the target object is obtained by Eq. 6.17. The detailed steps of BTEM algorithm can be seen in references [14, 15].

The central idea of BTEM method is to conduct PCA on the original spectral matrix of the mixture and decompose it into multiple loading vectors. The interested spectral features are identified from these feature vectors by visual observation, and select one of the interested spectral features to be retained during spectral reconstruction. The proposed method can compulsively preserve the selected spectral features and reconstruct the pure component spectra of the whole target object with the minimum entropy of the target band. Compared with TTFA, it does not require a priori information about the target (such as reference spectra), nor does it rely on the advantages of statistical test. The BTEM method has been successfully applied to various solid and liquid phase reaction systems, and successfully reconstructed complex spectra such as Raman spectroscopy, Fourier transform infrared spectroscopy (FT-IR), nuclear magnetic resonance (NMR), and mass spectrometry (MS) [16–18].

In addition to the MCR-ALS, TTFA, SIMPLISMA, and BTEM algorithms introduced in this book, interactive principal component analysis (IPCA) [19], orthogonal projection approach-alternating least squares (OPA-ALS) [20, 21], and target partial least square (TPLS) [22] can be used to identify the pure component spectra from the mixture spectral matrix.

6.2.7 Multilevel Simultaneous Component Analysis

Multilevel simultaneous component analysis (MSCA) method is proposed on the basis of PCA for data analysis with different types of variance in the data [23]. A two-level MSCA model can explain the inter-individual and intra-individual variances in the data, respectively. For an individual, this model can be expressed by Eq. 6.10 [24].

$$\mathbf{X}_{raw,i} = 1_{K_i} \mathbf{m}^T + 1_{K_i} \mathbf{t}_{b,i}^T \mathbf{P}_b^T + \mathbf{T}_{w,i} \mathbf{P}_w^T + \mathbf{E}_{MSCA,i} \quad (6.19)$$

The constraint condition is

$$\begin{cases} \sum_{i=1}^I K_i \mathbf{t}_{b,i}^T = 0 \\ 1_{K_i}^T \mathbf{T}_{w,i} = 0 \end{cases} \quad (6.20)$$

where $\mathbf{1}_{K_i}$ represents the column vector of size K_i , \mathbf{m}^T represents the overall mean value, $\mathbf{t}_{b,i}^T$ and $\mathbf{T}_{w,i}$ represent the scores of between-individual and within-individual models, respectively, and \mathbf{P}_b and \mathbf{P}_w represent the loadings of two models, respectively. $\mathbf{E}_{MSCA,i}$ represents the residual matrix and 0 represents a zero vector. By imposing constraint conditions on the score, the three parts of the model are guaranteed to be orthogonal to each other, that is, models at different levels can be explained by different types of variance in the data, respectively.

Since the MSCA method can analyze the data at different levels, it can be used to distinguish the temperature-controlled near-infrared (NIR) spectral data and investigate the effects of temperature and concentration on the spectrum at the same time. If the volume ratio of water and ethanol is the same for each group of samples, the volume concentration of isopropanol is 10, 20, 30, 40, 50, 60, 70, 80, and 90% of the mixed solution (i.e., $K_i = 9$), each group of sample data contains spectra measured at 7 temperatures, i.e., 7 individuals $\mathbf{X}_{raw,i}$ ($i = 1, 2, \dots, 7$).

The calculation of MSCA model includes the following two main steps [25]:

- (1) Firstly, the overall centralization of \mathbf{X}_{raw} is carried out, and then the one-level model is obtained by PCA analysis to the matrix composed of the mean vector of each individual. Since this model only explains the variance between individuals (temperature), it is called the between-temperature model. The model only considers the temperature effect and excludes the concentration effect. Therefore, the score calculated by the model between temperatures is called the temperature coefficient, and the quantitative relationship between spectrum and temperature (QSTR) can be obtained by the temperature coefficient at different temperatures.
- (2) Conduct local centralization of each individual after overall centralization, eliminate temperature effect through this step, and then PCA decomposition of matrix composed of all individuals is carried out to get the model. Since this model only considers the changes in the spectrum generated by changes in the individual or concentration, it is called the within-temperature model. The model only accounts for the concentration effect and excludes the temperature effect, so the score of the model is called the concentration coefficient. The quantitative relationship between the spectrum and the concentration can be obtained by the concentration coefficient at different concentrations.

The MSCA method is not only used for the establishment of two-level models (such as temperature and concentration), but also used for the establishment of three-level models. For example, the influence of temperature, concentration, and pH value on the spectrum can be investigated simultaneously [26].

6.3 Non-negative Matrix Factorization

Non-negative matrix factorization (NMF) and PCA are both linear data analysis methods. The basic idea of linear data analysis is to express the high-dimensional original data vector as a linear combination of a set of low-dimensional vectors through some appropriate transformation or decomposition. NMF requires the coefficients of these linear combinations are non-negative, that is, for any given non-negative matrix \mathbf{V} ($n \times m$), the goal of NMF algorithm is to find a non-negative matrix \mathbf{W} ($n \times r$) and a non-negative matrix \mathbf{H} ($r \times m$), such that $\mathbf{V} = \mathbf{W}\mathbf{h}$ is satisfied, so as to decompose a non-negative matrix into the product of two non-negative

matrix [27–29], where n is the number of samples, m is the number of wavelength variables, and r is the number of PCs of the system.

If the evaluation function F is constructed by the square of the residual Euclidean distance:

$$F = \sum_{i,j} (V_{ij} - (WH)_{ij})^2 \quad (6.21)$$

Then, matrices \mathbf{W} and \mathbf{H} can be obtained through a simple iterative process. The main steps of the algorithm are as follows:

- (1) Random initial values are assigned to non-negative matrices \mathbf{W} and \mathbf{H} .
- (2) Calculate \mathbf{W} from \mathbf{H} :

$$W'_{ia} = W_{ia} \sum_{\mu} \frac{V_{i\mu}}{(WH)_{i\mu}} H_{a\mu} \quad (6.22)$$

where $i = 1, 2, \dots, n$, $a = 1, 2, \dots, r$, $\mu = 1, 2, \dots, m$.

- (3) Column normalization:

$$W_{ia} = \frac{W'_{ia}}{\sum_j W'_{ja}} \quad (6.23)$$

Among them, $i = 1, 2, \dots, n$, $a = 1, 2, \dots, r$, $j = 1, 2, \dots, n$. \mathbf{W}'_{in} represents the new iteration value of \mathbf{W}_{in} .

- (4) Calculate \mathbf{H} from \mathbf{W} :

$$H'_{a\mu} = H_{a\mu} \sum_i W_{ia} \frac{V_{i\mu}}{(WH)_{i\mu}} \quad (6.24)$$

Among them, $i = 1, 2, \dots, n$, $a = 1, 2, \dots, r$, $\mu = 1, 2, \dots, m$.

When the iteration reaches the maximum number of iterations or the sum of squares of residuals between original data and reconstructed data is less than a given threshold, the iteration terminates.

It can be seen from the above calculation steps that NMF is performed based on each element in the matrix, rather than each vector in the matrix as PCA calculates, and the decomposition results of NMF can better represent the local characteristics of data. When the variables overlap with each other seriously, usually NMF can still find the “basis function” that characterizes the data structure, so NMF can directly extract the pure chemical composition information from the complex mixed systems. In addition, the “basis functions” of NMF are combined into individual variables by linear summation, which is more consistent with the combined characteristics of response spectra of different chemical components. In order to further improve the

separation ability of highly overlapping spectra, Gan et al. [30] proposed a non-negative matrix decomposition method based on pure variable initialization. Yin et al. [31] proposed a non-negative matrix decomposition method based on spectral feature constraints.

6.4 Independent Component Analysis

The purpose of independent component analysis (ICA) is to separate statistically independent source signals from multi-dimensional mixed signals of unknown source signals by linear transformation [32, 33]. For spectral matrix \mathbf{X} with $n \times m$ dimensions, where n is the number of samples and m is the number of wavelength points, supposing that the number of ICA components and the number of samples are the same as n , $n < m$, ICA model can be expressed as

$$\mathbf{X} = \mathbf{A}\mathbf{S} \quad (6.25)$$

where \mathbf{A} is the mixed matrix, dimension $n \times n$, \mathbf{S} is the component matrix, and dimension $n \times m$. Since both \mathbf{A} and \mathbf{S} are unknown, ICA is the optimal solution for finding the \mathbf{S} , so that

$$\mathbf{S} = \mathbf{W}\mathbf{S} \quad (6.26)$$

That is, to find a suitable separation matrix \mathbf{W} ($n \times n$), and then to obtain the independent component \mathbf{S} . The diversity of ICA objective function and optimization algorithm determine the complexity of ICA algorithm. In the specific application process, many typical algorithms have been formed, mainly including the following: fast ICA algorithm for maximum likelihood function estimation, mutual information minimization, non-Gaussian maximization, information maximization, negative entropy maximization, etc. Fast ICA algorithm is commonly used [34, 35]. The specific steps are as follows:

- (1) The spectral matrix \mathbf{X} is processed by mean centering.
- (2) \mathbf{X} was whitened. By SVD decomposition of the covariance matrix of \mathbf{X} , the eigenvalue diagonal matrix \mathbf{D} and eigenvalue vector matrix \mathbf{Q} were obtained. $\mathbf{U} = \mathbf{D}^{1/2}\mathbf{Q}^T$ was calculated, and the whitened matrix \mathbf{Z} was

$$\mathbf{Z} = \mathbf{U}\mathbf{X} \quad (6.27)$$

- (3) Set the number of components that need to be estimated n , let the number of iterations p , and randomly initialize the transformation matrix \mathbf{W} .
- (4) Calculate the separation matrix \mathbf{W} .

$$W_p = E\{Zg(W_p^T Z)\} - E\{g'(W_p^T Z)\}W \tag{6.28}$$

where $E\{\cdot\}$ is the mean value operation function; $g(\cdot)$ is a nonlinear function, usually using the tanh function; and $g'(\cdot)$ is the first derivative of the function $g'(\cdot)$.

- (5) The orthogonal matrix W .

$$W_p = W_p - \sum_{j=1}^{p-1} (W_p^T W_j) W_j \tag{6.29}$$

- (6) Standardization matrix W .

$$W_p = W_p / \|W_p\| \tag{6.30}$$

Judge whether W_p converges. If it converges, separate an independent component and proceed to the next step. If it does not converge, then return to step (4) to continue iterative calculation.

Let $p = p + 1$. If $p \leq n$, go back to step (4) until n independent components are calculated.

Calculate the composition matrix

$$S = WX \tag{6.31}$$

For the new spectrum x ($l \times m$), the vector a after dimensionality reduction by ICA is

$$a = xS^{-1} \tag{6.32}$$

where the dimension of a is $1 \times n$.

6.5 Multi-dimensional Scaling Transformation

Multi-dimensional scaling (MDS) transformation, also known as multi-dimensional scaling analysis, is a visualization method to display high-dimensional multivariate data in low-dimensional space. When the similarity (or distance) among each pair of n samples is fixed, the basic goal of multi-dimensional scaling is to determine the representation of these samples in low-dimensional (Euclidian) space (called perceptual mapping), and make it “roughly match” with the original similarity (or distance) as far as possible. This method can minimize any deformation caused by dimensionality reduction, so it is also called “Similarity structure analysis” [36].

If \mathbf{X} is an $n \times m$ -dimensional matrix composed of sample x_i , n is the number of samples, m is the number of wavelength variables, $i = 1, 2, \dots, n$, the steps of the multi-dimensional scaling algorithm are as follows:

Calculate the Euclidean distance between two samples in the distance \mathbf{X} matrix d_{ij} , $i, j = 1, 2, \dots, n$, forms the distance matrix \mathbf{D} .

The centralized inner product matrix \mathbf{B} is further calculated from the distance matrix \mathbf{D} , $\mathbf{B} = (b_{ij})_{n \times n}$,

$$b_{ij} = \frac{1}{2} \left(-d_{ij}^2 + \frac{1}{n} \sum_{j=1}^n d_{ij}^2 + \frac{1}{n} \sum_{i=1}^n d_{ij}^2 - \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^n d_{ij}^2 \right) \quad (6.33)$$

Conduct orthogonal decomposition of matrix \mathbf{B} , $\mathbf{B} = \mathbf{U}\mathbf{S}\mathbf{U}^t$; select f largest eigen-roots and their corresponding eigenvectors; and get the fitting composition \mathbf{Z} in f -dimensional space, $\mathbf{Z} = \mathbf{S}_f^{1/2} \mathbf{U}_f^t$, where $\mathbf{S}_f = \text{Diag}(\lambda_1, \lambda_2, \dots, \lambda_f)$ are the first f eigen-roots of matrix \mathbf{B} and $\mathbf{U}_f = [u_1, u_2, \dots, u_f]$ is a matrix composed of corresponding eigenvectors. \mathbf{Z} is the final matrix of multi-dimensional scaling transformation and its dimension is $n \times f$.

In essence, the basic goals of multi-dimensional scaling transformation and PCA are consistent. They both transform the high-dimensional spatial data into the low-dimensional space through the mapping of spatial variables, so as to maintain the original relationship of the data of each research object, and minimize any deformation caused by dimensionality reduction. The difference is that MDS takes samples as the analysis object, while PCA takes variables as the analysis object. It is mathematically proved that the f -dimensional principal coordinate of \mathbf{Z} after MDS transformation is exactly the value of the first f PCs obtained by using PCA after the centralization of \mathbf{X} matrix.

Chen et al. [37] used multi-dimensional scaling to reduce spectral variables, combined with multi-linear regression, to establish a quantitative model of four clinical biochemical indicators (glucose, low-density lipoprotein cholesterol, triglycerides, and urea). Wang et al. [38] used the MDS method to reduce the infrared spectra of asphalt and established a discrimination model that quickly identifies different brands.

6.6 Isometric Mapping

Isometric mapping (ISOMAP) [39] is a nonlinear dimensionality reduction technique and belongs to manifold learning method. MDS is a linear dimensionality reduction method. The Euclidean distance matrix constructed by MDS cannot reflect the nonlinear relationship between the sample points of manifolds. However, some of the data in the spatial distribution is like a twisted strip or spherical, etc. A common example is about the globe, if the South Pole to the North Pole, and the Euclidean distance is the linear distance of point to point, but the ant cannot walk in this way,

the shortest distance can be reached only by walking in the direction of the meridian, and this distance is called geodesic distance [40].

In order to keep the intrinsic geometric properties of data points (geodesic distance between two points) unchanged, ISOMAP algorithm uses geodesic distance among sample points to replace Euclidean distance on the basis of MDS. The approximate value of geodesic distance measurement can be obtained by using shortest path algorithm to reconstruct the local geodesic distance in the neighborhood. The samples in Fig. 6.10a are distributed on a Swiss-roll. The Euclidean distance between two points linked by the dashed line cannot represent the true distance between two points. The curve distributed on the manifold surface is the geodesic of the two points, which cannot be obtained under the condition that the manifold is unknown. The geodesic distance between the two points can be approximated by piecing the distance in the neighborhood through the shortest path algorithm, as shown in the curve in Fig. 6.10b. Figure 6.10c is the projection of two points and two paths (corresponding to geodesic distance and short-range distance splicing, respectively) in the space after dimensionality reduction using ISOMAP.

The ISOMAP algorithm first uses the shortest path in the nearest neighbor graph to get the approximate geodesic distance, inputs the distance into MDS instead of the Euclidean distance, and then finds the low-dimensional coordinates embedded in high-dimensional space. If is an $n \times m$ -dimensional matrix composed of samples, n is the number of samples, m is the number of wavelength variables, $i = 1, 2, \dots, n$, set the dimensionality reduction number d and the adjacent number k , the ISOMAP algorithm is as follows:

- (1) Construct K -neighborhood graph \mathbf{G} . Calculate the Euclidean distance \mathbf{D}_{ij}^E between each sample x_i and the rest sample x_j . When x_j is one of the k samples nearest to x_i , it is considered that x_i and x_j are adjacent, that is, graph \mathbf{G} has edge E_{ij} , and the weight of edge E_{ij} is set as d_{ij}^E .
- (2) Calculate the shortest path. When graph \mathbf{G} has edge E_{ij} , set shortest path $d_{ij}^G = d_{ij}^E$; otherwise, d_{ij}^G is equal to infinity. In Figure G, the shortest path distance matrix \mathbf{D}^G is obtained according to Dijkstra algorithm or Floyd algorithm.

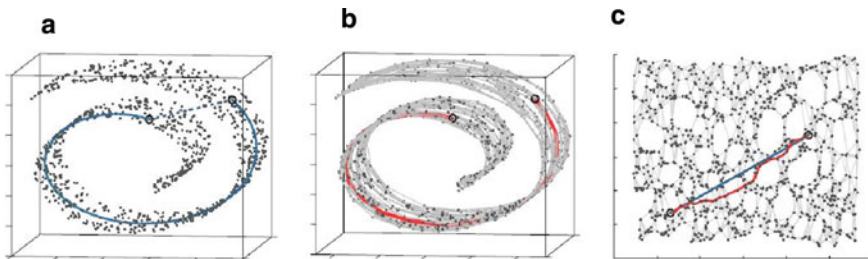


Fig. 6.10 The “Swiss-roll” data set, illustrating how ISOMAP exploits geodesic paths for nonlinear dimensionality reduction. **a** For two arbitrary points (circled) on a nonlinear manifold. **b** An approximation (red segments) to the true geodesic path **c** The two-dimensional embedding recovered by ISOMAP in step three low-dimensional projection results [39]

- (3) Calculated dimensional embedding. MDS is applied to the short-path distance matrix \mathbf{D}^G :

- ① Calculate the matrix \mathbf{R} ($n \times n$)

$$\mathbf{R} = (r_{ij}) = \left((d_{ij}^G)^2 \right) \quad (6.34)$$

- ② Calculation matrix \mathbf{H} ($n \times n$)

$$\mathbf{H} = (h_{ij}) = (\delta_{ij} - 1/n) \quad (6.35)$$

where $\delta_{ij} = 0$ ($i = j$), $\delta_{ij} = 1$ ($i \neq j$)

- ③ Calculate the matrix

$$\mathbf{L}^G = -\mathbf{H}\mathbf{R}\mathbf{H}/2 \quad (6.36)$$

Perform SVD decomposition on \mathbf{L}^G , because the matrix \mathbf{L}^G is symmetric, that is,

$$\mathbf{L}^G = \mathbf{U}^T \mathbf{S} \mathbf{U} \quad (6.37)$$

- ④ Calculate the projection matrix \mathbf{M} ($n \times d$) of low-dimensional space. d row and n column before matrix \mathbf{U} are taken to form matrix \mathbf{U}_d , d row and d column before matrix \mathbf{S} are taken to form matrix \mathbf{S}_d , then

$$\mathbf{M} = \mathbf{S}_d \mathbf{1}/2\mathbf{U}_d \quad (6.38)$$

Yang et al. [41] applied ISOMAP algorithm to the dimensionality reduction of NIR spectra, and then established a quantitative model by using PLS. The results showed that the prediction error could be significantly reduced when there was a nonlinear relationship between the property data and NIR spectra. Yu et al. [42] used ISOMAP algorithm to reduce the dimension of the NIR spectra of knots in solid wood plates, and then realized the effective modeling of the angle of the knot edge by using wavelet neural network. Lu et al. [43] applied ISOMAP algorithm to the dimensionality reduction of soil hyperspectral data, used random forest method to build the calibration model of copper content in mine tailings soil, and obtained better results than PCA dimensionality reduction. Ding et al. [44] also applied ISOMAP algorithm to dimensionality reduction of hyperspectral data, and the branching of similar categories could be greatly improved by using fewer feature dimensions. Li et al. [45] proposed an improved supervised dimensionality reduction method of ISOMAP, which used the correlation of spectral data itself to guide the construction of neighborhood map and reduce the sensitivity to noise and neighborhood parameters. Zhou et al. [46] proposed an ISOMAP algorithm based on matrix partitioning and automatic map adjustment to reduce the complexity of calculation and improve the calculation rate.

6.7 Local Linear Embedding

Local linear embedding (LLE) is a nonlinear dimensionality reduction method. Compared with the traditional PCA dimensionality reduction method, which focuses on the sample variance, LLE focuses on maintaining the local linear characteristics of the sample during dimensionality reduction, and can effectively realize the mapping of data from high-dimensional space to low-dimensional space [39].

The basic idea of LLE algorithm is to reveal the nonlinear dimensionality reduction of the global nonlinear structure through the joint of local linear relations in the sample space. The specific description of the algorithm is as follows: let $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N\}$ be the spectra of N input samples, and its low-dimensional mapping is $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_n\}$, and \mathbf{y}_i are eigenvectors of spectra \mathbf{x}_i . Any \mathbf{X}_i can be expressed as a linear combination of the spectra of its k adjacent samples:

$$\mathbf{x}_i = \sum_{j=1}^k \mathbf{w}_{ij} \mathbf{x}_{ij} \quad (6.39)$$

Among them, \mathbf{x}_{ij} is the j spectrum closest to \mathbf{x}_i and \mathbf{w}_{ij} is the linear reconstruction coefficient. LLE realizes the mapping of samples from high-dimensional space in low-dimensional space through local linear relations. The specific algorithm is as follows:

Input: Spectral matrix of sample set $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$, the nearest neighbor k , the dimension reduced to d .

Output: Low-dimensional sample set matrix $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_d\}$.

- (1) Take Euclidean distance as a measure to calculate the spectra of k samples closest to \mathbf{X}_i $\{\mathbf{x}_{i1}, \mathbf{x}_{i2}, \dots, \mathbf{x}_{ik}\}$.
- (2) Calculate the local variance matrix \mathbf{Z}_i and calculate the corresponding weight coefficient vector: \mathbf{w}_i

$$\mathbf{Z}_i = (\mathbf{x} - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^T \quad (6.40)$$

$$\mathbf{w}_i = \frac{\mathbf{Z}_i^{-1} \mathbf{1}_k}{\mathbf{1}_k^T \mathbf{Z}_i^{-1} \mathbf{1}_k} \quad (6.41)$$

Among them, $\mathbf{1}_k$ is the k dimension full $\mathbf{1}$ vector, $\mathbf{w}_i = (w_{i1}, w_{i2}, \dots, w_{ik})^T$.

- (3) According to the weight coefficient matrix \mathbf{W} of all sample spectra in the sample set, the calculation matrix

$$\mathbf{M} = (\mathbf{I} - \mathbf{W})(\mathbf{I} - \mathbf{W})^T \quad (6.42)$$

where \mathbf{I} is the identity matrix.

- (4) Calculate the second eigenvector of matrix \mathbf{M} to the eigenvector corresponding to $\mathbf{D} + 1$ minimum eigenvalue, that is, the output low-dimensional sample set matrix $\mathbf{Y} = \{\mathbf{y}_2, \mathbf{y}_3, \dots, \mathbf{y}_{d+1}\}$.

Duan et al. [47] used LLE to conduct nonlinear dimensionality reduction of the visible NIR spectra of eggs, and then used SVR to establish a model to predict egg freshness. The dimensionality reduction effect of LLE was better than that of PCA. Kang et al. [48] combined LLE with SVR to establish an analysis model for predicting COD in water samples by UV-Vis spectra, which can effectively extract nonlinear features in the spectra. Xu et al. [49] divided the NIR spectra of tobacco samples into regions, then performed LLE dimensionality reduction, and constructed a similarity measurement model. The accuracy rate was 93.3%, which improved the robustness and accuracy of the similarity measurement of NIR spectra. Zhang et al. [50] combined LLE with Gaussian process regression (LLE-GPR) to detect the quality of red pine nuts by NIR spectroscopy, which could accurately distinguish normal pine nuts from moldy pine nuts. Fan et al. [51] based on the fluorescence spectrum characteristics of TCM properties combined the LLE algorithm with random forest algorithm to construct the LLE-RF fluorescence spectrum classification model of cold and warm TCM, which has a good classification and recognition results.

6.8 T-Distributed Stochastic Neighborhood Embedding

T-distributed stochastic neighborhood embedding (t-SNE) algorithm is a popular manifold learning method for visual data dimensionality reduction. This algorithm can not only map the nearby points in the flow pattern to the nearby points in the low-dimensional representation, but also preserve the geometry of all scales, that is, map the nearby points to the nearby points and map the distant points to the distant points. T-SNE algorithm is a method of dimensionality reduction analysis using probability. It converts the Euclidean distance between any two data points in a high-dimensional space into the similar probability. In addition, the conditional probability of the stochastic neighbor embedding algorithm (SNE) is replaced by the joint probability between the data points in the high-dimensional space and the data points simulated in the low-dimensional space, so as to solve the problem of asymmetry in the SNE algorithm [52]. In addition, the algorithm adopts t-distribution in the low-dimensional space, which is a typical long-tail distribution. It can make the data points with medium and low equal distances in the high-dimensional space have a larger distance after the mapping, thus effectively solving the problem of data points crowding in the low-dimensional space.

For $n \times m$ -dimensional spectral matrix \mathbf{X} , where n is the number of samples and m is the number of wavelength points, the t-SNE steps are as follows:

- (1) The joint probability \mathbf{P}_{ef} in the m -dimensional space is calculated. Similar conditional probability \mathbf{P}_{ef} and $\mathbf{p}_{f|e}$ of two spectra in m -dimensional space can be calculated.

$$\mathbf{p}_{e|f} = \frac{\exp(-\|x_f - x_e\|^2 2\sigma_f^2)}{\sum_{f=1}^n \sum_{g=1}^n \exp(-\|x_f - x_g\|^2 2\sigma_f^2)} \quad (f \neq g) \quad (6.43)$$

$$\mathbf{p}_{f|e} = \frac{\exp(-\|x_e - x_f\|^2 2\sigma_e^2)}{\sum_{e=1}^n \sum_{g=1}^n \exp(-\|x_e - x_g\|^2 2\sigma_e^2)} \quad (e \neq g) \quad (6.44)$$

where \mathbf{P}_{eff} represents the probability that the f sample is distributed around sample e , $\mathbf{p}_{e/e} = 0$, and σ_f represents the variance of the Gaussian distribution in the center of \mathbf{x}_f .

The high-dimensional joint probability PEF is expressed as follows:

$$\mathbf{pef} = \frac{\mathbf{p}_{f|e} + \mathbf{p}_{e|f}}{2n} \quad (6.45)$$

- (2) Calculate the joint probability \mathbf{q}_{ef} in the low-dimensional space:

T-SNE algorithm adopts t-distribution in low-dimensional space, and the joint probability \mathbf{q}_{ef} of low-dimensional space \mathbf{Z} ($n \times d$, where d is the dimension after dimensional reduction) is expressed as follows:

$$\mathbf{q}_{ef} = \frac{(1 + \|z_e - z_f\|^2)^{-1}}{\sum_{g=1}^n \sum_{l=1}^n (1 + \|z_g - z_l\|^2)^{-1}} \quad (g \neq l) \quad (6.46)$$

- (3) Calculate the KL divergence between \mathbf{p}_{ef} and \mathbf{q}_{ef} , and set it as the objective function \mathbf{C} , i.e.,

$$\mathbf{C} = \mathbf{KL}(P||Q) = \sum_{e=1}^n \sum_{f=1}^n \mathbf{p}_{ef} \log_2 \frac{\mathbf{p}_{ef}}{\mathbf{q}_{ef}} \quad (6.47)$$

KL divergence is used to measure the similarity of two spatial distributions of high and low dimensions. The goal of SNE algorithm is to minimize the KL distance for all data points in the sample set.

- (4) Take the derivative of the low-dimensional expression corresponding to the input data with the objective function \mathbf{C} :

$$\frac{\delta \mathbf{C}}{\delta \mathbf{Z}_e} = 4 \sum_{f=1}^n (\mathbf{p}_{ef} - \mathbf{q}_{ef})(z_e - z_f)(1 + \|z_e - z_f\|^2)^{-1} \quad (6.48)$$

The low-dimensional expression is taken as an optimizable variable to be optimized, and the optimal simulation point of the input matrix \mathbf{X} in the low-dimensional space is obtained.

- (5) Define the perplexity:

$$\mathit{Perp}(\mathbf{p}_e) = 2^{H(\mathbf{p}_e)} \quad (6.49)$$

Among them,

$$H(\mathbf{p}_e) = - \sum_{f=1}^n p_{f|e} \log_2 p_{f|e} \quad (6.50)$$

The perplexity can be interpreted as the number of valid nearest neighbors near a point, which is usually selected between 5 and 50. It is a global parameter that controls the fitting and affects the complexity of the Gaussian distribution in higher-dimensional space. It is necessary to adjust the perplexity continuously in order to get the optimal dimensionality reduction result.

- (6) In order to obtain the minimum objective function \mathbf{C} , multiple iterations of the input matrix \mathbf{X} are needed. By adjusting the parameters such as the perplexity, learning rate η , and momentum $\alpha(t)$, the specific iteration steps are as follows:

- ① The perplexity is calculated, and the iteration number \mathbf{T} , learning rate η , and momentum $\alpha(t)$ are set.
- ② Calculate the \mathbf{p}_{ef} at a given perplexity.
- ③ Initialize $\mathbf{Z}(\mathbf{0})$ with the normal distribution $N(0, 10^{-4}I)$.
- ④ Iterate from $t = 1$ to T .
- ⑤ Calculate QEF in low dimensions.
- ⑥ Calculate the gradient $\frac{\delta \mathbf{C}}{\delta \mathbf{Z}}$.
- ⑦ Update all landowners

$$\mathbf{Z}^{(t)} = \mathbf{Z}^{(t-1)} + \eta \frac{\delta \mathbf{C}}{\delta \mathbf{Z}} + \alpha(t)(\mathbf{Z}^{(t-1)} - \mathbf{Z}^{(t-2)}) \quad (6.51)$$

- ⑧ Determines whether t is equal to T , otherwise $t = t + 1$, returns ⑤.

Wang et al. [53] established a random forest model to identify egg origin using the short-wave NIR spectral feature extracted by t-SNE, and the result was better than that of PCA. Li et al. [54] used t-SNE to reduce the dimension of NIR spectra of, and then realized the identification of different wood species through cluster analysis. Li et al. [55] used t-SNE to map sample points of high-dimensional terahertz time domain spectra to low-dimensional space, and realized visual observation of sample features in low-dimensional space. Li et al. [56] used t-SNE to reduce the dimension of NIR spectra of pine nuts, which was used as the input of SVM classification model, and the accuracy of identification of pine nuts storage period could reach 97.5%.

6.9 Other Algorithms

In addition to PCA, ICA, and MDS, there are also projection pursuit (PP), minimum noise fraction rotation (MNF Rotation), etc. [57, 58] for linear dimensionality reduction. Among the nonlinear dimensionality reduction methods, there is kernel PCA (KPCA) and kernel ICA (KICA) based on kernel function.

In the nonlinear manifold dimensionality reduction, besides LLE, ISOMAP, and t-SNE, there are also Laplacian eigenmaps (LE) [59–61], locality preserving projection (LPP) [62], diffusion maps (DM) [63], Hessian locally linear embedding method (HLLE) [64], linear local tangent space alignment (LLTSA) [65], etc. [66].

This chapter mainly discusses spectral dimensionality reduction methods based on spectra. If concentration or classification is considered to participate in data dimensionality reduction, which is called supervised data dimensionality reduction, methods such as partial least squares (PLS), canonical correlation analysis (CCA), Fisher linear discriminant analysis (LDA), supervised locality preserving projection (SLPP) can be considered [67, 68].

References

1. Hu YZ. Computational drug analysis. Beijing: Science Press; 2006.
2. Hughes G. On the mean accuracy of statistical pattern recognizers. *IEEE Trans Inf Theory*. 1968;14(1):55–63.
3. Wold S, Esbensen K, Geladi P. Principal component analysis. *Chemom Intell Lab Syst*. 1987;2(1–3):37–52.
4. Wang GZ, Ye H. Principal component analysis and partial least squares. Beijing: Tsinghua University Press; 2012.
5. Zhu YN, Yang P, Yang X, et al. Classification of fresh meat species using laser-induced breakdown spectroscopy with support vector machine and principal component analysis. *Chin J Anal Chem*. 2017; 45(3):336–41.
6. Zang J, Yang XH. Multivariate statistical process control. Beijing: Chemical Industry Press; 2000.
7. Pan LD. Advanced control and on-line optimization technology and its application. Beijing: China Machine Press; 2009.
8. Windig W, Guilment J. Interactive self-modeling mixture analysis. *Anal Chem*. 1991;63(14):1425–32.
9. Chen G, Harrington PDB. Real-time interactive self-modeling mixture analysis. *Appl Spectrosc*. 2001;55(5):621–9.
10. Azzouz T, Tauler R. Application of multivariate curve resolution alternating least squares (MCR-ALS) to the quantitative analysis of pharmaceutical and agricultural samples. *Talanta*. 2008;74(5):1201–10.
11. Lyndgaard LB, Frans VDB, De JA. Quantification of paracetamol through tablet blister packages by Raman spectroscopy and multivariate curve resolution-alternating least squares. *Chemom Intell Lab Syst*. 2013;125:58–66.
12. Oliveira RR, Lima KM, Tauler R, et al. Application of correlation constrained multivariate curve resolution alternating least-squares methods for determination of compounds of interest in biodiesel blends using NIR and UV-visible spectroscopic data. *Talanta*. 2014;125:233–41.

13. Garrido M, Rius FX, Larrechi MS. Multivariate curve resolution-alternating least squares (MCR-ALS) applied to spectroscopic data from monitoring chemical reactions processes. *Anal Bioanal Chem.* 2008;390(8):2059–66.
14. Gao Q, Lu F. The principle and application of band-target entropy minimization. *Comput Appl Chem.* 2011;28(10):127–30.
15. Tan ST, Zhu HH, Chew W. Self-modeling curve resolution of multi-component vibrational spectroscopic data using automatic band-target entropy minimization (AUTOBTEM). *Anal Chim Acta.* 2009;639(1–2):29–41.
16. Chew W, Widjaja E, Garland M. Band-target entropy minimization (BTEM): an advanced method for recovering unknown pure component spectra. Application to the FTIR spectra of unstable organometallic mixtures. *Organometallics* 2002; 21(9):1982–90.
17. Widjaja E, Garland M. Pure component spectral reconstruction from mixture data using SVD, global entropy minimization, and simulated annealing. Numerical investigations of admissible objective functions using a synthetic 7-species data set. *J Comput Chem.* 2002; 23(9):911–9.
18. Yu LL, Shao LM. Qualitative analysis of open-path fourier transform infrared spectra. *Chin J Anal Chem.* 2015;43(2):226–32.
19. Bu DS, Brown CW. Self-modeling mixture analysis by interactive principal component analysis. *Appl Spectrosc.* 2000;54:1214–21.
20. Sanchez FC, Toft J, Massart DL, et al. Orthogonal projection approach applied to peak purity assessment. *Anal Chem.* 1996;68:79–85.
21. Frenich AG, Zamora DP, Vidal JLM, et al. Resolution (and Quantitation) of mixtures with overlapped spectra by orthogonal projection approach and alternating least squares. *Anal Chim Acta.* 2001;449(1–2):143–55.
22. Feudale RN, Brown SD. An inverse model for target detection. *Chemom Intell Lab Syst.* 2005;77(1–2):75–84.
23. Timmerman ME. Multilevel component analysis. *Br J Math Stat Psychol.* 2006;59(2):S301–320.
24. Cui XY, Liu XW, Yu XM, et al. Water can be a probe for sensing glucose in aqueous solutions by temperature dependent near infrared spectra. *Anal Chim Acta.* 2017;957:47–54.
25. Shan RF. Modeling methods and temperature effects for near-infrared spectra. Tianjin: Nankai University; 2014.
26. Han L, Cui XY, Cai WS, et al. Three-level simultaneous component analysis for analyzing the near-infrared spectra of aqueous solutions under multiple perturbations. *Talanta* 2020; 217:121036.
27. Liu P, Li B, Yu DY, et al. Analysis of option waveguide spectroscopy by non-negative matrix factorization. *J Huazhong Univ Sci Technol (Nat Sci Ed).* 2013;41(8):6–9.
28. Gao JL, Li TH, Gao HT, et al. Analysis of pKa of Mixed Acid with NMF. *Comput Appl Chem.* 2007;24(5):604–9.
29. Wang GZ. The re-research of non-negative matrix factorization and its application in chemical spectra resolution. Qingdao University of Science and Technology; 2007.
30. Gan JZ, Qin BY, Li Y, et al. Resolution of overlapping terahertz spectra using non-negative matrix factorization base on pure variables initialization. *Optik.* 2019;176:600–10.
31. Yin XH, Liu Y, Feng ML, et al. Separation of tire rubber overlapping terahertz spectra using non-negative matrix factorization of spectral feature constraints. *Spectrosc Spectr Anal.* 2020;40(12):3736–42.
32. Chen J, Wang XZ. A new approach to near-infrared spectral data analysis using independent component analysis. *J Chem Inf Comput Sci.* 2001;41(4):992–1001.
33. Kassouf A, Ruellan A, Bouveresse DJR, et al. Attenuated total reflectance-mid infrared spectroscopy (ATR-MIR) coupled with independent components analysis (ICA): a fast method to determine plasticizers in polylactide (PLA). *Talanta.* 2016;147:569–80.
34. Yu SH, Zhang YJ, Zhao NJ, et al. Analysis of three-dimensional fluorescence overlapping spectra using differential spectra and independent component analysis. *Spectrosc Spectr Anal.* 2013;33(1):111–5.

35. Wang J, Jin AD. ICA-based dimensionality reduction and segmentation of hyperspectral image. *Geomat Spatial Inf Technol*. 2018;41(6):86–90.
36. He XQ. *Multivariate statistical analysis*. 4th ed. Beijing: China Renmin University Press; 2015.
37. Chen HZ, Song QQ, Shi K, et al. Multidimensional scaling linear regression applied to FTIR spectral quantitative analysis of clinical parameters of human blood serum. *Spectrosc Spectr Anal*. 2015;35(4):914–8.
38. Wang K. Rapid identifying bitumen produced by different manufacturers with IR and multidimensional scaling. *Phys Test Chem Anal (Part B: Chem Anal)* 2019; 55(2):141–6.
39. Tenenbaum JB, De Silva V, Langford JC. A global geometric framework for nonlinear dimensionality reduction. *Science*. 2000;290(5500):2319–23.
40. Lei M. *Machine learning principles, algorithms and applications*. Beijing: Tsinghua University Press; 2019.
41. Yang HH, Tan F, Wang YM, et al. Isomap-PLS nonlinear modeling method for near infrared spectroscopy. *Spectrosc Spectr Anal*. 2009;29(2):322–6.
42. Yu HL, Zhang M, Hou HY, et al. The inversion of knots in solid wood plates based on near-infrared spectroscopy. *Spectrosc Spectr Anal*. 2019;39(8):2618–23.
43. Lv J, Hao NY, Shi XL. Extraction of hyperspectral characteristics of soil based on manifold learning. *J Arid Land Resour Environ*. 2015;29(7):176–80.
44. Ding L, Tang P, Lin HY. Dimensionality reduction and classification for hyperspectral remote sensing data using ISOMAP. *Infrared Laser Eng*. 2013;42(10):2707–11.
45. Lin QB, Jia ZH. A dimension reduction method applied in spectrum analysis. *Spectrosc Spectr Anal*. 2013;33(3):780–4.
46. Zhou SY, Tan K, Wu LX. Hyperspectral image classification based on ISOMAP algorithm using neighborhood distance. *Remote Sens Technol Appl*. 2014;29(4):695–700.
47. Duan YF, Wang QH, Ma MH, et al. Study on non-destructive detection method for egg freshness based on LLE-SVR and visible/near-infrared spectrum. *Spectrosc Spectr Anal*. 2016;36(4):981–5.
48. Kang B, Ma J. Study of UV visible spectrum-based COD detection method for water quality monitoring based on LLE-SVR. *Sens World*. 2018;24(9):11–5.
49. Xu BD, Ding XQ, Qing YH, et al. Similarity measurement method of near infrared spectrum based on grid division local linear embedding algorithm. *Laser Optoelectron Prog*. 2019;56(3):251–7.
50. Zhang DY, Jiang DP, Zhou BL, et al. Near-infrared detection of outer pine nuts by LLE manifold learning. *J Northeast For Univ*. 2019;47(6):45–8.
51. Fan FJ, Xuan FL, Bai Y, et al. Pattern recognition of traditional Chinese medicine property based on three-dimensional fluorescence spectrum characteristics. *Spectrosc Spectr Anal*. 2020;40(6):1763–8.
52. Yu HL, Huo JY, Zhang YZ, et al. Urban vegetation identification method based on PCA-t-SNE-SVM. *Res Explor Lab*. 2019;38(12):135–40.
53. Wang B, Wang QH, Xiao Z, et al. Discrimination of origin of eggs using visible-near-infrared spectroscopy and random forest. *Sci Technol Food Ind*. 2017;38(24):243–7.
54. Li Y. Study of non-destructive detection of wood species and density based on visible/near infrared spectroscopy. Northeast Forestry University;2019.
55. Li TJ. Research on sample feature recognition algorithm based on terahertz time domain spectroscopy. Chongqing University;2018.
56. Li HB, Cao J, Jiang DP, et al. Identification of new and old *Pinus Koraiensis* seeds by near-infrared spectroscopy (NIRS) with t-SNE dimensionality reduction. *Spectrosc Spectr Anal*. 2020;40(9):2918–24.
57. Yang RX, Yang Y, Yuan JJ. Research on hyper-spectral image feature extraction and feature selection. *J Guangxi Teach Educ Univ Nat Sci Ed*. 2015;2:39–43.
58. He RY, Jiang JB, Guo HQ, et al. Using projection pursuit dimension reduction to estimate canopy chlorophyll density of winter wheat. *J Triticeae Crops*. 2014;34(10):1447–52.
59. Liu P, Ai SR, Yang PX, et al. Nonlinear manifold dimensionality reduction methods for quick discrimination of tea at different altitude by near infrared spectroscopy. *J Tea Sci*. 2019;39(6):715–22.

60. Lin P, Chen YM, Zou ZY. Quick discrimination of rice storage period based on manifold dimensionality reduction methods and near infrared spectroscopy techniques. *Spectrosc Spectr Anal.* 2016;36(10):3169–73.
61. Li X, Lv Y. A weighted naive bayes hyperspectral classification algorithm combined with laplacian eigen mapping. *J Instrum Anal.* 2020;38(10):1293–8.
62. Liu WJ, Li WJ, Tan H, et al. Research on identifying maize haploid seeds using near infrared spectroscopy based on kernel locality preserving projection. *Spectrosc Spectr Anal.* 2019;39(8):2574–7.
63. Ni JP, Shen T, Zhu Y, et al. Terahertz spectroscopic identification with diffusion maps. *Spectrosc Spectr Anal.* 2017;37(8):2360–4.
64. Jin R, Li XY, Yan YY, et al. Detection method of multi-target recognition of potato based on fusion of hyperspectral imaging and spectral information. *Trans Chin Soc Agric Eng.* 2015;31(16):258–63.
65. Ma YJ, Guo JX, Guo ZM, et al. Origin tracing of red Fuji apple based on near infrared transmission spectrum and various dimension reduction methods. *Modern Food Sci Technol.* 2020;36(6):303–9.
66. Guo JX, Ma YJ, Guo ZM, et al. Watercore identification of Xinjiang Fuji apple based on manifold learning algorithm and near infrared transmission spectroscopy. *Spectrosc Spectr Anal.* 2020;40(8):2415–20.
67. He KX, Cheng H, Du WL, et al. Online updating of NIR model and its industrial application via adaptive wavelength selection and local regression strategy. *Chemom Intell Lab Syst.* 2014;134:79–88.
68. Lee S, Kim K, Lee H, et al. Improving the classification accuracy for IR spectroscopic diagnosis of stomach and colon malignancy using non-linear spectral feature extraction methods. *Analyst.* 2013;138(14):4076–82.

Chapter 7

Linear Calibration Methods



7.1 Univariate Linear Regression

Unary linear regression is the simplest linear regression with the formula of $y = b_0 + bx + \varepsilon$, where x is an observable and controllable variable, and it is often called an independent variable or a controlled variable (absorbance of NIRS), y is the dependent variable (such as the benzene content of gasoline, the protein content of wheat, etc.), b_0 and b are regression coefficients, ε is measurement error.

In regression analysis, the main purpose is to find the best estimated values of b_0 and b based on a set of n measured values (x_i, y_i) , so as to achieve the closest degree of \hat{y} and y . Once \hat{b}_0 and \hat{b} are calculated, they can be used for predictive analysis [1].

Estimated values of b_0 and b are often obtained by least square method as considering sum of square $Q(b_0, b) = \sum_{i=1}^n (y_i - b_0 - bx_i)^2$, find b_0 and b that minimize $Q(b_0, b)$ as its estimation, that is, $Q(\hat{b}_0, \hat{b}) = \min_{b_0, b} Q(b_0, b)$.

$$\text{As for, } \begin{cases} \frac{\partial Q}{\partial b_0} = 2 \sum_{i=1}^n [y_i - b_0 - bx_i] = 0 \\ \frac{\partial Q}{\partial b} = 2 \sum_{i=1}^n (y_i - b_0 - bx_i)x_i = 0 \end{cases} \quad (7.1)$$

$$\text{Solution } \begin{cases} \hat{b} = \frac{L_{xy}}{L_{xx}} \\ \hat{b}_0 = \bar{y} - \hat{b}\bar{x} \end{cases} \quad (7.2)$$

where \hat{b}_0 and \hat{b} are the least square estimate values of b_0 and b , respectively.

$$L_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n x_i^2 - \frac{1}{n} \left(\sum_{i=1}^n x_i \right)^2$$

$$L_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^n x_i y_i - \frac{1}{n} \left(\sum_{i=1}^n x_i \right) \left(\sum_{i=1}^n y_i \right) \quad (7.3)$$

where \bar{x} and \bar{y} are the average values of n measurement data x_i and y_i , respectively.

In spectral analysis, unary linear regression is often used to evaluate the correlation between the spectral prediction results of a set of samples and the results of the reference method.

7.2 Multiple Linear Regression

In practice, multiple regression methods are used in many cases to better describe the relationship between variables. Actually, the method of dealing with multivariate is basically the same as that of single variate, except that the calculation of multiple linear regression (MLR) is much larger, and computers are generally involved for processing [2].

Suppose there is a relationship between the dependent variable y and the independent variables x_1, x_2, \dots, x_m : $y = b_0 + b_1 x_1 + \dots + b_m x_m + \varepsilon$;

For n sets of measurement data:

$$\begin{aligned} &(y_1; x_{11}, x_{12}, \dots, x_{1m}) \\ &(y_2; x_{21}, x_{22}, \dots, x_{2m}) \\ &\dots \dots \dots \\ &(y_n; x_{n1}, x_{n2}, \dots, x_{nm}) \end{aligned}$$

where x_{ij} is the i th observation value of the independent variable x_j , y_i is the i th value of the dependent variable y , and m is the number of independent variables (such as m spectral wavelengths participating in the regression). Data structure of the model is as follows:

$$\begin{aligned} y_1 &= b_0 + b_1 x_{11} + b_2 x_{12} + \dots + b_m x_{1m} + \varepsilon_1 \\ y_2 &= b_0 + b_1 x_{21} + b_2 x_{22} + \dots + b_m x_{2m} + \varepsilon_2 \\ \dots &\dots \dots \dots \\ y_n &= b_0 + b_1 x_{n1} + b_2 x_{n2} + \dots + b_m x_{nm} + \varepsilon_n \end{aligned}$$

Above equation can be written in matrix form as $y = Xb + \varepsilon$ where

$$\mathbf{y} = \begin{bmatrix} y_1 \\ \dots \\ y_n \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & \dots & x_{1m} \\ 1 & x_{21} & \dots & x_{2m} \\ \dots & \dots & \dots & \dots \\ 1 & x_{n1} & \dots & x_{nm} \end{bmatrix} \quad \mathbf{b} = \begin{bmatrix} b_0 \\ \dots \\ b_m \end{bmatrix} \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \dots \\ \varepsilon_n \end{bmatrix} \quad (7.4)$$

The estimated value of \mathbf{y} obtained by the least square method is $\hat{\mathbf{y}}$, and residual sum of squares is $S_{res} = \boldsymbol{\varepsilon}^T \boldsymbol{\varepsilon} = (\mathbf{y} - \mathbf{X}\mathbf{b})^T (\mathbf{y} - \mathbf{X}\mathbf{b}) = \mathbf{y}^T \mathbf{y} - \mathbf{b}^T \mathbf{X}^T \mathbf{y} - \mathbf{y}^T \mathbf{X}\mathbf{b} + \mathbf{b}^T \mathbf{X}^T \mathbf{X}\mathbf{b}$.

To obtain the minimum value of S_{res} , \mathbf{b} must satisfy the equation: $\frac{\partial S_{res}}{\partial \mathbf{b}} = \frac{\partial}{\partial \mathbf{b}} (\mathbf{y} - \mathbf{X}\mathbf{b})^T (\mathbf{y} - \mathbf{X}\mathbf{b}) = 0$, that is, $-2\mathbf{X}^T (\mathbf{y} - \mathbf{X}\mathbf{b}) = 0$.

Formal equation is obtained as $\mathbf{X}^T \mathbf{X}\mathbf{b} = \mathbf{X}^T \mathbf{y}$.

Solve the above equation and get the estimated value of regression coefficient as $\hat{\mathbf{b}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$.

MLR is a basic algorithm for early quantitative analysis of NIRS that is suitable for simple systems with particularly good linear relationships. The formula is very clear and simple, without considering the influence of mutual interference between components. But MLR has many limitations. Firstly, due to the dimension limit of equation, the number of variables (wavelength points) involved in the regression cannot exceed the number of samples in calibration set. The limited wavelengths would inevitably lose many useful spectral information. Secondly, spectral matrix \mathbf{X} often has a collinearity problem, that is, at least one column or one row in \mathbf{X} can be expressed by a linear combination of other columns or rows, leading to $|\mathbf{X}^T \mathbf{X}|$ to be equal to or close to zero. This kind of ill conditioned matrix cannot find its inverse matrix or the obtained inverse matrix is unstable. Thirdly, because the noise of the \mathbf{X} matrix is not considered in the regression process, it often leads to the occurrence of over-fitting, which will greatly reduce the predictive ability of model.

7.3 Concentration Residual Augmented Classical Least Squares

Concentration residual augmented classical least squares (CRACLS) is an improved algorithm proposed on the basis of multiple linear regression [3, 4].

If the calibration set spectral matrix is $\mathbf{X}(n \times m)$, n is the number of samples, m is the number of wavelength points; concentration matrix is $\mathbf{Y}(n \times p)$, and p is the number of components, then the CRACLS algorithm steps are as follows:

- (1) Calculate the absorption coefficient matrix \mathbf{S} : $\mathbf{S} = (\mathbf{Y}^T \mathbf{Y})^{-1} \mathbf{Y}^T \mathbf{X}$
- (2) Calculate the concentration prediction matrix $\hat{\mathbf{Y}}$: $\hat{\mathbf{Y}} = \mathbf{X}\mathbf{S}^T (\mathbf{S}\mathbf{S}^T)^{-1}$
- (3) Calculate the concentration residual matrix \mathbf{E} : $\mathbf{E} = \hat{\mathbf{Y}} - \mathbf{Y}$
- (4) Amplify a column in the concentration residual matrix \mathbf{E} to the concentration matrix \mathbf{Y} to obtain a new concentration matrix \mathbf{Y}_+
- (5) Replace \mathbf{Y} with \mathbf{Y}_+ , repeat Steps (7.1)–(7.4) until the error \mathbf{E} meets the requirements.

This method maintains the advantages of the least square method and can partially solve the problem of spectral overlap, improve the utilization of spectral information, and obtain a model with better predictive ability. Similarly, the above strategy can also be used to augment the spectral residuals as spectral residual augmented classical least squares (SRACLS) [5, 6].

7.4 Stepwise Linear Regression

Stepwise linear regression is a method of solving multicollinearity through variable selection. All possible variable combinations of m wavelengths would be listed to relate variable y to establish an MLR regression equation, and then the optimal equation according to the screening criterion is selected. In practice, the stepwise variable selection method is commonly used as three methods including stepwise backward method, stepwise forward method, and stepwise regression analysis (SRA) [7, 8].

- (1) Stepwise backward method (reverse screening): From the regression equation containing all m variables, according to the criterion, it eliminates a variable that does not have a significant effect on y at a time until it cannot be eliminated.
- (2) Stepwise forward method (forward screening): Start with a first variable, it introduces a variable that has a significant impact on y each time, until it cannot be introduced.
- (3) Stepwise regression analysis (reverse and forward combining screening): Start with a first variable, it introduces a variable that has a significant impact on y each time, until it cannot be introduced. Then, it eliminates a variable that does not have a significant effect on y at a time until it cannot be eliminated. Repeat the above forward and reverse screening until it is unable to be selected or excluded.

When using stepwise regression, the problem often encountered is that there are multiple interactions between input variables, which are not only related to output, but also related to each other. In this case, an input variable in the model may shield the influence of other variables on the result. Thus, the variables selected by the stepwise regression method are not optimal in most cases. Besides, the work of screening more than a dozen variables from hundreds or even thousands of actual wavelengths is extremely overloaded. As a result, principal component regression and partial least squares methods developed based on factor analysis have well solved the above problems and become the common algorithms in modern spectroscopy analysis.

7.5 Ridge Regression

For the multicollinearity problems, Hoerl proposed an improved least squares estimation method named Ridge Regression in 1962. When $|X^T X| \approx 0$, a matrix of normal numbers λI ($\lambda > 0, I$ is the identity matrix) to $X^T X$, then the probability that the matrix $(X^T X + \lambda I)^{-1}$ is close to singularity will be much smaller than that of $(X^T X)^{-1}$. The ridge regression estimate of the regression coefficient is expressed as $b(\lambda) = (X^T X + \lambda I)^{-1} X^T y$.

When using regularization, attention needs to be paid to the choice of regularization parameter λ [9]. If λ is too large, all regression coefficients will be minimized, and the final model will be almost a horizontal straight line, causing under-fitting problems. If λ is too small, then the regular term is almost ineffective, which will lead to improper solutions to over-fitting or multicollinearity problems.

7.6 Lasso Regression

Least absolute shrinkage and selection operator (Lasso) introduces a norm penalty term in the least square regression estimation, that is, adds a L1 regular term, which is generally used to calculate the sum of absolute errors between two vectors, as is to calculate the absolute value. The L1 norm has natural advantages in terms of sparse solution. It compresses the regression coefficients of some less contributing variables to 0, thereby removing useless information features to achieve the purpose of sparseness and feature selection. Formula for solving the Lasso regression coefficient is as follows:

$$\hat{\beta}(Lasso) = \underset{\beta}{\operatorname{argmin}} \left[\sum_{i=1}^n \left(y_i - \sum_{j=1}^m x_{i,j} \beta_j \right)^2 + \lambda \sum_{j=1}^m |\beta_j| \right] \quad (7.5)$$

$$\sum_{j=1}^p |\beta_j| \leq t \quad (7.6)$$

The above can be rewritten in matrix form as

$$\hat{\beta}_{Lasso}(\lambda_1) = \underset{\beta}{\operatorname{argmin}} \left\{ \beta^T (X^T X) \beta - 2y^T X \beta + \lambda_1 \sum_{j=1}^p |\beta_j| \right\} \quad (7.7)$$

where t is constraint constant and $t \geq 0$, λ is the regularization parameter, also known as the penalty coefficient; $i = 1, 2, \dots, n$ as sample number, $j = 1, 2, \dots, m$, m is the number of wavelength points of the spectrum. As λ increases, the optimal solution $\sum_{j=1}^p |\beta_j|$ will decrease, the coefficients of some independent variables will

be compressed to 0, so as to achieve reduction of high-dimensional data, which can better solve many problems in high-dimensional data modeling. The essence of the Lasso algorithm is to minimize the residual sum of squares under the constraint that the sum of the absolute values of the regression coefficients is not greater than λ to generate some regression coefficients strictly equal to 0, and finally obtain the estimated values of the parameters. Because a penalty term is added, the Lasso algorithm is a biased estimation method compared to the classical least square method, which improves the predictive ability of the model by sacrificing part of the deviation and also makes the model more stable.

There have been many reports on Lasso regression, such as the use of Lasso method in NIRS to predict the physical properties of pulped wood [10], rapid determination of eucalyptus extract content [11], Lasso combined with Boosting method to analyze the high-concentration zinc ions and trace cobalt ions by UV spectroscopy [12], hand-held LIBS analyzer for prediction of element content in soil [13], and Lasso combined with Just-in-time (JIT) framework to solve the problems of online NIRS nonlinearity and multiple operating conditions [14].

Also, Lasso is an effective variable selection method. Study revealed the accuracy and stability of the Matsutake authenticity discrimination model and the edible fungus classification model selected by Lasso were higher than those of the PCA method [15]. The combined interval PLS (siPLS) and Lasso were combined to select the characteristic wavelength of NIRS for monitoring the pH value during the solid-state fermentation of straw feed protein [16]. Lasso combined with logistic regression was used for the selection of UV spectral characteristics of fiber dye classification [17]. Lasso along with a weighted voting strategy was proposed for selecting characteristic variables of NIRS [18].

7.7 Least Angle Regression

Least angle regression (LARS) is a method of solving linear regression and variable selection proposed by Efron in 2004, which is similar to the form of forward stepwise regression. From the perspective of the solution process, it is an efficient solution of the Lasso method [19, 20]. Variables of forward stepwise regression are increasing by one at a time, until there are no variables to introduce. This method has an obvious disadvantage, that is, because there may be a correlation between the respective variables, the selection of subsequent variables may make the previously selected independent variables unimportant. It does not consider removing unimportant variables from the selected variables, and the final “optimal” subset may contain some independent variables that have little effect on the dependent variable. Forward stagewise method is more cautious than the forward stepwise method. The algorithm increases or decreases a small amount on the corresponding coefficient of the selected variable each time, and the other coefficients remain unchanged. This process can be repeated until all residuals are zero or the coefficients are equal to zero. Therefore, this algorithm may require thousands of steps to arrive at the final

model. The least angle regression combines the advantages of these two algorithms, and the amount of calculation is not that large.

For example, the algorithm takes the largest step along the direction of x_1 until another variable x_2 has as much correlation with the current residual. Next, calculation is not along the direction of x_2 but along the equiangular line of the two vectors, until the third variable has as much correlation with the current residual. Then, the algorithm continues along the direction equiangular with the three vectors, that is, the “least angle direction”, until the fourth variable enters the “most relevant set”, and so on. Its equiangularity makes it easier to calculate the step length of the iteration compared to the forward stepwise method.

Suppose a spectral array X composed of n samples, where each spectrum x is composed of p wavelength variables, denoted as $a_i, I = 1, 2, \dots, p$, the corresponding concentration vector is y (dimension is n), b is the regression coefficient vector (dimension is p), and the concentration regression residual vector is r (dimension is n), then the calculation of least angle regression is as follow:

- (1) Center the mean of X and y , and the residual $r = y - \bar{y}$, \bar{y} is the mean value of the concentration vector, and all elements of the regression coefficient vector b are set to 0;
- (2) Select the variable a_i that is most relevant to the residual r ;
- (3) Change the coefficient of the variable a_i from 0 to the least squares coefficient $\langle a_i, r \rangle$, Where $\langle a_i, r \rangle$ is the inner product of a_i and r , until the residual correlation of the new variable a_j is greater than the residual correlation of the variable a_i ;
- (4) The coefficients b_i and b_j corresponding to a_j and a_i are updated together in the direction of adding the least squares estimation of the new variable until a new variable is selected according to the above rules;
- (5) Repeat the operations from (7.2) to (7.4) until all variables are selected, and the final estimate is the solution of the least square method.

Detailed mathematical algorithm of least angle regression can be found in the references [21].

Yan et al. used the LAR to eliminate the collinearity between the variables in the full spectrum to obtain the initially selected wavelengths, then used the GA-PLS method to further optimize the final model [22]. LAR was used to select the characteristic wavelengths of NIRS of citrus leaves, and the accurate detection of citrus yellow dragon disease was realized through the nuclear extreme learning machine [23].

7.8 Elastic Net

Ridge regression cannot make any regression coefficient zero, but can only make it infinitely approach zero, so the model is more difficult to interpret. After the L1 regularization term is introduced, Lasso can not only shrink variables like ridge

regression, but also shrink certain regression coefficients to zero accurately, which greatly improves the explanatory nature of the model. However, in the case of high collinearity, Lasso may forcefully delete a certain predictive feature, which will lose the predictive ability of the model [24].

Based on the ridge regression and the Lasso regression, Zou et al. proposed an elastic network by combining the two regularization methods of L1 norm and L2 norm.

$$\hat{\beta}(\text{Elastic Net}) = \arg \min \left\{ Y - X\beta^2 + \lambda_1 \sum_{j=1}^p |\beta_j| + \lambda_2 \sum_{j=1}^p \beta_j^2 \right\} \quad (7.8)$$

If $\alpha = \lambda_1/(\lambda_1 + \lambda_2)$, $\lambda = \lambda_1 + \lambda_2$,
Above formula can be written as

$$\hat{\beta}(\text{Elastic Net}) = \arg \min \left\{ Y - X\beta^2 + \lambda \left[\alpha \sum_{j=1}^p |\beta_j| + (1 - \alpha) \sum_{j=1}^p \beta_j^2 \right] \right\} \quad (7.9)$$

s.t. $(1 - \alpha)\|\beta\|^2 + \alpha\|\beta\|_1 \leq t, t \geq 0$.

where $t \geq 0$, is the constraint constant, $(1 - \alpha)\beta^2 + \alpha\beta_1$ is the penalty for elastic net, also a convex combination of Lasso penalty and ridge regression penalty. When $\alpha = 0$, the elastic net becomes ridge regression, when $\alpha = 1$, it becomes Lasso regression. Elastic net can effectively handle the situation when the dimension of the feature vector is much larger than the sample size, and automatically select the feature vector with group effect from it.

Zheng et al. used the elastic net in the NIRS modeling process, when the number of variables is much larger than the samples, the elastic net can compress the number of variables to an appropriate degree, select important independent variables that have a significant impact on the response variable, and establish a linear model with better performance [25]. Zhao et al. adopted LASSO and elastic net to reduce the dimensionality of IR spectra of the mixed gas. In the bands where the absorption peaks overlapped severely, the characteristic wavelengths selected by the elastic net were more advantageous [26].

7.9 Principal Component Regression

7.9.1 Theory

The first f score vectors obtained by the principal component analysis of the spectrum matrix X are used to form a matrix $T = [t_1, t_2, \dots, t_f]$, instead of absorbance variables

for MLR regression, then a principal component regression (PCR) model is obtained as $y = T\mathbf{b} + E$. The least square solution of regression coefficient \mathbf{b} is $\mathbf{B} = (T^T T)^{-1} T^T Y$.

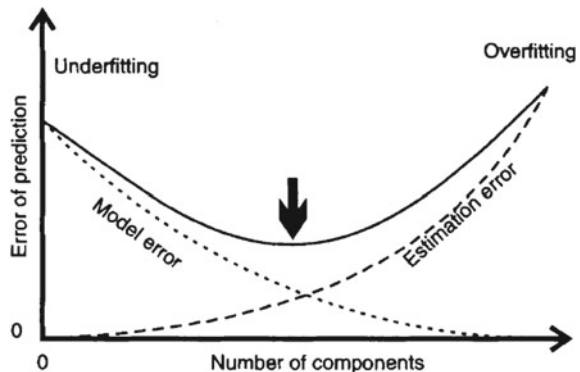
For the spectra \mathbf{x} of the tested samples, first, the score vector is obtained from the loading matrix by principal component analysis: $\mathbf{t} = \mathbf{x}P$. Then, the final result will be obtained through the PCR model \mathbf{b} as $\mathbf{y} = \mathbf{t}\mathbf{b}$.

PCR effectively overcomes the problem of unstable calculation results caused by MLR due to severe collinearity (pathological matrix) between input variables. On the premise of maximizing the useful information in the spectra, it also suppresses the influence of measurement noise on the model by ignoring those secondary principal components and further improves the predictability of models. This method can be applied to the complex systems, and the target components can be predicted more accurately without knowing the presence of specific interfering components.

7.9.2 Method for Selecting the Optimal PCs

In principal component regression (including the PLS introduced in the next section), it is particularly important to determine the optimal number of principal components (PCs) (also known as principal factors) participating in the regression. As shown in Fig. 7.1, if too few PCs were selected (too few features), certain useful information of the original spectra would be lost, as well, the regression fitting would be insufficient, which is called under-fitting. Performance of under-fitting is that the obtained model has poor predictability on the calibration set samples, and the potential rules within the spectral data have not been fully learned. In the regression, besides for the under-fitting caused by the lack of feature variables and the inability to correctly establish the mapping, the inappropriate selection of the regression algorithm (e.g., the linear regression algorithm selected for the nonlinear data set), or the unreasonable setting of modeling parameters can also lead to the occurrence of under-fitting.

Fig. 7.1 Selection of the optimal main components



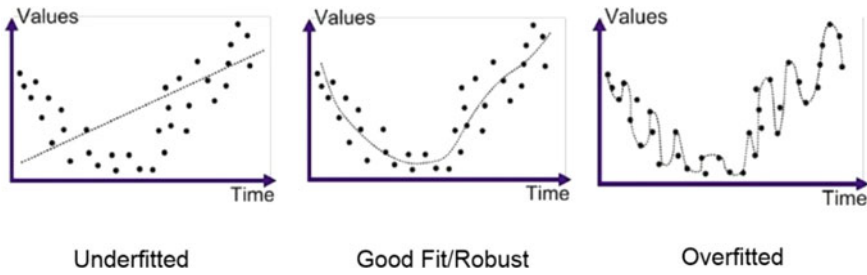


Fig. 7.2 Scheme of under-fitting, moderate fitting, and over-fitting

If too many PCs were selected, the measurement noise would be included too much, and the prediction error of model would increase significantly, which is called over-fitting (Fig. 7.2). Thus, a reasonable determination of the number of PCs participating in the establishment of the model is one of the effective methods to make full use of spectral information and filter out noise. In practice, over-fitting may occur if the number of calibration samples was too small or lacks representativeness, or the algorithm used is not appropriate, etc. That is, the model performs well on the calibration set, but poorly on the testing set.

(1) Leave-one-out cross validation

In spectral analysis, cross validation is the most choice, and prediction residual error sum of squares (PRESS) is the most commonly used criterion. Specific steps are as follows:

For a certain factor f , take only one sample as prediction from n calibration samples, which is leave-one-out cross validation (LOOCV), in other words, $n-1$ samples are used to establish a calibration model and predict the sample to be retained. After repeated modeling and prediction, until the n samples are predicted once and only once, the PRESS value corresponding to this factor f is obtained as $PRESS = \sum_{i=1}^n (y_i - \hat{y}_i)^2$. Relationship between the standard error of cross validation (SECV) and the PRESS value is $SECV = \sqrt{\frac{PRESS}{n-1}}$, where the smaller the PRESS or SECV value, the better the predictability of model.

Generally, the method of plotting PRESS value against the PCs (called the PRESS diagram as Fig. 7.3) is often used to establish the optimal number of PCs. The theoretical PRESS graph usually shows a decreasing trend with the increase of the PCs. When PRESS reaches the lowest point, a slight rise or fluctuation shows up, indicating that after this point, the added PCs is a noise component that has nothing to do with the measured component. Corresponding to the lowest point of PRESS graph is the optimal number of PCs. If there was no minimum value, the first point when PRESS value reaches a fixed level can be regarded as the optimal PCs. However, in some cases, such as narrow sample distribution, relatively weak information, or abnormal samples, a non-ideal PRESS graph may appear. In this case, the SECV can be compared with the repeatability standard deviation of the reference method. If

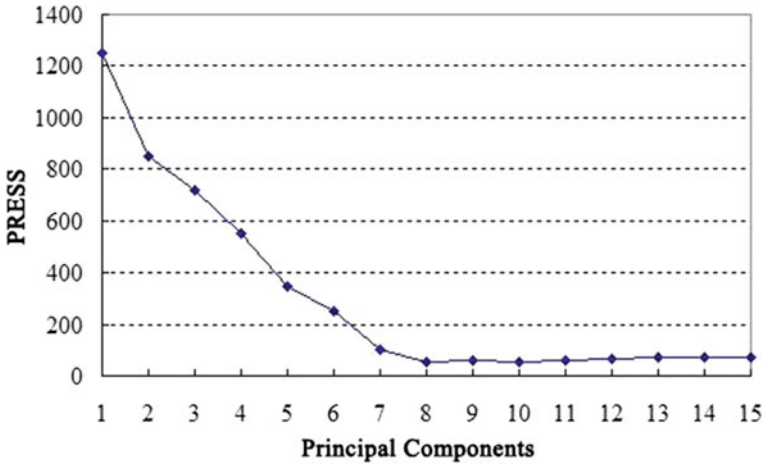


Fig. 7.3 PRESS graph obtained by leaving-one-out cross validation

the SECV is significantly below the repeatability standard deviation of the reference method, it indicates that the model is likely to be over-fitting.

In fact, the inflection point of PRESS graph is not very obvious, which brings difficulties to the selection of the optimal PCs. Thus, the F test can be introduced [27]. Suppose the PCs corresponding to the minimum PRESS value is r^* , and calculate $F(r) = \text{PRESS}(r)/\text{PRESS}(r^*)$, $r = 1, 2, \dots, r^*$, when $F(r) < F_{\alpha, n}$, the corresponding minimum r is the optimal PCs. Where, $F_{\alpha, n}$ is the critical value of F when the confidence degree is $(1-\alpha)$ and the degree of freedom is n (n is the number of samples in the calibration set). Usually, the value of α is 0.25.

(2) **Multi-fold cross validation**

In many circumstances, LOOCV probably overemphasizes the calibration samples and selects too many PCs, resulting in an over-fitting situation [28]. For the case where calibration samples are not adequate, the “leave more out” cross validation can be tried. Take the leave three out as an example, that is, three samples are selected for validation from n calibration samples for each modeling run, and the remaining $(n-3)$ samples are used to build the calibration model, that requires C_n^3 runs of modeling to complete all possible three validation samples combinations. Given 100 calibration samples, the number of cross validations required would be 161,700 runs. Moreover, leaving out one more sample, the number of cross validations will increase exponentially.

When the number of calibration samples is large (over 800 samples), the multi-fold cross validation (MFCV) can be practiced. It randomly breaks the order of calibration samples, and divides them into m groups, then use $(m-1)$ samples to predict the left out group samples when modeling. Optimized PCs are selected by the PRESS curve.

Although each sample has been traversed and predicted once in the multi-fold validation, this strategy still has many missing-test samples. For example, using the “ten-fold” validation, assuming there are 100 calibration samples, then the number of samples per fold is 10, and the number of cross-validation only needs to be 10 runs. But in fact, there are C_{100}^{10} ways to select 10 from 100 samples, and the “ten-fold” validation only selects 10 of them. Therefore, factors such as the way samples grouped and outliers will affect the final result.

(3) Monte Carlo method

In addition to LOOCV and MFCV, there are also Bootstrap and Monte Carlo cross validation for selecting the number of PCs [29]. Basic idea of the bootstrap is to randomly select samples from the entire calibration set with “send-back” replacement to form a new calibration set, the number of which is the same as the original set. Repeat this step several times to get several bootstrap calibrations, and use them to build models, respectively. Then the original calibration is employed for prediction, and the mean value of the corresponding prediction error is regarded as the parameter for selecting PCs.

Monte Carlo method randomly selects m samples from the calibration samples to establish the calibration model, uses the remaining $(n-m)$ samples as the prediction set. Sampling is repeated several times, mean value of the prediction error is taken as the parameter for selecting PCs. It is generally believed that the sampling frequency of Monte Carlo method is n^2 (n is the number of calibration samples), which can ensure the representativeness and unbiasedness of sampling. Compared with LOOCV, the Monte Carlo method pays more attention to prediction (usually 75% for calibration, 25% for prediction), which effectively avoids the possible over-fitting problem (too many PCs). Meanwhile, it avoids the exponential increase in modeling times caused by the leave more cross validation.

In order to further evaluate the influence of prediction on the selection of PCs, Filzmoser et al. proposed a strategy of repeated double cross validation [30]. As shown in Fig. 7.4, this method first divides the calibration samples into new calibration samples and test samples through Monte Carlo sampling, and then performs Monte Carlo again on the new calibration samples to obtain training samples and validation samples. It predicts the test samples based on the optimal model established by the training set, repeats sampling several times, and selects the optimal number of PCs according to the distribution of RMSEP of the obtained test sets.

It is worth noting that, for calibration sets with high spectral quality, high accuracy of concentration, uniform sample distribution, sufficient number, and no outlier, the number of optimal PCs selected by the above method usually does not have significant difference.

(4) Sum of ranking differences

In order to overcome the problem of over-fitting due to the selection PCs and other modeling parameters, Gowen et al. suggested that in the process of determining model parameters, not only indicators that characterize model bias (as RMSECV)

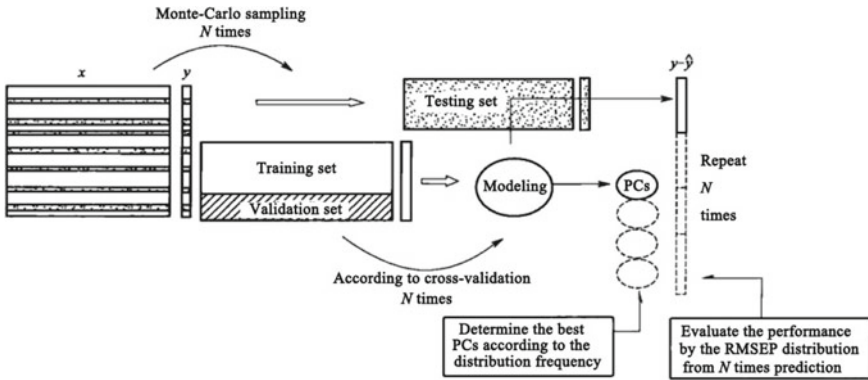


Fig. 7.4 Scheme of the number of PCs selected by the repeated double cross-validation strategy

should be considered, but also indicators that characterize model variances (2-norm of regression coefficient) should be involved [31–33].

Kalivas et al. proposed to use the sum of ranking differences (SRD) algorithm to determine parameters of multivariate calibration model by combining the indicators that characterize the model deviation and model variance [34]. It takes all the possible parameters and evaluation indicators (bias or variance) as the input of the SRD matrix, and then selects a model that is agreed upon by the evaluation indicators of each model according to the SRD algorithm, whose corresponding parameters are the final modeling parameters. In recent years, the SRD algorithm has been used to compare the pros and cons of calibration models, the identification of outliers, and the evaluation of spectral experiments [35, 36], etc.

7.9.3 Partial Least Squares Regression

In PCR, only the spectral array X is decomposed to eliminate noise information. Similarly, the concentration matrix Y also contains useless information, which should be treated in the same way, and the influence of the concentration matrix Y should be considered when decomposing the spectral matrix X . Partial least squares (PLS) is a multiple factor regression method based on the above ideas [37].

PLS first decomposes the spectral matrix X and the concentration matrix Y , and its model is as follows:

$$Y = UQ^T + E_Y = \sum_{k=1}^f u_k q_k^T + E_Y \tag{7.10}$$

$$X = TP^T + E_X = \sum_{k=1}^f t_k p_k^T + E_X \tag{7.11}$$

where $t_k(n \times 1)$ is the score of the k -th PCs of the absorbance matrix X ; $p_k(1 \times m)$ is the loading of the k -th PCs of the absorbance matrix; $u_k(n \times 1)$ is the score of the k -th PCs of the concentration matrix Y ; $q_k(1 \times m)$ is the loading of the k -th PCs of the concentration matrix Y ; f is the number of PCs. Thus, T and U are the score matrices of X and Y , P and Q are the loading matrices of X and Y , E_X and E_Y are the PLS fitting residual matrix of X and Y , respectively.

The second step of PLS is to conduct linear regression of T and U .

$$U = TB$$

$$B = (T^T T)^{-1} T^T Y$$

When predicting, first the score $T_{(\text{unknown})}$ of the unknown sample spectral matrix $X_{(\text{unknown})}$ needs to be solved according to P , and then the predicted concentration value is obtained from the following equation: $Y_{(\text{unknown})} = T_{(\text{unknown})} BQ$.

In the actual PLS calculation, PLS combines matrix decomposition and regression into one step, in other words, the decomposition of X and Y matrices are performed at the same time. Before calculating each new PC, the score T of X is exchanged with the score U of Y , so that the PC of X is directly related to Y . Thus, when PLS calculates the PCs, it needs to consider the calculated variance of PCs to be as large as possible, while also making PCs and concentration related to the greatest extent. The maximization of variance is to extract as much useful information as possible, and the maximum correlation with the concentration is to make the best use of the linear relationship between spectral variables and concentration. So it covers the shortcomings of PCR only decomposing X .

PLS is calculated by the nonlinear iterative partial least squares algorithm (NIPALS) proposed by Wold. The specific algorithm is as follows. In the calibration, the residual matrix E is ignored, and the PC is as 1, then:

For $X = tp^T$, multiply the left side by t^T to get $p^T = t^T X / t^T t$; multiply the right side by p to get $t = Xp / p^T p$.

For $Y = uq^T$, multiply the left side by u^T to get $q^T = u^T Y / u^T u$; divide both sides by q^T to get $u = Y / q^T$.

- (1) Find the weight vector w of the absorbance matrix X

Take a column of the concentration matrix Y as the initial iterative value of u , replace t with u , and calculate w .

$$\text{Equation is } X = uw^T, \text{ the solution is } w^T = u^T X / u^T u.$$

- (2) Normalize the weight vector w

$$w^T = w^T / \|w^T\|$$

- (3) Find the factor score t of the absorbance matrix X

Calculate t from the normalized w by equation: $X = tw^T$.

$$T = Xw / w^T w.$$

- (4) Find the loading q value of the concentration matrix Y

Use t instead of u to calculate q by equation of $Y = tq^T$.

$$q^T = t^T Y / t^T t.$$

- (5) Normalize the loading vector
- q

$$q^T = q^T / \|q^T\|$$

- (6) Find the factor score
- u
- of the concentration matrix
- Y

Calculate u from q^T by equation of $Y = uq^T$.

$$U = Yq/q^Tq.$$

- (7) Then replace
- t
- with this
- u
- and return to step (1) to calculate
- w^T
- , and calculate
- t_{new}
- from
- w^T
- , iteratively. If
- t
- has converged as
- $\|t_{\text{new}} - t_{\text{old}}\| \leq 10^{-6}\|t_{\text{new}}\|$
- , go to step (8) for calculation, otherwise return to Step (1).

- (8) Find the loading vector
- p
- of absorbance matrix
- X
- from the converged
- t
- by equation
- $X = tp^T$
- .

$$p^T = t^TY/t^Tt$$

- (9) Normalize the loading
- p

$$p^T = p^T / \|p^T\|$$

- (10) Standardized the factor score
- t
- of
- X

$$t = t / \|p\|$$

- (11) Standardized the weight vector
- w

$$w = w / \|p\|$$

- (12) Calculate the relationship between
- t
- and
- u

$$b = u^T t / t^T t$$

- (13) Calculate the residual matrix
- E

$$E_X = X - tp^T$$

$$E_Y = Y - btq^T$$

- (14) Replace
- X
- with
- E_X
- , replace
- Y
- with
- E_Y
- , and return to step (1), by analogy, find the
- w
- ,
- t
- ,
- p
- ,
- u
- ,
- q
- , and
- b
- of
- X
- and
- Y
- . Finally, the optimal PCs
- f
- is determined by cross validation.

As for the unknown sample x_{unknown} , the prediction process is as follows:

- (a) Let
- $h = 0$
- ,
- $y_{\text{unknown}} = 0$
- ;

- (b) Set
- $h = h + 1$
- , and calculate

$$t_h = x_{\text{un}} w_h^T$$

$$y_{\text{unknown}} = y_{\text{unknown}} + b_h t_h q_h^T$$

$$x_{\text{unknown}} = x_{\text{unknown}} - t_h p_h^T$$

- (c) If
- $h < f$
- , go to step (2), otherwise, stop the calculation, and the final
- y_{unknown}
- is the predicted value.

Besides, the unknown sample x_{unknown} can also directly calculate the predicted value through the following equation.

$y_{un} = \mathbf{b}_{PLS}x_{un}$, where $\mathbf{b}_{PLS} = \mathbf{w}^T (\mathbf{p}\mathbf{w}^T)^{-1} \mathbf{q}$, and it is the regression coefficient of PLS.

The PLS regression can often be grouped in PLS1 and PLS2. In fact, PLS1 and PLS2 share the same algorithm. The difference is that PLS1 only calibrates one component at a time, while PLS2 can simultaneously calibrate regression for multiple components.

PLS2 uses the same set of score T and loading matrix P when calibrating all components, obviously, the T and P obtained in this way are not optimal for the concentration vector used in Y . Especially for complex systems, it will significantly reduce the prediction accuracy.

In PLS1, the calibrated T and P are optimized for each concentration vector in Y . When the concentration of different components in the calibration set varies greatly, for example, the concentration range of one component is 50% to 70%, and other one is 0.1–1.0%, prediction results of PLS1 are usually better than PLS2 and PCR because it is optimized for each tested component. In spectral analysis, if not specifically noted, PLS usually refers to the PLS1 method.

From the above introduction, it can be concluded that MLR, PCR, and PLS are actually connected and coherent, and there is a gradually developed course of the linear multivariate calibration. Overcoming the weaknesses of MLR sub-rank inversion and insufficient use of spectral information, PCR uses PCA to decompose the spectral array X , and performs MLR regression by the score vector, significantly enhancing the model prediction. PLS decomposes the spectra X and concentration Y simultaneously, strengthens the corresponding calculation relationship between the two matrices, assuring the best calibration model established. Up to now, PLSR is a perfect combination of MLR, canonical correlation analysis (CCA), and PCA. This is why PLS is most widely used in spectral multivariate calibration analysis.

References

1. Mark H, Workman J. Chemometrics in spectroscopy, 2nd ed. Elsevier;2018.
2. Adams MJ. Chemometrics in analytical spectroscopy, 2nd ed. UK, Cambridge: RSC;2004.
3. Melgaard DK, Haaland DM, Wehlburg CM. Concentration residual augmented classical least squares (CRACLS): a multivariate calibration method with advantages over partial least squares. Appl Spectrosc. 2002;56(5):615–24.
4. Darwish HW, Metwally FH, Bayoumi AE, et al. Artificial neural networks and concentration residual augmented classical least squares for the simultaneous determination of diphenhydramine, benzonatate, guaifenesin and phenylephrine in their quaternary mixture. Trop J Pharm Res. 2014;13(12):2083–90.
5. Saeys W, Beullens K, Lammertyn J, et al. Increasing robustness against changes in the interferent structure by incorporating prior information in the augmented classical least-squares framework. Anal Chem. 2008;80(13):4951–9.
6. Hegazy MA, Abdelwahab NS, Ali NW, et al. Comparison of two augmented classical least squares algorithms and PLS for determining nifuroxazide and its genotoxic impurities using UV spectroscopy. J Chemom. 2019;33:e3190.
7. Gemperline P. Practical guide to chemometrics, 2nd ed. CRC Press;2006.

8. Brown SD, Tauler R, Walczak B. *Comprehensive chemometrics*, 2nd ed. Elsevier;2020.
9. Zhang M, Liu XH, He XK. Study on the application of ridge regression to near-infrared spectroscopy quantitative analysis and optimum wavelength selection. *Spectrosc Spectr Anal.* 2010;30(5):1214–7.
10. Wu T, Fang GG, Liang L, et al. Four kinds of algorithms used for the determination of pulpwood properties by near infrared spectroscopy. *Chem Indus For Prod.* 2016;36(6):63–70.
11. Zhu H, Wu T, Fang GG, et al. Analysis of extractives content of guangxi fast-growing eucalyptus and models optimization based on near-infrared technique. *Spectrosc Spectr Anal.* 2020;40(3):793–8.
12. Zhu HQ, Zhou T, Li YG, et al. An ultraviolet-visible absorption spectrometric method for detection of zinc(II) and cobalt(II) ions concentration based on boosting modeling. *Chin J Anal Chem.* 2019;57(4):576–82.
13. Erler A, Riebe D, Beitz T, et al. Soil nutrient detection for precision agriculture using handheld laser-induced breakdown spectroscopy (LIBS) and multivariate regression methods (PLSR, Lasso and GPR). *Sensors.* 2020;20(2):418.
14. Liu J, Luan XL, Liu F. Adaptive jit-lasso modeling for online application of near infrared spectroscopy. *Chemom Intell Lab Syst.* 2018;183:90–5.
15. Li YQ, Pan TH, Li HR, et al. NIR spectral feature selection using lasso method and its application in the classification analysis. *Spectrosc Spectr Anal.* 2019;39(12):3809–15.
16. Mei CL, Chen Y, Yin L, et al. Wavelength selection by siPLS-LASSO for NIR spectroscopy and its application. *Spectrosc Spectr Anal.* 2018;38(2):436–40.
17. Rich DC, Livingston KM, Morgan SL. Evaluating performance of lasso relative to PCA and LDA to classify dyes on fibers. *Forensic Chem.* 2020;18:100213.
18. Zhang RQ, Zhang FY, Chen WC, et al. A variable informative criterion based on weighted voting strategy combined with lasso for variable selection in multivariate calibration. *Chemom Intell Lab Syst.* 2019;184:132–41.
19. Hesterberg T, Choi NH, Meier L, et al. Least angle and L1 regression: a review. *Stat Surv.* 2008;2:61–93.
20. Hastie T, Taylor J, Tibshirani R, et al. Forward stagewise regression and the Monotone Lasso. *Electron J Stat.* 2007;1:1–29.
21. Efron B, Hastie T, Johnstone I, et al. Least angle regression. *Ann Stat.* 2004;32(2):407–99.
22. Yan SK, Yang HH, Hu BC, et al. Variable selection method of NIR spectroscopy based on least angle regression and GA-PLS. *Spectrosc Spectr Anal.* 2017;37(6):1733–8.
23. Chen WL, Wang QB, Lu HX, et al. Identification of citrus huanglongbing by near infrared spectroscopy with least angle regression and kernel extreme learning machine. *J Instrum Anal.* 2020;38(10):1267–74.
24. Zou H, Hastie T. Regularization and variable selection via the elastic net. *J Roy Stat Soc B.* 2005;67:301–20.
25. Zheng NN, Luan XL, Liu F. Elastic net modeling for near infrared spectroscopy. *Spectrosc Spectr Anal.* 2018;38(10):114–8.
26. Zhao AX, Tang XJ, Song Y, et al. Spectral wavelength selection and dimension reduction using elastic net in spectroscopy analysis. *Infrared Laser Eng.* 2014;43(6):1977–81.
27. Haaland DM, Thomas EV. Partial least-squares methods for spectral analyses. 1. relation to other quantitative calibration methods and the extraction of qualitative information. *Anal Chem.* 1988;60(11):1193–202.
28. Martens HA, Dardenne P. Validation and verification of regression in small data sets. *Chemom Intell Lab Syst.* 1998;44:99–121.
29. Xu QS, Liang YZ, Du YP. Monte Carlo cross-validation for selecting a model and estimating the prediction error in multivariate calibration. *J Chemom.* 2004;18(2):112–20.
30. Filzmoser P, Liebmann B, Varmuza K. Repeated double cross validation. *J Chemom.* 2009;23:160–71.
31. Gowen AA, Downey G, Esquerre C, et al. Preventing over-fitting in PLS calibration models of near-infrared (NIR) spectroscopy data using regression coefficients. *J Chemom.* 2011;25(7):375–81.

32. Kalivas JH, Palmer J. Characterizing multivariate calibration tradeoffs (bias, variance, selectivity, and sensitivity) to select model tuning parameters. *J Chemom.* 2014;28(5):347–57.
33. Faber NM. A Closer Look at the Bias-Variance Trade-off in Multivariate Calibration. *J Chemom.* 2015;13(2):185–92.
34. Kalivas JH, Heberger K, Andries E. Sum of ranking differences (SRD) to ensemble multivariate calibration model merits for tuning parameter selection and comparing calibration methods. *Anal Chim Acta.* 2015;869:21–33.
35. Heberger K. Sum of ranking differences compares methods or models fairly. *Trends Anal Chem.* 2010;29(1):101–9.
36. Brownfield B, Kalivas JH. Consensus outlier detection using sum of ranking differences of common and new outlier measures without tuning parameter selections. *Anal Chem.* 2017;89:5087–94.
37. Geladi P, Kowalski BR. Partial least-squares regression: a tutorial. *Anal Chim Acta.* 1985;185(1):1–17.

Chapter 8

Nonlinear Calibration Methods



8.1 Artificial Neural Network

8.1.1 Introduction

Artificial neural network (ANN) is based on simulations of the structure of the human brain. It stores both the storage and calculation of information in neural units. From certain perspective, neural network can simulate the activity process of human brain nervous system. It has the ability of self-learning, self-organization, self-adaptation, strong fault-tolerant ability, distributed storage and parallel processing of information, and highly nonlinear expression, which is not available in other traditional multivariate calibration methods. More and more analytical chemists begin to use the ANN method to solve problems in analytical chemistry, such as nonlinear multivariate calibration, pattern recognition, QSAR, and spectral library retrieval.

(1) Biological neural network

There are approximately 10^{11} – 10^{12} biological neurons in the human brain. Each neuron is connected to approximate 10^3 – 10^5 neurons. These neurons are connected by about 10^{15} connections into a vast and complex network system. The neuron has an independent ability to receive, process, and transmit electrochemical signals, which is through the neural pathways that constitute the brain's transmission system. Figure 8.1 shows the typical structure of biological neurons and their interconnection.

Biological neurons are structurally composed of cyton, dendrite, axon, and synapse, which are used to receive, transmit and process information between neurons. Dendrites receive input from other neurons, and axons provide output to other neurons. Electrochemical signals between neurons pass through their surface, and these neuron-to-neuron connections are called synapses. On the receiving side of the synapse, signals are sent into the cyton, where they are combined. Some of the input signals act as stimuli and some act as inhibitions. When the cumulative stimuli received in the cyton exceed a threshold, the cyton is excited, and it sends signals along the axon through the branches and tendrils to other neurons.

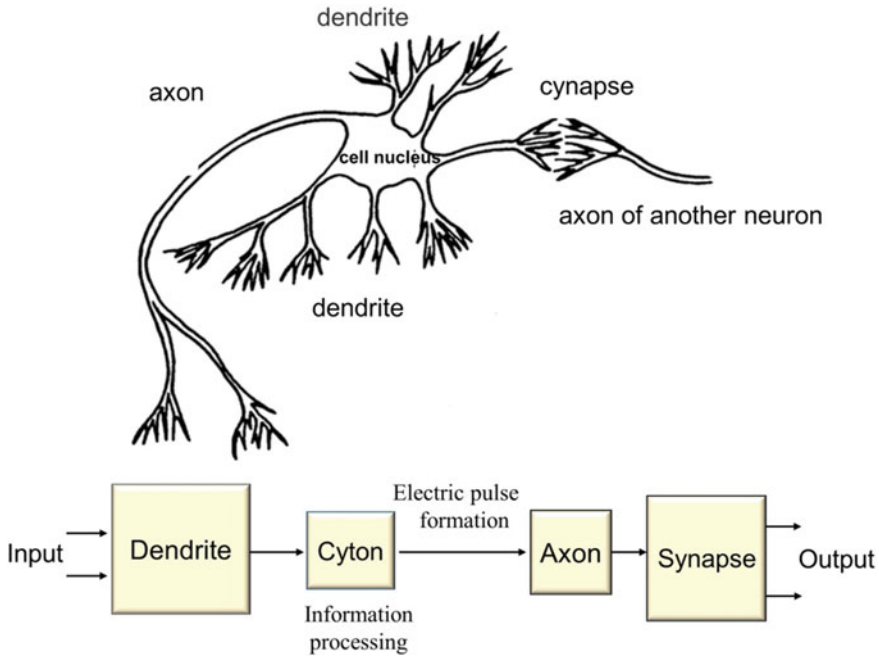


Fig. 8.1 Schematic diagram of typical biological neurons

In this system, each neuron is connected with synapses to many other neurons in the system. It is thought that the same signal that the same neuron bursts through its axis may have different effects on different neurons that receive it, depending on the corresponding synapse. The greater the synaptic “connection strength”, the stronger the received signal, and vice versa, the weaker the synaptic “connection strength”, the weaker the received signal. Synaptic neurons combine the input signals of each synaptic point in some way and trigger the output signals under certain conditions, which is transmitted to other neurons through the axon. It can be seen that the basic structural and functional unit of the biological nervous system is the biological neuron (i.e., nerve cell), which has the functions of receiving, processing, and outputting information. A large number of biological neurons are connected with each other through synapses, forming a complex information transmission network system. ANN is designed according to the inspiration obtained from the biological neural network.

(2) Neuron

Neuron, also known as node, is the most basic unit of neural network. Artificial neuron is a kind of approximation of biological neuron in function. It simulates the processing process of the input signal of biological neuron to some extent, and its characteristics determine the overall characteristics of neural network to a certain extent. For each artificial neuron, it can receive a set of input signals from other

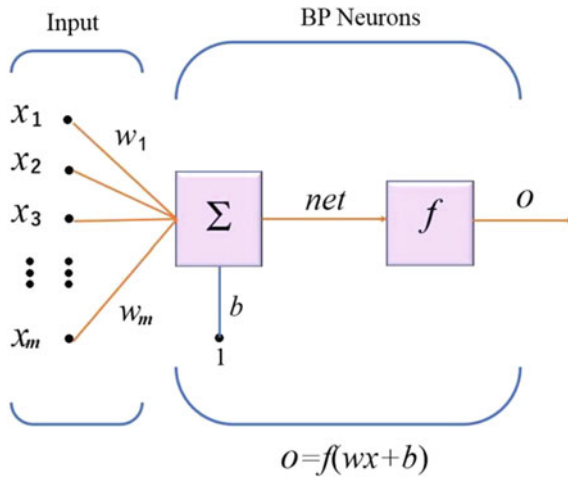


Fig. 8.2 Model of an artificial neuron

neurons in the system. Each input corresponds to a weight, and the weighted sum of all the inputs determines the activation state of the neuron. Here, each weight corresponds to the “connection strength” of the synapse. The interconnection of a large number of simple neurons constitutes the neural network system, which has powerful functions of information processing and computation.

The artificial neuron model can be described in Fig. 8.2, which is mainly composed of the following five parts.

① Import. $x_1, x_2, x_3, \dots, x_m$ represents m input variables of the neuron.

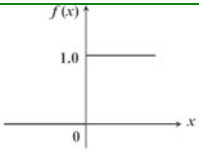
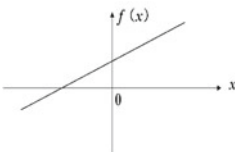
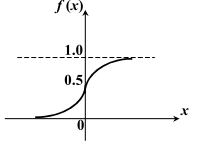
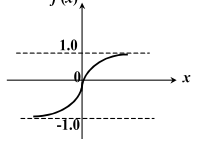
② Network weights and thresholds. $w_1, w_2, w_3, \dots, w_m$ are the network weights, which represents the connection strength between the input variable and the neural network. b is the threshold value or the bias value of the neuron. The introduction of the bias value can make the transfer function move left and right, and improve the possibility and ability to solve practical problem. These two parameters are dynamically tunable.

③ Sum units. The summation unit performs the weighted summation of input variables, which is the first step in the processing of input signals by a neuron.

$$net = \sum_{i=1}^m x_i w_i + b \tag{8.1}$$

④ Transfer function. f represents the transfer function of the neuron, or the excitation function, the transmission function, the action function, etc. It is used to perform functional operation on the calculation result of the summation unit to get the output of the neuron. This is the second process of processing the input variable by the neuron. Table 8.1 shows several typical neuron transfer functions.

Table 8.1 Forms of several typical neuron transfer functions

function	expression	curve
threshold function	$f(x) = \begin{cases} 1 & x \geq 0 \\ 0 & x < 0 \end{cases}$	
linear function	$f(x) = kx + b$	
logarithmic Sigmoid function	$f(x) = \frac{1}{1 + e^{-x}}$	
tangent Sigmoid function	$f(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$	

⑤ Output. After the weighted sum of the input variables and the transformation of the transfer function, the final output is

$$o = f(wp + b) \tag{8.2}$$

In ANN, neurons are often referred to as “processing units” or sometimes “nodes” from the point of view of the network.

(3) The main connection patterns of neural networks

Neural network system is a highly interconnected complex nonlinear system. There are many connections between neurons. According to the topological structure of the network, the neural network structure can be divided into two categories: hierarchical structure and interconnected structure. The hierarchical structure of neural network divides neurons into input layer, intermediate layer (hidden layer), and output layer according to their function and order. Each neuron of the input layer is responsible for receiving input information from the outside world and transmitting it to the neurons of the hidden layer. The hidden layer is the internal information processing

layer of the neural network, which is responsible for information transformation. It can be designed as one or more layers according to the needs. The last hidden layer transmits information to the neuron of the output layer for further processing and then outputs the result of information processing to the outside world. However, in the interconnected network structure, there may be a connection path between any two nodes, so the interconnected network can be subdivided into three types according to the connection degree of nodes in the network: fully interconnected, local interconnected, and sparse connected.

According to the connection orientation (or information flow direction), neural networks can be divided into two types: feedforward network and feedback (recursive) network. The structure of the simple feedforward network is the same as that of the hierarchical network. The feedforward network is named because the direction of network information processing is from the input layer to the hidden layer and then to the output layer. In the feedforward network, the output of the previous layer is the input of the next layer, and the information processing has the directivity of transferring layer by layer, and there is a general no feedback loop. Therefore, this kind of network can easily be linked together to build a multi-layer feedforward network. Feedforward networks include multi-layer perceptron (MLP) and learning vector quantization (LVQ) networks. The structure of feedback network is the same as that of single-layer fully interconnected network. The output of the neuron is fed back to the same or anterior neuron, and the signal can flow forward and backward. Therefore, all nodes in the feedback network have information processing function, and each node can not only receive input from the outside world, but also output to the outside world. Hopfield network and Elman network are representative recursive networks.

(4) Learning methods of neural networks

Learning method is the main symbol of ANN intelligent characteristics. Because of the learning algorithm, ANN has the ability of self-adaptation, self-organization and self-learning. Currently, there are many learning methods of neural network, which can be classified into supervised learning, unsupervised learning, reinforcement learning, and other categories according to whether there is a tutor or not. In the supervised learning, the output of the network is compared with the expected output (namely the teacher signal), and then the weight of the network is adjusted according to the difference between the two, and the difference is finally reduced. The typical representative of this method is BP-ANN. In the unsupervised learning, the input mode is put into the network, and the network automatically adjusts the weight according to the preset rules (such as competition rules). There is no need to train the known samples, so that the network finally has the function of pattern classification, such as Kohonen neural network and Hopfield model. Reinforcement learning is a kind of learning style in between the above two.

8.1.2 Back Propagation-Artificial Neural Network

Among many neural networks, back propagation-artificial neural network (BP-ANN) is the most widely used method. According to statistics, BP network is used in more than 80% of neural network applications. It is the most representative and widely used network model. BP network is a kind of feedforward multi-layer neural network composed of nonlinear transformation neural units. The transfer function used by its neurons is usually Sigmoid differentiable function, which can realize arbitrary nonlinear mapping between input and output, and has excellent nonlinear mapping approximation and generalization (prediction) ability. In spectral analysis, BP neural network has been used to establish nonlinear calibration model with large sample size.

BP neural network consists of three parts: input layer, hidden layer, and output layer. Figure 8.3 depicts the topology of a typical BP neural network, with circles representing neurons. The data is input from the input layer, processed by standardization, and transmitted to the second layer with weights, namely, the hidden layer. The hidden layer transmits the data to the output layer after calculating the weight, threshold, and excitation function. The output layer gives the predicted value of the neural network and compares it with the expected value. If there is an error, the error will be propagated back from the output and the weight and threshold will be adjusted to make the network output gradually consistent with the expected output.

BP algorithm consists of four processes: the input mode from the input layer through the middle layer to the output layer of the “mode forward propagation” process. The error signals between the expected output and the actual output are corrected layer by layer from the output layer through the middle layer to the input layer in the process of “error backward propagation”. The network “memory training” process is carried out repeatedly alternately by “mode forward propagation” and

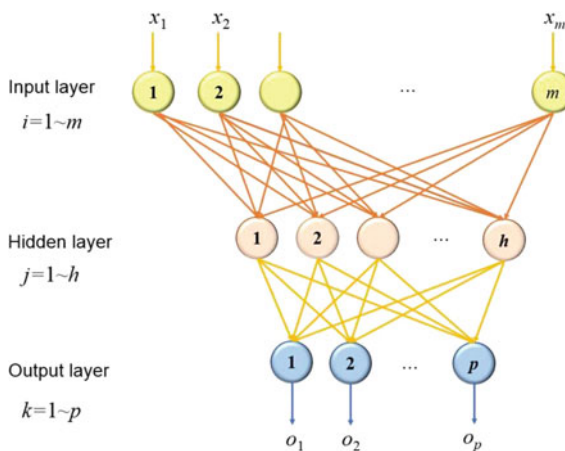


Fig. 8.3 Topological structure of typical BP neural network

“error backward propagation”. The network tends to converge, that is, the process of “learning convergence” in which the global error of the network tends to the minimum.

The standard BP learning algorithm is a gradient descent algorithm, that is, the weight and threshold of the network are adjusted along the negative gradient direction of network error change. Finally, the network error reaches a minimum value (the error gradient at this point is zero).

Generally speaking, in network training, the least square function is used as the error function or the objective function, i.e.,

$$E = \sum_{r=1}^n \sum_{k=1}^p (y_{rk} - o_{rk})^2 \tag{8.3}$$

where o_k is the output value of node k , y_k is its corresponding expected output value, p is the number of nodes in the output layer, and n is the number of training samples. As shown in Fig. 8.4, in the BP algorithm, the weight correction calculation is first carried out from the output layer, and then the weight correction of the hidden layer is carried out.

The algorithm of standard BP network is as follows:

- (1) The initial weight in the range (0, 1) is given by a random number.
- (2) Input the vector of the sample into the input layer.
- (3) Forward information transmission is calculated from Eqs. 8.4–8.6 below.

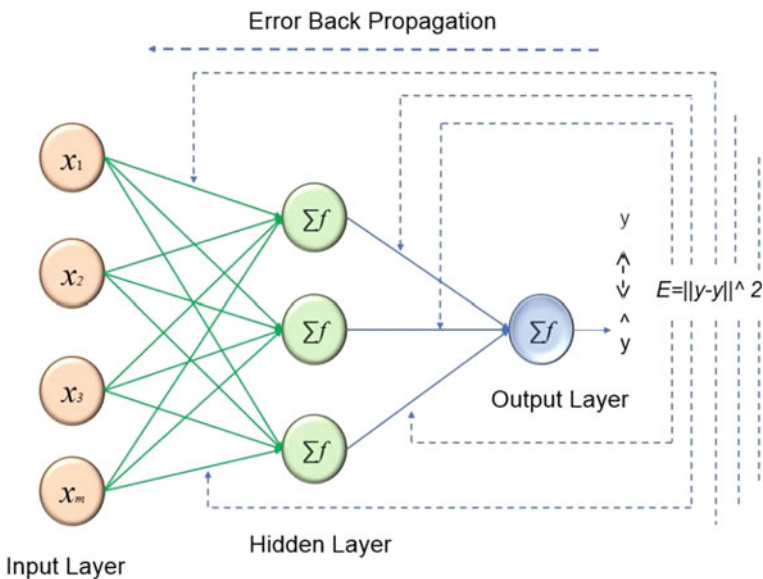


Fig. 8.4 Schematic diagram of error back propagation algorithm

①: Output of the hidden layer

$$g_j = \frac{1}{1 + e^{-net_j}} \quad (8.4)$$

Among them

$$net_j = \sum_{i=1}^m w_{ij}x_i + b_j \quad (8.5)$$

where $i = 1, 2, \dots, m$, m is the number of nodes in the input layer, $j = 1, 2, \dots, h$, h is the number of nodes in the hidden layer, w_{ij} is the connection weight between input layer node i and hidden layer node j .

②: Output of the output layer

$$o_k = \frac{1}{1 + e^{-net_k}} \quad (8.6)$$

Among them

$$net_k = \sum_{j=1}^h v_{jk}g_j + b_k \quad (8.7)$$

where $k = 1, 2, \dots, p$, p is the number of nodes in the output layer, v_{jk} is the connection weight between hidden layer node j and output layer node k .

③: Error

$$E = \sum_{i=1}^n \sum_{j=1}^p (y_{ij} - o_{ij})^2 \quad (8.8)$$

where n is the sample size.

(4) Calculate the error parameters δ of the output layer and the hidden layer.

$$\delta_k = (y_k - o_k)f'(net_k) \quad (8.9)$$

Among them, if the transfer function is logarithmic sigmoid function,

$$f'(net_k) = f(net_k)[1 - f(net_k)] \quad (8.10)$$

$$\delta_k = (y_k - o_k) \cdot o_k \cdot (1 - o_k) \quad (8.11)$$

where $k = 1, 2, \dots, p$, p is the number of nodes in the output layer.

Error parameters of hidden layer

$$\delta_j = \left(\sum_k \partial_k w_{kj} \right) f'(net_j) \tag{8.12}$$

If the transfer function is logarithmic sigmoid function

$$f'(net_j) = f(net_j)[1 - f(net_j)] \tag{8.13}$$

then

$$\delta_j = \left(\sum_k \delta_k v_{kj} \right) \cdot g_j \cdot (1 - g_j) \tag{8.14}$$

where $j = 1, 2, \dots, h$, h is the number of nodes in the hidden layer, $k = 1, 2, \dots, p$, p is the number of nodes in the output layer.

(5) The connection weight between hidden layer node j and output layer node k is adjusted.

$$v_{jk}(l + 1) = v_{jk}(l) + \eta \delta_k g_j \tag{8.15}$$

Connection weight between input layer node i and hidden layer node j is calculated.

$$w_{ij}(l + 1) = w_{ij}(l) + \eta \delta_j x_i \tag{8.16}$$

where η is learning rate (i.e., step length), which determines the rate of training (iteration), $(l + 1)$ is the number of iterations in training.

(6) Repeat Steps 2–5 to calculate the next training sample.

(7) For the training samples used, the iteration is stopped when the error reaches a predetermined value. One weight training for all samples in the training set is called one iteration. Generally, it takes hundreds of iterations (100–5000) to minimize the error, and it is better to randomly select training samples for each iteration.

In order to speed up the iterative process and prevent the oscillation of the iterative process, a learning algorithm with momentum factor can be adopted to add a “momentum” item to the weight modification value.

$$\Delta w(l + 1) = \eta \delta o + \alpha \Delta w(l) \tag{8.17}$$

where $\alpha \Delta w(l)$ is the momentum term (or inertia term), and the initial value of the momentum factor α is usually set to 0.9.

The standard BP learning algorithm is gradient descent algorithm, that is, the weight and threshold of the network are adjusted along the negative gradient direction of network error change, and finally the network error reaches a minimum value (the error gradient at this point is zero). Gradient descent learning algorithm has inherent disadvantages such as slow convergence speed and easy to fall into local minimum. Therefore, there are many improved fast algorithms, which can be divided into two main categories in terms of improvement ways. One is to use heuristic learning method, such as the learning algorithm with momentum factor mentioned above, variable learning rate learning algorithm, “elastic” learning algorithm, etc. The other is to adopt more effective numerical optimization algorithms, such as conjugate gradient learning algorithm, Quasi-Newton algorithm, and Levenberg-Marquardt (L-M) optimization algorithm. At present, in the establishment of spectral quantitative model, most of the L-M optimization algorithm is used. This learning algorithm can effectively prevent the network from falling into local minimum and increase the reliability of BP algorithm.

8.1.3 Design of BP-ANN

When BP-ANN is used for modeling, the following network parameters need to be selected and set.

- (1) Input and output variables: the scores of principal component analysis (PCA) or partial least squares (PLS) are usually used as input variables, which not only greatly reduces the training time and network size, but also the input variables are orthogonal, and the noise is eliminated under the premise of almost no loss of major spectral information. The quantitative calibration method combining PCA or PLS with ANN is called PCA-ANN or PLS-ANN method. The input variables can also adopt the wavelength variables selected by simulated annealing algorithm and genetic algorithm (GA), as well as the coefficients obtained by Fourier transform and wavelet transform. Since the node value in the neural network is defined as between 0 and 1, the input variable x is often preprocessed by the following equation:

$$x_i^p = 0.8 \frac{x_i - x_{\min}}{x_{\max} - x_{\min}} + 0.1 \quad (8.18)$$

where x_i is the i th variable, x_{\min} and x_{\max} are the minimum value and the maximum value of the variables, respectively.

- (2) Number of hidden layer networks: usually choose a hidden layer, namely, three layers BP network, which can be used to solve the nonlinear quantitative calibration of most problems. For more complex problems, at most two hidden layers can be selected. In the BP algorithm, the error propagates from the output

layer to the input layer, and the more layers there are, the less reliable the back propagation error will be when it is close to the input layer. The effect can be imagined by using such an unreliable error to correct the weight.

- (3) Number of hidden layer nodes: generally speaking, the more hidden layer nodes can store more information. However, with the number of hidden layer nodes increasing, the weight increases square, which leads to longer training time. Moreover, for more hidden layer nodes, more training samples are usually needed. Otherwise, the resulting mathematical model is unstable, that is, the results on the training set appear to be good, but the result of the prediction set may be poor, which is often referred to as “over-fitting”. “Over-fitting” is caused by fitting the noise in the test. Less hidden layer nodes store less information, which cannot fully reflect the complex functional relationship between input and output variables. Moreover, it is easy to fall into local minimum in the training process (Fig. 8.5), and the established model cannot correctly transmit nonlinear information. The number of nodes h in the hidden layer can be selected as the initial value through some empirical formulas, and then the optimal value can be finally determined by gradually increasing or decreasing the number of nodes by 1–3.
- (4) Initial weight: in neural network, initial weight has a great influence on whether learning reaches local minimum, whether convergence can be achieved and the length of training time. If initial weights are different, the output is generally not the same. How to select the optimal weights is still no rule to follow. At present,

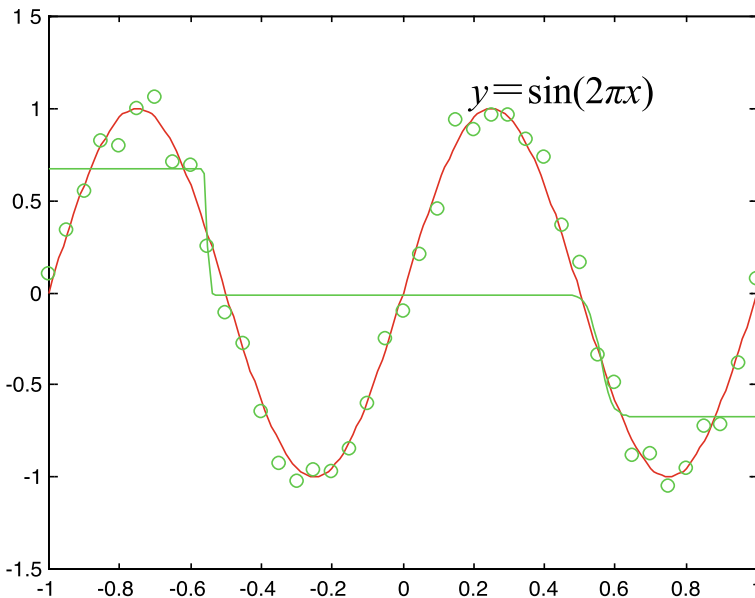


Fig. 8.5 Phenomenon of trapping in local minima

the common method is to try different initial weights by experiment (such as random numbers). Sometimes, if the initial weight is not selected properly, the BP algorithm will not be able to obtain satisfactory results. In this case, it is suggested to reinitialize the weight and let the network learn again. In order to prevent accidental correlation or local optimization, it is suggested to repeat the calculation at least 50–200 times under the same network structure and take the average value. Genetic algorithm (GA) or particle swarm optimization (PSO) can also be used to optimize the initial weights and thresholds of the neural network.

- (5) Transfer function: for nonlinear problems, the input layer and hidden layer mostly adopt nonlinear transmission function, while the output layer adopts linear transmission function such as Purelin function to maintain the range of output. In terms of nonlinear transfer function, logarithmic sigmoid function is used when the sample output is greater than zero, otherwise tangent sigmoid function is used.
- (6) Learning rate: the effectiveness and convergence of BP network depend on the learning rate to a large extent. In the initial stage of learning, a larger learning rate is expected to accelerate the learning process and convergence speed. However, when the training process is close to the optimal weight value, the learning rate must be quite small, otherwise it will oscillate and cannot converge. The method of variable learning rate can be adopted, which is generally selected as the values between 0.001 and 0.8, and then dynamically changed according to the gradient change and the change value of mean square error in the training process.
- (7) Learning algorithms: when selecting learning algorithms for training BP network, it is necessary to consider the complexity of the problem, size of sample set, network size, error target and problem type (function fitting or pattern recognition). Table 8.2 compares the several typical fast learning algorithms, which can be used as a reference when selecting algorithms.
- (8) Termination condition: once the training reaches the maximum number of training, or the sum of the square of the network error drops below the expected error, the network will stop learning. In addition, in order to solve the “over-training” problem, that is, although the error of the training set can continue to decrease during the iteration process, the deviation of the prediction set starts to rise, which is caused by the model being built to “fit” individual sample (Fig. 8.6).

Usually, the training set is divided into two parts. One part is the calibration set or training set, whose prediction error is transmitted in reverse to adjust the weight. The other part is the test set or validation set, which does not participate in the training directly, but its prediction residual error sum of squares (PRESS) is used for the training of the control network. As shown in Fig. 8.7, in the initial stage of training, the error of the test set usually decreases with the decrease of the network training error. However, when the network begins to be “over-trained”, the

Table 8.2 Comparison of several typical BP fast learning algorithms

Learning algorithm	Apply to problems	Convergence	Memory	Other characteristics
L-M optimization algorithm	Function fitting	Fast convergence and small convergence error	Big	Performance deteriorates as the network size increases
L-M optimization algorithm for Bayesian regularization	Function fitting	Convergence is slow	Medium	It is suitable for function fitting of small-scale network and has good generalization ability
Quasi-Newton algorithm	Function fitting	Convergence faster	Larger	The amount of computation increases geometrically with the increase of network size
Elastic learning algorithm	Pattern recognition	Convergence is the fastest	Smaller	The performance becomes worse with the decrease of network training error
Conjugate gradient algorithm	Function fitting Pattern recognition	Fast convergence and stable performance	Medium	Especially suitable for the large scale of the network

test error will gradually increase. When the test error increases to a certain extent, the network training will stop in advance, and then the training function will return to the network object with the minimum verification error. Test error does not participate in the network training, but it can be used to evaluate the rationality of the network of training results and training set composition. If the training error and validation error reached the minimum training step difference is very big, or the change trend of error curve difference is quite different, it indicates that the sample composition of the training set is not very reasonable and needs to be redivided. This method is simple and effective, often gets good results, and the training time is greatly reduced. This approach is commonly referred to in the literature as “early-stopping”.

8.1.4 Other Types of Neural Networks

Radial basis function networks (RBF) is a single-hidden layer feedforward network proposed by Moody and Darken, in which input layer nodes directly transmit input signals to the hidden layer. The function of hidden layer node is a radial basis function

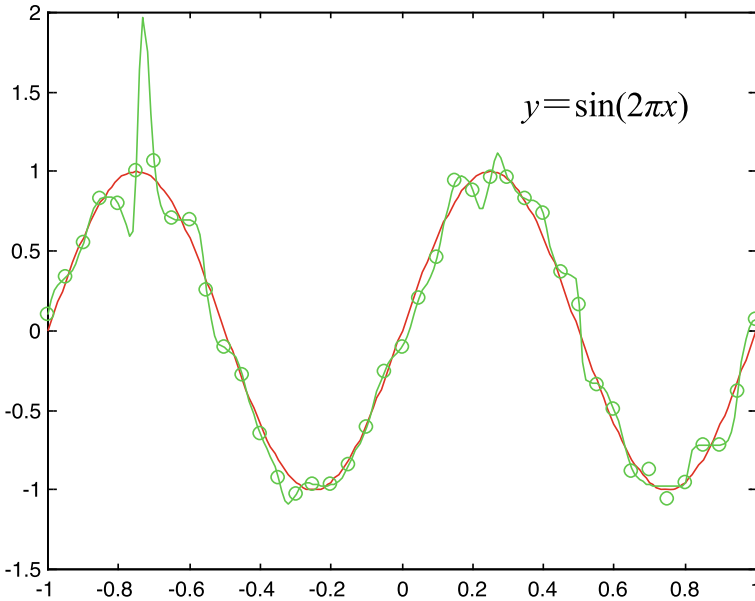


Fig. 8.6 Phenomenon of over-training

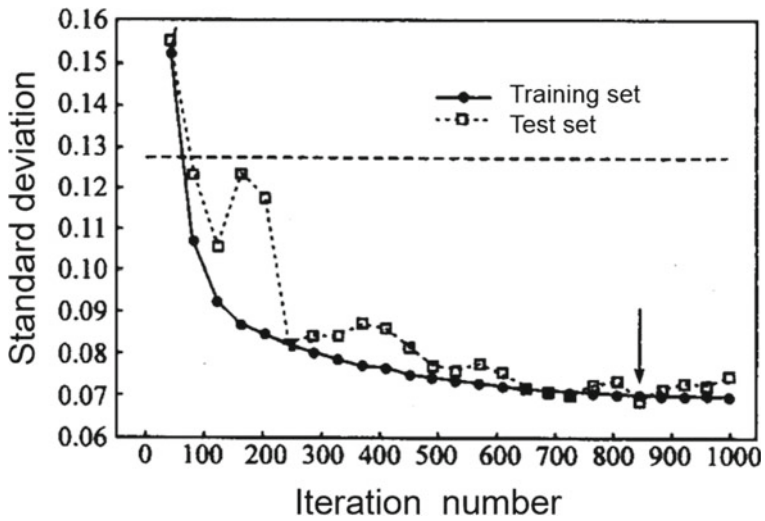
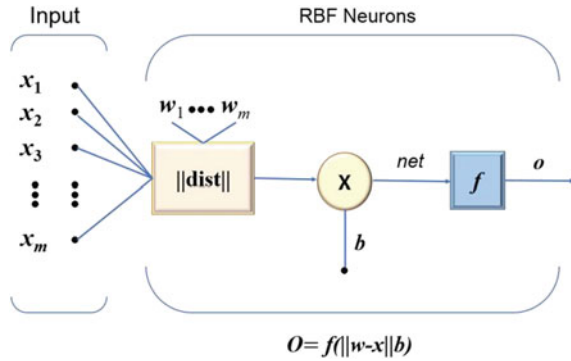


Fig. 8.7 Test set monitoring the training process of training set [3]

Fig. 8.8 RBF neuron model



such as Gaussian function, while the output layer node is usually a simple linear function.

RBF network can determine the corresponding network topology according to the problem, which has the characteristics of fast learning speed, no local minimum problem and easy convergence for iterative training. Neuron model of RBF neural network as shown in Fig. 8.8, `lstdist` module for calculating the Euclidean distance between the input vector x and the weight vector w , and the transfer function of RBF neurons usually uses Gaussian function: $f(net) = \exp(-net^2)$, its `net` input value is the Euclidean distance between the input vector x and the weight vector w multiplied by the threshold value of b (`lstdist`· b), which can also be written $\frac{\|x-w\|}{\sqrt{2}\sigma}$, namely, $b = \frac{1}{\sqrt{2}\sigma}$, σ is the variance of Gaussian function.

Center and width are two important parameters of RBF neurons. The weight vector w of the neuron determines the center of the radial basis function. When the input vector x coincidences with the weight vector w , the output of the RBF neuron reaches the maximum. The farther the distance between the input vector x and w is, the smaller the neuron output will be. This feature makes RBF network very suitable for approximating fuzzy rules. Neuron threshold b determines the width of the radial basis function. When b is larger, the attenuation range of the function will be larger when the input vector x is far away from w .

The structure of RBF network is similar to that of BP network, which is a three-layer forward network. The first layer is the input layer, which is composed of input nodes. The input layer does not process information. The second layer is the hidden layer, the number of units depends on the need of the description problem, and each neuron in the hidden layer represents a group of radial basis functions. The third layer is the output layer, which responds to the action of the input pattern. The basic idea of forming RBF network is that the radial basis function is used as the “basis” of the hidden unit to form the hidden layer space, so that the input vector is directly mapped to the hidden space (without weight connection). When the center point of the radial basis function is determined, this mapping relationship is also determined. The mapping from hidden layer space to output space is linear, that is, the output of the network is linear weighted sum of the output of hidden units.

The training and learning method of RBF network is similar to that of BP network. Theoretically, RBF network and BP network can approximate any continuous nonlinear function as well. The main difference between them is that they use different transfer functions. The hidden layer in BP network usually uses sigmoid function, which is non-zero over an infinite range of input space, while the transfer function of RBF network is local.

Another commonly used method in the spectral analysis is the Kohonen network. It belongs to self-organizing neural network and will be introduced in the section of pattern recognition.

8.1.5 Optimization of Neural Network Parameters

In the traditional BP neural network, the weights and thresholds are determined randomly and easily fall into the local minimum, which has a great influence on the accuracy of the prediction results. GA is self-adaptive and global optimal, and it is easy to search the global optimal solution. Therefore, the GA can be used to optimize the initial weights and thresholds of the neural network. After optimization, the BP network has a fast convergence speed and is not limited to local optimum.

GA to optimize the BP neural network in order to get better network initial weights and threshold value, as shown in Fig. 8.9, the basic framework is using the GA individuals on behalf of the network's initial weights and threshold. The individual value of initialization of the BP neural network prediction error as the individual fitness values, and through the selection, crossover, and mutation operation to find

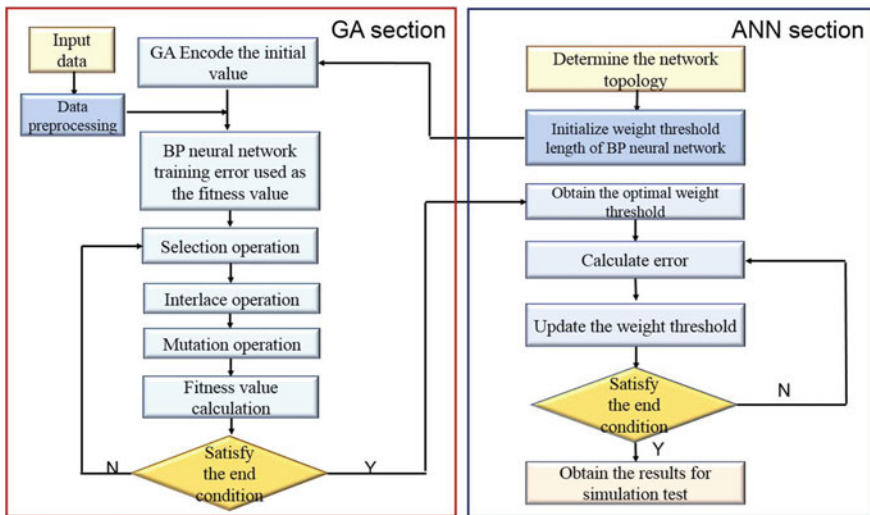


Fig. 8.9 Flowchart of BP neural network optimization by GA

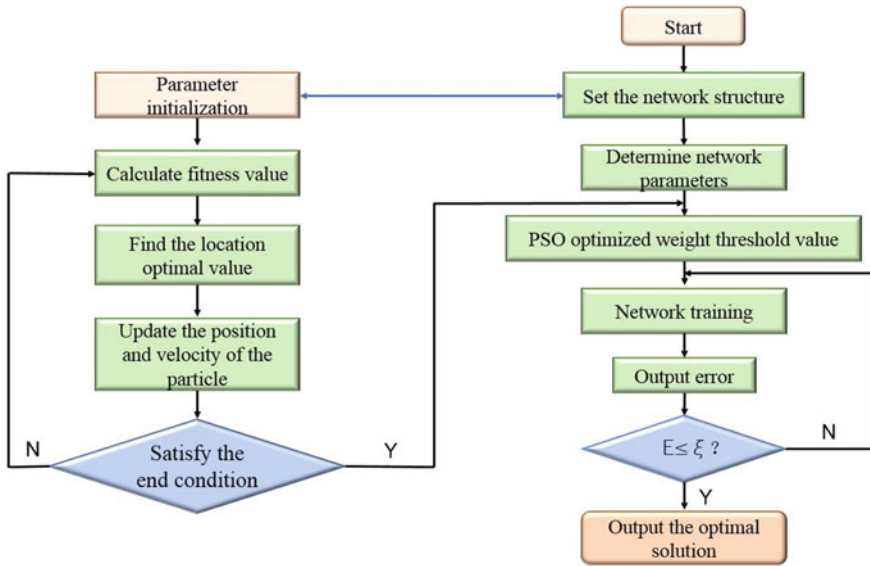


Fig. 8.10 Flowchart of BP neural network optimization by PSO

the optimal individual, that is, the initial weight of the optimal BP neural network [4, 5].

In addition to GA, PSO and ant colony algorithm can also be used to optimize the initial weight of BP neural network. Figure 8.10 shows the calculation flow of optimizing BP neural network with PSO [6, 7].

8.2 Support Vector Machine

8.2.1 Introduction

Support vector machine (SVM) is a new pattern recognition method developed on the basis of statistical learning theory (SLT) established by Vapnik after the 1990s [8, 10]. Since SLT is a statistical learning method, established specifically for small samples, SVM can effectively overcome the disadvantages of neural network, including difficult convergence, unstable solution, and poor generalization (i.e., predictive ability). Furthermore, SVM has advantages compared with many traditional pattern recognition algorithms for pattern recognition of small-sample, nonlinear and high-dimensional data spaces. At present, SVM has been widely used in pattern recognition, signal processing, signal communication, etc.

The basic idea of SVM comes from the optimal classification surface of linear discrimination, which is to require the classification surface to not only separate the

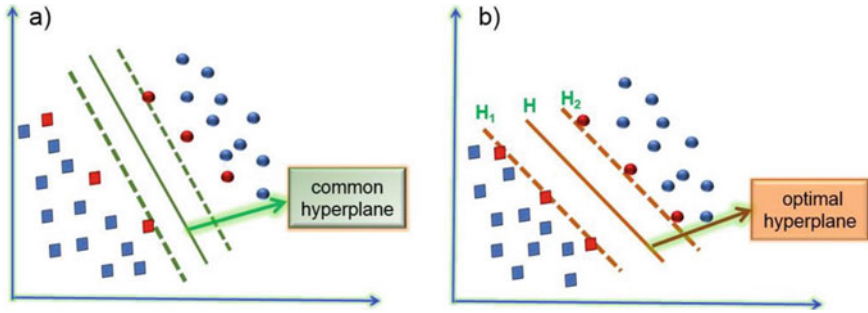


Fig. 8.11 Schematic diagram of common (a) and optimal (b) hyperplane

two samples without error, but also to maximize the classification interval (Fig. 8.11). Through the realization of the optimal classification surface, a direct advantage is to improve the prediction ability and reduce the classification error rate.

Let two classes of linear separable total training sets be $(x_i, y_i), i = 1, \dots, n$, training set with a total of n samples, $\mathbf{x} \in \mathbb{R}^d$, d as the number of characteristic variables, and $y \in \{+1, -1\}$ is the class label. The general form of the linear discriminant function in d -dimensional space is the $g(x) = \mathbf{w}^T \mathbf{x} + b$, classification surface equation for $\mathbf{w}^T \mathbf{x} + b = 0$, which normalizes the discrimination function so that all samples of both classes satisfy $|g(x)| \geq 1$. When the $|g(x)| = 1$, of the sample nearest to the classification surface so that the classification interval is equivalent to $2/\|\mathbf{w}\|$. Thus, making the classification interval maximum is equivalent to making $\|\mathbf{w}\|$ or $\|\mathbf{w}\|^2$ minimum. If it is required that the classification surface to classify all the samples correctly, it must be met

$$y_i(\mathbf{w}^T x_i + b) - 1 \geq 0, i = 1, \dots, n \tag{8.19}$$

Therefore, the classification surface that meets the above conditions and minimizes the $\|\mathbf{w}\|^2$ is the optimal classification surface, and passing the training samples on the H_1, H_2 closest to the classification surface and parallel to the optimal classification surface, even those with the above equality are called support vectors, because they support the optimal classification surface.

From the above analysis, it is concluded that the problem of obtaining the optimal classification surface can be expressed as a constrained optimization problem, that is, finding the $\|\mathbf{w}\|^2/2$ minimum under the constraint condition $y_i(\mathbf{w}^T x_i + b) - 1 \geq 0$. To do this, the following Lagrange function can be defined:

$$L(\mathbf{w}, b, \alpha) = \frac{1}{2} \mathbf{w}^T \mathbf{w} - \sum_{i=1}^n \alpha_i [y_i(\mathbf{w}^T x_i + b) - 1] \tag{8.20}$$

where $\alpha_i \geq 0$ is the Lagrange coefficient, the goal of the problem is to obtain the minimum of the Lagrange function for \mathbf{w} and b .

Find the partial differential of the above equation for w and b , respectively, and make it equal to 0, turn the original problem into a simple convex quadratic planning dual problem:

Under the constraints $\sum_{i=1}^n \alpha_i y_i = 0$ and $\alpha_i \geq 0$, the maximum value for the following function can be solved:

$$Q(\alpha) = \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j (x_i^T x_j) \quad (8.21)$$

This is a problem of finding the extreme value of quadratic function under inequality constraints, there is a unique optimal solution, and the optimal solution of this optimization problem must be met

$$\alpha_i [y_i (w^T x_i + b) - 1] = 0, i = 1, \dots, n \quad (8.22)$$

If α_i^* is the optimal solution obtained, $w^* = \sum_{i=1}^n \alpha_i^* y_i x_i$, for most samples α_i^* will be 0. The samples whose α_i^* value is not 0 are support vectors, which are usually only a small part of the whole training set samples.

After solving the above problems, the optimal classification function can be obtained:

$$f(x) = \text{sgn}(w^{*T} x + b^*) = \text{sgn}\left(\sum_{i=1}^n \alpha_i^* y_i x_i^T x + b^*\right) \quad (8.23)$$

The $\text{sgn}()$ is a symbolic function, and since the α_i^* corresponding to the non-support vector is 0, the sum in the equation actually sums only for the support vectors. While b^* is the threshold for classification, it can be obtained by either one support vector by the constraint $\alpha_i [y_i (w^T x_i + b) - 1] = 0$, or by taking the median of any pair of support vectors of the two classes.

When the two classes of samples cannot be completely separated with a superplane, and a few samples are misclassified, at this time, the relaxation variable ξ_i , $\xi_i \geq 0, i = 1, \dots, n$ can be introduced to make the superplane $w^T x + b = 0$ to satisfy $y_i (w^T x_i + b) \geq 1 - \xi_i$. When $0 < \xi_i < 1$, the sample x_i is still correctly classified, and when $\xi_i \geq 1$, the sample x_i is misclassified. To do this, the following objective function is introduced:

$$\phi(w, \xi) = \frac{1}{2} w^T w + C \sum_{i=1}^n \xi_i \quad (8.24)$$

where C is a constant greater than zero, called the penalty factor. It plays a role in controlling the degree of penalty for the misclassified samples, realizing a compromise between the proportion of the misclassified samples and the complexity of the algorithmic. This optimization problem can be solved by the same method as solving

the optimal classification surface, obtaining a quadratic function extremal problem, and also obtaining the almost exact same result, just the constraint of α_i changes to $0 \leq \alpha_i \leq C$.

If a simple hyperplane in the original space cannot get a satisfactory classification effect, a complex hypersurface must be used as the interface. For this linear nonseparable issues, the SVM algorithm introduces the kernel space theory in which the low-dimensional input space data is mapped to the high-dimensional feature space (Hilbert space) through the nonlinear mapping function $\varphi(x)$, and then seeks the optimal linear classification surface in this new space. It is shown that if the appropriate mapping function $\varphi(x)$ is selected, the input space linear nonseparable issues will be transformed into a linear separable problem in the feature space (Fig. 8.12).

In the nonlinear case, the classification superplane is

$$w\phi(x) + b = 0 \tag{8.25}$$

The optimization function is

$$Q(\alpha) = \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j \langle \phi(x_i), \phi(x_j) \rangle \tag{8.26}$$

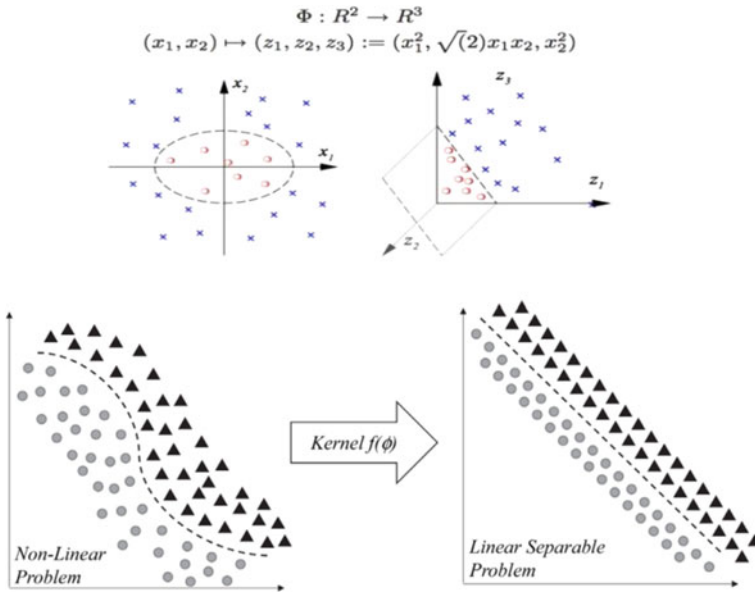


Fig. 8.12 Mapping from low-dimensional to high-dimensional feature space by nonlinear functions

where $\langle \phi(x_i), \phi(x_j) \rangle$ represents the inner product (or point product) of $\phi(x_i)$ and $\phi(x_j)$.

The optimal classification function obtained is

$$f(x) = \text{sgn} \left(\sum_{i=1}^n \alpha_i^* y_i \langle \phi(x_i), \phi(x) \rangle + b^* \right) \quad (8.27)$$

However, if classification or regression are conducted directly in high-dimensional space, there are problems such as determining the form and parameters of the nonlinear mapping function, the dimension of feature space (high dimension, or even infinite dimension), while the biggest obstacle is the “dimensional disaster” in high-dimensional feature space operations. These problems can be effectively solved by using the kernel function, which $K(\cdot)$ is defined as $K(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle$, that is, the kernel function transforms the inner product operation of high-dimensional space into the kernel function $K(\cdot)$ calculation of low-dimensional input space. It solves the problems of “dimensional disaster” computed in the high-dimensional feature space and lays the theoretical foundation for solving complex classification or regression problems in a high-dimensional feature space.

Instead of $\langle \phi(x_i), \phi(x_j) \rangle$ with kernel function $K(x_i, x_j)$, the optimization function changes to

$$Q(\alpha) = \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j K(x_i, x_j) \quad (8.28)$$

The corresponding discriminant function also changes to

$$f(x) = \text{sgn} \left(\sum_{i=1}^n \alpha_i^* y_i K(x_i, x) + b^* \right) \quad (8.29)$$

This is the SVM, where x_i is a support vector and x is an unknown vector. Since the final discriminant function contains only a linear combination of the inner product of the unknown vectors and the support vectors, the computational complexity at recognition depends on the number of support vectors.

After adopting the kernel function, there is no need to know the specific form of the nonlinear mapping function $\phi(x)$. There are mainly several common forms of nuclear functions, all of which are corresponding to the existing algorithms.

- (1) The kernel function of the polynomial form

$$K(x_i, x_j) = [(x_i^T x) + 1]^q \quad (8.30)$$

At this time, the corresponding SVM is an order q polynomial classifier.

- (2) The kernel function of the radial basis form

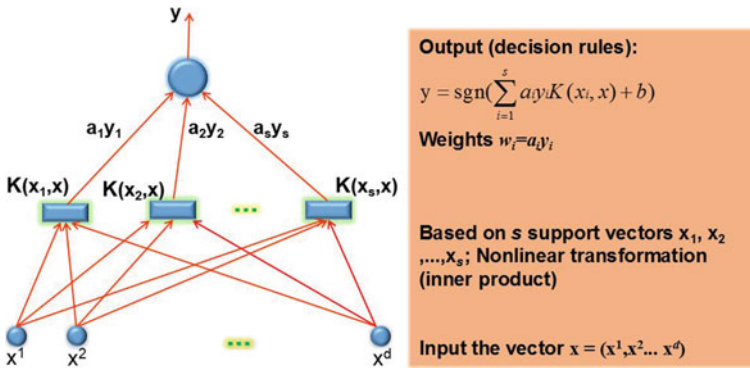


Fig. 8.13 Schematic diagram for decision rules of SVM

$$K(x_i, x_j) = \exp(-\|x_i - x_j\|^2 / 2\sigma^2) \tag{8.31}$$

At this time, the corresponding SVM is a radial basis function classifier distinguished from the traditional radial basis RBF function in that the center of each basis function here corresponds to a support vector. Those support vectors and their output weights are determined automatically by the algorithm.

- (3) S-shaped kernel function,

$$K(x_i, x_j) = \tanh(\beta_0(x_i^T x_j) + \beta_1) \tag{8.32}$$

Moreover, there are exponential radial kernel function, Fourier series, spline function, B spline function, etc.

As shown in Fig. 8.13, the discriminant function of the SVM is formally similar to a neural network whose output can be seen as a linear combination of several hidden layer nodes, and each hidden layer node corresponds to the inner product of a input sample and a support vector. Therefore, the SVM is also called the support vector network. SVM implements a two-layer perceptron neural network, where both the network weights and the number of hidden layer nodes are automatically determined by the algorithm.

For classification learning problems, traditional pattern recognition methods emphasize dimensionality reduction, while SVM is the opposite of this. For the nonlinear problem that the two types of samples in the feature space cannot be separated by the hyperplane, SVM adopts the mapping method to map it to the higher-dimensional space, and obtains the hyperplane equation which best distinguishes the two types of sample points as the criterion to distinguish the unknown samples. Since the inner product operation is only changed by the kernel function after the dimension increase, the complexity of the algorithm does not increase with the increase of the dimension, thus limiting over-fitting. Although with a small number of known samples, it can still effectively make statistical forecasts. The specific steps

to apply SVM are: selecting the appropriate kernel function \rightarrow solving the optimization equation to obtain the support vector and the corresponding Lagrange operator \rightarrow obtaining the optimal classification surface discriminant equation.

Introduction above is a binary classifier. The SVM-based construction of multi-valued classifiers can be realized by combining multiple binary sub-classifiers, and the specific construction includes one-to-one and one-to-many ways. The implementation steps for mode recognition of SVM are relatively simple, without a long training process. It only needs to solve the optimal hyperplane according to the initial sample to find the support vector, and then determine the discrimination function, and then it can be generalized to identify other unknown samples. The accuracy of the SVM is greatly affected by the parameters of the kernel function itself. How to select these parameters, such as the width of the radial basis function, and the order of the polynomial kernel function, there are no mature methods and generally determined by multiple attempts.

8.2.2 Support Vector Regression

The SVM method was first proposed for the pattern recognition problem. With the introduction of ε insensitive function, SVM has extended for nonlinear regression and function approximations, and shows good learning performance, especially for solving the regression problem of small samples [11]. The following is a brief introduction to the support vector regression (SVR) method based on SVM.

For a linear regression system, $f(x) = w^T x + b$, and its calibration set sample (x_i, y_i) , $i = 1, 2, \dots, n$, n is the number of calibration samples. If all the calibration data can be fitted with a linear function with precision ε , without error, ε is a normal number,

$$\begin{aligned} y_i - w^T x_i - b &\leq \varepsilon \\ w^T x_i + b - y_i &\leq \varepsilon \end{aligned} \quad (8.33)$$

Considering the allowable fitting error, the relaxation factors ξ_i and ξ_i^* , $\xi_i, \xi_i^* \geq 0$ are introduced, and the above equations become

$$\begin{aligned} y_i - w^T x_i - b &\leq \varepsilon + \xi_i \\ w^T x_i + b - y_i &\leq \varepsilon + \xi_i^* \end{aligned} \quad (8.34)$$

$$\xi_i, \xi_i^* \geq 0$$

Similar to the SVM problem for pattern recognition, the problem can be transformed into finding the minimum of the following function under the above constraints:

$$L(w, \xi, \xi^*) = \frac{1}{2} w^T w - C \sum_{i=1}^n (\xi_i + \xi_i^*) \quad (8.35)$$

where item 1 is to make the regression function flatter in order to improve the generalization ability, while item 2 is to reduce the error. The constant C is a constant greater than zero, called the penalty factor or the regularization coefficient, which represents the degree of penalty for a sample that exceeds the error ε . The dual problem can be obtained by using the Lagrange optimization method, that is, maximizing the following objective function for Lagrange factors α_i and α_i^* under the constraints of $\sum_{i=1}^n (\alpha_i + \alpha_i^*) = 0$, $0 \leq \alpha_i \leq C$ and $0 \leq \alpha_i^* \leq C$:

$$W(\alpha, \alpha^*) = -\varepsilon \sum_{i=1}^n (\alpha_i + \alpha_i^*) + \sum_{i=1}^n y_i (\alpha_i^* - \alpha_i) - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n (\alpha_j^* - \alpha_j) (\alpha_i^* - \alpha_i) (x_i^T x_j) \quad (8.36)$$

Get a regression function of

$$f(x) = \sum_{i=1}^n (\alpha_i^* - \alpha_i) (x_i^T x_j) + b \quad (8.37)$$

In the equation, only a few parts of the $(\alpha_i^* - \alpha_i)$ are not zero, and their corresponding samples are called support vectors. If the fitted mathematical model is expressed as a curve in multi-dimensional space, the result obtained on the ε insensitive function is the “ ε pipeline” containing the curve and the training point. Of all the samples, only that portion of the sample points distributed on the “pipe wall” determines the location of the “pipeline”, and this part of the training samples is the support vector.

As shown in Fig. 8.14, in the case of a nonlinear problem, the main idea of the SVR method is to transform the original problem into a linear problem in a high-dimensional space through nonlinear transformation and solve it linearly in the high-dimensional space. As the SVM method of pattern recognition, nonlinear regression can be achieved as long as the kernel function $K(x_i, x_j)$ replaces the point product $x_i^T x_j$ operation in the regression function. In this way, the nonlinear solving problem becomes to maximize the following objective function for the Lagrange factors α_i and α_i^* under the constraints of $\sum_{i=1}^n (\alpha_i + \alpha_i^*) = 0$, $0 \leq \alpha_i \leq C$ and $0 \leq \alpha_i^* \leq C$:

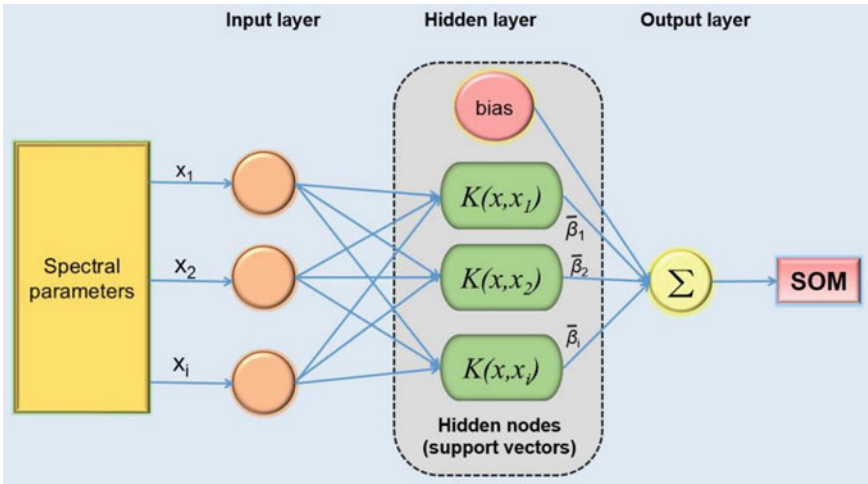


Fig. 8.14 Schematic diagram of SVR topological structure

$$\begin{aligned}
 W(\alpha, \alpha^*) &= -\varepsilon \sum_{i=1}^n (\alpha_i + \alpha_i^*) + \sum_{i=1}^n y_i (\alpha_i^* - \alpha_i) \\
 &\quad - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n (\alpha_j^* - \alpha_j) (\alpha_i^* - \alpha_i) K(x_i, x_j) \quad (8.38)
 \end{aligned}$$

Using the same optimization method, the nonlinear regression function is obtained as

$$f(x) = \sum_{i=1}^n (\alpha_i^* - \alpha_i) K(x_i, x_j) + b \quad (8.39)$$

Kernel functions such as polynomial, radial basis, and S-form are also mostly used in SVM regression methods.

Using the ε insensitive function, the above optimization algorithm is expressed by the matrix as

$$\min_p \frac{1}{2} p^T H p + c^T p \quad (8.40)$$

Among them,

$$p = \begin{bmatrix} \alpha \\ \alpha^* \end{bmatrix}, H = \begin{bmatrix} X & -X \\ -X & X \end{bmatrix}, X = \begin{bmatrix} K(x_1, x_1) & \cdots & K(x_1, x_k) \\ \vdots & & \vdots \\ \vdots & & \vdots \\ K(x_k, x_1) & \cdots & K(x_k, x_k) \end{bmatrix}, c = \begin{bmatrix} y_1 \\ \vdots \\ y_k \end{bmatrix}.$$

Its constraints are $p \cdot (1, \dots, 1, -1, \dots, -1) = 0, 0 \leq \alpha_i \leq C, 0 \leq \alpha_i^* \leq C$, where $i = 1, \dots, n$, n is the number of samples in the calibration set. The above algorithm can be easily implemented by using the MATLAB language.

It shows the advantage of SVR is specifically clear for the finite sample case, with the purpose to obtain the optimal solution under the existing information, instead of just optimal value when the sample number tends to infinity. SVR solution algorithm can be transformed into a quadratic optimization (quadratic programming) problem. In theory, the global optimum can be obtained for SVR. It adopts kernel function to realize the transformation from nonlinearity in low-dimensional space to linearity in high-dimensional space, which ensures that the algorithm has good generalization ability and solves the problem of dimension disaster. However, due to the dimension of the \mathbf{H} matrix is twice the number of samples, the number of samples that can be processed by this method cannot be too large.

8.2.3 Least Squares Support Vector Regression

To reduce training time, reduce computational complexity, and improve generalization abilities, some improved SVM algorithms such as least squares SVM (LS-SVM) and weighted SVM were proposed. The LS-SVM uses least squares linear systems as a loss function, reduces the computational complexity, and accelerates the solution by solving a set of linear equations instead of the more complex quadratic programming method adopted by the traditional SVM. LS-SVM method has been applied in the qualitative and quantitative spectral analysis [12–14].

The objective optimization function of the LS-SVM algorithm is

$$\min J(w, e) = \frac{1}{2}w^T w + \frac{1}{2}\gamma \sum_{i=1}^n e_i^2 \quad (8.41)$$

Constraint: $y_i = w^T \varphi(x_i) + b + e_i$.

where w is the weight vector; γ is the regularization parameter; e_i is error; x_i and y_i are input and output variables of calibration set, respectively; $i = 1, \dots, n$, n is the number of samples of calibration set.

The following Lagrange function can be defined:

$$L(w, b, \alpha, e) = J(w, e) - \sum_{i=1}^n \alpha_i [w^T \phi(x_i) + b + e_i - y_i] \quad (8.42)$$

where α_i is the Lagrange coefficient. The above optimization problems can be transformed into solving the linear equations:

$$\begin{bmatrix} 0 & l^T \\ l & \Omega + \frac{1}{\gamma} I \end{bmatrix} \begin{bmatrix} b \\ \alpha \end{bmatrix} = \begin{bmatrix} 0 \\ y \end{bmatrix} \quad (8.43)$$

where $l = [1, 1, \dots, 1]^T$; The \mathbf{I} is a unit matrix; $\Omega = \langle \varphi(x_i), \varphi(x_j) \rangle = K(x_i, x_j)$, $i, j = 1, \dots, n$; $\alpha = [\alpha_1, \alpha_2, \dots, \alpha_n]^T$; $y = [y_1, y_2, \dots, y_n]^T$.

Making $A = \Omega + \frac{1}{\gamma} I$, the matrix equation can be solved:

$$b = \frac{l^T A^{-1} y}{l^T A^{-1} l} \quad (8.44)$$

$$\alpha = A^{-1} (y - bl) \quad (8.45)$$

For unknown sample x , the predicted value of LS-SVM is

$$y(x) = \sum_{i=1}^n \alpha_i K(x, x_i) + b \quad (8.46)$$

The standard SVM solves a convex quadratic programming. Its solution is unique and optimal, without local extremal problem for general neural networks. When the linear equations are solved with LS-SVM, its solution satisfies the extremal conditions, but it is not guaranteed to be the global optimal solution. However, it is solved faster and has less computational resources required for solving.

8.2.4 Optimization of Support Vector Regression Parameters

In SVR methods, the selection of the penalty factor C and the kernel radius parameter σ is crucial to the construction of the regression function. The penalty factor C represents the importance of the SVM algorithm to the abnormal points and affects the prediction accuracy of the model. The larger the C , the smaller the training set

error, and the larger the C will easily lead to over-fitting; the smaller the C , the larger the training set error, and the smaller the C is easy to underfitting. Too large or too small the C will weaken the model generalization ability. σ represents the distribution of data mapped to high-dimensional feature space, affecting the training speed of the model. The faster speed will be obtained with larger σ and fewer support vectors. Otherwise, smaller σ and more support vectors will induce slower speed.

For the selection of these two parameters, the common method is to have C and σ take values within a certain range, and then adopt the interactive verification method to find the best parameters [15] by grid search method based on a certain step length. However, this method is time consuming, especially when looking for the best point in a larger range, as it needs to traverse all the parameter points in the grid.

To search for optimal parameters over a wider range, heuristic algorithms such as GA, PSO, and gray wolf optimization (GWO) to select penalty factor C and kernel parameter σ , tending to be more [16, 17] efficient than grid search methods. Particle optimization group algorithm is a group intelligence algorithm that simulates bird group foraging, each particle represents a possible solution vector. The quality of particles is judged according to the fitness function value and realizes the particle position and speed continuously updated by learning from the global and individual optimal solutions, and finally realizes the purpose of global optimization [18, 19]. Figure 8.15 shows the algorithm process for optimizing the SVR parameters with PSO.

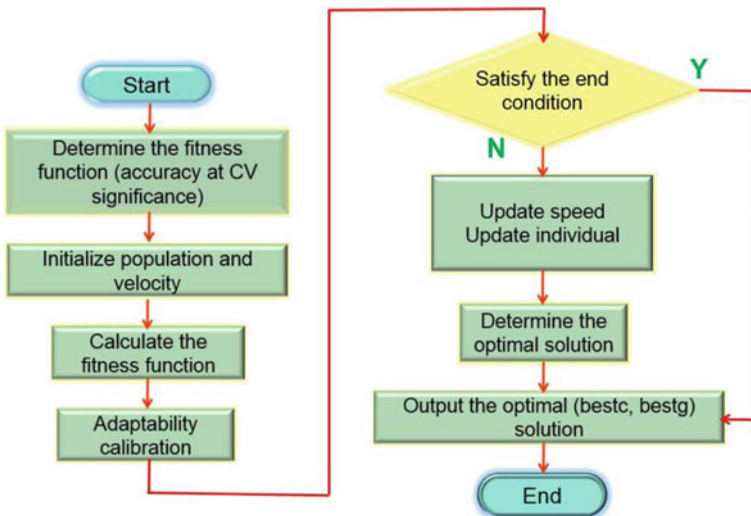


Fig. 8.15 Flowchart of parameters optimization of support vector machine by PSO

8.3 Relevance Vector Machine

Relevance vector machines (RVM) is a new supervised learning approach similar to SVM, based on kernel function mapping that transforms the nonlinear problem in low-dimensional space into the linear problem in high-dimensional spaces. It is trained in the Bayesian framework and builds a sparse model [20] based on active correlation decision theory under prior conditions. In the iterative learning of the training sample set, the posterior distribution of the parameters independent of the predicted values gradually tends to zero, and the points corresponding to the non-zero parameters can reflect the core characteristics of the data, known as the relevance vectors, reflected in the data the most core characteristics. Compared to the SVM, RVM reduces the operation amount of the kernel function and outperforms the SVM [21, 22] in terms of sparsity and generalization abilities.

Given the training set input vector $X = \{x_1, x_2, \dots, x_n\}^T$ and the corresponding output $y = \{y_1, y_2, \dots, y_n\}^T$, n as the number of training set samples. The purpose of learning is to apply these training data and prior knowledge to design a system that predicts the system to output y^* for the new input x^* .

Suppose that the target value is a combination of an unknown function and some noise:

$$y = f(X, w) + \varepsilon \quad (8.47)$$

where w is the model weight, $w = \{w_1, w_2, \dots, w_m\}^T$, m is the number of wavelength variables, ε is the noise with zero mean and σ^2 variance. $f(X, w)$ is the family of functions, given by the following equation:

$$f(X, w) = \sum_{i=1}^m w_i \Phi(x) + w_0 \quad (8.48)$$

where $\Phi(x)$ is a set of nonlinear basis functions (kernel functions):

$$\Phi(x) = \{\varphi_1(x), \varphi_2(x), \dots, \varphi_m(x)\}^T \quad (8.49)$$

The Gaussian function centered on each training sample is usually chosen as the base function, and the weight w can be trained by maximum likelihood methods in the Bayesian framework:

$$p(y|w, \sigma^2) = (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{\|y - w\Phi\|^2}{2\sigma^2}\right\} \quad (8.50)$$

To avoid over-fitting, sparse Bayesian learning method gives a priori conditional probability distribution to the weight w :

$$p(w|\alpha) = \prod_{i=0}^n N(w_i|0, \alpha_i^{-1}) \quad (8.51)$$

where superparametric $\alpha = \{\alpha_0, \alpha_1, \dots, \alpha_n\}^T$, each weight of the w_i corresponds to a unique superparametric α_i . Parameters are affected by the prior distribution. The training set samples are constantly trained, most of the superparametric α_i will tend to infinity, and the corresponding weight w_i will go to 0, thus ensuring the sparsity of the RVM.

According to the Bayesian rule, the posterior probability on the weights can be obtained:

$$p(w|y, \alpha, \sigma^2) = (2\pi)^{-\frac{n+1}{2}} |\Sigma|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(w - \mu)^T \Sigma^{-1}(w - \mu)\right\} \quad (8.52)$$

Posterior covariance $\Sigma = (\Phi^T \mathbf{B} \Phi + \mathbf{A})^{-1}$

Mean $\mu = \Sigma \Phi^T \mathbf{B} y$.

where $\mathbf{A} = \text{diag}(\alpha_0, \alpha_1, \alpha_2, \dots, \alpha_n)$, $\mathbf{B} = \sigma^2 \mathbf{I}_n$.

Based on the maximum expectation hyperparameter estimation, after multiple iterative calculations can obtain:

$$(\alpha_i)^{new} = \gamma_i / \mu_i^2 \quad (8.53)$$

$$(\sigma^2)^{new} = \frac{|y - \Phi \mu|^2}{n - \sum_i \gamma_i} \quad (8.54)$$

where μ_i is the i th posterior mean weight, $\gamma_i = 1 - \alpha_i \Sigma_{ii}$.

For the prediction of the weight posterior probability distribution, the constraints of both α_{MP} and σ_{MP}^2 take the maximum value, and according to the normal distribution properties, $p(y^*|y)$ is conformed to the normal distribution:

$$p(y^*|y, \alpha_{MP}, \sigma_{MP}^2) = N(\mu^*, \sigma_*^2) \quad (8.55)$$

where

$$\sigma_*^2 = \sigma_{MP}^2 + \Phi(x^*)^T \Sigma \Phi(x^*) \quad (8.56)$$

$$y^* = \mu^T \Phi(x^*) \quad (8.57)$$

Ying et al. [23] adopted RVM to establish an calibration model for predicting the content of elements in soil by laser-induced breakdown spectroscopy, and the results outperformed the SVM model and the least squares vector model in stability and

prediction accuracy. Wang et al. [24] selected characteristic variables through random forest, and then established a model of infrared (IR) spectroscopy for predicting the acid value of diesel engine oil by using the RVM, and the results were satisfactory. Zhu et al. [25] combined IR spectroscopy with RVM to identify the origin of mushroom, and the results were comparable to KNN and SVM, and the identification correct rate was higher than 90%. Zhu et al. [26] established a recognition model for the detection of infertile eggs and fertilized eggs based on the hyperspectral information fusion and RVM. It outperforms the SVM in both computational speed and recognition accuracy. Fu et al. [27] used near-infrared (NIR) spectroscopy combined with RVM to establish a model to judge the *Tetrastigma Hemsleyanum* of Chinese medicinal materials, with the recognition accuracy of 100%.

8.4 Kernel Partial Least Squares

The success of SVM in the field of machine learning has led people to “kernel-ization” various traditional linear methods to nonlinear ones by inner products. The idea of kernel function has gradually developed into kernel method, which provides a unified framework for dealing with many problems [28, 29]. As shown in Fig. 8.16, a variety of kernel function-based methods is derived by combining the kernel function with different chemometric algorithms, such as kernel principal component analysis (KPCA), kernel Fisher discriminant analysis (KFDA), kernel principal component regression (KPCR), kernel partial least squares (KPLS), and kernel ridge regression (KRR). The design of kernel function and chemometric algorithm can be carried out separately. To solve different problems, different kernel functions and chemometric algorithms can be selected. These methods have shown good performance in applications in many fields, among which the KPLS method is being adopted more and more [30–33].

For the training sets \mathbf{X} and \mathbf{Y} , given the kernel function type and the maximum number of principal factors f , the KPLS algorithm is as follows:

(1) Calculate the kernel matrix \mathbf{K} ($n \times n$) of matrix \mathbf{X} ($n \times m$, n is the number of samples in the training set, and m is the number of variables) by the kernel function.

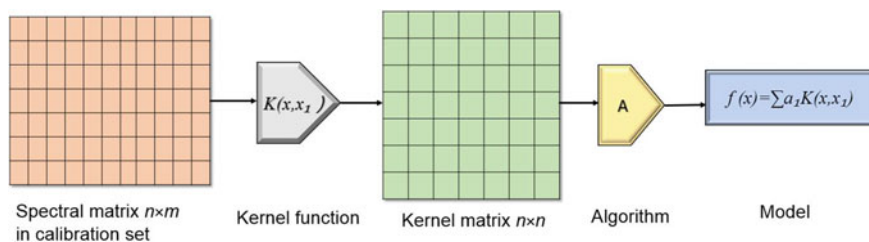


Fig. 8.16 Framework of implementation steps for calibration method based on kernel function

(2) Centralized processing of the kernel matrix \mathbf{K} by the following equation:

$$\tilde{\mathbf{K}} = \left(\mathbf{I} - \frac{1}{n} \mathbf{l} \mathbf{l}^T \right) \mathbf{K} \left(\mathbf{I} - \frac{1}{n} \mathbf{l} \mathbf{l}^T \right) \quad (8.58)$$

where \mathbf{I} is the unit matrix, and \mathbf{l} is the n -dimension full 1 column vector.

(3) Initialize variable \mathbf{u} .

(4) $\mathbf{t} = \tilde{\mathbf{K}} \mathbf{u}$, $\mathbf{t} = \mathbf{t} / \|\mathbf{t}\|$.

(5) $\mathbf{c} = \mathbf{Y}^T \mathbf{t}$.

(6) $\mathbf{u} = \mathbf{Y} \mathbf{c}$, $\mathbf{u} = \mathbf{u} / \|\mathbf{u}\|$.

(7) Repeat (3)–(6), until it converges.

(8) Calculate $\tilde{\mathbf{K}}$ and \mathbf{Y}

$$\tilde{\mathbf{K}} = (\mathbf{I} - \mathbf{t} \mathbf{t}^T) \tilde{\mathbf{K}} (\mathbf{I} - \mathbf{t} \mathbf{t}^T) \quad (8.59)$$

$$\mathbf{Y} = \mathbf{Y} - \mathbf{t} \mathbf{t}^T \mathbf{Y} \quad (8.60)$$

Back to (4) until all f , \mathbf{u} and \mathbf{t} vectors are obtained.

(9) The predicted value of the training set samples

$$\hat{\mathbf{Y}} = \tilde{\mathbf{K}} \mathbf{U} (\mathbf{T}^T \tilde{\mathbf{K}} \mathbf{U})^{-1} \mathbf{T}^T \mathbf{Y} \quad (8.61)$$

where $\mathbf{T} = [t_1, t_2, \dots, t_f]$, $\mathbf{U} = [u_1, u_2, \dots, u_f]$.

For the validation set \mathbf{X}_{test} ($p \times m$, p is the number of samples for the validation set, m is the number of variables), its kernel matrix \mathbf{K}_{test} is calculated by the kernel function. The kernel matrix \mathbf{K}_{test} is centralized by the following equation.

$$\tilde{\mathbf{K}}_{\text{test}} = \left(\mathbf{K}_{\text{test}} - \frac{1}{n} \mathbf{l} \mathbf{l}^T \mathbf{K} \right) \left(\mathbf{I} - \frac{1}{n} \mathbf{l} \mathbf{l}^T \right) \quad (8.62)$$

The predicted value of the samples in the validation set is calculated.

$$\hat{\mathbf{Y}} = \tilde{\mathbf{K}}_{\text{test}} \mathbf{U} (\mathbf{T}^T \tilde{\mathbf{K}} \mathbf{U})^{-1} \mathbf{T}^T \mathbf{Y} \quad (8.63)$$

Compared with ANN and SVM, KPLS has fewer parameters, faster calculation speed, and easier implement. It is promising to be a commonly used nonlinear multivariate calibration method.

8.5 Extreme Learning Machine

The extreme learning machine (ELM) is a single-hidden layer feedforward neural network, which overcomes the shortcomings of traditional BP-ANN, including slow training speed, easy to fall into local minima, and selection sensitivity of learning rate [34]. ELM randomly generates the connection weight between the input layer and the hidden layer and the threshold value of the hidden layer neurons, and obtains the output layer weight with a very small 2-norm through the Moore-Penrose generalized inverse. It is no need to adjust during the training process, the only need is to set the number of hidden layer neurons and the activation function of hidden layer neurons to obtain the unique optimal solution. Therefore, compared with traditional SLFNN, ELM has the characteristics of easy selection of parameters, fast learning speed, and strong generalization abilities. Figure 8.17 shows the schematic diagram of the ELM topology.

For a training set containing N calibration samples, \mathbf{x}_i is the spectral vector of the i th sample ($n \times 1$), containing n wavelength variables. \mathbf{y}_i is the concentration vector of the i th sample ($m \times 1$), containing m concentration samples. The standard SLFNN algorithm with H hidden nodes is as follows.

$$\sum_{i=1}^H \beta_i f_i(x_j) = \sum_{i=1}^H \beta_i f(a_i x_j + b_i), j = 1, 2, \dots, N \tag{8.64}$$

where $a_i = [a_{i1}, a_{i2}, \dots, a_{in}]^T$, is the weight of n -dimensional input layer and hidden layer, b_i is the threshold of i th node. $\beta_i = [\beta_{i1}, \beta_{i2}, \dots, \beta_{im}]^T$ is the weight of the m -dimensional output layer and hidden layer. $f_i(x_j)$ is the activation function.

The above equation can be expressed as

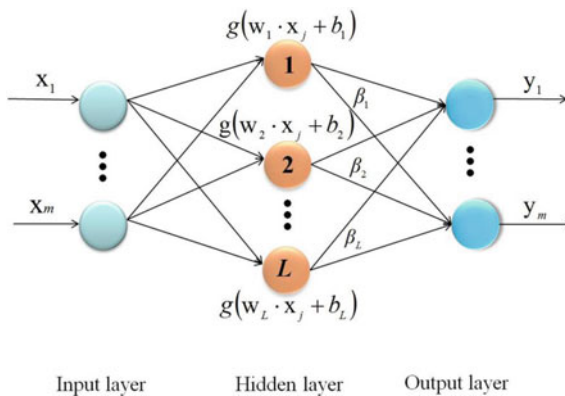


Fig. 8.17 Schematic diagram of ELM topological structure

$$\sum_{i=1}^H \beta_i f_i(x_j) = H\beta \quad (8.65)$$

where

$$\mathbf{H} = \begin{bmatrix} f(\mathbf{a}_1\mathbf{x}_1 + b_1) & \dots & f(\mathbf{a}_H\mathbf{x}_1 + b_H) \\ \dots & \dots & \dots \\ f(\mathbf{a}_1\mathbf{x}_N + b_1) & \dots & f(\mathbf{a}_H\mathbf{x}_N + b_H) \end{bmatrix} \quad (8.66)$$

The parameters of the ELM do not need to be fully adjusted throughout the training process, only the connection weights between the hidden layers and output layers need to be set, and their values can be obtained by the following equation:

$$\min_{\beta} \|\mathbf{H}\beta - \mathbf{Y}^T\| \quad (8.67)$$

The solution is

$$\hat{\beta} = \mathbf{H}^+ \mathbf{Y}^T \quad (8.68)$$

where \mathbf{H}^+ is the Moore-Penrose generalized inverse of the hidden layer output matrix \mathbf{H} , and \mathbf{Y}^T is the transpose of the output matrix \mathbf{Y} .

Kernel extreme learning machine (KELM) introduces kernel functions into ELM. It replaces random mapping in ELM with kernel mapping and uses kernel functions to map all input samples from N -dimensional space to high-dimensional hidden layer feature space, which effectively improve the unsatisfactory generalization ability and stability problems caused by randomization of hidden layer parameters. By solving it at once, the method can obtain the least square solution of the weight, which is faster and the generalization performance is more stable than the ELM algorithm. The type and parameters of kernel function are the main factors of the performance of the KELM, once the parameters are selected. The results are stabilized and no longer mixed with random. The parameters of the KELM can be optimized selection by using optimization algorithms such as cuckoo search (CS) algorithm, PSO algorithm, chaos particle swarm optimization (CPSO), and GA [35, 36].

Zhu et al. [37] used ultraviolet-visible spectroscopy technology and ELM to identify the sex of embryonated chicken eggs in the early incubation period, and adopted GA to optimize the weight variables of the ELM model and the threshold of hidden layer neurons. The accuracy of the recognition rate was above 85%. Han et al. [38] combined NIR spectroscopy with ELM to quickly identify adulterated pork in beef, and obtained satisfactory identification results. Lu et al. [39] adopted the compressed self-encoding network (CAE) in deep learning to extract the deep features in the NIR spectral data of citrus leaves, and then sent the extracted deep features into the ELM model for identification. The established citrus yellow-shoot disease identification model (CAE-ELM) had great robustness and scalability. Pan et al. [40] had carried

out the study of PCA combined with ELM assisted laser-induced breakdown spectroscopy (LIBS) in the classification and identification of aluminum alloys. Results showed that the PCA-ELM classification model has high accuracy and stability and can finely classify waste aluminum according to their respective component brands. Based on laser-induced fluorescence spectroscopy (LIF), Zhou et al. [41] successfully identified the types of edible oil using KELM. Liang et al. [42] used laser-induced breakdown spectroscopy combined with KELM to classify the origin of the root of red-rooted salvia, and the results were better than the least squares support vector machine and random forest.

Rao et al. [43] combined stack auto-encoders (SAE) with ELM to establish a depth neural network prediction model (SAE-ELM) for hyperspectral imaging to predict apple hardness, which had better prediction performance than traditional ELM model. Xia et al. [44] used ELM to build models for the IR spectral data of lubricating oil, which can effectively identify the types and predict the additives content in lubricating oil. Wang et al. [45] optimized the connection weight and threshold of ELM network through GA, improved the instability of the prediction results caused by the randomness of ordinary ELM connection weight and threshold, and established a model of NIR spectroscopy to predict the moisture content of jujube. Wei et al. [46] adopted KELM optimized by GA to establish a model for hyperspectral nondestructive detection of the total number of bacteria on the surface of cooled mutton.

In order to further improve the prediction accuracy and stability of ELM, Bian and Chen [47, 48] proposed ensemble ELM modeling methods for quantitative analysis of NIR spectroscopy, respectively. Shan et al. [49] proposed stacked ensemble ELM (SE-ELM), where NIR spectra were divided into segments and multiple ELM models were built, and then these models were combined with different weights (Fig. 8.18) to further improve the generalization performance of the model. Hu et al. [50] combined the stacked partial least square regression based on the variable importance in the projection (VIP-SPLS) with the ELM algorithm to propose an improved extreme learning machine (iELM). This method used VIP-SPLS algorithm to establish a regression model between the hidden layer output matrix \mathbf{H} and the concentration to be measured, which replaced the calculation process of matrix \mathbf{H} generalization inverse, and solved the problem of high dimensionality and highly collinearity of the hidden layer output matrix due to the large number of NIR spectral variables to a certain extent.

8.6 Gaussian Process Regression

Gaussian process regression (GPR) is a machine learning method that has been developing continuously in the past decade and has received more and more attention. It combines the related theories and methods of kernel-based machine learning and Bayesian-based machine learning and has the advantages of the above two machine learning methods. It has a strict theoretical basis of statistics and is suitable

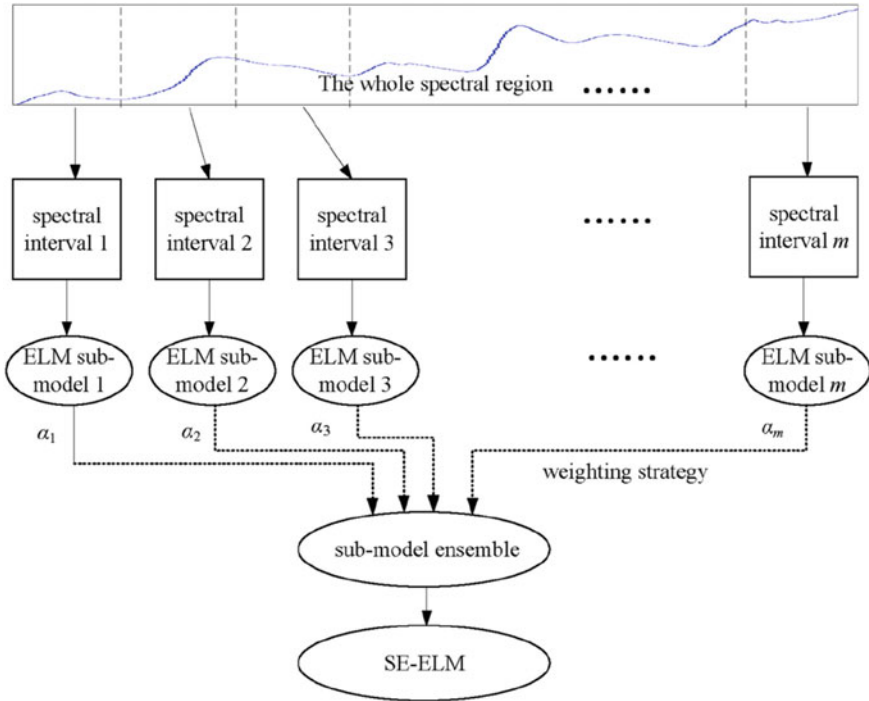


Fig. 8.18 Schematic diagram for the framework of SE-ELM model

for dealing with complex learning problems, such as nonlinearity, small samples, and high dimensions, and has strong generalization ability. Compared with neural networks, SVM, and other methods, this method has the advantages of easy implementation, self-adaptive hyperparameters, flexible non-parametric inference, and statistical significance of the prediction results [51, 52]. At present, GPR has been applied to the regression and classification and other fields and has become a research hotspot in the field of machine learning at home and abroad [53, 54].

GPR treats $f(x)$ values directly in function space as random variables and a priori distribution of $f(x)$ as Gaussian distributions, based on the following basic principles:

Given a set of training samples $D = \{(x_i, y_i) | i = 1, 2, \dots, n\} = (X, y)$, the regression model can be represented as

$$y = f(x) + \varepsilon, \varepsilon \sim N(0, \sigma^2) \tag{8.69}$$

where \mathbf{X} is the $n \times d$ -dimensional matrix composed of the input vector \mathbf{x}_i , \mathbf{y} is the n -dimensional vector consisting of output scalar y_i , n is the number of training set samples, d is the number of spectral variables and ε is Gaussian white noise with σ_n^2

variance. A Gaussian process is determined entirely by its mean function $m(\mathbf{x})$ and covariance function $k(\mathbf{x}, \mathbf{x}')$. For easy to calculate, the mean function $m(\mathbf{x})$ is usually treated as 0,

$$f(x) \sim GP(0, k(x, x')) \tag{8.70}$$

The priori distribution of the training set output value \mathbf{y} is

$$y \sim N(0, K + \sigma_n^2 I) \tag{8.71}$$

where \mathbf{I} is an unit matrix, $\mathbf{K} = \mathbf{K}(\mathbf{X}, \mathbf{X}) = k(x_i, x_j)_{n \times n}$ is the symmetrical covariance matrix, representing the correlation between \mathbf{x}_i and \mathbf{x}_j .

For the input vector \mathbf{x}^* of the sample to be tested, the joint Gaussian distribution composed by the corresponding output value y^* and the sample output \mathbf{y} of the training set is

$$\begin{bmatrix} y \\ y^* \end{bmatrix} \sim N(0, \begin{pmatrix} \mathbf{K}(\mathbf{X}, \mathbf{X}) + \sigma_n^2 I & \mathbf{K}(\mathbf{X}, \mathbf{x}^*) \\ \mathbf{K}(\mathbf{x}^*, \mathbf{X}) & k(\mathbf{x}^*, \mathbf{x}^*) \end{pmatrix}) \tag{8.72}$$

where $\mathbf{K}(\mathbf{X}, \mathbf{x}^*) = \mathbf{K}(\mathbf{x}^*, \mathbf{X})$ is the $n \times 1$ -dimensional covariance matrix between the training set sample \mathbf{X} and the sample \mathbf{x}^* to be tested, and the $k(\mathbf{x}^*, \mathbf{x}^*)$ is the autocovariance of the sample \mathbf{x}^* to be tested.

The posterior distribution of y^* can be obtained by the Bayesian principle:

$$y^* | \mathbf{X}, y, \mathbf{x}^* \sim N(\widehat{y}^*, \text{cov}(y^*)) \tag{8.73}$$

where

$$\widehat{y}^* = \mathbf{K}(\mathbf{x}^*, \mathbf{X})[\mathbf{K}(\mathbf{X}, \mathbf{X}) + \sigma_n^2 I]^{-1} \mathbf{y} \tag{8.74}$$

$$\text{cov}(y^*) = k(\mathbf{x}^*, \mathbf{x}^*) - \mathbf{K}(\mathbf{x}^*, \mathbf{X})[\mathbf{K}(\mathbf{X}, \mathbf{X}) + \sigma_n^2 I]^{-1} \mathbf{K}(\mathbf{X}, \mathbf{x}^*) \tag{8.75}$$

where \widehat{y}^* and $\text{cov}(y^*)$ represent Gaussian regression models for the predicted output values and predicted variances of the sample \mathbf{x}^* to be tested, respectively.

In GPR, the selection of covariance function (also known as kernel function) and related parameters determines the fundamental performance of Gaussian process model being built. The most commonly used covariance function is the square exponential covariance function:

$$k(x_p, x_q) = \sigma_f^2 \exp\left(-\frac{1}{2l^2}(x_p - x_q)^2\right) \quad (8.76)$$

where σ_f^2 is the overall measure of prior knowledge, and l is the degree to control the local relevance.

The value of the hyperparameter $\Theta = (\sigma_f, l)$ has a great influence on the prediction effect of the model, and the negative logarithm likelihood function $L(\Theta)$ is generally used as the optimization target function of the hyperparameter. $L(\Theta)$ is

$$L(\Theta) = -\frac{1}{2}y^T(K + \sigma_n^2I)^{-1}y - \frac{1}{2}\lg|K + \sigma_n^2I| - \frac{n}{2}\lg 2\pi \quad (8.77)$$

Calculate the partial derivative of each parameter of the negative logarithm likelihood function $L(\Theta)$, and then use the conjugate gradient iteration method to obtain the optimal hyperparameter $\hat{\Theta}$. By obtaining the optimal hyperparameter, the results of GPR can be obtained by calculating the equations of \hat{y}^* and $\text{cov}(y^*)$.

GPR model parameters are usually obtained by conjugate gradient method. However, the optimization effect of conjugate gradient is strongly dependent on the initial value, and there are disadvantages of difficult to determine the number of iterations and easy to fall into the local optimal. Therefore, the parameters of GPR model can be optimized by using optimization algorithms such as PSO.

Martinez-Espana et al. [55] adopted GPR methods to establish a quantitative model for the prediction of soil critical properties by portable infrared spectroscopy, and the results were better than random forests and PLS. Ying et al. [56] used microwave plasma torch atomic emission spectroscopy (MPT-AES) combined with GPR to establish a model for predicting the content of five chemical elements in ginseng, and obtained better results than SVR. Li et al. [57] used NIR spectroscopy combined with the PLS, LS-SVR, and GPR method to establish a calibration model to predict the content of active ingredients in Tanreqing injections, and the results showed that LS-SVR and GPR method gave a better prediction result. Xu et al. [58] combined the wavelength selection strategy of synergy interval with GPR (SiGPR) and established a model of NIR spectroscopy to predict the moisture content and pH value of the solid-state fermentation process of the monascus. This method could effectively select the wavelength range and improve the accuracy of the NIR calibration model.

At present, there are some shortcomings in GPR, such as limitation to the assumption of Gaussian noise distribution. Some improved algorithms that reduce the amount of calculation and break through the assumption of Gaussian noise distribution have been proposed, prompting the continuous development of GPR models. For example, Liu et al. [59] proposed a new noise-level-penalizing robust Gaussian process regression method (NLP-RGP), which can better deal with training set with outliers.

References

1. Lopes MB, Calado CRC, Figueiredo MAT, et al. Does nonlinear modeling play a role in plasmid bioprocess monitoring using fourier transform infrared spectra? *Appl Spectrosc.* 2017;71(6):1148–56.
2. Cui C, Fearn T. Hierarchical mixture of linear regressions for multivariate spectroscopic calibration: an application for NIR calibration. *Chemom Intell Lab Syst.* 2018;174:1–14.
3. Balabin RM, Safieva RZ. Near-infrared (NIR) spectroscopy for biodiesel analysis: fractional composition, iodine value, and cold filter plugging point from one vibrational spectrum. *Energy Fuels.* 2011;25:2373–82.
4. Du Y, Meng XC, Zhu LQ. Overlapping spectral analysis based on genetic algorithms and BP neural networks. *Spectrosc Spect Anal.* 2020;40(7):2066–72.
5. Li YN, Xu ZB, Liu SL. Determination of lead and arsenic in iron ore by X-ray fluorescence spectrometry based on genetic neural network. *Metallurg Anal.* 2017;37(10):22–6.
6. Zou HM, Li XC, S X, et al. Hyperspectral estimation of soil organic matter based on particle swarm optimization neural network. *Sci Surv Map.* 2019;44(5):146–50.
7. Wang XC, S F, YL, et al. MATLAB neural network analysis of 43 cases. Beihang University Press. 2013.
8. Zhang XG, Bian ZQ. Mode recognition. 2nd ed. Beijing: Tsinghua University Press; 2004.
9. Li H, Liang Y, Xu Q. Support vector machines and its applications in chemistry. *Chemom Intell Lab Syst.* 2009;95(2):188–98.
10. Brereton RG, Lloyd GR. Support vector machines for classification and regression. *Analyst.* 2010;135(2):230–67.
11. Luca F, Conforti M, Castrignano A, et al. Effect of calibration set size on prediction at local scale of soil carbon by vis-NIR spectroscopy. *Geoderma.* 2017;288:175–83.
12. Wu D, He Y, Feng S, et al. Study on infrared spectroscopy technique for fast measurement of protein content in milk powder based on LS-SVM. *J Food Eng.* 2008;84(1):124–31.
13. Ferrão MF, Godoy SC, Gerbase AE, et al. Non-destructive method for determination of hydroxyl value of soybean polyol by LS-SVM using HATR/FT-IR. *Anal Chim Acta.* 2007;595(1–2):114–9.
14. Chauchard F, Cogdill R, Roussel S, et al. Application of LS-SVM to non-linear phenomena in NIR spectroscopy: development of a robust and portable sensor for acidity prediction in grapes. *Chemom Intell Lab Syst.* 2004;71(2):141–50.
15. Chen Y, Yan X, Zhang X, et al. Surface-enhanced raman spectroscopy quantitative analysis of polycyclic aromatic hydrocarbons based on support vector machine algorithm. *Chin J Lasers.* 2019;46(3):298–305.
16. Cao L, Ouyang XY. Parameters optimization of SVM based on genetic algorithm. *Comput Digit Eng.* 2016;44(4):575–8.
17. Wang ST, Liu N, Cheng Q, et al. Classification and identification of polycyclic aromatic hydrocarbons by three-dimensional fluorescence spectroscopy combined with GA-SVM. *Spectrosc Spect Anal.* 2020;40(4):1149–55.
18. Hu XH, Liu W, Liu CH, et al. Rapid identification of producing area of coffee bean based on terahertz spectroscopy and support vector machine. *Trans Chinese Soc Agricult Eng.* 2017;33(9):302–7.
19. Zhou HM, Chen TB, Liu MH, et al. Quantitative analysis of chromium in rice husks by laser induced breakdown spectroscopy based on particle swarm optimization-support vector machine. *Chin J Anal Chem.* 2020;48(6):811–6.
20. Hernández N, Talavera I, Dago A, Biscay RJ, et al. Relevance vector machines for multivariate calibration purposes. *J Chemom.* 2008;22(11–12):686–94.
21. Caesarendra W, Widodo A, Yang BS. Application of relevance vector machine and logistic regression for machine degradation assessment. *Mech Syst Signal Process.* 2010;24:1161–71.
22. Hou MM, Yu QB, Jiao SJ, et al. Near infrared spectroscopy measurement of moisture content in turbine oil using relevance vector machines. *Chinese J Chongqing Ind Commer Univ.* 2012;29(3):94–8.

23. Ying LN, Zhou WD. Comparative analysis of multiple chemometrics methods in application of laser-induced breakdown spectroscopy for quantitative analysis of soil elements. *Acta Opt Sin.* 2018;38(12):1214002.
24. Wang JX, Wang K, Han X. Rapid determination of acid value of diesel engine oil by infrared spectroscopy with application of algorithm of RF-RVM. *Phys Test Chem Anal.* 2019;55(1):26–30.
25. Zhu ZY, Zhang C, Liu F, et al. Study on mushroom origin identification based on mid-infrared spectroscopic analysis technology. *Spectrosc Spect Anal.* 2014;34(3):664–7.
26. Zhu ZH, Liu T, Ma MH. Hatching eggs nondestructive detection based on hyperspectral-imaging information and RVM. *Trans Chinese Soc Agricult Eng.* 2015;15:293–300.
27. Fu CL, Li Y, Wang W, et al. Use of fourier transform near-infrared spectroscopy combined with a relevance vector machine to discriminate *Tetrastigma Hemsleyanum* (Sanyeqing) from other related species. *Anal Methods.* 2017;9:4023–7.
28. Wang HZ, Yu JS. Application of kernel based methods to soft sensor modeling of selectivity to acrylonitrile. *Process Autom Instrum.* 2005;31(3):367–70.
29. Yang HH, Wang XY, Wang Y, et al. PCA /KPCA feature extraction approach to SVM for anomaly detection. *Control Decis.* 2005;20(3):251–6.
30. Rosipal R. Kernel partial least squares for nonlinear regression and discrimination. *Neural Netw World.* 2003;13(3):291–300.
31. Rosipal R, Trejo LJ. Kernel partial least squares regression in reproducing kernel hilbert space. *Mach Learn Res.* 2002;2(2):97–123.
32. Kim K, Lee JM, Lee IB. A novel multivariate regression approach based on kernel partial least squares with orthogonal signal correction. *Chemom Intell Labor Syst.* 2005;79(1/2):22–30.
33. Shinzawa H, Jiang JH, Ritthiruangdej P, et al. Investigations of bagged kernel partial least squares (KPLS) and boosting KPLS with applications to near-infrared (NIR) spectra. *J Chemom.* 2007;20(8–10):436–44.
34. Zheng WB, Shu HP, Tang H, et al. Spectra data classification with kernel extreme learning machine. *Chemom Intell Labor Syst.* 2019;192:103815.
35. Zhang SY, Tan WA, Wang N. Combined kernel extreme learning machine based on cuckoo search algorithm parameter optimization. *J Jilin Univ (Sci Ed).* 2019;57(5):1185–92.
36. Miao FJ, Sun TR, Tao BR, et al. Algorithmic research on kernel extreme learning machine for speaker recognition based on PSO and PCA optimization. *Sci Technol Eng.* 2019;19(21):195–9.
37. Zhu ZH, Hong Q, Wu LF, et al. Early identification of male and female embryos based on UV/Vis transmission spectroscopy and extreme learning machine. *Spectrosc Spect Anal.* 2019;39(9):2780–7.
38. Han FK, Liu C, Huang Y, et al. Rapid identification of beef adulteration with pork by near-infrared combined limit learning machine. *Anhui Agricult Sci.* 2019;47(5):183–182.
39. Lu HX, Xu MC, Zhang WD, et al. Identification of citrus huang long bing based on contractive auto-encoder combined extreme learning machine. *Chin J Anal Chem.* 2019;47(5):652–60.
40. Pan LJ, Chen WF, Cui RF, et al. Application of laser-induced breakdown spectroscopy assisted by principal component analysis and extreme learning machine in the classification recognition of aluminum alloy. *Metallurg Anal.* 2020;40(1):1–6.
41. Zhou MR, Wang JG, Song HP, et al. Application of kernel extreme learning machine and laser induction fluorescence technique in edible oil identification. *Laser Optoelectron Progr.* 2020;57(17):173002.
42. Liang J, Yan CH, Zhang Y, et al. Rapid discrimination of *Salvia Miltiorrhiza* according to their geographical regions by laser induced breakdown spectroscopy (LIBS) and particle Swarm optimization-kernel extreme learning machine (PSO-KELM). *Chemom Intell Labor Syst.* 2020;197:103930.
43. Rao LB, Pang T, Ji RS, et al. Firmness detection for apples based on hyperspectral imaging technology combined with stack autoencoder-extreme learning machine method. *Laser Optoelectron Progr.* 2019;56(11):113001
44. Xia YQ, Xu DY, Feng X, et al. Identification and content prediction of lubricating oil additives based on extreme learning machine. *Tribology.* 2020;40(1):97–106.

45. Wang WX, Ma BX, Luo XZ, et al. Study on the moisture content of dried Hami big jujubes by near-infrared spectroscopy combined with variable preferred and GA-ELM model. *Spectrosc Spect Anal.* 2020;40(2):543–9.
46. Wei J, Guo ZH, Xu J. Measurement of total viable count on chilled mutton surface based on hyperspectral technique and extreme learning machine. *Jiangsu Agricult Sci.* 2018;46(24):211–4.
47. Bian XH, Zhang CX, Tan XY, et al. Boosting extreme learning machine for near-infrared spectral quantitative analysis of diesel fuel and edible blend oil samples. *Anal Methods.* 2017;9(20):2983–9.
48. Chen H, Tan C, Lin Z. Ensemble of extreme learning machines for multivariate calibration of near-infrared spectroscopy. *Spectrochimica Acta Part A: Molec Biomolec Spectrosc.* 2020;229:117982.
49. Shan P, Zhao YH, Wang QY, et al. Stacked ensemble extreme learning machine coupled with partial least squares-based weighting strategy for nonlinear multivariate calibration. *Spectrochim Acta Part A Mol Biomol Spectrosc.* 2019;215:97–111.
50. Hu BX, Zhang HG, Lu JG, et al. Novel modeling method based on improved extreme learning machine algorithm for gasoline octane number detection by near infrared spectroscopy. *J Nanjing Univ Sci Technol.* 2017;41(5):660–5.
51. Feng AM, Fang LM, Lin M. Gaussian process regression and its application in near-infrared spectroscopy analysis. *Spectrosc Spect Anal.* 2011;31(6):1514–15171.
52. Ni WD, Norgaard L, Morup M. Non-linear calibration models for near infrared spectroscopy. *Anal Chim Acta.* 2014;813:1–14.
53. Chen T, Morris J, Martin E. Gaussian process regression for multivariate spectroscopic calibration. *Chemom Intell Lab Syst.* 2007;87(1):59–71.
54. Cui C, Fearn T. Comparison of partial least squares regression, least squares support vector machines, and gaussian process regression for a near infrared calibration. *J Near Infrared Spectrosc.* 2017;25(1):5–14.
55. Ying YW, Jin W, Yan YW, et al. Gaussian process regression coupled with MPT-AES for quantitative determination of multiple elements in Ginseng. *Chemom Intell Lab Syst.* 2018;176:82–8.
56. Martinez-Espana R, Bueno-Crespo A, Soto J, et al. Developing an intelligent system for the prediction of soil properties with a portable mid-infrared instrument. *Biosys Eng.* 2019;177:101–8.
57. Li WL, Yan X, Pan JC, et al. Rapid analysis of the Tanreqing injection by near-infrared spectroscopy combined with least squares support vector machine and gaussian process modeling techniques. *Spectrochim Acta Part A Mol Biomol Spectrosc.* 2019;218:271–80.
58. Xu C, Yin YY, Liu F. Near infrared spectroscopy wavelength selection method and the application based on synergy interval gaussian process. *Spectrosc Spect Anal.* 2016;36(8):2437–41.
59. Liu C, Yang SX, Li XF, et al. Noise level penalizing robust gaussian process regression for NIR spectroscopy quantitative analysis. *Chemom Intell Labor Syst.* 2020;201:104014.

Chapter 9

Method of Selecting Calibration Samples



9.1 Introduction

In the process of establishing the calibration model, it is very necessary to select the samples that participate in the calibration set to establish a robust model. As shown in Fig. 9.1, the selection of samples is to select the row vector of spectral matrix \mathbf{X} and the row vector of corresponding concentration matrix \mathbf{y} , and the selection of wavelength is to select the column vector of spectral matrix \mathbf{X} .

The established methods combined spectra and chemometrics are used to analyze most complex analytical systems, such as gasoline, wheat, and tobacco. For this kind of calibration samples that cannot be obtained by manual preparation, actual samples must be collected. With routine laboratory analysis of samples, thousands of samples can be obtained in a few months, but it is possible that more than 80% of these samples are duplicates. Therefore, it is necessary to select strong representative samples to establish a calibration model, which can not only improve the speed of model establishment but also reduce the storage space of the model library. More importantly, when the sample outside the model boundary is encountered, the application range of the model can be expanded through fewer samples, which is convenient for model update and maintenance [1, 2]. In addition, if the collected samples do not have the corresponding basic concentration data, the cost will be huge if all samples are analyzed and tested without screening.

The ideal calibration set should meet the following conditions: (1) The samples in calibration set should include all possible composition of samples to be tested in the future; (2) Its concentration (or property) range should exceed the situation that may be encountered in the sample to be tested in the future (generally, its standard deviation should be greater than 5 times of the reproducibility of the reference method); (3) The physicochemical parameters of the samples in the calibration set should be evenly distributed as shown in Fig. 9.2a; (4) The calibration set should have enough number of samples to statistically determine the mathematical relationship between the spectral variable and the concentration (or properties) (usually the number is not less than $6(f + 1)$, and f is the number of PLS factor). In the actual production

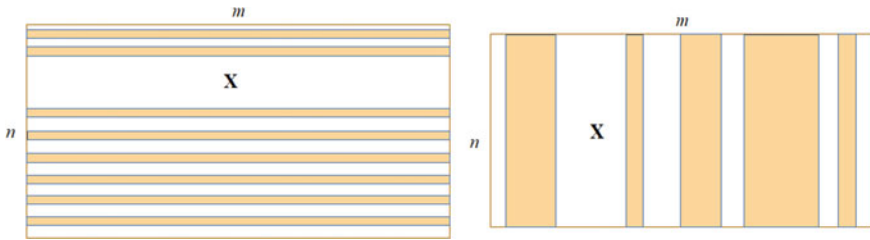


Fig. 9.1 Schematic diagram of selection of calibration samples (left) and spectral variables (right) for spectral matrix X

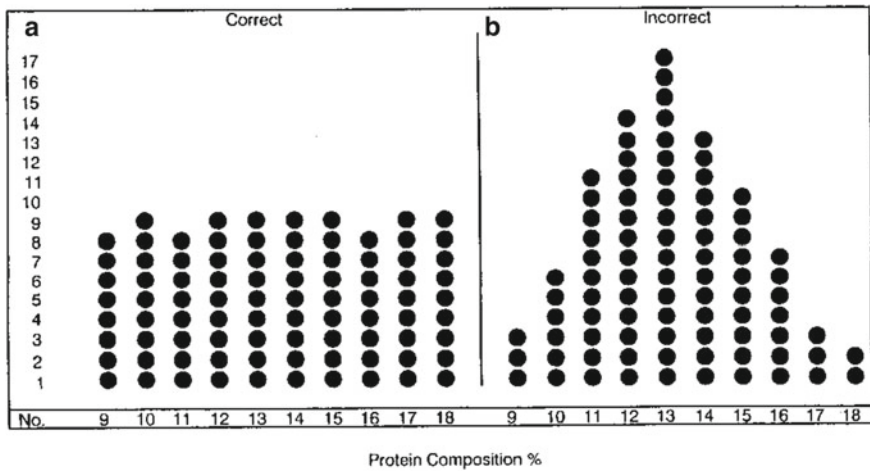


Fig. 9.2 Schematic diagram of uniform distribution (a) and Gaussian distribution (b) of sample concentration in calibration set [3]

process, especially in the large-scale process industry, the composition concentration of the collected samples is mostly Gaussian as shown in Fig. 9.2b. If these samples are not selected to directly participate in the establishment of the calibration model, the “Dunne effect” phenomenon is likely to occur in the prediction, that is, the regression prediction results tend to the central value as Fig. 9.3 displays [3]. In addition, the number of samples in the calibration set should be large enough to statistically determine the quantitatively functional relationship between the spectral variables and the physicochemical parameters to be calibrated.

It is difficult to obtain a relatively ideal sample set by random selection, and it is often not satisfactory to select calibration samples only according to the concentration distribution, because the spectra of two samples with the same concentration may be quite different. Currently, the most commonly used method is the Kennard-Stone selection method based on spectral variables [4–7].

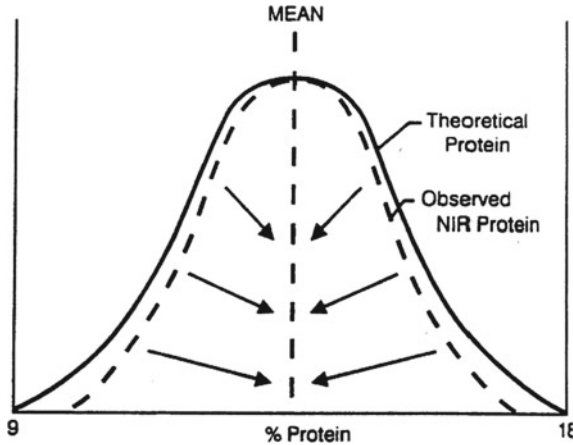


Fig. 9.3 Schematic diagram of Dunne effect [3]

Before selecting calibration samples, outliers should be eliminated first. These abnormal samples may contain abnormal chemical components or have extreme concentrations that are significantly different from other samples. If these abnormal samples participate in the establishment of the model, the accuracy and robustness of the calibration model will be affected.

In addition to the optimization selection of calibration set, test set and validation set also need to select samples with strong representativeness. In fact, the representativeness requirement of set test or validation set samples is not lower than that of calibration set. The method of selecting calibration samples can be applied to the selection of samples for validation set and test set. For example, the Duplex method use Kennard-Stone (K-S) methods alternately to divide samples into calibration and validation set [8].

As shown in Fig. 9.4, the uniformity and representativeness of the selected samples can be checked through the concentration value distribution map of the samples of calibration set or validation set. As shown in Fig. 9.5, the concentration value distribution of the samples of calibration set or validation set can also be checked by the violin plots, which are a combination of box plots and kernel density plots, and can show the concentration and dispersion of data.

In addition, the consistency of sample distribution of calibration and validation set can be evaluated by skewness and kurtosis.

Skewness is a measure of the direction and degree of the distribution skew of a set of data, namely the numerical characteristics of the degree of the asymmetry of variables. The greater the skewness, the stronger the data asymmetry. As shown in Fig. 9.6, the definition of skewness includes a normal distribution (skewness = 0), a right skewness distribution (also called a positive skewness distribution, with skewness > 0), and a left skewness distribution (also called a negative skewness distribution, with skewness < 0).

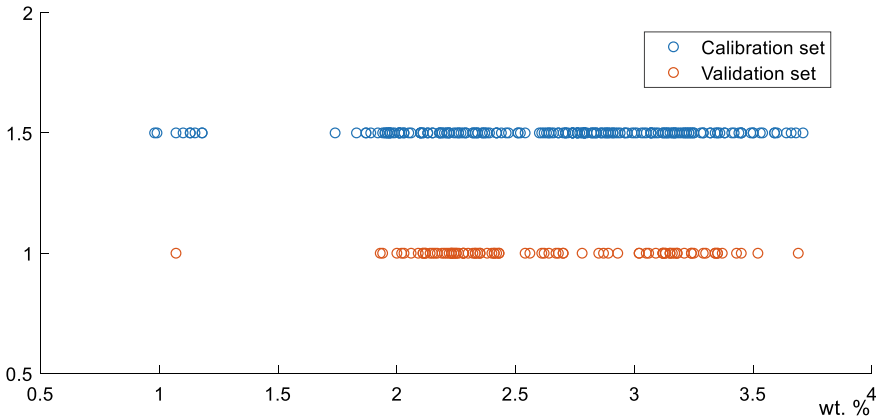


Fig. 9.4 Concentration value distribution diagram of calibration set and validation set samples

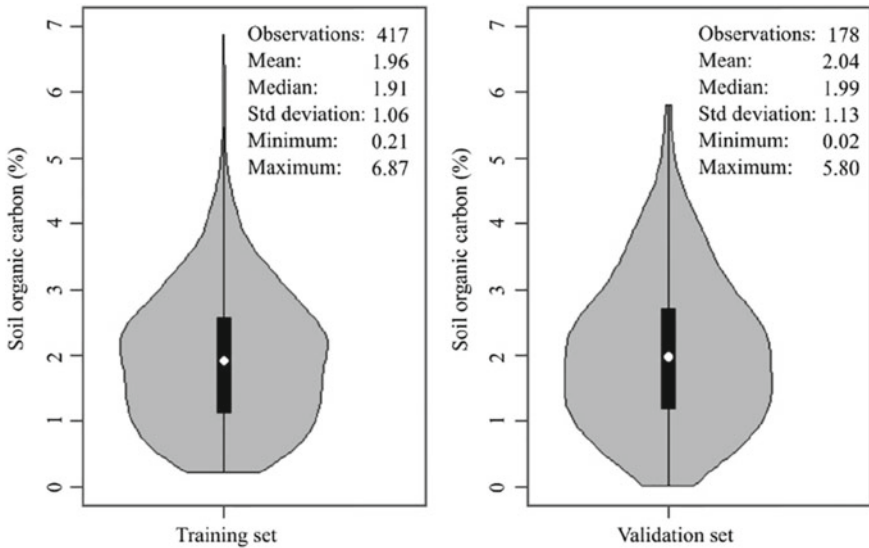


Fig. 9.5 Violin diagrams of calibration set and validation set samples [9]. The dark area indicates the inter-quartile range and the white dot indicates the median value of the dataset

The calculation formula of skewness is as follows:

$$I_{skewness} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^3 \left(\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \right)^{3/2} \tag{9.1}$$

Kurtosis, also known as kurtosis coefficient, is a statistical measure to describe the steepness of the distribution pattern of a set of data. Kurtosis reflects the sharpness

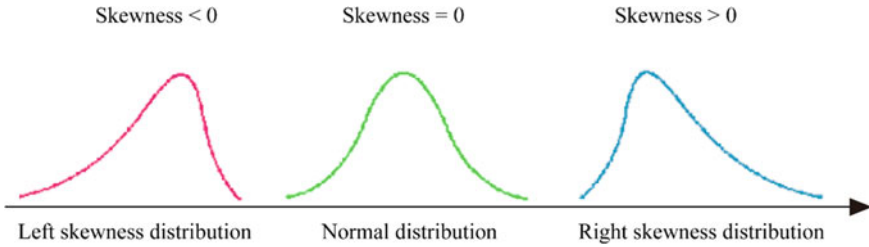


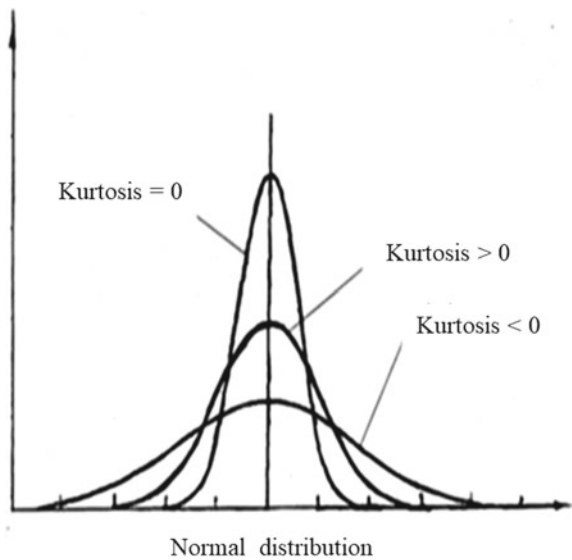
Fig. 9.6 Skewness characteristics of the distribution of a set of data

of the peak. The greater the kurtosis, the steeper the data distribution curve, which is compared to a normal distribution. As shown in Fig. 9.7, kurtosis includes normal distribution (kurtosis value = 0), positive kurtosis (kurtosis value > 0), and negative kurtosis (kurtosis value < 0). The calculation formula of kurtosis is as follows:

$$I_{kurtosis} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^4 / \left(\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \right)^2 - 3 \tag{9.2}$$

Homogeneity test of variance for the samples of calibration and validation set is evaluated by Levene test which can be used for the sample of normal distribution and also for the sample of non-normal distribution. At the same time, the size of two groups of samples for comparison cannot be equal.

Fig. 9.7 Kurtosis characteristics of the distribution of a set of data



9.2 Kennard-Stone Method

Based on the Euclidean distance between variables, the Kennard-Stone (K-S) method [4] uniformly selects samples in the feature space. The spectrum can be directly used as the characteristic variable, or the principal component score can be selected as the characteristic variable after the analysis of the spectrum by PCA. The selection process of calibration samples by K-S method is as follows:

Assume that there is a total of Z samples, from which n samples should be selected into calibration set.

- (1) First, calculate the Euclidean distance d_{ij} between pairs of all samples, and select the two samples with the longest distance, $Z1$ and $Z2$, to enter the calibration set.
- (2) Calculate the distance between the remaining ($Z2$) samples and the two selected samples $Z1$ and $Z2$ and take their minimum values as $\min(d_{i, Z1}, d_{i, Z2})$, and then select a sample $Z3$ corresponding to the maximum value as $\max(\min(d_{i, Z1}, d_{i, Z2}))$ to enter the calibration set.
- (3) Calculate the distance between the remaining ($Z3$) samples and the three selected samples $Z1$, $Z2$, and $Z3$ and take their minimum value as $\min(d_{i, Z1}, d_{i, Z2}, d_{i, Z3})$, and then select a sample $Z4$ corresponding to the maximum value as $\max(\min(d_{i, Z1}, d_{i, Z2}, d_{i, Z3}))$ to enter the calibration set.
- (4) Repeat the above process until n samples are selected into calibration set.

Figure 9.8 shows PCA results of NIR spectra of 210 calibration samples selected from 300 tobacco samples by K-S method.

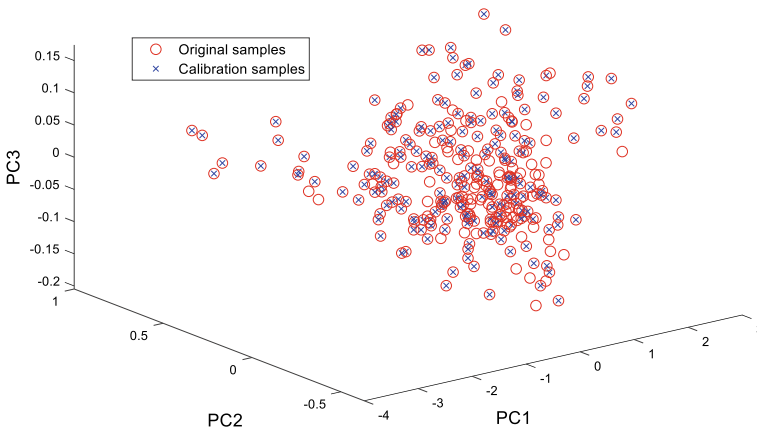


Fig. 9.8 Result of Selecting Calibration Samples (O-Original samples, × -Selected calibration samples) by K-S method

K-S method is usually based on spectral variables for distance calculation. In order to obtain representative samples in concentration space, concentration array can be used to replace spectral matrix in distance calculation [10].

Duplex method alternately applies K-S method to the selection of samples into calibration set and validation set to ensure that both calibration set and validation set sample are representative.

9.3 Sample Set Partitioning Based on Joint X–Y Distances (SPXY) Method

K-S method selects samples based on the spectral characteristics without considering the influence of concentration array. For low content components, if the spectral characteristics are not significant, K-S method may not obtain satisfactory samples in calibration set. Galvao et al. proposed Sample set partitioning based on joint x–y distances (SPXY) method on the basis of K-S method [11]. The step-by-step selection process of SPXY method is the same as that of K-S method, except that the newly defined $d_{xy}(i, j)$ is used when calculating the distance between samples as follows:

$$d_{xy}(i, j) = \frac{d_x(i, j)}{\max_{i,j \in (1,z)}(d_x(i, j))} + \frac{d_y(i, j)}{\max_{i,j \in (1,z)}(d_y(i, j))}, i, j \in [1, Z] \quad (9.3)$$

In the formula, $d_x(i, j)$ is the distance between samples calculated with spectra as characteristic parameters, and $d_y(i, j)$ is the distance between samples calculated with concentration as characteristic parameters. In order to make the samples have the same weight in the spectral space and the concentration space, they are divided by their respective maximum values for normalization processing, respectively. In order to highlight the role of spectral space or concentration space, the weighted method can be used to select samples as follows [12]:

$$d_{xy}(i, j) = \alpha \frac{d_x(i, j)}{\max_{i,j \in (1,z)}(d_x(i, j))} + \frac{d_y(i, j)}{\max_{i,j \in (1,z)}(d_y(i, j))} (1 - \alpha), i, j \in [1, Z] \quad (9.4)$$

where α is the weighting factor, and $0 \leq \alpha \leq 1$.

9.4 Optimizable K-dissimilarity Selection Method

When selecting calibration samples, both representativeness and diversity of samples need to be taken into consideration. The so-called representativeness means that the selected samples should reflect the attributes of all the samples in the whole dataset

as much as possible. Whereas, diversity means that the differences between selected samples should be as large as possible, so that they can be easily distinguished from each other. Optimizable K-dissimilarity selection (OptiSim) is a method that can select both representative and diversified samples [13, 14].

OptiSim algorithm involves three parameters: K defines the size of the subset of samples in each iteration; R defines the minimum similarity that is allowed between a valid candidate sample and a selected sample; M is the total number of samples of the selected representative subset. The algorithm is described as follows:

- (1) Select a sample randomly from the sample set, create a candidate sample buffer pool in the remaining dataset, and create an empty recycle bin and subset of samples.
- (2) A sample is randomly taken from the candidate buffer pool. If it is more similar to any selected sample than R , it will be discarded and put into the recycling bin. Otherwise, add it to the subsample set.
- (3) Repeat Step (2) until the subset includes K samples or the candidate buffer pool is exhausted.
- (4) If the number of samples in the subset is less than K and the candidate buffer pool is exhausted, all samples are taken from the recycle bin and put into the candidate buffer pool. Step (2) is returned.
- (5) If the subset is empty, exit.
- (6) Scan the subset and find out the “best” sample, which refers to the sample with the largest difference with other selected samples.
- (7) Take out the “best” sample from the subset and add it to the selection set.
- (8) Take out the samples that are not selected from the subset and put them into the recycling bin.
- (9) Determine whether the number of selected samples has reached M . If so, quit; Otherwise, go back to Step (2) and start a new subset.

The balance between representativeness and diversity of selected samples can be controlled by K value. Low K value produces more representative selection, while large K value can select samples with more diversity. If K is equal to the total number of samples in the dataset, that is, all objects as candidates in each step are considered, and the first two selected objects are not put back into the candidate pool, then OptiSim is a special case of the maximum-heterogeneity algorithm. If $K = 1$, it is a special case of the minimum-heterogeneity algorithm.

9.5 Other Methods

In order to solve the shortcomings of the K-S method, Liu et al. proposed the Rank-K-S method, which firstly sorted the samples according to the concentration value and divided the whole concentration interval into multiple intervals, and then used the K-S method to select the representative samples into calibration set for each

interval [15]. According to this idea, the SPXY method can also be improved to the Rank-SPXY method.

Selecting samples into calibration set can also be carried out by the way of eliminating samples. The basic principle is to calculate the Euclidean distance between each sample and the adjacent samples with the characteristics of the spectra (or the score matrix of PCA), and determine the threshold according to the density of sample distribution. As shown in Fig. 9.9, for each sample, all the samples whose distance is less than the threshold value are eliminated, so as to eliminate the redundant samples, and the remaining ones are used as calibration samples.

Another method is to condense the sample to obtain a representative sample. This kind of method is characterized by spectra (or the score matrix of PCA) for cluster analysis (such as Kohonen network method), and the cluster number is the number of samples to be selected into calibration set. One or several of each category are selected as calibration samples (Fig. 9.10) [16, 17]. Spectral and concentration data of all samples in each category can also be averaged and used as a calibration sample.

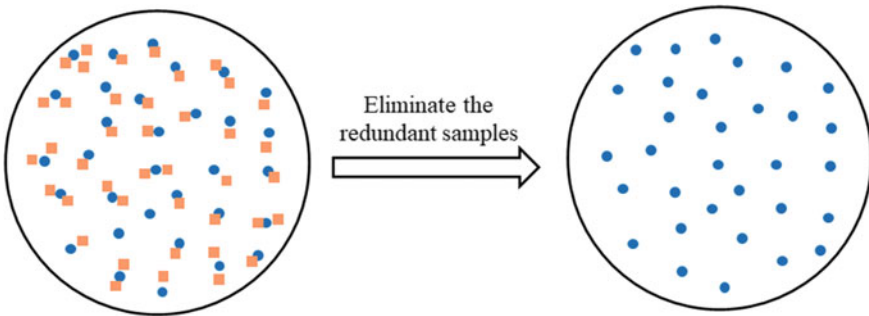


Fig. 9.9 Schematic diagram of eliminating the redundant samples. ○ represents the selected samples into calibration set. □ represents the redundant samples whose Euclidean distance between it and its adjacent calibration samples is less than the threshold value

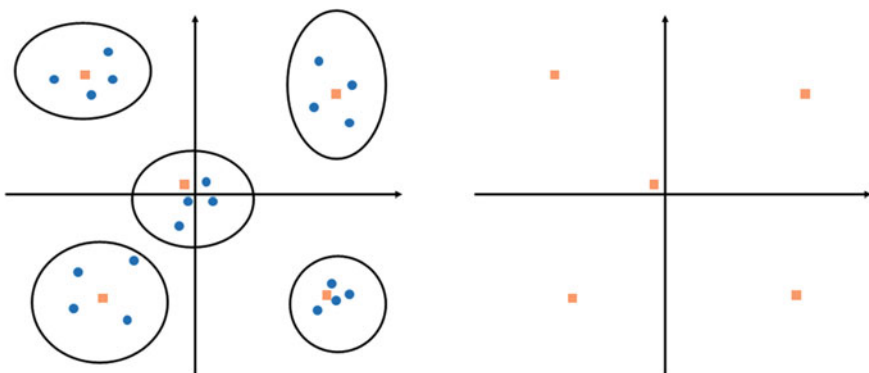


Fig. 9.10 Cluster analysis for selection of calibration samples

The advantage of this method is that the accuracy of the basic data can be improved to some extent through data averaging. Chen et al. applied isolation forest (IForest) algorithm to detect outliers and select representative subsets [18]. Isolation forest, an ensemble of isolation trees (ITree), can provide a ranking of samples which reflects the degree of outliers and representativeness. All samples are sorted according to their scores obtained by IForest. The outliers are ranked at the top of list and their scores are significantly larger than normal samples, and then excluded from the representative samples. Further, the samples with scores which are larger than 0.5 are selected as normal uncommon samples. Finally, the required number of samples from the remaining samples is uniformly selected as representative normal common samples as shown in Fig. 9.11.

Wavelength selection methods such as successive projection algorithm (SPA) can also be used for the selection of calibration samples. As shown in Fig. 9.12, after the transpose of spectral matrix \mathbf{X} , samples can be selected by conducting SPA [19].

Based on the different spatial distributions of samples in different spectra, Li et al. adopted the consensus strategy of combining different derivative spectral spaces to select representative calibration samples [20]. As shown in Fig. 9.13, the K-S method was firstly adopted to select the intersection samples from the zeroth derivative, the first derivative and the second derivative space as the basic calibration set, and then the extended calibration samples were selected from the samples with large prediction errors. In addition, Rowland-Jones et al. adopted the method of design of experiments (DoE) to select 20 representative samples in the design space from 957 samples in the historical sample base for calibration set [21]. Rius et al. used the Fedorov algorithm in D-optimal design method to select representative samples [22, 23].

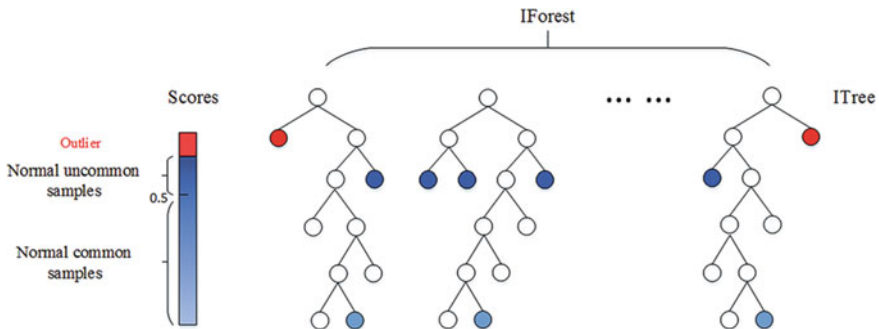


Fig. 9.11 Overview of isolation forest on detecting outliers and selecting representative subsets. Red circles in the ITree represent outliers in samples, dark blue circles represent normal uncommon samples, and light blue circles represent normal common samples [18]

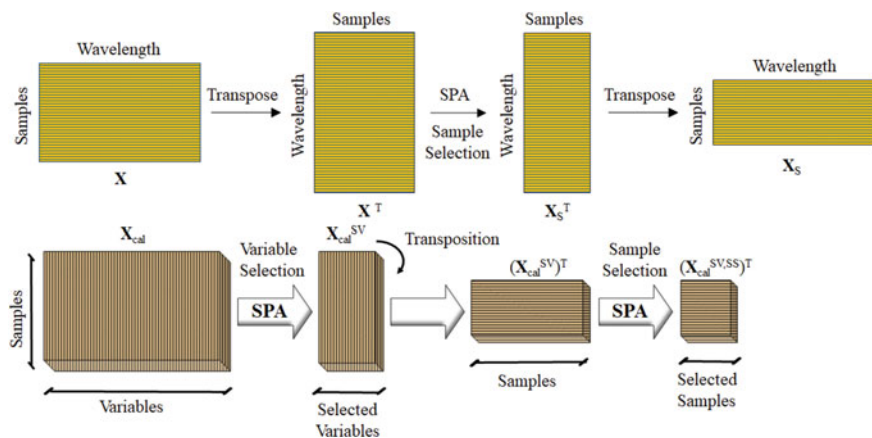


Fig. 9.12 Schematic diagram of the SPA method used in the process of selecting samples for calibration set [19]

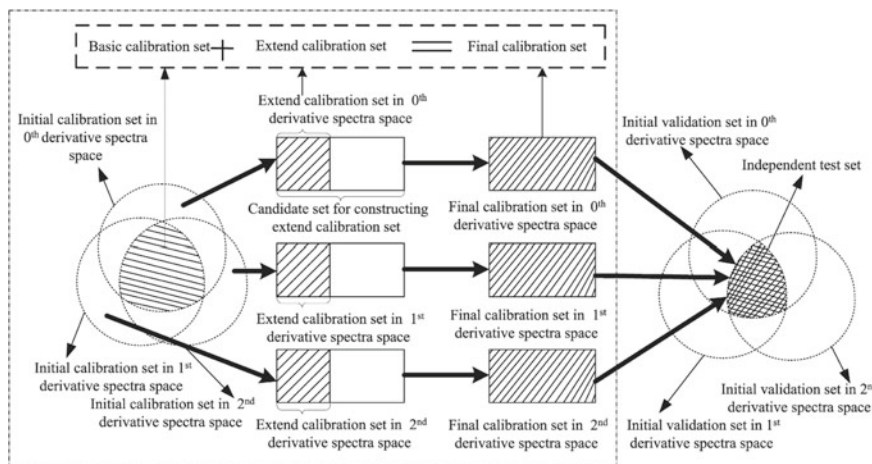


Fig. 9.13 Strategy diagram for selecting calibration samples based on spatial distribution of different spectra [20]

References

1. Honigs DE, Hieftje GM, Mark HL, et al. Unique-sample selection via near-infrared spectral subtraction. *Anal Chem.* 1985;57:2299–303.
2. Tominaga Y. Representative subset selection using genetic algorithms. *Chemom Intell Lab Syst.* 1998;43:157–63.
3. Williams PCNK. *Implementation of Near-Infrared technology.* 2nd Edition ed., Minnesota: American Association of Cereal Chemists;2001.
4. Kennard RW, Stone LA. Computer aided design of experiments. *Technometrics.* 1969;11:137–48.

5. Rajer-Kanduč K, Zupan J, Majcen N. Separation of data on the training and test set for modelling: a case study for modelling of five colour properties of a white pigment. *Chemom Intell Lab Syst.* 2003;65:221–9.
6. Wu W, Walczak B, Massart DL, et al. Artificial neural networks in classification of NIR spectral data: design of the training set. *Chemom Intell Lab Syst.* 1996;33:35–46.
7. de Groot PJ, Postma GJ, Melssen WJ, et al. Selecting a representative training set for the classification of demolition waste using remote NIR sensing. *Anal Chim Acta.* 1999;392:67–75.
8. Sneek RD. Validation of regression models: methods and examples. *Technometrics.* 1977;19:415–28.
9. Dotto AC, Dalmolin RSD, ten Caten A, et al. A systematic study on the application of scatter-corrective and spectral-derivative preprocessing for multivariate prediction of soil organic carbon by Vis-NIR spectra. *Geoderma.* 2018;314:262–74.
10. He Z, Li M, Ma Z. Design of a reference value-based sample-selection method and evaluation of its prediction capability. *Chemom Intell Lab Syst.* 2015;148:72–6.
11. Galvão RKH, Araújo MCU, José GE, et al. A method for calibration and validation subset partitioning. *Talanta.* 2005;67:736–40.
12. Tian H, Zhang L, Li M, et al. Weighted SPXY method for calibration set selection for composition analysis based on near-infrared spectroscopy. *Infrared Phys Technol.* 2018;95:88–92.
13. Clark RD. Optimisim: an extended dissimilarity selection method for finding diverse representative subsets. *J Chem Inf Comput Sci.* 1997;37:1181–8.
14. Siano GG, Goicoechea HC. Representative subset selection and standardization techniques. A comparative study using NIR and a simulated fermentative process UV data. *Chemom Intell Lab Syst.* 2007(88):204–12.
15. Liu W, Zhao Z, Yuan H-F, et al. An optimal selection method of samples of calibration set and validation set for spectral multivariate analysis. *Spectrosc Spectr Anal.* 2014;34:947–51.
16. Daszykowski M, Walczak B, Massart DL. Representative subset selection. *Anal Chim Acta.* 2002;468:91–103.
17. Tan C, Chen H, Wang C, et al. A multi-model fusion strategy for multivariate calibration using near and mid-infrared spectra of samples from brewing industry. *Spectrochim Acta Part A Mol Biomol Spectrosc.* 2013;105:1–7.
18. Chen W-R, Yun Y-H, Wen M, et al. Representative subset selection and outlier detection via isolation forest. *Anal Methods.* 2016;8:7225–31.
19. Filho HAD, Galvão RKH, Araújo MCU, et al. A strategy for selecting calibration samples for multivariate modelling. *Chemom Intell Lab Syst.* 2004;72:83–91.
20. Li Z, Liu J, Shan P, et al. Strategy for constructing calibration sets based on a derivative spectra information space consensus. *Chemom Intell Lab Syst.* 2016;156:7–13.
21. Rowland-Jones RC, van den Berg F, Racher AJ, et al. Comparison of spectroscopy technologies for improved monitoring of cell culture processes in miniature bioreactors. *Biotechnol Prog.* 2017;33:337–46.
22. Rius A, Callao MP, Ferré J, et al. Assessing the validity of principal component regression models in different analytical conditions. *Anal Chim Acta.* 1997;337:287–96.
23. Ferré J, Rius FX. Selection of the best calibration sample subset for multivariate regression. *Anal Chem.* 1996;68:1565–71.

Chapter 10

Detection Methods for Outlier Samples



The identification of outliers in spectral analysis is mainly used in two aspects: one is the identification of outliers in the process of model building; the other is the determination of whether the samples to be tested are the outliers of the model in predictive analysis. The second aspect is also regarded as the domain of applicability which is the third principle of five OECD principles as “a defined domain of applicability” [1, 2].

10.1 Detection of Outlier Samples During Calibration Process

Two types of outlier samples may appear in the calibration process. The first type is the sample with extreme composition, often called the high leverage point sample, which has a strong influence on the regression results. Such outlier samples are usually detected by the combination of principal component analysis (PCA) and Mahalanobis distance (MD) (PCA-MD) method [3]. The calibration samples with greater MD than $3f/n$ are excluded, where f is the number of principal components used in PCA and n is the number of samples in the calibration set.

Here the MD is defined as follows:

$$MD_i = [(t_i - \bar{t}) \cdot (\mathbf{T}_{cen}^T \mathbf{T}_{cen})^{-1} \cdot (t_i - \bar{t})^T] \quad (10.1)$$

where t_i is the score of the spectrum of the i th sample in the calibration set, \mathbf{T} is the score matrix of all the samples in the calibration set, \bar{t} is the average score vector of \mathbf{T} , \mathbf{T}_{cen} is the mean-centered matrix of \mathbf{T} , that is $\mathbf{T}_{cen} = \mathbf{T} - \bar{t}$, and MD_i is the Mahalanobis distance of the i th sample in the calibration set. The partial least squares (PLS) score can also be used to calculate MD.

The second type of outlier samples refers to samples in calibration set with statistical difference between the reference value and the predicted value. The existence of such outlier samples indicates that the reference data may be subject to large errors. The outlier samples can be eliminated by considering the reproducibility requirements specified by the corresponding reference method, that is, the calibration samples whose deviation between the predicted value of cross validation and the measured value of the reference method is greater than the reproducibility specified by the corresponding reference method can be eliminated. If the reference method does not provide reproducibility, the following formula can be used to eliminate calibration samples:

$$(y_i - \hat{y}_i) > 2 \times \text{SECV} \times \sqrt{\frac{n - f - 1}{n}} \quad (10.2)$$

where f is the number of the optimal principal component selected by PLS or PCR, and n is the number of samples in calibration set. SECV is the standard error of cross validation.

It can also be identified by t -test, which is defined as

$$t_i = \frac{e_i}{\text{SEC} \sqrt{1 - \text{MD}_i}} \quad (10.3)$$

where t_i is the t -test value of the i th sample of the calibration set, and e_i is the difference between the predicted value of the i th sample in the calibration set and the reference data. MD_i is the Mahalanobis distance of i th sample, while SEC is the standard error of calibration. The t value is compared with the critical value of the t distribution with $n-f-1$ degree of freedom. The samples whose t value is larger than the critical value are excluded. Sometimes, for simplicity, outlier samples of which with deviations are greater than 2.5–3 times SECV can be eliminated.

10.2 Detection of Outlier Samples During the Prediction Process

The identification of outlier samples in the prediction process is mainly used to check whether the samples to be tested are within the coverage of the established calibration model, so as to ensure the accuracy of the prediction results. According to ASTM E 1655–05 [4], outlier samples of model include three categories: (1) the outlier samples based on concentration, that is, the use of MD to detect whether the concentration of the unknown sample exceeds the concentration range of the calibration samples; (2) the outlier samples based on spectral residual, namely, using the root mean square of spectral residual (RMSSR) to detect whether the unknown sample contains the component which does not exist in calibration set; (3) the outlier

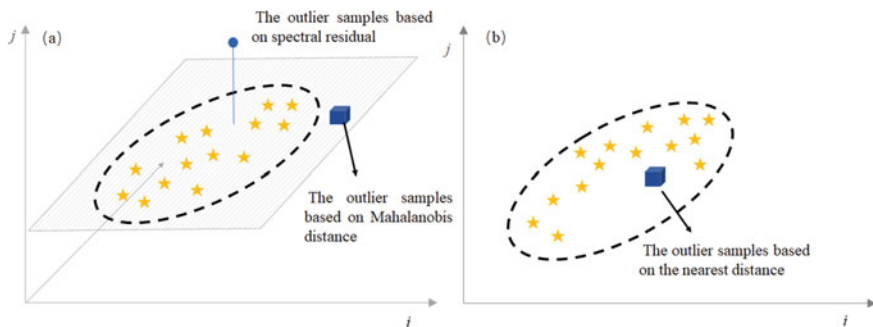


Fig. 10.1 Schematic diagram of three types of outlier samples

samples based on the nearest distance, that is, the nearest distance is used to detect whether the unknown sample is located in the area with sparse distribution of samples in calibration set [5–7]. When any of the spectral residual, MD and nearest neighbor distance of the unknown sample exceeds the corresponding threshold (Fig. 10.1), it indicates that the sample is an outlier sample of the model, and the accuracy of its prediction results will be greatly questioned.

(1) Identification of outlier samples based on concentration

The PCA-MD method, which combines PCA and MD, is usually used to identify outlier samples based on concentration. For the spectrum of unknown sample, the spectral loading matrix and score matrix obtained by the calibration samples is calculated, and then the MD is calculated. The outlier samples were determined according to the MD threshold defined during the prediction process. For example, for a system consisting of a mixture of three pure substances, A, B, and C, the concentration ranges of the three components in the calibration set are 0–10% of component A, 5–25% of component B, and 50–75% of component C. If a sample to be tested is composed of 5% of component A, 40% of component B and 55% of component C, the sample is identified as an outlier sample because the concentration of component B is beyond the concentration range of the calibration samples.

(2) Identification of outlier samples based on spectral residual

Spectral residuals can be used for detection when the unknown sample contains a component which does not exist in the calibration set. The selected principal component f is used to reconstruct the spectral matrix \mathbf{X} of the calibration set to obtain the reconstructed spectral matrix $\hat{\mathbf{X}}$ then the spectral residual matrix of the calibration set can be obtained as follows:

$$\mathbf{R} = \mathbf{X} - \hat{\mathbf{X}} \quad (10.4)$$

The RMSSR of the spectral residual of each sample of the calibration set can be calculated by the following formula:

$$\text{RMSSR}_i = \sqrt{\frac{\mathbf{r}_i \mathbf{r}_i^T}{f}} \quad (10.5)$$

where \mathbf{r}_i is the spectral residual of the i th sample in the spectral residual matrix \mathbf{R} of the calibration set. RMSSR_i is the root mean square of the spectral residual of the i th sample of the calibration set, and f is the optimal number of principal components selected in the calibration process of PLS. The threshold of spectral residual root mean square can be determined by the spectral repeatability.

For the unknown sample spectra \mathbf{x} , the spectral score is first calculated by the spectral loading matrix of the model with PLS, and the spectral residual matrix can be obtained by reconstructing $\hat{\mathbf{x}}$ as follows, and then its RMSSR was calculated:

$$\mathbf{r} = \mathbf{x} - \hat{\mathbf{x}} \quad (10.6)$$

If RMSSR value is greater than the defined threshold, it indicates that the sample is an outlier sample of spectral residual, that is, the sample may contain components that do not exist in the calibration set.

For example, for a system consisting of a mixture of three pure substances, A, B, and C, the concentration ranges of the three components in the calibration set are 0–10% of component A, 5–25% of component B, and 50–75% of component C. If a sample to be tested consists of 9% of component A, 10% of component B, 61% of component C and 61% of component D, then the sample is an outlier sample of spectral residual, because the sample contains component D which does not exist in the sample of calibration set.

(3) Identification of outlier samples based on the nearest distance

If the calibration samples are unevenly distributed in the variable space, an unknown sample to be tested may fall into a calibration space with less relative sample aggregation regardless of its MD and RMSSR values which are less than the setting threshold. In this case, it is necessary to use the nearest distance to detect whether the unknown sample falls into the blank area of the calibration space. PCA-MD method is usually used to calculate the nearest distance. The specific steps are as follows: Calculate the MD between all samples in the calibration set through the principal component score \mathbf{t} , and get the maximum NND_{\max} value, which represents the maximum distance between samples in the calibration set.

For unknown sample spectra, obtain the spectral loading matrix by the samples in the calibration set, then calculate its score and the MD between each sample, and calculate the minimum value. If the minimum value is greater than the NND_{\max} , indicating that the sample falls into the space with less distribution of calibration samples, and this kind of sample is called the outlier samples based on the nearest distance.

10.3 Other Detection Methods

In fact, there are many forms of the outlier sample of the model. As shown in Fig. 10.2, the outlier samples of the model can be divided into score distance outliers (samples 1 and 4), spectral residual distance outliers (samples 5), and bad leverage points that have a large spectral residual distance and a large score distance (samples 2 and 3) [8, 9]. Sometimes, the classical method cannot detect these outlier samples at the same time, and they will affect the accuracy and robustness of the model to a certain extent.

At this time, Monte Carlo cross-validation (MCCV) method can be used to diagnose the outlier samples. First, a certain proportion of samples (such as 80%) are selected from the calibration set through MC sampling as the training samples, and the remaining samples (20%) are taken as the samples of the independent test set. The process is repeated N times, and N training subsets and corresponding N test subsets can be obtained. The model is established with each training subset and the corresponding test subset samples are predicted. The outlier samples are diagnosed according to the statistical distribution characteristics of the prediction errors of each sample (such as the mean value and standard deviation of the error distribution) [9–11]. This kind of method is based on the MPA framework as we introduced it in Chap. 5 [12, 13].

Figure 10.3 shows the results obtained by MCCV [14]. Figure 10.3a shows the mean (X-axis) and standard deviation (Y-axis) of the error distribution, and Fig. 10.3b shows the distribution of the sample prediction error. Among this, three types of samples can be determined, the normal sample (sample A), the outlier sample in the X-direction (sample B), and the outlier sample in the Y-direction (samples C, D, and E). As can be seen from Fig. 10.3, for normal samples, the mean error is

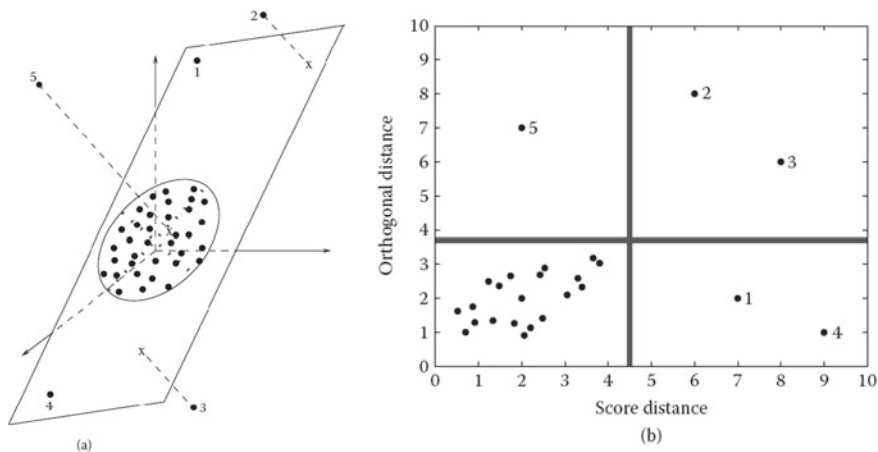


Fig. 10.2 **a** Different types of outliers when a three-dimensional dataset is projected on a robust two-dimensional PCA-subspace. **b** The corresponding outlier map [8]

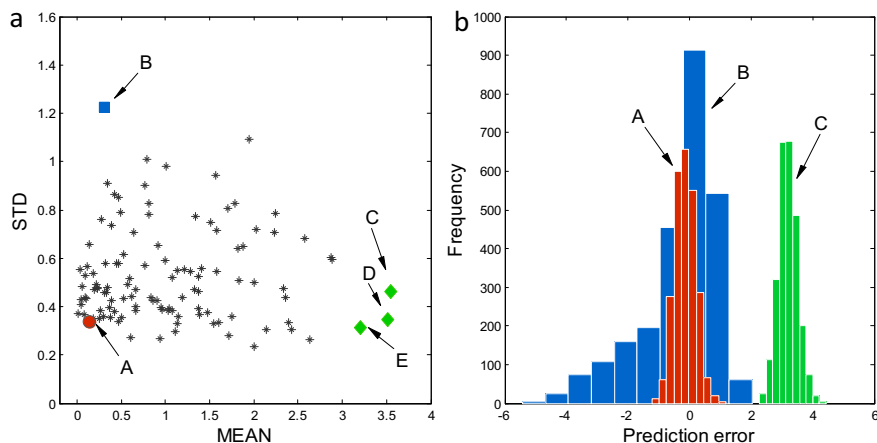


Fig. 10.3 **a** diagnosis plot of outlier samples based on the mean and standard deviation of prediction errors of Monte Carlo cross validation. Three kinds of samples that are most representative of (A) a normal sample, (B) an X-outlier, and (C, D, and E) Y-outliers are selected. **b** the distributions of prediction errors of A, B, and C outlier samples [14]

around 0, and the standard deviation of the error distribution is very small. For the outlier sample in the X-direction (sample B), the mean error is close to zero, but the standard deviation of the error distribution is large. For the outlier samples in the Y-direction (samples C, D, and E), not only the mean error deviates from zero, but also the standard deviation of the error distribution is large.

To further improve the identification efficiency of outlier samples, Zhang et al. proposed an enhanced MC outlier sample identification method based on the interactive prediction model established by normal samples and independent validation of suspected outlier samples [15, 16].

Based on cluster analysis algorithm of the model, Chen et al. proposed sampling error profile analysis (SEPA) algorithm [17], using a variety of statistical indicators for a comprehensive analysis of random sampling error, such as the median and standard deviation of the error distribution, distribution skewness, and distribution kurtosis. The indicator can be used not only for the screening of outlier samples, but also for the evaluation of spectral preprocessing methods and wavelength selection methods.

References

1. Sahigara F, Ballabio D, Todeschini R, et al. Defining a novel k-nearest neighbours approach to assess the applicability domain of a QSAR model for reliable predictions. *J Cheminform.* 2013;5:27.
2. Yun Y-H, Wu D-M, Li G-Y, et al. A strategy on the definition of applicability domain of model based on population analysis. *Chemom Intell Lab Syst.* 2017;170:77–83.

3. Mark H. Use of Mahalanobis distances to evaluate sample preparation methods for near-infrared reflectance analysis. *Anal Chem.* 1987;59:790–5.
4. Silva MAM, Ferreira MH, Braga JWB, et al. Development and analytical validation of a multivariate calibration method for determination of amoxicillin in suspension formulations by near infrared spectroscopy. *Talanta.* 2012;89:342–51.
5. Jouan-Rimbaud D, Bouveresse E, Massart DL, et al. Detection of prediction outliers and inliers in multivariate calibration. *Anal Chim Acta.* 1999;388:283–301.
6. Fernández Pierna JA, Wahl F, de Noord OE, et al. Methods for outlier detection in prediction. *Chemom Intell Lab Syst.* 2002;63:27–39.
7. Walczak B. Outlier detection in multivariate calibration. *Chemom Intell Lab Syst.* 1995;28:259–72.
8. Hubert M, Engelen S. Robust PCA and classification in biosciences. *Bioinformatics.* 2004;20:1728–36.
9. Gemperline P. *Practical Guide To Chemometrics* CRC Press, 2006.
10. Cao D-S, Liang Y-Z, Xu Q-S, et al. A new strategy of outlier detection for QSAR/QSPR. *J Comput Chem.* 2010;31:592–602.
11. Bian X, Cai W, Shao X, et al. Detecting influential observations by cluster analysis and Monte Carlo cross-validation. *Analyst.* 2010;135:2841–7.
12. Li H-D, Liang Y-Z, Cao D-S, et al. Model-population analysis and its applications in chemical and biological modeling. *TrAC, Trends Anal Chem.* 2012;38:154–62.
13. Li H-D, Liang Y-Z, Xu Q-S, et al. Model population analysis for variable selection. *J Chemom.* 2010;24:418–23.
14. Deng B-C, Yun Y-H, Liang Y-Z. Model population analysis in chemometrics. *Chemom Intell Lab Syst.* 2015;149:166–76.
15. Zhang L, Li P, Mao J, et al. An enhanced Monte Carlo outlier detection method. *J Comput Chem.* 2015;36:1902–6.
16. Zhang L, Wang D, Gao R, et al. Improvement on enhanced Monte-Carlo outlier detection method. *Chemom Intell Lab Syst.* 2016;151:89–94.
17. Chen W, Du Y, Zhang F, et al. Sampling error profile analysis (SEPA) for model optimization and model evaluation in multivariate calibration. *J Chemom.* 2018(32):e2933.

Chapter 11

Maintenance and Update of Calibration Model



11.1 Necessity

The maintenance and update of the calibration model is one of the main tasks of spectroscopy combined with chemometrics analysis. No matter how advanced the instruments or how large the model library is, the established calibration model is not permanent. In practice, there are usually samples that models cannot cover (Fig. 11.1). Therefore, model updates are necessary, even in many cases, this work has become a key factor affecting the successful application of this technology.

Wise et al. provided a routine for model maintenance (Fig. 11.2) [1], which involves the maintenance and transfer of models between instruments (refer to Chap. 16).

When there are samples outside the model boundary, we should figure out the reason why they fall outside. ① Chemical composition of the tested sample has changed. As shown in Fig. 11.3, there are two types of outlier samples with chemical composition, one is that the chemical composition has changed (spectral residual outlier samples); the other one is that the chemical composition has not changed, but the concentration range of one or more components has changed significantly (principal component scores are out of bounds) [2]. ② Samples without change of chemical composition, but with changes in the spectrometer caused by the environment, abnormal operation of the light source, significant changes in the temperature or particle size of the sample, etc. If in case ①, it is necessary to add these samples to the sample collection in time, update the calibration model, and expand the coverage of the model. If in case ②, it needs to eliminate hardware failures to ensure the consistency of analysis conditions.

Common problems of spectrometer hardware include the aging of light sources, lasers, electronic components, contamination of reference materials, and changes in wavelength accuracy and S/N ratio caused by other factors.

Changes in test samples include changes in natural products (such as grains, tobacco, and forages) due to climate, species evolution, genetic changes, etc., industrial products change due to changes in raw materials, production formulas,

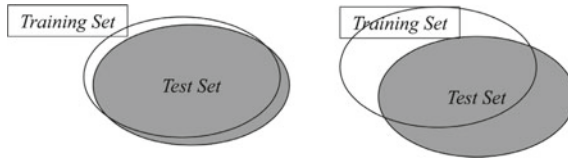


Fig. 11.1 Scheme of training and test set coverage. A: Basic coverage, no need to maintain the model; B: problem of noncoverage and the model needs to be maintained

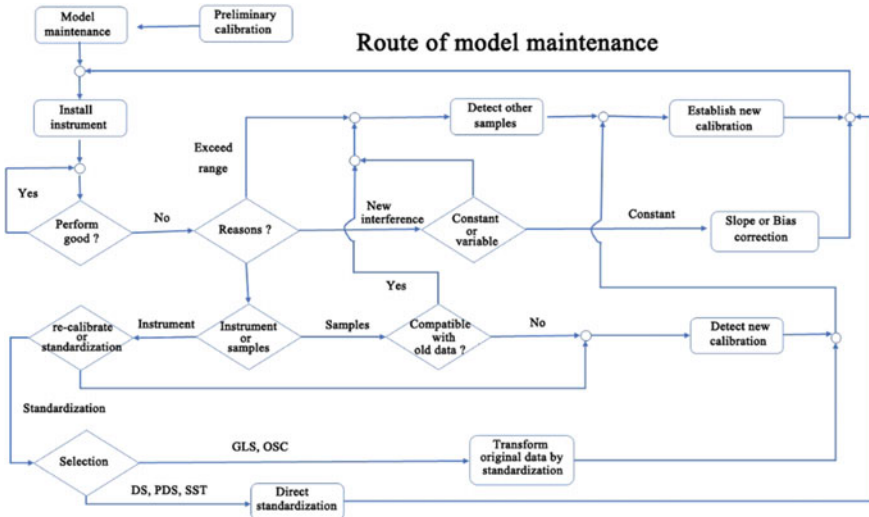


Fig. 11.2 Rout of model maintenance

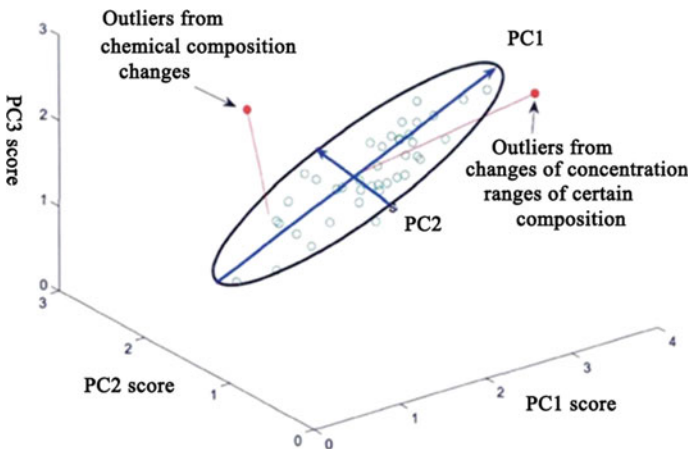


Fig. 11.3 Scheme of the spatial distribution of PCA scores of two types of outliers due to chemical composition variation

processing techniques, processing parameters, etc., as well as changes in sample preparation due to grating, mixing (homogenization), sieving, drying, pressure, density, and thickness.

For outlier samples caused by temperature, moisture, or particle size, these variable factors can also be introduced into the model, but it will reduce the accuracy of model by a certain extent [3, 4]. Therefore, it is very necessary to use the comparative data of regular validation samples to update the model in a fixed period (as 2 months) to improve the robustness of the model [5, 6].

We can use quality control samples (or actual test samples) to monitor the model or instrument status and other factors that affect the accuracy of prediction through the quality control chart. The monitoring frequency of quality control samples needs to be determined according to the actual situation. It can be once a day, or the quality control samples can be measured before each routine analysis. In the quality control chart, when a sample exceeds the action limit or is outside the alarm limit for 3 times in a row, 2 times in a row, or on the the same side of the zero line for 9 times in a row, the model update procedure should be started.

Figure 11.4 is a quality control chart for analyzing fat content in grain feed by NIRS. As shown, there are no monitoring points exceeding the action limit, however, from the 14th to the 22nd points, there have been 9 consecutive times on the same side of the zero line. Plus, from the 26th to 28th points, there are 2 out of 3 consecutive times outside the alarm limit. It is indicated that there are systematic errors in the prediction results, and it is necessary to update the model or check whether the test environment or instrument status is in a normal state [7].

As shown in Fig. 11.5, the quality control diagram of another model for the

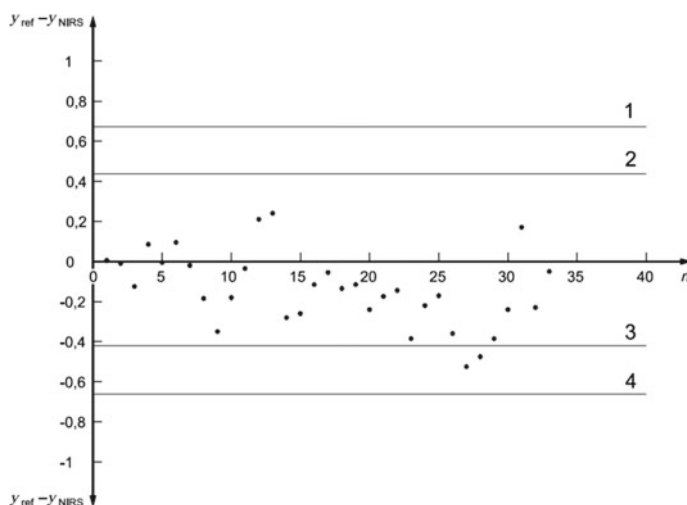


Fig. 11.4 Quality control chart for determination of crude fat content in grain feed ingredients. 1 Upper action limit, UAL, + 3SEP, 2 Upper warning limit, UWL, + 2SEP, 3 Lower warning limit, LWL, -2SEP, 4 Lower action limit, UAL, + 3SEP

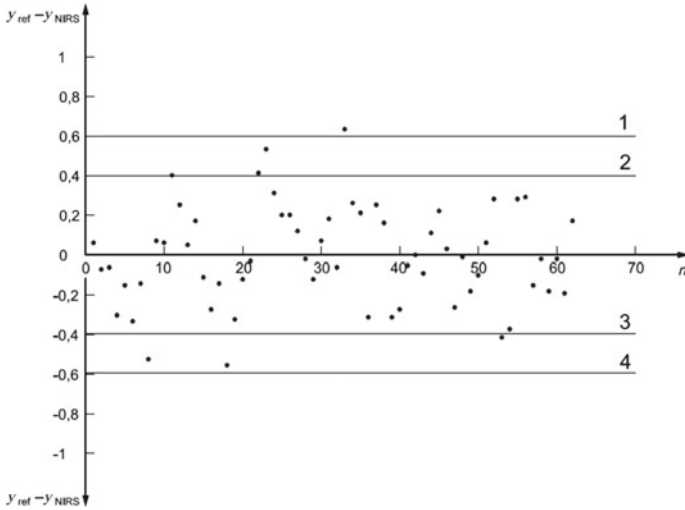
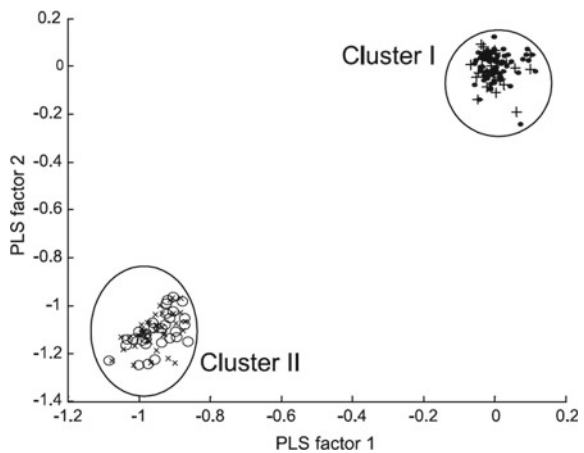


Fig. 11.5 Quality control chart for the determination of a parameter in a model. 1 Upper action limit, UAL, + 3SEP, 2 Upper warning limit, UWL, + 2SEP, 3 Lower warning limit, LWL, -2SEP, 4 Lower action limit, UAL, + 3SEP

determination of a certain index, before the first 35 monitoring points, there is a situation that two of the three consecutive detections are outside the alarm limit, and at the same time, one monitoring point exceeds the upper limit of the action limit, which indicates the model is not in an ideal status. After adding new samples to re-update the model (after 35 monitoring points), all quality control points are within the controllable range, indicating that performance of the updated model has been significantly improved.

Fig. 11.6 Two clusters of samples due to changes in the production process



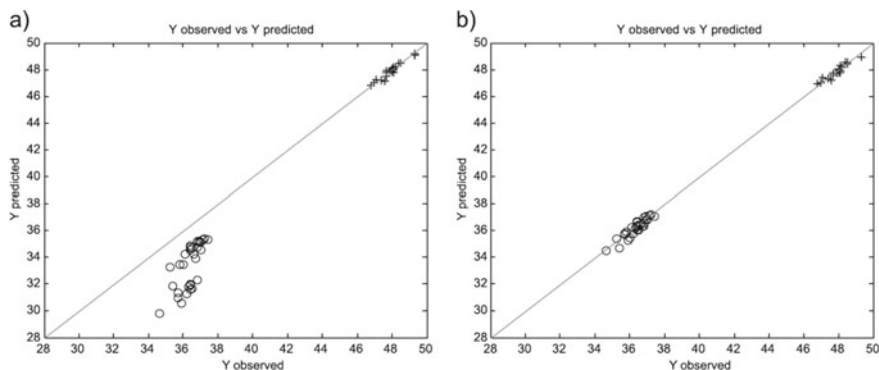


Fig. 11.7 **a** The model built by class I samples cannot correctly predict class II samples; **b** The updated model by adding class II samples predicts class II samples well

When updating the model, it is necessary to re-check the outliers of calibration process, because if only one sample representing a new range or new type was added, then the new sample may be kicked as an outlier. So it is required to add multiple new samples of each type to calibration model. After updated, model needs to be re-validated, and can be validated with the initial validation set samples, but the proportion of samples representing the new range or new type should not be less than the proportion of new samples in the calibration set.

The most direct way to update the model is to add new samples (spiked samples) to the old calibration set to form a new one, and use PLS and other calibration methods to recalculate the model [8, 9]. The number and representativeness of additional samples are related to the multivariate calibration method used [10, 11]. For example, when the old calibration set samples are relatively little, adding a few samples will have a great impact on the model. As shown in Fig. 11.6, due to changes in the production process, class I samples and class II samples have been produced. In Fig. 11.7, the model established by the class I cannot accurately predict the class II sample. After adding two representative samples, the new model can be well adapted to the class II sample [12].

Along with samples in the calibration set increases, the burden of calculations (cross validation tests, etc.) caused by the large scale of data will become more and more prominent. Accordingly, the recursive partial least square (RPLS) method can be used to update the model adaptively [13, 14].

11.2 Recursive Exponentially Weighted PLS

Recursive exponentially weighted PLS (REWPLS) was proposed by Dayal et al. in 1997. Given a new calibration sample, the regression coefficients of the model will be recursively updated [15, 16]. For the existing calibration set spectra matrix $X(n)$

$\times m$) and concentration matrix Y ($n \times p$), the spectrum of a newly added calibration sample is x , and the corresponding concentration is y . The steps of the recursive exponentially weighted PLS method are as follows.

- (1) Calculate the covariance matrix of the existing spectra and concentration matrix.

$$\mathbf{R}_{xx}^{\text{old}} = \mathbf{X}^t \mathbf{X}, \mathbf{R}_{xy}^{\text{old}} = \mathbf{X}^t \mathbf{Y} \quad (11.1)$$

- (2) Calculate the covariance matrix after adding a new sample.

$$\mathbf{R}_{xx} = \lambda \mathbf{R}_{xx}^{\text{old}} + x^t x \mathbf{R}_{xy} = \lambda \mathbf{R}_{xy}^{\text{old}} + x^t y \quad (11.2)$$

where X , Y , x , and y are all variables after the mean centralization or standardization process, and λ is the forgetting factor, usually $0 < \lambda < 1(0.95)$.

- (3) Set $k = 1$, the largest number of PCs is A .
 (4) Calculate the weight vector w_k .

$$w_k = \mathbf{R}_{xy} w_k = w_k / \|w_k\| \quad (11.3)$$

- (5) Calculate r_k .

$$r_k = w_k (k = 1), r_k = w_k - p_1^t w_k r_1 - p_2^t w_k r_2 \dots - p_{k-1}^t w_k r_{k-1} (k > 1) \quad (11.4)$$

- (6) Calculate the score and loading vector.

$$\begin{aligned} t_k^t t_k &= r_k^t \mathbf{R}_{xx} r_k \\ p_k^t &= r_k^t \mathbf{R}_{xx} / r_k^t r_k \\ q_k^t &= r_k^t \mathbf{R}_{xy} / t_k^t t_k \end{aligned} \quad (11.5)$$

- (7) Update the covariance matrix \mathbf{R}_{xy} .

$$\mathbf{R}_{xy} = \mathbf{R}_{xy} - p_k q_k^t (t_k^t t_k) \quad (11.6)$$

- (8) If $k = A$, go to the next step; otherwise, $k = k + 1$, go back to step (4).
 (9) Calculate regression coefficient b .

$$b = [r_1 \cdot r_2 \dots r_A] [q_1 \cdot q_2 \dots q_A]^t \quad (11.7)$$

For the new calibration sample x and y , the mean value is centered by the following equation.

Set \bar{x}^{old} and \bar{y}^{old} are mean vectors of X and Y , respectively.

Then,

$$\bar{\mathbf{x}} = (n - 1)/n \times \bar{\mathbf{x}}^{\text{old}} + 1/n \times \mathbf{x} \quad (11.8)$$

$$\bar{\mathbf{y}} = (n - 1)/n \times \bar{\mathbf{y}}^{\text{old}} + 1/n \times \mathbf{y} \quad (11.9)$$

The mean centered vectors $\tilde{\mathbf{x}}$ and $\tilde{\mathbf{y}}$ of the new calibration samples \mathbf{x} and \mathbf{y} are.

$$\tilde{\mathbf{x}} = \mathbf{x} - \bar{\mathbf{x}} \quad (11.10)$$

$$\tilde{\mathbf{y}} = \mathbf{y} - \bar{\mathbf{y}} \quad (11.11)$$

Since the dimension of covariance matrix \mathbf{R}_{xx} is $m \times m$, the dimension of \mathbf{R}_{xy} is $m \times p$, so the calibration process has nothing to do with the number of samples, only the number of wavelength variables. The forgetting factor λ reduces the influence of the historical calibration set, so it can better solve the problem of little influence on the model given the few new samples [17, 18].

11.3 Block-Wise Recursive PLS

In the above REWPLS, recursive calculation is performed when a new calibration sample is involved to obtain a new PLS model. However, in practice, there are usually a certain number of accumulated calibration samples that are expected to be combined with the original model to form a new PLS model. Therefore, a block-wise recursive PLS was proposed by Qin et al. in 1998 [19].

In order to be suitable for the recursive algorithm, Helland et al. modified the classic nonlinear iterative partial least squares algorithm (NIPALS) proposed by Wold [20]. It normalizes the score matrix \mathbf{T} of \mathbf{X} , instead of normalizing the weight \mathbf{W} and loading matrix \mathbf{P} , which can obtain a feature that is actually very important for various recursive PLS algorithms, i.e., $\mathbf{T}^t \mathbf{T} = \mathbf{I}$, where \mathbf{I} is the identity matrix. Specific algorithm is as follows.

- (1) Take a column \mathbf{y}_i of the concentration matrix \mathbf{Y} as the initial iteration value of \mathbf{u} , and generally take the column with the largest variance.
- (2) Calculate the weight vector of \mathbf{X} , $\mathbf{w} = \mathbf{X}^t \mathbf{u} / (\mathbf{u}^t \mathbf{u})$.
- (3) Calculate the score vector of \mathbf{X} , $\mathbf{t} = \mathbf{X} \mathbf{w} / \|\mathbf{X} \mathbf{w}\|$.
- (4) Calculate the weight vector of \mathbf{Y} , $\mathbf{c} = \mathbf{Y}^t \mathbf{t} / \|\mathbf{t}^t \mathbf{t}\|$, and score vector $\mathbf{u} = \mathbf{Y} \mathbf{c}$.
- (5) If the difference between the obtained \mathbf{t} and the last iteration result meets the set allowable error, proceed to the next step, otherwise return to Step (2).
- (6) Calculate the loading vector of \mathbf{X} , $\mathbf{p} = \mathbf{X}^t \mathbf{t}$, and the loading vector of \mathbf{Y} , $\mathbf{q} = \mathbf{Y}^t \mathbf{u} / \|\mathbf{u}^t \mathbf{u}\|$.
- (7) Calculate the regression coefficients of the internal model, $\mathbf{b} = \mathbf{u}^t \mathbf{t} / \|\mathbf{t}^t \mathbf{t}\|$.
- (8) Calculate the residual matrix, $\mathbf{E}_X = \mathbf{X} - \mathbf{t} \mathbf{p}^t$, $\mathbf{E}_Y = \mathbf{Y} - \mathbf{t} \mathbf{q}^t$.

- (9) Replace X with E_X , replace Y with E_Y , go back to step (1), by analogy, find the w, t, p, u, q , and b for PCs of X and Y .

Based on $T^t T = I$, the following features can be introduced.

$$X^t X = P T^t T P^t = P P^t \quad (11.12)$$

$$X^t Y = P T^t T B Q^t = P B Q^t \quad (11.13)$$

If the former calibration set matrix is X, Y , and X_1, Y_2 are the new sample calibration matrix, the new calibration set can be expressed as

$$X_{\text{new}} = \begin{bmatrix} X \\ X_1 \end{bmatrix}, Y_{\text{new}} = \begin{bmatrix} Y \\ Y_1 \end{bmatrix} \quad (11.14)$$

It can be concluded that

$$X_{\text{new}}^t Y_{\text{new}} = \begin{bmatrix} X \\ X_1 \end{bmatrix}^t \begin{bmatrix} Y \\ Y_1 \end{bmatrix} = \begin{bmatrix} P^t \\ X_1 \end{bmatrix}^t \begin{bmatrix} B Q^t \\ Y_1 \end{bmatrix} \quad (11.15)$$

As seen, performing PLS regression on $\begin{bmatrix} X \\ X_1 \end{bmatrix}$ and $\begin{bmatrix} Y \\ Y_1 \end{bmatrix}$ is the same as the model parameters obtained by performing regression on $\begin{bmatrix} P^t \\ X_1 \end{bmatrix}$ and $\begin{bmatrix} B Q^t \\ Y_1 \end{bmatrix}$.

From the above, the algorithm of the block-wise recursive PLS is as follows.

- (1) Perform mean centralization and standardization of former calibration data matrix X and Y .
- (2) Use the improved NIPALS to calculate the PLS model (take k PCs).

$$\{X, Y\} \xrightarrow{\text{PLS}} P, T, B, Q \quad (11.16)$$

- (3) After preprocessing, a new batch of calibration data matrix X_1 and Y_2 is acquired to form a new full calibration data matrix:

$$X_{\text{new}} = \begin{bmatrix} \lambda P^t \\ X_1 \end{bmatrix} Y_{\text{new}} = \begin{bmatrix} \lambda B Q^t \\ Y_1 \end{bmatrix} \quad (11.17)$$

where λ is the forgetting factor, usually $0 < \lambda < 1(0.95)$.

- (4) Perform PLS regression on X_{new} and Y_{new} and new model is generated.

$$X_{\text{new}}, Y_{\text{new}} \xrightarrow{\text{PLS}} P, T, B, Q \quad (11.18)$$

It can be seen that the block-wise recursive PLS only needs to retain the parameters of the former PLS model, other than the former calibration set. If n_1 new calibration samples are added and the model is updated with block-wise recursive PLS, the calculation amount is $k + n_1$, while the conventional PLS is $n + n_1$ when updated. In practice, sample n is usually far more than PCs k . This change of block-wise recursive PLS significantly reduces the computation burden and storage space, and with the more samples, the advantages of block-wise recursive PLS will be more prominent [21].

To weaken the influence of original data on the new model, the former model matrix can be weighted by the forgetting factor, and then combined with the new data to form the input and output data matrix of PLS regression. The best PCs of block-wise recursive PLS can be determined by cross validation.

11.4 Just-In-Time Learning and Active Learning

For online spectral analysis, people are increasingly using a combination of Just-in-time learning (JITL), moving windows, and recursive methods to update models. JITL is an online update method of local model based on database. Its basic idea is similar to the local weight regression strategy, which builds real-time model on the new samples to adapt to the latest process situation so as to improve the predictivity of modeling [22–24]. As reported, with the help of JITL modeling and Gaussian process regression methods, [31] proposed an automatic real-time model calibration strategy and realized the “intelligence” of model maintenance [25].

In the recent years, the idea of active learning (AL) in machine learning has been used to maintain the spectral calibration model [26–30]. AL uses a certain algorithm to locate the most useful unlabeled samples and hand them over to experts for labeling, and then use the located samples to train the classification model to improve the model accuracy. At present, the most popular application in spectral analysis is model update based on AL and SVM classification, which is called incremental support vector data description (ISVDD). It implements the uncertain sampling strategy of AL algorithm to select certain new samples closest to the optimal hyperplane for classification and add them to the old calibration set. It tries to make the old calibration model has all the information of the new tested samples, so as to update the model and improve the predictability of calibration model.

References

1. Wise BM, Roginski RT. A calibration model maintenance roadmap IFAC-PapersOnLine. 2015;48–8:260–5.
2. Setarehdan SK, Soraghan JJ, Littlejohn D, et al. Maintenance of a calibration model for near infrared spectrometry by a combined principal component analysis-partial least squares approach. *Anal Chim Acta*. 2002;452:35–45.

3. Guthrie JA, Reid DJ, Walsh KB. Assessment of internal quality attributes of mandarin fruit. 2. NIR Calibration Model Robustness. *Aust J Agric Res.* 2005; 56(4):417–426.
4. Wortel VAL, Hansen WG, Wiedemann SCC. Optimising multivariate calibration by robustness criteria. *J Near Infrared Spectrosc.* 2001;9(1):141–51.
5. Garcia-Mencia MV, Andrade JM, Lopez-Mahia P, et al. An empirical approach to update multivariate regression models intended for routine industrial use. *Fuel.* 2000;79(14):1823–32.
6. Dyrby M, Engelsen SB, Nørgaard L, et al. Chemometric quantitation of the active substance (Containing C≡N) in a pharmaceutical tablet using near-infrared (NIR) transmittance and NIR FT-IR Raman spectra. *Appl Spectrosc.* 2002;56:579–85.
7. ISO 12099 Animal feeding stuffs, cereals and milled cereal products. Guidelines for the application of near infrared spectrometry. ISO International Standard, 2010.
8. Nawar S, Mouazen AM. Optimal sample selection for measurement of soil organic carbon using on-line vis-nir spectroscopy. *Comput Electron Agric.* 2018;151:469–77.
9. Kuang B, Mouazen AM. Effect of spiking strategy and ratio on calibration of on-line visible and near infrared soil sensor for measurement in European farms. *Soil and Tillage Research.* 2013;128:125–36.
10. Guerrero C, Wetterlind J, Stenberg Bo, et al. Do we really need large spectral libraries for Local scale SOC assessment with NIR spectroscopy. *Soil Tillage Res.* 2016, 155:501–9.
11. Guerrero C, Stenberg B, Wetterlind J, et al. Assessment of soil organic carbon at local scale with spiked NIR calibrations: effects of selection and extra-weighting on the spiking subset. *Eur J Soil Sci.* 2014;65:248–63.
12. Capron X, Walczak B, Noord OED, et al. Selection and weighting of samples in multivariate regression model updating. *Chemom Intell Lab Syst.* 2005;76(2):205–14.
13. Jia SY. Research on the detection methods and instrumentation of soil properties using spectral analysis technology [D]. Hangzhou: Zhejiang University; 2015.
14. Chen LY, Zhao ZG, Liu F. An updating method of NIR model based on characteristic wavelength for yellow rice wine detection. *Spectrosc Spect Anal.* 2017;37(11):3414–8.
15. Dayal B, MacGregor JF. Recursive exponentially weighted PLS and its applications to adaptive control and prediction. *J Process Control.* 1997;7(3):169–79.
16. Dayal B, MacGregor JF. Improved PLS algorithms. *J Chemom.* 1997;11(1–2):73–85.
17. Mu SJ, Zeng YZ, Liu RL, et al. Online dual updating with recursive PLS model and its application in predicting crystal size of purified Terephthalic Acid (PTA) process. *J Process Control.* 2009;16(6):557–66.
18. Chen ML, Khare S, Huang B, et al. Recursive wavelength-selection strategy to update near-infrared spectroscopy model with an industrial application. *Ind Eng Chem Res.* 2013;52(23):7886–95.
19. Qin SJ. Recursive PLS algorithms for adaptive data modeling. *Comput Chem Eng.* 1998;22(4–5):503–14.
20. Helland K, Berntsen H, Borgen O, et al. Recursive algorithm for partial least squares regression. *Chemom Intell Lab Syst.* 1991;14(1–3):129–37.
21. Wang PL, Ye XF, Yang ZY. Quality prediction method based on adaptive updating of block-RPLS model. *Control Decision.* 2018;33(3):455–62.
22. He K X, Zhong M Y, Du W L. Weighted incremental minimax probability machine-based method for quality prediction in gasoline blending process. *Chemom Intell Lab Syst.* 2020, 196:103909.
23. He KX, Qian F, Cheng H, Du WL. A novel adaptive algorithm with near-infrared spectroscopy and its application in online gasoline blending processes. *Chemom Intell Lab Syst.* 2015;140:117–25.
24. Ren ML, Song YL, Chu W. An improved locally weighted PLS based on particle swarm optimization for industrial soft sensor modeling. *Sensors.* 2019;19(19):4099.
25. Tulsyan A, Wang T, Schorner G, et al. Automatic real-time calibration, assessment, and maintenance of generic raman models for online monitoring of cell culture processes. *Biotechnol Bioeng.* 2019;117(2):406–16.

26. Hu MH, Zhao Y, Zhai GT. Active learning algorithm can establish classifier of blueberry damage with very small training dataset using hyperspectral transmittance data. *Chemom Intell Lab Syst.* 2018;172:52–7.
27. Tang JY, Huang M, Zhu QB. Purity detection model update of maize seeds based on active learning. *Spectrosc Spect Anal.* 2015;35(8):2136–214.
28. Huang M, Tang JY, Yang B, et al. Classification of maize seeds of different years based on hyperspectral imaging and model updating. *Comput Electron Agric.* 2016;122:139–45.
29. Xie L, Yang Z, Tao D, et al. The model updating based on near infrared spectroscopy for the sex identification of silkworm pupae from different varieties by a semi-supervised learning with pre-labeling method. *Spectrosc Lett.* 2019;52(10):642–52.
30. Jin HP, Chen XG, Wang L, et al. Dual learning-based online ensemble regression approach for adaptive soft sensor modeling of nonlinear time-varying processes. *Chemom Intell Lab Syst.* 2016;151:228–44.
31. Tulsyan A, Garvin C, Undey C. Industrial batch process monitoring with limited data. *J. Process Control.* 2019;77:114–133.

Chapter 12

Pattern Recognition Methods



12.1 Introduction

In the practical application of molecular spectral analysis technology, there has always been a situation that we only need to know the class or grade level of the samples instead of knowing the certain components and their contents in the samples, which is the qualitative analysis problem. Wherein, the pattern recognition method in chemometrics is usually involved. Different samples can be classified and identified according to some common characteristics by spectral data, so as to find the internal relations between the measured samples and obtain decision-making information. Therefore, pattern recognition is a kind of important means to transform spectral data into information needed to solve practical problems.

Pattern recognition methods can be divided into supervised and unsupervised types according to the learning process (or training process). Supervised pattern recognition method is to use a group of known classes of samples as a training set and makes the computer learn from these already known samples. This method of calculating the classifier is also called “managed” or “teacher” of learning, in which the training set is “teacher”. The classification model is obtained by learning process so as to predict the class of unknown samples. The unsupervised method is a classification method that the class of samples is unknown in advance and the training process is not required.

The steps of establishing pattern recognition (or qualitative model) are generally composed of four parts, as shown in Fig. 12.1, including data acquisition, preprocessing, feature extraction and selection, and classification decision [1, 2].

Data acquisition includes collection of the sample and the measurement of spectral data, as well as identification and analysis of the class of training samples by traditional methods. The more samples and the stronger representativeness are, the more reliable the results are. Common spectral preprocessing methods include derivative, MSC, SNV, mean-centering and standardization, etc.

In pattern recognition, the spectral information is used as the original feature variable. The extraction of feature information is a crucial step. The selection of feature

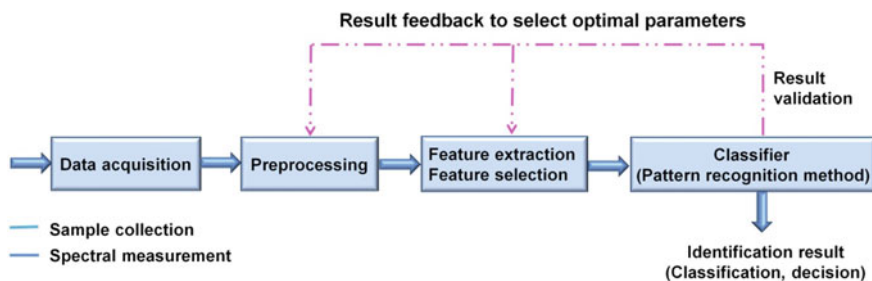


Fig. 12.1 Basic structure of pattern recognition system

variables will directly affect the results of classification or recognition. The purpose of feature selection is to make similar samples close to each other in the feature space and the heterogeneous samples far away. Feature extraction and compression are sometimes inseparable and they are generally carried out at the same time. The most commonly used method is principal component analysis (PCA) and the first several principal components scores with large eigenvalues are generally selected as feature variables to participate in pattern recognition.

Of course, in the practical application process, for some special analysis systems, the principal components with large eigenvalues are not the preferred feature variable. Thus, it is necessary to select variables from the principal components through optimization methods such as chemical knowledge or genetic algorithm, which is called feature selection. If these spectral features are only a part of the feature variables, other physicochemical parameters such as density are also involved in the feature variables. It is necessary to preprocess these feature variables by normalization or logarithmic transformation in order to eliminate the scale difference and increase the comparability among the variables. Other commonly used variable compression and extraction methods include: wavelength variables or their mathematical combination selected by chemical knowledge and optimization methods such as genetic algorithm; The wavelength intervals or their encompassing areas selected according to chemical knowledge; The coefficients or their mathematical combinations obtained by mathematical processing to spectra such as wavelet transform or Fourier transform, etc.

In spectral pattern recognition, the common supervised pattern recognition methods include minimum distance discriminant method, Bayes discriminant method, K-nearest neighbor method, BP neural network, soft independent modeling of class analogies (SIMCA), etc. The common unsupervised pattern recognition methods include cluster analysis and unsupervised neural networks.

12.2 Unsupervised Pattern Recognition Methods

In many practical problems of pattern recognition for samples, people often know nothing about the intrinsic classification of data in advance, and then unsupervised pattern recognition methods are needed. Clustering analysis is the representative of unsupervised methods. The main idea is to use similar samples to be similar to each other, which is often called “birds of a feather flock together”. Similar samples have small distances from each other in multi-dimensional space, while the distance between dissimilar samples should be larger. Clustering analysis is to make similar samples “together” so as to achieve the purpose of classification.

In spectral qualitative analysis, clustering analysis is widely used, such as clustering analysis of different classes of plant samples to study the genetic relationship between them. In addition, clustering analysis is often combined with quantitative multivariate calibration methods such as PLS or ANN. Firstly, the calibration samples are divided into several classes by clustering analysis. Secondly, models are established for each class of samples to improve the prediction ability of the model. This section mainly introduces the commonly used hierarchical cluster analysis (HCA) method, K-means clustering method, Fuzzy clustering method, and Kohonen neural network (KNN) for spectral qualitative analysis, etc.

It is worth noting that clustering analysis is actually a process that requires the participation of multiple parties. It cannot be separated from the participation of experts in this field. The clustering algorithm is only a part of the whole clustering process. Generally, satisfactory classification results cannot be achieved only by relying on pure mathematical clustering algorithms.

12.2.1 Similarity Coefficients and Distances

The important components of clustering analysis are the distance between samples, the distance between classes, the way of merging classes and the number of clusters. The first problem to be solved is the similarity between two samples. There are usually two definitions of intimacy between samples, which are similarity coefficient and distance. They regard each sample as a point in the m -dimensional space (m -variables), in which the degree of intimacy between samples is defined.

Similarity coefficient is expressed by cosine and correlation coefficient:

Cosine:

$$\cos \alpha_{ij} = \frac{\sum_{k=1}^m x_{ik}x_{jk}}{\sqrt{\sum_{k=1}^m x_{ik}^2 \sum_{k=1}^m x_{jk}^2}} \quad (12.1)$$

where x_{ik} represents the k th feature variables of the i th sample. If two samples are exactly the same, the angle cosine $\cos\alpha = 1$. Conversely, if they are completely different, $\cos\alpha = 0$.

Correlation coefficient;

$$r_{ij} = \frac{\sum_{k=1}^m (x_{ik} - \bar{x}_i)(x_{jk} - \bar{x}_j)}{\sqrt{\sum_{k=1}^m (x_{ik} - \bar{x}_i)^2 \sum_{k=1}^m (x_{jk} - \bar{x}_j)^2}} \quad (12.2)$$

\bar{x}_i, \bar{x}_j is the mean value of all feature variables in the i th and j th samples, respectively. The closer the two samples are, the closer the similarity coefficient between them is to 1 (or -1).

Distance is usually represented by Euclidean distance and Mahalanobis distance:
Euclidean distance:

$$D_{ij} = \sqrt{\sum_{k=1}^m (x_{ik} - x_{jk})^2} \quad (12.3)$$

Mahalanobis distance:

$$M_{ij} = \sqrt{(\mathbf{x}_i - \mathbf{x}_j) \mathbf{V}^{-1} (\mathbf{x}_i - \mathbf{x}_j)^T} \quad (12.4)$$

where \mathbf{x}_i and \mathbf{x}_j are the spectral row vectors of samples i th and j th. \mathbf{V}^{-1} is the inverse matrix of the class \mathbf{X} covariance matrix, i.e.

$$\mathbf{V}^{-1} = \left[\frac{1}{n-1} (\mathbf{X} - \bar{\mathbf{x}})^T (\mathbf{X} - \bar{\mathbf{x}}) \right]^{-1} = \left(\frac{1}{n-1} \mathbf{X}_{\text{cen}}^T \mathbf{X}_{\text{cen}} \right)^{-1} \quad (12.5)$$

The Mahalanobis distance between sample \mathbf{x}_i and a class of \mathbf{X} is

$$M_i = \sqrt{(\mathbf{x}_i - \bar{\mathbf{x}}) \left(\frac{1}{n-1} \mathbf{X}_{\text{cen}}^T \mathbf{X}_{\text{cen}} \right)^{-1} (\mathbf{x}_i - \bar{\mathbf{x}})^T} \quad (12.6)$$

where $\bar{\mathbf{x}}$ is the average spectrum of class \mathbf{X} , \mathbf{X}_{cen} is the spectral matrix after \mathbf{X} -means centralization.

In the actual calculation, the spectral data \mathbf{X} is usually replaced by the PCA score \mathbf{T} at this time.

$$M_i = \sqrt{(t_i - \bar{t}) \left(\frac{1}{n-1} \mathbf{T}_{\text{cen}}^T \mathbf{T}_{\text{cen}} \right)^{-1} (t_i - \bar{t})^T} \quad (12.7)$$

It can also be written as:

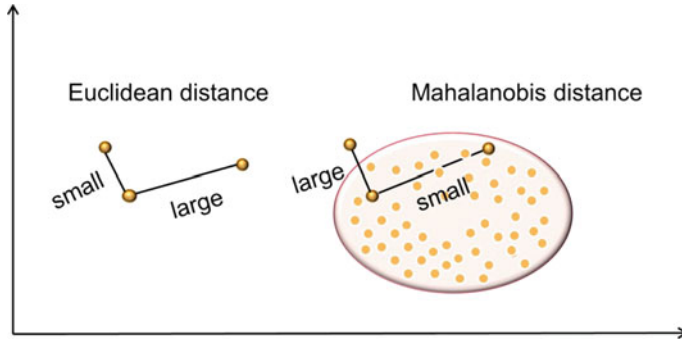


Fig. 12.2 Schematic diagram of Mahalanobis distance compared with Euclidean distance

$$M_i = \sqrt{(n-1) \sum_{j=1}^f \frac{(t_{ij} - \bar{t}_j)^2}{\lambda_j}} \quad (12.8)$$

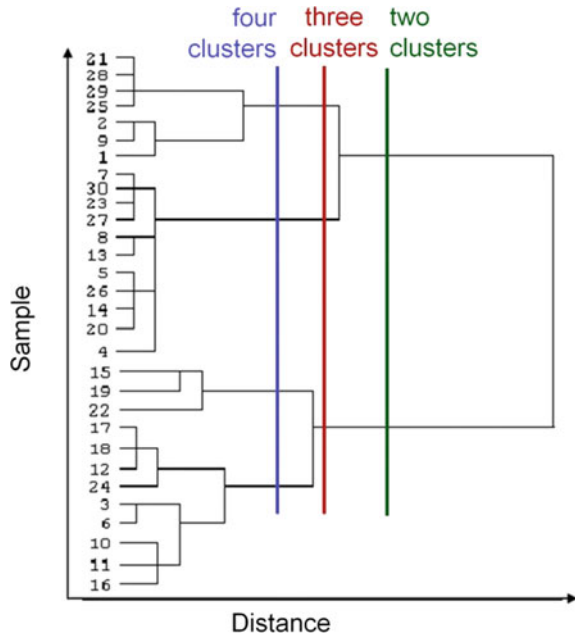
where t_{ij} is the j th principal component score of sample x_i , \bar{t}_j is an average score for the j th principal component of class \mathbf{X} , λ_j is the j th eigenvalue of a matrix $\mathbf{X}_{cen}^T \mathbf{X}_{cen}$, f is the selected number of principal components.

As can be seen from the above equation, compared with Euclidean distance, Mahalanobis distance takes the variation (variance) of the same feature variable in the same class and the variation (covariance) between different feature variables into account. Therefore, as shown in Fig. 12.2, for two samples in the same class, the Mahalanobis distance is small and the Euclidean distance may be large. On contrary, the two samples in different classes may have the large Mahalanobis distance and the Euclidean distance may be small. Since Mahalanobis distance takes the distribution of samples into account, it plays an important role in identifying samples outside the model.

12.2.2 Hierarchical Cluster Analysis

Hierarchical cluster analysis (HCA) is also known as pedigree clustering method, which is one of the most widely used clustering analysis methods. It adopts a non-iterative hierarchical clustering strategy. The basic idea is that each sample is considered to be self-classification, and then the distance between classes is specified. First of all, since each sample is a class of itself and the distance between classes is equivalent. The pair with the smallest distance is selected and merged them into a new class. The distance between the new class and other class is calculated. Second of all, the two classes with the smallest distance are merged into a class so that each class is

Fig. 12.3 Final result obtained by hierarchical cluster analysis



reduced until all samples are clustered into a class. According to the merging process of samples, we can get the pedigree figure of HCA (as shown in Fig. 12.3). It can show all the intermediate situations from the process of all samples being classified into one class to the whole being classified into one class in detail. It reflects the affinity of all samples from coarse to fine. Then according to certain principles, such as domain experts choose the appropriate classification threshold by experience or domain knowledge to determine the final classification results.

In the hierarchical clustering method, there are many definitions of distance between classes. Therefore, the hierarchical clustering method is divided into many methods according to the definition of distance between classes, such as single linkage, complete linkage, median method, centroid method, average linkage, flexible-beta method and Ward's minimum-variance method.

There are five most commonly used distance methods.

- (1) Single linkage: the distance between classes is equal to the distance between the nearest two samples of two classes.
- (2) Complete linkage method: the distance between classes is equal to the distance between the farthest two samples of two classes.
- (3) Median method: the distance between classes is neither the nearest distance between the two classes nor the furthest distance instead of taking the distance between the two.
- (4) Centroid method: It starts from the physical center. It represents class by the centeroid and uses the distance between two kinds of centeroids to describe the similarity between classes.

- (5) Ward's minimum-variance method: It is also called Ward method in some literature, which adopts uneven judgment rules. From the perspective of variance analysis, it is believed that the correct classification should make the intra-class variance as small as possible and the inter-class variance as large as possible.

The steps of HCA are as follows:

- (1) The beginning of the clustering analysis process is that each sample forms its own class (There are n classes of n samples), and then calculates the distance between each sample and merges the two samples with the nearest distance into one class.
- (2) Select and calculate the distance between classes, then merge the nearest two classes. If the number of classes is greater than 1, continue to merge classes until all samples are classified as one class.
- (3) Finally, draw the hierarchical clustering pedigree figure.

Hierarchical clustering method can get a complete clustering pedigree figure, which can explain all clustering schemes from class 1 to class n in detail. It is one of the most widely used methods in practice. However, using different calculation methods between inter-class, distance, the results are not exactly the same and sometimes get very different clustering results. Generally speaking, single linkage method is applicable to the class of long strip or S-shaped distribution. Complete linkage, centroid method and Ward method are applicable to the class of ellipsoidal distribution. In the preliminary clustering analysis, different distance methods should be investigated first and the optimal distance is determined by comparing their results.

12.2.3 *K-Means Clustering*

By using HCA, once a sample is divided into a certain class, it will be not changed. This requires the division must be very accurate. Moreover, since HCA needs to calculate the distance matrix, the storage cost is large when dealing with a dataset with a large sample size. MacQueen et al. proposed a dynamic clustering method based on iterative operation in 1967. Firstly, a rough preliminary classification were given, then the clustering results were dynamically modified according to certain principles until reasonable classification results were obtained. Dynamic clustering method usually requires artificially to give the number of classes k or some thresholds in advance.

K-means clustering method is a commonly used dynamic clustering analysis method. It divides the samples to be clustered into k classes according to the number of classes k determined in advance, so that the sum of squares of distances from all samples in the clustering domain to the clustering center is minimized.

The algorithm is an iterative process with the following steps.

- (1) Firstly, k samples are randomly selected from n clustering samples $\{x_1, x_2, \dots, x_n\}$ as the initial clustering centers.

- (2) Calculate the distance between each sample and the k clustering centers and divides them into the nearest class.
- (3) Calculate the mean value of each point in each class as a new central point.
- (4) Calculate the distance between each sample and these new centers then reclassify them according to the minimum distance principle.
- (5) Calculate square error function

$$J = \sum_{i=1}^k \sum_{j=1}^n d_{ij} \| \mathbf{x}_j - \mathbf{w}_i \|^2 \quad (12.9)$$

where \mathbf{w}_i is the clustering center of class i , k is the number of clusters, n is the number of samples and d_{ij} is used to indicate whether \mathbf{x}_j of sample j th belongs to class i . If \mathbf{x}_j belongs to class i , $d_{ij} = 1$. If \mathbf{x}_j does not belong to class i , $d_{ij} = 0$.

- (6) Repeat steps (3–5) above until J does not change significantly or reaches a pre-set maximum number of iterations.

K-means clustering algorithm has a clear idea, simple algorithm and fast convergence speed, which is more suitable for large sample size. Therefore, it has been widely used. However, this method requires domain experts to determine the number of clusters k in advance. If the selection is not appropriate, the final classification result will be affected. Moreover, this method is sensitive to the center point of the initial clustering and sometimes convergences to local optimal solution due to improper selection.

Aiming at resolving the weakness of K-means clustering algorithm, there are many improved algorithms. For example, the iterative self-organizing ISODATA algorithm proposed by the United States Bureau of Standards is one of the representative algorithms. The ISODATA algorithm has six parameters. When there are too many and too scattered elements in a certain class, it can be divided into two class. When there are few samples in a certain class, it performs the merging operation with another class. Such a self-organizing process is more flexible to control the number of classes and has better adaptability and flexibility than the K-means algorithm. However, there are many parameters in this algorithm, which make it difficult to optimize the whole algorithm.

At present, most of the global optimization methods (such as genetic algorithm, simulated annealing algorithm, ant colony algorithm, and particle swarm algorithm) are used to improve the K-means clustering algorithm [3–5] to obtain the optimal clustering numbers and clustering centers.

The following is a brief introduction about the K-means clustering method based on GA. This method tries to obtain the global optimal solution by GA and improves the convergence speed by K-means method. Firstly, the first generation of GA is randomly generated and evolved. In each generation of evolution, the K-means method is used to further optimize each individual and these local optimal results are used to replace the original individual and continue to evolve until the maximum

number of iterations or the results meet the requirements. Based on different coding, evolutionary strategies, and fitness functions, a variety of genetic K-means clustering algorithms can be designed. The general steps of such algorithms are as follows:

- (1) The fitness function is defined and the genetic parameters are set. Such as the number of clustering, population size, crossover probability, mutation probability, and the maximum number of iterations.
- (2) The initial population is generated randomly.
- (3) The fitness of each individual in the population is calculated.
- (4) Crossover, mutation, and K-means clustering operations are selected to generate a new generation of groups.
- (5) Steps (3–4) are repeated until the maximum number of iterations is reached.
- (6) The fitness of the new generation is calculated and the optimal individual with the maximum fitness is taken as the final K-means clustering result.

12.2.4 Fuzzy K-Means Clustering

Since the boundaries between objective things are not often very clear, it is undoubtedly appropriate to introduce fuzzy mathematics into clustering analysis to deal with the clustering problem of fuzzy things. In fact, fuzzy clustering analysis is one of the most rapidly developed clustering methods in recent years. Among them, fuzzy K-means clustering algorithm is one of the most popular algorithms in fuzzy clustering methods in current.

The clustering criterion function of classical K-means clustering algorithm is the sum of error squares function.

$$J = \sum_{i=1}^k \sum_{j=1}^n d_{ij} \| \mathbf{x}_j - \mathbf{w}_i \|^2 \quad (12.10)$$

where \mathbf{w}_i is the clustering center of class i , k is the number of clusters, n is the number of samples and d_{ij} is used to indicate whether \mathbf{x}_j of sample j belongs to class i . If \mathbf{x}_j belongs to class i , then $d_{ji} = 1$. If \mathbf{x}_j does not belong to class i , then $d_{ji} = 0$. d_{ji} is either 1 or 0. However, it is not so absolute in practice, \mathbf{x}_j belongs to a class of membership (μ_{ij}) that is often a number between 0 and 1. Therefore, the fuzzy K-means clustering algorithm changes d_{ij} to μ_{ij} , $\mu_{ij} \in [0, 1]$ and its clustering criterion function is changed to:

$$J = \sum_{i=1}^k \sum_{j=1}^n (\mu_{ij})^m \| \mathbf{x}_j - \mathbf{w}_i \|^2 \quad (12.11)$$

where μ_{ij} is the membership degree of sample \mathbf{x}_j to class i , and $\sum_{i=1}^k \mu_{ij} = 1$. The sum of membership degree of each sample is 1. m is the weighted index and $m > 1$,

which is to enhance the contrast of x_j of belonging to various degrees. The greater the value of m is, the greater the fuzzy degree of the classification matrix is. Generally, m is taken as 1.1–2.0. $\|x_j - w_i\|$ is the Euclidean distance between sample x_j and cluster center w_i .

It can be seen from the above that the objective function J represents the sum of squares of weighted distances between sample x_j and each cluster center w_i , and then its weight is the m th power of sample x_j belonging to the membership degree μ_{ij} of cluster w_i , while the optimal clustering is to minimize the objective function J . Therefore, in order to get the best clustering result, the appropriate membership degree μ_{ij} and clustering center w_i are required. It can be proved that when $m > 1$, $x_j \neq w_i$, the membership degree μ_{ij} and clustering center w_i can be calculated by the following two iterative Equations.

$$\mu_{ij} = \frac{\left(\frac{1}{\|x_j - w_i\|^2}\right)^{\frac{1}{m-1}}}{\sum_{h=1}^k \left(\frac{1}{\|x_j - w_h\|^2}\right)^{\frac{1}{m-1}}} \quad (12.12)$$

$$w_i = \frac{\sum_{j=1}^n (\mu_{ji})^m x_j}{\sum_{j=1}^n (\mu_{ji})^m} \quad (12.13)$$

The specific algorithm of fuzzy K-means clustering method is as follows [6].

- (1) Fixed classification k , weighted index m , and convergence threshold ε (generally 0.01). The initial membership degree matrix $U^{(0)}$ is selected and the element μ_{ij} meets the requirement:

$$\begin{aligned} 0 &\leq \mu_{ij} \leq 1, \forall i, j \\ \sum_{i=1}^k \mu_{ij} &= 1, \forall j \end{aligned} \quad (12.14)$$

- (2) According to the calculation equation of clustering center and $U^{(q)}$, the clustering center $w_i^{(q)}$ is calculated and q is the number of iterations.
- (3) $U^{(q+1)}$ is obtained from the obtained $w_i^{(q)}$ and the membership degree calculation equation.
- (4) If $\max \{|U^{(q)} - U^{(q+1)}|\} \leq \varepsilon$, then stop the iteration, $U^{(q+1)}$ and the corresponding $w_i^{(q)}$ are the results. Otherwise, return step (2) and continue iteration.
- (5) In the obtained membership degree matrix U , let the maximum element in each column be 1 and the rest be 0. A general classification matrix is obtained, which is the classification result.

12.2.5 Gaussian Mixture Model

Gaussian mixture model (GMM) is the sum of multiple single Gaussian models (as shown in Fig. 12.4), which represents the probability density function of data by linear combination of multiple Gaussian functions. Its expression ability is very strong, so any distribution can be expressed by GMM. The mathematical form of Gaussian mixture model is as follows.

$$p(\mathbf{x}) = \sum_{k=1}^K w_k g_k(\mathbf{x} | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \quad (12.15)$$

where K is the number of single Gaussian models. g_k is the single Gaussian model with mean value $\boldsymbol{\mu}_k$ and covariance matrix $\boldsymbol{\Sigma}_k$. w_k is the weight coefficient of g_k and satisfies the following constraints.

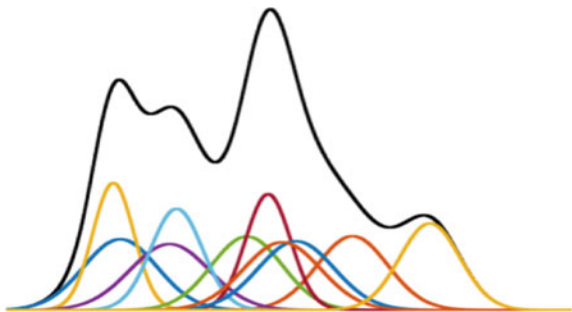
$$w_k > 0 \quad \sum_{k=1}^K w_k = 1 \quad (12.16)$$

The parameters $\boldsymbol{\mu}_k$, $\boldsymbol{\Sigma}_k$, and w_k of GMM are usually obtained by expectation maximization algorithm (EM). It is an iteration method that estimates the class of each sample and the probability distribution parameters of each class. EM algorithm is a local optimization algorithm, which is sensitive to the setting of initial parameters and easy to fall into local optimum. Therefore, intelligent optimization algorithms (such as particle swarm optimization algorithm) can also be used to obtain the optimal model parameters.

In addition to clustering analysis, GMM can also be used for regression calculation. Gaussian mixture regression (GMR) predicts the joint density of future objects by constructing a series of GMM and then obtains the probability density and regression function from each GMM.

Li et al. extracted the feature of GC-MS signals of tea by PCA and combined with 10 variables such as tea polyphenols measured by liquid chromatography. They

Fig. 12.4 Schematic diagram of Gaussian mixture model



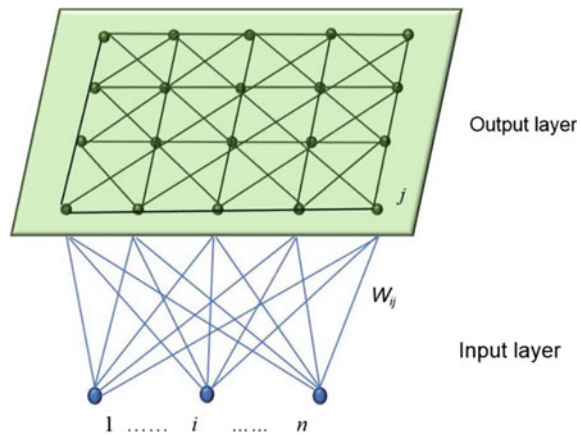
used GMM to classify the tea samples and the accuracy of prediction set reached 90% [7]. Sun et al. used GMM based on particle swarm optimization and GMR combined with mid-infrared spectroscopy to qualitatively and quantitatively analyze olive oil adulteration samples, and achieved good results [8]. Wang et al. used near infrared spectroscopy (NIR) combined with GMR method to predict and analyze the growth process of yeast. The results are superior to the methods of kernel partial least squares, support vector machine, and extreme learning machine [9].

12.2.6 Self-organizing Neural Network

Self-organizing neural network is a kind of learning neural network without teachers. It can simulate that human beings can automatically adapt to unpredictable environmental changes according to past experience. Through self-training, the network can automatically classify the input samples. Because there is no teacher signal, self-organizing neural network usually uses competition principle to conduct online study. In the competition network, the output layer is also called competition layer. The weights connect with input nodes and inputs are called the input layer together. The activation function of the competition network is called the binary $\{0,1\}$ function. The most typical self-organizing neural network is Kohonen self-organizing feature map (SOM) that was proposed by Kohonen in 1981, which is also called Kohonen network [10].

The Kohonen network structure is shown in Fig. 12.5. It is a simple two-layer network. The number of input layer neurons is m and competition layer consists of q^2 neurons. They form a two-dimensional plane array. The competition layer of the two-dimensional matrix is called the output layer. The input layer node and the competition layer node are fully interconnected. Sometimes the neurons in the competition layer are connected by lateral inhibition.

Fig. 12.5 Schematic diagram of Kohonen network structure



The learning process of Kohonen network is divided into two steps: competition learning process and the lateral interaction process of the neurons in the output layer. For each input vector, by comparing the input vector with the weight vector, there are competitions between the neurons. The neuron whose weight vector is closest to the input vector is considered as the strongest response of the input vector. The neuron is winning. The neuron is called the image of the input vector. Obviously, the same input vector produces the same image in competition layer. Lateral feedback process of neurons in the output layer for each input vector can cause the nearby neurons to generate lateral feedback according to the following rules. On the one hand, with the winning neuron as the center of the circle, excitatory lateral feedback was shown to the nearby neurons. On the other hand, with the winning neuron as the center of the circle, inhibitory lateral feedback is shown to the neurons of distant neighbors. The result of the lateral feedback is the formation of a clustering near each winning neuron. The result of studying makes the weight vectors of each neuron in the clustering area keep the trend of approaching to the input vector. Thus, the input vector with similar characteristics is gathered together and this process is self-organizing. Kohonen network realizes clustering by using the lateral feedback process of neurons in the output layer.

It can be seen that the working principle of Kohonen network is to map any dimensional input patterns into a one-dimensional or two-dimensional discrete graph at the output layer and keeps its topological structure unchanged. For example, samples that are closing in the high-dimensional space are still closing in the two-dimensional space. An advantage of this approach is that the result of the mapping is easy to visualize. In addition, the network can make the probability distribution of the weight vector space and the input pattern consistent through repeated learning of the input pattern. The weight vector space can reflect the statistical characteristics of the input pattern.

The self-organizing learning process of Kohonen networks can be summarized as follows.

- (1) All neurons in the competition layer are given an initial weight matrix W , whose element W_{ij} represents the weight between the i th feature variable of the input vector x and the neuron j in the competition layer.
- (2) The Euclidean distances d_j between the input sample vector x and the ownership weight vector w_j are calculated and then the winning neuron is determined with the shortest corresponding distance and it is marked as j^* .
- (3) The winning neuron is adjusted according to the following Eq. 12.17.

$$w_{j^*}(t+1) = w_{j^*}(t) + \eta[x - w_{j^*}(t)] \quad (12.17)$$

where $w_{j^*}(t)$ is the weight vector of the t th iteration number of the winning neuron j^* . η is the learning rate. In general, the initial value η_0 of η is relatively selected to be larger and generally $0.2 \sim 0.5$, which is used to accelerate the calibration speed of the connection weight. As the number of iterations increases, η gradually decreases and coarse tuning is replaced by fine tuning

to avoid the possible oscillation phenomenon in the network learning process. The typical functional form of η is

$$\eta(t) = \eta_0(1 - t/T) \quad (12.18)$$

where t is the current number of iterations and T is the total number of iterations.

- (4) The neurons in the nearby field with the winning neuron j^* as the center and the radius of $r(t)$ are also adjusted. The adjusted area is generally uniform and symmetrical. The most typical area is a square or circular area.

$$w_j(t+1) = w_j(t) + \eta N(t)[x - w_j(t)] \quad (12.19)$$

where $N(t)$ is a domain function that is presented. It is also called the neighborhood or near neighborhood function. A variety of domain functions can be chosen. The general choosing principle is that neurons close to the winning neuron j^* are adjusted to a great extent. Generally, in the initial stage of learning, the value of r is large. The range of the adjusted domain is large, which is generally 1/3 to 1/2 of the range of the matrix of the competition layer and can even cover the whole competition layer. With the deepening of learning, this range gradually decreases and finally only includes the neuron j^* . The commonly used neighborhood function is

$$N(t) = \text{int}[N_0(t)(1 - t/T)] \quad (12.20)$$

where $\text{int}(x)$ represents the integer symbol and $N_0(t)$ is the initial value of $N(t)$.

- (5) Set $t = t + 1$ and return to Step (2) until the weight vector has no significant change or $t = T$.

There are several other parameters that are very important in practical applications. The first is the determination of the number of neurons q in the competition layer. The node m of the input layer is determined by known input feature variable, but the neuron q of the competition layer is self-determined according to the actual problem. It represents the number of class that the input samples may be classified. If the value is selected too small, some input samples may not be classified as bad results. If the value is selected too big, many nodes may be idle after competition, resulting in wasting to some degree. The second is the determination of the initial value of the weight vector. Generally, the initial weight vector W_{ij} is assigned to a random value in the interval of $[0, 1]$. In practical application, this initial method takes much time to learn and even fail to converge. Since the ideal distribution of the initial state of the connection weight is consistent with the direction of each input sample. Therefore, during weight initialization, the initial state and the input sample should be as close as possible in mutually accessible. The common method is to assign the ownership to the same initial value or make it within a small range, such as $[0.5-0.05, 0.5 + 0.05]$. In this way, the selection of weights in the initial stage of the total input samples can

be reduced and the chance of each connection weight being selected can be increased so as to correct the direction deviation between the connection weight and the input sample as soon as possible.

The network model established after learning and training is the classification model. When the spectral data of the unknown sample are input into the network model, the class is represented by the neurons in the output layer that finally win the competition is the class of the sample.

12.3 Supervised Pattern Recognition Methods

The general idea of supervised pattern recognition methods is to use a group of samples with known classes as training set, and then let the computer “learn” from these known samples. Thus, this pattern recognition method for obtaining classifier is called “supervised learning”. The training set is the manager, and the discriminant model for unknown samples is obtained by the training set.

Common methods include minimum distance discriminate (MDD), Bayes linear discriminate (BLD), Fisher linear discriminate (FLD), linear learning machine (LLM), K-nearest neighbor method (KNN), classification with potential function (CPF), soft independent modeling of class analogy (SIMCA), artificial neural network (ANN), support vector machine (SVM), etc.

12.3.1 Minimum Distance Discriminant Method

The MDD is one of the simplest classifiers. If the covariance matrix of each class is similar and the prior probabilities of each class are equal, for the discriminant analysis of unknown samples x_{un} , we only need to calculate the square value of Euclidean distance between x_{un} and the mean of the given class \bar{x}_j :

$$d_{un,j}^2 = \left\| x_{un} - \bar{x}_j \right\|^2, \quad j = 1, \dots, k, \quad (12.21)$$

In Eq. 12.21, k is the number of classes and then x_{un} is judged to class with the smallest distance.

If the covariance matrix of various classes differs greatly (Fig. 12.6), for unknown sample x_{un} , we need to calculate the square \bar{x}_j of the Mahalanobis distance between x_{un} and the mean of the given class:

$$md_{un,j}^2 = \left(x_{un} - \bar{x}_j \right) H_j^{-1} \left(x_{un} - \bar{x}_j \right)^T, \quad j = 1, \dots, k, \quad (12.22)$$

In Eq. 12.22, k is the number of classes. H_j is the covariance matrix of class j th,

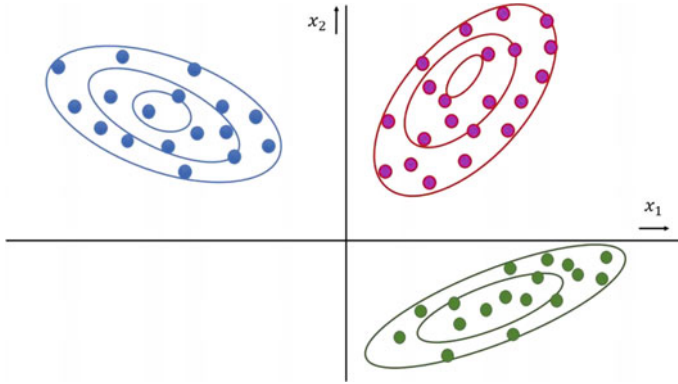


Fig. 12.6 Schematic diagram of multi-classes with different covariances

$$H_j = \frac{1}{g_j - 1} (X_j - \bar{x}_j)^T (X_j - \bar{x}_j) \quad (12.23)$$

In Eq. 12.23, g_j is the sample number of class j th.

If the prior probabilities of various classes are different, the Bayes discriminant analysis method is needed. In this case, the discriminant function from the unknown sample x_{un} to class j th is:

$$d_j(x_{un}) = (x_{un} - \bar{x}_j) H_j^{-1} (x_{un} - \bar{x}_j)^T + \ln |H_j| - 2 \ln P(j) \quad (12.24)$$

$P(j)$ is the prior probability of class j th,

$$P(j) \approx \frac{g_j}{n} \quad (12.25)$$

where n is the total number of samples of all classes; g_j is the sample number of class j th, $|H_j|$ is the determinant of the matrix. This method is also called quadratic discriminant analysis (QDA) in some literature.

12.3.2 Canonical Variate Analysis

The correlation coefficient is used to measure the correlation between two variables. However, if the correlation between two groups of variables (two matrices) is studied, it is necessary to transform the correlation between two groups of variables into the correlation between two variables for consideration. To investigate the correlation

between the linear combination of the first group of variables and the linear combination of the second group of variables, the linear coefficient is selected to make the linearization variables have the maximum correlation coefficient and to form the first pair of canonical variables. Then the second pair and the third pair of canonical variables can be formed and each pair of canonical variables is uncorrelated. In this way, the correlation between the two groups of variables is transformed into the correlation between several pairs of canonical variables.

Because a group of variables can have a number of linear combinations (the linear combinations are determined by the correlation coefficient), it is necessary to find the linear combinations that are both meaningful and determinable. Canonical correlation analysis (CCA) is also called canonical variate analysis (CVA) [11, 12]. In order to find coefficient of linear combination of the two groups of variables so that the correlation coefficient between the two variables generated by the linear combination (compared with other linear combinations) is the largest. CVA is a statistical analysis method of studying the correlation between two groups of variables and it is also a common data dimensionality reduction technique.

CVA can offer feature variable for multi-class discriminant analysis. The Fisher linear discriminant analysis (LDA) is often mentioned in literature. The spectral matrix X ($n \times m$) of the training set contains k classes of samples and there are g_i samples in each class.

$$n = \sum_{i=1}^k g_i \tag{12.26}$$

The inter-class and intra-class covariance matrices are calculated according to Fig. 12.7.

Intra-class covariance matrix

$$S_W = \frac{1}{n - k} \sum_{i=1}^k \sum_{j=1}^{g_i} \left(x_{ij} - \bar{x}_i \right) \left(x_{ij} - \bar{x}_i \right)^T \tag{12.27}$$

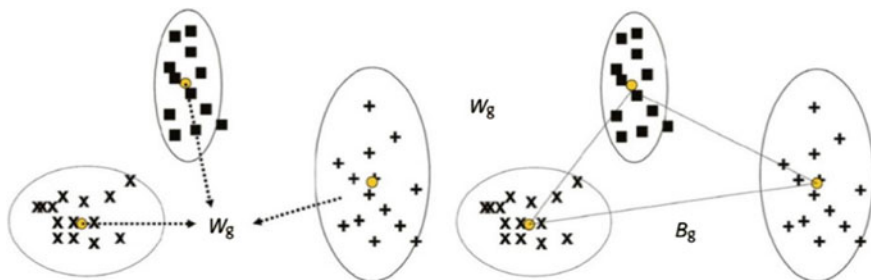


Fig. 12.7 Schematic diagram of inter-class and intra-class covariances

Inter-class covariance matrix

$$S_B = \frac{1}{k-1} \sum_{i=1}^k g_i \left(\bar{\mathbf{x}}_i - \bar{\mathbf{x}} \right) \left(\bar{\mathbf{x}}_i - \bar{\mathbf{x}} \right)^T \quad (12.28)$$

where \mathbf{x}_{ij} is the spectral vector of the j th sample of class i th.

$$\bar{\mathbf{x}}_i = \frac{1}{g_i} \sum_{j=1}^{g_i} \mathbf{x}_{ij} \quad (12.29)$$

The average spectra of class i th are calculated by Eq. 12.30.

$$\bar{\mathbf{x}} = \frac{1}{n} \sum_{i=1}^k \bar{\mathbf{x}}_i \quad (12.30)$$

where $\bar{\mathbf{x}}$ is the average spectra of n samples. S_W and S_B are both $m \times m$ matrices.

CVA aims to get the maximum of the objective function $J(\mathbf{w})$:

$$J(\mathbf{w}) = \frac{\mathbf{w}^T S_B \mathbf{w}}{\mathbf{w}^T S_W \mathbf{w}} \quad (12.31)$$

The solution can be transformed into the eigenvalue and eigenvector problem of the matrix.

$$S_B \mathbf{w} = \lambda S_W \mathbf{w} \quad (12.32)$$

i.e.

$$S_W^{-1} S_B \mathbf{w} = \lambda \mathbf{w} \quad (12.33)$$

In fact, CVA is to calculate the eigenvalue and eigenvector of $S_W^{-1} S_B$. It can simplify the multivariable data to each other into a few uncorrelated new variables data and the simplified data can keep most of the information of the original data. The discriminant function can be obtained by taking the first several canonical variables with high contribution rates as features.

The first linear discriminant function (the first score of CVA) can be given by the eigenvector \mathbf{w}_1 based on the maximum eigenvalue λ_1 : $S_{1i} = \mathbf{x}_i \mathbf{w}_1^T$. The second linear discriminant function (the second score of CVA) can be given by the eigenvector \mathbf{w}_2 based on the second eigenvalue λ_2 : $S_{2i} = \mathbf{x}_i \mathbf{w}_2^T$. This calculation can be continued

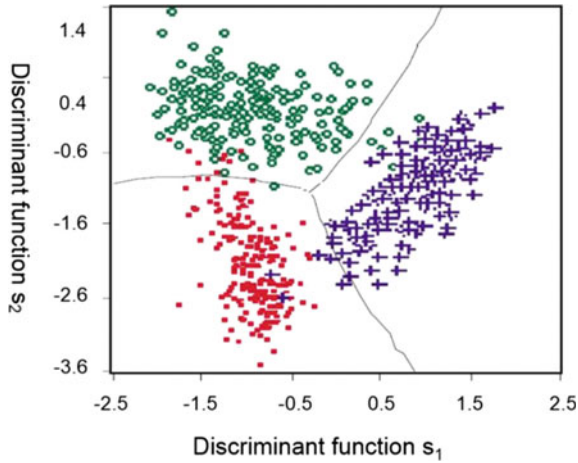


Fig. 12.8 Schematic diagram of sample distribution by using CAV transformation

until all the discriminant functions need to solve the identification problem that is found. The space distribution of all samples can be observed by drawing with different discriminant functions as the coordinate axis after CAV transformation (Fig. 12.8).

As shown in Fig. 12.9, in some cases, the principal component direction of PCA and the direction of CAV discriminant functions are basically the same. However, in some cases, the directions of the two methods are different. This is because PCA transforms by selecting the direction with the maximum variance of variable, while

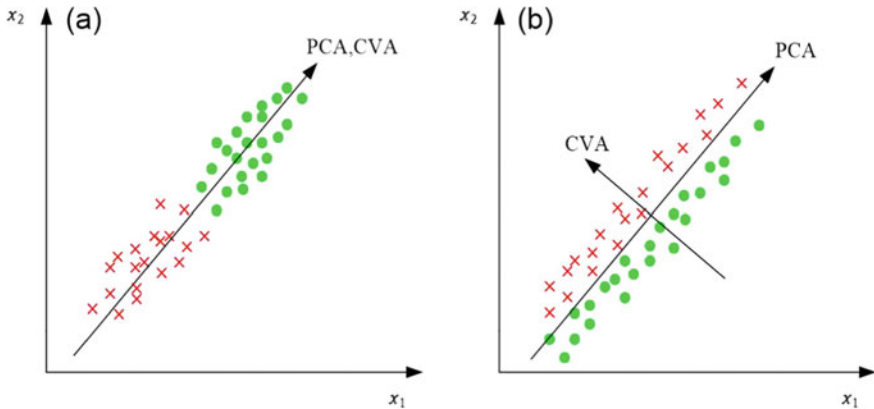


Fig. 12.9 The case for PCA and CVA basically the same (a) and significant differences (b)

CAV transforms by selecting the direction that can separate all known classes in the greatest degree.

For the discriminant analysis of unknown samples spectra x_{un} , only the discriminant function is needed to substitute to classify it as class of the smallest Euclidean distance from the center of the class:

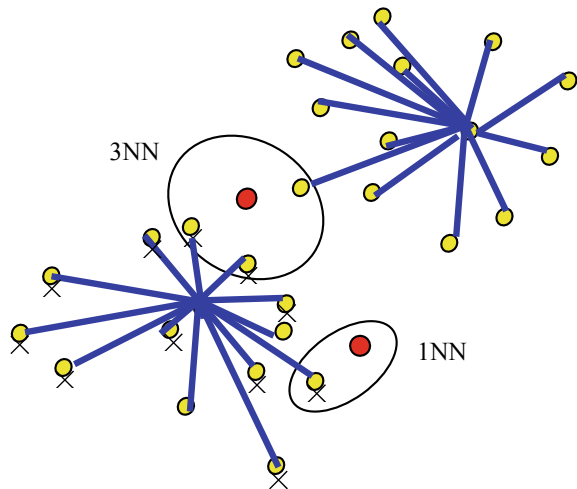
$$\min_j \left\| (x_{un} - \bar{x}_j) w^T \right\| \quad (12.34)$$

12.3.3 *K-Nearest Neighbor*

Different from other distance discriminant methods, the nearest neighbor method does not compare the distance between the samples to be measured and all kinds of mean values. Instead, the distance between it and all training samples is calculated. As long as the distance is the closest, it is classified into the class. In fact, the nearest neighbor method stores all the samples in training set and calculates the distance between the unknown samples and the training set samples one by one. In order to overcome the high error rate of the nearest neighbor method, k nearest neighbor samples (Fig. 12.10) is selected instead of only one nearest neighbor for classification. Then samples are classified into couples of classes with the largest proportion.

The final class is determined by using discriminant function method. For example, the discriminant function S can be calculated according to the following equation for the discriminant problem of two classes:

Fig. 12.10 Schematic diagram of K -nearest neighbor



$$S = \sum_{i=1}^k (S_i / D_i) \quad (12.35)$$

In Eq. 12.35, S_i is the value of the i th sample in the k samples of training set. If it belongs to the first class, the S_i takes “+1”. If it belongs to the second class, the S_i takes “-1”. D_i is the distance between the unknown samples and the i th sample. D_i can be understood as the weight. The training set sample with a smaller distance is given a larger weight, while the training set sample with a larger distance is given a smaller weight. Obviously, in the same number of samples, the larger D_i is, the smaller contribution to the total S value is. In the case of the same distance, the more samples of class 1, the more positive the total S value is. Therefore, if the calculation S value is positive, the unknown sample belongs to class 1. Instead, it belongs to class 2.

The advantage of KNN method is that it does not require several classes of samples of the training set to be linearly separable, nor does it require a separate training process. It is also easy to add samples of known classes into the training set and it can deal with multi-classes of problems. So it is convenient to apply. The main problem of this method is the selection of k value. Because the number and distribution of samples in each class are different, if different k values are selected, the discriminant results of unknown samples may be different. There is no certain rule to follow in the selection of k value, which can only be determined by specific circumstances or experience. It is usually inappropriate to choose a smaller k value.

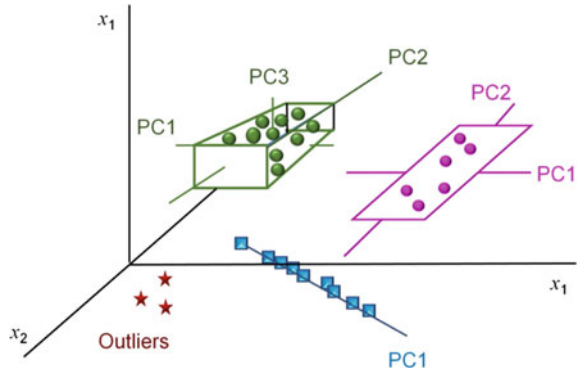
The KNN method is simple and effective. If an appropriate calculation method is defined to characterize the similarity between samples. It can achieve good performance. Therefore, the key to the application of KNN is to construct spectral feature variables and determine an appropriate distance function. In fact, spectral searching method is an extension of KNN.

12.3.4 Soft Independent Modeling of Class Analogy

Soft independent modeling of class analogy (SIMCA) is also called similarity analysis. It was proposed by Swedish chemist Wold in 1976. It has been widely used in chemical pattern recognition. SIMCA classification method is a supervised pattern recognition method, which is based on PCA. The basic idea of the algorithm is to use PCA on the spectral matrix of each class of samples in the training set, respectively. The PCA mathematical model is built for each class. Then the unknown samples are classified based on the models. The unknown samples are tried to fit each model to determine which class it belongs to or not.

The SIMCA method has two main steps to carry out discrimination and classification. The first step is to establish the PCA model for each class. The second step is to fit each class of PCA models of unknown samples one by one.

Fig. 12.11 Schematic diagram of SIMCA model with different PC numbers



NIPALS method can be adopted for PCA required in SIMCA method. The principle and algorithm of PCA have been introduced in detail in Chap. 6 and are not repeated here. For each class in the training set, the following PCA models are established, respectively.

$$\mathbf{X}_k = \mathbf{T}_k \mathbf{P}_k^t + \mathbf{E}_k \tag{12.36}$$

\mathbf{X}_k is the spectral matrix ($n \times m$) of all samples for k th class in the training set. n is the number of samples for k th class. m is the number of wavelength variables. \mathbf{T}_k is the score matrix ($n \times f$). f is the optimal principal component (PC) number. \mathbf{P}_k is the loading matrix ($m \times f$). \mathbf{E}_k is spectral residual matrix ($n \times m$).

The optimal number of PC f in each class model can be determined by cross validation. Each independent model can select different PC number. Thus, different classes of models may be expressed as line, plane, box, and superbox shape, as shown in Fig. 12.11.

If the spectral residual matrix \mathbf{E}_k conforms to normal distribution, the spectral residual variance s^2 can be calculated according to Eq. 12.37.

$$s^2 = \sum_{i=1}^n \sum_{j=1}^m \frac{e_{ij}^2}{(n - f - 1)(m - f)} \tag{12.37}$$

where e_{ij} is the spectral residual matrix of sample i at wavelength j .

For the unknown sample x_{new} , the score vector t_{new} and the residual spectrum e_{new} are first calculated by Eqs. 12.38 and 12.39.

$$t_{new} = x_{new} P_k \tag{12.38}$$

$$e_{new} = x_{new} - t_{new} P_k^T \tag{12.39}$$

Then the variance of spectral residual is calculated.

$$s_{new}^2 = \sum_{i=1}^m \frac{e_{ij}^2}{m-f} \quad (12.40)$$

If the variance s_{new}^2 and the total residual variance s_K^2 of class k have similar number of magnitude, the sample can be classified into class k . If s_{new}^2 is larger than s_K^2 significantly, the sample does not belong to class k .

F-significance test can also be adapted to conduct class analysis of unknown samples. F-statistic is defined as Eq. 12.41.

$$F = \frac{s_{new}^2}{s_k^2} \quad (12.41)$$

Comparing the calculation F statistic with the one-sided critical value $F_0 [\alpha, (m-f), (n-f-1)(m-f)]$, the confidence level α is generally 0.05 or 0.01. If $F < F_0$, the unknown samples belong to k th class. Otherwise, the samples are fitted to other classes until the class is determined. If the sample does not belong to any class in the training set, it can be classified into a new class.

The SIMCA method is based on PC spectral residual to identify unknown samples. There is a phenomenon in practical application that although the unknown samples conform to a certain class of PCA model, the samples may be far away from the training set samples of this class. Therefore, it is common to add a step in the SIMCA method that limits it by the PC score:

$$t_{\max} = \max(tk) + 0.5st \quad (12.42)$$

$$t_{\min} = \min(tk) - 0.5st \quad (12.43)$$

where $\max(t_k)$ and $\min(t_k)$ are the maximum and minimum element values of the score vectors and are obtained from the PCA of the k th sample in training set, respectively. Here s_t is the standard deviation of the corresponding PC score vectors. If the score vector t_{new} of the unknown sample is not in the range of $[t_{\max}, t_{\min}]$, the sample shall not be judged to belong to the k th class.

For discriminant analysis of one-class classification (normal or abnormal samples) [13] such as original identification of traditional Chinese medicine, food adulteration, drug authenticity, data-driven soft independent modeling of class analogy (DD-SIMCA) is commonly used [14–16]. As shown in Fig. 12.12, this method can provide the Chi-square acceptance area of normal samples. It can also provide the distribution area of extreme samples and abnormal samples through probabilistic statistical analysis [17, 18].

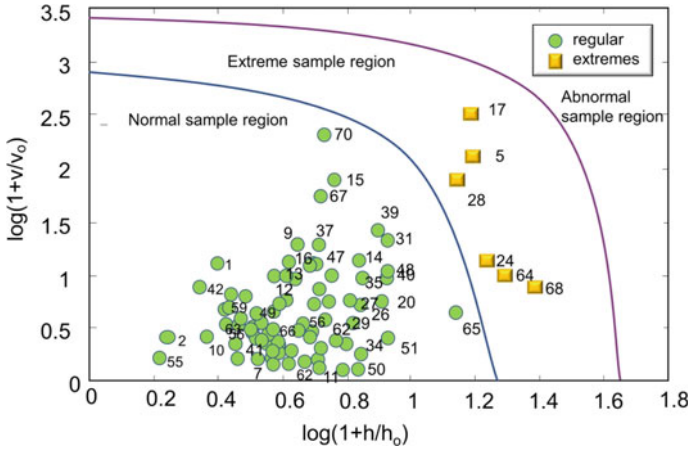


Fig. 12.12 Schematic diagram of chi-square acceptance area obtained by DD-SIMCA method

12.3.5 Logistic Regression

Although logistic regression (LR) is called the regression, it is actually a kind of classifier. It can handle with binary classification problems well [13]. Sigmoid function is used in the LR model of binary variables. It has good mathematical properties, which is a convex function. It can be differentiated in any order. The equation of sigmoid function is as follows.

$$g(z) = \frac{1}{1 + e^{-z}} \tag{12.44}$$

For a binary classification problem, it is specified that one class is positive and the other is negative. The corresponding class values y are 1 and 0, respectively. The spectral input vector is x and the dimension is $1 \times m$. The augmented vector is $x = [1 \ x]$ and its dimension is $1 \times (m + 1)$. Then the LR model is:

$$h_{\theta}(x) = P(y = 1|x; \theta) = \frac{1}{1 + e^{-(\theta^T x)}} \tag{12.45}$$

where $P(y = 1|x)$ represents the probability that x is a positive class ($y = 1$). θ is the model parameter and the dimension of θ is $1 \times (m + 1)$. The task of LR is to learn the θ value of the above model. Once θ value is determined, the prediction probability value is calculated for x of an unknown sample. If $h_{\theta}(x) > 0.5$, x is classified as positive class ($y = 1$). If $h_{\theta}(x) < 0.5$, x is classified as negative class ($y = 0$).

The learning algorithm of LR can adopt the maximum likelihood method and the probability function can be written as:

$$\begin{aligned} P(y = 1|x; \theta) &= h_{\theta}(x) \\ P(y = 0|x; \theta) &= 1 - h_{\theta}(x) \end{aligned} \quad (12.46)$$

The 0 or 1 can be taken as the value and the above equation can be written in the form of conditional probability distribution:

$$P(y|x; \theta) = h_{\theta}(x)^y (1 - h_{\theta}(x))^{(1-y)} \quad (12.47)$$

where the superscript i represents the number of the sample. Suppose there are n independent samples $(x^{(i)}, y^{(i)})$ in training set, the likelihood function of n samples is:

$$\begin{aligned} L(\theta) &= \prod_{i=1}^n P(y^{(i)}|x^{(i)}; \theta) \\ &= \prod_{i=1}^n h_{\theta}(x^{(i)})^{y^{(i)}} (1 - h_{\theta}(x^{(i)}))^{(1-y^{(i)})} \end{aligned} \quad (12.48)$$

To achieve the maximum value of the above equation $\theta = (\theta_0, \theta_1, \dots, \theta_{m+1})$, which is also the solution of model parameters. m is the number of spectral wavelength points.

For calculation convenience, logarithm of $L(\theta)$ is acquired. The maximum $L(\theta)$ is equivalent to maximizing the following logarithmic likelihood function:

$$\begin{aligned} l(\theta) &= \ln L(\theta) \\ &= \sum_{i=1}^n (y^{(i)} \ln h_{\theta}(x^{(i)}) + (1 - y^{(i)}) \ln(1 - h_{\theta}(x^{(i)}))) \end{aligned} \quad (12.49)$$

The solution of $\frac{dl(\theta)}{d\theta} = 0$ is the value of θ , which can be solved by gradient descent method. The steps of iteration calculation are as follows.

- (1) For known n independent training samples, the initial value of θ is assigned.
- (2) Calculate

$$h_{\theta}(x^{(i)}) = \frac{1}{1 + e^{-(\theta^T x^{(i)})}} \quad (12.50)$$

- (3) Update θ_j

$$\theta_j := \theta_j + \alpha (y^{(i)} - h_{\theta}(x^{(i)})) x_j^{(i)} \quad (12.51)$$

where $j = 0, 1, \dots, m$, α is the iteration step size.

- (4) Determine whether the convergence condition is met. If not, return to (2). If so, terminate the iteration. The convergence condition can be reached a certain number of iterations, or the difference of value is less than specified ε before and after the θ update.

12.3.6 Soft-Max Classifier

Soft-max classifier is a multi-classification method. It uses nonlinear functions to calculate the probability that the input variable X belongs to each class by comparing the probability value to determine the classification. It has strong classification ability for the class data of nonlinear structure. Soft-max classifier is a multi-class extension of LR. Sigmoid function is used LR. It is different from LR classification which only two-class labels can be taken. Soft-max classifier is suitable for multi-class problems. Soft-max classifier maps the input vector x from the n -dimensional space to the class, and the results are given in the form of probability. The equation is as follows:

$$p_j = \frac{e^{\theta_j^T x}}{\sum_{k=1}^K e^{\theta_k^T x}} \quad (12.52)$$

where $\theta_k^T = [\theta_k^1 \theta_k^2 \theta_k^3 \dots \theta_k^N]^T$ is weights, which is the corresponding classifier parameter of class k .

The total model parameter θ is trained by Soft-max classifier, which is used to calculate all possible class probabilities of the item to be classified. Then the class is determined. A data set is given that containing m training samples. $\{(\mathbf{x}^{(1)}, y^{(1)}), (\mathbf{x}^{(2)}, y^{(2)}), \dots, (\mathbf{x}^{(m)}, y^{(m)})\}$, x represents the input vector and y represents the class label of each x . For a test sample $\mathbf{x}^{(i)}$ is given, the probability of belonging to each class is calculated by Soft-max classifier. The function equation is as follows:

$$\mathbf{h}\theta(\mathbf{x}^{(i)}) = \begin{bmatrix} p(y^{(i)} = 1) | \mathbf{x}^{(i)}; \theta \\ p(y^{(i)} = 2) | \mathbf{x}^{(i)}; \theta \\ \vdots \\ p(y^{(i)} = K) | \mathbf{x}^{(i)}; \theta \end{bmatrix} = \frac{1}{\sum_{k=1}^K e^{\theta_k^T \mathbf{x}^{(i)}}} \begin{bmatrix} e^{\theta_1^T \mathbf{x}^{(i)}} \\ e^{\theta_2^T \mathbf{x}^{(i)}} \\ \vdots \\ e^{\theta_K^T \mathbf{x}^{(i)}} \end{bmatrix} \quad (12.53)$$

In Eq. 12.53, $\mathbf{h}\theta(\mathbf{x}^{(i)})$ is a vector. The element $p(y^{(i)} = k | \mathbf{x}^{(i)}; \theta)$ represents the probability that $\mathbf{x}^{(i)}$ belongs to class k . The sum of the elements in the vector is equal to 1. For $\mathbf{x}^{(i)}$, The maximum probability value of k is selected as the classification result.

The value of parameter θ can be obtained by minimizing the cost function, which is defined as:

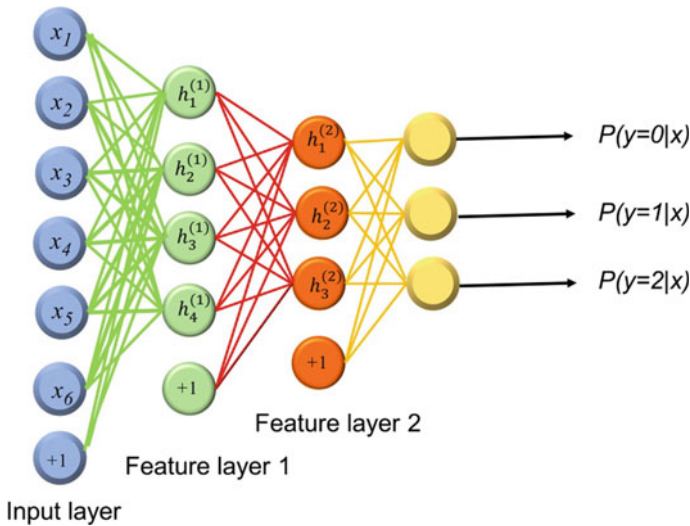


Fig. 12.13 Schematic diagram of Softmax classifier combined with deep learning algorithm for multi-class discriminant analysis

$$J(\theta) = -\frac{1}{m} \left[\sum_{i=1}^m \sum_{j=1}^K 1\{y^{(i)} = j\} \log \frac{e^{\theta_j^T x^{(i)}}}{\sum_{k=1}^K e^{\theta_k^T x^{(i)}}} \right] \quad (12.54)$$

where $1\{\cdot\}$ is an indicative function, with a value of true equal to 1 and a value of false equal to 0. Soft-max classifier can be regarded as a neural network without hidden layer. In the training process, the parameter θ can be adjusted continuously by gradient descent method, so that it can minimize the cost function $J(\theta)$ until it converges to the global optimal solution.

As shown in Fig. 12.13, Logistic and Soft-max classifiers are mostly combined with deep learning algorithms (such as auto-encoder network and convolutional neural network), which is used for discriminant analysis of two or more classes of problems, respectively.

Gan et al. [14] used stacked contractive auto-encoding to extract the NIR spectral features of drugs. Moreover, they used Logistic classifier and Soft-max classifier to carry out two-class and multi-class for drugs identification, respectively. The results are superior to BP network and SVM methods. Wang et al. [15] aimed at the near infrared hyperspectral imaging of *Lycium barbarum* from different habitats, zero-phase component analysis whitening preprocessing was used to remove the correlation of input features firstly, and then PLS-DA algorithm to extract the PC with the greatest correlation between input features and classes of PC to reduce the model complexity. Finally, Soft-max classifier is used to classify the input data from the perspective of probability, which can effectively identify the origin of *Lycium barbarum* in Ningxia. Liu et al. [16] compared the effects of logistic classifier, BP

network and K-means method on the classification of visible-near infrared spectra of soil, and the result of Logistic classifier was the best.

12.3.7 *Random Forest*

Random forest (RF) is a fusion classification algorithm that includes many decision trees and voting strategies. It belongs to the ensemble algorithm. As shown in Fig. 12.14, the essence of RF is to randomly select samples from the original training data set to form a new data set of the same size, and repeatedly replace the new data set with the samples from the original data set to continuously form a new data set of the same size. This process is called bootstrap aggregating. Several groups of new training data sets can be obtained through the bootstrap resampling process, which is classified by the decision trees algorithm, respectively. In this way, the number of new classifiers equal to the number of new data sets can be obtained. When spanning the tree, the variables for each node are generated only from a few randomly selected variables. Namely, the use of variables and samples is both randomized, and a large number of trees generated in this random way are used for classification or regression analysis, so it is called RF. The samples that out-of-bag samples (OOB) are not selected, which are used as validation set to test each tree model to get the OOB error rate, it is used to optimize the model parameters and evaluate the quality of the model.

When the unknown samples are classified and predicted, the RF obtains the multiple groups classifiers in the training process to make the prediction, respectively, and selects the class with the most votes from the classifier as the final result. Since RF combines the results of multiple binary decision trees, the number of decision trees, and the number of features used in each decision tree, they are important parameters that affect the output of RF. The number of decision trees refers to the total number of decision trees in the RF. The performance of the model can be improved by increasing the number of decision trees. However, at the cost of computation, it is necessary to consider the calculation efficiency to determine the most appropriate number of decision trees. The number of features refers to the maximum number of features used in each decision tree, which can select any integer value between 0 and all feature numbers. In general, it is most appropriate for the number of features of each decision tree to take the arithmetic square root of the total number of features. A large number of theoretical and experimental studies have proved that RF has high prediction accuracy with good tolerance for outliers and noise. Moreover, it is not easy to fall in overfitting problem.

RF can also measure the importance of features. Its basic principle is that if a certain feature variable is very important, the prediction result of the sample leads to large deviation if the feature variable of the sample is changed, that is, the feature variable is very sensitive to the prediction result. On the contrary, if a feature variable is not important, arbitrarily changing does not have much effect on the prediction results. Therefore, RF process can be used to select feature variables [17, 18].

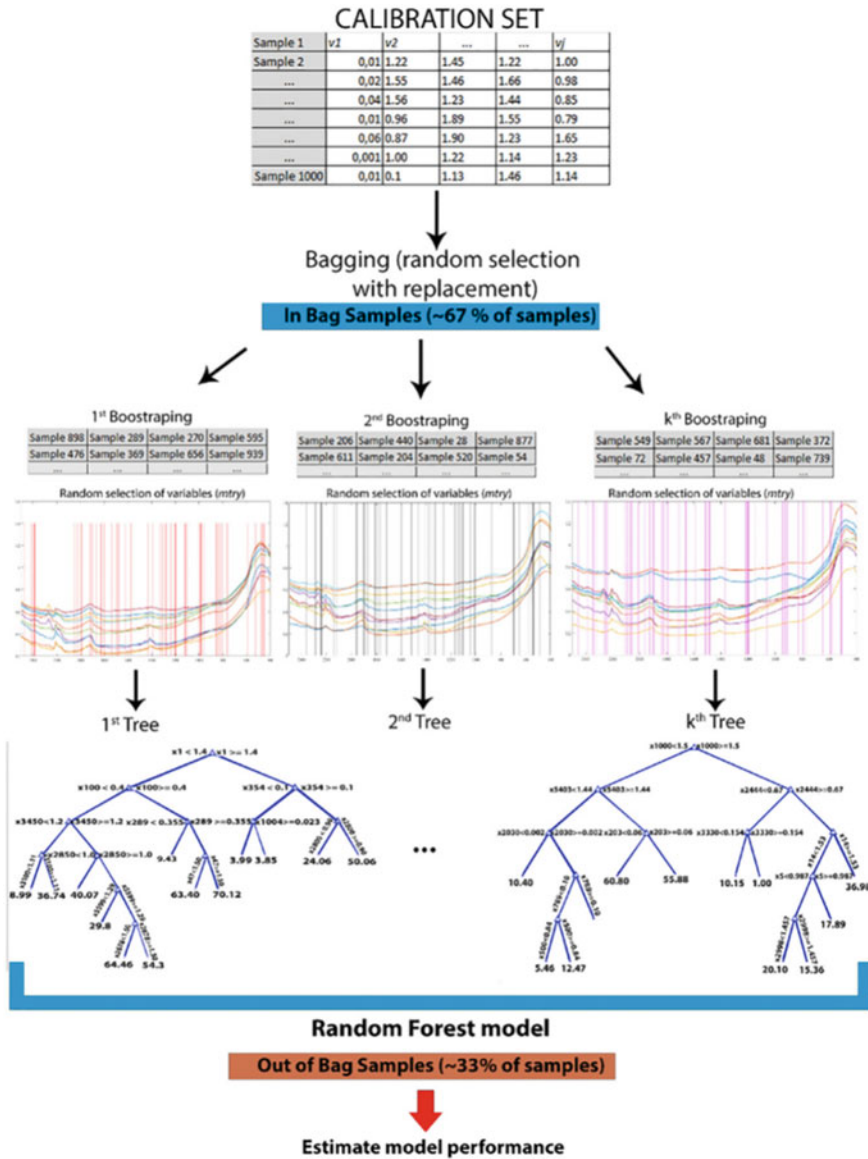


Fig. 12.14 Schematic diagram of modeling framework for random forest

The feature importance is usually measured by the prediction accuracy of OOB. OOB refers to the training samples that are not sampled during the training of each decision tree, and they do not participate in the establishment of the decision tree. Therefore, they can be used to evaluate the performance of the decision tree. The basic principle is to measure the importance of the features by rearranging the value

of features in OOB (i.e. the value of the feature exchanged between the OOB data) and uses the difference in the prediction accuracy of the OOB before and after rearrangement to measure the feature importance. The specific steps for calculating the importance of some features areas follows. Firstly, the OOB prediction is obtained for each decision tree. Then all values of the features on all OOB are sorted, and the OOB are predicted again. Finally, the importance values of the features are obtained by averaging the prediction accuracy difference before and after sorting in each decision tree.

RF is a natural nonlinear modeling tool that can be used for classification or regression analysis. Lai et al. adopted wavelet transform and RF to establish the recognition model of NIR spectroscopy for different mildewed tobacco leaves and achieved satisfactory results in the discrimination of the mildew degree of tobacco samples [19]. Li et al. used RF to identify the Raman spectra of precursor chemicals and flammable and explosive chemicals. The result of RF was equivalent to adaboost algorithm and superior to decision tree, SVM, and ANN algorithm [20]. Wang et al. used terahertz time domain spectroscopy (THz-TDS) and RF to classify and recognize five species of rosewood, and the classification accuracy reached more than 95% [21]. Zhou et al. fused mid-infrared spectroscopy with NIR spectroscopy and used RF method to identify *Panax notoginseng* in five areas, and the recognition accuracy was 95.6% [22]. Amjad et al. used Raman spectroscopy and RF to conduct discriminant analysis on four kinds of milk powder, and the average accuracy was about 94% [23].

Ma et al. used RF to construct a Vis–NIR spectra model for estimating soil salt, which could effectively extract the main ion information of soil salt in arid areas [24]. Li et al. used RF to establish a hyperspectral model for estimating soil organic matter content, and the results were superior than PLS method [25]. Zheng et al. classified the price grades of dendrobium based on laser-induced breakdown spectroscopy (LIBS) and RF and realized the rapid identification of the grades of dendrobium [26]. Li et al. used wavelet transform-random forest (WT-RF) to establish a model for predicting methanol content in gasoline by NIR spectroscopy, and the result was superior than WT-PLS and WT-LSSVM [27]. Santana et al. used Vis–NIR spectra combined with RF method to rapidly predict and analyze soil quality parameters and achieved superior prediction accuracy than PLS [28]. Teixeira et al. used portable X-ray fluorescence spectroscopy (XRF) combined with RF method to predict and analyze soil pH, base saturation percentage, cation exchange capacity, and aluminum saturation, etc., and then obtained satisfactory results [29]. Zhang et al. used NIR combined with RF to predict the content of food dye indigotine in cream, the results are superior than MLR and PLS methods [30].

12.3.8 Application of Regression Methods for Discriminant Analysis

ANN method has been introduced in detail in quantitative calibration method (Chap. 8) and clustering analysis above, and it can also be used in supervised pattern recognition, which classifies and predicts unknown samples by establishing recognition model based on the training set of known classes. The only difference with quantitative calibration is the difference in the output layer. For quantitative calibration, the output layer is usually a single node. For pattern recognition, multi-node output is generally used. If there are four classes, they can be represented by (1,0,0,0), (0,1,0,0), (0,0,1,0) and (0,0,0,1), respectively.

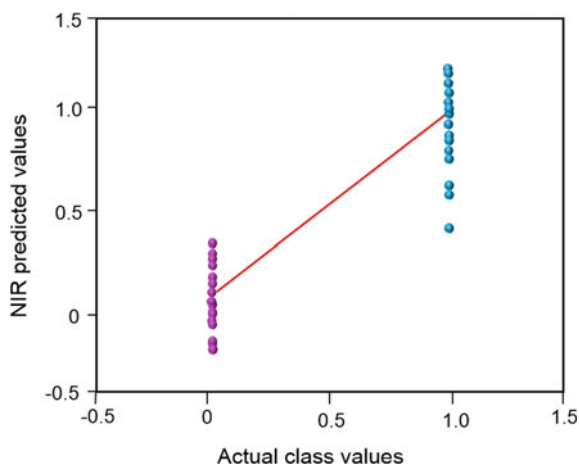
Similarly, PLS also can be used for discriminant analysis. PLS method is essential a regression method based on feature variables. However, if the concentration matrix of samples of known classes is set as 0, 1 (PLS1 method is used for two classes), -1, 0, +1 (PLS1 method is used for three classes), respectively. Or 0 1, 1 0 (PLS2 method is used for two classes), 0 0 1, 0 1 0, 0 0 0 (PLS2 method is used for three classes, as shown in Fig. 12.15, PLS method can be used for supervised discriminant analysis. It is often called Dummy partial least squares regression (D-PLS) or PLS discriminant analysis (PLS-DA).

Figure 12.16 shows the result regression of PLS for two-class samples. If the predicted PLS value of an unknown sample is between -0.5 and 0.5, it belongs to the first class, if it is between 0.5 and 1.5, it belongs to the second class. Similar to the quantitative calibration, since PLS method can decompose both spectral matrix and classes matrix at the same time, it strengthens the role of class information in spectral decomposition, so as to extract the most relevant spectral information to



Fig. 12.15 Schematic diagram of regression method used for discriminant analysis of three classes of samples

Fig. 12.16 Schematic diagram of the results for classification of two-class samples by using PLS-DA method



sample class, namely maximizing the different among different classes. Therefore, PLS method can usually get better classification and discrimination results than PCA method. At present, the application of PLS in pattern recognition receives more and more attention and applications. Many studies also have proved that the discriminant results of PLS are superior to the pattern recognition method based on PCA.

12.4 Spectral Searching Methods

12.4.1 Introduction

In recent years, with the continuous improvement of instrument manufacturing level and the popularization of chemometrics method, modern spectral analysis technologies, especially MIR, NIR, and Raman spectroscopy have been widely used in qualitative analysis in many fields because of its convenient test, fast speed, abundant information, and on-site application. Using pattern recognition methods, spectra can be used to cluster or identify samples of complex systems (such as oil, grain, fruit, and drugs). In chemometrics, as shown in Fig. 12.17, pattern recognition methods for spectral analysis include three classes [31–33]:

- (1) Unsupervised methods, such as PCA, hierarchical clustering method, K-means clustering, and self-organizing neural networks.
- (2) Supervised methods, such as LDA, SIMCA, KNN, PLS-DA, and SVM. Both of the above two methods are based on the class of samples for qualitative analysis, and each class must contain many representative samples. When new samples are added to the database, the recognition model needs to be calibrated again.

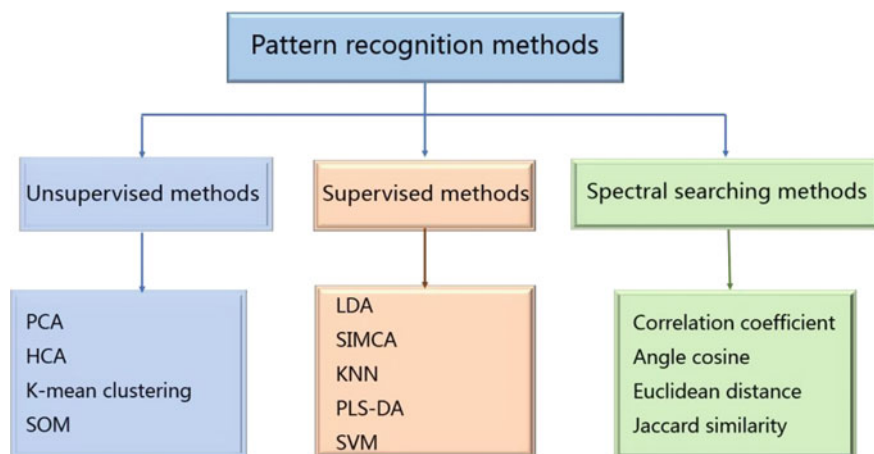


Fig. 12.17 Schematic diagram of classification for pattern recognition methods

- (3) Spectral searching methods, such as correlation coefficient, angle cosine, Euclidean distance, and spectral information divergence. According to the spectrum of the sample to be tested, this kind of algorithm retrieves one or more samples closest to the sample to be tested from the established spectral library, so as to achieve qualitative or even quantitative analysis. In fact, the spectral searching method can also be regarded as KNN method. The basic idea of the two methods is completely consistent, but the final display results are different.

Previously, spectral searching methods are mostly used for spectral recognition of pure compounds, such as infrared spectroscopy databases of Sadtler and Aldrich. In the past two decades, spectral databases of modern complex mixed systems are gradually established in many fields (such as soil, feed, minerals, drugs, and oils) [34–36]. For example, the drug product administration (DPA) of the food and drug administration (FDA) is developing the spectral databases of pharmaceutical excipients based on different Raman and NIR spectroscopy instruments such as laboratories, portable, and hand-held, so as to monitor possible problems such as contamination, adulteration, and tampering in pharmaceutical production and supply chain. Relevant departments in different fields in China are also gradually establishing and improving the corresponding spectral database. For another example, the Vis–NIR spectral database of soil has been established internationally, as shown in Fig. 12.18. More than 20,000 samples have been collected from more than 10,000 sites around the world. The database can be used for remote sensing and rapid analysis of soil physical properties on site (Fig. 12.19) [37].

The spectral searching method is one of the core technologies to make full use of these spectral databases. Therefore, spectral searching algorithms attracted more and more attention [38–40]. Some new searching algorithms and searching strategies have emerged, and the accuracy and reliability of spectral searching are significantly

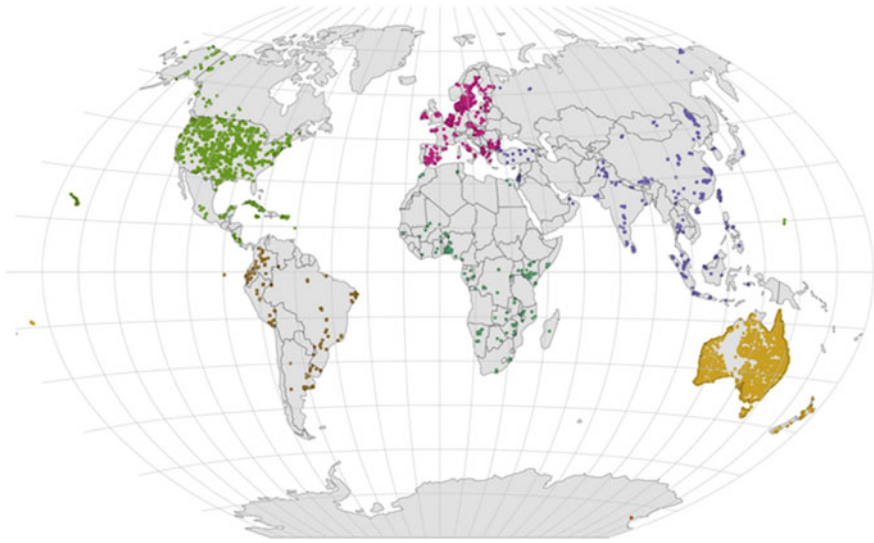


Fig. 12.18 Schematic diagram of more than 10,000 sampling points for building a global near infrared spectroscopic database of soil

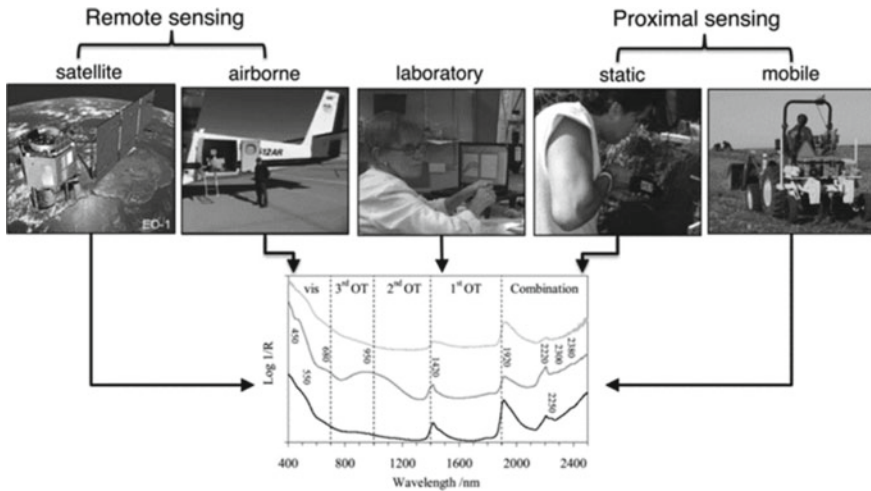


Fig. 12.19 Schematic diagram of variety acquisition ways for Vis-NIR spectra of soils (remote sensing and short-distance remote sensing)

improved. Compared with unsupervised and supervised pattern recognition methods, spectral searching method has many advantages such as simple operation, intuitive information, and convenient library maintenance, which plays an important role in practical applications.

12.4.2 Spectral Searching Algorithms

The goal of spectral searching is to find one or more samples closest to \mathbf{x} from the spectral library \mathbf{R} based on certain algorithms and rules for the spectrum \mathbf{x} of the sample to be measured. If there is a known property value \mathbf{Y} in the spectral library, the property value of the sample to be measured can be quantitatively predicted and analyzed. Where \mathbf{x} represents the spectrum of the sample to be tested, which is $l \times m$ vector, and m is the number of wavelength points. \mathbf{R} represents all spectra in the spectral library, which is $n \times m$ matrix, n is the number of samples spectra in the library. \mathbf{r}_j represents the spectrum of the j th sample in the spectral library, which is $l \times m$ vector, $j = 1, 2, \dots, n$. \mathbf{Y} represents the property value corresponding to all samples in the spectral library, which is the $n \times p$ matrix, and p is the number of properties. \mathbf{y}_j represents the property value of the j th sample in the spectral library, which is $l \times p$ vector.

In order to obtain satisfactory searching results, it is often necessary to carry out the necessary preprocessing and band selection of the spectrum before searching. The preprocessing methods include derivative, vector normalization, standardization, and wavelet transform. The band selection can find one or more spectral intervals with strong features, high signal-to-noise ratio, and small external influence according to chemical knowledge and mathematical methods. The commonly used spectral preprocessing and band selection methods can be referred to Chaps. 4 and 5, and other relevant literature [41].

(1) Distance-based algorithm

The basic principle of this algorithm is that the closer the spectral of the two samples are, the shorter the distance between them is [42]. There are many forms of distance between spectra, in which the simplest is absolute distance (L_1 norm). The absolute distance between the spectrum \mathbf{x} of the sample to be measured and the spectrum \mathbf{r}_i of the i th sample in the spectral library can be expressed as,

$$d(\mathbf{x}, \mathbf{r}_j) = \sum |\mathbf{x}, \mathbf{r}_j| \quad (12.55)$$

The most commonly used one is Euclidean distance, also called as least squares distance,

$$d(\mathbf{x}, \mathbf{r}_j) = \sqrt{(\mathbf{x} - \mathbf{r}_j)(\mathbf{x} - \mathbf{r}_j)^t} \quad (12.56)$$

Many calculation methods for distance are based on L_1 norm and L_2 norm and some weighted distance calculation methods are derived. Normalized Euclidean distance (NED) is often used, that is, the spectrum is normalized before calculating the spectral distance, and NED usually has stronger robustness.

(2) Similarity-based algorithm

There are two main parameters to evaluate similarity: angle cosine and correlation coefficient.

The cosine of the angle between x and r_j is expressed as

$$\cos(\mathbf{x}, \mathbf{r}_j) = \frac{\mathbf{x} \mathbf{r}_j^t}{\sqrt{\mathbf{x} \mathbf{x}^t} \sqrt{\mathbf{r}_j \mathbf{r}_j^t}} \quad (12.57)$$

The smaller the angle is, the closer the two samples are in the model space, and the greater the similarity is. If the two spectra are exactly the same, then $\cos(\mathbf{x}, \mathbf{r}_j) = 1$, which means the two samples are a point in the mode space. If the two spectra are completely different, $\cos(\mathbf{x}, \mathbf{r}_j) = 0$.

Sometimes, angle cosine is replaced by the spectral angle,

$$S(\mathbf{x}, \mathbf{r}_j) = \arccos\left(\frac{\mathbf{x} \mathbf{r}_j^t}{\sqrt{\mathbf{x} \mathbf{x}^t} \sqrt{\mathbf{r}_j \mathbf{r}_j^t}}\right) \quad (12.58)$$

The closer the angle $S(\mathbf{x}, \mathbf{r}_j)$ is to 0, the more similar the two spectra are. $S(\mathbf{x}, \mathbf{r}_j)$ is also called spectral angle metric (SAM), and one of the features of SAM method is the invariance of multiplicative factor. Since the amplitude and shape of the spectral curve correspond to the length and direction of the vector in Euclidean space, the multiplicative factor only causes the change of the length of the vector and does not change the direction of the vector. Therefore, SAM method is sensitive to spectral shape difference. However, it is not sensitive to spectral amplitude difference.

The correlation coefficient between x and r_j is expressed as the following equation.

$$R(\mathbf{x}, \mathbf{r}_j) = \frac{(\mathbf{x} - \bar{\mathbf{x}})(\mathbf{r}_j - \bar{\mathbf{r}}_j)^t}{\sqrt{(\mathbf{x} - \bar{\mathbf{x}})(\mathbf{x} - \bar{\mathbf{x}})^t} \sqrt{(\mathbf{r}_j - \bar{\mathbf{r}}_j)(\mathbf{r}_j - \bar{\mathbf{r}}_j)^t}} \quad (12.59)$$

where $\bar{\mathbf{x}}$ and $\bar{\mathbf{r}}_j$ are the average values of \mathbf{x} and \mathbf{r}_j , respectively. The closer R is to 1, the more similar the two spectra are, and the closer R is to 0, the greater the difference between the two spectra is.

Hit quality index (HQI) is also commonly used in spectral searching:

$$\text{HQI} = 1 - R(\mathbf{x}, \mathbf{r}_j)^2 \quad (12.60)$$

(3) Algorithm based on information theory

Spectral information divergence (SID) [43, 44] transforms the spectral similarity evaluation problem into the redundancy evaluation problem for probabilities of the two spectral vectors. In addition, the relative entropy of spectral information is used to evaluate the similarity of the two spectra.

$$\text{SID}(\mathbf{x}, \mathbf{r}_i) = D(\mathbf{x} \parallel \mathbf{r}_j) + D(\mathbf{r}_j \parallel \mathbf{x}) \quad (12.61)$$

where $D(\mathbf{x} \parallel \mathbf{r}_j)$ is the relative entropy of \mathbf{r}_j with respect to \mathbf{x} , $D(\mathbf{r}_j \parallel \mathbf{x})$ is the relative entropy of \mathbf{x} with respect to \mathbf{r}_j .

$$D(\mathbf{x} \parallel \mathbf{r}_j) = \sum_{i=1}^m q_i \log \left(\frac{q_i}{p_{j,i}} \right) \quad (12.62)$$

$$D(\mathbf{r}_j \parallel \mathbf{x}) = \sum_{i=1}^m p_{j,i} \log \left(\frac{p_{j,i}}{q_i} \right) \quad (12.63)$$

where \mathbf{q} and \mathbf{p}_j are the probability vectors of spectral \mathbf{x} and spectral \mathbf{r}_j , respectively, $\mathbf{q} = \frac{\mathbf{x}}{\sum_{i=1}^m x_i}$, $\mathbf{p}_j = \frac{\mathbf{r}_j}{\sum_{i=1}^m r_{j,i}}$.

SID method determines the similarity between the two spectra by measuring the mutual information between the two spectra. The smaller the SID value is, the higher the similarity between spectra is. On the contrary, the spectral similarity is low.

(4) Degree similarity algorithm

The degree similarity algorithm is evolved and simplified based on similar of system theory [45, 46], and the degree similarity Q can reflect the average relative difference between the two spectra,

$$Q = 1 - \frac{1}{m} \sum_{i=1}^m \left(1 - \frac{\min(x_i, r_{j,i})}{\max(x_i, r_{j,i})} \right) \quad (12.64)$$

The algorithm compares the spectral intensity at each wavelength point one by one, so it is sensitive to relative difference of wavelength position. The closer Q is to 1, the more similar the two spectra \mathbf{x} and \mathbf{r}_j are, and the closer Q is to 0, the greater the difference between the two spectra are.

(5) Jaccard similarity algorithm

Jaccard similarity is a spectral matching algorithm based on characteristic peaks. It is necessary to binarize the spectrum [47, 48] and calculate the proportion of intersection of two spectra characteristic peaks in the union set:

$$J = \frac{\mathbf{x} \cap \mathbf{r}_j}{\mathbf{x} \cup \mathbf{r}_j} = \frac{p}{p + q + r} \quad (12.65)$$

where p is the number of corresponding wavelength points that are 1 after the binarization of \mathbf{x} and \mathbf{r}_j . The wavelength point after \mathbf{q} is \mathbf{x} binarization is 1, and the number of wavelength point after \mathbf{r}_j binarization is 0. r is the number of wavelength points 0 after \mathbf{x} binarization and 1 after \mathbf{r}_j binarization.

12.4.3 Improvements of Spectral Searching Algorithms

On the basis of absolute distance (spectral difference), Meng et al. established a method to calculate the ultraviolet spectrum similarity S by taking the effect of absorption intensity on the difference into account.

$$S = 1 - \frac{1}{m} \sum_{i=1}^m \left| \frac{x_i - r_{j,i}}{x_i + r_{j,i}} \right| \quad (12.66)$$

where m is the number of wavelength points of the spectra, and the closer the S value is to 1, the more similar the two spectra are. The closer to 0, the more different the two spectra are. Compared with the angle cosine and correlation coefficient method, this method is sensitive to the difference between spectra. It overcomes the disadvantage of ultraviolet spectrum of broadband absorption to a certain degree. In addition, it can quickly and sensitively reflect the similarities and differences in the quality of traditional Chinese medicine, so as to quickly monitor the differences in the components in the production process of traditional Chinese medicine injection [49]. Li and Tang et al. improved the calculation method of similarity S by increasing the weight, and the sensitivity of similarity S was further improved by highlighting the spectral changes in the key wavelength range. They were used to investigate the stability of Danshen injection by ultraviolet spectroscopy and identify NIR abnormal spectroscopy, respectively [50, 51]. Khan et al. improved the Euclidean distance by weighting rules [52] in order to search the spectra of mixture by using the Raman spectral library of pure compounds, and the recognition results are superior to the traditional Euclidean distance and angle cosine.

Considering the repeatability of the samples in the spectra collection process, Plugge et al. proposed the conformity index (CI) method based on absolute distance. The essence of this method is a weighted absolute distance method. The library spectrum r_j is replaced by the average spectrum \bar{r}_j of a set of repetitive spectra. The weight of each wavelength point is the reciprocal of the standard deviation σ_j of the repetitive spectrum.

$$CI = \text{MAX} \left(\frac{x_i - \bar{r}_{j,i}}{\sigma_{j,i}} \right) \quad (12.67)$$

CI actually refers to the allowable spectral repeatability (or reproducibility) range, which is usually 3–5 times of the standard deviation [53]. Plugge et al. used it to detect the changes in physicochemical properties of ampicillin trihydrate, which could be used to control the production process and ensure the consistency of product quality. Ritchie et al. investigated the accuracy, precision, robustness, and consistency of this method, and the results showed that this method could meet the current verification standard and could be accepted by modern strict guiding principles [54]. Feng et al. used the CI method to quickly determine the authenticity of drug quality [55] by NIR

spectroscopy and established a NIR library for the CI of hundreds of drugs [56, 57], which are widely used in China.

For repetitive measurements or library spectra of a class of multiple samples, Thermo Company used PCA decomposition to calculate the spectral difference e between the spectra of the sample to be measured and the library spectra. Similarity match value (SMV) is defined by the improved Euclidean distance.

$$\text{SMV} = \left(1 - \frac{\|e\|}{\|x\|}\right) \times 100 \quad (12.68)$$

For the NIR spectra of complex mixtures, if the main components of the mixture are the same, it is difficult to identify the differences between samples by traditional spectral searching. Nie et al. used SMV method and NIR spectroscopy to quickly and conveniently identify the differences between Tongren Wuji Baifeng pills and other Wuji Baifeng pills [58]. Tao et al. used the SMV method to detect the stability of the quality of cut tobacco by taking the NIR diffuse reflectance spectrum as the feature [59], which can be used for rapid detection of tobacco filament in production, providing a new technical means for the quality control of cigarette processing. Lu et al. adopted the attenuated total reflection (ATR) measurement method to establish an infrared spectral library of textile fibers based on more than 1000 samples and realized the rapid detection of fiber types by using the spectral library searching method [60]. Wang et al. established an infrared spectral library of plastic resins containing 513 samples for 18 common classes of plastic resins, which can be used to quickly identify the classes of plastics [61].

Correlation coefficient and angle cosine are the most commonly used methods for database searching [62, 63]. For example, Chen et al. obtained 940 paint infrared spectra from 287 car body paint samples. The infrared spectral comparison database of car body paint was established by using the feature peak number method and the correlation coefficient method. The rapid investigation of vehicle paint left in the accident scene, the vehicle type of hit and run vehicles can be determined. [64]. He et al. established a terahertz spectrum database containing of 38 drugs with purity above 90%, which was expected to be a powerful supplement to the existing drug detection methods [65]. Guedes et al. established a database for identifying airborne pollen species by micro-Raman spectroscopy using correlation coefficient method [66]. In addition, in the ground object recognition of hyperspectral remote sensing image database, correlation coefficient and angle cosine are also the two most commonly used methods [67].

The correlation coefficient and angle cosine emphasize the overall similarity between the spectra. In order to improve the expression of the differences in the details of the spectra between the two algorithms, many methods have been tried for applications. It is a common strategy to build a secondary searching library (sub-library). Blanco et al. proposed to improve the accuracy of traditional correlation coefficient searching by establishing a sub-spectral library for the recognition of NIR of pharmaceutical raw materials [68]. Selecting feature intervals to calculate correlation coefficient is also an effective method to highlight the difference between

spectra. Wang et al. used the correlation coefficient method combined with characteristic spectral band for rapid identification of illegal addition of sildenafil citrate into capsules of traditional Chinese medicine. The overall accuracy rate of screening results was about 95.0% [69, 70]. Xu et al. [71] divided the whole spectral range into several regions and calculated the correlation coefficient or angle cosine of each region, respectively. This piecewise correlation coefficient method (called matrix correlation coefficient) improved the difference between spectra to some extent. The results showed that it was superior to the traditional correlation coefficient method. Park et al. also divided the Raman spectrum region into four sections according to the Hann window method and calculated the angle cosine value of each section. The final similarity was obtained by the weighted angle cosine of these four sections [72]. Griffiths et al. proposed a spectral matching algorithm based on self-weighted correlation coefficient, which can effectively overcome the interference signal in IR [73], and obtain good recognition results in the application of IR in open-path monitoring of atmospheric pollutants.

Chu et al. [74] proposed a method of moving window correlation coefficient based on the concept of moving window. The basic idea is to select a spectral window with width of w (w is odd) and move backward from the second wavelength point of the whole spectrum, each time moving a wavelength sampling interval until the final wavelengths. In each sub-wavelength region of the window, the correlation coefficient value is calculated by the traditional correlation coefficient equation, and then the moving window correlation coefficient value is plotted with the starting position of the corresponding window. This moving correlation coefficient method can distinguish the subtle differences between the two spectra and improve the accuracy of spectral recognition, and facilitate the extraction of hidden information. There are two threshold parameters in the moving window correlation coefficient method, one is the correlation coefficient of all windows, and the other is the sum of the correlation coefficients of all windows. According to different application objects, different window width and threshold should be set. For the spectra of two samples to be identified, only if the two parameters are greater than the corresponding threshold can be determined as the same sample.

Chu and Li et al. used the correlation coefficient of moving window to establish the NIR spectrum recognition library and the mid-infrared spectrum recognition library for crude oil, respectively [75, 76], which can accurately identify the classes of crude oil. On this basis, Li et al. also used the moving window correlation coefficient to identify the two-dimensional infrared spectra of crude oil and proposed the moving matrix window correlation coefficient method (Fig. 12.20), which can accurately identify the low proportion of mixed crude oil [77]. Guo et al. used the moving window correlation coefficient method to determine the end point of the extraction process of traditional Chinese medicine by NIR. Compared with the original moving standard deviation method, this method can greatly reduce the influence of baseline drift to a large degree and has better anti-interference ability [78]. Ramirez-Lopez et al. improved the traditional spectral difference and proposed a surface difference spectrum (SDS) based on the different order differential spectra of the difference spectrum. The SDS distance between the two spectra was calculated by the weight of

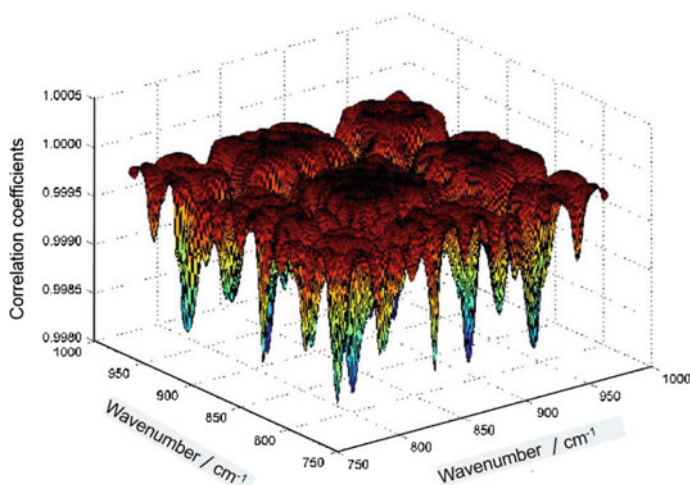


Fig. 12.20 Schematic diagram of correlation coefficients of two-dimensional correlation spectra moving window for two crude oils

the moving window correlation coefficient, which was used for the sample searching of the global soil Vis–NIR spectroscopy library [79].

Liu et al. introduced nonlinear kernel function into spectral angle mapper (SAM), which is called kernel spectral angle mapper (KSAM). The purpose of KSAM is to evaluate the similarity between spectra by using the high-order statistical characteristics of vectors [80, 81]. The KSAM method not only maintains the characteristics of the spectral vector in the original feature space but also extracts the nonlinear features between spectra. In addition, the KSAM method can set a variety of kernel functions and has high adaptability.

Cross-correlogram spectral matching (CCSM) was proposed by Meer et al. This method introduces the concept of relative sliding between spectra. The wavelength points of the spectra to be measured move one wavelength point to the left and one wavelength point to the right at a time, respectively. The correlation coefficient between the overlap regions of the library spectrum and the spectrum to be measured is calculated, and the cross-correlation characteristic curve between the matching position and the correlation coefficient is obtained (Fig. 12.21) [82]. By calculating the skewness of the cross-correlation characteristic curve, the similarity between the two spectra can be evaluated. The smaller the skewness is, the higher the similarity is. This method is not sensitive to spectral amplitude variation and has good anti-noise performance.

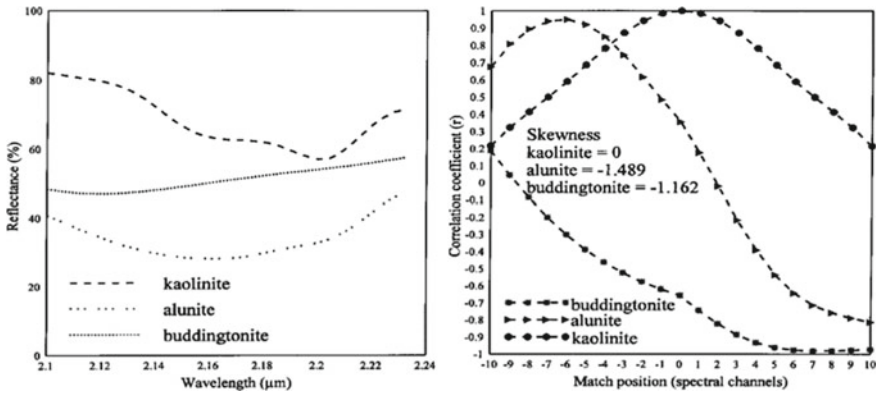


Fig. 12.21 Library and test spectrum (left) and corresponding cross correlogram (right) for kaolinite versus kaolinite, alunite, and buddingtonite [82]

12.4.4 Spectral Searching Strategies and Applications

In order to obtain accurate and fast spectra searching results, many new algorithms and searching strategies are proposed based on the above basic searching algorithms and improved algorithms for specific application objects.

Library searching method may sometimes lead to non-robust recognition results. The ensemble or consensus strategy is an effective means to solve this problem. The basic idea is to use a variety of searching algorithms to establish recognition rules, and at the same time to conduct discriminant analysis of the spectrum of the sample to be tested. The final hit rate or weighted value is taken as the recognition result. This searching strategy reduces the dependence of searching results on a certain algorithm, which can improve the stability of searching results. Himmelsbach et al. established an ATR mid-infrared spectroscopy library for recognition of foreign materials in cotton, which was used for the rapid recognition and analyzed of cotton pollutants. The library contains 601 samples of spectra, involving plant impurities (such as leaves, stems, shells, skins), compounds (plastic bags, films, and rubber), organic substance (other fibers, yarns, paper, feathers, cowhide, etc.), and inorganic substance (sand and rust) [83]. When the spectral library was used for the spectra of cotton in different regions, different picking periods, different spectrometers, or measurement accessories, the recognition accuracy was significantly reduced. Loudermilk et al. adopted the consensus or ensemble strategy to integrate the results of six common library spectral searching methods and achieved very satisfactory results [84].

Kong et al. integrated three algorithms of Euclidean clustering, correlation coefficient, and spectral information divergence and carried out experiments about airborne hyperspectral remote sensing images obtained by USGS mineral spectral library and operational modular imaging spectrometer system (OMIS) in China. The results

showed that it has the stronger spectral discrimination ability and the smaller spectral recognition uncertainty [85]. Zhao et al. fused the spectral information divergence with angle cosine to identify the classes of oil spills on the sea surface (light oil, medium oil, lubricating oil, and other oils) by airborne laser fluorescence radar, while heavy fuel oil and crude oil need a secondary recognition library for identification [85]. Feng et al. combined similarity algorithm, angle cosine algorithm, and correlation coefficient algorithm to identify pathogenic bacteria in water by ultraviolet–visible spectroscopy, which effectively improved the reliability and stability of the recognition results [86].

For the spectra of mixtures, how to use the spectra of pure compounds to resolve the qualitative and quantitative composition information of mixtures has attracted much attention. The commonly used methods include spectral peak fitting and non-negative least squares fitting, etc. [87]. Liu et al. proposed a spectral integration matching method by using the mathematical model of logistic regression to fuse the spectral peak matching coefficient, non-negative least squares matching coefficient, and angle cosine matching coefficient, which have a lower-misjudgment rate [88, 89].

Local calibration strategy, which combines spectral searching with multivariate calibration method, has been widely concerned and applied in recent years, especially with the expansion of large-scale near infrared spectral databases, such as soil, feed, and oil. The nonlinear relationship between spectra and concentration is aggravated by the sharp increase of samples from different sources, different years, and different classes. Local calibration strategy solves this problem by selecting the most similar set of samples from the spectral database to form the training set [90]. Aiming at how to select local samples and how to get the final prediction results, a variety of local modeling analysis strategies are proposed, such as CARNAC (comparative analysis using restructured near infrared and constituent data) method, LWR (locally weighted regression) method, LOCAL method [91], etc. The NIR spectroscopy libraries based on more than 3000 red grape samples by Dambergs et al., NIR libraries based on more than 20,000 feed samples by Fernandez-Ahumada, and Vis–NIR spectroscopy libraries based on more than 1000 soil samples by Genot, have been used to build the local calibrations for predicting the key physical, physicochemical properties, respectively, and more accurate results were obtained than traditional calibration methods [92–94]. Li et al. also selected representative samples from large gasoline NIR database based on spectral automatic searching algorithm to establish local models for different refineries, which improved the prediction accuracy of the model [95]. This modeling strategy is not only suitable for the calibration of nonlinear systems but also can make full use of the advantages of spectral database to avoid the disadvantages of traditional multivariate calibration methods that need to update the model frequently due to change in sample composition, etc. It is especially suitable for the qualitative and quantitative analysis of large network spectral database. In addition, Lee et al. applied the spectral searching algorithm to the discriminant analysis of classes (as shown in Fig. 12.22). The basic idea is similar to KNN method but used the HQI combined with voting strategy to discriminate classes [96].

It is also a common spectral searching strategy to reduce the dimension of spectrum and search in low-dimensional and more characteristic space. PCA, isometric

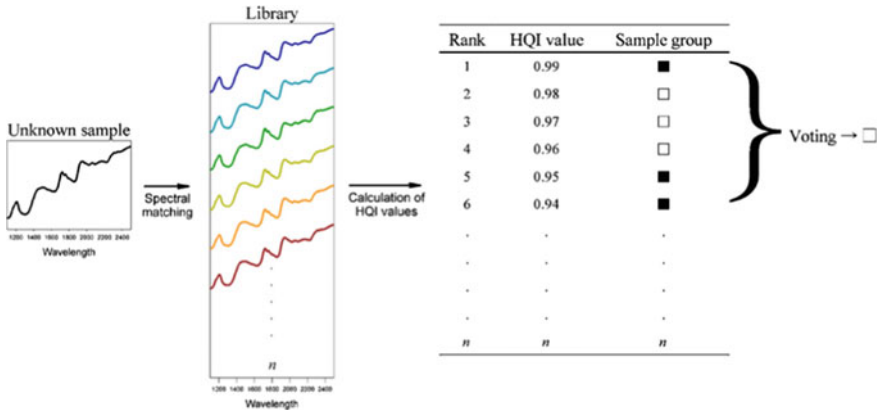


Fig. 12.22 Schematic diagram of class discrimination method based on spectral searching algorithm

mapping, local linear embedding (LLE), local preserving projection (LPP), neighborhood preserving projection (NPP), and other dimensionality reduction methods are mostly used in supervised and unsupervised pattern recognition, Fractal theory has attracted more attention in spectral searching [97–99]. For example, Lei et al. used the method of wavelet transform combined with fractal theory in NIR spectral recognition of lubricating oil [100]. In this method, the spectrum was transformed by wavelet transform firstly, and then the fractal dimension of wavelet approximation and detail spectrum was calculated. The fractal dimension was used as the feature for searching the spectrum, and good results were obtained. Xu et al. proposed a NIR spectral similarity measurement method based on grid division local linear embedding algorithm [101]. In this method, the high-dimensional spectral data are divided into multiple grid subspace, and the improved LLE algorithm is used to realize the feature mapping of each subspace from high-dimensional space to low-dimensional space, and the similarity matrix of the generated subspace is calculated. Finally, the subspace similarity matrix is normalized, and the similarity matrix of the accumulated and generated spectral sample set is solved to realize the spectral similarity measurement.

The combination of pattern recognition method and spectral searching algorithm can improve the speed and accuracy of spectral searching. The international forensic vehicle paint data query (PDQ) system is developed by the Royal Canadian mounted police forensic laboratory based on the original chemical and color information for searching information database. It can compare the paint data of crime scenes or suspect vehicles with the known paint samples in the database, and search for vehicles with similar information, thus quickly narrow the scope of investigation. The database contains more than 20,000 vehicle information, involving thousands of automobile manufacturers, with more than 80,000 painting layer information, and each year more than 500 samples are used to expand the PDQ database [102–104]. The database includes two text information, one is the vehicle model, manufacturer, year, and

other vehicle information, the other is the level of order, quantity, color, and chemical composition information, including each layer of paint infrared spectrum. Based on PDQ infrared spectroscopy database, Lavine et al. carried out systematic research work [105–107]. For example, they first transformed the infrared spectra of car paint by wavelet, and then selected the feature variables of wavelet coefficients by GA, and clustered the car paint by PCA. For the samples to be analyzed, they first quickly judged their classes, and then obtained accurate results through the spectral searching algorithm. They also carried out segmented autocorrelation transform and cross-correlation transform of the spectra to eliminate the influence of the spectral subtle differences between different classes of instruments on the spectral searching results. In the field of criminal investigation, spectral searching methods are increasingly combined with expert knowledge and experience, further evaluation and fusion are carried out through probability theory methods, such as likelihood ratio [108, 109].

The spectral searching algorithm has also made great progress in quantitative analysis of NIR. For example, Li et al. combined moving correlation coefficient with Monte Carlo (MC) method for quantitative analysis of NIR, and directly predicted the octane number and chemical composition of gasoline by spectral searching algorithm [110]. As shown in Fig. 12.23, this method firstly selects a group of samples with the closest spectra to the sample to be measured from the training set based on the moving correlation coefficient method, and uses the MC method to generate thousands of virtual spectra by using this group of the closest spectra. Then, it uses the moving correlation coefficient method to search all the virtual spectra that are completely consistent with the spectra of the sample to be measured. According to the basic principle of “the same sample, the same spectrum and the same property”, the quantitative analysis results are finally given by weighted average method. When

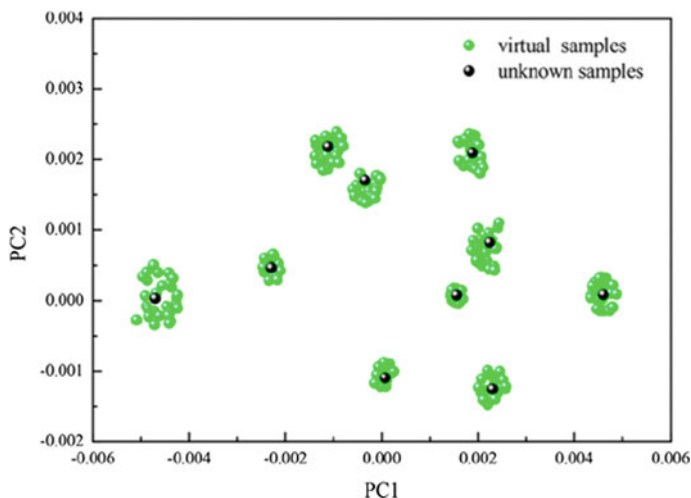


Fig. 12.23 Schematic diagram of spectral searching algorithm combined with virtual spectrum for quantitative analysis

encountering outliers in the spectral database, different from the traditional quantitative multivariate calibration method, the quantitative method of spectral searching method only needs to add outliers into the spectral database according to certain rules, which is extremely convenient to operate and does not require professional modeling personnel. Therefore, this strategy of applying spectral searching method for quantitative analysis is expected to become a common method. Bi et al. also proposed similar ideas and applied the spectral searching algorithm to the evaluation of tobacco quality, tobacco leaf substitution, and cigarette formulation maintenance [111, 112].

Differences between instruments in multivariate quantitative calibration are called calibration transfer, which still exist in spectral searching technology. There are not only certain differences between spectral instruments of different brands but also slight differences between instruments of the same type. The essence of the model transfer algorithm is the mathematical transformation between spectra. With the continuous expansion of the applications of spectral library, although this problem has received more and more attention in spectral searching [113–118], compared with quantitative multivariate calibration, the research is not systematic, and there are few reported cases of practical application in the spectral searching field, which need to be further studied.

Finally, it is worth mentioning that while paying attention to spectral retrieval algorithms and strategies, we should pay more attention to the experimental technology of establishing spectral database, that is, how to obtain high-quality standard library spectra (large amount of information, strong characteristics, high signal-to-noise ratio, excellent reproducibility, etc.), which involves many technical details, such as the instruments, sample preprocessing methods, measurement and accessories, optimization of measurement parameters, standardization of library construction process, etc. High-quality spectra is the basis of all searching methods, therefore, to some extent, the experimental technology of establishing spectral library is more important than the searching method.

References

1. Breton RG. Chemometrics for pattern recognition. Wiley;2008.
2. Zhang XG, Bian ZQ. Pattern recognition, 2nd ed. Beijing: Tsinghua University Press;2004.
3. Sun XJ, Liu XY. New genetic k-means clustering algorithm based on meliorated initial center. *Comput Eng Appl.* 2008;44(23):166–8.
4. Liu JM, Han LC, Hou LW. Cluster analysis based on particle swarm optimization algorithm. *Syst Eng Theory Pract.* 2005;6:54–8.
5. Yang X, Peng YQ. The k-means clustering analysis combined with ant colony. *J Hebei Univ Technol.* 2007;36(3):48–52.
6. Chu XL, Yuan HF, Lu WZ. Samples clustering and recognition with fuzzy clustering and principal component analysis method in spectral analysis. *Chin J Anal Chem.* 2000;28(4):421–7.
7. Li XH, Luo HY, Xu XQ, et al. The classification of tea based on PCA and GMM. *J Zhengzhou Univ (Natural Science Edition).* 2015;47(4):62–5.

8. Sun XD, Li XH, Shi WM, et al. An improved GMM and GMR based on particle swarm optimization with an application in extra virgin olive oil analysis. *J Pingdingshan Univ.* 2015;30(5):62–5.
9. Wang W, Jiang H, Liu GH, et al. Quantitative analysis of yeast growth process based on FT-NIR spectroscopy integrated with gaussian mixture regression. *RSC Adv.* 2017;7(40):24988–94.
10. Yang SY. Pattern recognition and intelligent computing technology realization in MATLAB, 3rd ed. Beijing: Publishing House of Electronics Industry;2015.
11. Nørgaard L, Bro R, Westad F, et al. A modification of canonical variates analysis to handle highly collinear multivariate data. *J Chemom.* 2006;20(8–10):425–35.
12. Canals T, Riba JR, Cantero R, et al. Characterization of paper finishes by use of infrared spectroscopy in combination with canonical variate analysis. *Talanta.* 2008;77(2):751–7.
13. Huang YC. Scikit-learn machine learning. Beijing: Machinery Industry Press;2018.
14. Gan BR, Yang ZH, Zhang WD, et al. Stacked contractive auto-encoders application in identification of pharmaceuticals. *Spectrosc Spectr Anal.* 2019;39(1):96–102.
15. Wang L, Qin F, Li J, et al. Geographical origin identification of *Lycium barbarum* using near-infrared hyperspectral imaging. *Spectrosc Spectr Anal.* 2020;40(4):1270–5.
16. Liu HJ, Meng XT, Wang X, et al. Soil classification model based on the characteristics of soil reflectance spectrum. *Spectrosc Spectr Anal.* 2019;39(8):2481–5.
17. Ke YC, Shi ZK, Li PJ, et al. Lithological classification and analysis using hyperion hyperspectral data and random forest method. *Acta Petrol Sinica.* 2018;34(7):2181–8.
18. Kong QQ, Ding XQ, Gong HL, et al. Research on application of feature selection algorithm based on combination of random forest and game theory in near infrared spectroscopy. *J Instrum Anal.* 2017;36(10):1203–7.
19. Lai YH, Lin Y, Tao H, et al. Rapid identification of tobacco mildew based on near infrared spectroscopy and random forest algorithm. *Acta Tabacaria Sinica.* 2020;26(2):36–43.
20. Li ZH, Shen J, Bian RH, et al. Accuracy comparison of the machine learning algorithm used to Raman real sample collection in the front line of public security. *Spectrosc Spectr Anal.* 2019;39(7):2171–5.
21. Wang Y, Zhe S, Zhou N, et al. Classification of terahertz rosewood based on continuous projection algorithm and random forest. *Spectrosc Spectr Anal.* 2019;39(9):2719–24.
22. Zhou YH, Zuo ZT, Xu FR, et al. Origin identification of *Panax notoginseng* by multi-sensor information fusion strategy of infrared spectra combined with random forest. *Spectrochim Acta Part A Mol Biomol Spectrosc.* 2020;226:117619.
23. Amjad A, Ullah R, Khan S, et al. Raman spectroscopy based analysis of milk using random forest classification. *Vib Spectrosc.* 2018;99:24–129.
24. Ma LF, Xiong HG, Zhang F. Prediction of major ions in soil salinity based on field VIS-NIR spectroscopy. *Soils.* 2020;52(1):188–94.
25. Li GW, Gao XH, Xiao NW, et al. Estimation soil organic matter contents with hyperspectra based on sCARS and RF algorithms. *Chin J Lumin.* 2019;40(8):1030–9.
26. Zheng PC, Zheng S, Wang JM, et al. Study on grade identification of dendrobium by LIBS. *Spectrosc Spectr Anal.* 2020;40(3):941–4.
27. Li MG, Yan CH, Xue J, et al. Rapid quantitative analysis of methanol content in methanol gasoline by near infrared spectroscopy coupled with wavelet transform-random forest. *Chin J Anal Chem.* 2019;47(12):1995–2003.
28. De Santana FB, De Souza AM, Poppi RJ. Visible and near infrared spectroscopy coupled to random forest to quantify some soil quality parameters. *Spectrochim Acta Part A Mol Biomol Spectrosc.* 2018;191:454–62.
29. Teixeira AFDS, Pelegrino MHP, Faria WM, et al. Tropical soil pH and sorption complex prediction via portable x-ray fluorescence spectrometry. *Geoderma.* 2020;(361):114132.
30. Zhang SP, Tan Z L, Liu J, et al. Determination of the food dye indigotine in cream by near-infrared spectroscopy technology combined with random forest model. *Spectrochim Acta Part A Mol Biomol Spectrosc.* 2020;227:117551.
31. Blanco M, Romero MA. Near-infrared libraries in the pharmaceutical industry: a solution for identity confirmation. *Analyst.* 2001;126(12):2212.2217.

32. Lavine B, Almirall J, Muehlethaler C, et al. Criteria for comparing infrared spectra—a review of the forensic and analytical chemistry literature. *Forensic Chem.* 2020;18:100224.
33. Araujo CF, Nolasco MM, Ribeiro AMP, et al. Identification of microplastics using Raman spectroscopy: latest developments and future prospects. *Water Res.* 2018;142:426–40.
34. Terhoeven-Urselmans T, Vagen TG, Spaargaren O, et al. Prediction of soil fertility properties from a globally distributed soil mid-infrared spectral library. *Sci Soc Am J.* 2010;74(5):1792–9.
35. Veij MD, Vandennebelee P, Beer TD, et al. Reference database of Raman spectra of pharmaceutical excipients. *J Raman Spectrosc.* 2008;40(3):297–307.
36. Lafuente B, Downs RT, Yang H, et al. The power of databases: the RRUFF project. *Highlights Mineral Crystallogr.* 2015;1:1–30.
37. Viscarra Rossel RA, Behrens T, Ben-Dor E, et al. A global spectral library to characterize the world's soil. *Earth Sci Rev.* 2016;155:198–230.
38. Zhou W, Ying Y, Xie L. Spectral database systems: a review. *Appl Spectrosc Rev.* 2012;47(8):654–70.
39. Wang L, Guo HL, Zhu J, et al. Application of statistical methods in micro material evidence data processing. *Forensic Sci Technol.* 2020;45(2):125–30.
40. Wang XF, Gao CF, Xu BQ, et al. Visual and rapid identification of sole materials by mid infrared spectroscopy. *China Plast.* 2019;33(8):101–5.
41. Leung AK, Chau FM, Gao JB, et al. Application of wavelet transform in infrared spectrometry: spectral compression and library search. *Chemom Intell Lab Syst.* 1998;43(1–2):69–88.
42. Zhao CH, Tian MH, Li JW. Research progress of spectral similarity measurement methods. *J Harbin Eng Univ.* 2017;38(8):1179–89.
43. Chang CI. An information-theoretic approach to spectral variability, similarity, and discrimination for hyperspectral image analysis. *IEEE Trans Inf Theory.* 2000;46(5):1927–32.
44. Yan Y, Zhang HG, et al. Local partial least squares modeling method of near infrared spectrum based on spectral information divergence. *Comput Appl Chem.* 2017;34(5):18–22.
45. Liu YS, Cao M, Wang YM, et al. Quantitative evaluation of similarity of chromatographic fingerprints of traditional Chinese medicine by similarity system theory. *Chin J Anal Chem.* 2006;34(3):333–7.
46. Zhang WL, Wang XP, Zhao Y, et al. Comparison method of infrared spectra based on similarity system theory. *Chin J Spectrosc Lab.* 2013;30(6):2742–6.
47. Varmuza K, Karlovits M, Demuth W. Spectral similarity versus structural similarity: infrared spectroscopy. *Anal Chim Acta.* 2003;490(1&2):313–24.
48. Zhou WH, Xie LJ, Ying YB. Application of all-optical spectrum matching algorithm in apple classification and recognition. *Trans Chin Soc Agric Eng.* 2013;19:285–92.
49. Meng QH, Wang WB, Hu YZ. UV spectral similarity and its application in quality control of traditional Chinese medicine injection. *China J Chin Materia Med.* 2007;32(3):206–10.
50. Tang TB, Yang HH, Liang XZ, et al. Weighted similarity measurement method and its application in spectral anomaly determination. *J Guilin Univ Electron Technol.* 2012;32(5):391–7.
51. Li Y, Lv JW, Chen L. Study on the stability of different concentrations of *Salvia miltiorrhiza* injection by UV spectral similarity method. *Pharm Care Res.* 2011;11(4):304–6.
52. Khan SS, Madden MG. New similarity metrics for Raman spectroscopy. *Chemom Intell Lab Syst.* 2012;114(1):99–108.
53. Plugge W, Vlies CJVD. Near-infrared spectroscopy as an alternative to assess compliance of ampicillin trihydrate with compendia specifications. *J Pharm Biomed Anal.* 1993;11(6):435–42.
54. Ritchie GE, Mark H, Ciurczak EW. Evaluation of the conformity index and the mahalanobis distance as a tool for process analysis: a technical note. *AAPS Pharm Sci Tech.* 2003;4(2):109–18.
55. Feng YC, Yang XL, Yang ZH, et al. Monitoring the quality of drugs in circulation using rapid NIR spectral comparison methods. *J Chin Pharm Sci.* 2011;20(3):290–6.
56. Zhang XB, Yi LH. Study on the consistency test method of near infrared spectroscopy for rapid judgment of drug quality. *Chin J Pharm Anal.* 2011;31(3):603–8.

57. Zhou W, Chen W. Consistency test of erythromycin film coated tablets by near infrared diffuse reflectance spectroscopy. *China Pharm.* 2009;12(4):451–2.
58. Nie LX, Wang GL, Li ZM, et al. Qualitative and quantitative analysis of Tongren Wu Ji Bai Feng Pills by near infrared spectroscopy. *J Infrared and Millim Waves.* 2008;27(3):205–9.
59. Tao Y, Dang LZ, Liu J, et al. Application of near infrared spectroscopy in quality stability control of cigarette silk. *Chin J Spectrosc Lab.* 2013;30(1):27–32.
60. Lu Y, Jiang L, Wu WW, et al. Establishment and application of textile fiber infrared spectrum database based on attenuated total reflection method. *China Fiber Insp.* 2013;1:71–3.
61. Wang Y, Ji L, Wang YJ, et al. Establishment and application of infrared standard spectrum library of plastic resin. *Eng Plast Appl.* 2005;33(9):47–51.
62. Howari FM. Comparison of spectral matching algorithms for identifying natural salt crusts. *J Appl Spectrosc.* 2003;70(5):782–7.
63. Reeves JB, Zapf CM. Spectral library searching: mid-infrared versus near-infrared spectra for classification of powdered food ingredient. *Appl Spectrosc.* 1999;53(7):836–44.
64. Chen T, Long XJ, Wei L, et al. Comparison of automobile body paint based on fourier infrared spectroscopy. *Spectrosc Spectr Anal.* 2013;33(2):367–70.
65. He T, Sheng JL. Application of terahertz spectroscopy in drug detection. *Spectrosc Spectr Anal.* 2013;33(9):2348–53.
66. Guedes A, Ribeiro H, Fernandez-Gonzalez M, et al. Pollen Raman spectra database: application to the identification of airborne pollen. *Talanta.* 2014;119:473–8.
67. Meer FVD. The effectiveness of spectral similarity measures for the analysis of hyperspectral imagery. *Int J Appl Earth Obs Geoinf.* 2006;8(1):3–17.
68. Blanco M, Eustaquio A, González JM, et al. Identification and quantitation assays for intact tablets of two related pharmaceutical preparations by reflectance near-infrared spectroscopy: validation of the procedure. *J Pharm Biomed Anal.* 2000;22(1):139–48.
69. Zhang XB, Ma JJ, Cao LM. Near infrared spectral correlation coefficient method for rapid detection of drug quality. *Chin J Spectrosc Lab.* 2013;30(4):2010–5.
70. Wang XL, Feng YC, Hu CQ. Determination of sildenafil citrate in traditional Chinese medicine capsules by near infrared characteristic band correlation coefficient method. *Chin J Anal Chem.* 2009;37(12):1825–8.
71. Xu YQ, Sun SQ, Xu JW. Rapid identification of traditional Chinese medicine by infrared fingerprint library and array correlation coefficient. *Chin J Spectrosc Lab.* 2002;19(5):606–10.
72. Park JK, Park A, Yang SK, et al. Raman spectrum identification based on the correlation score using the weighted segmental hit quality index. *Analyst.* 2017;142(2):380–8.
73. Griffiths PR, Shao LM. Self-weighted correlation coefficients and their application to measure spectral similarity. *Appl Spectrosc.* 2009;63(8):916–9.
74. Chu XL, Xu YP, Tian SB, et al. Rapid identification and assay of crude oils based on moving-window correlation coefficient and near infrared spectral library. *Chemom Intell Lab Syst.* 2011;107(1):44–9.
75. Chu XL, Tian SB, Xu YP, et al. Study on rapid evaluation of crude oil by near infrared. *China Petrol Process Petrochem Technol.* 2012;43(1):72–7.
76. Li JY, Chu XL, Tian SB, et al. The identification of highly similar crude oils by infrared spectroscopy combined with pattern recognition method. *Spectrochim Acta Part A Mol Biomol Spectrosc.* 2013;112(8):457–62.
77. Li JY, Chu XL, Tian SB. Application of infrared two-dimensional correlation spectroscopy in rapid identification of crude oil. *Acta Petrolei Sinica (Petroleum Processing Section).* 2013;29(4):655–60.
78. Guo ZF, Dai LK. Endpoint judgment of traditional Chinese medicine extraction process based on near infrared spectral shape analysis. *Chin J Spectrosc Lab.* 2013;30(5):2418–23.
79. Ramirez-Lopez L, Behrens T, Schmidt K, et al. Distance and similarity-search metrics for use with soil Vis–NIR spectra. *Geoderma.* 2013;199(1):43–53.
80. Liu XF, Yang CA. Kernel spectral angle mapper algorithm for remote sensing image classification. In: *International congress on image and signal processing*, Hangzhou; 2013. pp. 814–818.

81. Camps-Valls G. Kernel spectral angle mapper. *Electron Lett.* 2016;52(14):1218–20.
82. Van DMF, Bakker W. CCSM: cross correlogram spectral matching. *Int J Remote Sens.* 1997;18(5):1197–201.
83. Himmelsbach DS, Hellgeth JW, McAlister DD. Development and use of an attenuated total reflectance/fourier transform infrared (ATR/FT-IR) spectral database to identify foreign matter in cotton. *J Agric Food Chem.* 2006;54(20):7405–12.
84. Loudermilk JB, Himmelsbach DS, Barton FE, et al. Novel search algorithms for a mid-infrared spectral library of cotton contaminants. *Appl Spectrosc.* 2008;62(6):661–70.
85. Kong XB, Shu N, Tao JB. A new spectral similarity measure based on multi feature fusion. *Spectrosc Spectr Anal.* 2011;31(8):2166–70.
86. Zhao CF, Qi MJ, Ma YJ, et al. An oil spill type identification method based on fluorescence spectrum: ZL201010216725.6 [P]. 2010.
87. Feng C, Zhao NJ, Yin GF, et al. Identification method of pathogenic bacteria in water based on spectral similarity analysis. *Acta Optica Sinica.* 2020;40(3):0330002.
88. Zhang T, He FL, Jia EH, et al. A qualitative and quantitative identification method of dangerous liquid mixtures by Raman spectroscopy. *Spectrosc Spectr Anal.* 2019;39(11):3372–6.
89. Liu MH, Dong ZR, Xin GF, et al. Discrimination method of Raman spectrum peaks based on Voigt function fitting. *Chin J Lasers.* 2017;44(5):0511003.
90. Liu MH, Dong ZR, Xin GF, et al. Matching method of Raman spectrum library based on integrated features. *Chin J Lasers.* 2019;46(1):0111002.
91. Wei CL, Zhao YG, Li DC, et al. Prediction of soil organic matter and cation exchange capacity based on similar spectral matching. *Trans Chin Soc Agric Eng.* 2014;30(1):81–8.
92. Shi X, Cai WS, Shao XG. Research on local modeling method and application of near infrared spectrum based on wavelet coefficients. *Chin J Anal Chem.* 2008;36(8):1093–6.
93. Genot V, Colinet G, Bock L, et al. Near infrared reflectance spectroscopy for estimating soil characteristics valuable in the diagnosis of soil fertility. *J Near Infrared Spectrosc.* 2011;19(2):117–38.
94. Damberg R, Cozzolino D, Cynkar W, et al. The determination of red grape quality parameters using the LOCAL algorithm. *J Near Infrared Spectrosc.* 2006;14(1):71–9.
95. Fernandez-Ahumada E, Fearn T, Gomez-Cabrera A, et al. Evaluation of local approaches to obtain accurate near-infrared (NIR) equations for prediction of ingredient composition of compound feeds. *Appl Spectrosc.* 2013;67(8):924–9.
96. Li JY, Chu XL, Chen P, et al. Application of automatic spectral retrieval algorithm in rapid establishment of gasoline spectral database. *Acta Petrolei Sinica (Petroleum Processing Section).* 2017;33(1):131–7.
97. Lee S, Lee H, Chung H. New discrimination method combining hit quality index based spectral matching and voting. *Analytica Chimica Acta,* 758:58–65.
98. Zhang P, Wang XK, Li HT, et al. Terahertz spectrum recognition based on fractal theory. *Chin J Quantum Electron.* 2007;24(6):672–7.
99. Qin YH, Duan K, Wu LJ, et al. Similarity measure method based on spectra subspace and locally linear embedding algorithm. *Infrared Phys Technol.* 2019;100:57–61.
100. Song CJ, Ding XQ, Xu PM, et al. Research on spectral similarity measurement method based on adjacent set calculation. *Spectrosc Spectr Anal.* 2017;37(7):2032–5.
101. Lei M, Feng XL. Identification of internal combustion engine oil based on near infrared spectroscopy. *J Instrum Anal.* 2009;28(5):529–34.
102. Xu BD, Ding XQ, Qin YH, et al. Near infrared spectral similarity measurement method based on meshing local linear embedding algorithm. *Laser Optoelectron Progress.* 2019;56(3):251–7.
103. Lavine BK, Mirjankar N, Ryland S, et al. Wavelets and genetic algorithms applied to search prefilters for spectral library matching in forensics. *Talanta.* 2011;87:46–52.
104. Lavine BK, Nuguru K, Mirjankar N, et al. Development of carboxylic acid search prefilters for spectral library matching. *Microchem J.* 2012;103:21–36.
105. Lavine BK, Fasasi A, Mirjankar N, et al. Search prefilters for library matching of infrared spectra in the PDQ database using the autocorrelation transformation. *Microchem J.* 2014;113:30–5.

106. Lavine BK, White C, Allen M, et al. Pattern recognition-assisted infrared library searching of the paint data query database to enhance lead information from automotive paint trace evidence. *Appl Spectrosc.* 2016;71(3):480–95.
107. Lavine BK, White C, Allen M. Forensic analysis of automotive paints using a pattern recognition assisted infrared library searching system: ford (2000–2006). *Microchem J.* 2016;129:173–83.
108. Lavine BK, White CG, Ding T. Library search prefilters for vehicle manufacturers to assist in the forensic examination of automotive paints. *Appl Spectrosc.* 2018;72(3):476–88.
109. Martyna A, Michalska A, Zadora G. Interpretation of FTIR spectra of polymers and Raman spectra of car paints by means of likelihood ratio approach supported by wavelet transform for reducing data dimensionality. *Anal Bioanal Chem.* 2015;407(12):3357–76.
110. Muehlethaler C, Massonnet G, Hicks T. Evaluation of infrared spectra analyses using a likelihood ratio approach: a practical example of spray paint examination. *Sci Justice.* 2016;56(2):61–72.
111. Li JY, Chu XL. Rapid determination of physical and chemical parameters of reformed gasoline by near-infrared (NIR) spectroscopy combined with the Monte Carlo virtual spectrum identification method. *Energy Fuels.* 2018;32(12):12013–20.
112. Bi YM, Li ST, Zhang LL, et al. Quality evaluation of flue-cured tobacco by near infrared spectroscopy and spectral similarity method. *Spectrochim Acta Part A Mol Biomol Spectrosc.* 2019;215:398–404.
113. Li ST, Liao F, He WM, et al. Tobacco leaf substitution and cigarette formula maintenance based on near infrared spectrum similarity. *Tobacco Sci Technol.* 2020;53(2):88–93.
114. Yoon WL, Jee RD, Moffat AC, et al. Construction and transferability of a spectral library for the identification of common solvents by near-infrared transmittance spectroscopy. *Analyst.* 1999;124(8):1197–1203.
115. Yoon WL, Jee RD, Moffat AC. An interlaboratory trial to study the transferability of a spectral library for the identification of the solvents using near-infrared spectroscopy. *Analyst.* 2000;125(10):1817–22.
116. Ma LZ, Guan L, Feng XL, et al. Similarity evaluation method of infrared spectrum fingerprint area of lubricating oil based on Pruck analysis. *Acta Petrolei Sinica (Petroleum Processing Section).* 2013;29(5):891–898.
117. Genot V, Colinet G, Dardene P, et al. Transferring a calibration model and a spectral library to a soil analysis laboratory network. *Geophys Res Abstr.* 2009;11:2805.
118. Lavine BK, Fasasi A, Mirjankar N, et al. Development of search prefilters for infrared library searching of clear coat paint smears. *Talanta.* 2014;119:331–40.
119. Chen H, Zhang ZM, Miao L, et al. Automatic standardization method for Raman spectrometers with applications to pharmaceuticals. *J Raman Spectrosc.* 2014;46(1):147–54.

Chapter 13

Model Evaluation



13.1 Evaluation of Quantitative Calibration Model

13.1.1 Evaluation Parameters

Some statistical parameters, such as standard error of calibration (SEC), standard error of prediction (SEP), determination coefficient (R^2), or correlation coefficient (R), are often used in the process of model establishment and validation [1, 2].

(1) Bias or residual (d) and range (e)

The formula of bias or residual (d) is as follows:

$$d_i = y_{i,\text{predicted}} - y_{i,\text{actual}} \quad (13.1)$$

where $y_{i,\text{actual}}$ is the measured value of i th sample by the reference method; $y_{i,\text{predicted}}$ is the predicted value of the i th sample of the calibration set or the validation set. It is generally required that the deviation d_i should be less than the reproducibility specified in the reference measurement method. The mean bias is the mean value of d_i of all samples in the calibration set or validation set. Range e is the maximum bias of all samples in the calibration set or validation set, that is:

$$e = \max (d_i) \quad (13.2)$$

(2) Standard error of calibration (SEC)

The formula of SEC is as follows:

$$\text{SEC} = \sqrt{\frac{\sum_{i=1}^n (y_{i,\text{actual}} - y_{i,\text{predicted}})^2}{n - 1}} \quad (13.3)$$

where $y_{i,\text{actual}}$ is the measured value of the i th sample by the reference method; $y_{i,\text{predicted}}$ is the predicted value of the i th sample in the calibration set using the built model; n is the number of samples in the calibration set. In some literature, SEC is also called as RMSEC (root mean square error of calibration).

(3) Standard error of cross validation (SECV)

The formula of SECV is as follows:

$$\text{SECV} = \sqrt{\frac{\sum_{i=1}^n (y_{i,\text{actual}} - y_{i,\text{predicted}})^2}{n - 1}} \quad (13.4)$$

where $y_{i,\text{actual}}$ is the measured value of the i th sample by the reference method; $y_{i,\text{predicted}}$ is the predicted value of the i th sample in the cross-validation process of the calibration set; n is the number of samples in the calibration set. In some literature, SECV is also called as RMSECV (root mean square error of cross validation).

(4) Standard error of prediction (SEP)

The formula of SEP is as follows:

$$\text{SEP} = \sqrt{\frac{\sum_{i=1}^n (y_{i,\text{actual}} - y_{i,\text{predicted}})^2}{m - 1}} \quad (13.5)$$

where $y_{i,\text{actual}}$ is the measured value of the i th sample by the reference method; $y_{i,\text{predicted}}$ is the predicted value of the i th sample in the process of the prediction in the validation set; m is the number of samples in the validation set. In some literature, SEP is also called as RMSEP (root mean square error of prediction). The smaller the SEP is, the stronger the predictive power of the model is. In general, SEP is bigger than SEC and SECV.

The calculation formulas of SEC and SEP in different literature are also slightly different. For example, the number of PLS factors f is considered in SEC, while SEP is corrected by average bias as follows:

$$\text{SEC} = \sqrt{\frac{\sum_{i=1}^n (y_{i,\text{actual}} - y_{i,\text{predicted}})^2}{n - f - 1}} \quad (13.6)$$

$$\text{SEP}_{-b} = \sqrt{\frac{\sum_{i=1}^n (y_{i,\text{actual}} - y_{i,\text{predicted}} - \text{bias})^2}{m - 1}} \quad (13.7)$$

$$\text{SEP}^2 = \text{SEP}_{-b}^2 + \text{bias}^2 \quad (13.8)$$

(5) Ratio of standard deviation of the validation set to SEP (RPD)

The formula of RPD is as follows:

$$\text{RPD} = \frac{\text{SD}_v}{\text{SEP}} \quad (13.9)$$

where SD_v is the standard deviation of the concentration values of all samples in the validation set. The wider and the more uniform the property distribution of the samples in the validation set is, and the smaller the SEP is, then the larger the RPD value will be.

(6) Ratio of the SEP to the range (RER)

The formula of RER is as follows:

$$\text{RER} = R_n / \text{SEP} \quad (13.10)$$

where R_n is the property distribution range of samples in the validation set, that is, the range of concentration. If the concentration of samples in the validation set is normally distributed, there is a relationship of $\text{RER} = 2 \times \text{RPD}$.

(7) Ratio of performance to interquartile range (RPIQ)

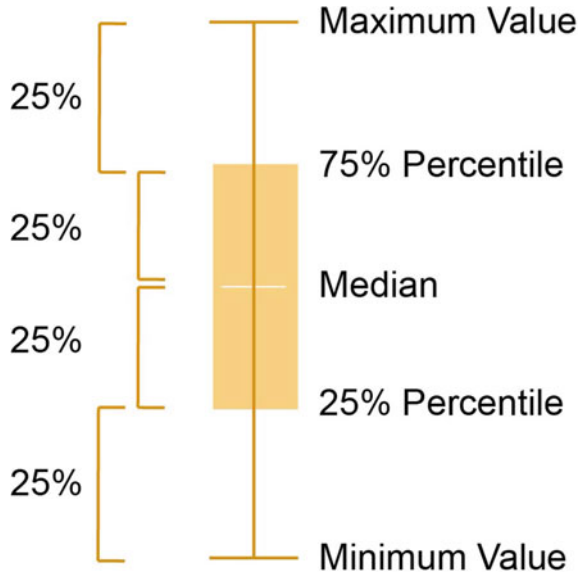
For a set of data with a normal distribution, the SD value can be used to express its distribution, that is, about 67% of the samples are distributed within $\pm\text{SD}$ range. However, for some data sets, such as the content of organic carbon in soil, the distribution is approximately lognormal, and about 93% of the samples are within the $\pm\text{SD}$ range. Then interquartile range (IQR) is used to replace the SD to calculate new evaluation parameters (ratio of performance to interquartile range, RPIQ) [3].

Quartile is a kind of quantile in statistics, that is, all the values are arranged from the smallest to the largest and divided into four equal parts. The values at the position of the three dividing points are the quartile (Fig. 13.1). The first quartile (Q1), also known as the “smaller quartile”, is equal to 25% of all the values in the data sets in order from smallest to largest; The second quartile (Q2), also known as the “median”, is equal to the 50% number of values in the data sets from smallest to largest; The third quartile (Q3), also known as the “higher fourth quartile”, is equal to the 75% of all values in the data sets in order from smallest to largest. The difference between the third and first quartile is called interquartile range (IQR).

The calculation formula of RPIQ is as follows:

$$\text{RPIQ} = (\text{Q3} - \text{Q1}) / \text{SEP} \quad (13.11)$$

Fig. 13.1 Schematic diagram of quarter segmentation



(8) Determination coefficient (R^2) or correlation coefficient (R)

The calculation formula of R^2 is as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_{i,\text{actual}} - y_{i,\text{predicted}})^2}{\sum_{i=1}^n (y_{i,\text{actual}} - \bar{y}_{\text{actual}})^2} \tag{13.12}$$

where $y_{i,\text{actual}}$ is the measured value of the i th sample by the reference method; \bar{y}_{actual} is the average value of all the measured values of the calibration set or the validation set by the reference method; $y_{i,\text{predicted}}$ is the predicted value of the i th sample in the prediction process of the calibration set or the validation set; n is the number of samples of the calibration set or the validation set. Under the premise of the same concentration range, the closer the R is to 1, the better the regression or prediction result should be.

For $R^2 < 0.9$, only one significant digit is required to represent R^2 , for example, $R^2 = 0.8$; For $R^2 < 0.99$, reserve two significant digits to represent R^2 , for example, $R^2 = 0.96$; For $R^2 < 0.999$, three significant digits are required to represent R^2 , for example, $R^2 = 0.994$; For $R^2 < 0.9999$, four significant digits are required to represent R^2 , for example, $R^2 = 0.9998$.

(9) Paired t test

Assuming that there is no systematic error between the spectroscopic method and the reference method, there should be no significant difference between the mean value \bar{d} and 0. \bar{d} is the difference between the measured results of the two methods, namely, $\bar{d} = 0$. The paired t -test statistic is as follows:

$$t = \frac{\bar{d}}{S_d/\sqrt{m}} \quad (13.13)$$

where \bar{d} is the mean deviation between the two results of the spectroscopic method and the reference method; S_d is the standard deviation of the deviation between the two analysis methods, and m is the number of measured samples.

For a given level of significance of α , if $|t| < t_{(\alpha, m-1)}$, indicating that there was no significant difference between the predicted value by the calibration model and the average value determined by the reference method.

13.1.2 Model Evaluation

In general, in the method of combining spectra with chemometrics, the modeling parameters and results that need to be reported should include: the size of calibration samples, the distribution range and standard deviation, reference methods and their repeatabilities and reproducibilities' requirements, spectral preprocessing methods and their parameters, wavelength selection methods and wavelength range, multivariate calibration methods for quantitative or qualitative analysis, the number of the outlier samples that are eliminated, regression and validation and its statistical parameters (such as the best number of principal component, SEC, SEP, and SECV.

For the establishment and validation of the spectral quantitative model, it is suggested that the following model parameters should be reported [4]:

- (1) Source of samples;
- (2) Sample preparation and storage methods;
- (3) Selection method of dividing samples into calibration set and validation set;
- (4) The number of samples in calibration set, the number of samples in test set, and the number of samples in validation set;
- (5) Reference method and its standard error of the test;
- (6) Mean and standard deviation of the reference values;
- (7) Modeling method and its parameters (such as the number of principal components of PLS method);

- (8) Data preprocessing method and its parameters;
- (9) Spectral wavelength interval or wavelength selection method and results;
- (10) Cross-validation method (such as leave one out cross validation, LOOCV);
- (11) Identification method of outlier samples and the results;
- (12) SECV (or RMSECV);
- (13) Regression coefficient, slope, and intercept;
- (14) RPD, R, or R^2 ;
- (15) Standard error of spectral method;
- (16) Software and its version used.

13.1.2.1 Number and Representativeness of Calibration Sets

Basically, there are two requirements for calibration samples: (1) the sample should be representative, and its composition should be included all of the chemical composition of the sample to be tested. The variation range of composition should be greater than the corresponding range of samples to be tested. Usually, the range of composition should be greater than five times of reproducibility of the reference measurement, and the composition is evenly distributed over the whole range. For example, the reproducibility of octane number of the gasoline determined by the standard method is 0.7 units, and the range of the calibration samples is at least 3.5 units; (2) the quantity should be large enough to effectively extract the quantitative mathematical relationship between the spectra and the components to be measured. For a simple analytical system, at least 60 representative samples are required, and for a complex measurement system, at least hundreds of representative samples are required.

As shown in Fig. 13.2, sometimes test set is used in the process of model establishment, also known as control set or optimization set in some literature. In fact, it should be a part of calibration set (or training set) and mainly used for determining and

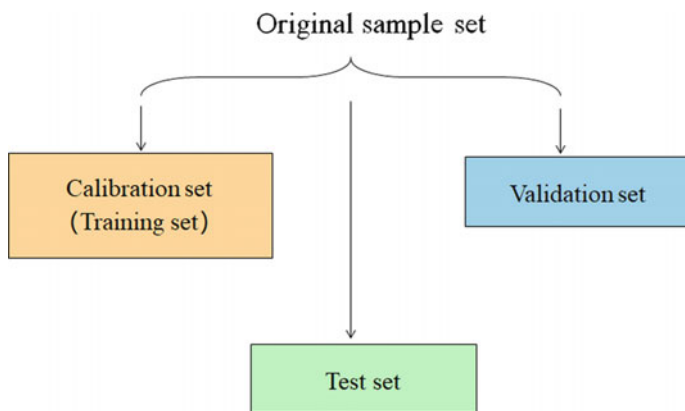


Fig. 13.2 Relationship between original sample set, calibration set, test set, and validation set

optimizing model parameters. For example, preprocessing methods and the number of PLS principal components can be determined through test set. In addition, “early stop” strategy in ANN is also based on the test set.

In general, 20% of the collected samples are used as test set, and 20% of them are used as independent validation set, and the remaining 60% are used for establishing the cross-validation model. After the model parameters have been optimized, the samples of test set and calibration set are combined to establish the final model. The representativeness of samples in calibration set, test set, and validation set should be considered simultaneously.

13.1.2.2 Evaluation of Model Establishment Process

In the process of model establishment, metrics such as SEC, SECV, and R^2 are used to evaluate the quality of the model to select the optimal modeling parameters.

The smaller the SEC, the better the regression of the model. Generally, the repeatability of SEC is equivalent to that specified by the reference method. If SEC is too small, it indicates that the calibration model may be overfitting, and SECV is usually greater than SEC.

The determination coefficient (R^2) of the calibration result can be also expressed

$$\text{as follows : } R^2 = 1 - \frac{SEC^2}{SD_c^2} \quad (13.14)$$

where SEC is the standard error of calibration, and SD_c is the standard deviation of concentration value of calibration samples. It can be seen that the size of R^2 is related to the concentration distribution range. For the same SEC, the wider the concentration distribution range (the larger the SD_c), the larger the R^2 .

The maximum R^2 value can be estimated as the following formula:

$$R_{\max}^2 = 1 - \frac{SEL^2}{SD_c^2} \quad (13.15)$$

where SEL is the repeatability of the reference method. If R^2 exceeds this maximum value, it is highly likely that the model is overfitting.

13.1.2.3 Model Validation

After the establishment of the model, it is necessary to use the validation set to verify the accuracy, repeatability, robustness, and transitivity of the model. The validation set is composed of a set of samples, which are completely independent of the calibration set. Only the validated model can be applied in practice.

The samples in validation set should contain all the chemical components contained in the samples to be tested, and the concentration or property range of the validation samples should cover at least 95% of the concentration or property range of the calibration samples, and the distribution should be uniform. In addition, the number of the validation samples should be large enough to conduct statistical test, usually requiring no less than 28 samples.

Accuracy: the spectra of the validation samples should be measured in the same way as that of the calibration samples, and the reference values should be determined in the same way as that of the calibration samples. The following parameters are usually used to evaluate the accuracy of the model:

- ① Standard error of prediction (SEP): the smaller the SEP, the more accurate the results. Some literature requires that the ratio of SEP to SEC should be less than 1.2, that is, SEP should not be greater than 1.2 times of SEC. According to probability statistics, the deviation between the predicted value and the actual value of the reference method can be estimated by SEP. If the predicted value of the spectral method is \hat{y} , the probability that the actual value of the reference method falls within the range of $[\hat{y} \pm \text{SEP}]$ is about 67% and that of $[\hat{y} \pm 2 \times \text{SEP}]$ is about 95%. For example, the SEP of wheat moisture determined by NIR spectroscopy is 0.5%, and if the predicted value of the sample is 20.0%, the probability that the actual value of the reference method falls between 19.0 and 21.0% is about 95%.
- ② The correlation coefficient R, or the determination coefficient R^2 : under the premise of the same standard deviation (SD_v) of the validation set, the greater the R, the higher the accuracy. The value of R^2 is greatly related to the distribution range (SD_v) of the properties to be measured. For properties with a wide distribution range, the value of R may be close to 1, but its accuracy may be poor.

As shown in Fig. 13.3, for the same RMSEP, the determination coefficient R^2 varies greatly in different concentration ranges. When the concentration range of validation set increases from 13.17% to 10–20%, its determination coefficient R^2 increases from 0.625 to 0.911 [5].

- ③ RPD value: under the premise of the same concentration range, the greater the RPD, the higher the accuracy. It is generally believed that if $\text{RPD} > 5$, it indicates that the prediction result of the model is acceptable. If $\text{RPD} > 8$, the prediction accuracy of the model is very high. If $\text{RPD} < 2$, it indicates that the prediction result is unacceptable.

RPD classification of models for predicting grain chemical composition content (Table 13.1) and for predicting forage, feed, soils, and functionality factors (Table 13.2) [6, 7] have been given in some literature.

In fact, RPD and R^2 are the same evaluation metric (Fig. 13.4), and their relationship is as follows:

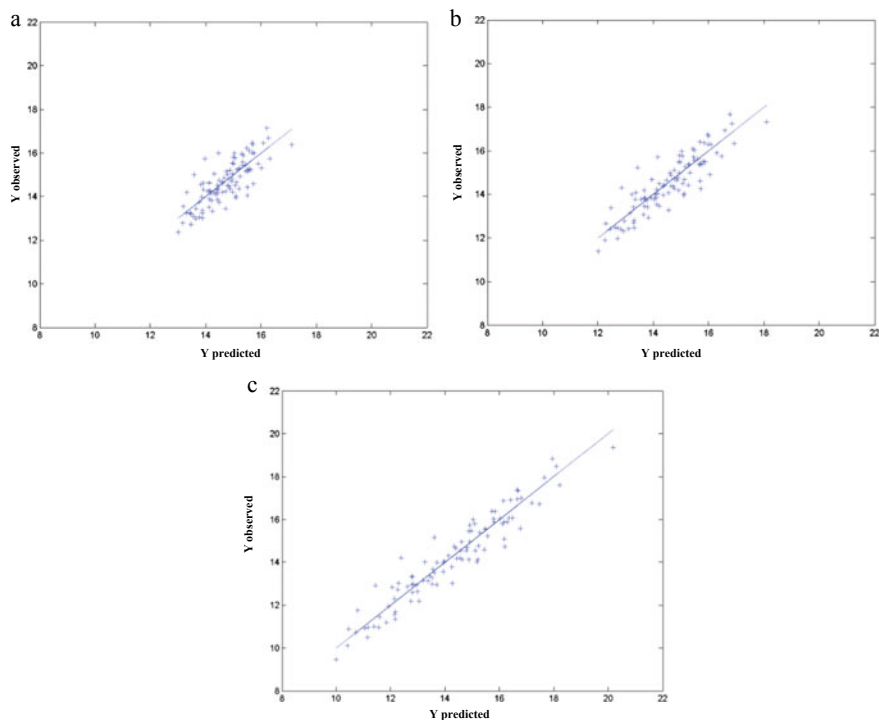


Fig. 13.3 Influence of concentration range on determination coefficient under the same RMSEP a: concentration range 12–18%, $R^2 = 0.786$; b: concentration range 13.17%, $R^2 = 0.625$; c: concentration range 10–20%, $R^2 = 0.911$ [5]

Table 13.1 The RPD statistic for predicting grain chemical composition content

RPD value	Classification	Application
0.0–2.3	Very poor	Not recommended
2.4–3.0	Poor	Rough screening
3.1–4.9	Fair	Screening
5.0–6.4	Good	Quality control
6.5–8.0	Very good	Process control
8.1+	Excellent	Any application

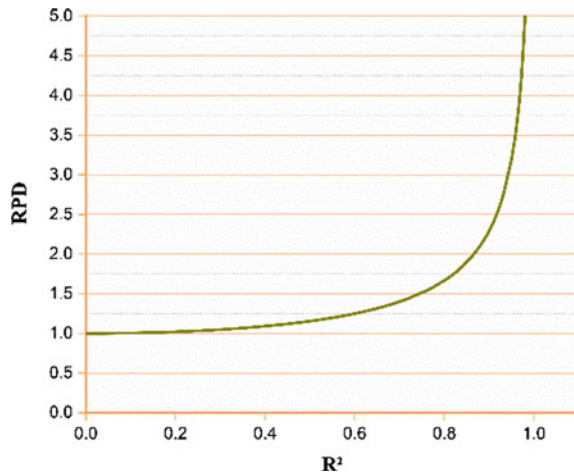
$$\text{RPD} = \frac{1}{\sqrt{1 - R^2}} \quad (13.16)$$

For example, if R^2 of the validation set is 0.90, its RPD is 2.29; If $R^2 = 0.98$, then RPD is 5.0. Therefore, the concentration range also has a significant effect on RPD.

Table 13.2 The RPD statistic for predicting forages, feeds, soils, and functionality factors

RPD value	Classification	Application
0.0–1.9	Very poor	Not recommended
2.0–2.4	Poor	Rough screening
2.5–2.9	Fair	Screening
3.0–3.4	Good	Quality control
3.5–4.0	Very good	Process control
4.1+	Excellent	Any application to this type of material

Fig. 13.4 Relationship between RPD and R^2



It should be noted that in some literature, when calculating RPD, the SD of calibration set is used to replace the SD of validation set. In this case, the above relationship between RPD and R^2 is not valid.

- ④ The *t* test is used to test whether there is a significant difference between the measured values by the spectral method and the reference method. If the *t*-test is passed, it can only show that there is no systematic error between the spectral method and the reference method, but it cannot completely explain the accuracy of its prediction results.

The above five parameters are all based on the evaluation results of statistical test. Another type of accuracy method is the validation method based on a single sample, which investigates whether the absolute deviation between the predicted value of the spectral method and the actual value of the reference method is less than the reproducibility required by the reference method. If 95% of samples in the validation set meet this requirement, the test passes. This validation method of accuracy is more suitable when the precision of the reference method is not uniformly distributed in the concentration or property range of the whole calibration set.

Repeatability: More than five samples are selected from the validation set to verify the repeatability of the spectral method. The concentrations of these samples must cover 95% of the calibration concentration range and be uniformly distributed. At least six times of continuous spectral measurements need to be carried out for each sample, and the samples should be reassembled during spectral collection. The results are calculated with the established calibration model, and the repeatability of the spectral method was evaluated by means of average value, range, and standard deviation. It is generally required that the standard deviation of repeatability of the measured results by the spectral method is not greater than 0.33 times of SEP, that is, the standard deviation of repeatability of the NIR spectral method accounts for about one third of SEP.

Robustness: The robustness of the model refers to its resistance to external interference factors. These factors mainly include the same type of the replacement of test sample device (such as color plate, optical fiber probe, and integral ball), the change of fiber bending degree, the replacement of light and reference material (such as ceramic chip or barium sulfate powder), the change of sample loading condition and temperature (environment temperature and sample temperature), and the physical state of particle (such as moisture content of grain, the polymer particle size, and residual solvent). The robustness of the model can be evaluated by examining the repeatability of samples. For example, when examining the influence of the color plate, you can choose more than one color plate of the same specification (material and optical path), such as the color plate of different manufacturers, as well as the same batch and different batches of the same manufacturer. The robustness can be evaluated by the average value, range, and standard deviation. It is generally required that the standard deviation of robustness of the measured results by the spectral method be no more than 0.5 times that of the SEP.

Transmissibility: the transmissibility of the analysis model mainly depends on the hardware differences between the instrument systems. Its essence is to examine the replaceability of the spectrometer and its key parts (optical systems such as interferometers). The transferability of the analysis model directly affects the generalization ability of the spectral method. If the spectrometer from the same manufacturer does not have the model transferability, it is difficult for users to share the rich model resources. It is possible to use the sample for the investigation of the repeatability to evaluate the transmissibility of the model. For example, multiple spectrometers of the same type are selected and used to collect the spectra of above samples. Prediction analysis of the spectrum of the same sample in a different spectrometer is carried out by the model built on one instrument. The average, range, and standard deviation values are employed to evaluate transmissibility. Typically, there are significant system errors that require spectral calibration, known as calibration transfer, to produce consistent results. However, there are a few instrumental manufacturers whose calibration models can be directly used for the same type of spectrometers without any modification, that is, model data can be directly transmitted, known as calibration transfer. It is usually required that the standard deviation of the transmissibility of the measured results of the spectral method is not greater than 0.7 times

of SEP, that is, the spectrum of the same sample is measured on different instruments, and the same model is used for prediction (before prediction, the spectrum of the secondary instrument can be transformed by calibration transfer). The standard deviation of the predicted results is not greater than 0.7 times of SEP.

13.2 Evaluation of Performance of Pattern Recognition Model

The following parameters are usually used to evaluate the performance of pattern recognition (mainly the supervised method). These parameters can be used to evaluate the performance of the discriminant analysis on both the calibration set (including the cross-validation process) and the validation set samples.

Confusion matrix is commonly used in the evaluation of classification results, which gives the corresponding relationship between the predicted category and the actual category of the sample [8]. For a classification problem of G classes, the confusion matrix is a $G \times G$ matrix as shown in Table 13.3. The rows of the confusion matrix represent the real classes and the columns represent the predicted classes. The element n_{gk} in the matrix means that there are n_{gk} samples of real class g that are predicted to be class k , and the element on the diagonal of the matrix represents the correct number of samples of predicted class. If the prediction class of each sample is correct, the confusion matrix is a diagonal matrix.

Based on the confusion matrix, the following parameters can be calculated:

Correct classification rate can be calculated in the formula (13.17):

$$NER = \frac{\sum_{g=1}^G n_{gg}}{n} \tag{13.17}$$

where n is the number of all samples in the calibration set or validation set.

The misclassification rate ER can be calculated as follows:

Table 13.3 Confusion matrix

Real class	Predicted class					
		1	2	3	...	G
1	n_{11}	n_{12}	n_{13}	...	n_{1G}	
2	n_{21}	n_{22}	n_{23}	...	n_{2G}	
3	n_{31}	n_{32}	n_{33}	...	n_{3G}	
...	
G	n_{G1}	n_{G2}	n_{G3}	...	n_{GG}	

$$ER = 1 - NER \quad (13.18)$$

The classification error rate ER can be compared with the NOMER value which can be calculated as follows:

$$NOMER = \frac{n - n_M}{n} \quad (13.19)$$

where n_M is the number of samples in the class with the largest number of samples in the calibration set, and NOMER represents the error rate of directly classifying the samples into class M without discriminant analysis, which obviously requires that $NER < NOMER$.

The misclassification rate ER can also be compared with the random classification rate (RER), which means the error rate of randomly assigning a sample to a certain class without discriminant analysis. The formula for calculating RER is as follows:

$$RER = \frac{\sum_{g=1}^G \left(\frac{n-n_g}{n} \right) n_g}{n} \quad (13.20)$$

where n_g is the number of samples in class g of the calibration set.

For each class of discriminant results, the following parameters can be used to evaluate. The sensitivity or recall rate can be calculated as follows:

$$Sn_g = \frac{n_{gg}}{n_g} \quad (13.21)$$

where Sn_g represents the ability of the discriminant model to correctly assign g samples to class g .

The precision parameter can be calculated as follows:

$$Pr_g = \frac{n_{gg}}{n'_g} \quad (13.22)$$

where n'_g represents the number of samples predicted for class g . Pr_g represents the ability of the discriminant model to assign only g samples to class g .

The specificity parameter can be calculated as follows:

$$Sp_g = \frac{\sum_{i=1}^G (n'_k - n_{gk})}{n - n_g} \quad (k \neq g) \quad (13.23)$$

Table 13.4 Confusion matrix of the discrimination results based on three classes and 30 samples

Real class	Predicted class				
	A	B	C		
A	9	1	0		10
B	2	8	2		12
C	1	2	5		8
	12	11	7		$n = 30$

where n'_k represents the number of samples predicted for class k . Sp_g represents the ability of the discriminant model to classify non- g class samples as non- g class samples.

Table 13.4 is the confusion matrix of three classes' discriminant results with 30 samples, and Table 13.5 is the statistical parameter value obtained by the evaluation of the discriminant results.

Kappa coefficient can be used to measure the classification effect. The calculation of Kappa coefficient is based on confusion matrix, and the calculation result of Kappa is between -1 and 1 , but usually the Kappa falls between 0 and 1 . It can be divided into five grades to represent the consistency of different classes: extremely low consistency from 0.0 to 0.2 (slight), general consistency from 0.2 to 0.4 (fair), moderate consistency from 0.4 to 0.6 (moderate), high consistency from 0.6 to 0.8 (substantial), almost complete consistency from 0.8 to 1.0 (almost perfect).

The Kappa coefficient (k) is calculated by the following formula:

$$k = \frac{p_o - p_e}{1 - p_e} \tag{13.24}$$

Table 13.5 Statistical parameters obtained by evaluating the discriminant results

Parameter	Value
NER	0.73
ER	0.27
NOMER	0.60
RER	0.66
Sn (A)	0.90
Sn (B)	0.67
Sn (C)	0.63
Sp (A)	0.85
Sp (B)	0.83
Sp (C)	0.91
Pr (A)	0.75
Pr (B)	0.73
Pr (C)	0.71

where p_o is the summary of the number of samples in each class divided by the total number of samples, that is, the overall classification accuracy.

The number of real samples for each class is a_1, a_2, \dots, a_C , respectively, and the predicted number of samples for each class is b_1, b_2, \dots, b_C , respectively. The total number of samples is n , then we have a following value.

$$p_e = \frac{a_1 \times b_1 + a_2 \times b_2 + \dots + a_C \times b_C}{n \times n} \tag{13.25}$$

For the confusion matrix in Table 2.10, the calculation process of its Kappa coefficient is as follows:

$$p_o = \frac{9 + 8 + 5}{30} = 0.7333$$

$$p_e = \frac{10 \times 12 + 12 \times 11 + 8 \times 7}{30 \times 30} = 0.3422$$

$$k = \frac{p_o - p_e}{1 - p_e} = \frac{0.7333 - 0.3422}{1 - 0.3422} = 0.5946$$

The Kappa coefficient (k) can also be calculated by the following formula:

$$k = \frac{n \sum_{i=1}^r (x_{ii}) - \sum_{i=1}^r (x_i + x_{+i})}{n^2 - \sum_{i=1}^r (x_i + x_{+i})} \tag{13.26}$$

where n is the total number of samples, r is the number of rows or columns of the confusion matrix, x_{i+} and x_{+i} represent the sum of all rows and columns respectively, and x_{ii} is the value of the diagonal of the confusion matrix, namely the number of samples correctly classified.

For true and false recognition problems, the confusion matrix can be simplified to a contingency table, as illustrated in Table 13.6.

If a true sample is identified as true, it is defined as TP (True positive); if a true sample is identified as false, it is defined as FN (False positive); if a false sample is identified as false, it is defined as TN (True negative); if a false sample is identified as true, it is defined as FP (false negative). Then the above parameters can be calculated according to the following formula:

Table 13.6 Confusion matrix of discriminant analysis based on two types (contingency table)

	Predicted class			
Real class		True (P)	False (N)	Total samples of real class
	True (P)	TP	FN	TP+FN
	False (N)	FP	TN	FP+TN
	Total samples of predicted class	TP+FP	FN+TN	TP+FP+FN+TN

$$\text{Precision, } \text{NER} = \frac{\text{TP} + \text{TN}}{n} \quad (13.27)$$

$$\text{Misclassification rate, } \text{ER} = \frac{\text{FP} + \text{FN}}{n} \quad (13.28)$$

True positive rate (TPR), also known as sensitivity, can be calculated as follows:

$$\text{TPR} = \text{Sn} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (13.29)$$

False negative rate (FNR), also known as missed diagnosis rate, can be calculated as follows:

$$\text{FNR} = 1 - \text{TPR} \quad (13.30)$$

True negative rate (TNR), also known as specificity, can be calculated as follows:

$$\text{TNR} = \text{Sp} = \frac{\text{TN}}{\text{FP} + \text{TN}} \quad (13.31)$$

Classification efficiency (EFF), also known as G Score in some literature, can be calculated as follows:

$$\text{EFF} = \sqrt{\text{TPR} \times \text{TNR}} \quad (13.32)$$

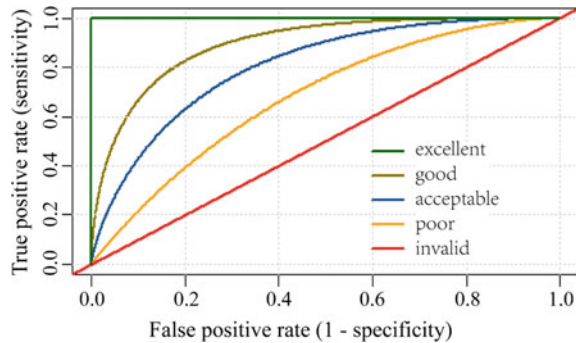
According to the above parameters, when $\text{TPR} = 0$ and $\text{FPR} = 0$, it indicates that all samples are predicted to be false class (negative class); when $\text{TPR} = 1$ and $\text{FPR} = 1$, it indicates that all samples are predicted to be true class (positive class); when $\text{TPR} = 1$ and $\text{FPR} = 0$, it indicates that all samples are correctly classified.

The $F1$ score is an indicator used in statistics to measure the accuracy of the binary classification (or multitasking binary classification) model. It takes both precision and sensitivity of the classification model into account. $F1$ score can be regarded as a weighted average value of model accuracy and sensitivity, with a maximum value of 1 and a minimum value of 0. The greater the value, the better the model. It can be calculated as follows:

$$F1 = \frac{2 \times \text{NER} \times \text{TPR}}{\text{NER} + \text{TPR}} \quad (13.33)$$

As shown in Fig. 13.5, ROC (receiver operating characteristic) curve can be obtained by drawing FPR (1-Specificity, X -axis) and TPR (Sensitivity, Y -axis). The ROC can be used to evaluate the quality of the classification model. A good classification model should be close to the upper left corner of the graph as possible, while a randomly guessed model should be located on its main diagonal. The closer the area under the ROC curve is to 1, the better the classification model is. If it is between

Fig. 13.5 Receiver operating characteristic (ROC) curve and different quality of the classification model based on the area under the ROC curve



0.9 and 1.0, the classification model is excellent. If it is between 0.8 and 0.9, the classification model is good. At 0.7–0.8, the classification model is acceptable; At 0.6–0.7, the classification model is very poor; At 0.5–0.6, the classification model is invalid.

Matthews correlation coefficient (MCC), also known as Phi coefficient, can be used to evaluate the advantages and disadvantages of the two types of discriminant results:

$$\text{MCC} = \varphi = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP} + \text{FN}) \times (\text{TN} + \text{FP}) \times (\text{TP} + \text{FP})(\text{TN} + \text{FN})}} \quad (13.34)$$

When φ is equal to 1, it indicates that the classification is completely correct. When φ is less than 0, it indicates that the classification effect is not as good as a random guess.

References

1. Siesler HW, Ozaki Y, Kawata S, et al. Near-infrared spectroscopy: principles, instruments, applications. Wiley;2001.
2. Liang Y-Z, Xu Q-S. Instrumental analysis of complex systems—white, gray and black analytical systems and their multivariate methods. Beijing: Chemical Industry Press; 2012.
3. Bellon-Maurel V, Fernandez-Ahumada E, Palagos B, et al. Critical review of chemometric indicators commonly used for assessing the quality of the prediction of soil attributes by NIR spectroscopy. *TrAC Trends Anal Chem.* 2010;29:1073–81.
4. Williams P, Dardenne P, Flinn P. Tutorial: Items to be included in a report on a near infrared spectroscopy project. *J Near Infrared Spectrosc.* 2017;25:85–90.
5. Davies AMC, Fearn T. Back to basics: calibration statistics. *Spectrosc Eur.* 2006;18:31–2.
6. Williams P, Antoniszyn J, Manley M. Near infrared technology: getting the best out of light. Stellenbosch: SUN Press; 2019.
7. Williams P. Tutorial: the RPD statistic: a tutorial note. *NIR News.* 2010;21:22–3.
8. Sun D-W. Infrared spectroscopy for food quality analysis and control. Academic Press;2008.

Chapter 14

Methods for Improving Prediction Ability of Model



Model prediction ability mainly refers to the performance of robustness and accuracy. They are unified in some cases but contradictory in others. For example, for liquid samples, the conditions of spectroscopic acquisition (such as temperature and pressure) can be strictly controlled. A quantitative calibration model can be established accordingly. The model has high prediction accuracy for samples within the bounds under the same conditions. However, if the spectral acquisition conditions change moderately, the prediction accuracy of the model will become significantly worse. A hybrid calibration model can be established on the spectra collected under different conditions to improve the robustness (or adaptability) of the model. Yet, the model accuracy would be decreased in this case. In practice, it is often necessary to seek a balance between robustness and accuracy [1].

14.1 Modeling Strategies for Improving the Robustness

There are two ways to improve the robustness of the model. One is to preprocess and select the spectral variables. The other is to establish a hybrid model.

Spectral preprocessing methods, such as derivative, multiplicative scatter correction (MSC), orthogonal signal correction (OSC), and wavelet transform (WT), would be used to eliminate the interference of external conditions on the spectra as much as possible. Wavelength variable selection methods such as genetic algorithm (GA) can select wavelength with strong information and insensitive to external influence factors, so as to establish a robust calibration model.

Another way to achieve the robustness of the analytical model is to establish a hybrid calibration model, also known as the global calibration model, incorporating expected intrinsic changes and external impacts into calibration set. For example, it can realize the robustness of the model to temperature by building a global temperature model with a temperature hybrid calibration set involved samples measured

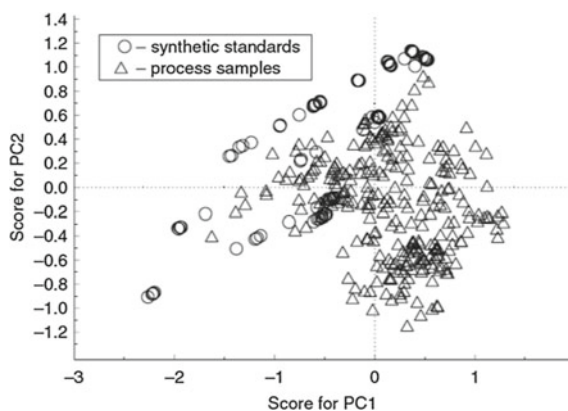
with different temperature. The method is simple and feasible. Other robust analysis models can also be established by adding spiked samples for other influencing factors such as sample types and measurement conditions. For example, Fig. 14.1 shows a hybrid calibration model with strong fitness and robustness is established by forming a calibration set with a wide range of laboratory synthesized samples and narrow distribution samples from the online process [3–6]. Mehdizadeh et al. [7, 8] monitored the cell culture process using online Raman spectra, the samples in calibration set were collected from cultures of different cell strains and cultures, and additional samples prepared from glucose and lactic acid were added to blank cultures to enhance the model robustness. In practice, in order to obtain robust calibration models, hundreds of batches of independently running samples need to be collected for many years [9].

However, in practical operation, nonlinear problems should be paid attention to. For example, wide temperature range or quite different sample types would have nonlinear influences on the spectra. It is difficult to establish a model to meet the accuracy requirements only by linear calibration methods such as PLS, which can be solved by nonlinear calibration methods such as ANN [10].

14.2 Modeling Strategies Based on Local Samples

The modeling strategy based on local samples was proposed in 1988, which is a method to improve the accuracy of modeling. However, due to the instrument hardware platform and other reasons, it has not attracted too much attention, until in recent years. With the continuous standardization of instrument manufacturing, this database and library search-based method is really practical. Contrary to the idea of establishing a global calibration model, the basic idea of the local modeling strategy is to select a group of samples most similar to the unknown samples from the database (i.e., samples in calibration set) based on the spectra (or its derived feature variables),

Fig. 14.1 The first two principal component distributions of the spectra of laboratory synthetic samples and field process samples by PCA



and then obtain the final results from these samples (i.e., the local samples) through statistical analysis or classical calibration methods (Fig. 14.2) [11].

How to select local samples and how to get the final prediction result in many local modeling methods, such as comparison analysis using restructured near infrared and constituent data (CARNAC), locally weighted regression (LWR), and LOCAL.

The CARNAC method processes the spectra using the Fourier coefficient as the characteristic variable for searching for local samples. To ensure the accurate determination of low content components, this method needs to select local samples for different analytical parameters, i.e., select the characteristic Fourier coefficient by gradual multivariate linear regression, and then select local samples according to the similar index s ($s = 1/(1-R^2)$, R is the correlation coefficient between the unknown sample and a sample in the database). The final predictions are given by the underlying data corresponding to the local samples via a similar exponential weighted average method. Davies et al. [12] improved the CARNAC method by replacing Fourier transform with the wavelet transform.

LWR method uses principal component analysis to compress the spectra of database samples and takes the principal component score as the characteristic variable combined with Euclidean or Mahalanobis distance to select local samples, and establishes a calibration model using principal component regression to predict the unknown samples [13]. Subsequently, several improvements are made on the selection and regression method of local samples, such as the calculation and selection of principal components and the calculation of distance [14].

The LOCAL method [15] uses the correlation coefficient between the unknown sample spectra and the database sample spectra to select the local samples and establishes the local calibration model (different principal factor weighting) by partial least squares to predict the unknown samples. The LOCAL has become a method in FOSS WINISI software and has been reported in several applications [16, 17]. TOPNIR method, which has certain applications in the petrochemical field, is also based on local samples. The method has been used for NIR spectral analysis of oil refining products. It selects local samples through neighboring indices of different properties constructed by absorbance of characteristic peaks of different chemical groups, and a weighted average method is adopted to calculate the final results.

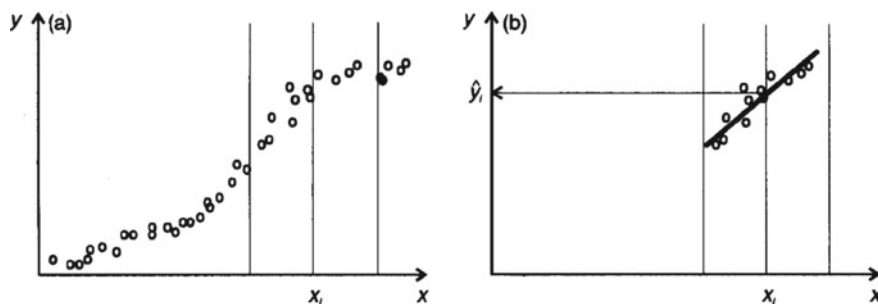


Fig. 14.2 Schematic diagram of the local modeling strategy

Based on the above methods, a variety of methods have emerged for modeling strategies based on local samples. Fearn et al. [18] proposed to select local samples with the predicted concentration values obtained by global regression and the scores obtained by OSC algorithm as characteristic variables. Chung et al. [19] used Fourier moment of the spectra as characteristic variables to select local samples, then established local partial least squares calibration models by using the differential spectra method to identify subtle differences between sample spectra.

He et al. [20] fused wavelength screening with local modeling strategies for the establishment of gasoline blending online NIR spectral models, and local samples were selected by supervised local preserving projection (SLPP) after dimension reduction of the spectra. Zhang and Yan et al. [21, 22] proposed local sample selection methods based on the net signal analysis and the spectral information divergence, respectively, in order to overcome the problems of large difference between calibration samples and nonlinearity between the properties to be measured and the spectra in quantitative analysis.

Based on the idea of just-in-time learning (JITL), Tulsyan et al. [23] selected 100 samples with the smallest sample Euclidean distance from the samples to be tested from 3800 samples for PLS modeling, showing a significant improvement in model performance compared to the use of full samples for modeling. Then, with the help of JITL modeling ideas and Gaussian process regression method, they put forward the automatic real-time modeling strategy to realize the “intelligent” of the model maintenance.

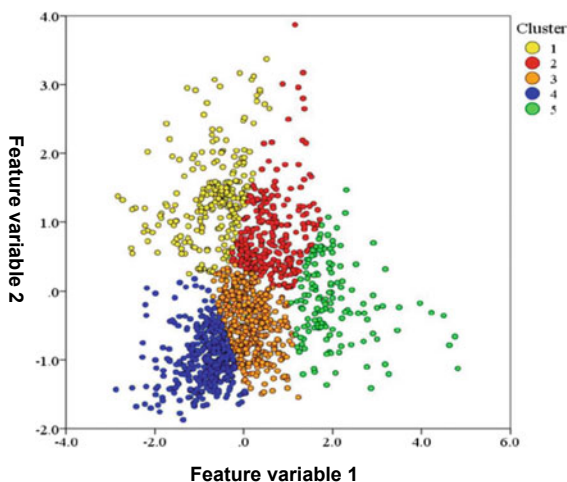
The modeling strategy based on local samples is suitable for the calibration of the nonlinear system. At the same time, it can take full advantage of the database to avoid the disadvantages of the traditional factor analysis methods that need to frequently update the model due to changes such as the sample compositions. However, for specific analysis projects, how to select the local samples most similar to the unknown samples, how many local samples to choose, and how to get the final prediction results still need to be further researched.

A similar idea for establishing the local calibration model is to establish a quantitative calibration model by classification. This method first conducts the samples into cluster analysis (as shown in Fig. 14.3) and divides the calibration set samples into multiple classifications, and then establishes the quantitative calibration model on each classification [24–26]. For the samples to be predicted, the classification is first judged based on the spectra, and then a quantitative calibration model of the corresponding classification is selected to predict the results.

14.3 Ensemble Modeling Strategies

Traditional multivariate calibration techniques such as PLS and ANN generally employ a single model, namely, establishing an optimal model by using established training sets for predictive analysis. However, when the number of samples

Fig. 14.3 Schematic diagram of cluster analysis for the samples in calibration set



in training set is limited or the calibration method is unstable, the prediction accuracy and stability of the model are often not satisfactory. The basic idea of ensemble or consensus strategy is to build multiple models (member models) in a random or combined manner with the simultaneous prediction of the different subsets of the same training set, making multiple predictions by simple or weighted average. It is characterized by reducing the dependence of the prediction results on a certain (or some) sample by repeatedly using the information in the training set, thus improving the prediction stability of the model.

Ensemble strategies were first applied to pattern recognition classification problems, especially some relatively unstable algorithms such as ANN, which have gradually attracted the attention in spectral analysis in recent years. It is combined with multiple algorithms such as PLS, SVM, and ANN to establish quantitative calibration models. The selection of member model samples in ensemble modeling is crucial. Bootstrap Aggregating (bagging) and boosting are the two main methods [27, 28].

14.3.1 Bagging Ensemble Strategy

In the classical bagging method, the sample selection adopts the bootstrap method. The sample size of the randomly selected member training set is the same as that of the original training set, but the sampling method is taken back. Thus, some samples in the original training set may occur multiple times in the member training set, while others may not appear once. The bagging method increases the divergence of model integration by re-selecting the training set to improve the generalization abilities. Stability is the key factor in whether bagging can play a role. Bagging can improve the prediction accuracy of the unstable calibration algorithm, but it does not work significantly on the stable calibration algorithm, and sometimes even

reduces the prediction accuracy. The stability of the calibration algorithm is that the calibration results will not change greatly if the training set has small changes. For the final predictions, the classical bagging method takes a simple averaging approach. Figure 14.4 shows the modeling strategy schematic of bagging combined with PLS.

Galvao et al. [29–31] improved the classical bagging method, such as using subbagging for sample selection, evaluation selection of member models, and prediction with weights. The composition or properties of soil, tobacco, and corn determined by near infrared (NIR) spectroscopy were modeled and verified using this method, and satisfactory results were obtained.

14.3.2 Boosting Ensemble Strategy

Boosting was first proposed by Schapire in 1990. In 1995, Freund and Schapire [32] improved the boosting algorithm and proposed boosting (adaptive boosting, AdaBoost) algorithm that can be very easily applied to practical problems. Therefore, the algorithm has become the most popular boosting algorithm at present.

The difference between boosting and bagging is that the selection of member training set for bagging is random, and each member training set is independent of each other. For boosting, the selection of member training set is not independent, which relates to the learning results of the previous iteration. Therefore, the predictive functions of bagging can be generated in parallel, while the prediction functions of boosting can only be generated sequentially and weighted.

The basic idea of the AdaBoost algorithm applied to classification is to gradually construct a set of classifiers. Each new classifier focuses on compensating the defects of the previous classifier, and finally integrates the classification results of all classifiers to achieve more ideal classification results. Zhang and Drucker et al. [33, 34] modified the boosting algorithm to solve the regression problem. The boosting regression algorithm proposed by Drucker which is commonly used in the literature is described below.

The boosting regression algorithm proposed by Drucker et al. was to produce a set of basic member models through an iterative process as shown in Fig. 14.5. Given the training set and the learning algorithm, first gave equal weights to each training sample, normalized to get the first sampling probability distribution P_1 of the training set. The sampling generated member training set 1. The learning algorithm was used to establish member regression model h_1 for member training set 1. Then, the weight of the sample was corrected according to the error generated by the member regression model h_1 on each sample, the weight of the sample with large error was increased, thus increasing its sampling probability. After normalization, the sampling probability distribution P_2 of the training set was obtained. Member training set 2 was generated by sampling, and member regression model h_2 was trained by learning algorithm on member training set 2. After that, the sample weight

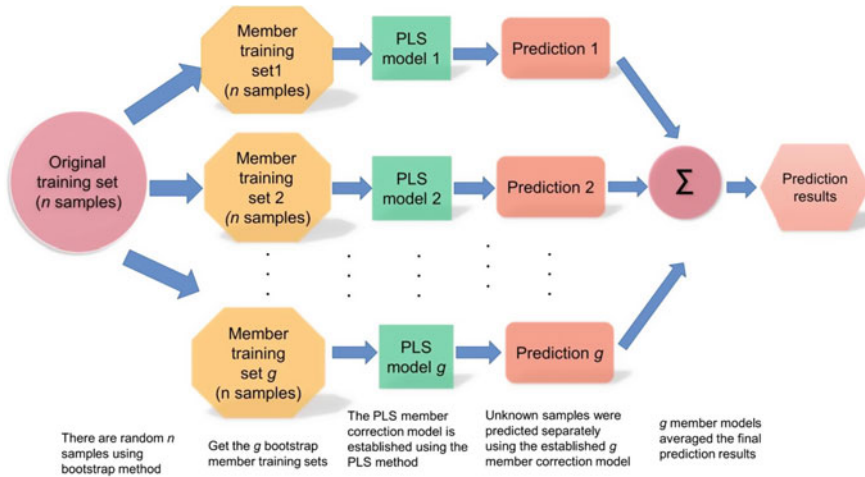


Fig. 14.4 Schematic diagram of modeling strategy for bagging combined with PLS method

was further adjusted according to the error, which was repeated to obtain a set of gradually modified member regression models $h_1, h_2, h_3, \dots, h_g$.

The implementation steps of the boosting regression algorithm are as follows.

For the original training set $\{(x_i, y_i), i = 1, \dots, n\}$ (n is the number of samples of the original training set), first given the basic learning algorithm (e.g., PLS, SVR or ANN, etc.) [35]. The maximum iteration number of boosting is g , and sample weights are initialized by Eq. 14.1.

$$\omega_i^{(1)} = 1/nv_i = 1v_1v_2\dots v_n \tag{14.1}$$

Take the number of iteration $t = 1, \dots, g$, repeat the following steps (1)–(7).

- (1) The sampling probability of each sample of the original training set is calculated using Eq. 14.2.

$$p_i^{(t)} = \omega_i^{(t)} / \sum_{j=1}^n \omega_j^{(t)} \tag{14.2}$$

Then part of the samples from the calibration set for member training set of the t iteration are picked up by roulette (allowing the repeated sampling).

- (2) Build the member regression model h_t by basic learning algorithm based on n samples in member training set of t iteration.
- (3) Predict each sample in the original training set by using the member regression model h_t . The predictive values for each sample $\hat{y}_i^{(t)}, i = 1, \dots, n$, can be obtained.

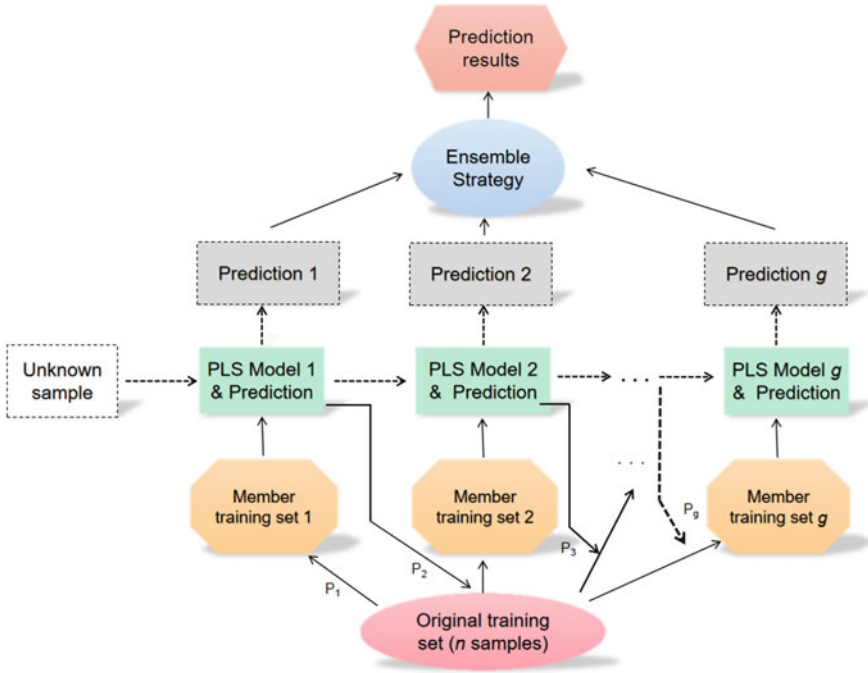


Fig. 14.5 Schematic diagram of modeling strategy for boosting combined with PLS method

- (4) Calculate the errors for each sample in the original training set by using Eq. 14.3.

$$L_i^{(t)} = \frac{|\hat{y}_i^{(t)} - y_i|}{\max |\hat{y}_i^{(t)} - y_i|}, i = 1, 2, \dots, n \tag{14.3}$$

- (5) Calculate the sum of weighted error for t iteration.

$$\bar{L}_t = \sum_{i=1}^n L_i P_i^{(t)} \tag{14.4}$$

- (6) Calculate the confidence indicator β_t

$$\beta_t = \frac{\bar{L}_t}{1 - \bar{L}_t} \tag{14.5}$$

- (7) Calculate the new sampling weights for the sample i in $t + 1$ iteration.

$$\omega_i^{(t+1)} = \omega_i^{(t)} \beta_i^{(1-L_i^{(t)})} \quad (14.6)$$

For an unknown sample, g predictive values are calculated by using the g member model, respectively. The final results are obtained by fusion calculations.

Drucker et al. calculated the final result by weighted median. First, g predictive values were sorted in ascending order.

$y^{(k_1)} \leq y^{(k_2)} \leq \dots \leq y^{(k_g)}$, k_g is 1, 2, ..., g to rearrange. Then the predictive value of the k_r member model corresponding to the following minimum r was the final predictive value. Then, the r th prediction that satisfies Eq. 14.7 is taken as the final result of the sample.

$$\sum_{t=1}^r \log(1/\beta_{kt}) \geq \frac{1}{2} \sum_{t=1}^g \log(1/\beta_{kt}) \quad (14.7)$$

In order to prevent over-fitting during the boosting PLS iteration, Wu et al. [36] proposed a new iterative stop criterion and was used to model the rapid prediction of total organic carbon (TOC) in water quality by the UV-visible spectroscopy. To reduce the influence of outliers in calibration set, Shao et al. [37] improved boosting PLS algorithm by adjusting sampling weight and defined a new loss function, which could obtain robust prediction results. Chen et al. [38] combined the variables with boosting PLS to further improve the NIR spectral quantitative prediction ability for ethanol precipitation process of *Lonicera japonica*.

14.3.3 Stacked Ensemble Strategy

Both boosting and bagging ensemble modeling strategies are based on sample selection from calibration set. Ensemble modeling strategies can also be based on wavelength selection. The wavelength-based ensemble modeling strategy is to select multiple different wavelength ranges (sub-feature) from the spectral matrix of the training set according to some rules for establishing member models. The sub-feature matrix used by different member models can be overlying or completely independent [39].

The stacked interval partial least squares (SPLS) method is a PLS model built by weight fusion on different spectral intervals [40–42]. The method divides the spectra into k intervals and establishes a PLS model for each spectral interval. The standard error of cross validation (SECV) obtained during the modeling process is used for calculating the fusion weight ω . The i spectral interval weight ω_i is calculated as follows:

$$\omega_i = \frac{S_i^2}{\sum_{i=1}^k S_i^2} \quad (14.8)$$

In Eq. 14.8, S_i is the reciprocal of $SECV_i$ for the model built by the i spectral interval, where $i = 1, 2, \dots, k$, k is the number of divided spectral intervals.

The concentration value y of the predicted sample spectrum \mathbf{x} , is calculated by the following equation.

$$y = \sum_{i=1}^k \mathbf{x}_i \omega_i \mathbf{b}_{i,PLS} \tag{14.9}$$

where \mathbf{x}_i is the spectrum of the i spectral interval and $\mathbf{b}_{i,PLS}$ is the PLS regression coefficient of the i spectral interval.

Using this idea, the strategy of the moving window can replace the interval spectrum. On the spectrum of a certain window width, a series of PLS models can be established through the movement of the window, and then the final integrated PLS model can be obtained by weighted fusion [43, 44].

Based on the stack modeling strategy, the dual stacked interval partial least squares (DSPLS) method was proposed by Bi et al. [45] As shown in Fig. 14.6, the method includes two steps, i.e., inner-stack step and outer-stack step. In the inner-stack process, the spectrum is divided into n intervals, establish PLS model in interval 1, interval 1–2, interval 1–3, ... and 1– n , respectively, and then stack into n sub-model. In the outer-stack process, the n sub-model is weighted and fused to obtain the final model.

The stack modeling strategy can be used in other quantitative calibration methods such as stack extreme learning machine algorithms [46–48]. Figure 14.7 shows the basic framework of stacked ensemble extreme learning machine (SE-ELM) model. It is actually an application of ELM in the frame of stacked generalization [47].

In quantitative multivariate calibration, in addition to the above ensemble modeling strategy, there is also an ensemble modeling strategy for adding noise based

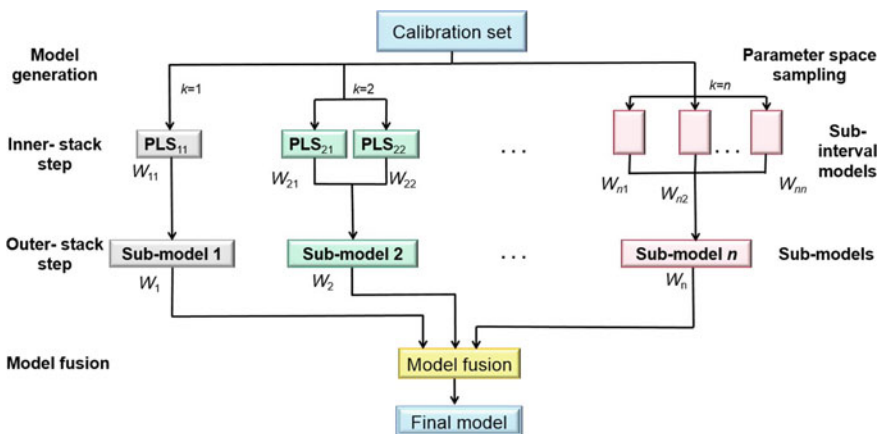


Fig. 14.6 Schematic diagram of ensemble strategy for dual stacked PLS

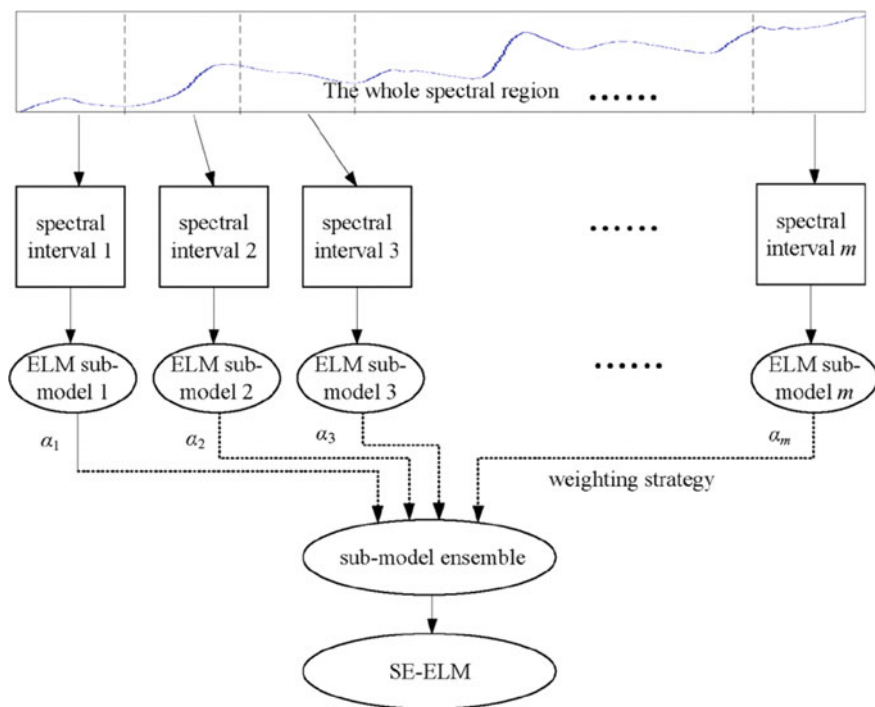


Fig. 14.7 Schematic diagram of stacked ensemble extreme learning machine [47]

on variables, which artificially adds noise to the spectral matrix \mathbf{X} or concentration vector \mathbf{y} in the training set to form multiple member training sets containing noise, so as to enhance the robustness of the member model [49–51]. In addition, there are ensemble modeling methods based on different data preprocessing (Fig. 14.8) and calibration algorithms [52–55], as well as ensemble methods combined with concentration classification and wavelength selection [56, 57].

14.3.4 Stacked Generalization Strategy

Similar to quantitative calibration, the strategy of ensemble modeling is also adopted in the field of pattern recognition, that is, the multiple classifier systems (MSC). The classification performance can be better than any single classifier by the selection and combination of base classifiers.

The construction of multi-classification systems mainly adopts parallel structure, parallel training with multiple classifiers, and then combines the results with some selections and weight strategies. For example, random forest is based on bagging generating multiple different sub-sample sets from the original sample set. It trains

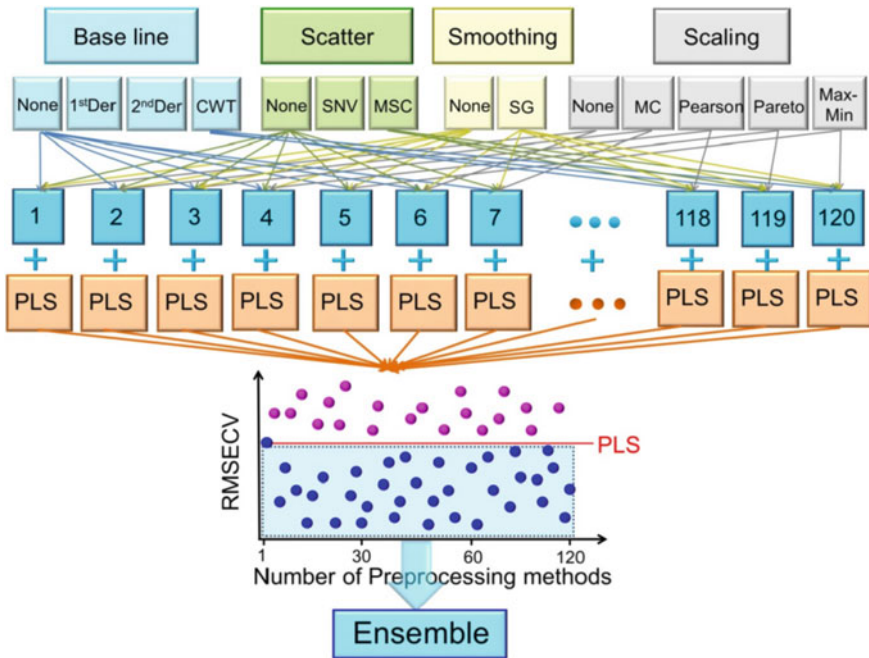


Fig. 14.8 Schematic diagram of ensemble modeling strategy based on different preprocessing methods

the binary decision tree to build classifiers using classification and regression trees (CART) algorithm. Then the classification results of each classifier adopt the method of majority voting to get the final results. Adaptive boosting (Adaboost) is the more representative algorithm in the boosting. It is achieved by changing the data distribution. The weight of each sample is determined according to whether the classification of each sample in each training set is correct and the accuracy of the last overall classification. The new data set with modified weights is sent to the lower classifier for training. Finally, integrate the classifier obtained from each training as the final decision classifier. In addition, there are other boosting algorithms such as gradient boosting decision tree (GBDT), extreme gradient boosting (XGBoost), and light gradient boosting machine (LightGBM). Adaboost, GBDT, XGBoost, and LightGBM can also be used for the integration of regression models [58–62].

Unlike bagging and boosting, stacked generalization, also known as stacking learning, is a multi-level integrated learning system with serial structures. To better describe the multi-level processing process of stack generalization, stack generalization introduces the base classifier and meta classifier concepts where base classifiers are trained using the original features with their output as secondary new features, while the meta classifier will retrain the secondary features and form the final judgment classifier [62–64]. As shown in Fig. 14.9, the stacked generalization structure framework is mainly divided into two levels: Level-0 and Level-1.

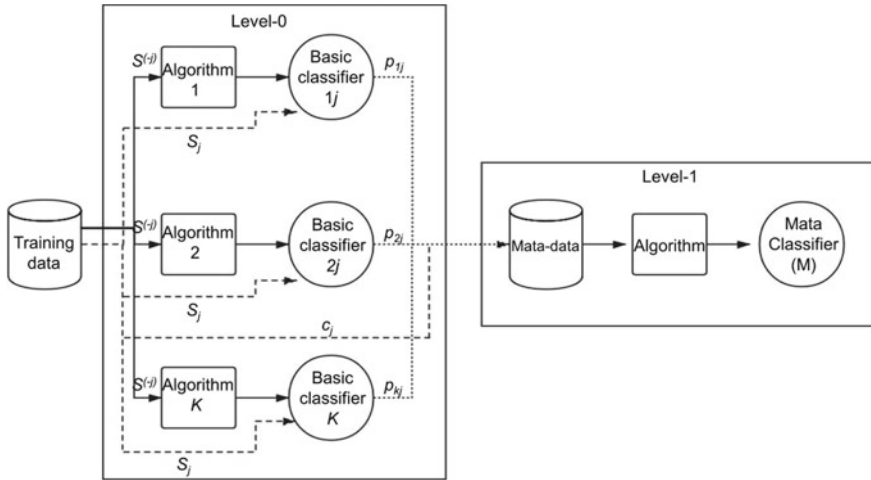


Fig. 14.9 Framework of stacked generalization algorithm

Using each basic classifier at the Level-0, the training set samples are trained and predicted by multiple-fold interactive verification to obtain various posterior probabilities of each training sample. If the training sample has N classifications, then each basic classifier will produce N new feature dimensions p_{kj} composed of posterior probabilities. K classifiers will constitute $K \times N$ new dimensions. These new feature dimensions will serve as training data in Level-1, called metadata. In the Level-0 stage, metadata is determined by the prediction and judgment of the base classifier on the original training set. Compared with the original spectral feature set, it belongs to strong features and also achieves the function of dimension reduction. In fact, this process can be considered as an efficient “dimensionality reduction” operation, which involves feature transformation of the original spectra to obtain new features composed of posterior-like probabilities. In the Level-1 stage, the meta classifier is trained by using new features, and any classifiers can be used as the meta classifier. Finally, a meta classifier model is obtained for the final classification judgment of the samples [63, 65].

14.4 Virtual Sample Modeling Strategy

Sufficient training samples are an important guarantee for improving the prediction accuracy and robustness of multivariate quantitative and qualitative models. The limited number and loosely distributed samples cannot fully describe the whole feature space, and there is an obvious information gap between the samples, which deteriorates the representation of the overall characteristics of small samples. Therefore, the conclusion of modeling directly using small sample data is one-sided and

biased. However, obtaining sufficient training samples usually consumes a lot of manpower and material resources. Therefore, how to improve the prediction ability of the model under a small amount of training data has become a topic worthy of research.

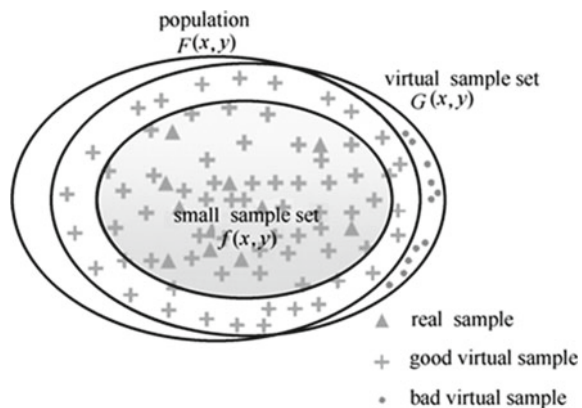
In order to improve the learning ability of small sample problems, semi-supervised learning strategies have been proposed in recent years. A common feature of these methods is that a large number of unlabeled samples are needed as auxiliary samples for learning, and this requirement is also difficult to meet in many cases. Therefore, in 1992, Poggio and Vetter [66] proposed the idea of virtual samples. Virtual samples can also be named by synthetic sample or artificial sample. It refers to the production of some reasonable samples in the sample space to be studied by using the prior knowledge of the research field and combining the existing training samples under the condition of unknown sample probability distribution function. Thus, they can be added to the original training sample set to expand the training sample set and improve the prediction ability of the model (Fig. 14.10).

So far, virtual sample generation (VSG) can be divided into the following three categories [67]

- (1) Construct virtual samples based on specific prior knowledge in the research field.
- (2) Construct virtual samples based on the idea of perturbation.
- (3) Construct virtual samples based on the distribution function in the research field.

Li et al. [68] proposed the mega-trend-diffusion (MTD) method. The method defined the global boundary of information diffusion, calculated the left boundary and right boundary of the corresponding virtual sample information by membership function, so as to generate the virtual sample information in this range. The sample information is expanded more evenly in the global boundary field, thus improving the prediction performance of the model. Zhu et al. [69] proposed multi-distribution mega-trend-diffusion (MD-MTD) method for further optimizing

Fig. 14.10 Relationship among population, small sample set and virtual sample set



the problems of virtual sample boundary establishment and sample screening. It improved the quality of generating virtual samples by multi-segmentation of the sample distribution areas.

Based on MTD, Gao et al. [70] proposed an improved multi-distribution mega-trend-diffusion method advanced-MTD (AD-MTD) to improve the balance of information distribution in diffusion regions. On this basis, the hybrid-MTD was further applied to virtual samples generated by MD-MTD and AD-MTD to improve the balance of information diffusion distribution near the boundary point of the information diffusion area and the center point of the original information area. This method effectively improved the prediction accuracy of PLS regression model for predicting total cholesterol and triglyceride content in blood by infrared spectroscopy. Gong et al. [71] proposed a new method for virtual sample generation based on Monte Carlo method and particle swarm algorithm, which could improve the prediction performance of extreme learning machine.

Aiming at the problem of insufficient NMR spectral data in the regression prediction of total hydrogen physical properties of crude oil, Yi et al. [72] generated virtual samples by adding random noise to the original spectra and established a model for predicting total hydrogen content of crude oil by using convolution neural network. It can not only solve the over-fitting problem in original data training, but also has more stability and accuracy than traditional PLS method. Ye et al. [73] produced virtual samples according to the specified proportion and proposed an automatic densification modeling method for spectral analysis. Li et al. [74] used Monte Carlo method to generate virtual spectra to densify the local database and predicted the chemical composition of oil according to the virtual spectra consistent with the sample to be predicted. The accuracy was higher than that of PLS method. Aiming at the problem of large classification error of heavy metal pollution caused by the high variability of soil heavy metal content and unbalanced samples in mining area, Qian et al. [75] used synthetic minority oversampling technique (SMOTE) to generate virtual samples to balance each pollution level sample, and then used random forest to regress and classify Cd and Pb, the classification accuracy of soil heavy metal Cd and Pb pollution was greatly improved compared with the original sample.

In recent years, with the research and application of deep learning in the field of data driven, generative adversarial networks (GAN) and transfer learning (TL) are increasingly used to generate virtual samples [76–78].

14.5 Semi-supervised Learning Methods

It is usually difficult, expensive, and time-consuming to obtain the label value of the samples. To solve this problem, semi-supervised learning (SSL) algorithms were developed. It uses a large number of unlabeled data and labeled data to train together to construct a better classifier or regression model (Fig. 14.11) [79, 80]. Semi-supervised learning can be used not only for classification and regression, but also for clustering and dimensionality reduction. Semi-supervised learning is a new research hot spot in

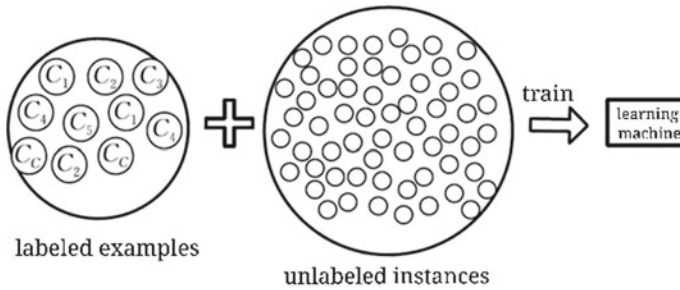


Fig. 14.11 Schematic diagram of semi-supervised learning

the field of machine learning. It is a learning method between supervised learning and unsupervised learning. The learning samples include both labeled and unlabeled classification samples. Under the guidance of supervised information provided by labeled classification samples, unlabeled samples are handled. Semi-supervised learning is based on the assumption that the unlabeled data and the labeled data of the same kind have a certain distance in the feature space. It only needs to provide a small number of labeled samples, and through the learning of all samples can obtain better learning effect than unsupervised learning.

The fundamental reason why unlabeled samples play a role in learner modeling is that they and labeled samples are independently and uniformly sampled from the same data source. In semi-supervised learning, using the gain information of unlabeled samples is mainly based on smoothing assumption, clustering assumption and manifold assumption. The essence of these assumptions is that similar samples have similar outputs. In recent years, semi-supervised learning has been increasingly combined with ensemble learning (consensus learning) to improve the generalization performance of classifiers [81].

Semi-supervised learning algorithms include [59] the following five types.

- (1) Generate semi-supervised models algorithm is used to cluster both labeled and unlabeled data sets, and then determines the labels of the whole cluster by any labeled data contained in each classification in the clustering results.
- (2) Self-training algorithm. Firstly, the labeled data is trained to obtain a classifier, which is used to classify the unlabeled data. According to the classification results, the unlabeled data with high credibility and their predictive markers are added to the training set, the scope of the training set is expanded, and the new classifier is obtained by relearning.
- (3) Joint (or collaborative) training algorithm (co-training). This kind of algorithm implicitly uses clustering assumption or manifold assumption. Firstly, the labeled data is divided into two different data sets, and then two classifiers are trained according to these two different data sets. Each classifier is used to classify unlabeled data sets. The samples with high confidence are selected and added to the training set of another model to continue training (Fig. 14.12).

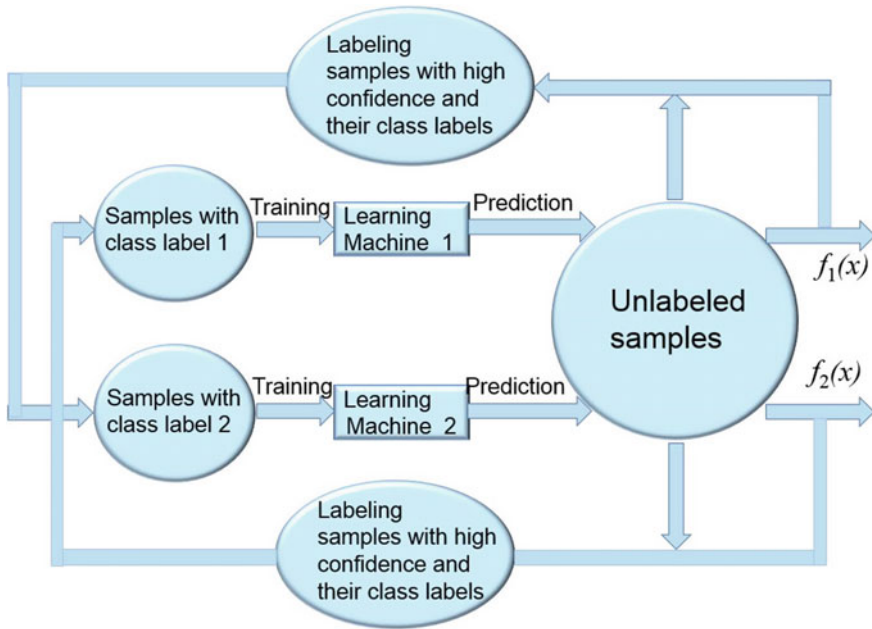


Fig. 14.12 Schematic diagram of co-training

- (4) Semi-supervised support vector machines (S3VM) is evolved from transductive support vector machines (TSVM). The S3VM algorithm uses both labeled and unlabeled data to find a classification surface with the largest class spacing. The algorithm uses the assumption of low-density segmentation, that is, the sample of the boundary region between two different classes is sparse. The classification boundary is located in the low-density region of the sample space.
- (5) Graph-based algorithm is a semi-supervised learning algorithm based on the graph regularization framework. This kind of algorithms directly or indirectly uses the manifold assumption. They usually first establish a graph according to the training example and a certain similarity measure. The vertex of the graph is a labeled or unlabeled sample, and the weight of the edge is the similarity among the samples. Then, the required optimization target function is defined, and the smoothness of the decision function on the graph is used as the regularization term to obtain the optimal model parameters. The core idea of this algorithm is that two samples are similar in the manifold, and their predicted label values are similar.

Most semi-supervised algorithms are used for classification problems, while co-training regressors (COREG) achieve semi-supervised regression in a relatively simple way [82]. The basic idea of the COREG algorithm is: in the training process, the regressor h_1 and h_2 select the data from the labeled data according to the nearest neighbor of K , and then select the samples with the highest confidence for labeling,

and add the labeled data to the regressor of the other party for learning, so as to achieve the purpose of collaborative training. The final predictive value is the average value of the updated regressor h_1 and h_2 .

Li et al. [83] proposed a semi-supervised least squares support vector regression machine which could simultaneously use chemical and non-chemical sample data. Its prediction accuracy was better than the traditional least squares support vector regression. Lv et al. [84] proposed an incremental semi-supervised support vector regression algorithm. Firstly, the incremental semi-supervised support vector regression model was established. The nearest neighbor algorithm was used to select the data with high confidence for collaborative labeling, and the support vector regression model was updated according to whether the labeled data could become a potential support vector [85]. Based on semi-supervised self-training algorithm, Liang et al. [86] proposed semi-supervised partial least squares (SS-PLS) method to optimize the sensory evaluation model of tobacco leaves predicted by NIR spectroscopy. The performance of the model was significantly improved compared with the original model. Guo et al. [87] proposed a method for updating the NIR prediction model of apple soluble solid content based on distance measurement and semi-supervised learning, which significantly improved the prediction ability of the model. Jing et al. [88] applied semi-supervised learning to extreme learning machine and proposed a semi-supervised extreme learning machine classification model for NIR spectroscopy classification of drugs and hybrid seeds. This method showed excellent performance in dealing with unbalanced data sets.

14.6 Multi-target Regression Strategy

Multi-target regression (MTR) is a regression analysis method for the simultaneous prediction of multiple interconnected continuous target variables, which is similar to the multi-label classification problem in pattern recognition. It improves the accuracy of prediction by mining and utilizing the correlation between multiple target variables [89].

In spectral quantitative analysis, the multi-target regression strategy uses the correlation between the variables of the target (concentration or physical property Y) to improve the prediction ability of the model. The most commonly used method is the stacked single-target (SST), also known as the multi-target regressor stack (MTRS) [90]. As shown in Fig. 14.13, this method is divided into two steps. Firstly, the single-target prediction model is established by using the traditional method. Then, the sample input variable space (X) is expanded by using the concentration prediction value, and the prediction model of each target is established. With the addition of multiple target prediction values, the predictive values of each target variable are dependent on the predictive values of other target variables, which makes it possible to improve the prediction ability of the model by using the correlation between target variables.

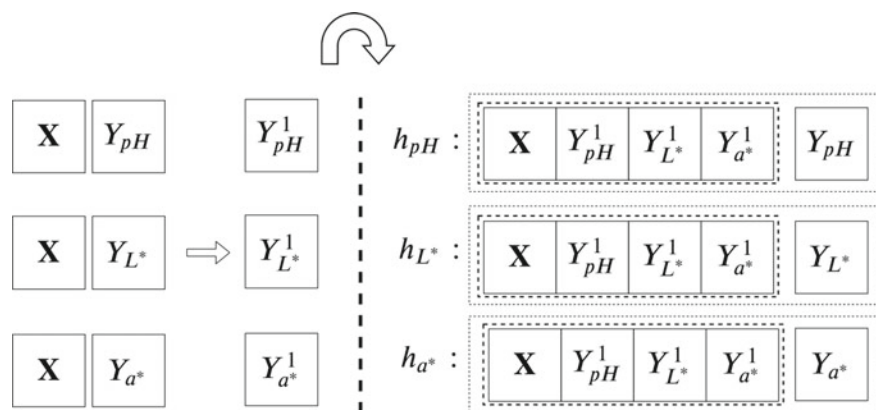


Fig. 14.13 Schematic diagram of modeling strategy for multi-target regressor stack (MTRA)

The ensemble of regressor chains (ERC) method is another multi-target regression strategy, which considers not only the dependence between target variables, but also the order between target variables. In addition, multi-target regression strategies also include multi-target SVR with max-correlation chain (SVRCC) and multi-target regression via target specific features (MTR-TSF) [91], etc.

Santana et al. [92] proposed a multi-target regressor stack (MTRA) method to predict multiple attributes of poultry breast muscle by NIR spectroscopy, including color attributes, pH value, chemical composition, water holding capacity, cooking loss, and tenderness, which improved the prediction ability of the model. Junior et al. [93] used multi-target regression strategy to establish a model for predicting Hectolitre weight, falling number, protein content, alveographic indexes and Farinograph stability of flour by NIR spectroscopy, and the prediction accuracy was improved by 7%.

References

1. Tulsyan A, Schorner G, Khodabandehlou H, et al. A machine-learning approach to calibrate generic raman models for real-time monitoring of cell culture processes. *Biotechnol Bioeng.* 2019;116(10):2575–86.
2. Chu XL, Yuan HF, Wang YB, et al. Developing robust near infrared calibration models. *Spectroscopy Spectral Anal.* 2004;24(6):666–71.
3. Hetrick E, Shi ZQ, Barnes L, et al. Development of near infrared (NIR) spectroscopy-based process monitoring methodology for pharmaceutical continuous manufacturing using an offline calibration approach. *Anal Chem.* 2017;89:9175–83.
4. Bakeev KA. *Process analytical technology: spectroscopic tools and implementation strategies for the chemical and pharmaceutical industries.* Oxford: Blackwell Publishing; 2005.
5. Blanco M, Coello J, Iturriaga H, et al. Strategies for constructing the calibration set in the determination of active principles in pharmaceuticals by near infrared diffuse reflectance spectrometry. *Analyst.* 1997;122:761–5.

6. Farrell JA, Higgins K, Kalivas JH. Updating a near-infrared multivariate calibration model formed with lab-prepared pharmaceutical tablet types to new tablet types in full production. *J Pharm Biomed Anal.* 2012;61:114–21.
7. Mehdizadeh H, Lauri D, Karry KM, et al. Generic raman-based calibration models enabling real-time monitoring of cell culture bioreactors. *Biotechnol Prog.* 2015;31(4):1004–13.
8. Santos RM, Kessler JM, Salou P, et al. Monitoring mAb cultivations with in-situ Raman spectroscopy: the influence of spectral selectivity on calibration models and industrial use as reliable PAT tool. *Biotechnol Prog.* 2018;34(3):659–70.
9. Zhang S, Xiong H, Zhou L, et al. Development and validation of in-line near-infrared spectroscopy based analytical method for commercial production of a botanical drug product. *J Pharm Biomed Anal.* 2019;174:674–82.
10. Shenk JS, Westerhaus MO. Near infrared reflectance analysis with single and multiproduct calibrations. *Crop Sci.* 1993;33:582–4.
11. Luo X, Ye ZZ, Xu HR, et al. Robustness improvement of NIR-based determination of soluble solids in apple fruit by local calibration. *Postharvest Biol Technol.* 2018;139:82–90.
12. Davies AMC, Fearn T. Quantitative analysis via near infrared databases: comparison analysis using restructured near infrared and constituent data-deux (CARNAC-D). *J Near Infrared Spectrosc.* 2006;14(6):403–11.
13. Næs T, Isaksson T, Kowalski BR. Locally weighted regression and scatter correction for near-infrared reflectance data. *Anal Chem.* 1990;62(7):664–73.
14. Centner V, Massart DL. Optimization in locally weighted regression. *Anal Chem.* 1998;70(19):4206–11.
15. Shenk JS, Westerhaus MO. Investigation of a LOCAL calibration procedure for near infrared instruments. *J Near Infrared Spectrosc.* 1997;5(4):223–32.
16. Damberg RG, Cozzolino D, Cynkar WU, et al. The determination of red grape quality parameters using the LOCAL algorithm. *J Near Infrared Spectrosc.* 2006;14(2):71–9.
17. Perez-Marin D, Garrido-Varo A, Guerrero JE. Implementation of LOCAL algorithm with near-infrared spectroscopy for compliance assurance in compound feeding stuffs. *Appl Spectrosc.* 2005;59(1):69–77.
18. Fearn T, Davies AMC. Locally-biased regression. *J Near Infrared Spectrosc.* 2003;11(6):467–78.
19. Chung H, Cho S, Toyoda Y, et al. Moment combined partial least squares (MC-PLS) as an improved quantitative calibration method: application to the analyses of petroleum and petrochemical products. *Analyst.* 2006;131(5):684–91.
20. He KX, Cheng H, Du WL, et al. Online updating of NIR model and its industrial application via adaptive wavelength selection and local regression strategy. *Chemom Intell Lab Syst.* 2014;134:79–88.
21. Zhang HG, Lu JG. Local regression algorithm based on net analyte signal and its application in near infrared spectral analysis. *Spectroscopy Spectral Anal.* 2016;36(2):384–7.
22. Yan Y, Zhang HG, Lu JG, et al. Spectral-information-divergence based local PLS modeling algorithm in near infrared spectroscopy. *Comput Appl Chem.* 2017;34(5):18–22.
23. Tulsyan A, Wang T, Schorner G, et al. Automatic real-time calibration, assessment, and maintenance of generic raman models for online monitoring of cell culture processes. *Biotechnol Bioeng.* 2019;117(2):406–16.
24. Chu XL, Yuan HF, Lu WZ. Determining four component contents in residues by partial least squares-ultraviolet-visible spectrophotometry. *Chin J Anal Chem.* 2000;28(12):1457–61.
25. Xu Y, Wu JZ, Wang YM, et al. Clustering method of unknown sort samples based on near infrared spectroscopy. *Trans Chinese Soc Agricult Eng.* 2011;27(8):345–9.
26. Ogen Y, Zaluda J, Francos N, et al. Cluster-based spectral models for a robust assessment of soil properties. *Geoderma.* 2019;340:175–84.
27. Fearn T. Bagging NIR news. 2006;17(8):15.
28. Boosting FT. NIR news. 2007;18(1):11–2.
29. Galvao RKH, Araujo MCU, Martins MD, et al. An application of subagging for the improvement of prediction accuracy of multivariate calibration models. *Chemom Intell Lab Syst.* 2006;81(1):60–7.

30. Viscarra Rossel RA. Robust modelling of soil diffuse reflectance spectra by bagging-partial least squares regression. *J Near Infrared Spectrosc.* 2007;15(1):39–47.
31. Li YK, Shao XG, Cai WS. Partial least squares regression method based on consensus modeling for quantitative analysis of near-infrared spectra. *Chem J Chinese Univ.* 2007;28(2):246–9.
32. Yao ZX, Yang JY, Zhang Q, et al. The application of boosting algorithm in chemical data mining. *J Guangxi Univ Technol.* 2006;17(4):13–8.
33. Zhang MH, Xu QS, Massart DL. Boosting partial least squares. *Anal Chem.* 2005;77(5):1423–31.
34. Drucker H. Improving regressors using boosting techniques. In: *Proceedings of the 14th international conference on machine learning*, 1997.
35. Luo RM, Tan SM, Zhou YP, et al. Quantitative analysis of tea using ytterbium-based internal standard near-infrared spectroscopy coupled with boosting least-squares support vector regression. *J Chemom.* 2013;27(7–8):198–206.
36. Wu XL, Li YJ, Wu TJ. A boosting-partial least squares method for ultraviolet spectroscopic analysis of water quality. *Chin J Anal Chem.* 2013;27(7–8):198–206.
37. Shao XG, Bian XH, Cai WS. An improved boosting partial least squares method for near-infrared spectroscopic quantitative analysis. *Anal Chim Acta.* 2010;666:32–7.
38. Chen Z, Wu ZS, Shi XY, et al. A study on model performance for ethanol precipitation process of *Lonicera Japonica* by NIR based on bagging-PLS and boosting-PLS algorithm. *Chin J Anal Chem.* 2014;42(11):1679–86.
39. Tan C, Li M, Qin X. Random subspace regression ensemble for near-infrared spectroscopic calibration of tobacco samples. *Anal Sci.* 2008;24(5):647–53.
40. Ni WD, Brown SD, Man RL. Stacked partial least squares regression analysis for spectral calibration and prediction. *J Chemom.* 2009;23(10):505–17.
41. Ni WD, Man RL. Stacked multivariate calibration analysis. *Chin J Anal Chem.* 2010;38(3):367–71.
42. Ji GL, Huang GZ, Yang ZJ, et al. Using consensus interval partial least square in near infrared spectra analysis. *Chemom Intell Lab Syst.* 2015;144:56–62.
43. Li YK, Jing J. A consensus PLS method based on diverse wavelength variables models for analysis of near-infrared spectra. *Chemom Intell Lab Syst.* 2014;130:45–9.
44. Liu K, Chen XJ, Li LM, et al. A consensus successive projections algorithm-multiple linear regression method for analyzing near infrared spectra. *Anal Chim Acta.* 2015;858:16–23.
45. Bi YM, Xie Q, Peng SL, et al. Dual stacked partial least squares for analysis of near-infrared spectra. *Anal Chim Acta.* 2013;792:19–27.
46. Cui JD. *A stacked extreme learning machine algorithm based on nir spectroscopy and its application.* Shenyang: Northeastern University; 2015.
47. Shan P, Zhao YH, Wang QY, et al. Stacked ensemble extreme learning machine coupled with partial least squares-based weighting strategy for nonlinear multivariate calibration. *Spectrochim Acta Part A Mol Biomol Spectrosc.* 2019;215:97–111.
48. Chen H, Tan C, Lin Z. Ensemble of extreme learning machines for multivariate calibration of near-infrared spectroscopy. *Spectrochimica Acta Part A: Molecul Biomolecul Spectr.* 2020; 229: 117982.
49. Mevik BH, Segtnan VH, Næs T. Ensemble methods and partial least squares regression. *J Chemom.* 2004;18(11):498–507.
50. Saiz-Abajo MJ, Mevik BH, Segtnan VH, et al. Ensemble methods and data augmentation by noise addition applied to the analysis of spectroscopic data. *Anal Chim Acta.* 2005;533(2):147–59.
51. Conlin AK, Martin EB, Morris AJ. Data augmentation: an alternative approach to the analysis of spectroscopic data. *Chemom Intell Lab Syst.* 1998;44(1):161–73.
52. Li ZG, Peng SL, Yang N, et al. Quantitative analysis method of infrared spectra based on derivative spectra fusion modeling. *Chin J Anal Chem.* 2016;44(3):437–43.
53. Li ZG, Lv JT, Si GY, et al. An improved ensemble model for the quantitative analysis of infrared spectra. *Chemom Intell Lab Syst.* 2015;146:211–20.

54. Bian X H, Wang K Y, Tan E X, et al. A selective ensemble preprocessing strategy for near-infrared spectral quantitative analysis of complex samples. *Chemom Intell Laborat Syst.* 2020; 197:103916.
55. Xu L, Zhou YP, Tang LJ, et al. Ensemble preprocessing of near-infrared (NIR) spectra for multivariate calibration. *Anal Chim Acta.* 2008;616(2):138–43.
56. Lascola R, O'Rourke PE, Kyser EA. A piecewise local partial least squares (PLS) method for the quantitative analysis of plutonium nitrate solutions. *Appl Spectrosc.* 2017;71(12):2579–94.
57. Tan C, Qin X, Li ML. Ensemble partial least squares algorithm mutual information-induced subspace for near-infrared quantitative calibration. *Chin J Anal Chem.* 2009;37(12):1834–8.
58. Xie JB. 20 Lectures on visual machine learning. Beijing: Tsinghua University Press; 2015.
59. Lei M. Machine learning: principles, algorithms and applications. Beijing: Tsinghua University Press; 2019.
60. Yu S, Liu GH, Xia SY, et al. State recognition of solid fermentation process based on near infrared spectroscopy with adaboost and spectral regression discriminant analysis. *Spectr Spectral Anal.* 2016;36(1):51–4.
61. Jin X, Zhu X Z, Li S W, et al. Predicting soil available phosphorus by hyperspectral regression method based on gradient boosting decision tree. *Laser Optoelectr Progr.* 2019; 56(13):131102.
62. Xu K, Cui Y. Application of stacking learning in hyperspectral image classification. *Appl Sci Technol.* 2018;45(6):42–6.
63. Tao YQ, Peng Y, Jiang Q, et al. Remote detection of critical growth stages in rapeseed using vegetation spectral and stacking combination method. *J Geomat.* 2019;44(5):20–3.
64. Shen T, Yu H, Wang YZ. Discrimination of gentiana and its related species using IR spectroscopy combined with feature selection and stacked generalization. *Molecules.* 2020;25(6):1442.
65. Shi RJ, Xia FZ, Zeng WD, et al. Raman spectroscopic classification of foodborne pathogenic bacteria based on PCA-stacking model. *Laser Optoelectr Progr.* 2019;56(4):20–3.
66. Yu X, Yang J, Xie ZQ. Research on virtual sample generation technology. *Comput Sci.* 2011;38(3):16–9.
67. Tang J, Qiao JF, Chai TY, et al. Multi-component mechanical signal modeling based on virtual sample generation technology. *Acta Autom Sin.* 2018;44(9):1569–89.
68. Li DC, Wu CS, Tsai TI, et al. Using mega-trend-diffusion and artificial samples in small data set learning for early flexible manufacturing system scheduling knowledge. *Comput Oper Res.* 2007;34(4):966–82.
69. Zhu B. Virtual sample generation technology and modeling application research. Beijing: Beijing University of Chemical Technology; 2017.
70. Gao KX, Li ZG, Xu CM, et al. Virtual sample construction and blood spectrum analysis of mixed overall trend diffusion. *Chinese J Sci Instrum.* 2019;40(8):94–101.
71. Gong HF. Research on virtual sample generation technology and application of industrial modeling. Beijing: Beijing University of Chemical Technology; 2018.
72. Yi L, Lv ZY, Ding JL, et al. Data amplification preprocessing method for prediction of total hydrogen properties of crude oil. *Control and Decision.* 2018;33(2):44–51.
73. Ye YF, Zhang XR, Mei B, et al. Research on modeling methods based on automatic densification technology. *Sci Technol Vis.* 2017;2:34–34.
74. Li JY, Chu XL. Rapid determination of hydrocarbon composition of LTAG raw materials and products by virtual spectral identification method. *Acta Petrol Sin (Petroleum Process Sect).* 2019;35(2):283–8.
75. Qian J, Guo YK, Zhang Q, et al. High spectral classification modeling of heavy metal Pb and Cd pollution in soil of mining area. *Bull Surv Map.* 2019;9:82–4.
76. Yang YN, Qi LH, Wang H, et al. Research on small sample data generation technology based on generative adversarial network. *Electric Power Construct.* 2019;40(5):71–7.
77. Zhi SS, Zhao QH, Jin DH, et al. The gait virtual sample generation method based on CNN and DLTL. *Appl Res Comput.* 2020;37(1):291–5.
78. Cui X W, Shen T, Liu Y L, et al. Small sample terahertz spectroscopy identification. *Laser Optoelectr Progr.* 2020.

79. Liu JW, Liu Y, Luo XL. Semi-supervised learning methods. *Chinese J Comput.* 2015;38(8):1592–618.
80. Chen WJ. Summarization of semi-supervised learning. *Comput Knowl Technol.* 2011;7(16):3887–9.
81. Cai Y, Zhu XF, Sun ZL, et al. Semi-supervised ensemble learning review. *Comput Sci.* 2017;44(6A):7–14.
82. Zhou ZH. *Machine learning and its application.* Beijing: Tsinghua University Press; 2007.
83. Li L, Xu S, An X, et al. New method for quantitative analysis of near infrared spectroscopy: semi-supervised least squares support vector regression machine. *Spectrosc Spectr Anal.* 2011;31(10):2702–5.
84. Zhang R. Incremental learning algorithm based on support vector regression. *J Shandong Univ Technol (Soc Sci Ed).* 2010;24(3):56–9.
85. Lv CC. *Research on ensemble learning algorithm for incremental NIR semi-supervised SVR.* Shenyang: Northeastern University; 2014.
86. Liang M, Cai JY, Yang K, et al. The application of semi-supervised partial least squares method in near infrared sensory evaluation model of tobacco leaves. *Chin J Anal Chem.* 2014;42(11):1687–91.
87. Guo DS. *Research on the updating method of agricultural product quality detection model.* Wuxi: Jiangnan University; 2018.
88. Jing SB, Yang LM, Li JH, et al. Semi-supervised extreme learning machine and its application in near infrared spectral data analysis. *J Comput Appl.* 2016;36(2):387–91.
89. Wang J, Gao XR, Zhang R, et al. Multi-objective regression combined with target-specific characteristics and target relevance. *Acta Electron Sin.* 2020;48(11):2092–100.
90. Spyromitros-Xioufis E, Tsoumakas G, Groves W, et al. Multi-target regression via input space expansion: treating targets as inputs. *Mach Learn.* 2016;104(1):55–98.
91. Shukla AK. *Spectroscopic techniques and artificial intelligence for food and beverage analysis.* Singapore: Springer; 2020.
92. Santana EJ, Geronimo BC, Mastelini SM, et al. Predicting poultry meat characteristics using an enhanced multi-target regression method. *Biosys Eng.* 2018;171:193–204.
93. Junior SB, Mastelini SM, Barbon APAC, et al. Multi-target prediction of wheat flour quality parameters with near infrared spectroscopy. *Inform Process Agricult.* 2019;7:342–54.

Chapter 15

Multi-spectral Fusion Technology



15.1 Fusion Strategies and Methods

Spectral fusion technology is used to optimize and integrate different types of spectra to achieve complementary advantages of single spectrum, in order to obtain more comprehensive, more reliable and richer characteristic data, and then combine with chemometrics method to build regression or classification model for quantitative and qualitative analysis of samples. As shown in Fig. 15.1, methcathinone and ephedrine in Raman spectra have higher overlaps on score plot of principal component analysis (PCA) (Fig. 15.1a), and the ion mobility spectral analysis on score plot of PCA has a certain tendency of clustering (Fig. 15.1b). But it can be seen from the score plot of PCA in Fig. 15.1c, the fusion of Raman spectra and ion mobility spectra can well realize classification discrimination [1]. As shown in Figs. 15.2 and 15.3, according to different data fusion strategies, the fusion of multi-spectral can be divided into low-level, middle-level, and high-level fusion [2–4].

As shown in Fig. 15.4, low-level fusion refers to spectral data level fusion, and data from different spectral sources are arranged into a matrix in a certain order, that is, the concatenation of spectral matrix [3]. The number of rows of the matrix is the same as the number of samples, and the number of columns is the same as the sum number of columns of the signals (spectral variables) measured by different instruments. Then, the chemometrics method is used to build the final single model. This method is often called concatenation method, such as concatenated PLS [5]. In the low-level fusion, the spectral interval can be selected and essential spectral preprocessing, such as spectral normalization, can be carried out.

As shown in Fig. 15.5, middle-level fusion, also known as feature level fusion, is to extract spectral data from different sources through feature extraction (such as principal component, wavelength ratio, wavelet coefficient), and vectorize the selected variables in a certain order to achieve data fusion. In addition to the traditional spectral feature extraction methods, the deep learning method can also be used to extract spectral features by extracting the NIR wavelength. As shown in Fig. 15.6, the two-channel convolutional neural network is used to extract the depth features of

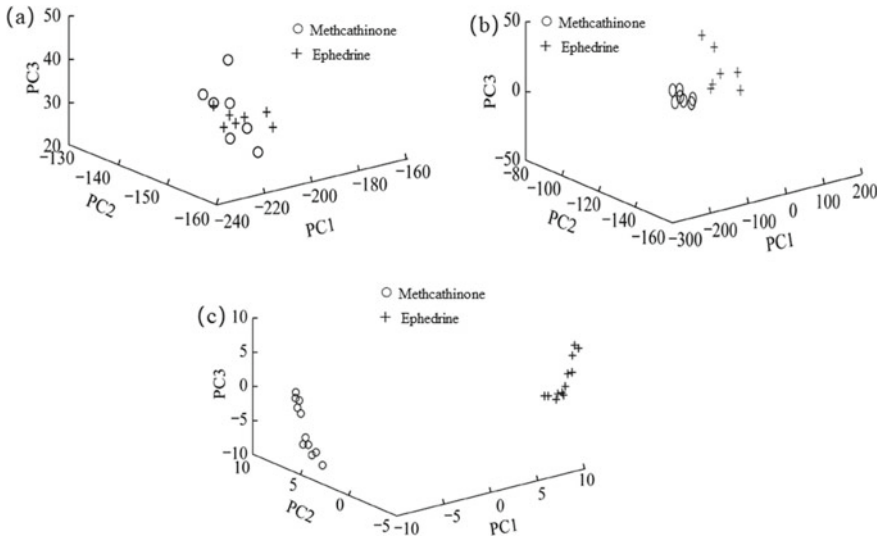


Fig. 15.1 **a** PCA diagram of the Raman spectra of methcathinone and ephedrine. **b** PCA diagram of ion mobility spectra of methcathinone and ephedrine. **c** PCA diagram of Raman and ion mobility spectral fusion of methcathinone and ephedrine [1]

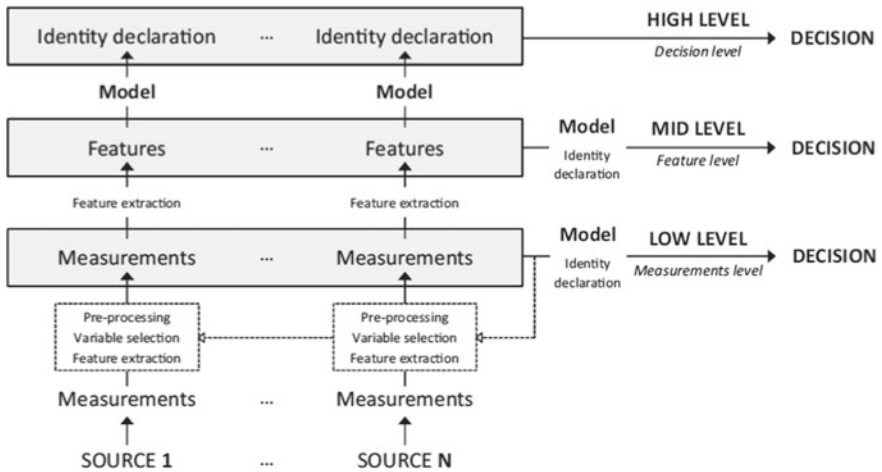


Fig. 15.2 Data fusion scheme at low-, mid-, and high levels [2]

hyperspectral and Lidar data, respectively, and then the extracted features are jointed, The features are trained by regressors or classifiers to obtain the final classification results.

High-level fusion, also known as decision-level fusion, establishes a classification or regression model from each spectral data source, respectively, and combines the

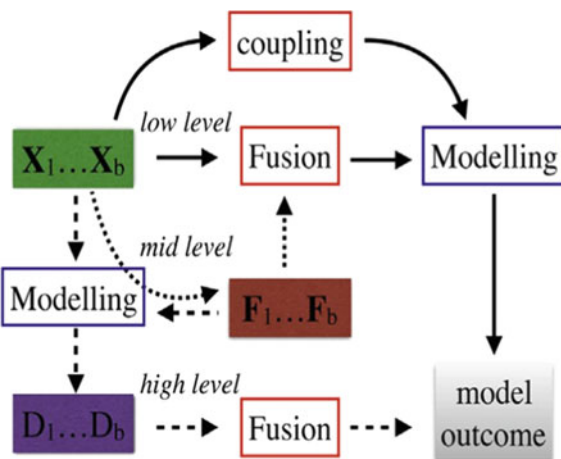


Fig. 15.3 Data fusion approaches. Each path, which corresponds to a different data fusion level, is identified by the arrow style: solid/low level; dotted/mid level; dashed/high level [4]

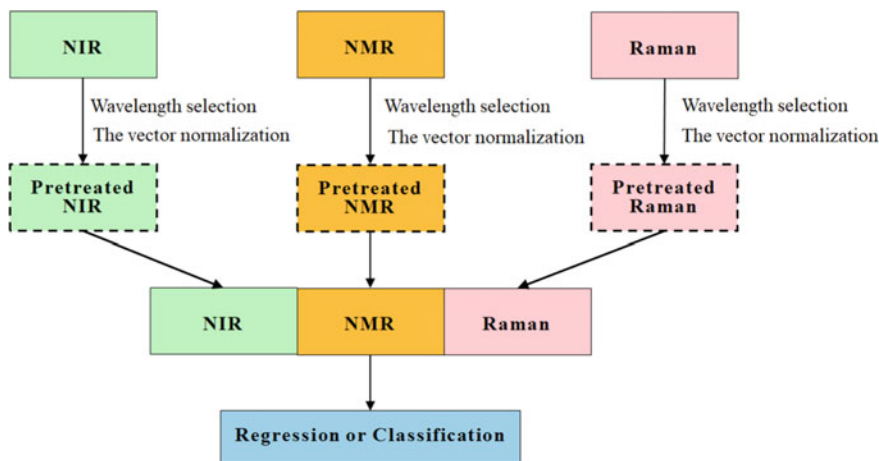


Fig. 15.4 Framework of low-level spectral data fusion

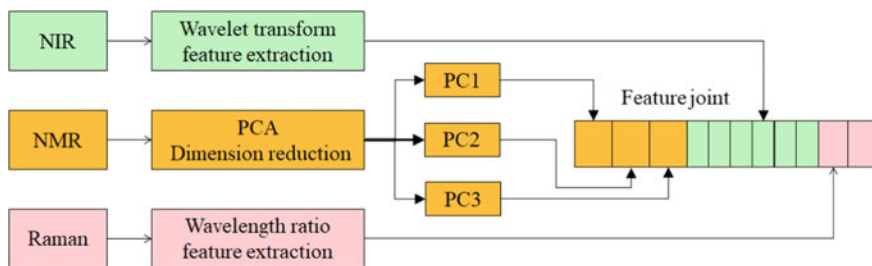


Fig. 15.5 Schematic diagram of mid-level spectral data fusion framework

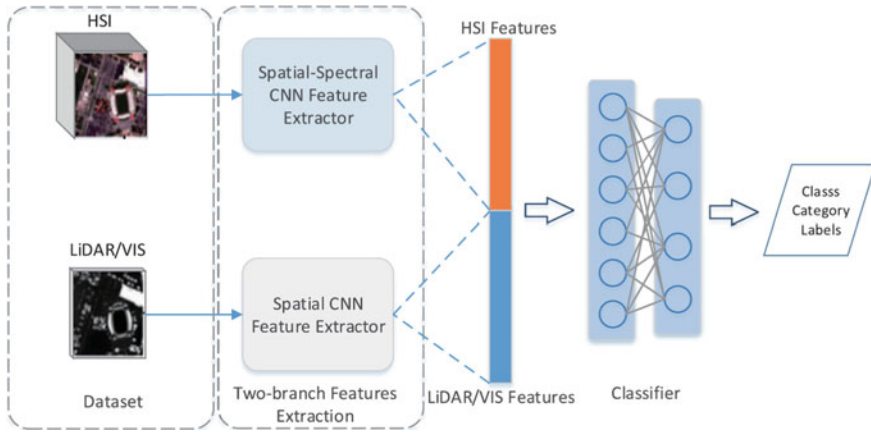


Fig. 15.6 Schematic diagram of feature-level fusion framework based on convolutional neural network [6]

prediction results of each individual model to get the final decision result. In fact, low-level and middle-level fusions of spectral data are often included in the high-level fusion. As shown in Fig. 15.7, using NIR, NMR, and Raman spectral information as well as the feature fusion information of the NIR spectrum and NMR spectrum build SVR model, respectively. The four predicted results can be used to conduct decision fusion in the way of a weighted fusion or voting mechanism, and the final prediction results are obtained.

For the low fusion of the spectral data, in addition to the concatenation of spectral vectors, there are the coaddition, outer sum, and outer product of spectral vectors [7].

Before the vectors coaddition of the spectral data, interpolation operation should be carried out for the fused multi-spectra to obtain vectors with the same dimension. The corresponding elements should be added to obtain the fused spectrum with the same dimension of vector. For example, spectrum A is the vector x of size $1 \times m$, and

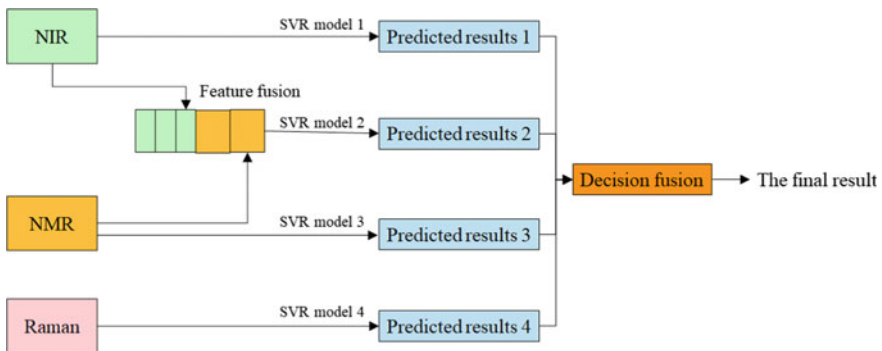


Fig. 15.7 Schematic diagram of high-level fusion

spectrum B is the vector \mathbf{r} of size $1 \times n$. When the spectra A and B are accumulated and fused, the vector \mathbf{r} of spectrum B is firstly interpolated to obtain the vector \mathbf{r}^d of size $1 \times m$. The accumulated fusion spectrum of spectrum A and B is the sum of vector \mathbf{x} and vector \mathbf{r}^d , and its dimension is $1 \times m$. During the accumulation fusion, the vector \mathbf{x} of spectrum A can also be interpolated to obtain the vector \mathbf{x}^d of size $1 \times n$, or the two vectors of size $1 \times n$ can be interpolated simultaneously between spectrum A and spectrum B, and then the vector sum can be carried out.

The spectral vector outer product is actually to find the outer product of two vectors. For a vector \mathbf{x} of size $1 \times m$ in spectrum A and a vector \mathbf{r} of size $1 \times n$ in spectrum B, the outer product $\mathbf{x} \otimes \mathbf{r} = \mathbf{x}^T \times \mathbf{r}$ is a matrix with $m \times n$ dimension. For k samples in calibration set, a $k \times m \times n$ three-dimensional matrix can be obtained. As shown in Fig. 15.8, the three-dimensional matrix obtained by outer product of X-ray fluorescence (XRF) and Vis-NIR spectral vectors can be quantitatively or qualitatively analyzed by multi-dimensional chemometrics [8]. Moreover, the obtained matrix with $m \times n$ dimension can be unfolded into $1 \times mn$ dimension vectors, and then data processing can be carried out by traditional chemometrics methods.

The outer sum of spectral vectors is similar to finding the outer product of two vectors. For the vector \mathbf{x} of spectrum A with a dimension of $1 \times m$, and for the vector \mathbf{r} of spectrum B is with a dimension of $1 \times n$, the outer sum $\mathbf{x} \oplus \mathbf{r}$ is as follows:

$$\mathbf{x} = (x_1, x_2, \dots, x_m), \mathbf{r} = (r_1, r_2, \dots, r_n) \tag{15.1}$$

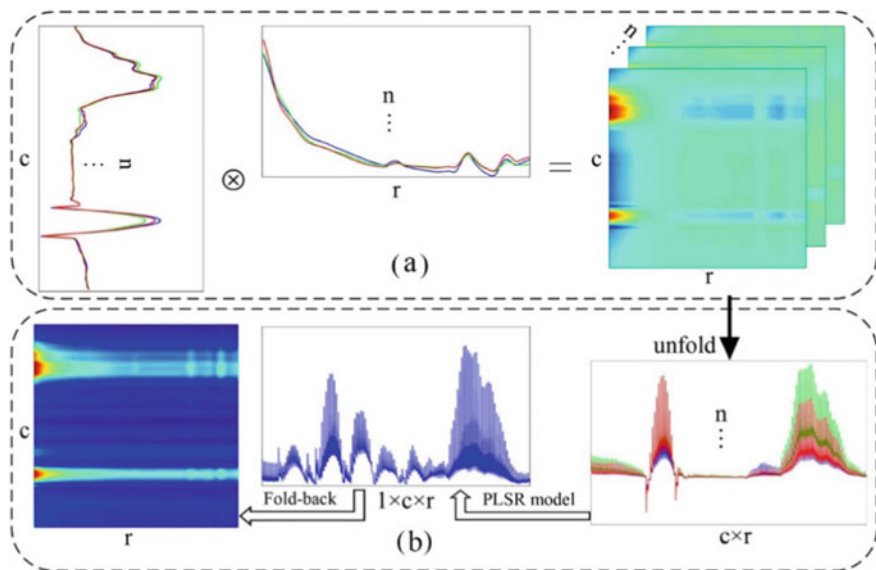


Fig. 15.8 Process of data fusion by outer product (a) and the unfolding process for modeling and refolding of the model results (b). c is the number of variables for X-ray fluorescence (XRF) and r is the number of wavelengths for Vis-NIR, n is the number of soil samples [8]

$$\mathbf{x} \oplus \mathbf{r} = \begin{bmatrix} x_1 \\ \vdots \\ x_m \end{bmatrix} \oplus [r_1 \cdots r_n] = \begin{bmatrix} x_1 + r_1 \cdots x_m + r_1 \\ \vdots \quad \ddots \quad \vdots \\ x_1 + r_n \cdots x_m + r_n \end{bmatrix} \quad (15.2)$$

Similar to the outer product of vectors, the three-dimensional matrix of the outer sum vectors can be processed by multi-dimensional chemometrics method or unfolded by traditional chemometrics method.

The outer product operation and the outer sum operation of the spectra are usually used for the fusion calculation of two kinds of spectra. For the fusion of multiple kinds of spectra, the pair-to-pair operation can be carried out, respectively, or the concatenation of spectral vectors can be conducted first, and then the outer product operation or the outer sum operation can be implemented.

15.2 Multi-block Partial Least Squares Method

For multi-spectral fusion technology, multi-block partial least squares method (Multi-block PLS) can be used to establish the calibration model. For example, for Raman spectral matrix (block \mathbf{X}_1) and NIR spectral matrix (block \mathbf{X}_2), the modeling strategy of Multi-block PLS method is shown in Fig. 15.9. Firstly, the PLS model of each

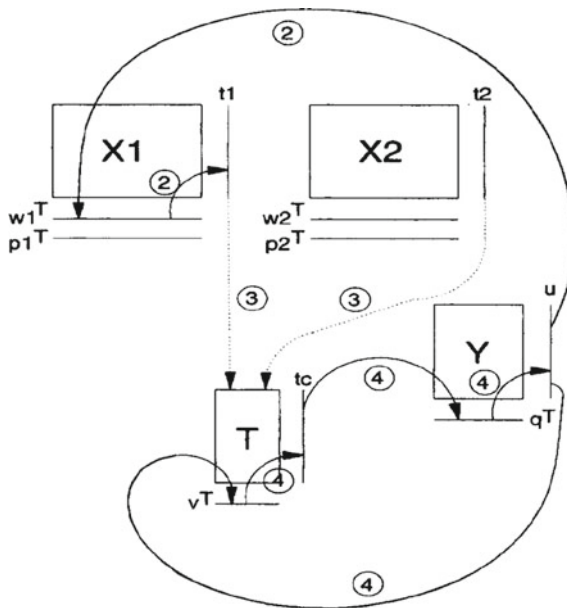


Fig. 15.9 Diagram of the iterative process of multi-block partial least squares [9]

block and concentration \mathbf{Y} is established respectively, and the corresponding PLS principal components (called the lower-layer model) are extracted. Then the PLS components and concentrations obtained from each block were used to establish the whole PLS model (called the upper layer model) [9]. In the above process, the results obtained by the multi-block PLS method have a stronger ability of comprehensive generalization of information, interpretation, and application value. Because the number of variables of each block is far less than the number of whole variables, and each block has a specific connotation meaning, Multi-block PLS method is also known as the hierarchical PLS regression [10].

The specific steps of multi-block PLS algorithm are as follows:

- (1) \mathbf{w}_1 and \mathbf{w}_2 are calculated by taking a column of concentration matrix \mathbf{Y} as the initial iteration value of \mathbf{u} , \mathbf{w}_1 , and \mathbf{w}_2 were calculated as the following formula:

$$\mathbf{W}_1^T = \mathbf{U}^T \mathbf{X}_1 / \mathbf{U}^T \mathbf{U}, \mathbf{W}_2^T = \mathbf{U}^T \mathbf{X}_2 / \mathbf{U}^T \mathbf{U} \quad (15.3)$$

- (2) Normalize \mathbf{w} as the following formula:

$$\mathbf{W}_1 = \mathbf{W}_1 / \|\mathbf{W}_1\|, \mathbf{W}_2 = \mathbf{W}_2 / \|\mathbf{W}_2\| \quad (15.4)$$

- (3) Calculate \mathbf{t}_1 and \mathbf{t}_2 .

$$\mathbf{t}_1 = \mathbf{X}_1 \mathbf{W}_1 / \mathbf{W}_1^T \mathbf{W}_1, \mathbf{t}_2 = \mathbf{X}_2 \mathbf{W}_2 / \mathbf{W}_2^T \mathbf{W}_2 \quad (15.5)$$

- (4) Construct the joint matrix \mathbf{T}_c , $\mathbf{T}_c = [\mathbf{t}_1 \mathbf{t}_2]$
- (5) Using the standard PLS algorithm, the regression model of \mathbf{T}_c and concentration \mathbf{Y} was established, and the vectors \mathbf{w} , \mathbf{t} , \mathbf{u} and \mathbf{q} were obtained.
- (6) Return to step (1) until \mathbf{u} converges.
- (7) Calculate \mathbf{p}_1 and \mathbf{p}_2 :

$$\mathbf{p}_1^T = \mathbf{t}_1^T \mathbf{Y} / \mathbf{t}_1^T \mathbf{t}_1, \mathbf{p}_2^T = \mathbf{t}_2^T \mathbf{Y} / \mathbf{t}_2^T \mathbf{t}_2 \quad (15.6)$$

- (8) Calculate the residual matrix:

$$\mathbf{E}_1 = \mathbf{X}_1 - \mathbf{t}_1 \mathbf{p}_1^T, \mathbf{E}_2 = \mathbf{X}_2 - \mathbf{t}_2 \mathbf{p}_2^T, \mathbf{F} = \mathbf{Y} - \mathbf{t} \mathbf{p}^T \quad (15.7)$$

- (9) Replace \mathbf{X}_1 , \mathbf{X}_2 , and \mathbf{Y} with \mathbf{E}_1 , \mathbf{E}_2 , and \mathbf{F} , respectively, and return to step (1) until the PLS principal components of the number of all major factors are calculated.

15.3 Sequential and Orthogonal Partial Least Squares Method

Multi-block PLS uses a parallel-type calibration mode, but sequential and orthogonal partial least squares (SO-PLS) is a tandem-type calibration method. For example, for the Raman spectral matrix (Block \mathbf{X}_1) and NIR spectral matrix (Block \mathbf{X}_2), the modeling strategy of SO-PLS is as follows: firstly, the PLS model of block \mathbf{X}_1 and concentration \mathbf{Y} was established to obtain the corresponding PLS principal components (such as fracted matrix $\mathbf{T}_{\mathbf{X}_1}$ and concentration residual matrix \mathbf{y}_R). The orthogonal spectral matrix $\mathbf{X}_{2\text{orth}}$ was obtained by orthogonalization of $\mathbf{T}_{\mathbf{X}_1}$ and block \mathbf{X}_2 , and then the PLS model of the orthogonal matrix $\mathbf{X}_{2\text{orth}}$ and concentration residual matrix \mathbf{y}_R was established. The final prediction results were given by the combination of the above two calibration models [11, 12]. Since the SO-PLS method adopts orthogonalization processing, the additional complementary spectral information of block \mathbf{X}_2 relative to block \mathbf{X}_1 can be effectively extracted [13].

Specific steps of the SO-PLS algorithm are as follows:

- (1) The standard PLS algorithm is adopted to establish the regression model of \mathbf{X}_1 and concentration \mathbf{y} , and the score matrix $\mathbf{T}_{\mathbf{X}_1}$, weight matrix $\mathbf{W}_{\mathbf{X}_1}$, loading matrix $\mathbf{P}_{\mathbf{X}_1}$, load matrix $\mathbf{Q}_{\mathbf{X}_1}$ of \mathbf{y} , and residual matrix \mathbf{y}_R of \mathbf{y} are obtained as the following formula:

$$\mathbf{y}_R = \mathbf{y} - \mathbf{T}_{\mathbf{X}_1} \mathbf{Q}_{\mathbf{X}_1}^T \quad (15.8)$$

- (2) Orthogonalize $\mathbf{T}_{\mathbf{X}_1}$ and block \mathbf{X}_2 to obtain the orthogonal spectral matrix $\mathbf{X}_{2\text{orth}}$:

$$\mathbf{X}_{2\text{orth}} = \mathbf{X}_2 - \mathbf{T}_{\mathbf{X}_1} (\mathbf{T}_{\mathbf{X}_1}^T \mathbf{T}_{\mathbf{X}_1})^{-1} \mathbf{T}_{\mathbf{X}_1}^T \mathbf{T}_{\mathbf{X}_2} \quad (15.9)$$

- (3) Based on the standard PLS algorithm, the regression model of $\mathbf{X}_{2\text{orth}}$ and concentration residual \mathbf{y}_R is established, and the score matrix $\mathbf{T}_{\mathbf{X}_{2\text{orth}}}$ of $\mathbf{X}_{2\text{orth}}$, the weight matrix $\mathbf{W}_{\mathbf{X}_{2\text{orth}}}$ of $\mathbf{X}_{2\text{orth}}$, the loading matrix $\mathbf{P}_{\mathbf{X}_{2\text{orth}}}$ of $\mathbf{X}_{2\text{orth}}$, and the loading matrix $\mathbf{Q}_{\mathbf{X}_{2\text{orth}}}$ of \mathbf{y}_R are obtained.
- (4) The predicted value of concentration \mathbf{y}^{pre} is given by the following formula:

$$\mathbf{y}^{\text{pre}} = \mathbf{T}_{\mathbf{X}_1} \mathbf{Q}_{\mathbf{X}_1}^T + \mathbf{T}_{\mathbf{X}_{2\text{orth}}} \mathbf{Q}_{\mathbf{X}_{2\text{orth}}}^T \quad (15.10)$$

The above equation can also be expressed as the following formula:

$$\mathbf{y}^{\text{pre}} = \mathbf{X}_1 \mathbf{V}_{\mathbf{X}_1} \mathbf{Q}_{\mathbf{X}_1}^T + \mathbf{X}_{2\text{orth}} \mathbf{V}_{\mathbf{X}_{2\text{orth}}} \mathbf{Q}_{\mathbf{X}_{2\text{orth}}}^T \quad (15.11)$$

where $\mathbf{V}_{\mathbf{X}_1} = \mathbf{W}_{\mathbf{X}_1} (\mathbf{P}_{\mathbf{X}_1}^T \mathbf{W}_{\mathbf{X}_1})^{-1}$, and $\mathbf{V}_{\mathbf{X}_{2\text{orth}}} = \mathbf{W}_{\mathbf{X}_{2\text{orth}}} (\mathbf{P}_{\mathbf{X}_{2\text{orth}}}^T \mathbf{W}_{\mathbf{X}_{2\text{orth}}})^{-1}$.

As shown in Fig. 15.10, the combination of multi-block PLS and SO-PLS and multi-point spectra or fusion technology of multi-spectra can be used in the production process control. It can be used in the process of system to predictively analyze

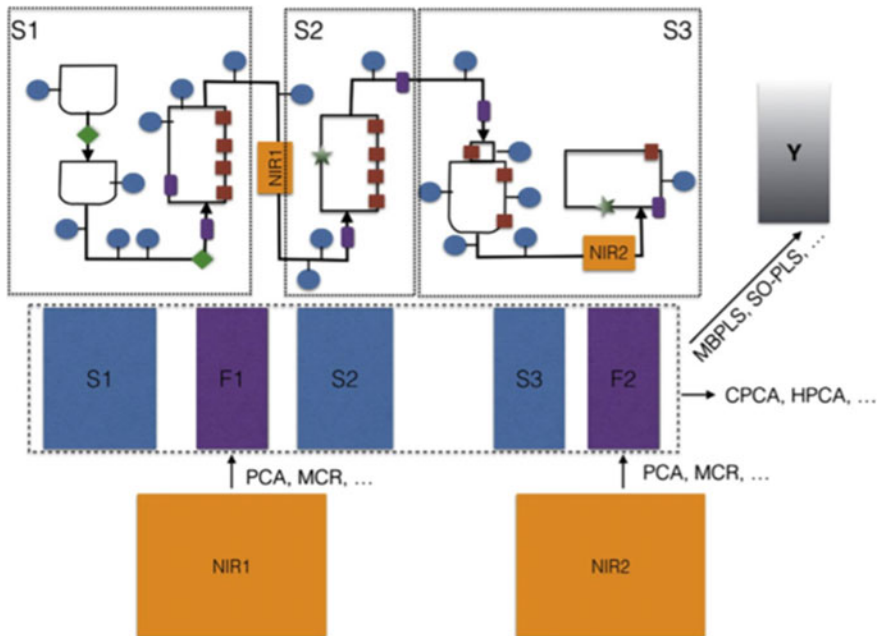


Fig. 15.10 An example of fusion of process sensors and NIR spectral data in a continuous process line. The top figure represents a schematic view of the process, indicating the location of the process sensors (represented by different shapes and colors) on the process line and the location of the two online NIR instruments [4]

key quality on each unit, deeply understand the causal relationship between various factors, identified the key quality control points, and improve the quality of the product stably.

In addition to multi-block PLS and SO-PLS methods introduced above, there are also some other methods such as common components and specific weights analysis (CCSWA) in the multi-spectral fusion technology, called ComDim (Common Dimension), a collective name for a series of algorithms [14–16] and parallel and orthogonalized PLS (PO-PLS) and so on [17–20].

15.4 Research on Application of Multi-Spectral Fusion

Dearing et al. performed low-level fusion of MIR, Raman, and NMR spectra to establish a quantitative model for predicting active pharmaceutical ingredient (API) degree of crude oil with PLS, and its prediction accuracy (RMSEP) was improved by more than 50% compared with using a single spectral technique [21]. Chen et al. used the multi-scale characteristics of the discrete wavelet transform (DWT) and competitive adaptive reweighted sampling-PLS discriminant analysis (CARS-PLSDA), to extract

the feature of Raman and IR spectra. The discriminant model of milk adulteration is established in the feature-level fusion, which could effectively combine the collaborative information and complementary characteristics of Raman and IR spectra and significantly improve the sensitivity and accuracy of adulteration detection of milk powder [22]. Marquez et al. used FT-Raman and NIR spectra to detect hazelnut kernel adulteration by using two data fusion strategies of middle-level fusion (feature-level) and high-level (decision-level) fusion. The results showed that the sensitivity and specificity of single spectral technique were between 75 and 100%, respectively, and the sensitivity and specificity of high-level fusion and middle-level fusion were between 96 and 100% and between 88 and 100%, respectively, indicating that its performance parameters were superior to that of single spectral technique [23]. Tao et al. employed MIR and NIR multi-spectral fusion technology combined with SO-PLS algorithm to predict and analyze various active components in the liquid extraction process of *Lonicera japonica* and *Artemisia annua*, and obtained satisfactory results [13].

For the adulterated sesame oil, Zhang et al. first used the two-dimensional correlation spectra technology to obtain the synchronous-asynchronous two-dimensional NIR correlation and MIR correlation spectra of the samples, respectively. Multi-way principal component analysis (MPCA) was carried out on the two-dimensional correlation spectra, and the score matrix was fused to identify the adulterated sesame oil through the PLS-DA model, the discrimination accuracy reaches 100%, which is higher than that of prediction model of single spectral technique [24]. Shen et al. used the middle-level data fusion based on wavelength selection and the high-level fusion based on stack generalization strategy to identify *Gentiana* plants by using MIR and NIR multi-spectral fusion technology. The results show that the strategy based on wavelength selection and stack generalization can improve the accuracy of discrimination and can prevent the occurrence of overfitting [25].

Rios-Reina et al. applied IR, NIR, three-dimensional fluorescence excitation-emission matrix (EEM), and ^1H NMR spectra to identify wine-protected designations of origin (PDO). As shown in Fig. 15.11, after low-level fusion of NIR and MIR spectra, PCA was conducted, and eight score variables of principal components were obtained as characteristic variable I; PARAFAC decomposition was performed on the EEM spectra, and five PARAFAC score variables were gained as characteristic variable II; multivariate curve resolution (MCR) was implemented on ^1H NMR spectra, and the 62 peak areas were differentiated and obtained as the characteristic variable III. Then data fusion of the above three characteristic variables was carried out to establish the PDO discriminant model with PLS-DA, and better discriminant results than single spectral method were obtained [26].

Yao et al. used UV spectra combined with IR spectra to distinguish the origin of the species of *Boletus tonientipes* Earle, extracted characteristic variables through wavelength selection method and performed data fusion, and the prediction accuracy of the discriminant model established by SVM reached 96.88% [27]. Comino et al. applied NIR and X-ray fluorescence spectra to conduct data fusion for the rapid analysis of nutrient elements in olive leaves. The feature-level fusion strategy based on PCA obtained good prediction results, and the deficiency of important

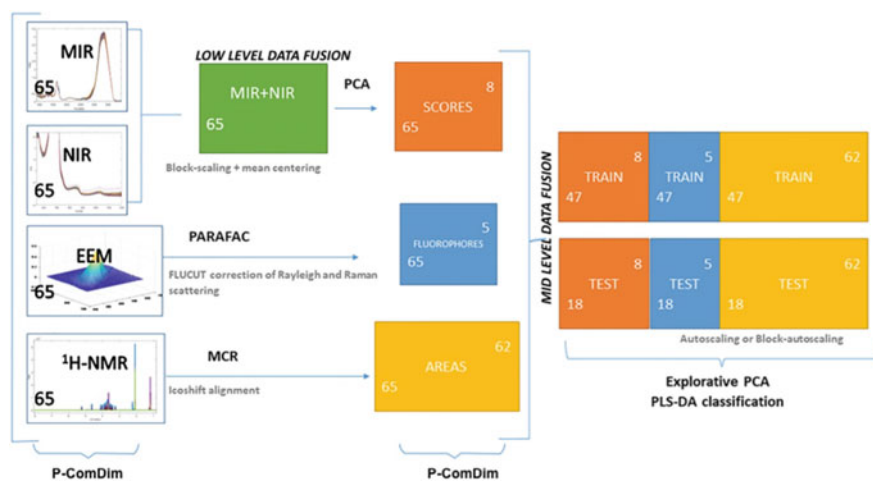


Fig. 15.11 Graphical representation of the datasets, data analysis flow, and data fusion process for identifying wine protected designations of origin [26]

elements such as nitrogen and potassium could be quickly detected [28]. They used the adaptive wavelet transform algorithm and CARS algorithm to remove the background noise and select the variables of LIBS and Raman spectra, respectively. Then, the characteristic variables were fused to establish the calibration model with PLS, which improved the precision and reliability of quantitative analysis of flour doping. Gibbons et al. also adopted the fusion technology of LIBS and Raman spectroscopy to classify and identify clay minerals by using the molecular structure information of Raman spectra and the element information of LIBS spectra [29].

After essential preprocessing of hyperspectral images for beef, Wang et al. extracted 22 characteristic wavelengths by using CARS, and then fused them with 48 features of image texture to establish a PLS-DA classification model. The prediction accuracy was 93.55%, which is higher than the classification rate of characteristic spectral data model. It shows that the fusion of texture features can make the expression of sample classification information more comprehensive [30]. They also used a similar technical route to establish a model for predicting saturated fatty acid content in mutton by the fusion of images and spectra and then obtained good results [31]. Zou et al. established the SVM identification model of wheat origin and drying degree after the fusion of the screened characteristic intervals of NIR and MIR spectra, and the discriminant result was better than that of the single spectral technique [32].

Casian et al. predicted the content of API in drugs with NIR spectroscopy, Raman spectroscopy, colorimetry and image analysis techniques. The prediction accuracy of the single technique was between 0.654 and 2.292%. After dimensionality reduction, the four types of data were used to conduct feature extraction and data fusion, and ANN is adopted to establish the quantitative calibration model, with the prediction accuracy of 0.153% [33]. Assis et al. used NIR spectroscopy and total reflection XRF spectroscopy to predict the component content of the two coffee mixtures and

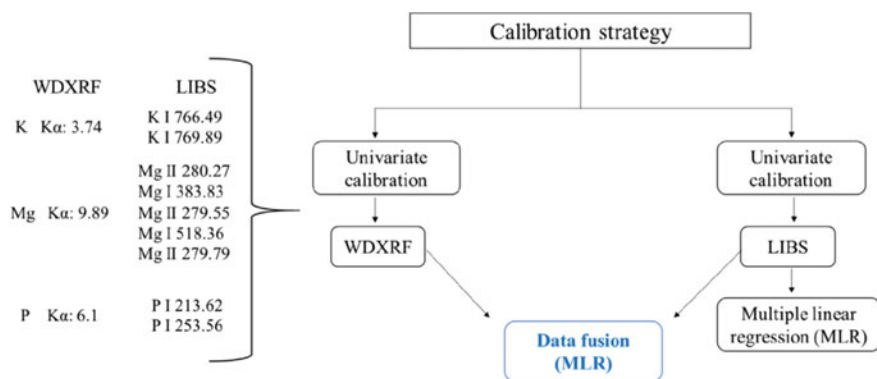


Fig. 15.12 Schematic diagram of wavelength dispersive X-ray fluorescence (WDXRF) and laser-induced breakdown spectroscopy (LIBS) for fused models [35]

fused the selected 75 characteristic variables of NIR spectra with the contents of 14 elements measured by total reflection XRF spectroscopy to establish a PLS quantitative analysis model. This model can be used to explain the differences between different coffee species from the atomic and molecular spectra [34]. As shown in Fig. 15.12, Gamela et al. extracted the characteristic variables of LIBS and wavelength dispersive XRF (WDXRF) spectra, respectively, and used MLR to establish the predictive model of the contents of potassium, magnesium, and phosphorus in soybean seeds. The prediction results were better than those based on the single spectral calibration model [35]. Oliveira et al. performed a low-level fusion of NIR and LIBS spectra for rapid analysis of trace elements and major elements in forage and obtained satisfactory results [36].

As shown in Fig. 15.13, Moro et al. employed MIR, ^1H NMR, and ^{13}C NMR spectra to make quantitative prediction analysis of seven kinds of physical properties of crude oil. When the scores extracted from the PLS model were used for data fusion, the best prediction results were obtained by PLS method [37]. Liu et al. fused IR and Raman spectra for feature-level fusion to predict oil peroxide and acid values in the process of thermal oxidation. The results showed that a highly correlated quantitative relationship existed between the C = O functional groups information and C = C functional information offered by IR and Raman spectra, respectively, and measured physical properties. The prediction results of fusion data from two kinds of spectra are better than that of the single spectra method [38]. Wang et al. established the PLS prediction model of puerarin content in the root of *pueraria DC.* by fusion of NIR and UV spectra and proved that the fusion of NIR and UV spectra had synergistic effect [39]. As shown in Fig. 15.14, Germany's art photonics company combined Raman, MIR, NIR, and molecular fluorescence spectra to monitor the chemical reaction process and used ComDim and SO-PLS algorithms to carry out fusion, discrimination, and modeling of multi-spectral data [40].

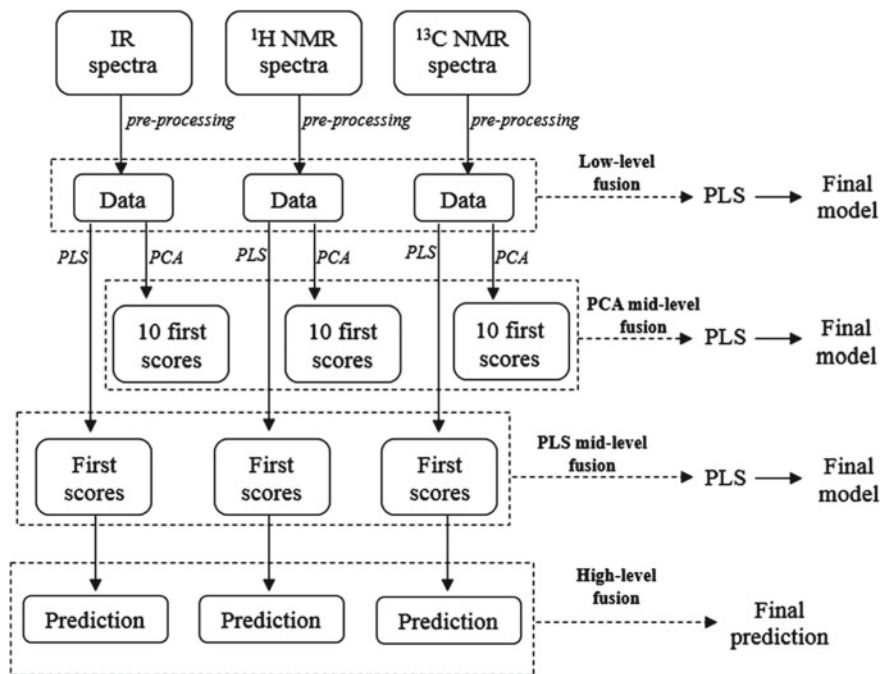


Fig. 15.13 Schematic diagram of various data fusion strategies based on MIR, ^1H NMR, and ^{13}C NMR spectroscopy [37]

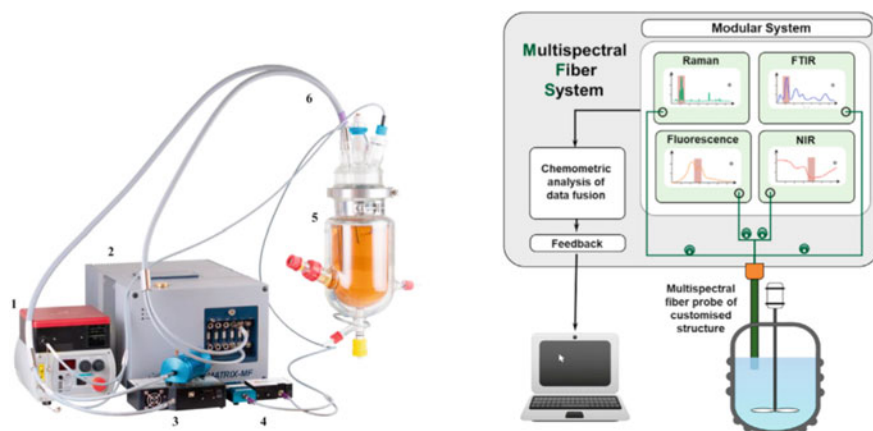


Fig. 15.14 Multi-spectral fusion for monitoring chemical reaction processes [40]. (Left panel) 1: Raman spectroscopy system; 2: Fourier transform mid-infrared spectroscopy (FTIR) system; 3: near-infrared (NIR) spectral reflection system; 4: molecular fluorescence spectroscopy system; 5: chemical reactor; 6: optical fiber probe

15.5 Future Prospect

Multi-spectral fusion technology can realize the synergistic information and complementary characteristics of each spectroscopy technology, which makes the qualitative or quantitative prediction results more accurate and reliable. The data processing of multi-spectral fusion technology requires appropriate chemometrics methods. Accurate algorithms and the improvement of modeling are conducive to improving the efficiency of data processing in the later period, which is helpful to develop corresponding software and provide a more convenient and effective platform for data processing. At present, multi-spectral fusion technology is developing vigorously, especially the development of multi-spectral all-in-one machine has received more and more attention [41, 42]. Currently, commercial or under development multi-spectral instruments include the combination of Raman and MIR spectrometer, LIBS and Raman spectrometer, XRF and Raman spectrometer, XRF and LIBS instrument, MIR and NIR spectrometer, Raman and terahertz instruments, deep UV-Raman and molecular fluorescence spectrometer, as well as a variety of spectroscopic imaging instruments. In this way, a miniature or small instrument can obtain more and richer information about the composition of substances [43, 44].

On this basis of this, the combination of multi-spectral instrument and hardware and multi-spectral data fusion algorithms is a development trend in the future. Through the cloud platform, multi-spectral data collection and data fusion processing can be integrated, which can further save manpower and material resources and improve the efficiency of analysis. Multi-spectral fusion technology is expected to be widely used in the fields of environment, biomedicine, pharmacy, geology, food, agriculture, and identification of physical evidence.

References

1. Jiang L, Shen J, Yu Z, et al. Drug identification method based on data fusion of ion mobility spectrometry and Raman spectroscopy by PCA-SVM analysis. *Opt Instrum.* 2018;40:31–7.
2. Borràs E, Ferré J, Boqué R, et al. Data fusion methodologies for food and beverage authentication and quality assessment – a review. *Anal Chim Acta.* 2015;891:1–14.
3. Yang Q-L, Deng X-J, Sun X-D, et al. Application and research progress of spectral data fusion technology in food testing. *Sci Technol Food Ind.* 2020;41:324–9.
4. Cocchi M. Chapter 1 - Introduction: Ways and means to deal with data from multiple sources. In: Cocchi M. editor. *Data handling in science and technology.* Elsevier; 2019, 1–26.
5. Dupuy N, Galtier O, Ollivier D, et al. Comparison between NIR, MIR, concatenated NIR and MIR analysis and hierarchical PLS model. Application to virgin olive oil analysis. *Anal Chim Acta.* 2010;(666):23–31.
6. Xu X, Li W, Ran Q, et al. Multisource remote sensing data classification based on convolutional neural network. *IEEE Trans Geosci Remote Sens.* 2018;56:937–49.
7. Moros J, Javier LJ. Unveiling the identity of distant targets through advanced Raman-laser-induced breakdown spectroscopy data fusion strategies. *Talanta.* 2015;134:627–39.
8. Xu D, Chen S, Viscarra Rossel RA, et al. X-ray fluorescence and visible near infrared sensor fusion for predicting soil chromium content. *Geoderma.* 2019;352:61–9.

9. MacGregor JF, Jaeckle C, Kiparissides C, et al. Process monitoring and diagnosis by multiblock PLS methods. *AIChE J.* 1994;40:826–38.
10. Wold S, Sjöström M, Eriksson L. PLS-regression: a basic tool of chemometrics. *Chemom Intell Lab Syst.* 2001;58:109–30.
11. Lauzon-Gauthier J, Manolescu P, Duchesne C. The sequential multi-block PLS algorithm (SMB-PLS): Comparison of performance and interpretability. *Chemom Intell Lab Syst.* 2018;180:72–83.
12. Næs T, Tomic O, Mevik B-H, et al. Path modelling by sequential PLS regression. *J Chemom.* 2011;25:28–40.
13. Tao L, Via B, Wu Y, et al. NIR and MIR spectral data fusion for rapid detection of *Lonicera japonica* and *Artemisia annua* by liquid extraction process. *Vib Spectrosc.* 2019;102:31–8.
14. Mazerolles G, Hanafi M, Dufour E, et al. Common components and specific weights analysis: A chemometric method for dealing with complexity of food products. *Chemom Intell Lab Syst.* 2006;81:41–9.
15. Cordella CBY, Bertrand D. SAISIR: A new general chemometric toolbox. *TrAC, Trends Anal Chem.* 2014;54:75–82.
16. El Ghaziri A, Cariou V, Rutledge DN, et al. Analysis of multiblock datasets using ComDim: Overview and extension to the analysis of (K+1) datasets. *J Chemom.* 2016;30:420–9.
17. Måge I, Mevik B-H, Næs T. Regression models with process variables and parallel blocks of raw material measurements. *J Chemom.* 2008;22:443–56.
18. Næs T, Tomic O, Afseth NK, et al. Multi-block regression based on combinations of orthogonalisation, PLS-regression and canonical correlation analysis. *Chemom Intell Lab Syst.* 2013;124:32–42.
19. Måge I, Menichelli E, Næs T. Preference mapping by PO-PLS: Separating common and unique information in several data blocks. *Food Qual Prefer.* 2012;24:8–16.
20. Biancolillo A, Marini F, Ruckebusch C, et al. Chemometric strategies for spectroscopy-based food authentication. *Appl Sci.* 2020;10:6544.
21. Dearing TI, Thompson WJ, Rechsteiner CE, et al. Characterization of crude oil products using data fusion of process Raman, infrared, and nuclear magnetic resonance (NMR) spectra. *Appl Spectrosc.* 2011;65:181–6.
22. Chen D, Luo W, Huang Z, et al. Adulterated milk powder diagnosis method based on multi-spectra fusion. *Nanotechnol Precis Eng.* 2017;15:384–8.
23. Márquez C, López MI, Ruisánchez I, et al. FT-Raman and NIR spectroscopy data fusion strategy for multivariate qualitative analysis of food fraud. *Talanta.* 2016;161:80–6.
24. Zhang J, Shan H-Y, Yang R-J, et al. Discrimination of adulterated sesame oil using fusion of near-mid infrared correlation spectra. *Acta Photon Sin.* 2019;48:56–62.
25. Shen T, Yu H, Wang Y-Z. Discrimination of *Gentiana* and its related species using IR spectroscopy combined with feature selection and stacked generalization. *Molecules.* 2020;25:1442.
26. Ríos-Reina R, Callejón RM, Savorani F, et al. Data fusion approaches in spectroscopic characterization and classification of PDO wine vinegars. *Talanta.* 2019;198:560–72.
27. Yao S, Li T, Liu H, et al. Identification of geographical origin of *Boletus Tomentipes* by multi-spectral data fusion. *Food Sci.* 2018;39:212–7.
28. Comino F, Ayora-Cañada MJ, Aranda V, et al. Near-infrared spectroscopy and X-ray fluorescence data fusion for olive leaf analysis and crop nutritional status determination. *Talanta.* 2018;188:676–84.
29. Gibbons E, Léveillé R, Berlo K. Data fusion of laser-induced breakdown and Raman spectroscopies: Enhancing clay mineral identification. *Spectrochim Acta Part B: Atom Spectrosc.* 2020;(170):105905.
30. Wang C-X, Wang S-L, He X-G, et al. The identification of beef varieties by fusing image information based on hypersepectral image technology. *Spectrosc Spect Anal.* 2020;40:911–6.
31. Wang C-X, Wwang S-L, He X-G, et al. Detection of saturated fatty acid content in mutton by using the fusion of hyperspectral spectrum and image information. *Spectrosc Spect Anal.* 2020;40:595–601.

32. Zou X-b, Feng T, Zheng K-y, et al. Simultaneous identification of wheat origin and drying degree using near-infrared and mid-infrared fusion techniques. *Spectrosc Spect Anal.* 2019;(39):1445–50.
33. Casian T, Farkas A, Ilyés K, et al. Data fusion strategies for performance improvement of a process analytical technology platform consisting of four instruments: an electrospinning case study. *Int J Pharm.* 2019;(567):118473.
34. Assis C, Gama EM, Nascentes CC, et al. A data fusion model merging information from near infrared spectroscopy and X-ray fluorescence. Searching for atomic-molecular correlations to predict and characterize the composition of coffee blends. *Food Chem.* 2020;(325):126953.
35. Gamela R R, Costa V C, Sperança M A, et al. Laser-induced breakdown spectroscopy (LIBS) and wavelength dispersive X-ray fluorescence (WDXRF) data fusion to predict the concentration of K, Mg and P in bean seed samples. *Food Res Int.* 2020;(132):109037.
36. de Oliveira DM, Fontes LM, Pasquini C. Comparing laser induced breakdown spectroscopy, near infrared spectroscopy, and their integration for simultaneous multi-elemental determination of micro- and macronutrients in vegetable samples. *Anal Chim Acta.* 2019;1062:28–36.
37. Moro MK, Neto ÁC, Lacerda V, et al. FTIR, 1H and 13C NMR data fusion to predict crude oils properties. *Fuel.* 2020;(263):116721.
38. Liu H, Chen Y, Shi C, et al. FT-IR and Raman spectroscopy data fusion with chemometrics for simultaneous determination of chemical quality indices of edible oils during thermal oxidation. *LWT.* 2020;(119):108906.
39. Wang Y, Yang Y, Sun H, et al. Application of a data fusion strategy combined with multi-variate statistical analysis for quantification of puerarin in *Radix puerariae*. *Vibrat Spectrosc.* 2020;(108):103057.
40. Mishra P, Roger JM, Rutledge DN, et al. MBA-GUI: A chemometric graphical user interface for multi-block data visualisation, regression, classification, variable selection and automated pre-processing. *Chemom Intell Labor Syst.* 2020;(205):104139.
41. Crocombe RA. Portable spectroscopy. *Appl Spectrosc.* 2018;72:1701–51.
42. Hashimoto K, Badarla VR, Kawai A, et al. Complementary vibrational spectroscopy. *Nat Commun.* 2019;10:4411.
43. Stuart MB, McGonigle AJS, Willmott JR. Hyperspectral imaging in environmental monitoring: A review of recent developments and technological advances in compact field deployable systems. *Sensors.* 2019;19:3071.
44. Deidda R, Sacre P-Y, Clavaud M, et al. Vibrational spectroscopy in analysis of pharmaceuticals: Critical review of innovative portable and handheld NIR and Raman spectrophotometers. *TrAC, Trends Anal Chem.* 2019;114:251–9.

Chapter 16

Multi-way Resolution and Calibration Methods



16.1 Introduction

In the field of chemometrics, tensor algebra is usually used to classify the data generated by the instrument into zeroth order, first order, second order, and higher order. As shown in Fig. 16.1, the instrument responses correspondingly to each sample may be zeroth-order (scalar), first-order (vector), second-order (matrix), third-order (three-dimensional array), or higher-order tensor. Analyzing a series of such samples, one-dimensional, two-dimensional, three-dimensional, four-dimensional, and N-dimensional data will be generated [1, 2]. The method to predict unknown samples by zeroth-order tensor data prediction system is called univariate calibration method, and all the analysis methods using non-scalar data are called multivariate calibration methods. Second-order and higher-order tensor data is called multi-way data, and the method of processing these data is called the multi-way calibration method.

With the rapid development of modern hyphenated analytical instrument technology, more and more instruments produce two-dimensional or higher-dimensional response data, such as excitation-emission fluorescence spectrometer, chromatography-mass spectrometry and gas chromatography-infrared spectrometer, etc. When these instruments are used to measure a set of samples, a three-dimensional matrix will be obtained. Therefore, multi-way chemometrics analysis and calibration methods have emerged, such as the Tucker3 method, parallel factor analysis (PARAFAC), and alternating trilinear decomposition (ALTD). These methods have strong resolution and analysis abilities. It can distinguish the response signals of multiple analytes with similar properties at the same time when the presence of interfering substances is unknown, and directly determine the analyte components of interest quantitatively [3–6].

The multi-way data matrix is usually generated in spectral analysis. For example, as shown in Fig. 16.2, for a group of samples, their spectra are measured under different measurement conditions (such as pH or temperature, etc.), and a three-dimensional data matrix \mathbf{X} ($\mathbf{I} \times \mathbf{J} \times \mathbf{K}$) is obtained. \mathbf{I} is the number of samples, \mathbf{J} is the

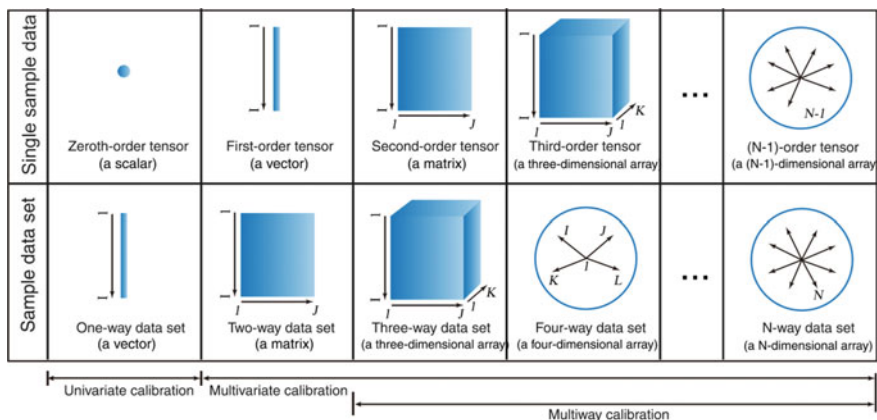


Fig. 16.1 Schematic diagram of expression “order” and “way” for analytical data

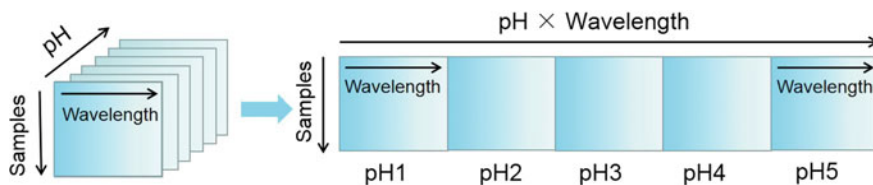
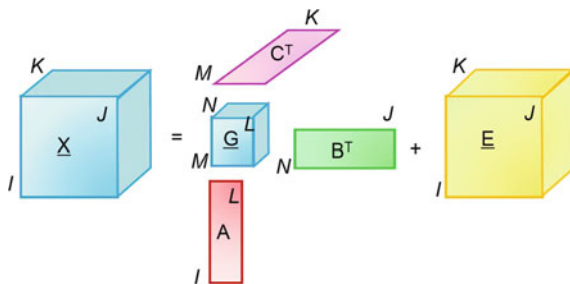


Fig. 16.2 Three-way spectra matrix obtained from a group of samples at six different pH values and its unfolded matrix

number of wavelength points, and K is the number of measurement conditions (such as six different pH values). Each element of \mathbf{X} can be expressed as x_{ijk} , which represents the absorbance of the i th sample at j wavelength under k condition. Likewise, the excitation-emission matrix fluorescence spectrometer (EEM) and spectral chemical imaging (infrared, near-infrared, Raman imaging, etc.) also obtain multi-way data matrices. The easiest way to deal with these data is to use unfolding strategy, that is, paving the cubic matrix \mathbf{X} ($I \times J \times K$) into an $I \times JK$ or $I \times KJ$ two-dimensional matrix, and then conduct PCA or PLS for analysis. However these methods tend to lose much information about a three-dimensional data structure.

In order to obtain more valuable results, people have improved the classical method for two-dimensional matrix resolution and calibration. For example, the multivariate curve resolution-alternating least squares (MCR-ALS) method is used to expand the three-dimensional data matrix into a two-dimensional data matrix that conforms to the bilinear structure and then analyzes it. It has a certain ability to overcome nonlinearity. However, in the iterative solution process, it is necessary to add non-negative constraints, unimodal constraints, and other constraints to obtain the analytical results with chemical significance. Unfolded partial least squares/residual bilinearization (U-PLS/RBL) expands the three-dimensional matrix into vector data, and then uses U-PLS and other methods to model the calibration sample to obtain

Fig. 16.3 Schematic diagram for decomposition of three-way data matrix by Tucker3 algorithm



the model parameters of each analyte. Finally the obtained model parameters are combined with residual bilinearity for the final quantitative analysis.

Tucker method is a classical three-dimensional data analysis method, which is an extension of traditional PCA. It was proposed by psychologist L. R. Tucker in 1963. Tucker3 decomposes the three-dimensional data matrix \mathbf{X} ($I \times J \times K$) into the product of three matrices \mathbf{A} ($I \times L$), \mathbf{B} ($J \times N$), and \mathbf{C} ($K \times M$) and a core matrix \mathbf{G} ($L \times N \times M$) (Fig. 16.3). L , N , and M are the number of factors, respectively.

$$x_{ijk} = \sum_{f=1}^L \sum_{g=1}^N \sum_{h=1}^M a_{if} b_{jg} c_{kh} g_{fgh} + e_{ijk} \quad (16.1)$$

Three load matrices \mathbf{A} ($I \times L$), \mathbf{B} ($J \times N$), and \mathbf{C} ($K \times M$) obtained by decomposition represent the number of rows, columns, and layers of \mathbf{X} , and the number of factors is L , N , and M , respectively. They are generally different, but they are less than the corresponding dimension of \mathbf{X} , which meets the purpose of data dimensionality reduction. In analytical chemistry, this algorithm does not have many practical application, so it is not often recommended.

The following mainly introduces three methods that are commonly used in trilinear decomposition. They are the PARAFAC method, alternating trilinear decomposition (ALTD), and multi-way partial least squares (N-PLS).

16.2 Parallel Factor Analysis

The parallel factor analysis (PARAFAC) method is a trilinear model, which was proposed by Harshman in 1970 and then used in psychology earlier [7]. The PARAFAC algorithm decomposes the three-dimensional data matrix \mathbf{X} ($I \times J \times K$) into the product of three two-dimensional matrices \mathbf{A} ($I \times N$), \mathbf{B} ($J \times N$), and \mathbf{C} ($K \times N$) (Fig. 16.4), and N is the factor number. For EEM data matrix, I represents the number of excitation wavelengths, J represents the number of emission wavelengths, K represents the number of samples, and N represents the component number of groups with response signals in the model, which includes the target

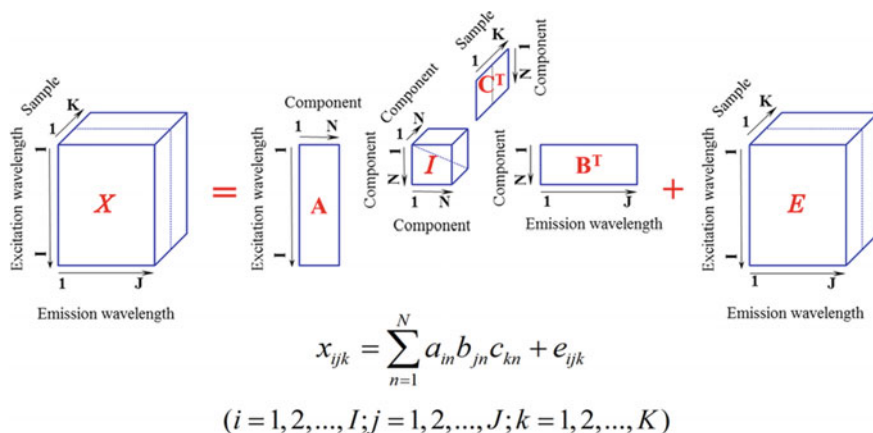


Fig. 16.4 Schematic diagram for decomposition of three-way data matrix by PARAFAC algorithm

analytes, the changing background, and other interferences. Matrix **A** is normalized excitation spectra, matrix **B** is normalized emission spectra, and matrix **C** is relative concentration matrix.

Comparing with the Tucker3 method, although the PARAFAC model is a special case of the Tucker3 model when $G = I$ and $L = N = M$. However, the essence of the Tucker model is the principal component model of a three-dimensional matrix, which is the result of calculating eigenvalues and principal components of the three-dimensional matrix, and its principal component have only mathematical significance. The PARAFAC model is a trilinear model, and the trilinear component model is the response summation model of the main components in the three-dimensional matrix. The simplest case conforms to the mathematical expression of Lambert-Beer's law, and the main components obtained have physical or chemical significance.

The trilinear decomposition model is very popular in analytical chemistry. One of the important reasons is that it is consistent with Beer's law in analytical chemistry. Thus the trilinear decomposition model has its corresponding chemical background. Another important reason is that the trilinear decomposition model of three-dimensional data generated under general analytical conditions is unique, and its decomposition results directly correspond to the qualitative (chromatographic or spectral) and quantitative information (concentration) of chemical components in the system. This method has a strong resolution ability. It has the so-called "second-order advantage", that is, it can simultaneously resolve the response signals of multiple analytes with similar properties in the presence of unknown interferences, and can directly determine the components of analytes of interest.

PARAFAC operation is realized by alternative least squares (ALS) algorithm, whose goal is to minimize the sum of squares of residuals (SSR):

$$\text{SSR} = \sum_{i=1}^I \sum_{j=1}^J \sum_{k=1}^K e_{ijk}^2 \quad (16.2)$$

The specific iterative process of the PARAFAC algorithm is as follows:

(1) Give the number of principal factors N , initialize \mathbf{B} and \mathbf{C} .

(2) Obtain \mathbf{d}_k :

$$\mathbf{d}_k = \{(\mathbf{B}^T \mathbf{B}) * (\mathbf{A}^T \mathbf{A})\}^{-1} \{(\mathbf{A}^T \mathbf{X}_{..k} \mathbf{B}) * \mathbf{I}\} \quad (16.3)$$

(3) Obtain matrix \mathbf{A} :

$$\mathbf{A} = \left(\sum_{k=1}^K \mathbf{X}_{..k} \mathbf{B} \mathbf{D}_k \right) \{(\mathbf{C}^T \mathbf{C}) * (\mathbf{B}^T \mathbf{B})\}^{-1} \quad (16.4)$$

(4) Obtain matrix \mathbf{B} :

$$\mathbf{B} = \left(\sum_{k=1}^K \mathbf{X}_{..k} \mathbf{A} \mathbf{D}_k \right) \{(\mathbf{C}^T \mathbf{C}) * (\mathbf{A}^T \mathbf{A})\}^{-1} \quad (16.5)$$

(5) Repeat steps (2)–(4) until convergence.

The convergence criterion is

$$\left| \frac{\text{SSR}^{(m)} - \text{SSR}^{(m-1)}}{\text{SSR}^{(m-1)}} \right| \leq \varepsilon \quad (16.6)$$

where m is the number of iterations and ε is the threshold (usually 1×10^{-6}).

The matrix \mathbf{C} obtained by PARAFAC decomposition can be correlated with the concentration vector \mathbf{y} to establish a quantitative correction model: $\mathbf{y} = \mathbf{C}\mathbf{b}$, and the regression coefficient \mathbf{b} can be obtained by the least square method. For unknown samples, matrix \mathbf{C} is first obtained from the loading matrices \mathbf{A} and \mathbf{B} , and the final result is calculated from the regression coefficient \mathbf{b} . The PARAFAC method is also used for the spectral analysis of the reaction process and the determination of the chemical reaction rate constant.

PARAFAC is based on the strict sense of the least square principle to optimize the fitting of trilinear data. In theory, the least and most stable model error should be obtained. However there are also some unsatisfactory places. For example, the estimation of the number of principle components that are too sensitive, vulnerable to the impact of random initialization values, and the convergence speed is slow.

16.3 Alternating Trilinear Decomposition

Alternating trilinear decomposition (ALTD) is an improvement of the PARAFAC method proposed by Wu et al. This method uses an iterative strategy based on improved tailed singular value decomposition to calculate Moore-Penrose generalized inverse and extract diagonal elements. In addition, ATLD uses the trilinear component model of slice matrix to calculate, which reduces the memory required for calculation, improves the efficiency of operation, and has the advantage of fast convergence.

The three objective functions of ATLD are as follows:

$$\sigma(a_{(i)}) = \sum_{i=1}^I \|X_{i..} - B \text{diag}(a_{(i)}) C^T\|_F^2 \quad (16.7)$$

$$\sigma(b_{(j)}) = \sum_{j=1}^J \|X_{.j.} - C \text{diag}(b_{(j)}) A^T\|_F^2 \quad (16.8)$$

$$\sigma(c_{(k)}) = \sum_{k=1}^K \|X_{..k} - A \text{diag}(c_{(k)}) B^T\|_F^2 \quad (16.9)$$

Based on the principle of the least square method, the following solutions of **A**, **B**, and **C** can be obtained by alternately minimizing the above objective function. The specific steps of the ATLD algorithm are as follows:

- (1) Determine the factors of the system.
- (2) Randomly initialize matrices **A** and **B**.
- (3) Calculate the matrix **C** by the following equation:

$$c_{(k)}^T = \text{diag}(A^+ X_{..k} (B^T)^+) \quad (k = 1, 2, \dots, K) \quad (16.10)$$

where $\text{diag}(\cdot)$ represents the construction of a diagonal matrix, whose elements are 0 except for the diagonal elements. $\text{diag}(\cdot)$ means that the elements on the diagonal of the matrix are extracted into a column vector.

- (4) Calculate **A** by the following formula and normalize **A** by column:

$$a_{(i)}^T = \text{diag}(B^+ X_{i..} (C^T)^+) \quad (i = 1, 2, \dots, I) \quad (16.11)$$

- (5) **B** is calculated by the following formula, and normalize **B** by column:

$$b_{(j)}^T = \text{diag}(C^+ X_{.j.} (A^T)^+) \quad (j = 1, 2, \dots, J) \quad (16.12)$$

- (6) Calculate **C** by matrices **A** and **B** according to the following equation:

$$c_{(k)}^T = \text{diagm}(A^+ X_{\cdot k} (B^T)^+) \quad (k = 1, 2, \dots, K) \quad (16.13)$$

(7) Repeat steps (4–6) until the convergence standard.

The convergence criterion is

$$\left| \frac{\text{SSR}^{(m)} - \text{SSR}^{(m-1)}}{\text{SSR}^{(m-1)}} \right| \leq \varepsilon \quad (16.14)$$

where m is the number of iterations and ε is the threshold (usually 1×10^{-6}). To avoid the slow convergence caused by falling into an exception, set the maximum number of iterations to 3000.

In each iteration, matrices \mathbf{A} and \mathbf{B} are normalized by column. By resolving the corresponding matrix, the concentration of analyte can be obtained by linear regression between the relative concentration and the real concentration of each analyte corresponding to the column in matrix \mathbf{C} .

Due to the advantages of being insensitive to the over-estimated factors and fast convergence, the ATLTD algorithm has been applied in many fields such as spectroscopy, chromatography, and electrochemistry to solve the problems of overlapping peaks and uncalibrated interferences. However, ATLTD is sensitive to noise, it should be used carefully in the analysis system with low signal-to-noise ratio (SNR).

On the basis of the ATLTD method, a series of derivative methods have been developed [9]. Among them, a representative method is self-weighted alternating trilinear decomposition (SWATLTD) proposed by Chen et al. This algorithm designs a unique objective function by introducing the idea of weight, which not only maintains the advantages of ATLTD but also has better stability, stronger ability to resist noise, and collinearity. The alternating penalty trilinear decomposition (APTLD) algorithm proposed by Charalin uses penalty factors to combine PARAFAC and SWATLTD. It has the advantages of both methods and is more flexible.

16.4 Multi-way Partial Least Squares

Multi-way partial least squares (N-PLS) is a three-dimensional matrix calibration algorithm based on trilinear decomposition and classical PLS proposed by Bro et al. [10]. It has been applied to establish quantitative calibration models for EEM, GC-MS and QSAR with satisfactory results [11-13]. The principle of the N-PLS algorithm is to decompose three-dimensional matrix \mathbf{X} ($I \times J \times K$) into a trilinear model:

$$\mathbf{X}_{ijk} = \sum_{f=1}^F \mathbf{t}_{if} \mathbf{W}_{if} \mathbf{W}^k k_f + \mathbf{e}_{ijk} \quad (16.15)$$

where \mathbf{t} is the score vector, \mathbf{w}^J and \mathbf{w}^K are the corresponding two loading vectors, F is the number of principal components, and \mathbf{e}_{ijk} is the residual matrix. Similar to the traditional PLS, N-PLS not only decomposes the spectral matrix \mathbf{X} but also decomposes the concentration vector \mathbf{y} , and the two decomposition processes of \mathbf{X} and \mathbf{y} are combined into one by iteration. The specific algorithm of the N-PLS is as follows:

(1) Calibration section.

\mathbf{X} ($I \times J \times K$) is the calibration set spectral matrix, I is the number of calibration set samples, J is the number of wavelength points, and K is the number of conditions for spectral measurement (such as pH or temperature). \mathbf{y} ($I \times 1$) is the concentration vector of the calibration set.

① Unfold \mathbf{X} into two-dimensional matrix $\mathbf{X}0$ ($I \times JK$), and determine the number of factors $F, f = 1, \dots, F$.

② Calculate \mathbf{Z} matrix, $\mathbf{Z}_f = \mathbf{X}_{f-1}^T \mathbf{y}$;

③ Perform SVD on \mathbf{Z} matrix, $[\mathbf{w}^K, \mathbf{s}, \mathbf{w}^J] = \text{svd}(\mathbf{Z}_f)$.

④ Calculate \mathbf{w} , $\mathbf{w}_f = \mathbf{w}^K \otimes \mathbf{w}^J$.

Symbol \otimes represents the Kronecker product of matrix, and the Kronecker product of matrix \mathbf{A} ($I \times J$) and matrix \mathbf{C} ($M \times N$) is expressed as

$$\mathbf{A} \otimes \mathbf{C} = \begin{bmatrix} a_{11}C & \cdots & a_{1J}C \\ \vdots & \ddots & \vdots \\ a_{I1}C & \cdots & a_{IJ}C \end{bmatrix} \quad (16.16)$$

⑤ Calculate \mathbf{t}

$$\mathbf{t}_f = \mathbf{X}_{f-1} \mathbf{W}_f \quad (16.17)$$

⑥ Calculate \mathbf{q}

$$\mathbf{q}_f = \mathbf{y}_{f-1}^T \mathbf{t}_f \quad (16.18)$$

⑦ Calculate \mathbf{u}

$$\mathbf{u}_f = \mathbf{y}_{f-1} \mathbf{q}_f \quad (16.19)$$

⑧ Calculate \mathbf{b}

$$\mathbf{b}_f = (\mathbf{T}_f^T \mathbf{T}_f)^{-1} \mathbf{T}_f^T \mathbf{u}_f. \quad (16.20)$$

where $\mathbf{T}_f = [\mathbf{t}_1, \dots, \mathbf{t}_f]$.

⑨ Update \mathbf{X} and \mathbf{y}

$$\mathbf{X}_f = \mathbf{X}_{f-1} - \mathbf{t}_f \mathbf{w}_f \tag{16.21}$$

$$\mathbf{y}_f = \mathbf{y}_{f-1} - \mathbf{T}_f \mathbf{b}_f \mathbf{q}_f^T \tag{16.22}$$

⑩ Let $f = f + 1$, and return 3, to get the F scores and loadings of \mathbf{X} and \mathbf{y} in turn.

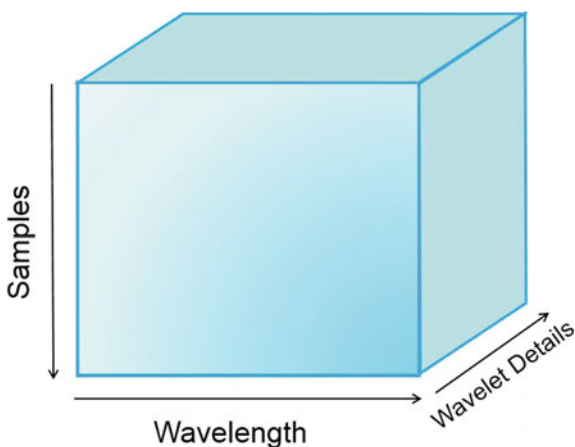
(2) Prediction section

For an unknown sample spectral matrix $\mathbf{X}^{\text{un}}(1 \times J \times K)$, the prediction results are calculated by the following steps:

- ① Unfold the \mathbf{X}^{un} into a two-dimensional matrix $\mathbf{X}^{\text{un}} (1 \times JK)$;
- ② Calculate $\mathbf{t}_f = \mathbf{X}^{\text{un}} \mathbf{w}_f$, $\mathbf{X}^{\text{un}}_f = \mathbf{X}^{\text{un}}_{f-1} - \mathbf{t}_f \mathbf{w}_f, f = 1, \dots, F$;
- ③ Calculate $\mathbf{y}_{\text{pred}} = \sum_{f=1}^F \mathbf{T}_f \mathbf{b}_f \mathbf{q}_f^T$, where $\mathbf{T}_f = [\mathbf{t}_1, \dots, \mathbf{t}_f]$.

Chu et al. combined wavelet transform (WT) with multi-way partial least squares (N-PLS) method and proposed a new method for establishing the quantitative calibration model of near-infrared spectroscopy [14]. The basic idea of the method is as follows: firstly, the spectrum of each sample in the calibration set is transformed by wavelet transform, and then a set of wavelet detail coefficients is selected according to the specific application. The three-dimensional spectral matrix \mathbf{X} is as shown in Fig. 16.5 ($I \times J \times K$, where I was the number of samples in the calibration set, J was the number of selected wavelet details, and K was the number of wavelength points). Finally, the calibration model was established by the N-PLS method. For unknown samples, the three-dimensional spectral matrix was first formed by wavelet transform, and then the established calibration model for prediction analysis. For example, in order to establish a robust temperature calibration model of near-infrared spectra, a set of wavelet detail coefficients with small temperature influence and strong information can be selected to form a three-dimensional matrix, and the quantitative calibration model can be established by the N-PLS method. The results show that

Fig. 16.5 Schematic diagram of three-way NIR spectral matrix decomposed by wavelet transform



the calibration model established by this method has a better prediction ability and robustness than the ordinary WT-PLS method.

For vibrational molecular spectroscopy, such as mid-infrared and near-infrared spectroscopy (Fig. 16.6), the fundamental frequency and different overtone absorp-

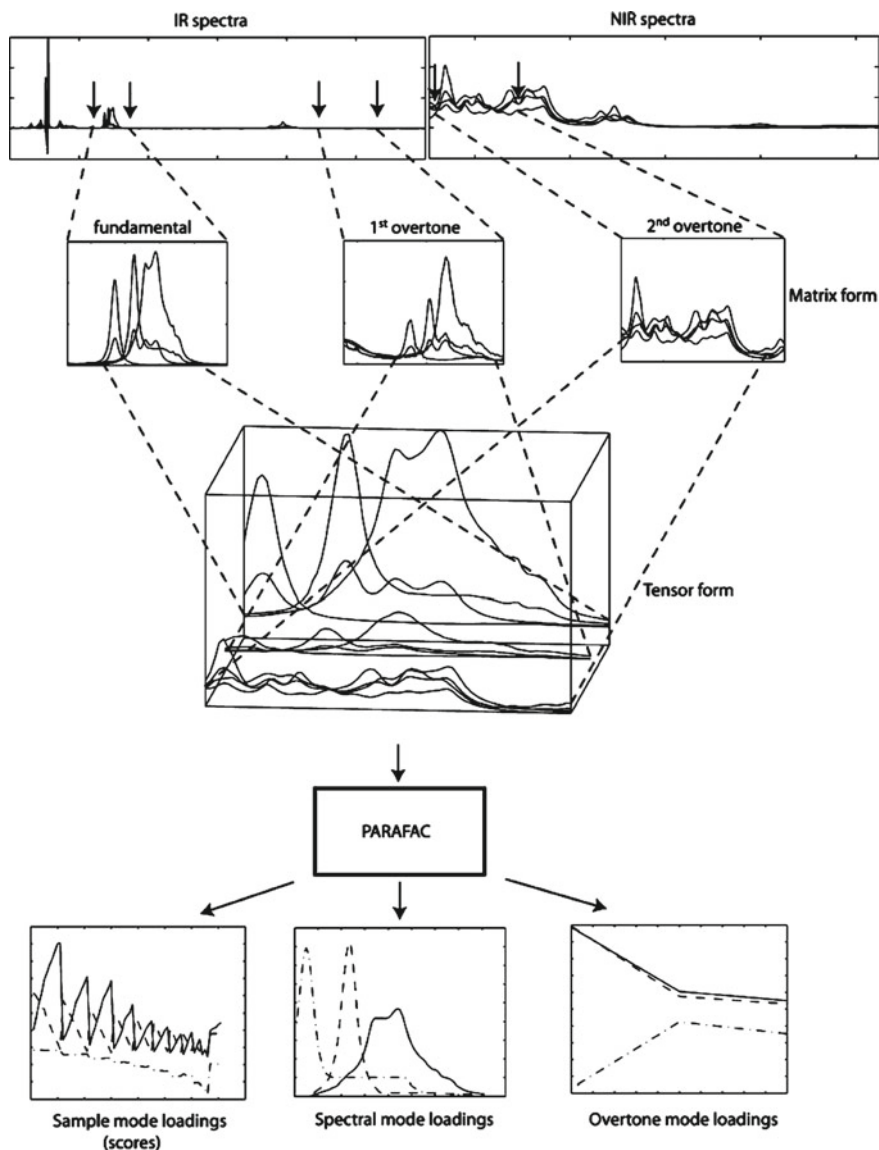


Fig. 16.6 Schematic diagram for multi-way calibration of vibrational overtone combination spectroscopy

tion bands can be stacked to form a three-dimensional spectral matrix, and then quantitative and qualitative analyses can be carried out by multi-way data resolution and calibration methods [15].

References

1. Wu H L, Wang T, Yu R Q. Recent Advances in Chemical Multi-Way Calibration with Second-Order or Higher-Order Advantages: Multilinear Models, Algorithms, Related Issues and Applications. *Trends in Analytical Chemistry*. 2020;130:115954.
2. Wu HL, Li Y, Kang C, et al. Research progress of three-dimensional fluorescence coupled with chemical multi-way calibration. *Chin J Anal Chem*. 2015;43(11):1629–37.
3. Liang YZ, Wu HL, Shen GL, et al. some new advances in analytical chemometrics. *Sci China Ser B Chem*. 2006;36(2):93–100.
4. Escandar GM, Olivieri AC, Faber NM, et al. Second-and third-order multivariate calibration: data, algorithms and applications. *Trends Anal Chem*. 2007;26(7):752–65.
5. Smilde A, Bro R, Geladi P. Multi-way analysis: applications in the chemical sciences. Wiley. 2004.
6. De La Pena AM, Goicoechea HC, Escandar GM, et al. Fundamentals and analytical applications of multi-way calibration. Elsevier. 2015.
7. Parafac BR. Tutorial and applications. *Chemom Intell Labor Syst*. 1997;38(2):149–71.
8. Wu HL, Xiao R, Hu Y, et al. The application of chemical multidimensional correction methods in drug analysis. *J Pharm Anal*. 2019;39(4):565–79.
9. Liang YZ, Wu HL, Yu RQ. Analytical chemistry manual 10-chemichitrics (Third edition). Beijing: Chemical Industry Press; 2016.
10. Moros J, Iñón FA, Garrigues S, et al. Determination of vinegar acidity by attenuated total reflectance infrared measurements through the use of second-order absorbance-ph matrices and parallel factor analysis. *Talanta*. 2008;74(4):632–41.
11. Sena MM, Poppi RJ. N-way PLS applied to simultaneous spectrophotometric determination of acetylsalicylic acid, paracetamol and caffeine. *J Pharm Biomed Anal*. 2004;34(1):27–34.
12. Matero S, Pajander J, Soikkeli AM, et al. predicting the drug concentration in starch acetate matrix tablets from ATR-FTIR spectra using multi-way methods. *Anal Chim Acta*. 2007;595(1–2):190–7.
13. Prazen BJ, Johnson KJ, Weber A, et al. Two-dimensional gas chromatography and trilinear partial least squares for the quantitative analysis of aromatic and naphthene content in Naphtha. *Anal Chem*. 2001;73:5677–82.
14. Chu XL, Tian GY, Yuan HF, et al. Quantitative analysis of near-infrared spectroscopy by combination of wavelet analysis and n-way partial least square. *Chinese J Anal Chem*. 2006;34 (Special issue):S175–S178.
15. Alm E, Bro R, Engelsen SB, et al. Vibrational overtone combination spectroscopy (VOCSY- a new way of using IR and NIR data. *Anal Bioanal Chem*. 2007;388(1):179–88.

Chapter 17

Calibration Transfer Methods



17.1 Introduction

A situation is often encountered in the process of spectral analysis that when the model established on one spectrometer (master, primary, parent instrument) is used on another spectrometer (slave, child instrument), the model cannot give correct prediction results due to the difference in spectra measured by different instruments [1, 2]. The first thing to solve this problem is to improve the standardization of instrument hardware processing, improve the level of processing technology, and reduce the differences between the master and the slaves in terms of devices. Instrument standardization makes the spectra from the same samples measured by different instruments as consistent as possible. There have been many reports on the standardization of spectroscopic instrument hardware such as calibrating the wavelength accuracy, absorbance accuracy, resolution, spectral response line type, and symmetry of the spectrometer through sharp-line emission light source, standard materials, etc. [3–9]. As well, there have been developments on instruments by optimizing optical components, assembly processes, and control strategies with performance indicators within the allowable range of variation [10, 11]. This is called the First Principle of instrument calibration, which is the most fundamental basis of modern spectroscopic analysis technology [12, 13].

For Fourier instruments of the same or even different types, it is basically possible to directly transfer the spectra through the instrument calibration method [14–19]. In recent years, some portable instruments can also transfer spectra between the same type [20, 21]. Surely, different applications have various requirements for the consistency of instruments [22–24]. Calibration models established by different methods and different systems also have different tolerances for divergence between instruments [25–29]. Due to the sharp peaks of the spectrum, Raman spectrometers can better solve the problem of consistency between instruments via hardware [30–33]. Besides spectroscopic instruments, other instruments like mass spectrometers have also encountered similar problems, but most of which can be solved by simple peak calibration [34, 35].

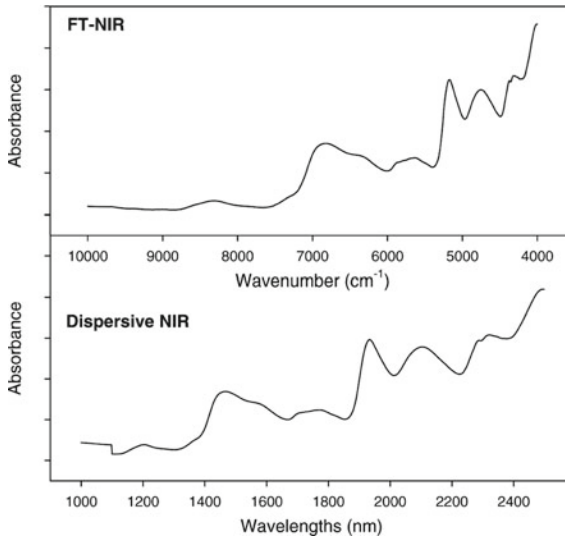


Fig. 17.1 Near-infrared spectra of barley samples on dispersive and Fourier transform instruments

Although there has been a development for decades, differences between different brands of instruments still exist, e.g., the difference between a grating-type and a Fourier transform-type spectrometers. Due to these differences, the inapplicability of the multivariate models would produce unacceptable systematic prediction bias [36–38]. Thus, the solution to this prediction bias is called calibration transfer or instrument transfer [39–42]. In the area of machine learning, the relevant keywords are transfer learning, domain adaptation, multi-task learning, etc. [43] (Fig. 17.1).

Calibration transfer mainly includes the following three categories of solutions [43–46]:

(1) Transfer in spectra

Transfer between the spectra establishes the functional relationship between the spectra measured by the master and slave through a mathematical method. There are two ways of spectral transfer. One is to convert the master’s calibration spectra and then re-establish a calibration model suitable for the slave spectra, which is called reverse standardization in the literatures [47–49]. The other is to convert the slave spectra, and directly use the master model to predict the result. The algorithm of two ways is essentially the same and needs to be selected according to different applications.

(2) Regression coefficient conversion

Regression coefficient conversion is conducted to the master model so as to make it suitable for the slave spectra [50]. Also, prediction results of the master model can be corrected, such as the slope/bias correction (SBC) method, etc., to eliminate the systematic deviation of the prediction results [51, 52].

(3) Robust calibration

Robust calibration is established through preprocessing, wavelength variable screening algorithms, etc. Or the master calibration set is expanded by adding spectra under different test conditions, spectra from the slave, etc., to establish a global calibration, or hybrid calibration [53, 54]. In some cases, this method is also called model updating or calibration maintenance. For instance, with the aging of the electronic components, detectors, and optical components of the instrument, as well as changes in the instrument hardware caused by the environment, these inevitable changes in the instrument will lead to changes in the spectra [55]. In practice, the most commonly used methods are the first and third ones [56, 57], the second is rarely used. This chapter mainly introduces the first and third types of calibration transfer methods.

17.2 Traditional Algorithms

Commonly used methods for converting spectra among different instruments (or under different conditions) include spectral subtraction correction (SSC), Shenk's algorithm, direct standardization (DS), piecewise direct standardization (PDS), etc. These methods usually require a representative set of standard samples (15–30 samples), which are called calibration transfer methods with standard samples.

17.2.1 Spectral Subtraction Correction

Calculate the average spectrum \bar{x}_{ms} and \bar{x}_{ss} of the master standard sample spectral matrix X_{ms} and slave matrix X_{ss} , respectively, and then calculate the difference between the average spectra of them $\Delta\bar{x} = \bar{x}_{ms} - \bar{x}_{ss}$.

For the unknown spectrum $x_{s,um}$ from the slave spectroscopy, convert with formula $x_{s,um}^p = x_{s,um} + \Delta\bar{x}$, acquire the spectrum $x_{s,um}^p$ consistent with $x_{m,um}$, and the final result is calculated by the calibration model established by the master [58].

17.2.2 Shenk's Algorithm

Shenk's algorithm consists of two main steps: wavelength correction and absorbance correction [59–61]. The following introduces this algorithm by taking the transfer of spectrum from the slave to master (slave \rightarrow master) as an example.

(1) Wavelength correction

- a. Corresponding to $X_{ms,i}$, choose the spectral part $X_{ss,j+k+1}$ with window size of $(k + j + 1)$ from the matrix X_{ss} in the slave instrument, and calculate the correlation coefficients of $X_{ms,i}$ and $X_{ss,i-j}, X_{ss,i-j+1}, \dots, X_{ss,i+k-1}, X_{ss,i+k}$, respectively. If the coefficient r_l of $X_{ms,i}$ ($i-j \leq l \leq i+k$) and $X_{ms,i}$ is max, it is indicated that the l wavelength from the slave instrument corresponds to the i wavelength from the master. To obtain more accurate results, wavelengths $l-1, l$, and $l+1$ and their corresponding correlation coefficients r_{l-1}, r_l , and r_{l+1} are selected to establish a univariate quadratic parabola model as $r = a_i + b_i i + c_i i^2$. The slave wavelength i^* corresponding to the master wavelength i would be obtained from this fitted parabola.
- b. Loop i to find all corresponding i^* .
- c. Establish the one-dimensional quadratic parabola wavelength calibration model $i^* = A + Bi + Ci^2$ using the obtained i^* and i .

(2) Absorbance correction

After wavelength correction, calculate the absorbance matrix X_{ss,i^*} of the slave wavelength i^* by the interpolation method, then find the regression coefficients sa_i and sb_i using the linear equation $X_{ms,i} = sa_i + sb_i X_{ss,i^*}$.

For the unknown spectrum $x_{s,un}$ from the slave instrument, wavelength fitting calibration curve $i^* = A + Bi + Ci^2$ is used to calibrate the wavelength, calculate $x_{s,un}^*$ by the interpolation method, and finally, the spectral transfer result consistent with the master is obtained by $x_{s,un}^p = sa_i + sb_i x_{s,un}^*$.

17.2.3 Direct Standardization

DS algorithm uses the transfer matrix F to convert the unknown sample spectrum $x_{s,un}$ measured from the slave machine to $x_{s,un}^p$. Transfer matrix F is calculated by $X_{ms} = X_{ss}F$ through the least square as $F = X_{ss}^+ X_{ms}$ [62, 63], where X_{ss}^+ is the generalized inverse matrix of X_{ss} , F is an $m \times m$ -dimensional matrix (m is the number of wavelength variables).

For the spectrum $x_{s,un}$ measured from the slave machine, the transfer is conducted by $x_{s,un}^p = x_{s,un}F$, and the final result is calculated by the calibration model established by the master.

17.2.4 Piecewise Direct Standardization

In the PDS algorithm, as shown in Fig. 17.2, standard sample spectral matrix $X_{ss,j+k+1}$ with the window width $(j + k + 1)$ on both sides of the i th wavelength point from the

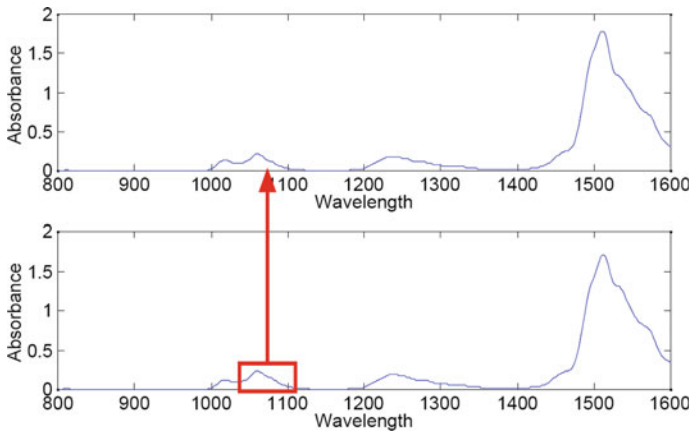


Fig. 17.2 PDS algorithm scheme

slave instrument $((i-j)$ th to $(i+k)$ th points) and standard sample spectral matrix $X_{ms,i}$ from the master i th wavelength point are used to calculate the transfer coefficient F_i of the i th wavelength point. Then, the transfer matrix F of all wavelengths is obtained by moving i point by point [64–67].

When calculating the transfer matrix F , besides the PLS method, as shown in Fig. 17.3, methods such as artificial neural networks can also be used [68, 69]. Moreover, spectra can also be transferred and transmitted in the Fourier transform domain or the wavelet domain, etc. [70, 71].

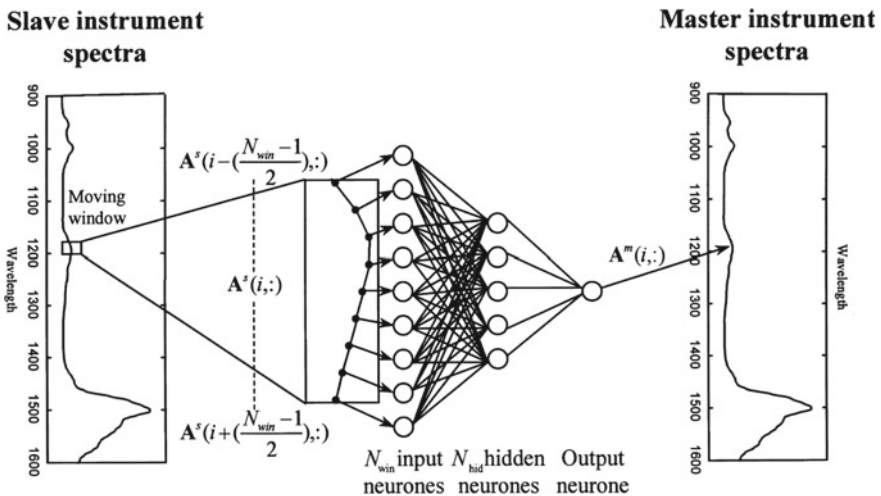


Fig. 17.3 Schematic diagram of PDS algorithm based on artificial neural network

17.2.5 Procrustes Analysis

Statistically, Procrustes analysis is used to compare two matrices $X_1(m \times p_1)$ and $X_2(m \times p_2)$, and to find the transfer matrix F between the matrices X_1 and X_2 , where m is the samples number, and p_1 and p_2 are the variables number.

A detailed algorithm is as follows [72, 73]:

- (1) Perform singular value decomposition on the matrices X_1 and X_2 , respectively, $X_1 = U_1 S_1 V_1^t$, $X_2 = U_2 S_2 V_2^t$, where X_1 and X_2 represent the spectral matrix measured by the master and slave (after averaging or standardization pretreatment), U is score matrix, V is loading matrix, and U and V matrices contain the rotation information between the spectral matrices. S matrix contains the stretching information between the spectral matrices.
- (2) Calculate $Z_1 = U_{1g} S_{1g}$ and $Z_2 = U_{2h} S_{2h}$, separately, where g and h represent the number of PCs used to find Z_1 and Z_2 , respectively.
- (3) Calculate transfer matrix F between Z_1 and Z_2 by $F = Z_2^+ Z_1$, where Z_2^+ is the generalized inverse matrix calculated by $Z_2^+ = S_{2g}^{-1} U_{2g}^t$.
- (4) A spectrum x_{un} measured from a slave can be transferred into spectrum x_{un}^P consistent with the master through the transfer matrix F and the loading matrix V .

17.2.6 Target Transformation Factor Analysis

Target transformation factor analysis (TTFA) is also a transfer method based on PCA. Its core concept is to use the target transfer method to make the principal component score (virtual component concentration) of the slave equal to that of the master. The main steps are as follows [74]:

- (1) Perform PCA on the standard sample spectral matrix of the master to obtain the load and score matrix $X_m = T_m P_m^t$
- (2) Perform PCA on the standard sample spectral matrix of the slave to obtain the load and score matrix $X_s = T_s P_s^t$
- (3) Establish the mathematical relationship between master and slave $T_m = T T_s$

$$T = T_m T_s^t (T_s T_s^t)^{-1}$$

Solve the transformation matrix by generalized inverse operation

$$T = T_m T_s^t (T_s T_s^t)^{-1}$$

- (4) Transfer of spectrum from master to slave can be expressed as $X_s^P = X_m P_m T^+ P_s^t$ where the transfer matrix $F = P_m T^+ P_s^t$.

17.2.7 Maximum Likelihood Principal Component Analysis

Maximum likelihood principal component analysis (MLPCA) treats calibration transfer as an issue of missing data [75].

Combine the standard sample spectra of master and slave: $\mathbf{x}_{i,comb} = [\mathbf{x}_{i,m}, \mathbf{x}_{i,s}]$, $i = 1, 2, \dots, n$, where n is the number of standard samples.

Besides, other samples of the calibration set on the master that has no corresponding spectra on the slave can be expressed as $\mathbf{x}_{i,comb}^\# = [\mathbf{x}_{i,m}, \mathbf{x}_{i,s}^\#]$, $i = 1, 2, \dots, m$, where m denotes the number of calibration set samples minus standard set samples, and $\mathbf{x}_{i,s}^\#$ denotes the missing spectra from the slave.

$\mathbf{x}_{i,comb}$ and $\mathbf{x}_{i,comb}^\#$ can be expressed as a matrix $\mathbf{X}_{comb} = \begin{bmatrix} \mathbf{X}^* \\ \mathbf{X}^\# \end{bmatrix}$, where \mathbf{X}^* represents combined spectral matrix without missing data in the standard sample set. $\mathbf{X}^\#$ represents combined spectral matrix of other samples in the calibration set with missing data in the slave.

Perform MLPCA on \mathbf{X}_{comb} : $\begin{bmatrix} \mathbf{X}^* \\ \mathbf{X}^\# \end{bmatrix} = \begin{bmatrix} \mathbf{U}^* \\ \mathbf{U}^\# \end{bmatrix} \mathbf{D} \mathbf{P}^t$.

Then, the spectrum $\mathbf{x}_{i,m}$ collected on the master can be transferred into the spectrum on the slave according to the following formula:

$$\hat{\mathbf{x}}_{i,s} = \mathbf{U}^\# (\mathbf{U}^{*t} \mathbf{U}^*)^{-1} \mathbf{U}^{*t} \mathbf{x}_{i,m} \quad (17.1)$$

Afterward, based on the MLPCA algorithm, Folch-Fortuny et al. [76] used the trimmed scores regression (TSR) method to deal with the issue of missing data and proposed the TSR method for calibration transfer.

17.2.8 Slope/Bias Correction

Different from the above methods that are based on the transfer between spectra, there is a method based on the transfer between the prediction results, that is, the functional relationship between the prediction results obtained by the master and the slave is established, which is named slope/bias correction (SBC) algorithm [77–79].

The calibration model established on the master is used to predict the analysis results \mathbf{y}_{mp} and \mathbf{y}_{sp} of the master and slave standard sample spectral arrays \mathbf{X}_{ms} and \mathbf{X}_{ss} , respectively. Assume that \mathbf{y}_{mp} and \mathbf{y}_{sp} have a relationship of $\mathbf{y}_{mp} = a \times \mathbf{y}_{sp} + b$. The least squares method can be used to obtain a and b .

As for the unknown spectrum $\mathbf{x}_{s,un}$ measured by the slave, firstly, the calibration model established by the master is used to calculate $\mathbf{y}_{sp,un}$, and then formula $\mathbf{y}_{sp,un}^P = b + a \times \mathbf{y}_{p,un}$ is used to calculate the corrected analysis result $\mathbf{y}_{sp,un}^P$.

Normally, the SBC algorithm is not recommended. Because if the spectral difference between instruments is significant, it would be difficult or even impossible to identify the outlier samples of the model.

17.3 Improvement of Traditional Algorithms

In the process of calibration transfer, the selection of standard samples is particularly important [80]. For the model update problem, Capron et al. [81] compared the effects of the weighting of calibration samples and the selection of representative samples, and the results showed that the selection effect of representative samples is better. Siano et al. [82] and Clark [83] compared the influence of different transfer standard selection methods on the spectrum transfer effect of PDS, etc., and the results showed that the optimal K-dissimilarity selection (OptiSim) method is better than the K-S method. Li et al. [84] replaced the Euclidean distance in the K-S algorithm with the Mahalanobis distance and selected the transfer standard sample through the improved K-S algorithm. In the PDS algorithm, the samples selected by the Mahalanobis distance are more representative, because the combination of concentration differences and spectral differences can better represent the differences between samples. Zhou et al. [85] proposed a transfer standard set selection method based on the Markov chain, and the result is better than the K-S method.

From the perspective of optimizing the selection method of the transfer standard set, Liang et al. [86] optimized the transfer matrix of the PDS algorithm and proposed the Rank-KS-PDS calibration transfer algorithm. When Rank-KS selects the transfer standard sample set, it comprehensively considers the influence of sample spectral space and sample concentration space, overcomes the shortcomings of the K-S algorithm insensitive to low concentration areas, and improves the accuracy of calibration transfer. Based on the idea of backward selection variables, Zheng et al. [87] proposed a backward selection iterative method to transfer the selection of the standard sample set. Through this method, the transfer effect of the standard sample set is better than the standard sample set selected by the KS algorithm. Aiming at the problem of difficulty in obtaining and storing standard samples of natural plant models, Ni et al. [88] proposed a method for preparing standard samples, and this kind of standard sample is stable under normal temperature and pressure, similar to the color of various natural plants, with the constant spectrum and good reproducibility, and can be used for a long time in the transfer process of the near-infrared model of a variety of natural plants.

Aiming at the limitations of the SBC algorithm in solving nonlinear problems, Xin et al. [89] established a linear function relationship between the prediction results of the master and the slave by introducing high powers, both Lagrange interpolation and Newton interpolation are used to find the parameters, and the nonlinear fitting of the two sets of data is realized. Cao et al. [90] proposed a method for selecting PDS algorithm parameters (number of standard samples, number of PLS main factors, and window width) based on the angle of the spectral space. Zhang et al. [91] used sampling error profile analysis (SEPA) to optimize the PDS algorithm's window width and PLS main factor number and other parameters, and proposed the SEPA-PDS method.

Blanco et al. [92, 93] used the spectral difference of a set of standard sample sets under different conditions combined with weighting to calculate a mutation matrix,

and randomly added it to the calibration set spectrum, which solved the problem of the laboratory-prepared drug calibration samples used for production process analysis. On this basis, Wang et al. [94] improved the SSC algorithm, the spectrum compensation between the master and the slave does not use the average spectral difference of the standard sample set spectrum array uniformly, but for each sample in the calibration set, select the most similar standard sample set to compensate for the spectral difference between the master and slave, and the compensation spectrum is weighted by the concentration ratio between the calibration set samples and the standard set samples. On the basis of the SSC algorithm, Li et al. [95] respectively compensated different correction vectors for different types of correction sets, and then through continuous model updates, the influence of different measurement environments on the identification of corn haploid grains by NIRS can be eliminated. Based on the traditional SBC algorithm, Wang et al. [96] proposed a dual-domain model delivery strategy. In this method, the master model is used to predict the spectra of the master and slave standard sample sets respectively, and the transfer model is established by using the ratio of the predicted value and the spectrum of the slave standard sample set. For the spectra collected by the slave, the initial value is calculated first using the master model, and then the ratio is calculated by the transfer model, and finally the final prediction result is obtained through the ratio. Li et al. [97] established a PLS model between the spectral difference and the predicted concentration difference between two NIRS instruments to achieve the correction of the predicted concentration from the slave machine. Tan et al. [98], Sum and Brown [99] improved the finite impulse response (FIR) transfer algorithm without standard samples, which eliminated the peak problem caused by the FIR algorithm and improved the transfer effect of the FIR method. Bouveresse et al. [100] used the local weight regression algorithm for the correction between spectral absorbance and improved Shenk's algorithm.

Before performing DS and PDS operations, Wang et al. [101] used additive background correction to improve the effect of spectral transmission. In response to the nonlinear transfer between spectra in the traditional PDS algorithm and the appearance of discontinuities and even abnormal peaks [102], Wang et al. [103] replaced the PLS regression in the PDS algorithm with a radial basis function neural network and obtained better results. Chen et al. [104] believe that the appearance of the peak in the PDS algorithm originates from the coefficient of the larger norm in the PLS model, and its essence is an over-fitting problem. They used a linear regression method with coefficient norm penalty to establish a spectral transfer model, and the transferred spectra were smoother and more robust.

Univariate correction is a special case where the PDS method takes a window width of 1. Yang et al. [105] proposed the simple linear regression direct standardization (SLRDS) method, and this univariate correction method is more suitable for the case where the spectrum has a small linear difference. Norgaard [106] also used this method in fluorescence spectroscopy, which is called the single wavelength standardization (SWS). Galvao et al. [107] calculated the covariance matrix based on the univariate spectral correction using the spectral residuals transferred from the master and slave of the standard sample set and then established the model through a robust

regression method. This method has certain advantages in the case of fewer wavelength variables. Lu et al. [108] used the minimum angle regression to first select the characteristic wavelength variable and then used the unary linear direct correction method to transfer the spectrum, which further improved the transfer effect. Wang et al. [109] used the dynamic time warping algorithm (DTW) in speech analysis to correct the spectral wavelengths on the two instruments, and then used the unary regression or multiple regression algorithm to correct the absorbance and obtain satisfactory results. In order to prevent excessive warping, Zou et al. [110] proposed a variable penalty dynamic time warping (VPdtw), which has a better transfer effect on the NIRS than the PDS algorithm.

Yan and Zhang [111] gave different weights to the wavelength variables in the moving window of the PDS algorithm and proposed a windowed PDS algorithm (WPDS) based on ridge regression and penalty terms. The PDS, a special case of the WPDS algorithm, can be regarded as all wavelength variables assigned the same weight. The double window PDS algorithm (DWPDS) is an extension of the traditional PDS algorithm, that is, a window of a certain width is taken from the master and slave spectra, and the spectrum transfer matrix is established window by window. Oliveri et al. [112] used the idea of the dual-window PDS algorithm to calculate the transfer coefficient between the average spectrum of the standard sample spectral matrix transferred from the master and the slave, and the transfer coefficient was calculated using the least square method (Fig. 17.4). Greensill et al. [70, 113] also used the dual-window PDS algorithm to transfer different array types of near-infrared spectra, but the best result is the wavelet transform combined with the DS algorithm. Ottaway and Kalivas [114] improved the DS and PDS algorithms by adding higher-order terms and derivative terms, which can solve the problem of nonlinear differences between different spectroscopic instruments to a certain extent.

On the basis of the DS algorithm, Chen et al. [115] used an extreme learning machine auto-encoder to establish the relationship between the master and slave standard sample spectral matrices and obtained stable spectral transfer results through an integrated strategy. Laref et al. [116] used SVM for the DS algorithm to obtain the signal transfer of different electronic nose instruments, and the standard set samples were obtained by the SPXY method. The stacked partial least squares (VIP-SPLS) method improved by the variable importance in the projection proposed by Li et al. [117] rearranges the wavelengths and divides them into a series of non-overlapping spectral intervals, and then transfers them through the DS algorithm.

Tan et al. [118] used wavelet transform spectroscopy to denoise and compress the signal, then reconstructed spectral signals of different scales through inverse wavelet transform, and then carried out transfer operations on the spectral signals of different scales, and proposed a wavelet hybrid direct standardization (WHDS). Chen et al. [119] also proposed a similar calibration transfer method. Yoon et al. [120] performed wavelet transform on the spectrum to obtain wavelet coefficients first and then used the DS algorithm to transfer. For large spectral matrices, this method can reduce the time for spectral transfer and modeling, which is called wavelet transform direct standardization (WTDS). Tan et al. [121] used wavelet packets to decompose the spectrum and realized the transfer of spectrum between different instruments

for $i = 1$ to N

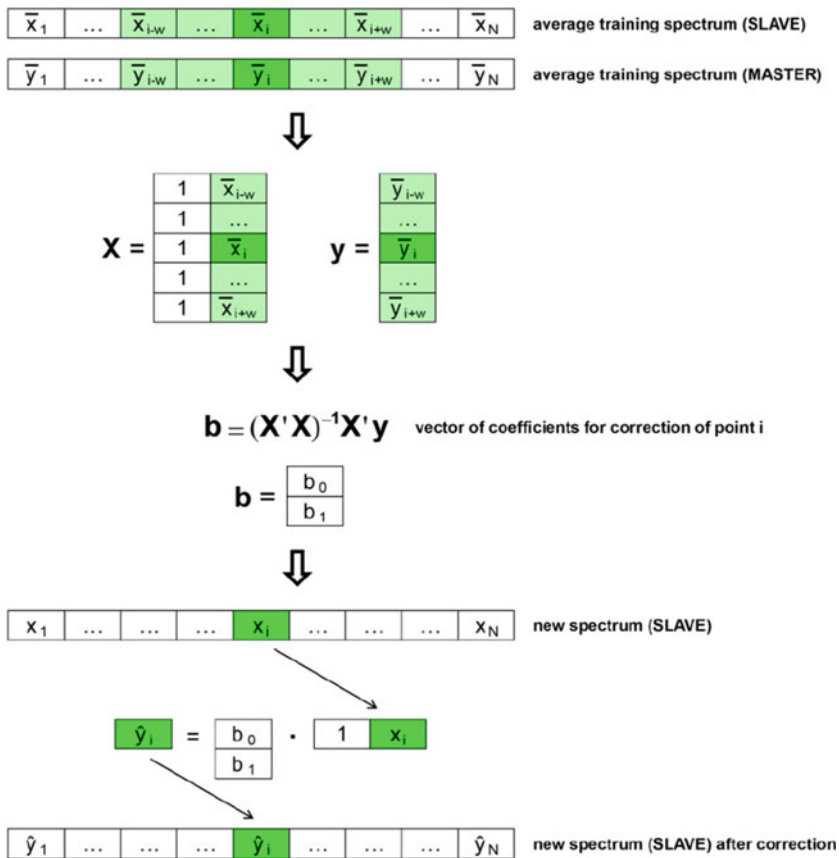


Fig. 17.4 Spectral transfer method based on the concept of double-window PDS

through wavelet packet coefficient transformation, which is called wavelet packet transform standardization (WPTS).

After Ni et al. [122] divided the spectrum by wavelet, used the PDS algorithm to transfer each sub-spectrum, and then used the consensus modeling strategy to establish a PLS model for each transferred sub-spectrum one by one, which is called stacked dual-domain piecewise direct standardization (SDDPDS), Poerio et al. [123] combined the dual-domain wavelet transform with orthogonal projection, and proposed the dual-domain transfer using orthogonal projection (DDTOP).

Based on the direct orthogonal signal correction (DOSC) algorithm, Lin et al. [124] and Wasng et al. [125] used the regression of the virtual standard average spectrum to eliminate the background differences between sample batches and proposed the orthogonal space regression (OSR) calibration transfer method, which can correct the systematic errors between the spectra of multiple batches of preparations, and

realize the calibration transfer of the chlorogenic acid quantitative model between batches during the water extraction process of the honeysuckle pilot-scale test. On this basis, Yang et al. [126] and Wang et al. [127] proposed a guided orthogonal projection technology combined with a calibration transfer method of SBC, which realized the transfer of the near-infrared quantitative moisture model of the small-scale test to the pilot-scale test. Wang et al. [128] used the random forest to select the wavelength of the NIRS and then used the DOSC algorithm to preprocess the spectrum to realize the sharing of calibration models between different instruments.

For the spectra collected at different temperature points, the transfer can be realized through the PDS algorithm. However, the traditional PDS algorithm cannot transfer the spectrum at any temperature between different temperature points [129]. Based on the PDS algorithm, Wulfert et al. [130] and Barring et al. [131] proposed a continuous piecewise direct standardization (CPDS) algorithm. The transfer matrix $F_{\Delta T}$ between two different temperatures and the temperature difference ΔT were subjected to polynomial regression to obtain two transfer matrix F_T at any temperature between two temperature points. In order to eliminate the influence of temperature on the online determination of electroplating bath composition content by AC voltammetry analyzer, Jaworski et al. [132] proposed continuous direct standardization (CDS) and used temperature as a variable to participate in the calibration model to eliminate the influence of temperature on the prediction results.

17.4 New Algorithms

17.4.1 Canonical Correlation Analysis

Canonical correlation analysis (CCA) is a multivariate analytical method to study the correlation between two sets of variables, which can reveal the linear dependence of them. CCA algorithm considers that the information of the measured object between two sets of spectra from the master and slave is consistent, and should be linearly related to each other, while the noise and interference information is random and independent. This method performs canonical correlation analysis on the master and slave spectral matrix and then converts the obtained canonical correlation variables. Transfer of canonical correlation variables can extract spectral transfer information from the overall spectra and can filter out noise and interference. The specific calculation is as follows [133]:

- (1) Execute CCA on the standard sample spectral matrix of the master and the slave, respectively.

$$L_m = X_m W_m \quad (17.2)$$

$$L_s = X_s W_s \quad (17.3)$$

L_m and L_s are the score matrix of the master standard sample spectral matrix X_m and the slave standard sample spectral array X_s , respectively, W_m and W_s are the loading matrix of X_m and X_s , respectively.

(2) Calculate the transfer matrix F .

$$F_1 = L_s^+ L_m \quad (17.4)$$

$$F_2 = L_m^+ X_m \quad (17.5)$$

$$F = W_s F_1 F_2 \quad (17.6)$$

(3) Transfer of the slave spectrum x_{in} to the master spectrum can be expressed as $x_{un}^P = x_{un} F$.

CCA algorithm only considers extraction of the maximum correlation of typical variables, which may introduce redundant information that has nothing to do with the target, thereby complicating the calibration transfer function. On this basis, Zheng et al. [135] proposed to use PLS to extract the factors that are related to the target value with the largest variance and then used CCA to perform the spectral transfer, which improves the pertinence of spectral transfer to a certain extent.

Before CCA transfer, Bin et al. [136] used wavelet transform to preprocess the original spectra and performed CCA by wavelet coefficients (WTCCA), which achieved a better transfer effect. Fan et al. [137] calculated the principal components of the master's standard sample spectral matrix X_m by the latent variables of the master PLS model, and then performed spectral transfer (PC-CCA). The validation results were better than the single CCA method.

Similar to CCA, based on spectral regression (SR), Peng et al. [138] decomposed on X_m and X_s , and proposed the spectral regression transfer algorithm. It puts the issue of solving the characteristic function in the regression model, avoids the frequent eigenvalue decomposition process of the dense matrix, and improves the calculation efficiency. Zhang et al. [139] decomposed the spectra based on multi-level simultaneous component analysis (MSCA) and proposed a two-level strategy for spectral transfer algorithm.

17.4.2 Spectral Space Transformation

Spectral space transformation (SST) obtained the combined spectrum matrix $X_{comb} = [X_m, X_s]$ by combining the standard set spectra X_m and X_s measured by the master and slave, respectively. The loading vectors of the combined spectrum matrix are obtained using PCA, and then the spectrum transfer matrix is calculated [140]. The calculation is as follows:

- (1) Combine the standard set spectra \mathbf{X}_m and \mathbf{X}_s measured by the master and slave to obtain a combined spectrum matrix.

$$\mathbf{X}_{comb} = [\mathbf{X}_m, \mathbf{X}_s] \quad (17.7)$$

- (2) PCA on the combined spectral matrix \mathbf{X}_{comb} .

$$\mathbf{X}_{comb} = \mathbf{T}[\mathbf{P}_m^t, \mathbf{P}_s^t] + \mathbf{E} \quad (17.8)$$

where \mathbf{P}_m^t and \mathbf{P}_s^t are loading of the master and slave matrices.

- (3) Calculate the transfer matrix \mathbf{F} .

$$\mathbf{F} = \mathbf{I} + (\mathbf{P}_s^t)^+ (\mathbf{P}_m^t - \mathbf{P}_s^t) \quad (17.9)$$

where \mathbf{I} is the identity matrix.

- (4) Transfer of the slave spectrum \mathbf{x}_{un} to the master spectrum can be expressed as

$$\mathbf{x}_{un}^p = \mathbf{x}_{un} \mathbf{F} \quad (17.10)$$

The structure of the SST algorithm is relatively simple to be calibrated, and it can still maintain good prediction results under a low standard sample number. Similarly, Liu et al. [141] used ICA to decompose the combined spectral matrix obtained from multiple instruments, and the expression of the transfer matrix is consistent with the SST algorithm.

17.4.3 Alternating Trilinear Decomposition

Alternating trilinear decomposition (ATLD) is an algorithm commonly used to decompose three-dimensional data. For a group of standard samples with collected spectra from different instruments, a three-dimensional matrix $\underline{\mathbf{X}}$ can be obtained, the dimension of which is $I \times J \times K$, where I is the number of standard samples, J is the number of spectral points, and K is the number of instruments. ATLD algorithm can decompose $\underline{\mathbf{X}}$ into three matrices as $\mathbf{A}(I \times N)$, $\mathbf{B}(J \times N)$, and $\mathbf{C}(K \times N)$, where N is the number of contributing factors, \mathbf{A} represents the relative concentration matrix of the standard sample, \mathbf{B} represents the relative spectral intensity matrix of the standard sample, and \mathbf{C} represents the instrument information matrix. The algorithm is as follows [142]:

- (1) Decompose $\underline{\mathbf{X}}$ by ATLD.

$$x_{ijk} = \sum_{n=1}^N a_{in} b_{jn} c_{kn} + e_{ijk}, \quad (i = 1, 2, \dots, I, \quad j = 1, 2, \dots, J, \quad k = 1, 2, \dots, K) \quad (17.11)$$

- (2) Calculate the transfer matrix F .

$$F_k = \text{diag}(c_k) \mathbf{B}^t \quad (17.12)$$

where c_k is the k th row of matrix C .

$$\mathbf{x}_{k2,\text{trans}} = \mathbf{x}_{k1,\text{new}} F \quad (17.13)$$

where $F = I + F_{kl}^+ (F_{k2} - F_{kl})$.

ATLD and SST algorithms have the same spectral transfer matrix formula, except that before the factor analysis, the SST algorithm expresses the standard sample spectral matrix collected from multiple instruments in an unfolded manner, while the ATLD algorithm is a cube matrix. As for cube data decomposition, besides ATLD, there are also PARAFAC and Tucker3 that can also be used. Kompany-Zareh et al. [143] used the Tucker3 algorithm to conduct the spectral transfer between different instruments as the missing data filling problem of the tensor array and realized the calibration transfer from the FT-Raman instrument to the CCD-Raman spectrometer.

17.4.4 Multi-task Learning

Multi-task learning (MTL) calculating the transfer matrix can be summarized as solving a convex optimization problem with a regular trace norm. This regular optimization method can extend the linear transfer between the spectra of different instruments to the nonlinear spectral transfer relationship. Compared with other transfer methods like neural networks, this method solves the final transfer matrix through a convex optimization problem. So it can efficiently and quickly obtain the global optimal solution while requiring fewer preset parameters. The calculation is as follows [144]:

- (1) Calculate the Gram matrix of the spectral matrix X_s of the slave standard samples.

$$K = X_s X_s^t \quad (17.14)$$

- (2) Perform eigenvalue decomposition on $n \times n$ order K matrix: $K = UDU^t$, where D is a diagonal matrix containing all eigenvalues, and each column of U is a corresponding eigenvector.
- (3) Solve a regular problem about the trace norm of matrix B of order $n \times p$ (n is the number of standard samples, and p is the number of wavelengths).

$$\min_B \left\| X_m - U D^{\frac{1}{2}} B \right\|_F^2 + \rho \|B\|_{\text{tr}} \quad (17.15)$$

where $\|\cdot\|_F$ represents the Frobenius norm of the matrix, $\|\cdot\|_{tr}$ represents the trace norm of the matrix, and ρ is the regular term coefficient. Method of the accelerated proximal gradient can be used to solve the convex optimization problem of trace norm regularization.

- (4) Calculate the transfer matrix F .

$$F = X_s^t U D^{-\frac{1}{2}} B \quad (17.16)$$

- (5) Transfer of the slave spectrum x_{un} to the master spectrum can be expressed as

$$x_{un}^P = F^t x_{un} \quad (17.17)$$

Boucher et al. [145] also proposed the proximal methods for spectral transfer between different instruments based on the regularization framework, which used the alternating direction method of multipliers (ADMM) to solve convex optimization problems. It has a good result on the transfer from a narrowband range spectrum to a wideband range spectrum.

As for the transfer of multiple qualitative models, Hu et al. [146] proposed an optimization framework using the maximum margin criterion (MMC).

$$\operatorname{argmin} \|X_m - X_s F\|_F^2 + \rho \|F^t (S_w - S_b) F\|_{tr} \quad (17.18)$$

where S_w is the intra-class scatter matrix, S_b is the inter-class scatter matrix, and F is the spectral transfer matrix. MMC algorithm has certain advantages in transferring spectra for qualitative analysis.

17.4.5 Generalized Least Squares

Generalized least squares (GLS) takes the difference matrix of the master and slave standard samples as a reference to establish a weighted filtering model to eliminate the influence of the difference between the instruments on the spectra. The main steps of the algorithm are as follows [147, 148]:

- (1) Calculate the mean-centered difference matrix of the master and slave standard sample spectra.

$$X_d = (X_m - \bar{X}_m) - (X_s - \bar{X}_s) \quad (17.19)$$

- (2) Calculate the covariance matrix of X_d .

$$C_d = \frac{X_d^t X_d}{n-1} + \alpha I \quad (17.20)$$

where n is the number of standard samples, α is the coefficient (usually $1 \times e^{-6}$), and I is the identity matrix.

- (3) Perform singular value decomposition on the covariance matrix C_d .

$$C_d = USV^t \quad (17.21)$$

- (4) Calculate the weighted filter matrix W .

$$W = VS_{adj}^+ V^t \quad (17.22)$$

where $S_{adj} = \text{sqrt}\left(\frac{S \times m}{\text{trace}(S)}\right)$, and m is the number of wavelength variables.

- (5) Transform the spectral matrix of the master calibration set and that of the slave prediction set separately.

$$X_{\text{trans}} = (X - \bar{X})W \quad (17.23)$$

17.4.6 Other Algorithms

Based on the principle of orthogonal projection, Andrew and Fearn [149] obtained the average spectrum of a set of standard samples on multiple instruments and performed PCA on the average spectra. The first few representative projection spaces formed by the different spectral loadings between the instruments can perform transfer by orthogonal projection (TOP) on the spectra on multiple instruments. From the concept of TOP, Zhu et al. [150] replaced the average spectral matrix with a change matrix of a set of sample repetitive spectra, performed PCA to obtain the projection space, and proposed an error removal by orthogonal subtraction (EROS). Subsequently, in view of EROS, Zeaiter et al. [151] calculated the virtual spectrum of the slave through the kernel function. According to the difference matrix between the measured and virtual spectrum, the PCA was performed to obtain the projection space, and the dynamic orthogonal projection (DOP) was proposed. Dabros et al. [152] used the DOP algorithm for the maintenance of the online infrared calibration model and achieved good application results. Igne et al. [153] reviewed and compared the effects of the above-mentioned orthogonal projection algorithms for the transfer of NIRS models. Siska and Hurburgh [154] used the wiener filter method to process the spectra for the transfer of fixed filter-type NIRS instruments and proposed a method of optimizing the master instrument.

Chen et al. [155] proposed a loading space standardization (LSS) algorithm for the influence of temperature on the spectrum. PCA was performed on the spectra collected from the standard set samples at different temperatures, and then the quadratic function relationship between the spectral loading and the temperature at each temperature was established. Spectrum measured at a certain temperature can be standardized in the loading space to obtain the corresponding spectrum at

standard temperature. Afterward, Chen and Morris [156] combined the LSS algorithm with the optical path length estimation and correction (OPLEC) and proposed the extended loading space standardization (ELSS), which was used to correct the effects of temperature and composition changes on the spectra. Similarly, Shi et al. [157] effectively eliminated the influence of temperature changes on the NIRS during the production of sugar and flavoring based on the LSS algorithm.

Zhang et al. [158] regarded calibration transfer as a global affine transformation problem:

$$\hat{\mathbf{x}}_{i,n} = a_i x_{i,n} + b_i, \quad i = 1, \dots, k, \quad n = 1, \dots, N$$

where $\mathbf{x}_{i,n}$ represents the spectrum of the n th standard sample on the i th slave instrument, k represents the number of slave instruments, N represents the number of standard samples, and $\hat{\mathbf{x}}_{i,n}$ represents the spectrum of the n th standard sample on the i th slave instrument transferred to the master. The transfer coefficients \mathbf{a}_i and \mathbf{b}_i are calculated by the robust weighted least square algorithm (RWLSA). Deshmukh et al. [159] also used a robust regression method to achieve inter-station signal transfer among the electronic nose systems for emission detection in paper mills.

Zhao et al. [160] proposed an algorithm of calibration transfer based on affine invariance (CTAI) for calibration transfer based on affine invariance. This method first establishes the PLS model to obtain the score matrix and prediction vector of the master spectra, as well as the pseudo-score matrix and pseudo-prediction vector of the slave spectra. Secondly, the regression coefficients of master and slave are obtained by least squares, respectively, and then the angle and deviation between master and slave are calculated by the regression coefficients. Finally, the prediction result of the slave spectra is obtained based on the affine transformation.

Folch-Fortuny et al. [76] and Munoz et al. [161] proposed a new spectral transfer method using Joint-Y partial least squares regression (JYPLS).

$$\mathbf{Y}_J = \begin{bmatrix} \mathbf{Y}_m \\ \mathbf{Y}_s \end{bmatrix} = \begin{bmatrix} \mathbf{T}_m \\ \mathbf{T}_s \end{bmatrix} \mathbf{Q}_J^t + \mathbf{E}$$

$$\mathbf{X}_m = \mathbf{T}_m \mathbf{P}_m^t + \mathbf{E}$$

$$\mathbf{X}_s = \mathbf{T}_s \mathbf{P}_s^t + \mathbf{E}$$

$$\mathbf{T}_m = \mathbf{X}_m \mathbf{W}_m$$

$$\mathbf{T}_s = \mathbf{X}_s \mathbf{W}_s$$

where \mathbf{Y}_m and \mathbf{Y}_s are the concentration matrices of the calibration set and standard set samples of the master; \mathbf{T}_m and \mathbf{T}_s are score matrices of \mathbf{X}_m and \mathbf{X}_s ; \mathbf{P}_m and \mathbf{P}_s

are loading matrices of X_m and X_s ; W_m and W_s are weighting matrices of X_m and X_s , respectively. Q_J is the loading matrix of the combined Y concentration matrix.

Spectrum $x_{i,m}$ in the master calibration set can be transferred into the spectrum $\hat{x}_{i,s}$ of the slave by the following formula:

$$\hat{x}_{i,s} = (Q_J Q_J^t)^{-1} Q_J x_{i,m} \quad (17.24)$$

Shan et al. [162] proposed a joint spectral subspace transfer method (JPCA) based on principal component analysis and kernel principal component analysis. The method combines the spectral matrices of the master and slave standard sample sets for PCA or kernel PCA so as to obtain the transfer matrix through least squares in the low-dimensional feature space.

Khaydukova et al. [163] proposed a standardization method with regularization coefficients (SRC) based on Tikhonov regularization (TR), which uses the following formula to obtain the transfer matrix:

$$F = (X_s^t X_s + a)^{-1} (X_s X_m) \quad (17.25)$$

In the formula, a is the regularization coefficient, usually from 1 to 30,000.

Spectrum $x_{i,m}$ in the master calibration set can be transferred into the spectrum $\hat{x}_{i,s}$ of the slave by the following formula:

$$\hat{x}_{i,m} = x_{i,s} F \quad (17.26)$$

Zhao et al. [164] used the PLS model of the master to project the spectra of master and slave standard sample sets, respectively, and then established the transfer between two instruments in the PLS projection space. For the spectrum collected from the slave, the projection vector on the master was obtained by transfer relationship in the projection space, and then the PLS model of the master produced the final prediction result. Zhang et al. [165] used X weight matrix of the PLS model built by the master to project the master standard sample spectrum to further obtain the matrix L , and then to obtain the transfer matrix F between L and the slave standard sample spectral matrix. It was a PLS-based weight matrix transfer method. Chen et al. [166] used the master model to predict the deviation of the concentration value of the slave standard sample set from the actual concentration and established a deviation prediction model for systematic prediction error correction (SPEC).

Mou et al. [167] aimed the minimization of the robust Cauchy estimator function of subspace learning and proposed a robust spectral transfer method. It calculated the shared base matrix of the master and slave spectra and its corresponding expression coefficients in an iterative manner and then established a transfer matrix based on the expression coefficients, which can reduce the impact of outlier samples and spectral noise on the spectral transfer. Seichte et al. [168] proposed a Bayesian calibration transfer method based on the Lagrangian multiplier method and hierarchical model, which used the Markov chain Monte Carlo method to estimate the error

bounds. It was successfully performed in calibration transfer of oxygen sensors. Subsequently, they used similar methods to eliminate the influence of oxygen on the determination of carbon dioxide content by mid-infrared spectroscopy [169]. Based on the multiple block orthogonal projections to latent structures (OnPLS), Skotare et al. [170] proposed a joint and unique multiblock analysis (JUMBA) for spectral transfer of multiple instruments.

Andries used the penalty matrix decomposition algorithm of domain adaptation in the transfer learning method such as transfer component analysis (TCA) and scatter component analysis (SCA) for spectral matrix transfer and model maintenance [171]. Similarly, Liu et al. [172] used the TCA to realize the NIRS transfer of edible oil on different instruments. In terms of soil research, TCA was successfully applied to transferring the NIRS among arsenic and available phosphorus models [173, 174]. Especially, TCA addresses the problem of different data distribution in the source domain and the target domain and maps the data from two domains to a high-dimensional regenerative kernel Hilbert space, where the data distance between the source and target are minimized, while their respective internal attributes to the greatest extent are retained.

Shi et al. [175] transferred the NIRS of the two types of wood by adaptive component analysis (ACA) and established a deep transfer learning model of oak wood defect classification with the color wood data as the source domain and oak wood data as the target domain. Shan et al. [176] proposed a joint spectral subspace method for calibration transfer based on PCA and KPCA. Based on the principle of domain adaptation, Nikzad-Langerodi et al. [177], Mishra and Nikzad-Langerodi [178] proposed a domain-invariant PLS (di-PLS) method, which can be used for unsupervised, semi-supervised, and supervised spectral model maintenance and calibration transfer. Huang et al. [179] also proposed a partial least squares method for domain adaptation. A transfer sample-based coupled task learning (TCTL) method was proposed based on transfer learning and multi-task learning, which could be used for transfer between electronic nose instruments and compensation for drift over time [180]. Based on the active learning algorithm from machine learning, Hu et al. [181] solved the problem of multivariate quantitative correction for hyperspectral imaging of different types of blueberries through iterative screening of standard samples.

Li et al. [182] proposed a double digital projection slit algorithm for Raman spectrometers with different resolutions to solve the problem of spectral consistency. The gradient descent method was adopted to obtain the optimal solution of the transfer matrix, and a better transfer result was achieved. In the study of Liu et al., the deep autoencoder (DAE) method was employed to establish a nonlinear mapping between spectra of different NIR instruments. An error function penalty term based on conditional probability and parameter maximum likelihood method was designed, and the network parameters of deep auto-encoding were optimized by combining with a gradient back propagation algorithm [183].

In addition to the transfer of the calibration spectra matrix, some new methods have also been applied in order to realize the transfer of the master calibration model. Liu et al. [184] proposed a linear model correction (LMC) that can realize the transfer of PLS model regression coefficients. Subsequently, the method was further improved

to obtain the globally optimized regression coefficients [185]. Kauppinen et al. [186] realized the transfer of the online NIR model of the moisture content of the drug freeze-drying process by converting the PLS regression coefficients. With regard to the evaluation issue of the transferred model, Eskildsen et al. [187] proposed to use the prediction results of the model instead of original reference data to evaluate the effect of the calibration transfer.

17.5 Global Calibration, Robust Calibration, and Model Update

Three keywords of global calibration, robust calibration, and model update (or model maintenance) are often referred to confusingly in spectral multivariate calibration analysis, and actually they have a lot in common. The global model calibration, also known as the augmented hybrid calibration, hybrid calibration, spiking method in the literature, usually refers to the expansion of the calibration set of the master by adding spectra under different test conditions, spectra measured from different instruments, etc. to build a model so as to achieve model sharing under different instruments, different measurement environments, or different sample types. Robust calibration often refers to the establishment of a model that is not sensitive to external influence factors through spectral preprocessing algorithms, wavelength variable screening algorithms, etc. [188, 189]. Therefore, it can be considered that the establishment of a global calibration is a means to achieve a robust model, and both two methods can also be combined to establish a robust global model. However, the model updating or calibration maintenance covers a broader range. In the traditional concept, when encountering samples outside of the model (chemical composition or physical state of the sample changes) or the instrument is aging over time, what we need is model update or model maintenance. In general concept, the process of establishing a global model or a robust calibration actually also belongs to the category of model update or model maintenance [190, 191]. Thus, global calibration, robust calibration, and model updates (or model maintenance) are usually employed to solve calibration transfer problems [192].

Koehler et al. [193, 194] introduced a self-designed FIR matrix to filter the MIR interference data, combined with the model update, accurately classifying the data obtained by two MIR remote sensing spectrometers based on piecewise linear discriminant analysis (PLDA). The FIR algorithm was used by Song et al. to eliminate the changes in the spectra from the same instrument at different times and under different environmental conditions, and a more robust NIRS model for predicting soil organic matter content was established [195, 196]. By means of spectral error analysis, Wang et al. [197] proved that the combination of the first derivative and SNV can significantly improve the calibration transfer results between FT-NIR spectrometers with integrating sphere diffuse reflectance measurement.

Milanez et al. [198] involved successive projections algorithm (SPA) to select characteristic wavelength variables for establishing NIRS and UV-vis models. The discriminant models can be used among different instruments, and their recognition accuracy was equivalent to DS and PDS methods. A similar study was reported that SPA and competitive adaptive reweighted sampling (CARS) were combined to select special wavelengths, and together with the SBC algorithm, the established NIRS model predicted the soluble solid content of apple spiked for consecutive years [197]. A double CARS strategy was proposed by Zheng et al. [200] for NIRS global modeling. Based on the concept of standard deviation, Ni et al. [201–203] proposed a stable and consistent wavelength variable selection method between spectrometers, which can establish a robust calibration model and realize the sharing of models on multiple instruments. Hong et al. [204] used the scale-invariant features transform (SIFT) algorithm to screen the stable characteristic wavelengths for establishing the NIRS model of the total alkaloids of tobacco leaves, which can realize the standard-free transfer of the model. Xu et al. [205] proposed a correlation-analysis-based wavelength selection (CAWS) based on the correlation coefficient between the master and slave spectra, which was used to establish a robust calibration model and obtained good results. Based on the multi-model consensus strategy, Zhang et al. [206, 207] proposed the guided model reoptimization (GMR) method to solve the problem of model update and calibration transfer, which selected wavelength variables to filter and weight by PLS regression coefficients, and screened calibration set samples through a method similar to stepwise multiple linear regression to select variables.

When establishing the global model between master and slave, it is necessary to collect the spectra of a certain number of representative samples from the slave, and the corresponding concentration value is granted to get better results [208]. To address this problem, Kalivas et al. [209, 210] proposed a strategy of models update and transfer based on the weighting method of slave samples and the Tikhonov regularization framework.

The common weighting modeling can be expressed as

$$\begin{pmatrix} \mathbf{y} \\ \lambda \mathbf{y}_L \end{pmatrix} = \begin{pmatrix} \mathbf{X} \\ \lambda \mathbf{L} \end{pmatrix} \mathbf{b} \quad (17.27)$$

where \mathbf{y} is the concentration vector of the master calibration set, \mathbf{X} is the spectra matrix of the master calibration set, \mathbf{y}_L is the concentration vector of the slave calibration, \mathbf{L} is the spectra matrix of the slave calibration, λ is the weighted value, and \mathbf{b} is the regression coefficient.

Under Tikhonov regularization (L2 regularization), the optimization framework is

$$\min \left(\|\mathbf{X}\mathbf{b} - \mathbf{y}\|_2^2 + \lambda^2 \|\mathbf{L}\mathbf{b} - \mathbf{y}_L\|_2^2 \right) \quad (17.28)$$

The solution of this equation is

$$\hat{\mathbf{b}} = (\mathbf{X}^t \mathbf{X} + \lambda^2 \mathbf{L}^t \mathbf{L})^{-1} (\mathbf{X}^t \mathbf{y} + \lambda^2 \mathbf{L}^t \mathbf{y}_L) \quad (17.29)$$

If \mathbf{L} represents the difference spectra matrix of a set of standard sample sets measured on the master and slave, respectively, or a set of spectral baseline background matrix under different test conditions, or a set of spectra matrix of blank samples, etc., its corresponding \mathbf{y}_L becomes a zero concentration vector. Then the optimization framework can be simplified as

$$\min(\|\mathbf{X}\mathbf{b} - \mathbf{y}\|_2^2 + \lambda^2 \|\mathbf{L}\mathbf{b}\|_2^2) \quad (17.30)$$

The solution of this equation is

$$\hat{\mathbf{b}} = (\mathbf{X}^t \mathbf{X} + \lambda^2 \mathbf{L}^t \mathbf{L})^{-1} \mathbf{X}^t \mathbf{y} \quad (17.31)$$

In order to obtain more stable regression coefficients, based on the idea of ridge regression, the weighting method modeling can be improved as

$$\begin{pmatrix} \mathbf{y} \\ \mathbf{0} \\ \lambda \mathbf{y}_L \end{pmatrix} = \begin{pmatrix} \mathbf{X} \\ \tau \mathbf{I} \\ \lambda \mathbf{L} \end{pmatrix} \mathbf{b} \quad (17.32)$$

Among which, τ is the penalty coefficient and \mathbf{I} is the identity matrix. The optimization framework is

$$\min(\|\mathbf{X}\mathbf{b} - \mathbf{y}\|_2^2 + \tau \|\mathbf{b}\|_2^2 + \lambda^2 \|\mathbf{L}\mathbf{b} - \mathbf{y}_L\|_2^2) \quad (17.33)$$

The solution of this equation is

$$\hat{\mathbf{b}} = (\mathbf{X}^t \mathbf{X} + \tau^2 \mathbf{I} + \lambda^2 \mathbf{L}^t \mathbf{L})^{-1} (\mathbf{X}^t \mathbf{y} + \lambda^2 \mathbf{L}^t \mathbf{y}_L) \quad (17.34)$$

Regarding the model update and transfer strategy of Tikhonov's regularization, Kunz et al. [211, 212] discussed the influence of standard design selection. Shah-bazikhah and Kalivas [213] proposed a consensus modeling strategy to optimize the selection of regularization parameters. While Tencate et al. [214] brought forward a method for selecting model update parameters based on the fusion strategy. Farrell et al. [215] used Tikhonov's regularization strategy to update the drug NIRS model under different conditions and obtained satisfactory results. A similar study was done for sharing the model on different NIRS instruments based on Tikhonov's regularization strategy [216].

Except for sample augmentation, the model update can also take samples and features to augment at the same time, which can be expressed as [217]

$$\begin{pmatrix} \mathbf{y} \\ \mathbf{0} \\ \lambda \mathbf{y}_L \end{pmatrix} = \begin{pmatrix} \mathbf{X} & \mathbf{0} \\ \mathbf{0} & \mathbf{X} \\ \lambda \mathbf{L} & \lambda \mathbf{L} \end{pmatrix} \begin{pmatrix} \mathbf{b}_m \\ \mathbf{b}_s \end{pmatrix} \quad (17.35)$$

Rudnitskaya et al. [218] compared the results of DS algorithm, Tikhonov regularization, and Joint-Y PLS on calibration transfer and update for the potential sensor array instrument, proving that Tikhonov regularization and Joint-Y PLS can get better results. On the basis of the least absolute shrinkage and selection operator (LASSO), Kunz and She [219] proposed a robust fused LASSO algorithm (RFL) for the maintenance and transfer of the calibration model. The genetic algorithm was performed by Guo et al. [220] to correct the wavelength variables from multiple Raman spectra, and the updated model by Tikhonov regularization obtained satisfactory results. Based on the ridge regression updating, Zhang et al. [221] combined the prediction optimization and the 2-norm constraint of model coefficients, realized the update of model coefficients, and solved the problem of deterioration of model prediction ability and reliability caused by instrument drift or sample changes.

Based on Lambert-Beer's law, Sulub and Small [222] proposed a spectral simulation calculation method. Spectra of the mixture can be generated by the absorption signal of the pure substance and the background signal of the instrument. In this way, whether for master or slave instrument, the calibration set can be quickly established by measuring the spectra of the pure substance in the mixture. Haaland [223] synthesized spectra by adding the temperature-affected background signal to the calibration spectra so as to solve the problem of the model's adaptability to temperature. The temperature-affected background signal was obtained by the least square expansion of the temperature spectra. By PDS, Sulub et al. [224] realized the spectral transfer of multiple NIRS instruments (including grating scan and FT instruments) to determine the content of active ingredients in medicines and updated the slave model by preparing a placebo. This method can analyze 30 medicines in 12 min, instead of 5 h by the HPLC method. Saiz-Abajo et al. [225] established a robust calibration model by adding different types of noise and interference to the calibration spectra, as well as the ensemble modeling strategy. Pierna et al. [226] added the spectral changes of different instruments and samples with different states (water content and particle size, etc.) to the spectra of 10 calibration samples, generating hundreds of virtual modeling spectra, and obtaining a robust prediction model.

Cooper et al. [227] took the spectra of several calibration samples from the master as the target and obtained the same number of virtual spectra through mathematical mixing calculation to form a virtual spectral matrix. The master model predicted the virtual spectral matrix of master and slave and established the SBC calibration curve based on the predicted value of master and slave. In the further study, they also used the spectra of 13 pure compounds to transfer the NIRS of jet fuel by the above method, which can accurately predict the properties and composition of aromatics, hydrogen content, density, viscosity, etc. [228]. Rauscher et al. [229] also referred to this method to transfer the spectra of the non-dispersive infrared spectrometer to monitor the quality of the oil. Similarly, Abdelkader et al. [230] used the spectra of

15 pure compounds and calculated the virtual spectra in segments, improving the transfer results of the method. Da Silva et al. [231] established the spectral transfer coefficient on the virtual spectral matrix using the DS algorithm and realized the transfer of oil from the desktop NIR spectrometer to the handheld instrument.

Ni et al. [232] combined the stacked PLS method with a model update to establish a robust calibration model for the sharing of multiple NIRS instruments. Honorato et al. [233] selected wavelength variables by the continuous projection algorithm (SPA) to establish a robust MLR model shared by multiple instruments, and the results were slightly better than PDS-PLS. With the establishment of a local calibration model, Igne and Hurburgh [234] realized the model sharing of four NIRS instruments from two brands. Liu et al. [235] first constructed the 3D spectral cube collected from multiple instruments and employed Tchebichef discrete orthogonal moments to extract chemical features. Finally, a stepwise regression method was conducted to establish a universal model on multiple instruments.

Moving window MSC was employed by Kramer et al. [236] for the preprocessing of jet fuel spectra on two NIRS instruments, and a better result was obtained by optimizing the size of the moving window. Also, Liu et al. [237] mainly used the MSC algorithm to standardize the line-scan NIR imaging system. Sahni et al. [238] compared the effects of MSC, OSC, FIR, PDS, and global models on the correction of optical path changes of optical fiber probes, and the results revealed that PDS and global models proved to be the better solutions. Extended multiplicative signal correction (EMSC) was used for the preprocessing of bacterial Raman spectra on multiple instruments. The PLS-DA algorithm can successfully establish a discriminant model for different bacterial species [239]. Preys et al. [240] combined orthogonal signal correction (OSC) with external parameter orthogonalization (EPO) to establish a robust calibration model, which solved the problem that OSC cannot consider the influence of external interference and the predictive performance degradation of EPO when the influence of external factors on the target value is too high.

Wijewardane et al. [241, 242] modeled the air-dried ground soil to predict the organic and inorganic carbon content at different moisture levels by EPO, DS, and a global model. Ackerson and Roudier et al. [243, 244] also solved similar problems through EPO and DS algorithms. Amat-Tosello et al. [245] used the EPO algorithm to simultaneously realize the sharing of gasoline short-wave NIRS on four instruments. EPO was employed by Hans and Allison [246] to reduce the influence of temperature and humidity on the NIRS to predict the calorific value of biomass, and to establish a model that was not sensitive to temperature and humidity. Similarly, the temperature mixing model strategy can eliminate the influence of temperature on NIRS prediction of reducing sugar and moisture content in longan honey. Thamasopinkul et al. [247], Thygesen and Lundqvist [248] successfully built a temperature mixing model for predicting moisture content in wood using NIRS, which was more effective than a single PDS algorithm.

Luoma et al. [249] attempted to use additive partial least squares modeling strategy for model maintenance. According to different measurement conditions, the residual PLS model was established to solve the problem of model inapplicability by addition. Elizalde et al. [250] also adopted a similar strategy to solve the problem of model

inapplicability caused by changes in the online Raman spectrometer. Nouri et al. [251] established a hybrid model of soil NIRS from lab and airborne hyperspectral images and updated the model with the difference spectra and zero concentration of the standard spectra. This model can be used for the prediction of airborne hyperspectral imaging.

A local model that is based on the local kernel function is often established in the SVR method, which may not work when the new spectrum from the new instrument is added to the calibration set. Based on the transfer learning idea, Yu and Ji [252] used regularized multi-task learning (RMTL) to estimate the relationship between the SVR model in the new condition and the original model and successfully improved SVM regression. A highlight of this method is that the most important support vectors of the SVR model can be selected as the sample for a model update.

17.6 Progress of Applications

17.6.1 SBC Method

Brito et al. [253] used the SBC algorithm to correct the model prediction values of the total suspended solids (TSS) and chemical oxygen demand (COD) of wastewater by UV spectroscopy, and the result was better than the SSC algorithm and the SLRDS algorithm. Brouckaert et al. [254] also used the SBC algorithm to correct the predicted values of the Raman spectroscopy established in the laboratory for the determination of the content of the two components of the liquid detergent, and used it for industrial-level online analysis, and obtained satisfactory results. Damberg et al. [255] used the SBC algorithm to quickly analyze the predicted value of the tannin content in red wine by ultraviolet spectroscopy and corrected it on multiple instruments, and obtained relatively consistent prediction results.

According to the PCA classification model of different types of coffee beans in the NIR spectroscopy, Myles et al. [256] used the SBC algorithm to transfer the PCA scores on the two instruments and obtained a better result. Wang et al. [257] conducted a similar study on the hyperspectral prediction models of lamb protein from different origins, and the result is that the SBC algorithm has a better effect. Li et al. [258] used the SBC algorithm to correct the acid value and peroxide value of edible oil predicted on two NIRS instruments.

17.6.2 SSC Method

Pierna et al. [259] used the spectral difference correction algorithm (SSC) to transfer the spectrum of more than 9000 feed samples on the dispersive desktop NIR spectrometer to the MEMS handheld spectrometer using 25 transfer standard samples

to establish the determination of feed fat, fiber, and protein, and starch PLS model. Zamora-Rojas et al. [260] used the SSC algorithm to transfer the pork spectrum collection on the desktop NIR spectrometer to the handheld instrument to realize the routine quality analysis beside the pork processing line. Daikos et al. [261] used a method similar to the SSC algorithm to subtract the background of the substrate, and transferred the near-infrared imaging spectrum of the coating on one substrate material to another substrate material, realizing the sharing of PLS models. Smith et al. [262, 263] used the SSC algorithm to realize the spectral transfer of different near-infrared spectroscopy instruments to predict the content of active ingredients in the whole paracetamol tablets. This article also tried to use six standard materials to establish a spectral response correction model through the SLRDS algorithm, which was slightly inferior to the SSC algorithm.

Hayes et al. [264] calibrated the wavelength of the array detector-type short-wave near-infrared spectrometer and then used the SSC algorithm to update the model. In predicting the soluble solid content of apples, the results are equivalent to PDS, but the implementation process is simple and easy. Xu et al. [265] aimed at the lack of universality in the multi-channel grading detection model for fruit quality, and adopted the DS algorithm (MSSC-DS) for correction of the average spectral difference to transfer the online detection spectrum of the crown pear sugar content between the two spectrometers, and the prediction accuracy of the model can meet the actual requirements of production ($<0.5^\circ$ Brix). Roggo et al. [266] used the SSC algorithm to transfer the near-infrared reflectance spectrum of sugar beet, and good results can be obtained no matter whether the actual sample or the general sample is used as the transfer standard sample. Saranwong and Kawano [267] used the SSC algorithm for the NIRS transfer of the online fruit screening machine, which greatly improved the MLR and PLS correction models, and the compensation of the average spectral difference of the standard sample set was better than the result of linear fitting or polynomial fitting.

Soldado et al. [268] combined the SSC algorithm with the transfer by orthogonal projection (TOP) method to transfer the silage spectrum from the desktop NIR spectrometer to a portable instrument. The built model can accurately predict the content of dry matter, neutral detergent fiber, and crude protein. Sun et al. [269] first screened the characteristic variables through regression coefficients, then compensated the absorbance through the SSC algorithm, and established a universal NIR analysis model for the quality of fresh jujubes between different instruments.

17.6.3 *Shenk's Method*

Qin and Gong [270] compared the transfer results of PDS, DWPDS, and Shenk's algorithms on the NIRS of tobacco leaf and smoke powder. Shenk's algorithm is superior to other methods. It is also a feasible way to establish a NIRS mixed model of tobacco leaf and powder. Based on Shenk's algorithm, Garcia-Olmo et al. [271] investigated the transfer results of four standard sample components on fatty acid

liquid near-infrared spectroscopy, and the results showed that the closer the composition of the sample to the test sample, the better the transfer result. De La Roza-Delgado et al. [272] used Shenk's algorithm to transfer the milk spectrum between desktop and portable near-infrared spectrometers and established a model that can quickly detect milk composition on site. Masahiro and Yukihiko [273] used a similar Shenk's algorithm to calibrate the wavelength and absorbance of the NIRS, respectively, to solve the problem of spectrum changes before and after the online instrument maintenance, and to achieve the correction and maintenance of the vinyl content prediction model in molten polypropylene. Perezmarin et al. [274] used Shenk's algorithm to successfully transfer the NIRS of feed on different scanning instruments.

17.6.4 DS Method

Milanez et al. [275] used the DS algorithm to transfer the NIRS of the adulterated ethanol-gasoline on two instruments, and a 100% recognition success rate can be obtained for PLS-DA discriminant analysis. Da Silva et al. [276] used the DS algorithm to transfer the FT-NIRS of drugs to multiple handheld spectrometers and established an analysis model for predicting the content of different crystal forms. Brito et al. [277] used the DS algorithm to transfer the NIRS of the flour from the desktop instrument to the handheld instrument, and then re-established the PLS model on the handheld instrument, which can obtain better results. Ji and Han [278] used the UVE-SPA wavelength screening method combined with the DS algorithm to realize the transfer of the NIRS of the apple on the same model and different models of fruit portable instruments. Hu and Xia [279] realized calibration transfer of navel orange total sugar prediction by near-infrared spectroscopy using the DS algorithm.

Chen et al. [280] used the DS algorithm to transfer the NIR hyperspectral spectra of soil under different humidity and realized that the model based on the air-dried soil spectrum could be used for soil samples with different moisture content. Wang et al. [281] used the DS algorithm to convert the NIRS of soil under different humidity to eliminate the interference of moisture on the prediction of soil organic matter content. Ji et al. [282] used the DS algorithm to transfer the NIRS of the water-bearing untreated soil and made predictions through a quantitative calibration model established by dry and ground soils. Silva et al. [283] used the DS algorithm to transfer the gasoline spectra measured by three mid-infrared spectrometers, and based on the global modeling strategy, the PLS-DA or SIMCA method can be used to distinguish gasoline from different origins. Liu et al. [284] used the DS algorithm to transfer two edible oil spectra above the NIRS and established an analysis model for predicting an acid value and peroxide value. Lopez-Moreno et al. [285] used the DS algorithm to transfer the LIBS spectrum at room temperature to the spectrum at high temperature and established a model for predicting the metal content in a high-temperature environment.

Khaydukova et al. [286] used the DS algorithm to transfer the volt-ampere signal and the potential signal in the electronic tongue sensor so that the regression model

based on the potential data can be used for the prediction of the volt-ampere signal. Weng et al. [287] used the DS algorithm to transfer citrus spectra between different models of hyperspectral imaging, and the extreme learning machine discriminant model of citrus canker has an accuracy of 86.2% on the slave spectra. Fonollosa et al. [288, 289] used DS, PDS, and other algorithms to transfer the signal of the metal oxide gas sensor matrix on multiple sensors and obtained satisfactory results. Panchuk et al. [290] used the DS algorithm to convert different types of spectra, such as the transfer between energy-dispersive XFS and vis-UV, and the transfer between different NIRS wavelength ranges. Vaughan et al. [291] used the DS combined with PLS algorithm to convert the spectra of two LC-MS, and initially solved the problem of metabolomics data fusion. De Moraes et al. [292] realized the transfer of digital imaging on two devices using the DS algorithm for predictive analysis of serum creatinine content. Khoshkam et al. [293, 294] embedded the DS algorithm in the multivariate resolution analysis of the extended matrix for the study of reaction kinetics and obtained good results. Surkova et al. [295] successfully transferred the spectrum from the UV-vis spectrometer to the optical multi-sensor system composed of four LEDs by using DS algorithm and realized the transfer between optical multi-sensor systems.

17.6.5 PDS Method

Barreiro et al. [296] used the PDS algorithm to transfer the spectrum from the desktop NIR spectrometer to the portable spectrometer and established an analytical model that can detect the olive breeding process in the field. Alamar et al. [297] used the PDS algorithm to transfer the FT-NIRS of 447 Jonagold apples to the array spectrometer, and before the spectrum transfer, the wavelength of the FT spectrum was normalized by the segmented cubic Hermite interpolation, and then the analytical model of the soluble solid content was established. Sulub et al. [298] used the PDS algorithm and the wavelet hybrid direct correction algorithm (WHDS) to achieve spectral transfer on five NIR spectrometers for rapid analysis of the uniformity of the active ingredient content of the tablet. Luo et al. [299, 300] combined wavelength selection and PDS algorithm for the transfer of bovine blood near-infrared spectroscopy, which can achieve a rapid diagnosis of bovine anemia on multiple instruments. Cen et al. [301] used PDS to transfer the hyperspectral spectrum of citrus canker and then established the least squares support vector machine discriminant model. The recognition rate of the model for the prediction set increased from 26% before transfer to 97%.

Pereira et al. [302] used the dual-window PDS algorithm (DWPDS) to transfer the NIRS of the drug powder to the spectrum of the tablet, which provides a feasible way to quickly obtain a sample of the drug spectrum calibration set. Sohn et al. [303] also used the DWPDS algorithm to transfer the NIRS for analyzing the cellulose content of flax between different instruments and achieved good results. Galvan et al. [304] used the DWPDS algorithm to transfer the low-field nuclear magnetic resonance spectrum of gasoline between different instruments, among which the resolution of low-field nuclear magnetic instruments is different.

Yang et al. [305] used DS and PDS algorithms to effectively transfer the surface-enhanced Raman prediction model of potassium sorbate in sweet-scented osmanthus wine to bayberry wine to realize the transfer of the same analyte prediction model among different species. Similar examples of transferring models between different sample types include the application of quality analysis of chilled pork and eggs [306–308]. Boiret et al. [309] used the PDS algorithm to transfer the transmission NIRS of the coated tablets between two Fourier transform spectrometers of the same model, and the predicted standard deviations of the active ingredient content before and after the transfer were 4.0% and 2.4%, respectively. Sales et al. [310] used the PDS algorithm to transfer the signal of the potential sensor under the two test conditions, selected the transfer standard sample based on the K-S algorithm, and obtained satisfactory results. Marchesini et al. [311] used the PDS algorithm to transfer the NIRS from the desktop instrument of the undried whole maize plant to two portable instruments.

Ge et al. [312] used the PDS algorithm to convert the diffuse reflectance spectra of the soil measured on multiple NIRS instruments of different types and established a model to predict the organic carbon content in the soil. Rodrigues et al. [313] used the PDS algorithm to transfer the MIRS of crude oil on two instruments, and then used the orthogonal projections to latent structures (OPLS) to establish a crude oil density prediction model, and got better results. Li et al. [314] used the PDS algorithm to transfer the spectrum of propylene glycol-water solution measured on multiple handheld NIR spectrometers and realized the rapid identification of ethylene glycol adulteration. Gryniiewicz-Ruzicka et al. [315] used the PDS algorithm to transfer the spectra of multiple Raman spectrometers measuring the content of propylene glycol in glycerol. Thygesen et al. [316] used algorithms such as DS and PDS to transfer excitation-emission three-dimensional fluorescence spectra, and then took advantage of PARAFAC to obtain satisfactory results. Sanlloriente [317] also used a similar method to transfer the three-dimensional fluorescence spectrum between the portable fluorescence spectrometer of the LED light source and the fluorescence spectrometer of the xenon light source. Sun et al. [318] used the PDS algorithm to transfer the three-dimensional fluorescence spectra on the two instruments, and then used the self-weighted alternating normalized residual fitting algorithm (SWANRF) to establish a three-linear decomposition multi-dimensional quantitative model. The results show that PDS can maintain the “second-order advantage” of the second-order tensor calibration methods.

Wang et al. [319] used PDS to realize the transfer of NIRS of leaves picked from different tree species and different periods and solved the problem of standard spectra based on chlorophyll content through linear interpolation. Watari and Ozaki [320] also used a similar method to transfer the NIRS of random polypropylene and block polypropylene in the molten state so that the model of ethylene content in one type of polypropylene can predict another type of polypropylene. Li et al. [321] discussed the influence of the number of PLS latent variables on the transfer performance of the model when the PLS model is established by the NIRS after PDS. The results show that too high number of latent variables can easily cause over-fitting, affect the robustness of the model, and make the error of calibration transfer larger.

Sun et al. [322] used the PDS algorithm to transfer the NIRS of the plasma alcohol precipitation process between two different types of instruments and established an analytical model for predicting the content of total protein, albumin, and globulin. Xiao et al. [323] combined linear interpolation with the PDS algorithm to transfer the FT-NIRS of a single grape to a portable grating spectroscopic spectrum and established a model for predicting the content of soluble solids. Fernandez [324] compared DS, PDS, OSC, GLSW, and other methods to transfer the signal of the gas sensor matrix at different temperatures, and the results of the PDS algorithm are better. Hoffmann et al. [325] used the PDS algorithm to transfer the FT-NIRS to the linear variable filter handheld instrument, and the quantitative and qualitative models obtained good results. Di Anibal et al. [326] used the PDS algorithm to transfer the UV spectrum used to determine whether illegal substances are added to the fragrance, and combined with the PLS-DA method to obtain satisfactory results. Zheng et al. [327] used the PDS method to transfer the NIRS of fish meal on the grating-type instrument to the Fourier-type instrument, and there was no significant difference in predicting the content of crude protein, crude fat, methionine, lysine, and other components.

Pu et al. [328] used the PDS algorithm to successfully transfer the banana spectrum from the handheld NIR spectrometer to the hyperspectral imaging instrument and established a model for predicting the soluble solid content. Xi et al. [329] used the PDS algorithm to transfer the NIRS of amoxicillin capsules and the NIRS of its contents so that the amoxicillin capsule quantitative model can accurately predict and analyze the powder spectrum of the content, and proposed an index to judge whether the spectrum is successfully transferred. Gislason et al. [330] used the PDS algorithm to realize the transfer of process nuclear magnetic resonance spectrum on different instruments, and compared the result of DS combined with the SSC algorithm. Monakhova et al. [331] used the PDS algorithm to transfer the spectrum of the sunflower lecithin and soybean lecithin mixture obtained by three high-resolution NMR instruments, and the result was better than the DS algorithm and the establishment of a hybrid model.

Chen et al. [332] used the PDS algorithm to transfer the UV-visible spectrum of the cuvette with a 10 mm optical path to a fiber optic probe with a 2 mm optical path. Before using the PDS algorithm for transfer, the spectrum was subjected to Fourier transform processing. Lin et al. [333] used the PDS algorithm to transfer the spectrum from the scanning NIRS instrument (cuvette) to the Fourier transform (fiber probe) instrument. Shi et al. [334] used the PDS algorithm to better solve the transfer of the NIRS of the mixture of fish meal and soybean meal on two different spectroscopic types of instruments. Tortajada-Genaro et al. [335] used the PDS algorithm for the transfer of chemiluminescence signals on two instruments and established a model for the rapid determination of Cr(III), Cr(VI), and total Cr content in water through the PLS method. Griffiths et al. [336] used the PDS algorithm combined with the variable selection method to solve the problem of the failure of the multivariate calibration model caused by the drift of the ICP-AES instrument over time.

Wang et al. [337] used the PDS to transfer the NIRS of *Poriacocos* samples measured on two different brands of NIR spectrometers and established a model to

predict alkali-soluble polysaccharides in *Poria cocos*. Morais et al. [338] used the DS and the PDS to standardize the MIR spectrochemical database of complex biological tissues and established a complete spectrum standardization process. Grelet et al. [339, 340] used the PDS algorithm to standardize the instruments in the European dairy MIR spectroscopy network, which can convert spectra on spectrometers of different brands into spectra on the master computer to realize the sharing of quantitative calibration models. Ji et al. [341] used the PDS algorithm to eliminate the influence of moisture and environment on the NIRS of the field soil, and the transferred spectrum can be accurately predicted by the model established in the laboratory. Pierna et al. [342] designed and produced a standard sample pool for the transfer of NIRS microscopic imaging instruments. Different parts of the standard sample pool are equipped with meat and bone meal of different animals, and the spectral transfer of multiple imaging instruments is realized through the PDS algorithm.

17.6.6 CCA Method

Zheng et al. [343] used the CCA algorithm to convert the NIRS between different times and different brands of milk, and the content of dimethyl fumarate in milk can be predicted by the sample enrichment-NIRS measurement method. Liu et al. [344, 345] used the DS algorithm and the CCA algorithm to realize the transfer of the near-infrared spectroscopy analysis model of wood lignin content between different types of portable spectrometers. Yang et al. [346] compared the transfer effects of algorithms such as CCA, SST, CTWM, ICA, and PDS on the transfer of tobacco NIRS on desktop, portable, and handheld instruments, and the results showed that the CCA algorithm is superior to other methods. Luo et al. [347] used the CCA algorithm to transfer the NIRS that predicts the content of polyester in textiles, and the results are better than the PDS method. Fan et al. [348] used the CCA or PDS algorithm combined with linear interpolation method for the transfer of soil near-infrared spectroscopy. A soil nutrient content model can be used to accurately predict soil nutrient content in different regions.

17.6.7 Establishment of Global Model

Eliaerts et al. [349] used the S/B algorithm, the PDS algorithm, and the method of establishing a hybrid global model to transfer the cocaine classification and quantitative SVM models on desktop and portable infrared spectrometers. The results showed that the method of establishing the hybrid model has better results. Yang et al. [350] transferred the model established in the laboratory to feed production enterprises for online application by using the method of hybrid modeling of near-infrared spectroscopy and online spectroscopy, and the predicted values of moisture content and crude protein content were in good agreement with the actual measured

values, which can meet the requirements of online analysis. Chen et al. [351] used the TrAdaBoost algorithm based on principal component analysis and weighted extreme learning machine to establish a global model to realize the sharing of models on multiple instruments. Ni et al. [352] used the establishment of a global hybrid model of multiple instruments to realize the sharing of tobacco models on different near-infrared spectroscopy instruments.

Pereira et al. [353] compared the transfer effects of DS, PDS, OSC, and model update methods based on the NIRS of gasoline on different instruments. The results show that DS combined with model update strategy can get better results. Fernandez-Ahumada et al. [354] used Shenk's algorithm and PDS algorithm to transfer the feed spectrum measured on the laboratory grating near-infrared spectrometer to the online array instrument, and then through the model update, the transfer of the model can be better realized. Debus et al. [355] used the method of establishing a hybrid model to solve the problem of sharing the multi-element calibration model of mid-infrared spectroscopy evaluation environment carbon-containing particulate matter among multiple instruments of the same type.

Krapf et al. [356] used the PDS algorithm to transfer the laboratory NIRS of the samples during the anaerobic digestion process of energy crops to the online analysis instrument, and the problem of online analysis model establishment was better solved through a model update. Li et al. [357] used the combination of Shenk's, PDS, and CCA algorithms and hybrid modeling technology to establish a hybrid model based on a homogeneous tobacco powder model, which was successfully applied to the prediction of nicotine content in heterogeneous cut tobacco samples and tobacco flake samples. Clavaud et al. [358] constructed a global calibration set of more than 3000 spectra of various types of freeze-dried drugs and their moisture content on two NIRS of the same model and established a global model by PLS, SVR, Bayesian Ridge regression, KNN, and other methods. The results show that the predictive ability of SVR was better. Ozdemir et al. [359] used a hybrid global model combined with genetic regression to establish a model of four vis-UV spectrometers, one of which is an array spectrometer, and the other three are dual-beam scanning instruments.

Kupyna et al. [360] used a global model to solve the problem of the application of acoustic multivariate quantitative correction models under different test conditions (temperature, flow rate, etc.). Igne and Hurburgh [361] compared the effects of multiple transfer methods on the same type and different types of NIRS instruments, and the results proved that establishing a stable and sound local model is a better strategy. On the basis of Shenk's algorithm, Fontaine et al. [362] can accurately predict and analyze the model of amino acid content in feed ingredients on dozens of instruments in the NIR network through the model update strategy. Steinbach et al. [363] established a hybrid calibration model for the drug transmission Raman spectra measured on the two instruments, and the model built can be accurately applied to the spectra obtained on the two instruments.

17.6.8 Other Methods

Xu et al. [364] realized the transfer between NIR spectra of rice single grains and rice flours of different varieties by spectral space transformation. Online monitoring of tobacco nicotine and total sugar was realized through the SST from the offline model [365]. In addition, calibration transfer by SST algorithm has also been applied to edible oil acid value and peroxide value [366], rice [367], the hyperspectral image on plant phenotype [368], and Terahertz spectral instruments [369].

Yang et al. [370] successfully transferred the apple spectra on two portable NIR spectrometers using simple linear regression direct standardization (SLRDS) algorithm. Salguero-Chaparro et al. [371] used the transfer by orthogonal projection (TOP) algorithm to transfer the olive spectra from the grating NIR spectrometer to the array portable spectrometer and established an analytical model to predict fat, free acid, and moisture content. Liu et al. [372] compared the transfer results of SBC, OSC, DS, PDS, and local centralization to the silage NIR spectra on the same and different types of spectrometers. They further studied the influence of different temperatures and measuring accessories on the NIRS of rice straw and found local centralization method can eliminate the influences on the spectra to a certain extent [373]. A similar study on pharmaceutical samples was done by Bergman et al. and the result revealed local centralization can be realized with fewer standard samples [374].

Li et al. [375] combined the wavelet multi-scale piecewise direct standardization (WMPDS) with the SBC algorithm to realize the transfer of NIR spectra of different types of soils and the correction of prediction for total nitrogen and total carbon content. Greensill et al. [376] and Walczak et al. [377] compared the effects of DS, PDS, DWPDS, OSC, FIR, and WT on the transfer of citrus spectra between micro-array detector NIR spectrometers, and the results showed WT and model update is better than others.

Martins et al. [378] combined the SPA wavelength selection algorithm with the multi-model consensus strategy to establish a calibration model by MLR, which was proved to be more effective than the PDS-PLS model. Yoon et al. [379] performed first-order derivative and OSC preprocessing on NIRS of gasoline and then transferred the spectra on the two instruments through the PDS algorithm, eventually establishing the model for predicting the benzene content in gasoline. Yahaya et al. [380] established an analytical method for rapid prediction of mango acidity by the variable screening coupled with MLR, realizing the application to multiple instruments by optimization of variable selection.

References

1. Shenk JS, Westerhaus MO, Templeton WC. Calibration transfer between near infrared reflectance spectrophotometers. *Crop Sci.* 1985;25:159–61.
2. Vogt F, Booksh K. Influence of wavelength-shifted calibration spectra on multivariate calibration models. *Appl Spectrosc.* 2004;58(5):624–35.

- Mann CK, Vickers TJ. Instrument-to-instrument transfer of Raman spectra. *Appl Spectrosc.* 1999;53(7):856–61.
- Blanco M, Coello J, Iturriaga H, et al. Wavelength calibration transfer between diode array UV-visible spectrophotometers. *Appl Spectrosc.* 1995;49(5):593–7.
- Fearn T, Eddison C, Withey R, et al. A method for wavelength standardisation in filter instruments. *J Near Infrared Spectrosc.* 1996;4(1):111–8.
- Busch KW, Soyemi O, Rabbe D, et al. Wavelength calibration of a dispersive near-infrared spectrometer using trichloromethane as a calibration standard. *Appl Spectrosc.* 2000;54(9):1321–6.
- Martinsen P, Jordan B, McGlone A, et al. Accurate and precise wavelength calibration for wide bandwidth array spectrometers. *Appl Spectrosc.* 2008;62(9):1008–12.
- Martinsen P, McGlone VA, Jordan RB, et al. Temporal sensitivity of the wavelength calibration of a photodiode array spectrometer. *Appl Spectrosc.* 2010;64(12):1325–9.
- Ray KG, McCreery RL. Simplified calibration of instrument response function for Raman spectrometers based on luminescent intensity standards. *Appl Spectrosc.* 1997;51(1):108–16.
- Yang H, Isaksson T, Jackson RS, et al. Effect of resolution on the wavenumber determination of a putative standard to be used for near infrared diffuse reflection spectra measured on fourier transform near infrared spectrometers. *J Near Infrared Spectrosc.* 2003;11(4):241–55.
- Isaksson T, Yang H, Kemeny GJ, et al. Accurate wavenumber measurements of a putative standard for near-infrared diffuse reflection spectrometry. *Appl Spectrosc.* 2003;57(2):176–85.
- Soyemi O, Rabbe D, Busch MA, et al. Design of a modular, dispersive spectrometer for fundamental studies in near-infrared spectroscopy. *Spectroscopy.* 2001;16(4):24–33.
- Workman JJ. First principles of instrument calibration. *NIR News.* 2016;27(3):12–5.
- Ridder TD, Steeg BJV, Price GL. Robust calibration transfer in noninvasive ethanol measurements, Part I: mathematical basis for spectral distortions in Fourier Transform Near-Infrared Spectroscopy (FT-NIR). *Appl Spectrosc.* 2014;68(8):852–64.
- Ridder TD, Ver Steeg BJ, Laaksonen BD, et al. Robust calibration transfer in noninvasive ethanol measurements, Part II: Modification of instrument measurements by incorporation of expert knowledge (Mimik). *Appl Spectrosc.* 2014;68(8):865–78.
- Xu JL, Dorrepaal RM, Martinez-Gonzalez J, et al. Near-infrared multivariate model transfer for quantification of different hydrogen bonding species in aqueous systems. *J Chemomet.* 2020;34:e3274.
- Terrell M. Two case studies of the transfer of near infrared methods for the analysis of pharmaceutical solid dosage forms. *NIR News.* 2015;26(5):8–9.
- Wang Q, De Jesus S, Conzen JP, et al. Calibration transfer in near infrared analysis of liquids and solids. *J Near Infrared Spectrosc.* 1998;6(A):A201–5.
- Cinier R, Guilment J. High precision measurements: from the laboratory to the plant. *J Near Infrared Spectrosc.* 1998;6(1):291–7.
- Sun L, Hsiung C, Smith V. Investigation of direct model transferability using miniature near-infrared spectrometers. *Molecules.* 2019;24(10):1997.
- Hacisalihoglu G, Gustin JL, Louisma J, et al. Enhanced single seed trait predictions in Soybean (Glycine max) and robust calibration transfer with near-infrared reflectance-spectroscopy. *J Agric Food Chem.* 2016;64:1079–86.
- Aldridge PK, Evans CL, Ward HW, et al. Near-IR detection of polymorphism and process-related substances. *Anal Chem.* 1996;68:997–1002.
- Barnes SE, Thurston T, Coleman JA, et al. NIR diffuse reflectance for on-scale monitoring of the polymorphic form transformation of pazopanib hydrochloride (GW786034); Model Development and method transfer. *Anal Methods.* 2010;2:1890–9.
- Isabelle M, Dorney J, Lewis A, et al. Multi-centre raman spectral mapping of oesophageal cancer tissues: a study to assess system transferability. *Faraday Discuss.* 2016;187:87–103.
- Pissarda A, Marques EJM, Dardenne P, et al. Evaluation of a handheld ultra-compact NIR spectrometer for rapid and non-destructive determination of apple fruit quality. *Postharvest Biol Technol.* 2021;172:111375.

26. Rodgers JE, Ghosh S, Cardwell WD. Measuring nylon carpet yarn heat history by remote NIR spectroscopy. Part II: Applying remote fiber optic NIR techniques to the manufacturing environment. *Text Res J.* 2001;71(2):135–44.
27. Sun L, Hsiung C, Pederson CG, et al. Pharmaceutical raw material identification using miniature near-infrared (MicroNIR) spectroscopy and supervised pattern recognition using support vector machine. *Appl Spectrosc.* 2016;70(5):816–25.
28. Via BK, So CL, Shupe TF, et al. Prediction of wood mechanical and chemical properties in the presence and absence of blue stain using two near infrared instruments. *J Near Infrared Spectrosc.* 2005;13(4):201–12.
29. Bakeev KA, Kurtyka B. Sources of measurement variability and their effect on the transfer of near infrared spectral libraries. *J Near Infrared Spectrosc.* 2005;13(6):339–48.
30. Hutsebaut D, Vandenabeele P, Moens L. Evaluation of an accurate calibration and spectral standardization procedure for Raman spectroscopy. *Analyst.* 2005;130(8):1204–14.
31. Choquette SJ, Etz ES, Hurst WS, et al. Relative intensity correction of Raman spectrometers: NIST SRMS 2241 through 2243 for 785 nm, 532 nm, and 488 nm/514.5 nm excitation. *Appl Spectrosc.* 2007;61(2):117–29.
32. Rodriguez JD, Westenberger BJ, Buhse LF, et al. Standardization of Raman spectra for transfer of spectral libraries across different instruments. *Analyst.* 2011;136(20):4232–40.
33. Chen H, Zhang ZM, Miao L, et al. Automatic standardization method for Raman spectrometers with applications to pharmaceuticals. *J Raman Spectrosc.* 2015;46(1):147–54.
34. Coleman MD, Brewer PJ, Smith IM, et al. Calibration transfer strategy to compensate for instrumental drift in portable quadrupole mass spectrometers. *Anal Chim Acta.* 2007;601(2):189–95.
35. Pavón JLP, Sánchez MDN, Pinto CG, et al. Calibration transfer for solving the signal instability in quantitative headspace-mass spectrometry. *Anal Chem.* 2003;75(22):6361–7.
36. Bergman EL, Brage H, Leion H, et al. Transfer of NIR calibrations between sites and different instruments. *NIR News.* 2003;14(4):6–7.
37. Drennen J. Calibration transfer: a critical component of analytical method validation. *NIR News* 2003;14(5):14–5.
38. Sohn M, Himmelsbach DS, Barton FE, et al. Transfer of calibrations for barley quality from dispersive instrument to Fourier transform near-infrared instrument. *Appl Spectrosc.* 2009;63(10):1190–6.
39. De Noord OE. Multivariate calibration standardization. *Chemomet Intell Lab Syst.* 1994;25(2):85–97.
40. Chu XL, Yuan HF, Lu WZ. Model transfer in multivariate calibration. *Spectrosc Spect Anal.* 2001;21(6):881–5.
41. Fearn T. Standardization and calibration transfer for near infrared instruments: a review. *J Near Infrared Spectrosc.* 2001;9(4):229–44.
42. Zhang J, Cai WS, Shao XG. New algorithms for calibration transfer in near infrared spectroscopy. *Progr Chem.* 2017;29(8):101–9.
43. Malli B, Birlutiu A, Natschlager T. Standard-free calibration transfer—an evaluation of different techniques. *Chemomet Intell Lab Syst.* 2017;161(1):49–60.
44. Feudale RN, Woody NA, Tan H, et al. Transfer of multivariate calibration models: a review. *Chemomet Intell Lab Syst.* 2002;64(2):181–92.
45. Zhang XB, Feng YC, Hu CQ. Progress in calibration transfer of near-infrared multivariate model. *Chin J Pharmaceut Anal.* 2009;29(8):1390–9.
46. Shi YY, Li JY, Chu XL. Progress and applications of multivariate calibration model transfer methods. *Chin J Anal Chem.* 2019;47(4):479–87.
47. Lima FSG, Borge LEP. Evaluation of standardisation methods of near infrared calibration models. *J Near Infrared Spectrosc.* 2002;10(4):269–78.
48. Leion H, Folestad S, Josefson M, et al. Evaluation of basic algorithms for transferring quantitative multivariate calibrations between scanning grating and FT NIR spectrometers. *J Pharm Biomed Anal.* 2005;37(1):47–55.

49. RukundoI R, Danao MGC, Weller CL, et al. Use of a handheld near infrared spectrometer and partial least squares regression to quantify metanil yellow adulteration in turmeric powder. *J Near Infrared Spectrosc.* 2020;28(2):81–92.
50. Forina M, Drava G, Armanino C, et al. Transfer of calibration function in near-infrared spectroscopy. *Chemomet Intell Lab Syst.* 1995;27:189–203.
51. Dardenne P. Calibration transfer in near infrared spectroscopy. *NIR News* 2002;13(4):3–7.
52. Hopkins DW. Shoot-out 2002: transfer of calibration for content of active in a pharmaceutical tablet. *NIR News* 2003;14(5):10–3.
53. Ozdemir D, Mosley M, Williams R. Hybrid calibration models: an alternative to calibration transfer. *Appl Spectrosc.* 1998;52(4):599–603.
54. Dardenne P, Welle R. New approach for calibration transfer from a local database to a global database. *J Near Infrared Spectrosc.* 1998;6(1):55–60.
55. Kramer KE, Small GW. Blank augmentation protocol for improving the robustness of multivariate calibrations. *Appl Spectrosc.* 2007;61(5):497–506.
56. Swierenga H, Haanstra WG, Weijer APD, et al. Comparison of two different approaches toward model transferability in NIR spectroscopy. *Appl Spectrosc.* 1998;52(1):7–16.
57. Workman JJ. A review of calibration transfer practices and instrument differences in spectroscopy. *Appl Spectrosc.* 2018;72(3):340–65.
58. Smith MR, Jee RD, Moffat AC, et al. A procedure for calibration transfer between near-infrared instruments—a worked example using a transmittance single tablet assay for piroxicam in intact tablets. *Analyst* 2004;129(9):806–16.
59. Bouveresse E, Massart D, Dardenne P. Calibration transfer across near-infrared spectrometric instruments using Shenk’s algorithm: effects of different standardisation samples. *Anal Chim Acta.* 1994;297(3):405–16.
60. Shenk JS, Westerhaus MO. New standardization and calibration procedures for NIRS analytical systems. *Crop Sci.* 1991;31:1694–6.
61. Hoffmann U, Zanier-Szydowski N. Portability of near infrared spectroscopic calibrations for petrochemical parameters. *J Near Infrared Spectrosc.* 1999;7(1):33–45.
62. Wang YD, Veltkamp DJ, Kowalski BR. MultivariateInstrument standardization. *Anal Chem.* 1991;63(23):2750–6.
63. Dreassi E, Ceramelli G, Perruccio PL, et al. Transfer of calibration in near-infrared reflectance spectrometry. *Analyst* 1998;123:1259–64.
64. Wang Y D, Kowalski BR. Calibration transfer and measurement stability of near-infrared spectrometers. *Appl Spectrosc.* 1992;46(5):764–71.
65. Wang Y D, Kowalski BR. Temperature-compensating calibration transfer for near-infrared filter instruments. *Anal Biochem.* 1993;65:1301–3.
66. Bouveresse E, Hartmann C, Massart DL, et al. Standardization of near-infrared spectrometric instruments. *Anal Chem.* 1996;68:982–90.
67. Wang YD, Lysaght MJ, Kowalski BR. Improvement of multivariate calibration through instrument standardization. *Anal Chem.* 1992;64(5):562–5.
68. Despaigne F, Walczak B, Massart DL. Transfer of calibrations of near-infrared spectra using neural networks. *Appl Spectrosc.* 1998;52(5):732–45.
69. Duponche L, Ruckebusch C, Huvenne JP, et al. Standardisation of near-IR spectrometers using artificial neural networks. *J Mol Struct.* 1999;480–1: 551–6.
70. Greensill CV, Wolfs PJ, Spiegelman CH, et al. Calibration transfer between PDA-Based NIR spectrometers in the NIR assessment of melon soluble solids content. *Appl Spectrosc.* 2001;55(5):647–53.
71. Igne B, Hurburgh CR. Using the frequency components of near infrared spectra: optimising calibration and standardisation processes. *J Near Infrared Spectrosc.* 2010;18(1):39–47.
72. Anderson CE, Kalivas JH. Fundamentals of calibration transfer through procrustes analysis. *Appl Spectrosc.* 1999;53(10):1268–76.
73. Chu XL, Yuan HF, Lu WZ. Calibration transfer of spectra from near infrared spectrometers. *Chin J Anal Chem.* 2002;30(1):114–9.

74. Wang YB, Yuan HF, Lu WZ. A new calibration transfer method based on target factor analysis. *Spectrosc Spect Anal.* 2005;25(3):398–401.
75. Andrews DT, Wentzell PD. Applications of maximum likelihood principal component analysis: incomplete data sets and calibration transfer. *Anal Chim Acta.* 1997;350(3):341–52.
76. Folch-Fortuny A, Vitale R, De Noord OE, et al. Calibration transfer between NIR spectrometers: new proposals and a comparative study. *J Chemom.* 2017;31(3):e2874–84.
77. Bouveresse E, Massart DL. Standardisation of near-infrared spectrometric instruments: a review. *Vib Spectrosc.* 1996;11(1):3–15.
78. Sales F, Callao MP, Rius FX. Multivariate standardization techniques on ion-selective sensor arrays. *Analyst.* 1999;124:1045–51.
79. Tillmann P, Reinhardt TC, Paul C. Networking of near infrared spectroscopy instruments for rapeseed analysis: a comparison of different procedures. *J Near Infrared Spectrosc.* 2000;8(2):101–7.
80. Hong TL, Tsai SJ, Tsou SCS. Development of a Sample Set for Soya bean calibration of near infrared reflectance spectroscopy. *J Near Infrared Spectrosc.* 1994;2(4):223–7.
81. Capron X, Walczak B, De Noord OE, et al. Selection and weighting of samples in multivariate regression model updating. *Chemom Intell Lab Syst.* 2005;76(2):205–14.
82. Siano GG, Goicoechea HC. Representative subset selection and standardization techniques. a comparative study using NIR and a simulated fermentative process UV data. *Chemom Intell Lab Syst.* 2007;88(2):204–12.
83. Clark RD. Optimism: an extended dissimilarity selection method for finding diverse representative subsets. *J Chem Inform Comput Sci.* 1997;37(6):1181–8.
84. Li H, Wang JX, Xing ZN, et al. Influence of improved Kennard/Stone algorithm on the calibration transfer in near-infrared spectroscopy. *Spectrosc Spect Anal.* 2011;31(2):362–5.
85. Zhou ZK, Li CX, Wang Z, et al. Study on cefradine granules component analysis and calibration transfer method based on near-infrared spectroscopy. *Spectrosc Spect Anal.* 2020;40(11):3562–6.
86. Liang C, Yuan HF, Zhao Z, et al. A new multivariate calibration model transfer method of near-infrared spectral analysis. *Chemom Intell Lab Syst.* 2016;153(1):51–7.
87. Zheng KY, Feng T, Zhang W, et al. Refining transfer set in calibration transfer of near infrared spectra by backward refinement of samples. *Anal Methods.* 2020;12:1495–503.
88. Ni LJ, Dong XX, Zhang LG, et al. Standard sample preparation method for near-infrared model transfer of natural plants and its application. *ZLCN201711264567.X [P]*, 17 July 2018.
89. Xin XW, Gong HL, Ding XQ, et al. Study on calibration model transfer for the near infrared spectrum based on improved S/B algorithm. *Spectrosc Spect Anal.* 2017;37(12):3709–13.
90. Cao YT, Yuan HF, Zhao Z. A new spectra transfer method for multivariate calibration model of molecular spectroscopy analysis. *Spectrosc Spect Anal.* 2018;38(3):973–81.
91. Zhang FY, Chen W, Zhang RQ, et al. Sampling error profile analysis for calibration transfer in multivariate calibration. *Chemom Intell Lab Syst.* 2017;171(1):234–40.
92. Blanco M, Peguero A. Analysis of pharmaceuticals by NIR spectroscopy without a reference method. *Tr AC: Trends Anal Chem.* 2010;29(10):1127–36.
93. Blanco M, Cueva MR, Peguero A. NIR analysis of pharmaceutical samples without reference data: improving the calibration. *Talanta.* 2011;85(4):2218–25.
94. Wang JJ, Zhe W, Liu Y, et al. A calibration transfer method for NIR model based on extended spectrum. *Acta Tabacaria Sinica.* 2014;20(6):1–6.
95. Li J, Yu XN, Ge WZ, et al. Qualitative analysis of maize haploid kernels based on calibration transfer by near-infrared spectroscopy. *Anal Lett.* 2019;52(2):249–67.
96. Wang JJ, Li ZF, Wang Y, et al. A dual model strategy to transfer multivariate calibration models for near-infrared spectral analysis. *Spectrosc Lett.* 2016;49(5):348–54.
97. Li XY, Cai WS, Shao XG. Correcting multivariate calibration model for near infrared spectral analysis without using standard samples. *J Near Infrared Spectrosc.* 2011;23(5):285–91.
98. Tan H, Sum ST, Brown SD. Improvement of a standard-free method for near-infrared calibration transfer. *Appl Spectrosc.* 2002;56(8):1098–106.

99. Sum ST, Brown SD. Standardization of fiber-optic probes for near-infrared multivariate calibrations. *Appl Spectrosc.* 1998;52(6):869–77.
100. Bouveresse E, Massart DL, Dardenne P. Modified algorithm for standardization of near-infrared spectrometric instruments. *Anal Chem.* 1995;67:1381–9.
101. Wang ZY, Dean T, Kowalski BR. Additive background correction in multivariate instrument standardization. *Anal Chem.* 1995;67:2379–85.
102. Gemperline PJ, Cho JH, Aldridge PK, et al. Appearance of discontinuities in spectra transformed by the piecewise direct instrument standardization procedure. *Anal Chem.* 1996;68:2913–5.
103. Wang H, Lin ZX, Wu BL, et al. Spectral analysis model transfer technology based on radial basis neural network. ZL201610396494.9[P], 2 Nov 2016.
104. Chen XS, Jiao YP, Su M, et al. A solution to strange peaks in near-infrared spectroscopy calibration transfer. ZL201811189178.X [P], 8 March 2019.
105. Yang HH, Zhang XF, Fan YX, et al. Near infrared spectroscopic model transfer based on simple linear regression. *Chin J Anal Chem.* 2014;42(9):1229–34.
106. Norgaard L. Direct standardisation in multi wavelength fluorescence spectroscopy. *Chemom Intell Lab Syst.* 1995;29(2):283–93.
107. Galvao RKH, Soares SFC, Martins MN, et al. Calibration transfer employing univariate correction and robust regression. *Anal Chim Acta.* 2015;864(1):1–8.
108. Lu HX, Wu PF, Yang HH, et al. A NIR model transfer method based on least angle regression combined with simple linear regression direct standardization. *J Instr Anal.* 2019;38(1):39–45.
109. Wang QB, Yang HH, Pan XP, et al. A near infrared spectroscopy model transfer method based on wavelet transform combined with dynamic time warping. *J Instr Anal.* 2019;38(12):1423–9.
110. Zou CM, Zhu HM, Shen JR, et al. Scalable calibration transfer without standards via dynamic time warping for near-infrared spectroscopy. *Anal Methods.* 2019;35(11):4481–93.
111. Yan K, Zhang D. Improving the transfer ability of prediction models for electronic noses. *Sens Actuat B Chem.* 2015;220(1):115–24.
112. Oliveri P, Casolino MC, Casale M, et al. A spectral transfer procedure for application of a single class-model to spectra recorded by different near-infrared spectrometers for authentication of Olives in Brine. *Anal Chim Acta.* 2013;761(1):46–52.
113. Greensill CV, Walsh KB. Calibration transfer between miniature photodiode array-based spectrometers in the near infrared assessment of Mandarin soluble solids content. *J Near Infrared Spectrosc.* 2002;10:27–35.
114. Ottaway J, Kalivas JH. Feasibility study for transforming spectral and instrumental artifacts for multivariate calibration maintenance. *Appl Spectrosc.* 2015;69(3):407–16.
115. Chen WR, Bin J, Lu HM, et al. Calibration transfer via an extreme learning machine auto-encoder. *Analyst.* 2016;141(6):1973–80.
116. Laref R, Losson E, Sava A, et al. Support vector machine regression for calibration transfer between electronic noses dedicated to air pollution monitoring. *Sensors.* 2018;18(11):3716.
117. Li XY, Zhang HG, Lu JG, et al. A new method of model transfer in near infrared spectral quantitative analysis. *Comput Appl Chem.* 2018;35(1):27–36.
118. Tan HW, Brown SD. Wavelet hybrid direct standardization of near-infrared multivariate calibrations. *J Chemom.* 2001;15(8):647–63.
119. Chen D, Lu F, Li QF. Development of multi-scale modeling methods for calibration transfer in near infrared spectroscopy. *Nanotechnol Precis Eng.* 2017;15(2):121–6.
120. Yoon J, Lee B, Han C. Calibration transfer of near-infrared spectra based on compression of wavelet coefficients. *Chemom Intell Lab Syst.* 2002;64(1):1–14.
121. Tan C, Li ML. Calibration transfer between two near-infrared spectrometers based on a wavelet packet transform. *Anal Sci.* 2007;23(2):201–6.
122. Ni WD, Brown SD, Man RL. Data fusion in multivariate calibration transfer. *Anal Chim Acta.* 2010;661(2):133–42.
123. Poesio DV, Brown SD. Dual-domain calibration transfer using orthogonal projection. *Appl Spectrosc.* 2017;72(3):378–91.

124. Lin ZZ, Xu B, Li Y, et al. Application of orthogonal space regression to calibration transfer without standards. *J Chemom.* 2013;27(11):406–13.
125. Wang AD, Wu ZS, Jia YF, et al. Model transfer of on-line pilot-scale near infrared quantitative model based on orthogonal signal regression. *Spectrosc Spect Anal.* 2018;38(4):1082–8.
126. Yang P, Chen J, Wu CY, et al. Achievement of moisture transfer of near infrared quantitative model from small-test preparation process to pilot-test by directed direct orthogonal signal correction combined with slope/bias correction. *J Instr Anal.* 2019;38(9):1044–50.
127. Wang AD, Yang P, Chen J, et al. A new calibration model transferring strategy maintaining the predictive abilities of NIR multivariate calibration model applied in different batches process of extraction. *Infrared Phys Technol.* 2019;103:103046.
128. Wang QB, Yang HH, Pan XP, et al. A model transfer method based on random forest-direct orthogonal signal correction. *Laser Infrared.* 2020;50(9):1081–7.
129. Lin J. Near-IR calibration transfer between different temperatures. *Appl Spectrosc.* 1998;52:1591–6.
130. Wulfert F, Kok WT, Noord OED, et al. Correction of temperature-induced spectral variation by continuous piecewise direct standardization. *Anal Chem.* 2000;72(7):1639–44.
131. Barring HK, Boelens HFM, De Noord OE, et al. Optimizing meta-parameters in continuous piecewise direct standardization. *Appl Spectrosc.* 2001;55(4):458–66.
132. Jaworski A, Wikiel K, Wikiel K. Temperature Compensation by calibration transfer for an AC voltammetric analyzer of electroplating baths. *Electroanalysis.* 2017;29(1):67–76.
133. Jaworski A, Wikiel H, Wikiel K. Temperature compensation by embedded temperature variation method for an AC voltammetric analyzer of electroplating baths. *Electroanalysis.* 2018;30(1):1–12.
134. Wei F, Liang YZ, Yuan DL, et al. Calibration model transfer for near-infrared spectra based on canonical correlation analysis. *Anal Chim Acta.* 2008;623(1):22–9.
135. Zheng KY, Zhang X, Iqbal J, et al. Calibration transfer of near-infrared spectra for extraction of informative components from spectra with canonical correlation analysis. *J Chemom.* 2014;28(10):773–84.
136. Bin J, Li X, Fan W, et al. Calibration transfer of near-infrared spectroscopy by canonical correlation analysis coupled with wavelet transform. *Analyst.* 2017;142(12):2229–38.
137. Fan XQ, Lu HM, Zhang ZM. Direct calibration transfer to principal components via canonical correlation analysis. *Chemom Intell Lab Syst.* 2018;181(1):21–8.
138. Peng JG, Peng SL, Jiang A, et al. Near-Infrared calibration transfer based on spectral regression. *Spectrochim Acta Part A Mol Biomol Spectrosc.* 2011;78(4):1315–20.
139. Zhang J, Guo C, Cui XY, et al. A two-level strategy for standardization of near infrared spectra by multi-level simultaneous component analysis. *Anal Chim Acta.* 2019;1050(1):25–31.
140. Du W, Chen ZP, Zhong LJ, et al. Maintaining the predictive abilities of multivariate calibration models by spectral space transformation. *Anal Chim Acta.* 2011;690(1):64–70.
141. Liu Y, Xu H, Xia ZZ, et al. Multi-spectrometer calibration transfer based on independent component analysis. *Analyst.* 2018;143(5):1274–80.
142. Liu Y, Cai WS, Shao XG. Standardization of near infrared spectra measured on multi-instrument. *Anal Chim Acta.* 2014;836(1):18–23.
143. Kompany-Zareh M, Berg FVD. Multi-way based calibration transfer between two Raman spectrometers. *Analyst.* 2010;135(6):1382–8.
144. Yu BF, Ji HB, Yu K. Standardization of near infrared spectra based on multi-task learning. *Spectrosc Lett.* 2016;49(1):23–9.
145. Boucher T, Dyar MD, Mahadevan S. Proximal methods for calibration transfer. *J Chemom.* 2017;31(4):e2877–85.
146. Hu Y, Peng SL, Bi YM, et al. Calibration transfer based on maximum margin criterion for qualitative analysis using fourier transform infrared spectroscopy. *Analyst.* 2012;137(24):5913–8.
147. Cogdill RP, Anderson CA, Drennen JK. Process analytical technology case study, Part III: Calibration monitoring and transfer. *AAPS Pharm Sci Tech.* 2002;6(2):284–97.
148. Martens H, Hoy M, Wise BM, et al. Pre-whitening of data by covariance-weighted pre-processing. *J Chemom.* 2003;17(3):153–65.

149. Andrew A, Fearn T. Transfer by orthogonal projection: making near-infrared calibrations robust to between-instrument variation. *Chemomet Intell Lab Syst.* 2004;72(1):51–6.
150. Zhu Y, Fearn T, Samuel D, et al. Error removal by orthogonal subtraction (EROS): a customised pre-treatment for spectroscopic data. *J Chemom.* 2008;22(1):130–4.
151. Zeaiter M, Roger JM, Bellon-Maurel V. Dynamic orthogonal projection. A new method to maintain the on-line robustness of multivariate calibrations. Application to NIR-based monitoring of wine fermentations. *Chemomet Intell Lab Syst.* 2006;80(2):225–36.
152. Dabros M, Amrhein M, Gujral P, et al. On-line recalibration of spectral measurements using metabolite injections and dynamic orthogonal projection. *Appl Spectrosc.* 2007;61(5):507–13.
153. Igne B, Roger JM, Roussel S, et al. Improving the transfer of near infrared prediction models by orthogonal methods. *Chemom Intell Lab Syst.* 2009;99(1):57–65.
154. Siska JJ, Hurburgh CR. The standardisation of near infrared instruments using master selection and wiener filter methods. *J Near Infrared Spectrosc.* 2001;9:107–16.
155. Chen ZP, Morris J, Martin E. Correction of temperature-induced spectral variations by loading space standardization. *Anal Chem.* 2005;77(5):1376–84.
156. Chen ZP, Morris J. Improving the linearity of spectroscopic data subjected to fluctuations in external variables by the extended loading space standardization. *Analyst.* 2008;133(7):914–22.
157. Shi XZ, Wang ZG, Du W, et al. On-line quantitative monitoring and control of tobacco flavors by near infrared spectroscopy combined with advanced calibration transfer method. *Chin J Anal Chem.* 2014;42(11):1673–8.
158. Zhang L, Tian F, Kadri C, et al. On-line sensor calibration transfer among electronic nose instruments for monitoring volatile organic chemicals in indoor air quality. *Sens Actuat B: Chem.* 2011;160(1):899–909.
159. Deshmukh S, Kamde K, Jana A, et al. Calibration transfer between electronic nose systems for rapid in situ measurement of pulp and paper industry emissions. *Anal Chim Acta.* 2014;841(1):58–67.
160. Zhao YH, Zhao ZH, Shan P, et al. Calibration transfer based on affine invariance for nir without transfer standards. *Molecules.* 2019;24(9):1802.
161. Munoz SG, Macgregor JF, Kourti T. Product transfer between sites using joint-Y PLS. *Chemom Intell Lab Syst.* 2005;79(1–2):101–14.
162. Shan P, Zhao YH, Wang QY, et al. Principal component analysis or kernel principal component analysis based joint spectral subspace method for calibration transfer. *Spectrochimica Acta Part A: Mol Biomol Spectrosc.* 2020;227:117653.
163. Khaydukova M, Panchuk V, Kirsanov D, et al. Multivariate calibration transfer between two potentiometric multisensor systems. *Electroanalysis.* 2017;29(9):2161–6.
164. Zhao YH, Yu JL, Shan P, et al. PLS subspace-based calibration transfer for near-infrared spectroscopy quantitative analysis. *Molecules.* 2019;249(7):1289–306.
165. Zhang FY, Zhang RQ, Ge J, et al. Calibration transfer based on the weight matrix (CTWM) of PLS for near infrared (NIR) spectral analysis. *Anal Methods.* 2018;10(18):2169–79.
166. Chen ZP, Li LM, Yu RQ, et al. Systematic prediction error correction: a novel strategy for maintaining the predictive abilities of multivariate calibration models. *Analyst.* 2011;136(1):98–106.
167. Mou Y, Zhou L, Yu S, et al. Robust calibration model transfer. *Chemom Intell Lab Syst.* 2016;156(1):62–71.
168. Seichter F, Vogt J, Radermacher P, et al. Nonlinear calibration transfer based on hierarchical Bayesian models and lagrange multipliers: error bounds of estimates via Monte Carlo E Markov chain sampling. *Anal Chim Acta.* 2017;951(1):32–45.
169. Seichter F, Vogt J, Radermacher P, et al. Response-surface fits and calibration transfer for the correction of the oxygen effect in the quantification of carbon dioxide via FTIR spectroscopy. *Anal Chim Acta.* 2017;972(1):16–27.
170. Skotare T, Nilsson D, Xiong S, et al. Joint and unique multiblock analysis for integration and calibration transfer of NIR instruments. *Anal Chem.* 2019;91(5):3516–24.

171. Andries E. Penalized eigendecompositions: motivations from domain adaptation for calibration transfer. *J Chemom.* 2017;31(4):e2818–31.
172. Liu CL, Zhou ZY, Li TR, et al. Application of migration learning in transfer of oil spectral model. *J Food Sci Technol.* 2019;37(4):95–102.
173. Tao C, Wang Y, Cui W, et al. A transferable spectroscopic diagnosis model for predicting arsenic contamination in soil. *Sci Total Environ.* 2019;669:964–72.
174. Zheng WR, Li SW, Han YL, et al. Study on transfer learning prediction methods for soil available phosphorus NIR. *J Instr Anal.* 2020;38(10):1274–81.
175. Shi GY, Cao J, Zhang YZ. Near infrared wood defects detection based on transfer learning. *Electr Mach Control.* 2020;24(10):159–66.
176. Shan P, Zhao YH, Wang QY, et al. Principal component analysis or kernel principal component analysis based joint spectral subspace method for calibration transfer. *Spectrochimica Acta Part A: Mol Biomol Spectrosc.* 2020;7:117653.
177. Nikzad-Langerodi R, Zellinger W, Lughofer E, et al. Domain-invariant partial-least-squares regression. *Anal Chem.* 2018;90(11):6693–701.
178. Mishra P, Nikzad-Langerodi R. Partial least square regression versus domain invariant partial least square regression with application to near-infrared spectroscopy of fresh fruit. *Infrared Phys Technol.* 2020;111:103547.
179. Huang GG, Chen XJ, Li LM, et al. Domain adaptive partial least squares regression. *Chemomet Intell Lab Syst.* 2020;201:103986.
180. Yan K, Zhang D. Calibration transfer and drift compensation of E-noses via coupled task learning. *Sens Actuat B Chem.* 2016;225(1):288–97.
181. Hu M, Li QL. An efficient model transfer approach to suppress biological variation in elastic modulus and firmness regression models using hyperspectral data. *Infrared Phys Technol.* 2019;99(1):140–51.
182. Li QF, Sun XQ, Ma XY, et al. A calibration transfer methodology for standardization of Raman instruments with different spectral resolutions using double digital projection slit. *Chemomet Intell Lab Syst.* 2019;191(1):143–7.
183. Liu ZW, Xu LJ, Chen XJ. Near infrared spectroscopy transfer based on deep autoencoder. *Spectrosc Spect Anal.* 2020;40(7):2313–8.
184. Liu Y, Cai WS, Shao XG. Linear model correction: a method for transferring a near-infrared multivariate calibration model without standard samples. *Spectrochim Acta Part A Mol Biomol Spectrosc.* 2016;169(1):197–201.
185. Zhang J, Cui XY, Cai WS, et al. Modified linear model correction: a calibration transfer method without standard samples. *NIR news.* 2018;29(8):24–7.
186. Kauppinen A, Toiviainen M, Lehtonen M, et al. Validation of a multipoint near-infrared spectroscopy method for in-line moisture content analysis during freeze-drying. *J Pharm Biomed Anal.* 2014;95(1):229–37.
187. Eskildsen C, Hansen P, Skov T, et al. Evaluation of multivariate calibration models transferred between spectroscopic instruments: applied to near infrared measurements of flour samples. *J Near Infrared Spectrosc.* 2016;24(2):151–6.
188. Fearn T. Calibration transfer without standards. *NIR News.* 1997;8(5):7–8.
189. Adhietty IS, McGuire JA, Wangmaneerat B, et al. Achieving transferable multivariate spectral calibration models: demonstration with infrared spectra of thin-film dielectrics on Silicon. *Anal Chem.* 1991;63:2329–38.
190. Zeaiter M, Roger JM, Bellon-Maurel V, et al. Robustness of models developed by multivariate calibration. Part I: The assessment of robustness. *Trends Anal Chem.* 2004;23(2):157–70.
191. Zeaiter M, Roger JM, Bellon-Maurel V. Robustness of models developed by multivariate calibration: Part II: The influence of pre-processing methods. *Trends Anal Chem.* 2005;24(5):437–45.
192. Hong YS, Chen Y, Zhang Y, et al. Transferability of vis-NIR models for soil organic carbon estimation between two study areas by using spiking. *Soil Sci Soc Am J.* 2018;82(5):1231–42.
193. Koehler FW, Small GW, Combs RJ, et al. Calibration transfer algorithm for automated qualitative analysis by passive fourier transform infrared spectrometry. *Anal Chem.* 2000;72(7):1690–8.

194. Koehler FW, Small GW, Combs RJ, et al. Calibration transfer in the automated detection of acetone by passive fourier transform infrared spectrometry. *Appl Spectrosc.* 2000;54(5):706–14.
195. Small GW, Harms AC, Kroutil RT, et al. Design of optimized finite impulse response digital filters for use with passive fourier transform infrared interferograms. *Anal Chem.* 1990;62(17):1768–77.
196. Song HY, Qin G. Study on the calibration transfer of near infrared spectroscopy model for soil organic matter content prediction by using FIR. *Spectrosc Spect Anal.* 2015;35(12):3360–3.
197. Wang YH, Hu WY, Song PF, et al. Model transfer between different fourier instruments and the analysis of error. *Spectrosc Spect Anal.* 2019;39(3):308–12.
198. Milanez KDTM, Nobrega TCA, Nascimento DS, et al. Selection of robust variables for transfer of classification models employing the successive projections algorithm. *Anal Chim Acta.* 2017;984(1):76–85.
199. FanS X, Li JB, Xia Y, et al. Long-term evaluation of soluble solids content of apples with biological variability by using near-infrared spectroscopy and calibration transfer method. *Postharvest Biol Technol.* 2019;151(1):79–87.
200. Zheng KY, Feng T, Zhang W, et al. Variable selection by double competitive adaptive reweighted sampling for calibration transfer of near infrared spectra. *Chemom Intell Lab Syst.* 2019;191:109–17.
201. Ni LJ, Han MY, Luan SR, et al. Screening wavelengths with consistent and stable signals to realize calibration model transfer of near infrared spectra. *Spectrochim Acta Part A Mol Biomol Spectrosc.* 2019;206(1):350–8.
202. Ni LJ, Xiao LX, Zhang LG, et al. Calibration transfer of near infrared spectral models without standards based on spectrum ratio analysis. *J Instr Anal.* 2018;37(5):539–46.
203. Zhang LG, Li YQ, Huang W, et al. The method of calibration model transfer by optimizing wavelength combinations based on consistent and stable spectral signals. *Spectrochimica Acta Part A: Mol Biomol Spectrosc.* 2020;27(1):117647.
204. Hong SJ, Huang W, Zhang LG, et al. A near infrared spectroscopy calibration model transfer method based on scale invariant feature transform to select stable characteristic wavelengths. *J Instr Anal.* 2020;38(10):1260–6.
205. Xu ZP, Fan S, Cheng WM, et al. A correlation-analysis-based wavelength selection method for calibration transfer. *Spectrochimica Acta Part A: Mol Biomol Spectrosc.* 2020;230:118053.
206. Zhang L, Small GW, Arnold MA. Calibration standardization algorithm for partial-least-squares regression: application to the determination of physiological levels of glucose by near-infrared spectroscopy. *Anal Chem.* 2002;74(16):4097–108.
207. Zhang L, Small GW, Arnold MA. Multivariate Calibration Standardization across Instruments for the Determination of Glucose by Fourier transform near-infrared spectrometry. *Anal Chem.* 2003;75(21):5905–15.
208. Stork CL, Kowalski BR. Weighting schemes for updating regression models—a theoretical approach. *Chemom Intell Lab Syst.* 1999;48(2):151–66.
209. Kalivas JH, Siano GS, Andries E, et al. Calibration maintenance and transfer using tikhonov regularization approaches. *Appl Spectrosc.* 2009;63(7):800–9.
210. Stout F, Kalivas JH. Tikhonov regularization in standardized and general form for multivariate calibration with application towards removing unwanted spectral artifacts. *J Chemom.* 2006;20(1–2):22–33.
211. Kunz MR, Kalivas JH, Andries E. Model updating for spectral calibration maintenance and transfer using 1-norm variants of Tikhonov regularization. *Anal Chem.* 2010;82(9):3642–9.
212. Kunz MR, Ottaway J, Kalivas JH, et al. Impact of standardization sample design on Tikhonov regularization variants for spectroscopic calibration maintenance and transfer. *J Chemom.* 2010;24(3–4):218–29.
213. Shahbazikhah P, Kalivas JH. A consensus modeling approach to update a spectroscopic calibration. *Chemom Intell Lab Syst.* 2013;120(1):142–53.
214. Tencate AJ, Kalivas JH, White AJ. Fusion strategies for selecting multiple tuning parameters for multivariate calibration and other penalty based processes: a model updating application for pharmaceutical analysis. *Anal Chim Acta.* 2016;921(1):28–37.

215. Farrell JA, Higgins K, Kalivas JH. Updating a near-infrared multivariate calibration model formed with lab-prepared pharmaceutical tablet types to new tablet types in full production. *J Pharm Biomed Anal.* 2012;61(1):114–21.
216. Hu Y, Li BY, Zhang J, et al. A new NIR calibration transfer method based on parameter correction. *Spectrosc Spect Anal.* 2020;40(6):1804–8.
217. Andries E, Kalivas JH, Gurung A, et al. Sample and feature augmentation strategies for calibration updating. *J Chemomet.* 2018;33(1):e3038.
218. Rudnitskaya A, Costa AMS, Delgado I. Calibration update strategies for an array of potentiometric chemical sensors. *Sens Actuat, B Chem.* 2017;238(1):1181–9.
219. Kunz MR, She YY. Multivariate calibration maintenance and transfer through robust fused LASSO. *J Chemom.* 2013;27(9):233–42.
220. Guo SX, Heinke R, Stockel S, et al. Towards an improvement of model transferability for raman spectroscopy in biological applications. *Vib Spectrosc.* 2017;91(1):111–8.
221. Zhang FY, Zhang RQ, Wang WM, et al. Ridge regression combined with model complexity analysis for near infrared (NIR) spectroscopic model updating. *Chemomet Intell Lab Syst.* 2019;195:103896.
222. Sulub Y, Small GW. Spectral simulation methodology for calibration transfer of near-infrared spectra. *Appl Spectrosc.* 2007;61(4):406–13.
223. Haaland DM. Synthetic multivariate models to accommodate unmodeled interfering spectral components during quantitative spectral analyses. *Appl Spectrosc.* 2000;54(2):246–54.
224. Sulub Y, LoBrutto R, Vivilecchia R, et al. Near-infrared multivariate calibration updating using placebo: a content uniformity determination of pharmaceutical tablets. *Vibr Spectrosc.* 2008;46(2):128–34.
225. Saiz-Abajo MJ, Mevik BH, Segtnan VH, et al. Ensemble methods and data augmentation by noise addition applied to the analysis of spectroscopic data. *Anal Chim Acta.* 2005;533(2):147–59.
226. Pierna JAF, Chauchard F, Preys S, et al. How to build a robust model against perturbation factors with only a few reference values: a chemometric challenge at ‘Chimométrie 2007.’ *Chemom Intell Lab Syst.* 2011;106(2):152–9.
227. Cooper JB, Larkin CM, Abdelkader MF. Calibration transfer of Near-IR partial least squares property models of fuels using virtual standards. *J Chemom.* 2011;25(9):496–505.
228. Cooper JB, Larkin CM, Abdelkader MF. Virtual standard slope and bias calibration transfer of partial least squares jet fuel property models to multiple near infrared spectroscopy instruments. *J Near Infrared Spectrosc.* 2011;19(2):139–50.
229. Rauscher MS, Krump M, Schardt M, et al. Multivariate calibration methods for a non-dispersive infrared sensor for engine oil condition monitoring. *Tech Mess.* 2018;85(6):395–409.
230. Abdelkader MF, Cooper JB, Larkin CM. Calibration transfer of partial least squares jet fuel property models using a segmented virtual standards slope-bias correction method. *Chemom Intell Lab Syst.* 2012;110(1):64–73.
231. Da Silva NC, Cavalcanti CJ, Honorato FA, et al. Standardization from a benchtop to a handheld NIR spectrometer using mathematically mixed NIR spectra to determine fuel quality parameters. *Anal Chim Acta.* 2017;954(1):32–42.
232. NiW D, Brown SD, Man RL. Stacked PLS for calibration transfer without standards. *J Chemom.* 2011;25(3):130–7.
233. Honorato FA, Galvao RKH, Pimentel MF, et al. Robust modeling for multivariate calibration transfer by the successive projections algorithm. *Chemom Intell Lab Syst.* 2005;76(1):65–72.
234. Igne B, Hurburgh CR Jr. Local chemometrics for samples and variables: optimizing calibration and standardization processes. *J Chemom.* 2010;24(2):75–86.
235. Liu JJ, Li BQ, Zhai HL, et al. The common quantitative model for the determination of multiple near infrared spectrometers. *Chemom Intell Lab Syst.* 2018;182(1):117–23.
236. Kramer KE, Morris RE, Rose-Pehrsson SL. Comparison of two multiplicative signal correction strategies for calibration transfer without standards. *Chemom Intell Lab Syst.* 2008;92(1):33–43.

237. Liu Z, Yu HL, MacGregor JF. Standardization of line-scan NIR imaging systems. *J Chemom.* 2007;21(3–4):88–95.
238. Sahni NS, Isaksson T, Næs T. Comparison of methods for transfer of calibration models in near-infrared spectroscopy: a case study based on correcting path length differences using fiber-optic transmittance probes in in-line near-infrared spectroscopy. *Appl Spectrosc.* 2005;59(4):487–95.
239. Guo SX, Achim K, Boris Z, et al. Extended multiplicative signal correction based model transfer for Raman spectroscopy in biological applications. *Anal Chem.* 2018;90(16):9787–95.
240. Preys S, Roger JM, Boulet JC. Robust calibration using orthogonal projection and experimental design. Application to the correction of the light scattering effect on turbid NIR spectra. *Chemomet Intell Lab Syst.* 2008;91(1):28–33.
241. Wijewardane NK, Ge Y, Morgan CLS. Prediction of soil organic and inorganic carbon at different moisture contents with air-dry ground VNIR: a comparative study of different approaches. *Eur J Soil Sci.* 2016;67(5):605–15.
242. Wijewardane NK, Ge Y, Morgan CLS. Moisture insensitive prediction of soil properties from VNIR reflectance spectra based on external parameter orthogonalization. *Geoderma.* 2016;267(1):92–101.
243. Ackerson JP, Morgan CLS, Ge Y. Penetrometer-mounted VisNIR spectroscopy: application of EPO-PLS to in situ VisNIR spectra. *Geoderma* 2017;286(1):131–8.
244. Roudier P, Hedley CB, Lobsey CR, et al. Evaluation of two methods to eliminate the effect of water from soil Vis-NIR spectra for predictions of organic carbon. *Geoderma.* 2017;296(1):98–107.
245. Amat-Tosello S, Dupuy N, Kister J. Contribution of external parameter orthogonalisation for calibration transfer in short waves-near infrared spectroscopy application to gasoline quality. *Anal Chim Acta.* 2009;642(1–2):6–11.
246. Hans G, Allison B. Temperature and moisture insensitive prediction of biomass calorific value from near infrared spectra using external parameter orthogonalization. *J Near Infrared Spectrosc.* 2019;27(4):1–11.
247. Thamasopinkul C, Ritthiruangdej P, Kasemsumran S, et al. Temperature compensation for determination of moisture and reducing sugar of Longan Honey by Near Infrared Spectroscopy. *J Near Infrared Spectrosc.* 2017;25(1):36–44.
248. Thygesen L, Lundqvist SP. NIR measurement of moisture content in wood under unstable temperature conditions. Part 2. Handling temperature fluctuations. *J Near Infrared Spectrosc.* 2000;8(1):191–9.
249. Luoma P, Natschläger T, Malli B, et al. Additive partial least squares for efficient modelling of independent variance sources demonstrated on practical case studies. *Anal Chim Acta.* 2018;1007:10–5.
250. Elizalde O, Asua JM, Leiza JR. Monitoring of emulsion polymerization reactors by raman spectroscopy: calibration model maintenance. *Appl Spectrosc.* 2005;59(10):1280–5.
251. Nouri M, Gomez C, Gorretta N, et al. Clay content mapping from airborne hyperspectral VIS-NIR data by transferring a laboratory regression model. *Geoderma.* 2017;298(1):54–66.
252. Yu BF, Ji HB. Near-infrared calibration transfer via support vector machine and transfer learning. *Anal Methods.* 2015;7(6):2714–25.
253. Brito RS, Pinheiro HM, Ferreira F, et al. Calibration transfer between a bench scanning and a submersible diode array spectrophotometer for in situ wastewater quality monitoring in Sewer systems. *Appl Spectrosc.* 2016;70(3):443–54.
254. Brouckaert D, Uyttersprot JS, Broeckx W, et al. Calibration transfer of a Raman spectroscopic quantification method for the assessment of liquid detergent compositions from at-line laboratory to in-line industrial scale. *Talanta.* 2018;179:386–92.
255. Damberg RG, Mercurio MD, Kassara S, et al. Rapid measurement of Methyl cellulose precipitable Tannins using ultraviolet spectroscopy with chemometrics: application to red wine and inter-laboratory calibration transfer. *Appl Spectrosc.* 2012;66(6):656–64.

256. Myles AJ, Zimmerman TA, Brown SD. Transfer of multivariate classification models between laboratory and process near-infrared spectrometers for the discrimination of Green Arabica and Robusta Coffee Beans. *Appl Spectrosc.* 2006;60(10):1198–203.
257. Wang J, Guo ZH, He FJ, et al. Maintenance methods of protein detection model for mutton of different areas of Ningxia based on hyperspectral. *Food Indus.* 2018;39(6):118–21.
258. Li TR, Liu CL, Wei LN, et al. Near-infrared spectral model transfer of acid value and peroxide value of edible oil by Slope Intercept Correction algorithm. *J Chin Cereals Oils Assoc.* 2018;33(1):118–24.
259. Pierna JAF, Vermeulen P, Lecler B, et al. Calibration transfer from dispersive instruments to handheld spectrometers. *Appl Spectrosc.* 2010;64(6):644–8.
260. Zamora-Rojas E, Perez-Marin D, Pedro-Sanz ED, et al. Handheld NIRS analysis for routine meat quality control: database transfer from at-line instruments. *Chemom Intell Lab Syst.* 2012;114(1):30–5.
261. Daikos O, Heymann K, Scherzer T. Development of a PLS approach for the determination of the conversion in UV-cured white-pigmented coatings by NIR chemical imaging and its transfer to other substrates. *Prog Org Coat.* 2019;132:116–24.
262. Smith MR, Jee RD, Moffat AC. Transfer between instruments of a reflectance near-infrared assay for paracetamol in intact tablets. *Analyst.* 2002;127(12):1682–92.
263. Smith MR. Calibration transfer in pharmaceutical near infrared spectroscopy. *NIR News.* 2004;15(6):13–5.
264. Hayes CJ, Walsh KB, Greensill CV. Improving calibration transfer between shortwave near infrared silicon photodiode array instruments. *J Near Infrared Spectrosc.* 2016;24(1):59–68.
265. Xu HR, Li QQ. Calibration model transfer between visible/nir spectrometers in sugar content on-line detection of crown pears. *Trans Chin Soc Agricult Mach.* 2017;9:317–22.
266. Roggo Y, Duponchel L, Noe B, et al. Sucrose content determination of sugar beets by near infrared reflectance spectroscopy. comparison of calibration methods and calibration transfer. *J Near Infrared Spectrosc.* 2002;10(1):137–50.
267. Saranwong S, Kawano S. A simple method of instrument standardisation for a near infrared sorting machine: the utilisation of average spectra as input vectors. *J Near Infrared Spectrosc.* 2004;12(1):359–65.
268. Soldado A, Fearn T, Martinez-Fernandez A, et al. The transfer of NIR calibrations for undried grass silage from the laboratory to on-site instruments: comparison of two approaches. *Talanta.* 2013;105(1):8–14.
269. Sun HX, Zhang SJ, Xue JX, et al. Model transfer method of fresh jujube soluble solids detection using variables optimization and correction algorithms. *Spectrosc Spect Anal.* 2019;39(4):1041–6.
270. Qin YH, Gong HL. NIR models for predicting total sugar in tobacco for samples with different physical states. *Infrared Phys Technol.* 2016;77:239–43.
271. Garcia-Olmo J, Garrido-Varo A, De Pedro E. The transfer of fatty acid calibration equations using four sets of unsealed liquid standardization samples. *J Near Infrared Spectrosc.* 2001;9(1):49–62.
272. De La Roza-Delgado B, Garrido-Varo A, Soldado A, et al. Matching portable NIRS instruments for in situ monitoring indicators of milk composition. *Food Control.* 2017;76(1):74–81.
273. Masahiro W, Yukihiro O. Practical calibration correction method for the maintenance of an on-line near-infrared monitoring system for Molten polymers. *Appl Spectrosc.* 2005;59(4):487–95.
274. Pérezmarín D, Garrido-Varo A, Guerreroginél J. Remote near infrared instrument cloning and transfer of calibrations to predict ingredient percentages in intact compound feedstuffs. *J Near Infrared Spectrosc.* 2006;3:81–91.
275. Milanez KDTM, Silva AC, Paz JEM, et al. Standardization of NIR data to identify adulteration in ethanol fuel. *Microchem J.* 2016;124(1):121–6.
276. Da Silva VH, Da Silva JJ, Pereira CF. Portable Near-Infrared Instruments: application for quality control of polymorphs in pharmaceutical raw materials and calibration transfer. *J Pharm Biomed Anal.* 2017;134(1):287–94.

277. Brito ALB, Santos AVP, Milanez KTM, et al. Calibration transfer of flour NIR spectra between benchtop and portable instruments. *Anal Methods*. 2017;9(21):3184–90.
278. Ji NY, Han DH. Study on near-infrared prediction model transfer for apples. *J Food Saf Qual*. 2014;5(3):712–7.
279. Hu RW, Xia JF. Transfer of NIRS calibration model for determining total sugar content in navel orange. *Food Sci*. 2012;33(3):28–32.
280. Chen YY, Qi K, Liu YL, et al. Transferability of hyperspectral model for estimating soil organic matter concerned with soil moisture. *Spectrosc Spect Anal*. 2015;35(6):1705–8.
281. Wang SF, Han P, Song HY, et al. Application of slope/bias and direct standardization algorithms to correct the effect of soil moisture for the prediction of soil organic matter content based on the near infrared spectroscopy. *Spectrosc Spect Anal*. 2019;39(6):1986–92.
282. Ji WJ, Li S, Chen SC, et al. Prediction of soil attributes using the Chinese soil spectral library and standardized spectra recorded at field conditions. *Soil Tillage Res*. 2016;155:492–500.
283. Silva NC, Pimentel MF, Honorato RS, et al. Classification of Brazilian and Foreign gasoline adulterated with alcohol using infrared spectroscopy. *Forensic Sci Int*. 2015;253(1):33–42.
284. Liu CL, Li TR, Wei LN, et al. Research on application of direct standardization algorithm in near-infrared spectrum calibration transfer of acid value and peroxide value of edible oil. *Spectrosc Spect Anal*. 2017;37(10):3042–50.
285. Lopez-Moreno C, Palanco S, Laserna JJ. Calibration transfer method for the quantitative analysis of high-temperature materials with stand-off laser-induced breakdown spectroscopy. *J Anal At Spectrom*. 2005;20:1275–9.
286. Khaydukova M, Medina-Plaza C, Rodriguez-Mendez ML, et al. Multivariate calibration transfer between two different types of multisensor systems. *Sens Actuat, B Chem*. 2017;246(1):994–1000.
287. Weng HY, Cen HY, He Y. Hyperspectral model transfer for citrus canker detection based on direct standardization algorithm. *Spectrosc Spect Anal*. 2018;38(1):235–9.
288. Fonollosa J, Fernández L, Gutiérrez-Gálvez A, et al. Calibration transfer and drift counteraction in chemical sensor arrays using direct standardization. *Sens Actuat B*. 2016;236(1):1044–53.
289. Fonollosa J, Neftci E, Huerta R, et al. Evaluation of calibration transfer strategies between metal oxide gas sensor arrays. *Procedia Eng*. 2015;120(1):261–4.
290. Panchuk V, Kirsanov D, Oleneva E, et al. Calibration transfer between different analytical methods. *Talanta*. 2017;170(8):457–63.
291. Vaughan AA, Dunn WB, Allwood JW, et al. Liquid chromatography-mass spectrometry calibration transfer and metabolomics data fusion. *Anal Chem*. 2012;84(22):9848–57.
292. De Moraes CDLM, De Lima KMG. Determination and analytical validation of creatinine content in Serum using image analysis by multivariate transfer calibration procedures. *Anal Methods*. 2015;7:6904–10.
293. Khoshkam M, Van Den Berg F, Kompany-Zareh M. Achieving bilinearity in non-bilinear augmented first order kinetic data applying calibration transfer. *Chemom Intell Lab Syst*. 2012;115(1):1–8.
294. Khoshkam M, Kompany-Zareh M. Calibration transfer in model based analysis of second order consecutive reactions. *Chemom Intell Lab Syst*. 2013;120(1):15–24.
295. Surkova A, Bogomolov A, Legin A, et al. Calibration transfer for LED-based optical multisensor systems. *ACS Sens*. 2020;5:2587–95.
296. Barreiro P, Herrero D, Hernández N, et al. Calibration transfer between portable and laboratory NIR spectrophotometers. *Acta Hort*. 2008;802:373–8.
297. Alamar MC, Bobelyn E, Lammertyn J, et al. Calibration transfer between NIR diode array and FT-NIR spectrophotometers for measuring the soluble solids contents of Apple. *Postharvest Biol Technol*. 2007;45(1):38–45.
298. Sulub Y, Lobrutto R, Vivilecchia R, et al. Content uniformity determination of pharmaceutical tablets using five near-infrared reflectance spectrometers: a process analytical technology (PAT) approach using robust multivariate calibration transfer algorithms. *Anal Chim Acta*. 2008;611(2):143–50.

299. Luo X, Ikehata A, Sashida K, et al. Transfer of calibration model between near-infrared spectrometers for hematocrit measurement of grazing cattle. *NIR News*. 2017;28(7):16–21.
300. Luo X, Ikehata A, Sashida K, et al. Calibration transfer across near infrared spectrometers for measuring hematocrit in the blood of grazing cattle. *J Near Infrared Spectrosc*. 2017;25(1):15–25.
301. Cen HY, Weng HY, He Y. A method for delivering hyperspectral models of citrus canker: ZL 201610260903.2 [P], 21 Sept 2016.
302. Pereira LSA, Carneiro MF, Botelho BG, et al. Calibration transfer from powder mixtures to intact tablets: a new use in pharmaceutical analysis for a known tool. *Talanta*. 2016;147(1):351–7.
303. Sohn M, Barton FE, Himmelsbach DS. Transfer of near-infrared calibration model for determining fiber content in flax: effects of transfer samples and standardization procedure. *Appl Spectrosc*. 2007;61(4):414–8.
304. Galvan D, Bona E, Borsato D, et al. Calibration transfer of partial least squares regression models between desktop nuclear magnetic resonance spectrometers. *Anal Chem*. 2020;92:12809–16.
305. Yang Y, Peng YK, Li YY, et al. Calibration transfer of surface-enhanced raman spectroscopy quantitative prediction model of potassium sorbate in Osmanthus wine to other wine. *Spectrosc Spect Anal*. 2018;38(3):824–9.
306. Liu J, Li XY, Guo XX, et al. Transfer method among water content detection models for different breeds of pork by hyperspectral imaging technique. *Trans Chin Soc Agricult Eng*. 2014;30(17):276–84.
307. Liu J, Li XY, Jin R, et al. Extending hyperspectral detecting model of pH in fresh pork to new breeds. *Spectrosc Spect Anal*. 2015;35(7):1973–9.
308. Dong XG, Dong J, Li YL, et al. Maintaining the predictive abilities of egg freshness models on new variety based on VIS-NIR spectroscopy technique. *Comput Electron Agric*. 2019;156:669–76.
309. Boiret M, Meunier L, Ginot YM. Tablet potency of tianeptine in coated tablets by near infrared spectroscopy: model optimisation, Calibration transfer and confidence intervals. *J Pharm Biomed Anal*. 2011;54(1):510–6.
310. Sales F, Callao MP, Rius FX. Multivariate standardization for correcting the ionic strength variation on potentiometric sensor arrays. *Analyst*. 2000;125(5):883–8.
311. Marchesini G, Serva L, Garbin E, et al. Near-infrared calibration transfer for undried whole maize plant between laboratory and on-site spectrometers. *Ital J Anim Sci*. 2017;17(1):66–72.
312. Ge YF, Morgan CLS, Grunwald S, et al. Comparison of soil reflectance spectra and calibration models obtained using multiple spectrometers. *Geoderma*. 2011;161(3–4):202–11.
313. Rodrigues RRT, Rocha JTC, Oliveira LMSL, et al. Evaluation of calibration transfer methods using the Atr-Ftir technique to predict density of crude oil. *Chemom Intell Lab Syst*. 2017;166(1):7–13.
314. Li X, Arzhantsev S, Kauffman JF, et al. Detection of diethylene glycol adulteration in propylene glycol-method validation through a multi-instrument collaborative study. *J Pharm Biomed Anal*. 2011;54(5):1001–6.
315. Gryniewicz-Ruzicka CM, Arzhantsev S, Pelster LN, et al. Multivariate calibration and instrument standardization for the rapid detection of diethylene glycol in glycerin by raman spectroscopy. *Appl Spectrosc*. 2011;65(3):334–41.
316. Thygesen J, Van Den Berg FWJ. Calibration transfer for excitation-emission fluorescence measurements. *Anal Chim Acta*. 2011;705(1–2):81–7.
317. Sanllorente S, Rubio L, Ortiz MC, et al. Signal transfer with excitation-emission matrices between a portable fluorimeter based on light-emitting diodes and a master fluorimeter. *Sens Actuat, B Chem*. 2019;285:240–7.
318. Sun XD, Wu HL, Chen Y, et al. Chemometrics-assisted calibration transfer strategy for determination of three agrochemicals in environmental samples: solving signal variation and maintaining second-order advantage. *Chemomet Intell Lab Syst*. 2019;194:103869.

319. Wang M, Zheng KY, Yang GJ, et al. A robust near-infrared calibration model for the determination of chlorophyll concentration in tree leaves with a calibration transfer method. *Anal Lett.* 2015;48(11):1707–19.
320. Watari M, Ozaki Y. Prediction of ethylene content in melt-state random and block polypropylene by near-infrared spectroscopy and chemometrics: comparison of a new calibration transfer method with a Slope/Bias correction method. *Appl Spectrosc.* 2004;58(10):1210–8.
321. Li YQ, Hong SJ, Huang W, et al. Effect of number of latent variables for partial least square model based on near infrared spectroscopy on models transfer performance. *J Instr Anal.* 2020;38(10):1231–8.
322. Sun ZY, Wang JY, Nie L, et al. Calibration transfer of near infrared spectrometers for the assessment of plasma ethanol precipitation process. *Chemom Intell Lab Syst.* 2018;181(1):64–71.
323. Xiao H, Sun K, Sun Y, et al. Comparison of benchtop fourier-transform (FT) and portable grating scanning spectrometers for determination of total soluble solid contents in single grape berry (*Vitis vinifera* L.) and calibration transfer. *Sensors* 2017;17(11):2693.
324. Fernandez L, Guney S, Gutierrez-Galvez A, et al. Calibration transfer in temperature modulated gas sensor arrays. *Sens Actuat, B Chem.* 2016;231(1):276–84.
325. Hoffmann U, Pfeifer F, Hsuing C, et al. Spectra transfer between a fouriertransform near-infrared laboratory and a miniaturized handheld near-infrared spectrometer. *Appl Spectrosc.* 2016;70(5):852–60.
326. Di Anibal CV, Ruisánchez I, Fernández M, et al. Standardization of UV-visible data in a food adulteration classification problem. *Food Chem.* 2012;134(4):2326–31.
327. Zheng YH, Song T, Zhang S, et al. Spectral transfer of near-infrared spectrometric model for fish meal. *J Instr Anal.* 2020;38(11):1378–84.
328. Pu YY, Sun DW, Riccioli C, et al. Calibration transfer from micro NIR spectrometer to hyperspectral imaging: a case study on predicting soluble solids content of bananito fruit (*Musa Acuminata*). *Food Anal Methods.* 2017;11(4):1021–33.
329. Xi CC, Feng YC, Hu CQ. Evaluation of piecewise direct standardization algorithm for near infrared quantitative model updating. *Chin J Anal Chem.* 2014;42(9):1307–13.
330. Gislason J, Chan H, Sardashti M. Calibration transfer of chemometric models based on process nuclear magnetic resonance spectroscopy. *Appl Spectrosc.* 2001;55(11):1553–60.
331. Monakhova YB, Diehl BWK. Transfer of multivariate regression models between high-resolution NMR instruments: application to authenticity control of sunflower lecithin. *Magn Reson Chem.* 2016;54(9):712–7.
332. Chen CS, Brown CW, Lo SC. Calibration transfer from sample cell to fiber-optic probe. *Appl Spectrosc.* 1997;51:744–8.
333. Lin J, Lo SC, Brown CW. Calibration transfer from a scanning near-ir spectrophotometer to a FT-Near-IR spectrophotometer. *Anal Chim Acta.* 1997;349(1–3):263–9.
334. Shi GT, Han LJ, Yan ZL, et al. Near infrared calibration transfer for quantitative analysis of fish meal mixed with Soybean meal. *J Near Infrared Spectrosc.* 2013;18(3):509–22.
335. Tortajada-Genaro LA, Campíns-Falcó P, Bosch-Reig F. Calibration transfer in chemiluminescence analysis: application to chromium determination by luminol-hydrogen peroxide reaction. *Anal Chim Acta.* 2001;446(1):383–90.
336. Griffiths ML, Svozil D, Worsfold P, et al. The application of piecewise direct standardisation with variable selection to the correction of drift in inductively coupled atomic emission spectrometry. *J Anal At Spectrom.* 2006;21(10):1045–52.
337. Wang WH, Huck CW, Yang B. NIR model transfer of alkali-soluble polysaccharides in poria cocos with piecewise direct standardization. *NIR News.* 2019;30(5–6):6–14.
338. Morais CLM, Paraskevaidi M, CuiL, et al. Standardization of complex biologically derived spectrochemical datasets. *Nat Protocols* 2019;14(5):1546–77.
339. Grelet C, Fernández Pierna JA, Dardenne P, et al. Standardization of milk mid-infrared spectra from a European dairy network. *J Dairy Sci.* 2015;98(4):2150–60.
340. Grelet C, Fernández Pierna JA, Dardenne P, et al. Standardization of milk mid-infrared spectrometers for the transfer and use of multiple models. *J Dairy Sci.* 2017;100(10):7910–21.

341. Ji W, ViscarraRosell RA, Shi Z. Improved estimates of organic carbon using proximally sensed Vis-NIR spectra corrected by piecewise direct standardization. *Eur J Soil Sci.* 2015;66(4):670–8.
342. Pierna JAF, Sanfeliu AB, Slowikowski B, et al. Standardization of NIR microscopy spectra obtained from inter-laboratory studies by using a standardization cell. *Biotechnol Agron Soc Environ.* 2013;17(4):547–55.
343. Zheng KY, Xiang CL, Cao P, et al. Correcting NIR spectra of dimethyl fumarate in milk measured for different brands and in different dates. *Eur Food Res Technol.* 2013;237(5):787–94.
344. Liu YY, Xiong ZX, Wang Y, et al. Study on the transform of near-infrared calibration models for lignin determination between different types of portable near-infrared spectrometers. *J Forest Eng.* 2019;4(4):93–8.
345. Liu YY, Yang H, Xiong ZX, et al. Study on near-infrared calibration model transfer for lignin content in pulpwood. *Trans China Pulp Pap.* 2019;34(3):43–9.
346. Yang JX, Lou XP, Yang H, et al. Improved calibration transfer between near-infrared (NIR) spectrometers using canonical correlation analysis. *Anal Lett.* 2019;52(14):2188–202.
347. Luo J, Nie MF, Wu SH, et al. Research on the transfer and sharing of fast and nondestructive calibration model for textiles. *China Fib Inspect.* 2016;10(1):79–81.
348. Fan PP, Li XY, Lv MR, et al. Vis-NIR model transfer of total nitrogen between different soils. *Spectrosc Spect Anal.* 2018;38(10):3210–4.
349. Eliaerts J, Meert N, Dardenne P, et al. Evaluation of a calibration transfer between a benchtop and portable mid-infrared spectrometer for cocaine classification and quantification. *Talanta* 2020;209:120481.
350. Yang ZL, Yang QK, Shen GH, et al. Online application of Soybean meal NIRS quantitative analysis model from laboratory to factory. *Trans Chin Soc Agric Mach.* 2019;50(8):358–65.
351. Chen YY, Wang ZB. Cross components calibration transfer of NIR spectroscopy model through PCA and weighted ELM-based tradaboost algorithm. *Chemomet Intell Lab Syst.* 2019;192:103824.
352. Ni LJ, XiaoL X, Yao HM, et al. Construction of global and robust near-infrared calibration models based on hybrid calibration sets using partial least squares (PLS) regression. *Anal Lett.* 2019;52(7):1177–94.
353. Pereira CF, Pimentel MF, Galvao RKH, et al. A comparative study of calibration transfer methods for determination of gasoline quality parameters in three different near infrared spectrometers. *Anal Chim Acta.* 2008;611(1):41–7.
354. Fernandez-Ahumada E, Garrido-Varo A, Guerrero JE, et al. Taking NIR calibrations of feed compounds from the laboratory to the process: calibration transfer between predispersive and postdispersive instruments. *J Agric Food Chem.* 2008;56(21):10135–41.
355. Debus B, Takahama S, Weakley AT, et al. Long-term strategy for assessing carbonaceous particulate matter concentrations from multiple fourier transform infrared (FT-IR) instruments: influence of spectral dissimilarities on multivariate calibration performance. *Appl Spectrosc.* 2019;73(3):271–83.
356. Krapf LC, Nast D, Gronauer A, et al. Transfer of a Near infrared spectroscopy laboratory application to an online process analyser for in situ monitoring of anaerobic digestion. *Biores Technol.* 2013;129(1):39–50.
357. Li X, Bin J, Fan W, et al. Near infrared spectral hybrid model quantitative analysis on samples with different physical states. *Chin J Anal Chem.* 45(7):958–64.
358. Clavaud M, Roggo Y, Degardin K, et al. Global regression model for moisture content determination using near-infrared spectroscopy. *Eur J Pharm Biopharm.* 2017;119(1):343–52.
359. Ozdemir D, Williams R. Multi-instrument calibration with genetic regression in UV-visible spectroscopy. *Appl Spectrosc.* 1999;53(2):210–7.
360. Kupyna A, Rukke EO, Schüller RB, et al. The effect of flow rate in acoustic chemometrics on liquid flow: transfer of calibration models. *Chemom Intell Lab Syst.* 2010;100(2):110–7.
361. Igne B, Hurburgh CR. Standardisation of near infrared spectrometers: evaluation of some common techniques for intra- and inter-brand calibration transfer. *J Near Infrared Spectrosc.* 2008;16(6):539–50.

362. Fontaine J, Hörr J, Schirmer B. Amino acid contents in raw materials can be precisely analyzed in a global network of near-infrared spectrometers: collaborative trials prove the positive effects of instrument standardization and repeatability files. *J Agric Food Chem.* 2004;52(4):701–8.
363. Steinbach D, Anderson CA, McGeorge G, et al. Calibration transfer of a quantitative transmission Raman PLS model: direct transfer vs. global modeling. *J Pharm Innov.* 2017;12(4):347–56.
364. Xu Z P, Fan S, Liu J, et al. A calibration transfer optimized single kernel near-infrared spectroscopic method. *Spectrochimica Acta Part A: Mol Biomol Spectrosc.* 2019;220:117098.
365. Wu JZ, Li J, Du W, et al. Establishment and application of NIR models for online cut tobacco quality monitoring in primary processing. *Tobacco Sci Technol.* 2017;50(10):69–73.
366. Liu CL, Liu HY, Sun XR, et al. Transfer of near-infrared spectroscopy model of edible oil acid value and peroxidation value. *Trans Chin Soc Agricult Mach.* 2020;51(9):344–9.
367. Xu ZP. Study on calibration transfer methods of rapid and nondestructive near infrared detection for individual crop kernels. Beijing: University of Science and Technology of China; 2020.
368. Rehman TU, Zhang LB, Ma DD, et al. Calibration transfer across multiple hyperspectral imaging-based plant phenotyping systems: I—Spectral space adjustment. *Comput Electron Agric.* 2020;176:105685.
369. Zhou SL, Zhu SP, Wei X. Improving the transfer ability of calibration model for terahertz spectroscopy. *Spectrosc Lett.* 2020;53(6):448–57.
370. Yang H, Xiong ZX, Chen T. Study on near-infrared calibration model transfer for soluble solid content in apple. *Chin J Anal Lab.* 2018;37(2):163–7.
371. Salguero-Chaparro L, Palagos B, Pena-Rodríguez F, et al. Calibration transfer of intact Olive NIR spectra between a pre-dispersive instrument and a portable spectrometer. *Comput Electron Agric.* 2013;96:202–8.
372. Liu X, HanL J, Yang ZL. Transfer of near infrared spectrometric models for Silage crude protein detection between different instruments. *J Dairy Sci.* 2011;94(11):5599–610.
373. Liu X, Huang CJ, Han LJ. Calibration transfer of near-infrared spectrometric model for calorific value prediction of straw using different scanning temperatures and accessories. *Energy Fuels.* 2015;29(10):6450–5.
374. Bergman EL, Brage H, Josefson M, et al. Transfer of NIR calibrations for pharmaceutical formulations between different instruments. *J Pharm Biomed Anal.* 2006;41(1):89–98.
375. Li XY, Liu Y, Lv MR, et al. Calibration transfer of soil total carbon and total nitrogen between two different types of soils based on visible-near-infrared reflectance spectroscopy. *J Spectrosc.* 2018;1–10.
376. Greensill CV, Walsh KB. Calibration transfer between miniature photodiode array-based spectrometers in the near infrared assessment of Mandarin soluble solids content. *J Near Infrared Spectrosc.* 2002;10(1):27–35.
377. Walczak B, Bouveresse E, Massart DL. Standardization of near-infrared spectra in the wavelet domain. *Chemometr Intell Lab Syst.* 1997;36(1):41–51.
378. Martins MN, Galvao RKH, Pimentel MH. Multivariate calibration transfer employing variable selection and subgating. *J Braz Chem Soc.* 2010;21(1):127–34.
379. Yoon J, Chung H, Han C. Calibration transfer algorithm for NIR spectroscopy as an online analyzer. *IFAC Proc Vol.* 2001;34(27):303–8.
380. Yahaya OKM, MatJafri MZ, Aziz AA, et al. Visible spectroscopy calibration transfer model in determining pH of Sala Mangoes. *J Instrum.* 2015;10(5):T05002.

Chapter 18

Deep Learning Methods



Deep learning (DL) is a specific type of machine learning. It has strong ability and flexibility by learning features from complicated data. The core of DL is feature learning, starting from the original input data, the features of each layer are transformed into a higher-level and more abstract representation layer by layer. Useful information in data is extracted during classification and prediction process. It has the potential ability of automatic learning features. The term “deep” usually refers to hidden layers in neural networks. The network will be deeper with more layers. Traditional neural networks only contain two or three layers, while deep networks may contain dozens or even hundreds of hidden layers.

For the traditional neural network, it is difficult to train the entire network if the number of hidden layers in the network is simply increased. Back Propagation (BP) algorithm, which adjusts weights reversely using gradient descent algorithm according to the output error, plays an extremely crucial role in artificial neural network (ANN). However, in the back propagation of BP algorithm, the gradient becomes more and more diffused with the increase of hidden layer numbers, which leads to the relatively small weights closing to the input layer. Thus, only the weights close to the output layer play the real decision role, which leads to over-fitting of the model. This is often referred to “gradient dispersion (GD)”. Two strategies are usually used to address this problem, i.e., improving the training mechanism and improving the network structure. From the above two strategies, two typical DL models, auto-encoder (AE) and convolution neural networks (CNN), are developed. AE adopts a layer-by-layer pre-training method to alleviate the problems of gradient dispersion and local minima. CNN introduces the concepts of “weight sharing” and “local connection” from the structure to effectively reduce the parameter space and model training difficulty.

DL is different from traditional shallow network learning in two aspects. First, DL emphasizes the depth of the model structure, which usually has many hidden layer nodes. Second, the importance of feature learning is clearly highlighted. It transforms feature representation of samples in the original space into a new feature space by layer-by-layer feature transformation, which makes classification or regression

easier. The essence of DL is to learn more useful features by constructing machine learning models with many hidden layers and massive training data, so as to improve the accuracy of classification or regression. DL is a framework that contains several important algorithms, such as AE, CNN, restricted Boltzmann machine (RBM), deep belief network (DBN), etc. [1]. This chapter mainly introduces AE and CNN and their applications in spectral classification and regression.

18.1 Stacked Auto-encoder

Auto-encoder (AE) is an algorithm for data compression. As shown in Fig. 18.1, AE is composed of two parts as encoder and decoder. The former encodes the input signal to obtain the encoded signal, and the latter decodes the encoded signal to obtain the output signal. AE belongs to self-supervised learning algorithm, which realizes the reproduction of input data by expecting output equal to input. For the AE, what is concerned is the representation after coding, i.e., mapping from the input layer to the coding layer, which carries the main drivers and implicit relationships in the original information.

It is called auto-encoder network (DAE) when the compression and decompression of AE are realized by neural network. DAE is an unsupervised learning algorithm, which can easily code richer and higher-order network structures. As shown in Fig. 18.2, in the DAE, the input of the hidden layer is the encoding of the input layer. In fact, the output of the upper layer is nonlinear transformation, named “non-linear mapping”. The output of the hidden layer is actually the feature representation learned after mapping the input, reflecting the implicit correlation in the input. In the auto-encoder neural network, AE uses the encoding and decoding operations to achieve the reconstruction of the original information. The process of AE information reconstruction seems meaningless. In fact, a set of basis vectors can be obtained by hiding the sparse limitation of neuron links, and the intrinsic structure of input vectors can be reflected by the set of basis vectors.

The number of neurons in the auto-encoder hidden layer can be more than that in the input layer. In order to achieve the effective compression of the input variables, the

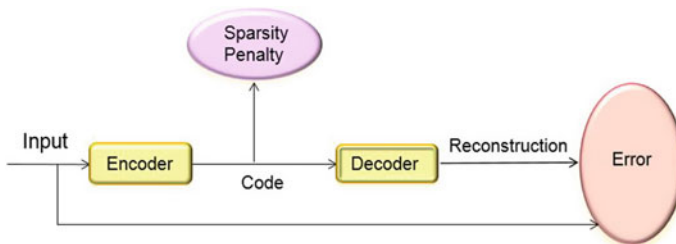


Fig. 18.1 Schematic diagram of stacked auto-encoder

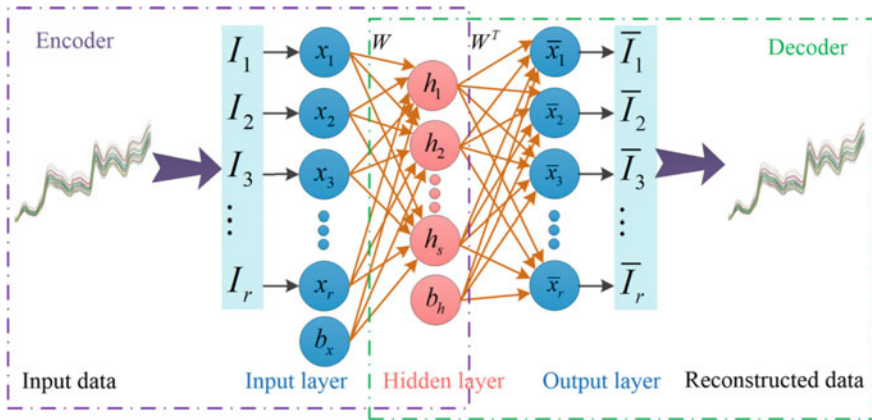


Fig. 18.2 Schematic diagram of self-encoder neural network structure

L1 regularization constraint term is added to the construction of the loss function of the neural network for sparse constraint, which evolves into a sparse auto-encoder. It is obtained by adding some sparsity constraints on the basis of the traditional auto-encoder (including two parameters of the regularization term and the weight coefficient of the sparse penalty term). This sparsely is aimed at the hidden layer neurons of AE. Suppressing most of the output of the hidden layer neurons (i.e., most of the nodes in the constraint hidden layer are 0, only a few are not 0), makes the network sparse.

In addition to sparse auto-encoder, there is also de-noising auto-encoder. Its main improvement is to add random noise to the spectra of training samples, and the target of reconstruction is the spectrum without noise. That is, the data reconstructed from the model learned by the AE can remove the noise, indicating that the AE can learn the characteristics from the noisy data.

Stacked auto-encoder (SAE) is an unsupervised learning network composed of multiple automatic encoders stacked layer by layer, and it is one of the DL networks. Compared with the shallow neural network, the expression of data features is more powerful, and has various advantages of traditional neural network. The programming method of stacked auto-encoder neural network is to perform the AE of each layer in order from the beginning to the end, and the output of the former AE is the output of the latter AE. Similarly, the decoding process of SAE is to perform each automatic encoder in reverse order. The training mode of SAE is unsupervised greedy training, and only one hidden layer is trained each time. This layer begins to train the next layer after the encoder is optimized until the last hidden layer is trained. Finally, the weight and deviation of each layer of the parameters are fine-tuned. Fine-tuning is that the parameters in the model are corrected by error back propagation, which is suitable for any multi-layer stack DAE.

Stacked sparse auto-encoder (SSAE) is stacked by multiple sparse auto-encoder networks. The feature expression of the original data obtained by learning is more abstract with the increase of the number of sparse AE layers.

Auto-encoder neural network is a kind of unsupervised learning field. It can automatically learn the corresponding features from unlabeled data. It is a neural network with the goal of reconstructing input signal. It can reconstruct better data features than the raw data to describe the categories represented by the raw data. The ability of learning features is strong. In DL, the data features generated by auto-encoder neural network training are often used to replace raw data in order to have better results in the subsequent regression, recognition, and classification.

Zhang et al. [2] combined stack auto-encoder with extreme learning machine (ELM) to identify cefixime tablets produced by different manufacturers by near infrared (NIR) spectroscopy, which has high classification accuracy and stability. Lu et al. fused stacked demising auto-encoders (SDAE) with random forest (RF) for NIR spectroscopy detection of Citrus yellow shoot. Firstly, SDAE was used to extract the deep features form NIR of Citrus, and then RF voting ensemble strategy was used to realize classification and identification. This method has an excellent performance in calibration training time, accuracy and stability [3]. Liu et al. [4] applied the five-layer DAE for NIR spectral feature extraction of tobacco samples, which reduced the 2760-dimensional spectra to 3-dimensional spectra. The classification effect of DAE on tobacco was significantly better than that of PCA method (Fig. 18.3). Combining stack auto-encoder with Softmax classifier, Hang et al. [5] established a method for nondestructive identification of radish seed varieties by visible-near-infrared (Vis-NIR) spectroscopy.

Wang et al. [6] used stack noise reduction auto-encoding to carry out further features from the NIR spectra of ethanol solid-state fermentation process. PLS algorithm was used as the final regressor of the depth framework to establish a model for predicting the content of alcohol and glucose in fermentation substrate, which improved the prediction accuracy of the model. Yu et al. [7] combined SAE with full-connected neural network (FNN) to predict the hardness and soluble solid content of Korla fragrant pear by Vis-NIR spectroscopy imaging. As shown in Fig. 18.4, the

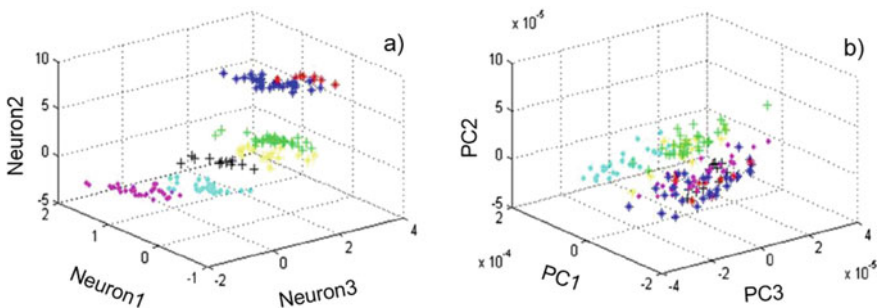
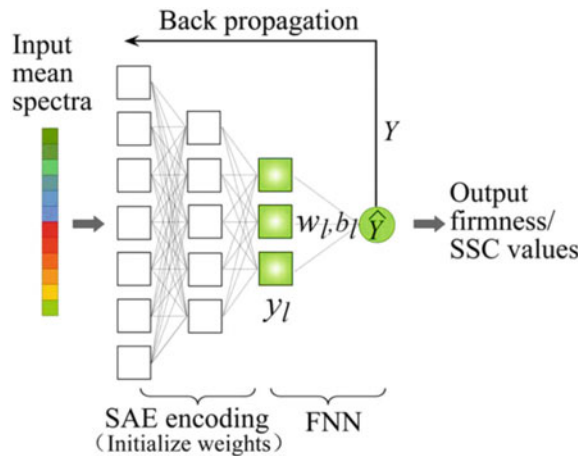


Fig. 18.3 Three-dimensional neurons and PCs of near infrared spectral feature extraction by DAE (a) and PCA (b) for tobacco samples

Fig. 18.4 Schematic diagram of SAE-FNN network topology structure



SAE network is pre-trained first, and the output is used as the initial input value of FNN. Then, the weight of the whole SAE-FNN is fine-tuned through back propagation, and the final prediction model is obtained. Yu et al. [8, 9] also used similar methods to predict nitrogen content in rape leaves and TVB-N content in *Penaeus vannamei*. Ran et al. [10] used SAE to extract the features of Vis-NIR spectra of soil, and constructed a prediction model of soil organic matter (SOM) combined with back propagation artificial neural network (BP-ANN). The results showed that the feature extraction effect of SAE was better than that of successive projections algorithm (SPA) and principal component analysis (PCA). Ni et al. [11] used the variable weighted stacked auto-encoder to extract the NIR spectral features of masson pine seedling roots, and then combined with support vector regression (SVR) to establish the NIR spectral prediction model of water content in masson's pine seedling root, and the prediction results were more accurate than PLS and SVR methods.

18.2 Convolution Neural Network

18.2.1 Basic Structure of CNN

Convolution neural network (CNN) is a multi-layer feed forward neural network, which is one of the mainstream DL algorithms. It is composed of multi-layer neurons such as the input layer, the hidden layer and the output layer. As shown in Fig. 18.5, the hidden layer is usually composed of alternating convolution layers, activation functions, pooling layers, and fully connected layers. According to actual needs, batch normalization and dropout can also be added to optimize the model. In order to prevent the model from over-fitting of the training set, regularization items can be added to the model. Conventional regularization operations include L1 norm and

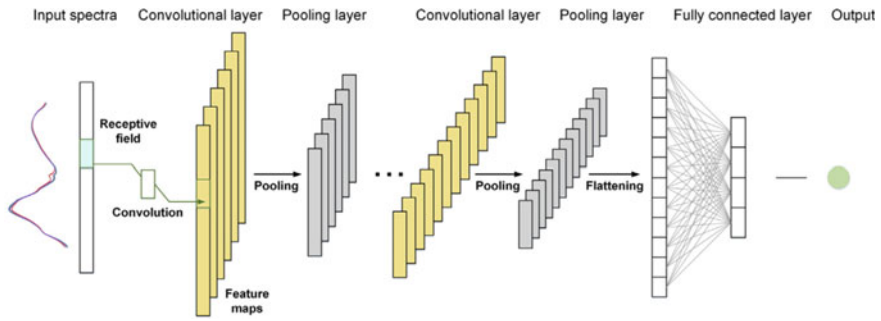


Fig. 18.5 Topological structure of spectral analysis model for CNN

L2 norm. The convolution layer can extract a variety of local features in the input information layer by layer. The pooling layer merges multiple adjacent feature points to simplify the data volume, improving the computational efficiency and robustness. The full connection layer can complete nonlinear regression or classification tasks. The activation function in DL is used to add nonlinear factors to the model to improve the ability of the model to express more advanced features.

(1) Convolution layer: Convolution layer is the core algorithm module of CNN, which is usually located behind the input layer and before the pooling operation layer; it is the most important part of CNN. The convolution layer is composed of a set of filters with parameters that can be trained. These filters usually have small perceptual regions which are also called convolution kernels. In the process of network forward propagation, each convolution kernel will slide in a certain direction on the input data, and perform convolution operation on the covered area. The values (weights) in the convolution kernel are initially set randomly. The essence of convolution operation is the weighted sum of the values in the convolution kernel and the local receptive field. After multiple operations, the parameters in the convolution kernel are continuously optimized and updated, and finally tend to converge. The convolution layer is mainly used to extract features and mine useful information. The convolution operation can extract the local relationship between adjacent pixels, and has certain robustness to translation, rotation, and scale transformation on the image.

Figure 18.6 is a simple example of convolution operation. \mathbf{I} denotes the original image, \mathbf{K} is a 3×3 convolution kernel, $*$ denotes convolution operation, and the sliding step length (stride) of the convolution kernel is 1. The feature map is obtained by convolution operation. The convolution result of each convolution kernel on the original data form a feature map with a specific meaning, corresponding to a certain type of features in the original data. The more convolution layers are, the more overall and representative feature data can be extracted. Figure 18.7 shows the process of convolution operation of one-dimensional spectral data.

The larger the kernel size, the less features are extracted by the convolution kernel, because the wider the convolution kernel window, the fewer times the convolution kernel moves across the spectral interval. Conversely, the smaller the size of the

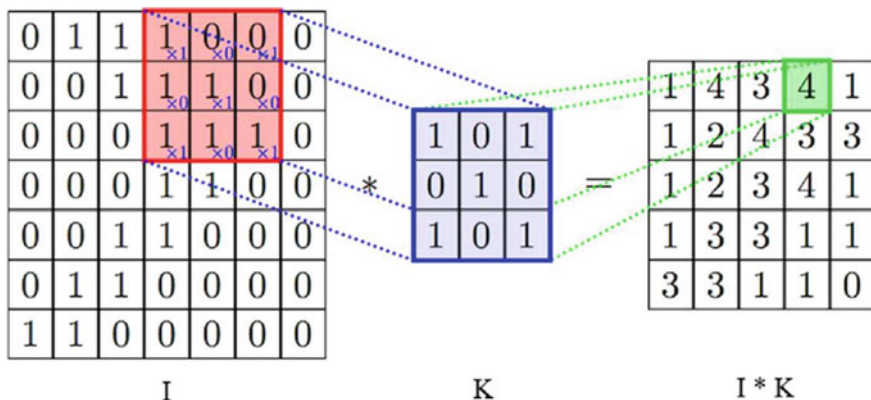


Fig. 18.6 Process of convolution calculation for two-dimensional data

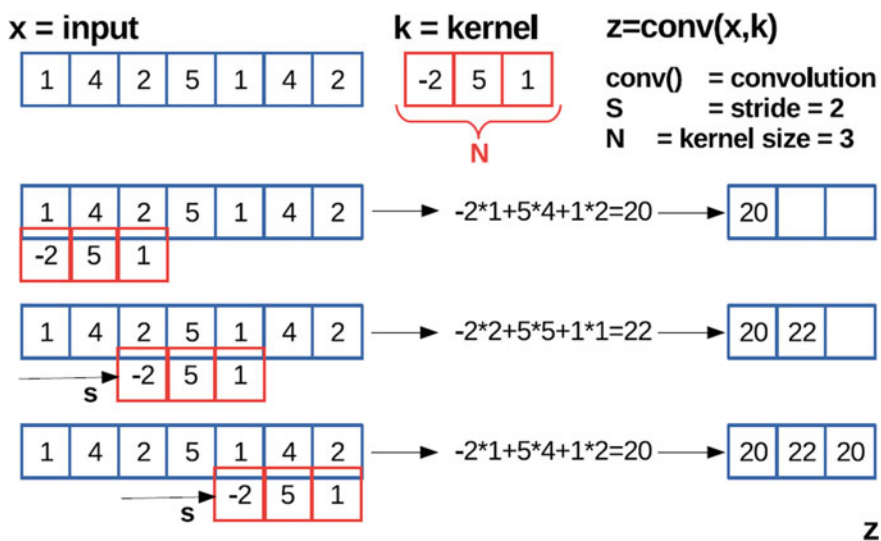


Fig. 18.7 Process of convolution calculation for one-dimensional spectral data

convolution, the more features will be extracted by the convolution kernel. The greater the step length (Stride) of the convolution kernel moves, the less features the convolution kernel extracts. On the contrary, if the moving step of convolution kernel is reduced, the features extracted by convolution kernel will increase. Usually, in the same convolution layer, there will be multiple different convolution kernels. Each convolution kernel will extract features of interest from a specific perspective.

(2) Pooling layer: The pooling layer is usually located behind the convolution layer, and its function is to sample the feature map generated by the convolution layer

operation, so it can also be called the lower sampling layer. The operation of pooling layer does not reduce the number of feature maps, but reduces the dimension of each feature map and the amount of data, improves the operation speed, and enhances the robustness of neural network model. The common pooling sampling methods are max-pooling and average-pooling. In the sampling window, the maximum value of all values is extracted as the eigenvalue by maximum pooling, and the average value of all values is calculated as the eigenvalue by mean pooling. The size of the sampling window area and the moving step size can be adjusted according to the actual applications. Figure 18.8 is an example of pooling operation. The sampling window size is 2×2 and the sliding window step size is 2. The pooling layer mainly compresses and simplifies the results after convolution, and expands the perception field and simplifies the complexity of network computing by reducing the dimension of feature expression.

(3) Activation layer: The main function of activation function in neural network is to provide nonlinear modeling ability of the network. Assuming that there is no active layer before the convolution layer and the full connection layer in the neural network, the network can only express the linear mapping between input and output. Even by increasing the depth of the network, it is still linear mapping, and it is difficult to express the nonlinear relationship between input and output. Therefore, the activation layer is often added to the DL network to make the network have hierarchical nonlinear mapping learning ability.

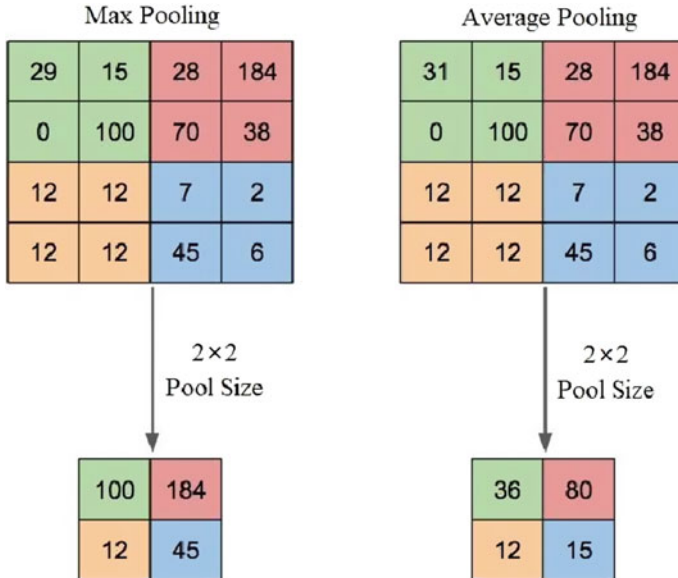


Fig. 18.8 Schematic diagram of max-pooling and average-pooling process

(4) Flatten layer: The Flatten layer is used to “flattening” the input data, that is, make the multidimensional input to one-dimensional data, which commonly used in the transition from convolution to dense. For example, M spectral matrix \mathbf{X} with N wavelength points is convoluted by K convolution kernels with S size, and the dimension of output Z is $M \times (N - S + 1) \times K$. The data Z after the convolution layer cannot be directly connected to the full connection layer, and the full connection layer needs to be connected after flattening. The data after flattening is two-dimensional data, and the dimension is $M \times ((N - S + 1) \times K)$. The function of flattening is equivalent to extending the features extracted by different convolution kernels for the next layer of calculation. Figure 18.9 shows the schematic diagram of flattening 3D data into 1D data.

(5) Full connection layer: After the operation of convolution layer and pooling layer, the neural network extracts local and global features that cannot be directly obtained at first in the input data. Full connection layer, also known as dense layer, is composed of one or more layers of neurons, such as BP network or Softmax network, where neurons are usually connected with all neurons in the adjacent layer. Its role is to receive the output results of convolution and pooling layer, classify or regress local and global features, which plays the role of “classifier” or “regressor”.

As long as the network structure design is reasonable and the training data is sufficient and effective, the ideal network model can be obtained. The trained network can learn the features in the training data, abstract, and filter the features automatically, so as to obtain the ideal feature extraction model. Compared with the artificial feature extraction, the CNN eliminates the subjective factors of human beings and makes the feature extraction more accurate and reasonable.

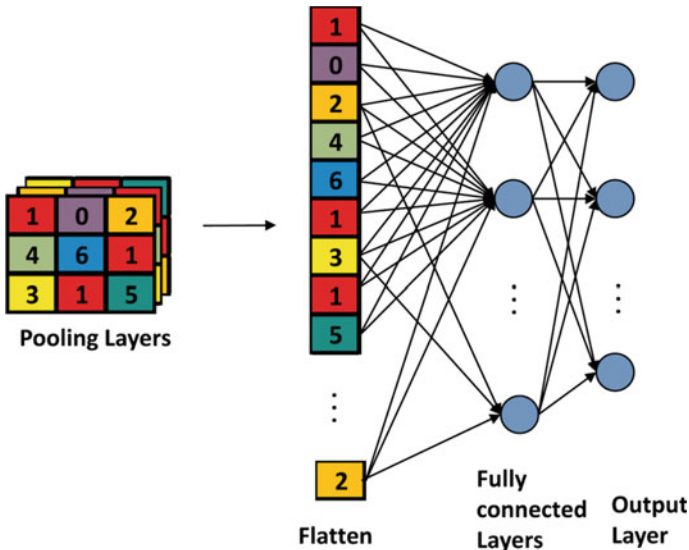


Fig. 18.9 Schematic diagram of flatten layer process

Traditional ANNs are usually fully connected neural networks, that is, all neurons between adjacent neural network layers are connected to each other. The connections among multiple levels of the fully connected neural network are numerous and complicated, which determines that it needs to train a large number of parameters, and it is easy to produce gradient diffusion and dimension disaster during training. CNN uses local connection and weight sharing to alleviate this problem. Local connection and weight sharing are the two most important features of CNN so that CNN can deal with more complex problems. Compared with the traditional full-connected neural network, its training and learning efficiency is higher.

(1) Local connection: Local connection is a typical structural feature of CNN. When the network level increases, local connection can greatly reduce the number of connections among network levels and reduce the complexity of network structure. It refers to that when the CNN learns large target data; it follows the cognitive process from the local to the whole. First, it establishes local small-scale connection, and then gradually enhances the understanding of the whole data from the training process. Figure 18.10a is the connection structure of CNN. The neurons in the convolution layer are only connected with some data in the region of interest in the previous layer, rather than to all input data. This local connection makes the neurons in the hidden layer only perform convolution operations on their local connection regions, without calculating other regions. Through local connection, the convolution kernel can fully extract the local features of the data, and the features of each local region are characterized as an element in the output feature graph.

Multiple convolution kernels can be used to extract a variety of local features (i.e., multi-core convolution) and output multiple feature maps. These local feature maps will be perceived by neurons at higher level of the network to extract global features. For example, for 18 convolution cores with different 10×10 dimensions, 18 feature maps can be obtained, which can be regarded as different channels of the original image. The convolution layer contains $10 \times 10 \times 18 = 1800$ parameters.

(2) Weight sharing: Weight sharing is another feature of CNN which is different from traditional neural network. It effectively reduces the number of parameters that need to be trained in the neural network model and improves the learning efficiency of

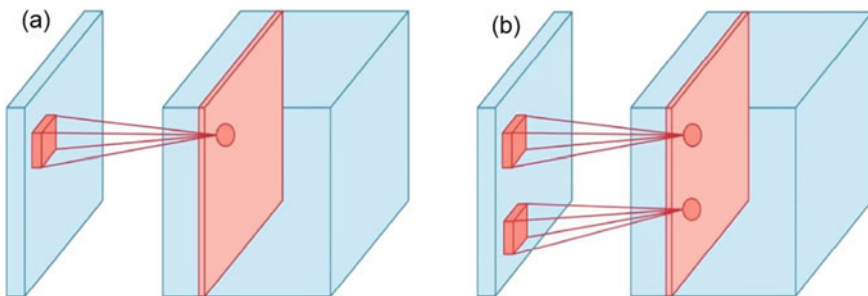


Fig. 18.10 Schematic diagram of local connection (a) and weight sharing (b) for CNN

the model. As shown in Fig. 18.10b, the main performance of weight sharing is that the parameter values of convolution kernel in the convolution layer are trained at the same time, and the convolution kernel with the same parameters act on the input data, which is equivalent to the process of extracting certain local characteristics from the input data. Each feature map generated represents the extraction results of a specific feature extraction method on the input data. The two neurons in the same sampling layer have the same parameters, and the weight parameters used in the feature extraction process of the same feature map are the same. It can be regarded as the translation of the same convolution kernel on the feature plane of the input layer. The method of extracting features from the convolution kernel in a certain region is also applied to other regions.

The weight sharing principle greatly reduces the number and complexity of neural network model parameters. When extracting each feature map, the hidden layer only needs to train a convolution kernel and a set of parameters, which significantly reduces the difficulty of network training and improves the training speed. Each different convolution kernel will convolute with all the input data, rather than only act locally, which can make the neural network more robust.

18.2.2 *Optimistic Algorithm*

Weight update is one of the most important processes in neural networks. At present, the most commonly used updating algorithm is random gradient descent method. In addition to random gradient descent method, momentum method, Adagrad, RMSProp, Adadelta, Adam, and other optimization methods are all used to solve various problems in the optimization process, so as to accelerate the training of network model and improve the performance of the model.

- (1) Stochastic gradient descent (SGD): The advantage of this optimization algorithm is that it can disperse the amount of training data, reduce computer load, and improve computational efficiency. Especially when the training data is repeated, the efficiency of SGD will not reduce due to the block training mode. The learning rate is generally selected through experience and error. Excessive learning rate will cause severe oscillation of the target curve, resulting in the failure of the neural network to update the parameters normally, making the training model unable to converge normally. Too small learning rate is easy to limit the system to local minimum and cannot jump out. This minimum value often makes the loss value of the system larger and cannot complete the optimization of neural network. The Mini-batch Gradient Descent (Mini-batch SGD) method randomly selects a part of the sample (Mini-batch) for gradient calculation and updates the parameters each time, which ensures the fast calculation speed and converge quickly.
- (2) Momentum: On the one hand, the momentum method is introduced to solve the “canyon” and “saddle point” problems. It can also be used to accelerate

the convergence of SGD, especially for high curvature, small amplitude but consistent direction gradient. The introduction of momentum into SGD can solve the instability of its random update to a certain extent and reduce the oscillation, because it retains the direction of the previous gradient to a certain extent during the update, and uses the current block to fine-tune the update direction.

- (3) Adagrad: The idea of the algorithm is to adapt to each parameter of the model independently, that is, the parameter with large deviation has a larger learning rate, and the parameter with small deviation corresponds to a small learning rate. Specifically, the learning rate of each parameter will scale the square root of each parameter inversely proportional to the sum of its historical gradient square values. The disadvantage is that the learning rate is monotonically decreasing. If the learning rate in the later stage of training is too small, the training will be difficult and even early termination. In addition, a global initial learning rate needs to be set.
- (4) Root mean square prop (RMSProp): RMSProp is mainly to solve the problem of excessive attenuation of learning rate in Adagrad method, that is, the learning rate becomes too small to continue training before reaching the local minimum. RMSProp uses exponential decay averaging to enable it to converge quickly after finding a “convex” structure. In addition, a hyper parameter is added to RMSProp to control the attenuation rate. RMSProp has been proved to be an effective and practical deep neural network optimization algorithm. RMSProp still needs to set a global learning rate.
- (5) Adaptive moment estimation (Adam): Adam algorithm is an algorithm that combines Momentum algorithm and RMSProp algorithm. It can dynamically regulate the learning rate and make it change in a stable direction. In the process of optimizing network parameters, it can efficiently find the global optimal solution according to the input data.

In practical training and application, no optimization algorithm can solve all problems perfectly. Therefore, according to the actual application requirements, it is very important to select the appropriate optimization algorithm and parameters on the basis of understanding the principle of the algorithm.

18.2.3 Loss Function

The loss function is the reflection of the neural network model on the degree of data fitting. The worse the fitting is, the greater the value of the loss function is. At the same time, when the loss function is relatively large, its corresponding gradient should be relatively large, so that the variables can be updated faster. Therefore, there are two requirements for the loss function. Firstly, the real error of solving the problem should be reflected. Secondly, the loss function should have a reasonable

gradient, which is conducive to solving the gradient and updating the weights and parameters.

Loss function is an important factor in the design of neural network. In the face of specific problems, different loss functions need to be selected or designed. Common loss functions include:

(1) Mean square error (MSE)

The mean square deviation is a commonly used loss function to evaluate the difference between the test data and the target data, that is, the mean of the square sum of the errors of the corresponding points between the predicted data and the original data. MSE performs well in linear regression and can effectively calculate the reverse gradient propagation. When the activation function is sigmoid function, it is easy to cause the loss of the gradient, which leads to the weight of the shallow layer is not updated, there is a problem of gradient vanishing. Therefore, the selection of MSE in logical regression needs to consider the loss of gradient.

(2) Cross-entropy

Cross entropy is a concept in information entropy theory, which is originally used to estimate the average coding length. In DL, it can be seen as the degree of difficulty that the probability distribution $p(x)$ (the distribution of real markers) is represented by the probability distribution $q(x)$ (the distribution of prediction markers of the trained model). Cross entropy depicts the distance (or similarity) between two probability distributions, that is, the smaller the cross entropy is (the smaller the relative entropy is), the closer the two probability distributions are. One advantage of cross-entropy as a loss function is that the sigmoid function can avoid the problem of decreasing the learning rate of the mean square error loss function when the gradient decreases, because the learning rate can be controlled by the output error.

(3) Log-likelihood cost

The essence of the log-likelihood loss function is that the likelihood value of a set of parameters under a pile of data is equal to the product of the conditional probability of each data under this set of parameters, and the loss function is generally the sum of the losses of each data. In order to change the product into sum, logarithm is taken and a negative sign is added to make the maximum likelihood value corresponding to the minimum loss. Log-likelihood loss function is generally used for multi-classification problems. Softmax activation is added to the output layer, and then the log-likelihood loss is calculated.

18.2.4 Activation Function

Activation function activates a part of neurons in the neural network and transmits the activation information back to the next layer neural network at running time, the reason why the neural network can solve the nonlinear problems is essentially that

nonlinear factors are added by activation function. They not only make up expressive ability of the linear model but also save and map the characteristics of activated neurons to the next layer through the function. The commonly used activation functions include sigmoid, tanh, and ReLU.

(1) Sigmoid function

The shape of sigmoid function is S-shaped, which is the most commonly used activation function, and its function form is

$$f(x) = \text{sigmoid}(x) = \frac{1}{1 + e^{-x}} \quad (18.1)$$

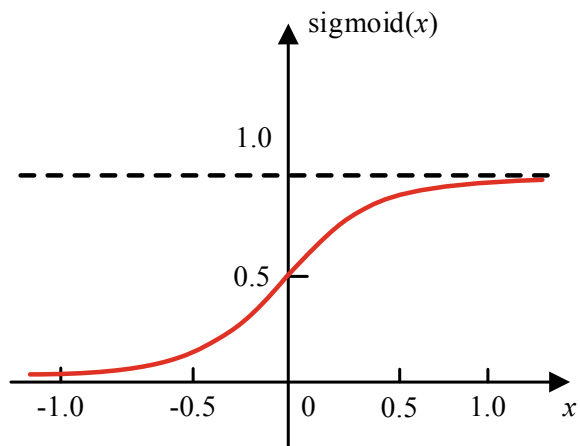
The image of sigmoid function is shown in Fig. 18.11. It is a strictly increasing function. It shows linearity near the value of 0 for x and nonlinearity far away from the value of 0. Therefore, this function can better balance the linear and nonlinear characteristics, and the function is differentiable. The trend of the gradient can be seen from the figure. When the input is very large or very small, the gradient of the neurons is close to 0, which makes the back propagation approach to 0 in the back propagation algorithm, resulting in little update of the final weight.

(2) Tanh function

Tanh is hyperbolic tangent function, the curves of tanh function and sigmoid function are relatively close. The difference between them is the output interval. As shown in Fig. 18.12, the output interval of tanh is between $(-1, 1)$, and the entire function is centered on 0. Its function form is

$$f(x) = \tanh(x) = \frac{\sinh x}{\cosh x} = \frac{e^x - e^{-x}}{e^x + e^{-x}} \quad (18.2)$$

Fig. 18.11 Sigmoid activation function



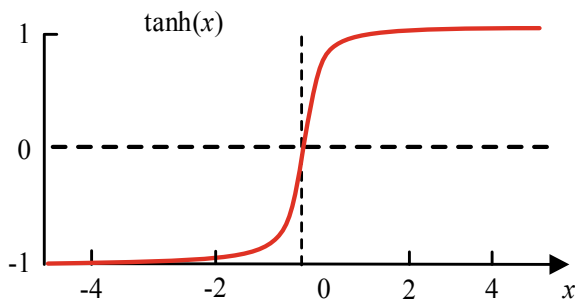


Fig. 18.12 Tanh activation function

Tanh function is an odd function whose function image is a strictly monotonic increasing curve passing through the origin. It allows the activation function to take negative value which can sometimes produce better practical results, but it still has the problem of gradient saturation.

(3) Rectified Linear Unit (ReLU) Function

The ReLU activation function, namely the rectified linear unit, is also known as the rectifier linear unit. It is the default activation function most used in CNN at present. Its form is as follows:

$$f(x) = \max(0, x) \quad (18.3)$$

The function and its derivative are shown in Fig. 18.13.

The output of the ReLU function is nonlinear. At $x = 0$, the function is not differentiable, but it usually has little effect on the gradient descent algorithm, because the numerical calculation hardly reaches the point which the gradient is 0, and the left derivative and the right derivative are usually defined in the network training. In the region $x > 0$, the function is a linear function with 1 at the first derivative. ReLU is in an active state, so it retains the good characteristics of linear model and is easy to use the optimization method based on gradient descent. The gradient values in this

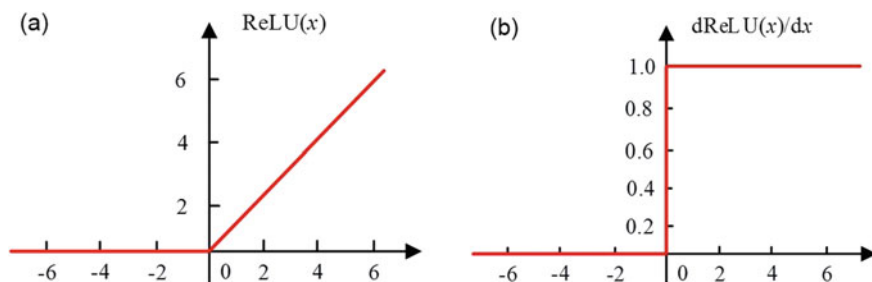


Fig. 18.13 ReLU activation function (a) and its derivative (b)

region are consistent, and it is not too large or too small, which is conducive to avoid the gradient vanishing or gradient explosion problems of the neural network due to the increase of depth, and its convergence rate is much faster than the sigmoid and tanh activation functions. The form of the ReLU function is relatively simple, and the memory consumption is less in the gradient calculation process. In addition, the ReLU function changes the output of some neurons into 0, which makes the network sparse and alleviates the over fitting problem to some extent.

One drawback of ReLU is that when $x < 0$, the function is in an inactive state, and the gradient is 0, that is to say, it cannot learn the samples that make them active to 0 by the gradient based method. In order to ensure that the gradient can be received everywhere in x , a variety of extended ReLU functions appear, such as leaky ReLU activation function, which does not set 0 when the input is less than 0, instead replaced by multiplying a small constant value. The function expression is $f(x) = \max(ax, x)$, and a is a small constant, such as 0.01. For example, Parametric ReLU activation function which approximates the leaky ReLU function, but a is not a preset constant which obtained by data learning. In addition, there are also Randomized ReLU activation functions.

The loss function is often selected by combining the activation function, because the activation function is inevitable in the process of chain derivation of the back propagation algorithm. Activation function is one of the most important designs in the forward propagation, which increases the complexity of the model and provides more nonlinear operations.

In the actual training process, MSE loss function is generally combined with sigmoid function for linear regression problems. The greater the difference between the real results is, the greater the difference is, and the higher the regression accuracy is. MSE is prone to gradient vanishing in classification problems, and cross-entropy function can solve the problem of gradient vanishing in classification models. The cross-entropy function only cares about whether the classification results are correct, while the MSE function focuses on the size of each category, which is unnecessary in the actual classification problems. Therefore, cross-entropy function is more suitable for logistic regression problems.

In addition, when the input data features are significantly different, the effect of using tanh will be very good, and the feature effect will continue to expand and show in the cycle. If the feature difference is not obvious, sigmoid works well. At the same time, when sigmoid and tanh are used as activation functions, the input needs to be standardized. Otherwise, all the activated values enter the flat area, and the output of the hidden layer will all converge and lose the original feature expression. ReLU is much better, and sometimes input normalization is not required to avoid the above situation. Therefore, most convolution neural networks now use ReLU as activation function.

18.2.5 *Methods to Avoid Over-Fitting*

In the deep neural network, when the difference among the parameters of the model and the number of training samples is too large, the over-fitting problem often occurs. It is also one of the main difficulties in network training. Over-fitting means that good fitting results can be obtained on the training data, but the data set outside the training data cannot be well predicted. To avoid over-fitting of CNN model, the following methods can be used.

(1) **Norm Regularization Methods**

Norm regularization is a way to reduce the complexity of the model by adding a penalty term to the loss function. The most common techniques are L1 and L2 regularization.

L2 regularization is the most norm regularization method, which is also called ridge regression or Tikhonov regularization in multiple linear regression. It adds a penalty term directly to the loss function. That is, the L2 norm of each weight w in the neural network is calculated, and then added it to the loss function, which is expressed as $0.5*\lambda*w^2$. λ is the regularization intensity. The greater the value is, the stronger the regularization is, and the more over-fitting can be prevented. However, if the value is too large, under-fitting will occur. Generally, the verification set is used to determine the hyper-parameters.

L1 regularization is another common regularization method, which is also called Lasso regularization in multiple linear regressions. The L1 norm is calculated for each weight w in the neural network, and then add it to the loss function, expressed as $\lambda*|w|$. The L1 regularization method will make the weight sparse, that is, the maximum value of the weight is close to 0, which is equivalent to part of the input data to participate in the network calculation, and it is robust to noise or redundant part of the input. This property can be used for feature selection.

(2) **Ensemble Methods**

Some ensemble learning strategies can prevent over-fitting. Common ensemble learning strategies include bagging, boosting, and RF and so on, which can reduce the generalization error by combining multiple models. This method can also be used in deep learning, but it increases the cost of computing and storage.

(3) **Dropout Method**

As shown in Fig. 18.14, the Dropout method randomly sets the weights of some neurons to 0 in each network training process, that is, to let some neurons fail which is equivalent to training different neural networks, so that the diversity of models can be enhanced, and the effect of similar multiple model ensembles can be obtained, so as to avoid over-fitting. In addition, Dropout leads to scarcity and reduces the complexity of network structure, which makes the difference of local data clusters more obvious, which is also the reason that it can prevent over-fitting.

Dropout generally appears in the full connection layer, commonly used to optimize the training network. During the training of network, each iteration randomly

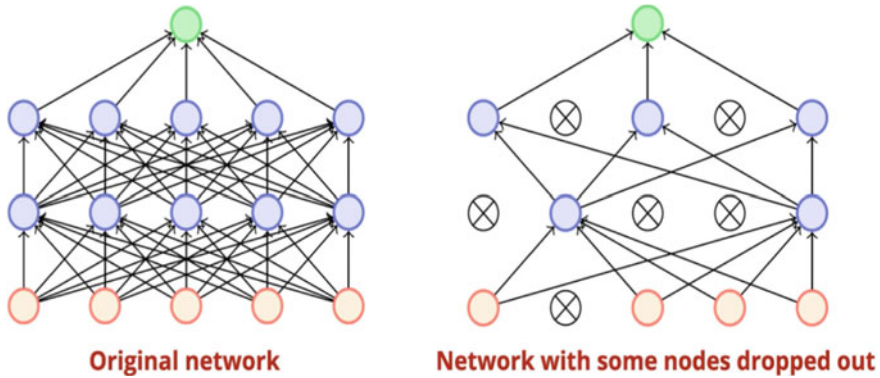


Fig. 18.14 Schematic diagram of Dropout method

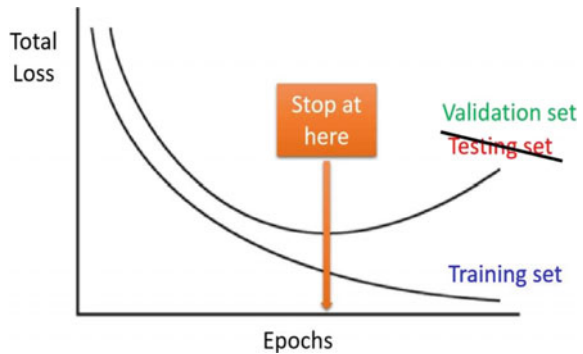
suppresses some neurons in the hidden layer with probability p , and the remaining neurons connect to the next layer of neurons. After back propagation, the uninhibited neuronal parameters are updated. The next iteration restores the parameters of the inhibited neurons, while the other neurons maintain the parameters updated after the last iteration, and then continue to randomly inhibit some neurons. Repeating the above process can generate different neural networks. Finally, the comprehensive averaging strategy is used to combine these different neural networks as the final output model. The hyper parameter p is called the discard rate, which is usually set to 50%.

(4) Batch normalization (BN) Method

The traditional deep neural network, with the deepening of the number of layers, the model will become difficult to train and fit, because the deep neural network between different layers will be nonlinear transformation. The purpose is to make the network more representatives through nonlinear transformation, multi-layer linear superposition is meaningless. The result of multi-layer nonlinear transformation makes the distribution of training data offset or change, which is called internal covariate shift. The reason why the training convergence is slow is that the overall distribution of the data gradually approaches the upper and lower bounds of the range of nonlinear function, where the gradient is very small, which can be said to be close to 0.

The essential idea of batch standardization (BN) method is to force the input distribution of the gradient vanishing interval that is gradually mapped to the nonlinear function back to the normal distribution with approximate mean value of 0 and variance of 1, so that the input value of the nonlinear transformation function falls into the region sensitive to the input. The learning rate can be increased many times to avoid the problems of gradient vanishing and difficult fitting, and delay the occurrence of over-fitting to a certain extent. BN forcibly changes the distribution of input data, avoiding the diffusion of input data distribution to the nonlinear activation function to the gradient vanishing region, and offsetting the nonlinear expression ability of the nonlinear activation function. Therefore, BN adds two parameters of scale and

Fig. 18.15 Schematic diagram of early stop training



shift after pulling the input data distribution back to the normal distribution with an approximate mean value of 0 and variance of 1, so as to scale and shift the data to maintain the nonlinear learning ability of the network at the same time.

(5) Data Augmentation Method

Most of the reasons for over-fitting are the lack of training samples and the increase of training parameters. If the training samples lack diversity, more amount of training parameters is meaningless, because it will cause over-fitting. One way to make the generalization ability of the model better is to use more training data for training. Feature diversity brought by large amounts of data helps to make full use of all training parameters. For image data, the commonly used methods of data enhancement include flipping transform, random pruning, color dithering, translation transform, scale transform, contrast transform, noise disturbance, rotation transform, and reflection transform. Data enhancement can also be obtained by generative adversarial networks (GAN).

(6) Early Stopping Method

Early Stopping is applicable when the expression ability of the model is strong. In this case, the general training error will gradually decrease with the increase of training times, and the test error will first decrease and then increase again. In order to avoid over-fitting of the training set, a good solution is to stop in advance and interrupt training when its performance on validation set begins to decline (Fig. 18.15).

18.2.6 Classical Convolution Neural Network Architecture

The first successful application of CNN is the LeNet-5 architecture developed by Cun et al. in 1998, which was used for the identification of handwritten numerals in postal code. The great development of deep convolution network started from the AlexNet network in 2012, which was proposed by Krizhevsky et al. in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC). AlexNet has a similar architecture

to LeNet-5, but its network is deeper and implemented by using multiple convolution layers. In the 2014 ILSVRC competition, GoogLeNet and VGGNet are two excellent architectures. GoogLeNet creatively proposed the Inception structure to solve the problem of gradient dissipation. The main contribution of VGGNet is that the depth of the network is the key reason for good performance. In 2015, ResNet developed by He et al. [12] solved the problem of network degradation through residual structure, greatly improved the depth of the network. Several classical convolution neural network models are introduced, including LeNet-5, AlexNet, VGGNet, GoogLeNet, and ResNet.

(1) LeNet-5 Network

LeNet-5 is the cornerstone of modern CNN. Its network structure is shown in Fig. 18.16, including three convolution layers, one full connection layer, and one Gaussian connection layer.

Layer 1: input layer is batch size 32×32 black and white resolution image;
 Layer 2: C1, convolution layer, there are six feature maps, convolution kernel size is 5×5 , depth is 6, without full 0 filling and step size is 1, a total of $28 \times 28 \times 6$ neurons ($32 - 5 + 1 = 28$), the number of parameters is 156 ($5 \times 5 \times 6 + 6 = 156$, 6 is the bias parameter), each unit is connected with 25 units in the input layer;
 Third layer: S2, pooled subsampling layer, there are six feature maps, each feature map size is 14×14 , pooled kernel size is 2×2 , long and wide step sizes are 2;
 Layer 4: C3, convolution layer, convolution kernel size is 5×5 , 16 feature maps, each feature map size is 10×10 ($14 - 5 + 1 = 10$), with a fixed connection to the third layer;
 Layer 5: S4, pooled subsampling layer, with 16 feature maps, each feature map size is 5×5 ($10/2$), pooled kernel size is 2×2 , long and wide step sizes are 2;
 Layer 6: C5, convolution layer, batchsize $\times 120$ feature maps;
 Layer 7: F6, full connection layer, batch size $\times 84$ feature maps;
 Layer 8: output layer, with batch size $\times 10$ feature maps.

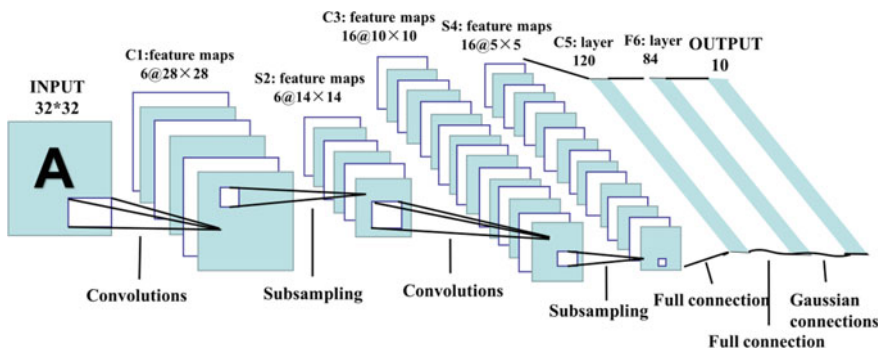


Fig. 18.16 LeNet-5 model of CNN

(2) AlexNet network

As shown in Fig. 18.17, AlexNet has five convolution layers, three of which are connected to the maximum pooling layer. The last three layers of AlexNet are full connection layers. The method established the dominant position of deep learning (deep convolution network) in computer vision, and also promoted the expansion of deep learning in speech recognition, natural language processing, reinforcement learning, and other fields.

AlexNet uses ReLU as the activation function of CNN, and verifies that its effect exceeds Sigmoid in deeper networks, which solves the gradient diffusion problem of Sigmoid in deep networks. Dropout is used during training to randomly ignore some neurons to avoid over-fitting the model. AlexNet all uses overlapping maximum pooling. Previously, CNN generally uses mean pooling to avoid the blurring effect of mean pooling. The step size in AlexNet is smaller than the size of the pooling kernel so that the outputs of the pooling layer will be overlap and coverage, which improves the richness of features.

AlexNet randomly intercepts 224×224 size regions (and horizontal flip mirrors) from 256×256 original images using data enhancement, which increases the amount of data by 2048 times, significantly reduces over-fitting and improves generalization ability. In addition, AlexNet uses the powerful parallel computing ability of GPU to deal with a large number of matrix operations in neural network training.

(3) VGGNet Network

VGGNet divides the network into five segments. Each segment connects several 3×3 convolution networks in series. Each segment is followed by a 2×2 maximum pooling layer and the last are three full connection layers and a softmax layer. VGGNet uses multiple convolution layers with smaller convolution kernels (3×3) to replace a convolution layer with larger convolution kernels (e.g., 5×5). On the one hand, it can reduce parameters; on the other hand, it is equivalent to more nonlinear mapping, which can increase the fitting/expression ability of the network.

VGGNet has different architectures such as VGG-11, VGG-16, and VGG-19, and constructs a CNN with 16–19 layers. All the small convolution kernels of 3×3 and the maximum pooling kernel of 2×2 are used to improve the performance by

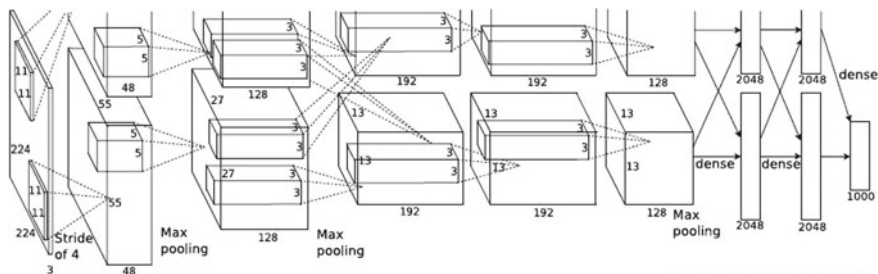


Fig. 18.17 AlexNet model of CNN

deepening the network structure, so as to achieve a larger receptive field (such as 5×5) similar effect to extract more complex features.

(4) GoogLeNet Network

GoogLeNet (also known as Inception) is a new deep learning structure. Previous structures such as AlexNet and VGGNet have achieved better training results by increasing the depth (number of layers) of the network, but the increase of the number of layers will bring many negative effects, such as too many parameters, and if the training data set is limited, it is easy to over-fitting. The larger the network, the greater the computational complexity, it is difficult to apply. The deeper the network is, the easier the gradient passes, and it is difficult to optimize the model. Inception is proposed to improve the training results from another perspective: more efficient use of computing resources, more features can be extracted under the same amount of calculation thereby enhancing the training results.

As shown in Fig. 18.18, the basic structure of Inception has four branches: the first branch carries out 1×1 convolution on the input, and 1×1 convolution is an excellent structure, which can realize cross-channel interaction and information integration, improving the expression ability of the network. Furthermore, the dimension of the output channel can be increased and reduced simultaneously. The second branch first uses 1×1 convolution and then connects 3×3 convolution, which is equivalent to two feature transformations. The third branch is similar, first 1×1 convolution, and then connect 5×5 convolution. The last branch is directly using 1×1 convolution after 3×3 maximum pooling. Four branches of Inception are merged at the end by an aggregation operation (aggregate on the dimension of output channels).

The Inception structure uses 1×1 convolution to carry out the lifting dimension, which reduces the computational complexity and obtains a more compact network structure. Although GoogLeNet has 22 layers in total, the number of parameters

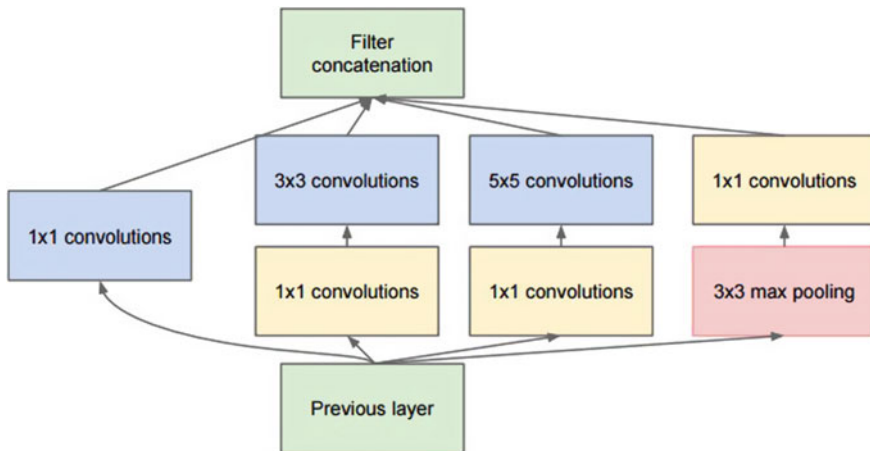


Fig. 18.18 Schematic diagram for basic structure of inception

is only one twelfth of the eight-layer AlexNet. The Inception structure performs convolution and re polymerization on multiple dimensions at the same time, and convolution on multiple scales at the same time. The characteristics of different scales can be extracted. The principle of sparse matrix decomposition into dense matrix calculation is used to accelerate the convergence speed.

GoogLeNet has a deeper network structure and less parameters and computation, which is mainly due to the extensive use of 1×1 convolution in the convolution network and the replacement of the full connection layer in the traditional network architecture with the AveragePool. This requires careful design of the Inception architecture to achieve excellent results.

(5) ResNet Network

For traditional deep learning networks, a simple increase in depth can lead to gradient diffusion or gradient explosion. The solution to this problem is regularization initialization and batch normalization in the middle (Batch Normalization), which can train dozens of layers of networks. Although the above method can be trained, but there is another problem, which is degradation, that is, as the number of network layers increasing, the accuracy rate on the training set is saturated or even decreased. Deep residual network (ResNet) solves the degradation problem of deep network by residual learning.

For an accumulation layer structure (stacked by several layers), when the input is x , the learned feature is denoted as $H(x)$. Now we hope it can learn the residual $F(x) = H(x) - x$, so the original learning feature is $F(x) + x$. This is because residual learning is easier than direct learning of original features. When the residual is 0, the accumulation layer only makes identity mapping at this time, at least the network performance will not decline, and in fact the residual will not to be 0, which will also make the accumulation layer learn new features based on input features, so as to have better performance. This can solve the problem gradient vanishing of the deep network, so that the network can be done very deep.

The structure of residual learning is shown in Fig. 18.19. It's kind of like a "short circuit" in the circuit, so it is a shortcut connection. In the ResNet, the input and

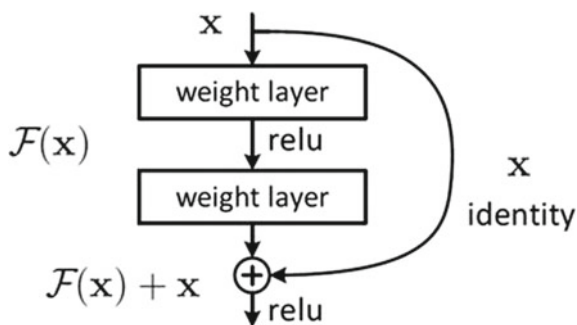


Fig. 18.19 Schematic diagram for structure of residual learning

output are added to a corresponding element-wise by Shortcut. This simple addition does not add additional parameters or computation to the network, but can greatly increase the training speed and improve the training effect of the model, and when the number of layers of the model deepened, this simple structure can well solve the problem of gradient vanishing.

(6) DenseNet Network

In deep learning networks, with the deepening of network depth, the problem of gradient vanishing will become more obvious. DenseNet network is separated from the thinking of deepening network layers number (ResNet) and widening network structure (Inception) to improve network performance. From the perspective of features, it is set through feature reuse and bypass set. It not only greatly reduces the number of network parameters but also alleviates the problem of gradient vanishing to a certain extent. The basic idea is to directly connect all layers on the premise of ensuring the maximum information transmission among layers of the network.

DenseNet is composed of Dense blocks, using the structure of Batch Normalization (BN) + ReLU + 3×3 Conv (Fig. 18.20). In these blocks, each layer is closely connected, and each layer obtains input from the output feature map of the previous layer. DenseNet architecture maximizes the use of residual mechanism so that each layer is closely connected to its subsequent layers. The compactness of the model makes the learned features non-redundant, because they are shared through collective knowledge. In addition, due to the short connection, the gradient is easier

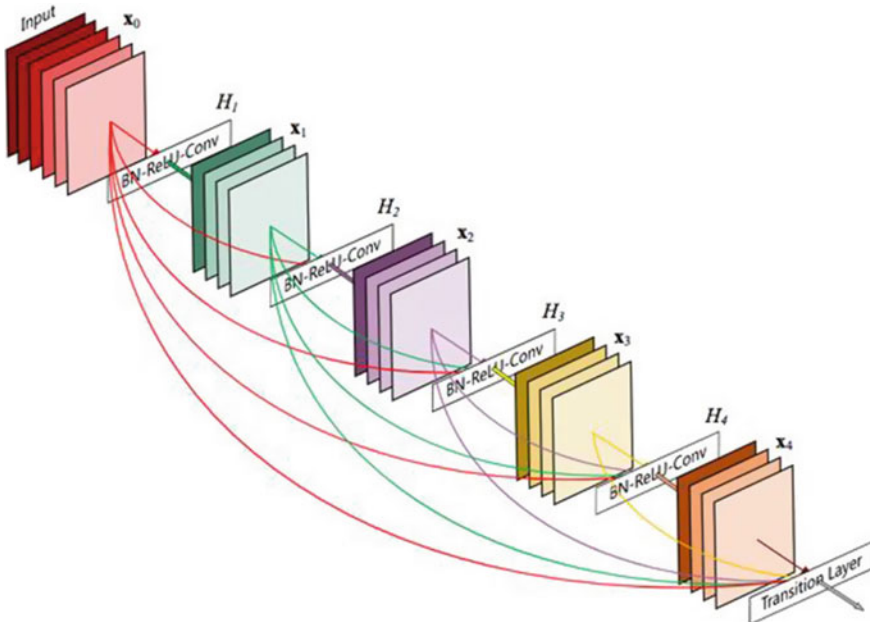


Fig. 18.20 Schematic diagram for structure of DenseNet network

to reverse flow. This high reusability of residuals creates deep oversight, since each layer receives more oversight from the previous layer, so the loss function responds accordingly.

The core idea of DenseNet is to establish the connection between different layers, make full use of the characteristics, and further reduce the gradient vanishing problem. With the deepening of the network, the training effect is gradually improved. In addition, the bottleneck layer, the translation layer and the smaller growth rate are used to narrow the network and reduce the parameters, which inhibit the over-fitting effectively and reduce the amount of calculation.

18.2.7 Popular Deep Learning Software Framework

The software framework of deep learning provides the necessary basis for the realization of deep learning architecture. These software frameworks can realize the rapid construction of training, testing, and tuning models by modularizing and encapsulating deep learning algorithms, and provide strong support for the prediction of technical applications and the decision-making of landing. There are many frameworks available for deep learning projects. The current popular deep learning frameworks mainly include TensorFlow, PytorchKeras, Caffe, Caffe2, MXNet, CNTK, Deeplearning4J, etc. These frameworks have their own advantages and disadvantages. It is very important to choose a suitable deep learning software framework for achieving the goal.

(1) TensorFlow

TensorFlow is an open source machine learning library of Google in 2015. It is one of the most popular deep learning frameworks. It supports distributed training, extensible production deployment options and Android and other devices. Tensorflow has a multi-level structure, which can be deployed on various servers, PC terminals and web pages, and supports GPU and TPU high-performance numerical computing. It is widely used in Google's internal product development and scientific research in various fields. TensorFlow is relatively mature, stable, and focuses on the industrial field as a whole, which is suitable for the development of large and medium-sized projects.

(2) PyTorch

In 2017, Facebook launched the PyTorch platform. Because of its dynamic computing diagram and efficient memory, it is suitable for rapid prototyping or small-scale projects, so it has become the preferred framework for a large number of academic researchers. It has the advantages of simple and transparent modeling process, many pre training models, easy combination of modular components, and support for distributed training.

(3) **Keras**

Francois Chollet who worked in Google developed Keras. As the top wrapper of Theano, Keras is mainly used for rapid prototyping. Although it is once one of the most popular deep learning libraries, Theano is now out of service. Later, several frameworks, such as TensorFlow, MXNet, CNTK, were extended from Keras and used as the back end. It supports various neural network layers, such as convolution layer, circulation layer or dense layer, which can be applied in translation, image recognition, speech recognition, and other fields. Keras is one of the fastest growing deep learning libraries at present. Characterized by simple prototype design, simple, and intuitive interface, Keras supports multi-GPU training, and is suitable for beginners to get started quickly.

(4) **Caffe and Caffe2**

Caffe is a Python deep learning library developed by Yangqing Jia from the University of Berkeley to monitor computer vision problems. It is suitable for CNN, image processing, and fine-tuning pre training networks. It can fine-tune the network with little or no code. On the basis of Caffe, Caffe2 introduced in 2017 is a lightweight modular framework designed for mobile and large-scale deployment in the production environment. Caffe2 is more scalable and lightweight. In 2018, Caffe2 project has been merged with PyTorch.

(5) **MXNet**

MXNet is a deep learning framework created by Apache Software Foundation, supported by Microsoft, Intel, and Amazon. MXNet supports multiple languages such as Python, C++, Julia, R, and JavaScript. For large-scale industrial projects, MXNet is a good software framework, which is very fast, flexible, and efficient, and can run on any device, providing rich support for a variety of programming languages.

(6) **CNTK**

CNTK is an open source deep learning framework developed by Microsoft to process big data sets and supports Python, C++, C# and Java. It is applicable to almost all types of tasks from voice, text to vision. Its characteristics are good performance and scalability, with more highly optimized components, in terms of resource use is also very effective.

(7) **DeepLearning4J**

Deep Learning4J is a commercial open source framework, which is an open source deep learning library for Java and Java virtual machines. It is a computing framework that widely supports various deep learning algorithms. The deep learning framework has great potential in image recognition, natural language processing, and text mining. It is flexible, efficient, and can process large amounts of data without sacrificing speed.

(8) **MatConvNet**

MatConvNet is a MATLAB toolbox for implementing CNN launched by Cambridge University. Because of its pure MATLAB development environment, it may be the most easily available software framework. MatConvNet provides a friendly and efficient environment for researchers, including many CNN computing blocks, such as convolution, normalization, and pooling. Most of them are written in C++ or CUDA, which means that it allows users to write new blocks to improve computational efficiency.

In addition, MathWorks has launched MATLAB and Simulink of 2018b version, which contains important deep learning enhancement functions, as well as new functions and bug fixes in each product series. The new Deep Learning Toolbox replaces the Neural Network Toolbox and provides a framework for engineers and scientists to design and implement deep neural networks. Image processing, computer vision, signal processing, and system engineers can use MATLAB to design complex network architecture more easily and improve the performance of its deep learning model. Using the ONNX converter in 2018b, the model can be imported and exported from the supporting frameworks (such as PyTorch, MXNet, and TensorFlow). With this interoperability, the model trained in MATLAB can be used in other frameworks. Similarly, models trained in other frameworks can be imported into MATLAB to perform debugging, validation, and embedded deployment tasks.

Currently, platforms such as MATLAB, Python, C++ , Java, and Go are usually used for the specific implementation of convolutional neural network. In order to facilitate the use of algorithms, frameworks such as Tensorflow, Theano, Caffe, and Pytorch are generally used to build classification or regression models. These open source development languages and learning frameworks provide very convenient conditions for researchers. For example, the TensorFlow programming interfaces are based on the graphical interfaces, which can easily run on the Python platform. The algorithm structure is simple and clear, and Python can use GPU for parallel computing, combined with the high-performance library unit of deep learning provided by Nvidia, cuDNN, etc. It greatly improves the training speed and model performance of deep learning. In addition, there is Sickest-Learn based on the commonly used machine learning algorithm.

18.2.8 Design of Convolution Neural Networks

In the design process of CNN, more parameters need to be selected, as shown in Fig. 18.21, the main parameters include [13, 14] as follows.

- (1) The layer structure of the network (the number of convolution layers, the number of full connection layers.)
- (2) Convolution kernel size, number of convolution kernels, moving steps (Stride) in convolution layers
- (3) The type of activation function

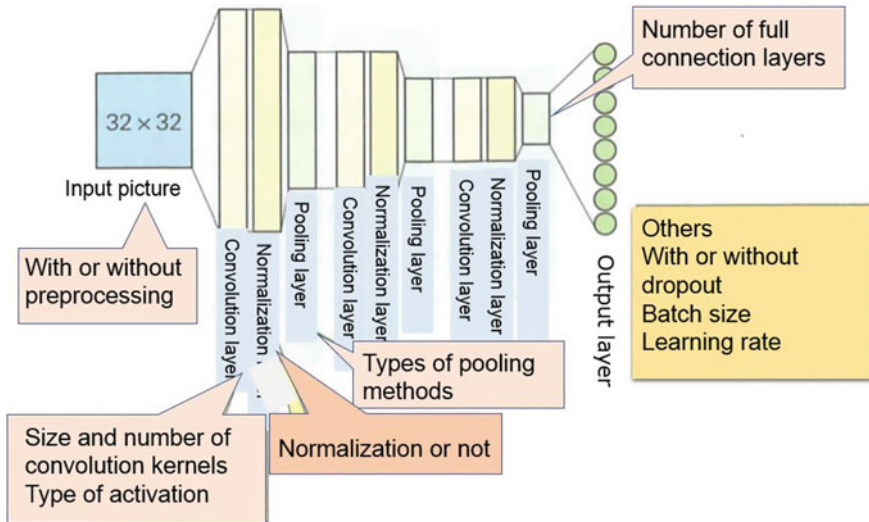


Fig. 18.21 Schematic diagram of training parameters for CNN

- (4) Types of pooling methods
- (5) With batch normalization or not
- (6) The probability of Dropout
- (7) The size of mini-Batch size
- (8) The type of loss function and its parameters (regularization coefficient, etc.)
- (9) The type of optimization algorithms and its parameters (learning rate, momentum, etc.)
- (10) Number of iterations.

In the CNN structure, the deeper the depth and the more number of feature faces, the greater the feature space that the network can represent and the stronger the network learning ability. However, it also makes the network calculation more complex, and is easy to over-fitting. Therefore, in practice, the network depth, the number of feature surfaces, the size of the convolution kernel, and the steps of sliding should be appropriately selected so that a good model can be obtained and the training time can be reduced.

In the field of spectral analysis, the relationship between the size of convolution kernel and the moving step of convolution kernel has clear physical significance. As shown in Fig. 18.22, when the moving step of convolution kernel is less than the size of convolution kernel, the convolution kernel will overlap during the movement which means that more features can be extracted. Dimensions are similar to uniform interval division in the interval partial least squares (iPLS), when convolution kernel moves faster than convolution dimensions, convolution kernel will skip some spectral sub-intervals and do not extract features.

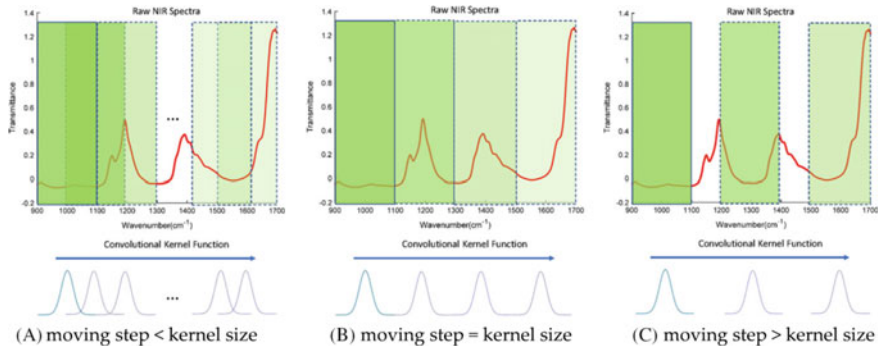


Fig. 18.22 Relationship between convolutional kernel size and moving steps

The size of convolution kernel, the number of convolution kernel, moving steps are not completely independent. There is a coupling relationship between them. Chen et al. [15] summed up the general design principles of CNN parameters in the field of NIR spectral analysis.

- (1) The size of convolution kernel should not be too small. When convolution kernel is small (10, 25), the convolution extracts features on sub-intervals that are not near the absorption peak, and using these features to model, the generalization performance of the model is usually poor. Conversely, when the size of the convolution kernel is large (50, 100), basically each feature extracted by the convolution kernel will contain the spectral information near the absorption peak. Using these features for modeling, the generalization performance of the model is usually better.
- (2) The number of convolution kernel does not need to be too many. When the size of convolution kernel is small, the number of features extracted by a single convolution is relatively large. In this case, continuing to increase the number of convolution kernels will double the total number of features extracted by all convolution kernels. It correspondingly results in “the number of features much more than the number of samples”, that is, the phenomenon of “over-fitting” will occur. The prediction performance of the model will be gradually reduced. Conversely, when the size of the convolution is large, the prediction performance of the model shows an upward trend, and when the number of convolution kernel reaches a certain value, and continues to increase, the prediction performance of the model will not continue to rise, but will decline slightly. Therefore, the number of convolution kernel is not the more the better. Given the appropriate size of convolution kernel, the number of convolution kernels does not need to be too large, as it is sufficient if it is not greater than 5.
- (3) Convolution kernel moving step should be less than convolution window width. When convolution moves at smaller steps, more features can be extracted to help to improve the generalization performance of the model.

For the selection of parameters, the ideal state is to select the optimal combination from these combinations of parameters for training. But the number of combinations is too large, when setting parameters, only the experience and more trials can constantly explore more optimized combinations. For example, when the background of input variables is complex in context, the number of convolutions can be increased so that the network can extract more features. While the number of input samples is large, the convolution and pooling layers need to be appropriately increased to form a deep convolution neural network. In addition, batch normalization is often added between the convolution layer and the activation layer, and dropout algorithms are introduced into the full connection layer to increase the robustness and convergence of the CNN model to a certain extent.

18.2.9 Training of Convolution Neural Networks

Similar to traditional neural networks, the training process of CNN is divided into two stages. The first stage is the stage of data propagation from low level to high level, the input data through the convolution and pooling of multi-layer convolution layer processing, proposed feature vectors, the feature vector into the full connection layer, to obtain classification or regression results, that is, forward propagation stage. Another stage is that when the results of the current propagation do not match the expectations, the error is calculated from the high level to the lower level of the propagation training stage, calculated the error of each layer, and then carried out the weight update, that is, the back propagation algorithm in the convolution neural network in the back propagation stage, like the shallow neural network, its essence is a chain-based process of guidance. In practice, mini-Batch-based training is often used, i.e., a fixed number of training samples are entered as a mini-Batch at a time, and each iteration calculation begins with the bias of each sample in mini-Batch, and then the average of the bias is calculated as a gradient to update the network weights [16].

The training process, as shown in Fig. 18.23, consists of the following steps.

1. The weights of the network are initialized first. The common method for initialization is random approach;
2. The output is obtained by forward propagation the input data through the convolution layer, down sampling layer and full connected layer.
3. The errors between the output value of the network and the target value are calculated, which is called loss function. The purpose of training is to minimize the loss function.
4. When the error is greater than expected, the error is transmitted back to the network by derivative, and the errors of the full connected layer, down sampling layer and convolution layer are obtained in turn. The training will be ended when the error is equal to or less than expected.
5. The weights are updated based on to the calculated error, and then goes to step 2 and repeat until convergence.

The training process for CNN is divided into two stages.

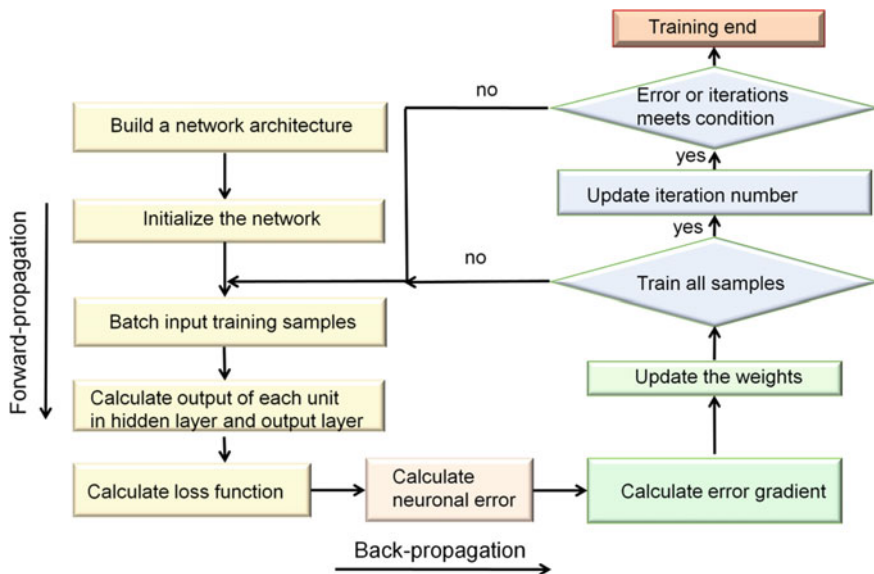


Fig. 18.23 Training flowcharts for CNN

(1) Forward propagation

The process of propagation the input from the previous layer to the next layer, progressing layer by layer, and finally outputting the result is called forward propagation. Forward propagation requires initialization of weights and offsets. Since initialization to zero causes the output to be the same for iteration that cannot be learned, parameters are generally initialized to random numbers in 0–1. The common output equation for a network is

$$a^i = f^i(w^i \times a^{i-1} + b^i) \tag{18.4}$$

where a^{i-1} , b , w , and f represent input, offset, weight, activation function, respectively. The output a^i is passed as input to the next layer, and the final output is y . In order to train deep neural network reasonably, it is necessary to quantify the difference between output y and real value using loss function to find the best parameters and reduce the loss function, which needs to be achieved by back propagation.

(2) Back propagation and gradient descent method

Back propagation refers to the process of transmitting the output error layer by layer to the input layer through the network, updating the weight and bias using the gradient

descent method, and iterating many times to minimize the loss function. The loss function quantifies the error between the actual value and the predicted value of the network. Assuming that there are m samples in the training sample set, the calculation formula of the loss function of a single sample is as follows:

$$J(W, b; x, y) = \frac{1}{2} \|h_w, b(x) - y\|^2 \quad (18.5)$$

where x, y, b, w , and $h_{w,b}(x)$ represent input, actual value, bias, weight, and prediction, respectively.

The loss function for m samples is

$$\begin{aligned} J(W, b) &= \frac{1}{m} \sum_{i=1}^m J(W, b; x^i, y^i) + \frac{\lambda}{2} \sum_{l=1}^{nl-1} \sum_{i=1}^{sl} \sum_{j=1}^{sl+1} (W_{ij}^l)^2 \\ &= \frac{1}{m} \sum_{i=1}^m \frac{1}{2} \|h_w, b(x^i) - y^i\|^2 + \frac{\lambda}{2} \sum_{l=1}^{nl-1} \sum_{i=1}^{sl} \sum_{j=1}^{sl+1} (W_{ij}^l)^2 \end{aligned} \quad (18.6)$$

The first item in Eq. 18.6 represents the mean variance, and the second indicates regularization. Wherein, nl represents the number of network layers, sl represents the number of neurons, W_{ij}^l represents the connection parameter of the i neuron of the l layer with the j neuron of the $l + 1$ layer, and the λ is the weight attenuation parameter (regularization parameter).

The gradient is calculated from the back forward, the gradient of the last layer is calculated first, and then the gradient of the previous layer is calculated, and the calculation uses a partial calculation from the gradient of the previous layer, and the information flows backward. The update formula for parameters w and b is as follows:

$$\begin{aligned} W_{ij}^{(l)} &= W_{ij}^{(l)} - \alpha \frac{\partial}{\partial w_{ij}^{(l)}} J(W, b) \\ b_i^{(l)} &= b_i^{(l)} - \alpha \frac{\partial}{\partial b_i^{(l)}} J(W, b) \end{aligned} \quad (18.7)$$

where α represents the learning rate and $b_i(l)$ represents the bias of the i neuron in the $l + 1$ layer.

18.2.10 Advantages and Disadvantages of Convolution Neural Network

CNN has the characteristics of low requirements for spectral preprocessing methods and wavelength variable selection methods, strong learning ability, and the ability to train and model multiple components at the same time, and the model performance is better when processing nonlinear and large sample data. However, CNN also has the following disadvantages.

- (1) CNN relies heavily on a large number of training samples, the larger the amount of data is, and the better the quality and performance it is. If there are not enough training samples, the training process will not converge and over fitting will occur.
- (2) The design of CNN is very complex, with a large number of hyper parameters, and the process of artificial parameter tuning is difficult and slow, and the time cost would be evitable.
- (3) The explanatory process of CNN is poor, the mechanism of action is explained by lack of clear theory, and the algorithm analysis is relatively difficult.
- (4) CNN put forward higher demand for hardware computing power, and ordinary computing hardware equipment is difficult to meet the speed and cost of huge computation.

18.2.11 Applications of Convolution Neural Network

Acquarelli et al. [17] have designed a CNN framework for molecular spectroscopic classification, consisting of a one-layer convolution layer and a one-layer full-connection layer, without a pooling layer. The activation function of the convolution layer uses ReLU, and the output layer uses the softmax activation function. Its loss function is a double-regularized cross-entropy function as shown in Eq. 18.8.

$$\begin{aligned}
 OBJ(w) = & \underbrace{-\frac{1}{N} \sum_{n=1}^N [yn \log \hat{y}_n + (1 - yn) \log(1 - yn)]}_{\text{Cross-entropy Error Loss}} \\
 & + \underbrace{\lambda_1 \|w\|^2 + \lambda_2 \cdot \|w - Shift(w)\|^2}_{\text{proximityL2norm}} \quad (18.8) \\
 & \text{Regularization term}
 \end{aligned}$$

where \mathbf{y}_n and $\hat{\mathbf{y}}_n$ is the target classification and CNN output classification values of the n th sample, respectively; w is the weight, λ_1 and λ_2 is the regularization parameters, $\text{Shift}(w)$ is the operation that moves the elements of w one position to the left. In addition to the standard L2 norm, the loss function uses an “approximate L2 norm”, which helps the network maintain the correlation between adjacent input variables (the number of wavenumbers in the vibration spectra) to punish the large differences between adjacent weights. For vibration spectra, these changes are not expected because the value of the spectrum on a wave number depends on the adjacent wave values.

Based on ten different vibration spectral databases (including MIR, NIR, and Raman spectroscopy), the results of the CNN and the classification effects of PLS-DA, logistic regression, and k-nearest neighbor (kNN) methods were compared. The results show that the CNN has the best results whether the spectral preprocessing method is used or not before classification. Compared with other CNN, the method has no pooled layer, and by inverse calculation, the characteristic wavelength range of effective spectral information can be extracted to better explain the training and learning process of convolution.

Le et al. [18] used stacked sparse auto-encoding networks (SSAE) in deep learning for the extraction of features of grain NIR spectra, and then use the affine transformation with extreme learning machine (AT-ELM) to establish quantitative models, and the prediction results are better than the PLS and ELM methods. Based on 6987 training set samples and 618 validation set samples, Cui et al. [19] studied CNN combined with NIR spectra to predict flour ash content. The effects of different activation functions, learning rate, random drop rate, regularization parameters, etc., were discussed, and the network training results were investigated and compared with PLS results. As shown in Fig. 18.24, the quality of regression coefficient (mainly

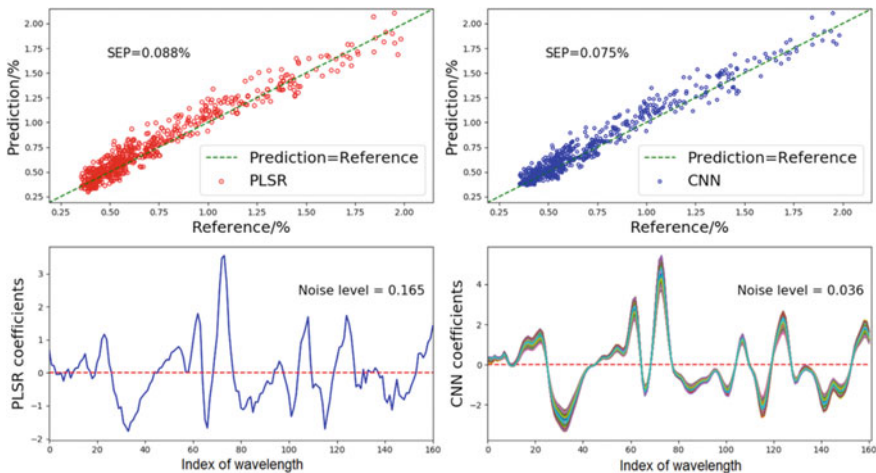


Fig. 18.24 Comparison of the quality of CNN with PLS models

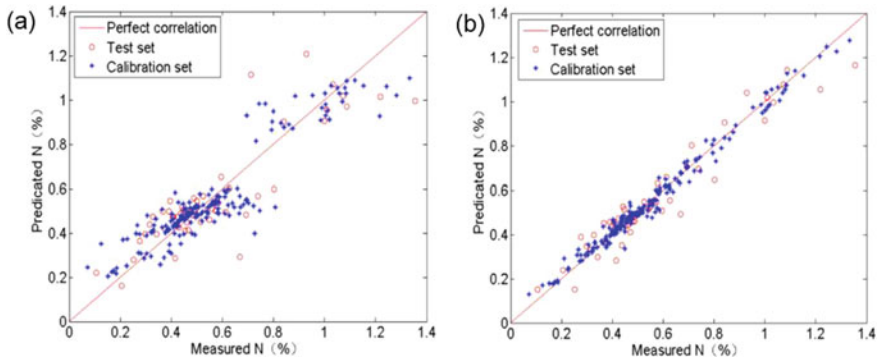


Fig. 18.25 Prediction results of nitrogen contents in the leaves of horsetail pine seedlings by SVR (a) and VWCNN (b) models

noise level) obtained by CNN is significantly better than the results of PLS, and CNN do not require spectral pretreatment methods, reducing the amount of modeling effort to a certain extent. Malek et al. [20] proposed the training of CNN using PSO algorithm, and established quantitative models by SVR and Gaussian process regression (GPR) as the last layer of CNN. Results showed that the prediction performance of the proposed method is significantly improved by analyzing three molecular spectral data sets.

Ni et al. [21] proposed a variable-weighted convolution neural network (VW-CNN), adding an important factor block similar to an auto-encoding network before the input layer, making the network more focused on important wavelength variables, thereby improving the generalization ability of CNN. As shown in Fig. 18.25, the prediction ability of the NIR spectroscopic prediction model in the leaves of horsetail pine seedlings established by VW-CNN is significantly better than that of the traditional PLS and SVR methods, and better than that of the classic CNN model. Padarian et al. [22] used tens of thousands of soil Vis-NIR spectroscopic samples to build CNN that simultaneously predicted the physical properties of the soil, and their input was spectrogram which was obtained by using short time Fourier transform of the original spectrum to convert the spectral one-dimensional vector of 4,200 wavelength points into a 2D matrix of 51×83 . The results of this study also show that the use of CNN to process large data set samples is more advantageous than small data sets. On this basis, Ng et al. [23] fused the Vis-NIR spectra of soil with the infrared (IR) spectra, and used the two-dimensional spectral map obtained by outer-product analysis (OPA) as the input variable of CNN, obtained satisfactory quantitative analysis results, and were able to resolve the characteristic spectral interval associated with the properties to be measured.

Bjerrum et al. [24] augmented the data of the training sample (Data Augmentation) by adding random fluctuation variables (offset, multiplication, and slope) to the NIR spectra, combined with the extended multiplicative scatter correction (EMSC) method, which can effectively improve the performance of CNN model for predicting

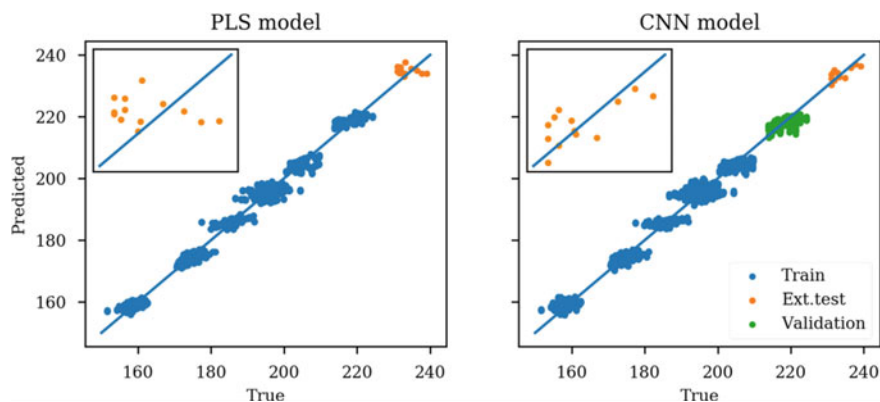


Fig. 18.26 Comparison of prediction ability of CNN model with PLS model for the external test set

the content of active components in tablets. Moreover, as shown in Fig. 18.26, CNN models have stronger extrapolation prediction abilities and better predictive consistency in the spectra between different instruments than PLS. Jernelv et al. [25] based on five vibration spectra, with sample numbers ranging from dozens to nearly a thousand, including regression and classification, compared CNN with traditional quantitative and pattern recognition methods. The results show that the influence of spectral preprocessing and variable selection methods on traditional quantitative and qualitative methods is much greater than that of CNN model, but the performance of CNN model can be improved by proper spectral preprocessing and variable selection method.

Using a GoogleNet model with a 22-layer deep network, Liu et al. [26] used hyper spectral imaging technology to quickly and non-destructively classify different varieties of peanuts, with significantly better classification results than the PLS-DA method. Based on CNN and macadamia nut of Vis-NIR spectra, Du et al. [27] established a macadamia nut quality identification model, and the identification accuracy of better nuts, worse nuts, and moldy nuts reached 100%. Lu et al. [28] proposed a convolution neural network topology for Raman spectroscopic diagnosis of hepatitis B virus (Fig. 18.27), which contains a multi-scale convolution layer that uses different

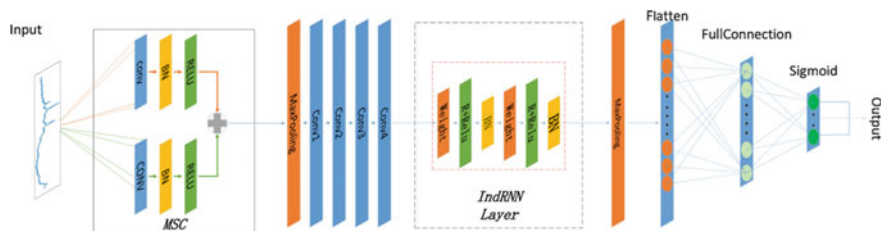


Fig. 18.27 A topological structure of CNN for Raman spectroscopic diagnosis of hepatitis B virus

convolution kernels to extract spectral features from multiple scales and then integrates them. An independent recurrent neural network layer (IndRNN) has also been added between the convolution and full-connection layers to avoid problems such as gradient vanishing and gradient explosion. Erzina et al. [29] realized the accurate detection of cancer by CNN combined with surface enhanced Raman spectroscopy. Ho et al. [30] used the CNN architecture DenseNet to identify Raman spectra of 30 common pathogens, and even in the case of high spectral noise, accurate prediction results can be obtained.

Lu et al. [31] improved the classical CNN architecture LeNet-5 for origin identification of tobacco based on NIR spectra, with a discrimination accuracy of 95%. Li et al. [32] applied CNN with NIR in identification of multi-variety and multi-manufacturer pharmaceuticals, and its classification performance is better than SVM and ELM algorithms. Zhao et al. [33] designed three Raman spectral data augmentation methods to construct Raman spectral database of three different estrogen powders. Then a CNN classification model was established. The Results indicate that the method is less affected by spectral measurement noise and has strong robustness, which is suitable for analyzing high-noise Raman spectra with more complex field measurements. Meng et al. [34] firstly used kernel principal component analysis (KPCA) to compress the NIR spectra of Chinese zither wood, and then established a classification model using CNN. The grade classification accuracy of the wood for Chinese zither panels was 95.5%. Dong et al. [35] used CNN model to feature the hyperspectral image of bacon with cross entropy as the optimization target. Multiplicative scattering correction (MSC) and PCA were used to preprocess spectra and extract spectral feature, and then the two features were fused and classified by SVM, the classification accuracy could reach 99.2%. Tuan et al. [36] used CNN combined with ELM (CNN-ELM) to establish a coal classification model first, and then used improved swarm optimization algorithm to further optimize CNN-ELM. The results show that the model has a good classification effect for coal species, and the recognition accuracy is more than 96%.

Zong et al. [37] established a CNN model for predicting total sugar, total nicotine, and chloride ions in tobacco leaves by NIR spectra, and obtained satisfactory results. Wang et al. [38] compressed the NIR spectra of soil by PCA, and then transformed it into a two-dimensional matrix by outer product and used as input variables for CNN. Compared with PLS, BP-ANN, and SVR, CNN model has more powerful prediction ability to predict soil moisture content. Tsakiridis et al. [39] pretreated the Vis-NIR spectra of soil samples by different preprocessing methods. Six different spectra can be obtained for each soil sample, forming a six-channel spectral matrix as input of CNN (Fig. 18.28). A quantitative model that can simultaneously predict multiple physicochemical properties was established. Moreover, the model was fine-tuned by using local spectral neighborhoods to perform an adaptive error-correction mechanism for prediction results where it exhibited the best performance compared with the existing methods. Based the near infrared spectra and organic carbon data of 17,272 mineral soil samples in LUCAS soil database, Shi et al. [40] used CNN with 6–7 convolution layers to extract better nonlinear features than PCA. The results show that the root mean square error of prediction for organic carbon content could reach

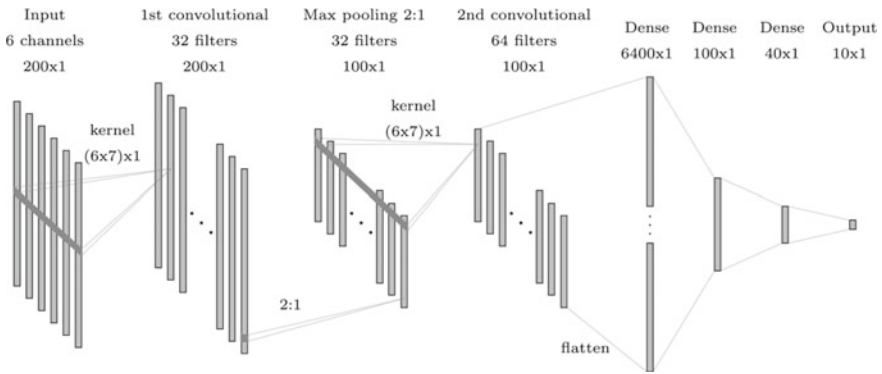


Fig. 18.28 A schematic diagram of CNN for multi-channel inputs and multi-parameter outputs

9.69 g/kg, which is more accurate than that of commonly used modeling methods. The proposed model is implemented by Keras toolbox in Python language. Zhang et al. [41] designed a CNN for one-dimensional spectral quantitative analysis based on GoogLeNet network incorporated Inception module. It can directly model the original spectra end-to-end, with better predictive accuracy than other CNN models.

Liu et al. [42] proposed a method of classifying multi-class Raman spectral data using CNNs (Fig. 18.29), which outputted thousands of classes and was actually a spectral searcher. The method has achieved excellent classification results in RUFF mineral Raman spectral database. This research also compared the influence of baseline correction methods on traditional pattern recognition and CNN. It proved that CNN can achieve end-to-end direct discrimination analysis without spectral preprocessing. Fan et al. [43] applied CNNs for component identification in Raman spectra

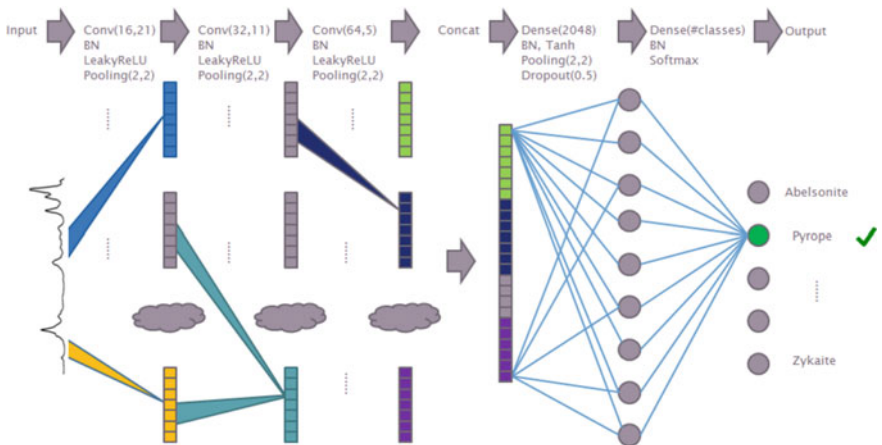


Fig. 18.29 A schematic diagram of CNN for multi-class classification of Raman spectral data

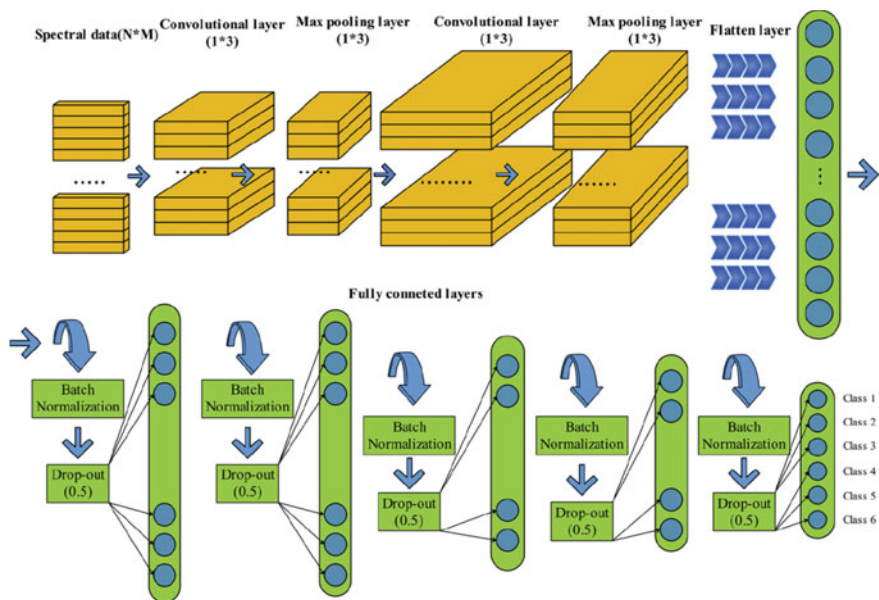


Fig. 18.30 A schematic diagram of CNN for classification of hyperspectral imaging data

of mixtures. Compared with the logistic regression (LR) with L1-regularization, kNN, RF, and BP-ANN models, the proposed method showed better identification accuracy for ternary mixture spectral data sets.

Ni et al. [44] designed a CNN for NIR hyperspectral imaging to identify six kinds of hybrid okra seeds and six hybrid loofah seeds, as shown in Fig. 18.30. This network consists of two convolution layers and five fully connected layers. Each fully connected layer uses batch normalization and dropout strategies. The results show that the advantage of CNN is more obvious for the data set with more classes compared with PLS-DA and SVM methods. Zhang et al. [45] converted the one-dimensional vector into two-dimensional spectral matrix (Fig. 18.31) by equally dividing and folding the NIR spectra of tobacco samples. Then a prediction model to identify the origin of tobacco using CNN was established, with an accuracy of 93%. Weng et al. [46] used a similar two-dimensional spectral matrix generation method for quantitative and qualitative modeling of CNN with surface enhanced Raman spectroscopy.

Tan et al. [47] constructed serial fusion spectra for the original, first derivative, second derivative spectra, which combined the advantages of original spectra containing all the feature information and derivative spectra removing interference, as shown in Fig. 18.32. The fusion spectra combined with one-dimensional CNN learning algorithm were further used to predict component content in corn samples and the satisfactory results were obtained. Shi et al. [48] proposed a hyperspectral image classification algorithm based on manifold spectral features and CNN. The method firstly used t-distribution stochastic neighbor embedding (t-SNE) algorithm

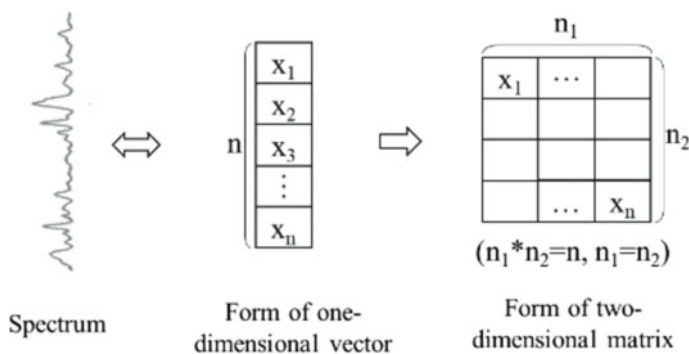


Fig. 18.31 A schematic diagram of construction two-dimensional spectral matrix from one-dimensional vector

for dimensionality reduction of hyperspectral image, and then employed CNN to extract spatial deep features. Finally, the spatial-spectral features of hyperspectral images mapped from hidden feature space to the sample marker space for classification. Zhang et al. [49] used CNN Inception to quantitatively analyze and model the chemical composition of LIBS spectra. The method does not only require the preprocessing of the original spectra but also does not need dimensionality reduction of the original spectra. The original spectral information is retained to the greatest extent, and the influence of matrix effect on quantitative results can be obviously eliminated. Lai et al. [50] established a method for quickly and accurately recognition the brand and degree of liquors by combining CNN with laser induced fluorescence technology (Fig. 18.32).

Yang et al. [51] combined CNN with recurrent neural networks (RNN) to establish an analytical model for predicting the physical properties of soils by Vis-NIR spectra, which has a better resistance toward noise and better transferability among different soil types. Fang et al. [52] used CNN for NIR spectral analysis of apple chips quality. The established model was used to predict the moisture, total sugar, and total acid

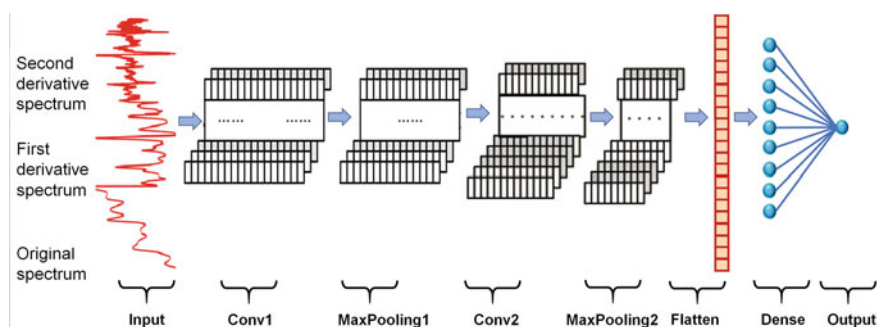
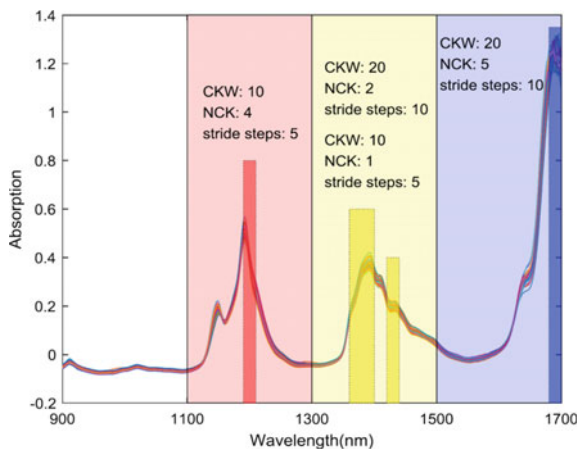


Fig. 18.32 A schematic diagram of one-dimensional CNN based on spectral fusion

Fig. 18.33 Heuristic selection results for feature mapping based on genetic algorithms



of apple chips, which had better stability and generalization ability than traditional modeling methods such as PLS, BP-ANN, LS-SVM. Weng et al. [53] fused rice spectra and shape feature variables, using CNN to establish a model to identify the high-quality rice species, with classification accuracy of 95%. Assadzadeh et al. [54] used CNN to establish a global NIR spectral calibration model that could simultaneously predict protein and moisture content in grains such as wheat, barley, field pea, and lentil, which can improve modeling efficiency and reduce model maintenance. Yang et al. [55] used CNN combined with NIR for classifying softwood species, with an accuracy of 100%. Hu et al. [56] used NIR fractional derivative spectra and CNN to model the nitrogen content level of rubber tree foliage.

Chen et al. [57] used genetic algorithms (GA) for optimized heuristic selection of CNN parameters, as shown in Fig. 18.33. In the region near the absorption peak of the original spectrum, the selection of relatively more convolution kernels, smaller convolution window width and moving step size will be more conducive to extract the key information contained in the original spectrum. However, in the non-absorption peak area of the original spectra, the model parameters can be effectively reduced and model generalization performance can be improved by using less convolution kernels, larger convolution window size and moving step size. Chen et al. [58, 59] also used ensemble modeling strategy to CNN, as shown in Fig. 18.34. Compared to a single CNN model, the generalization performance and robustness of ensemble CNN was improved. The disadvantage of ensemble CNN is that the computation and complexity of modeling will increase significantly.

18.3 Deep Belief Network

Deep belief network (DBN) is also a major framework of deep learning algorithms. It can be used for unsupervised learning, similar to an auto-coder. It also can be used for supervised learning, as a regressor or classifier.

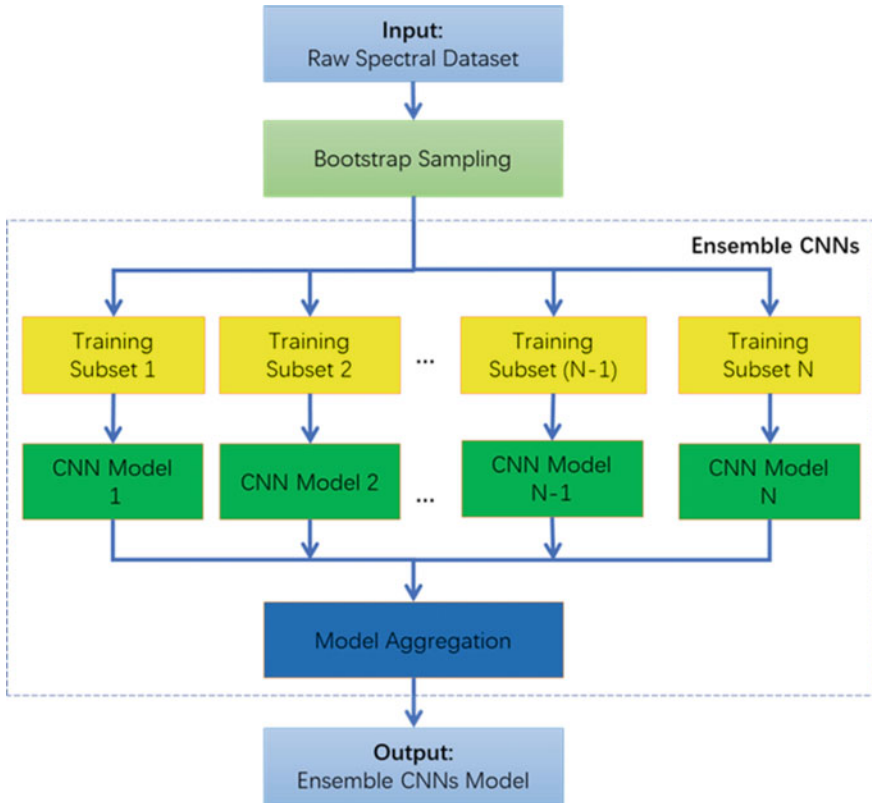


Fig. 18.34 Process of ensemble CNN model

In terms of unsupervised learning, the purpose of DBN is to retain the characteristics of the original features as much as possible and reduce the dimension of the features. In terms of supervised learning, the purpose of DBN is to make the classification or regression error as small as possible. Regardless of supervised learning or unsupervised learning, the essence of DBN is the process of feature learning, that is, how to get a better feature expression.

Restricted Boltzmann machines (RBM) are the components of DBN. As shown in Fig. 18.35, RBM has only two layers of neurons. The first layer is called visible layer and consists of visible units, which is used for the input of training data. The other layer is called the hidden layer, made up of hidden units, which is used as feature detectors. In fact, the essence of RBM is an unsupervised learning method, which can be used for dimensionality reduction, feature extraction, auto-encoders, etc. The specific algorithm of RBM can be found from relevant literature.

From the perspective of structure, DBN is composed of multi-layer unsupervised RBM and one layer supervised BP network or softmax classifier. As shown in Fig. 18.36, v is the node value of the visible layer, h is the node value of the hidden

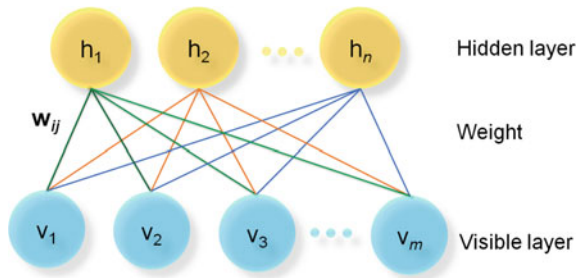


Fig. 18.35 Schematic diagram of deep belief network structure

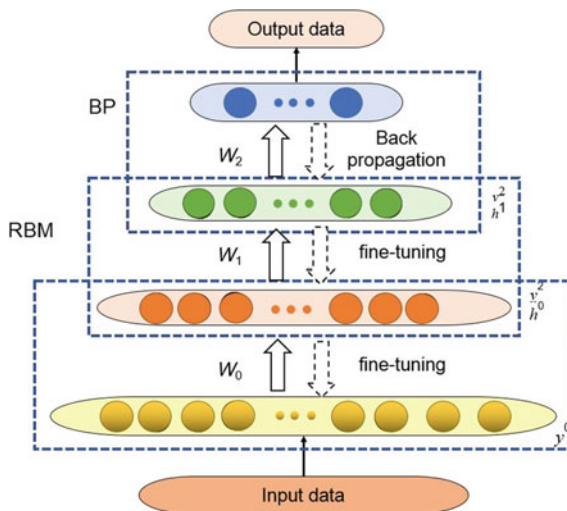


Fig. 18.36 Process of deep belief network

layer, and W represents the weight between the visible layer and the hidden layer. The original data is used as the input data of the lowest layer of RBM, which is transmitted from the bottom to the top. The feature vectors are transformed gradually from concrete to abstract. The neural network at the top forms a combination feature vector which is easier for classification. DBN is a neural network composed of multi-layer RBM.

The training process of DBN model mainly contains two stages.

(1) Pre-training stage

The DBN network parameters are initialized firstly, and then each layer of RBM network is trained unsupervised and individually layer by layer. The training result of the former layer is used as input of the next layer. The weight and offset of each

layer are retained, ensuring when the feature vector map to different feature space, feature information can be kept as much as possible.

(2) Fine-tuning stage

Firstly, forward propagation is carried out. The pre-trained parameters are assigned to each layer of neural network. The training is carried out according to the set network structure, and training values are output. Then back propagation is carried out. The actual output result of BP algorithm is compared with the expected label and then the error value is obtained. The error is back propagated layer by layer from output end to input end. The optimized parameters are constantly adjusted to minimize the error.

Thus, it can be seen that the training process of RBM network model can be regarded as the initialization of weight parameters of a deep BP network, which makes DBN overcome the shortcoming of BP network, which is easy to fall into local optimum and needs long training time due to random initialization of weight parameters.

Fu et al. [60] used the DBN as the feature extractor and random forest as the classifier to solve the problem of the high feature dimension of spectra and the lack of learning ability of the traditional shallow feature extraction method in NIR drug identification. Zhang et al. [61] proposed a DBN quantitative model construction method based on Dropout for the problems of small sample, high dimension, and non-linearity of NIR spectra. Wu et al. [62] proposed an age identification method of ancient ceramics based on DBN and Vis-NIR spectroscopy, which realized the classification of ancient ceramics of different dynasties and avoided the local optimum caused by random initialization of weight parameters by BP neural network. Huang [63] used DBN to extract features from ultraviolet (UV) absorption spectra of SO₂ gas, and then established a quantitative model by using ELM, and the problems of overlapping absorption spectra, difficult feature information extraction and insufficient extraction accuracy are solved. Zhang et al. [64] applied PLS to the DBN training process of NIR spectra. The DBN structure was improved and prediction accuracy of the model was increased.

18.4 Transfer Learning

Traditional machine learning methods can only be carried out under a common assumption: the training and test data set come from the same feature space and the same distribution. When the distribution changes, the statistical model needs to use newly collected training data to restart the training model. In many real-world applications, it takes time and effort to rebuild a model. If the model can be reused after some transformation, or knowledge transfer between the data sets, repeated modeling can be avoided (Fig. 18.37). In the field of artificial intelligence, transfer learning is a method to study the ability of model or data reuse and let the model apply the learned knowledge to new fields. Transfer learning attempts to find the

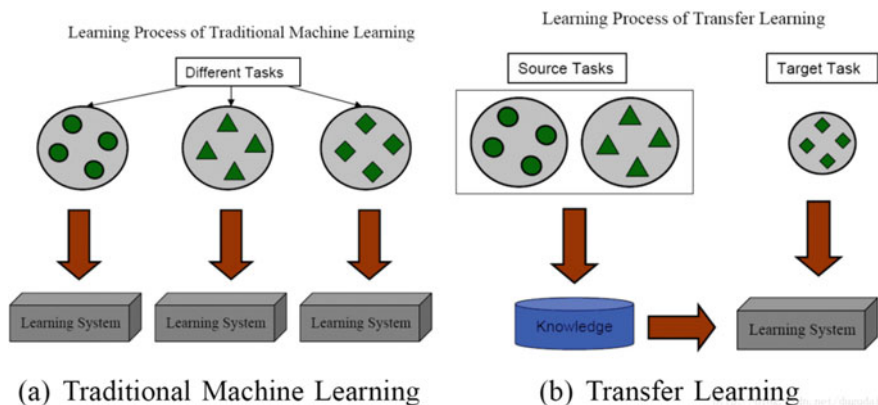


Fig. 18.37 Schematic diagram of the differences between transfer learning and traditional machine learning

common points between data sets, the relationship between model parameters and different tasks, and use old knowledge to deal with new problems. Generally, transfer learning is used in the following situations. ① There is a lack of large available data set in new task; ② There is a model trained by a large number of samples. If the target task only has very limited samples, training a new model from the beginning is prone to over-fitting. Using the pre-trained weights to train the new model can speed up convergence and improve the generalization ability of the network.

In the supervised training of deep convolution networks, a large number of labeled (or reference data) samples need to be fully trained so that the network can achieve excellent classification or regression results. However, in actual tasks, the cost of obtaining a large number of labeled (or reference data) samples is very high. Insufficient labeled samples will lead to over-fitting, and ultimately reduce the classification effect of the model in the test data set. In order to use as few labeled samples as possible and avoid over-fitting, a training strategy based on deep transfer learning method was proposed to improve the classification or regression accuracy of deep networks in the case of small amount of samples (Fig. 18.38). The method turned the

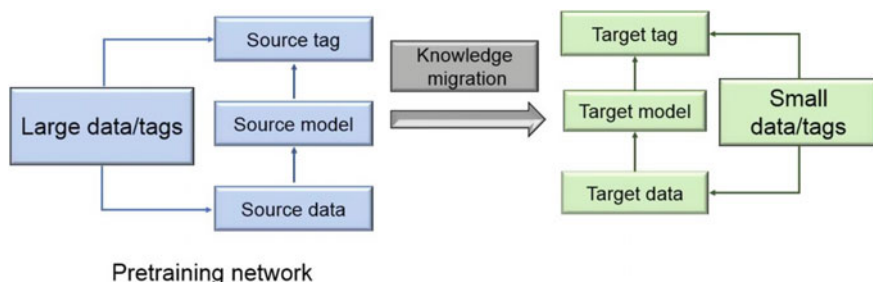


Fig. 18.38 Schematic diagram of application scenario of transfer learning

problem into using a relevant data set with sufficient labeled to pre-train the network, and then used the existing small sample data set to fine-tune the parameters of the deep convolution layer in the network to make the network learning deep features of target samples. The model combined the shallow features of the two data sets at the same time, so as to achieve better classification or regression results.

In the field of deep learning, transfer learning is a learning method that re-trains the pre-trained model and applies it to other tasks. The pre-trained network model in a large-scale data set can be used as a feature extractor in other tasks. These pre-trained models usually have consumed significant time and computational resources during development. In transfer learning, as shown in Fig. 18.39, relevant data set is called source domain, and the existing data sets that need to be classified are called target domain [65–68].

For deep convolution network model, because the network of shallow convolution kernels mainly captures shallow features such as edges and contours, which are universal and existed both in source data set and target data set samples, a large number of available source data samples can be fully pre-trained the network. After the network parameters are well trained, shallow convolution kernels will be fixed and no longer be optimized. However, the deep features extracted by the network top-level convolution layer are specific to the target data set. In order to ensure the classification or regression accuracy of the network model in the target data set, the kernel parameters of the network deep convolution layer are fine-tuned in the target data set. The parameters of the deeper network and the final parameters of the output layer are randomly initialized, and these parameters are continuously trained through a small amount of labeled data in target data set. The whole process can be regarded as the network transfers the prior knowledge learned from the source data set to the target data set, which avoids over-fitting to a certain extent and ensures the learning of unique features of the target data set.

As shown in Fig. 18.40, Lian et al. [69] carried out transfer learning on the CNN Inception-V3 model based on ImageNet data set. When the pre-trained CNN model was transferred to a small target set, the original convolution layer structure is

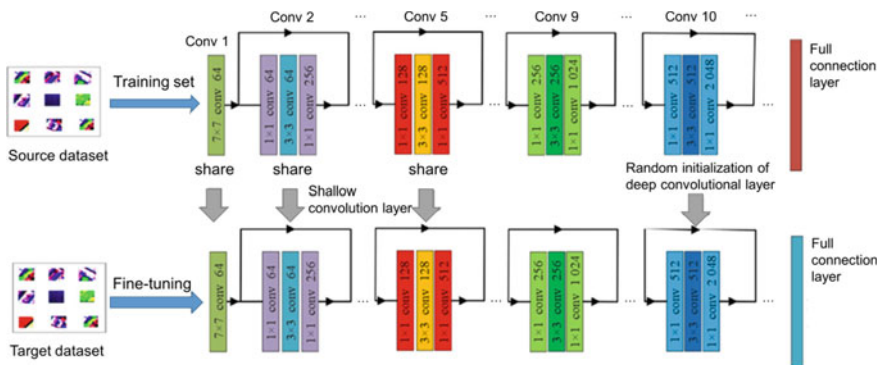


Fig. 18.39 Schematic diagram for training of transfer learning strategy

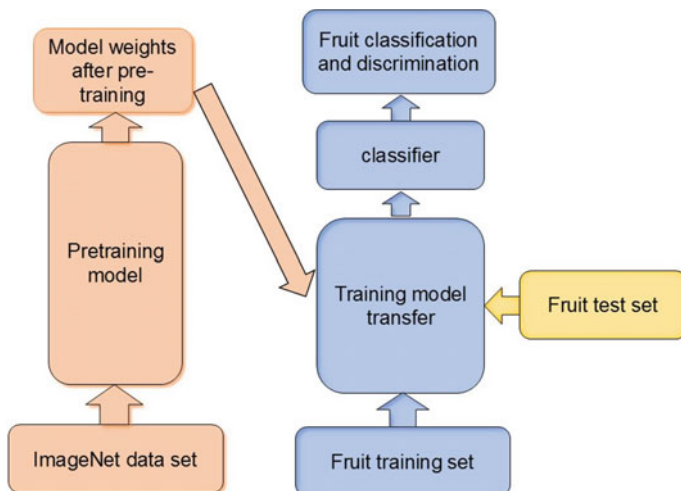


Fig. 18.40 Structure diagram of image classification for fruit based on transfer learning

retained and a new softmax classifier is built to classify data. The advantages of this model in image classification can be applied to fruit image recognition, so that the recognition of fruit images can be faster and more accurate. Because the traditional method requires rich human experience in the process of feature extraction, great uncertainty existed in feature extraction for the traditional methods. Moreover, the traditional method also has a complex parameter adjustment process, which greatly increases the training time. However, the Inception-V3 model based on transfer learning could fine-tune the parameters of deep convolution kernel for an excellent classification model, which improves the classification effect of CNN in the case of a small number of labeled samples. The fruit recognition accuracy was significantly improved in case of small number labeled samples compared with traditional fruit classification algorithm.

As a feature extractor, pre-trained deep residual network was generalized by Wang et al. [70] from the ImageNet data set into the hyperspectral classification task. Deep features of sample spatial features in hyperspectral image classification were extracted. The experiment proved that these features had stronger discriminant ability and could produce good complementarity with the original spectral features. The results show that the adequately trained deep convolution network on the common image data set is helpful for the hyperspectral classification task. By fine-tuning network high-order layer convolution kernel parameters through the target data set, the model could achieve better classification accuracy even if with small number of labeled samples.

Liu et al. [71] established a CNN model for predicting clay content in soil by using high-quality and large-scale soil Vis-NIR spectra obtained in the laboratory and their corresponding reference values. On this basis, through a small number of

field hyperspectral samples combined with the transfer learning strategy, the pre-trained CNN was transferred and applied for the prediction analysis of field hyper spectra, and satisfactory results were obtained. Padarian et al. [72] built a local model based on the near-infrared spectra and convolution neural network by more than 20,000 global soils through the transfer learning strategy, which is used for the prediction and analysis of local soil samples. Due to full use of big data pre-trained results, the proposed method obtained better results than individual model built by local samples.

Kraub et al. [73] applied the pre-trained AlexNet network based on ImageNet database to the recognition of cancer cells by confocal Raman microscopy through the transfer learning strategy, which significantly saved the training time of the network and improved the recognition accuracy. Sun et al. [74] constructed chemical images with two-dimensional correlation spectra and transferred image recognition model GoogLeNet through transfer learning method. The proposed method was used for NIR spectral classification and recognition of cashmere fabrics and cashmere/wool blended fabrics, as well as pure cotton and mercerized cotton fabrics, achieving high precision recognition of fabrics.

References

1. Lei M. Machine learning principles, algorithms and applications. Beijing: Tsinghua University Press; 2019.
2. Zhang WD, Lu HX, Gan BR, et al. Drug identification based on stacked auto encoders fusing extreme learning machine. *Comput Eng Design*. 2019;40(2):545–60.
3. Lu HX, Wei MM, Yang HH, et al. Detecting Huanglongbing by stacked denoising auto-encoders combined random forest. *Laser Infrared Sens*. 2019;49(9):460–6.
4. Liu T, Li ZR, Yu CX, et al. NIRS feature extraction based on deep auto-encoder neural network. *Infrared Phys Technol*. 2017;87:124–8.
5. Hang YY, Li YT, Sun MJ. Classification of radish seeds using hyperspectral imaging and deep learning method. *Agric Eng*. 2020;10(5):29–33.
6. Wan W. Research and application on process detection of solid-state fermentation of bioethanol using near-infrared spectroscopy (NIRS) technique. ZhenJiang: Jiangsu University; 2018.
7. Yu XJ, Lu HD, Wu D. Development of deep learning method for predicting firmness and soluble solid content of Postharvest Korla fragrant pear using Vis/NIR hyperspectral reflectance imaging. *Postharvest Biol Technol*. 2018;141:39–49.
8. Yu XJ, Lu HD, Liu QY. Deep-learning-based regression model and hyperspectral imaging for rapid detection of nitrogen concentration in oilseed rape (*Brassica Napus L.*) Leaf. *Chemometr Intell Labor Syst*. 2018; 172:188–93.
9. Yu XJ, Wang JP, Wen ST, et al. A deep learning based feature extraction method on hyperspectral images for nondestructive prediction of TVB-N content in pacific white shrimp (*Litopenaeus Vannamei*). *Biosys Eng*. 2019;178:244–55.
10. Ran S, Ding JL, Ge XY et al. Estimation method of VIS-NIR spectroscopy for soil organic matter based on sparse networks. *Laser Optron Progr*. 2020;57(21):212802.
11. Ni C, Zhang Y, Gao HD. Prediction model of moisture content of Masson Pine Roots based on near infrared spectroscopy. *J Nanjing Forestry Univ (Nat Sci Ed)*. 2019;43(6):91–6.
12. Tian QC, Wang ML. Research progress on deep learning. *Algor Comput Eng Appl*. 2019;55(22):25–33.

13. Zhang M, Shanxia LY. Illustrated guide to deep learning. Beijing: The Posts & Telecom Press; 2018.
14. Yang J, Xu JF, Zhang XL, et al. Deep learning for vibrational spectral analysis: recent progress and a practical guide. *Anal Chim Acta*. 2019;1081:6–17.
15. Chen YY, Wang ZB. End-to-end quantitative analysis modeling of near-infrared spectroscopy based on convolutional neural network. *J Chemom*. 2019;33:e3122.
16. Yang RLZ, Yongjing LX, Yongjing ZM. Mathematics in deep learning. Beijing: The Posts & Telecom Press; 2019.
17. Acquarelli J, Van Laarhoven T, Gerretzen J, et al. Convolutional neural networks for vibrational spectroscopic data analysis. *Anal Chim Acta*. 2017;954:22–31.
18. Le BT. Application of deep learning and near infrared spectroscopy in cereal analysis. *Vibrat Spectr*. 2020;106:103009.
19. Cui CH, Fearn T. Modern practical convolutional neural networks for multivariate regression: applications to NIR calibration. *Chemom Intell Lab Syst*. 2018;182:9–20.
20. Malek S, Melgani F, Bazi Y. One-dimensional convolutional neural networks for spectroscopic signal regression. *J Chemom*. 2018;32:e2977.
21. Ni C, Wang D, Tao Y. Variable weighted convolutional neural network for the nitrogen content quantization of masson pine seedling leaves with near infrared spectroscopy. *Spectrochim Acta Part A Mol Biomol Spectrosc*. 2019;209:32–9.
22. Padarian J, Minasny B, McBratney AB. Using deep learning to predict soil properties from regional spectral data. *Geoderma Region*. 2019;16:e00198.
23. Ng W, Minasny B, Montazerolghaem M, et al. Convolutional neural network for simultaneous prediction of several soil properties using visible/near-infrared, mid-infrared, and their combined spectra. *Geoderma*. 2019;352:251–67.
24. Bjerrum EJ, Glahder M, Skov T. Data augmentation of spectral data for convolutional neural network (CNN) based deep chemometrics. 2017. arXiv:171001927.
25. Jernelv IL, Hjelme DR, Aksnes A, et al. Convolutional neural networks for classification and regression analysis of one-dimensional spectral data. 2020. arXiv: 2005.07530.
26. Liu CL, Lin L, Yu CC, et al. Research on peanut hyperspectral image classification method based on deep learning. *Comput Simul*. 2020;3:189–92.
27. Du J, Hu BL, Liu YZ, et al. Study on quality identification of Macadamianut based on convolutional neural networks and spectral feature. *Spectr Spect Anal*. 2018;38(5):1514–9.
28. Lu HC, Tian SW, Yu L, et al. Diagnosis of hepatitis b based on raman spectroscopy combined with a multiscale convolutional neural network. *Vibrat Spectr*. 2020;107:103038.
29. Erzina M, Trelin A, Guselnikova O, et al. Precise cancer detection via the combination of functionalized SERS surfaces and convolutional neural network with independent inputs. *Sens Actuat B: Chem*. 2020;308:127660.
30. Ho CS, Jean N, Hogan CA, et al. Rapid identification of pathogenic bacteria using Raman spectroscopy and deep learning. *Nat Commun*. 2019;10:4927.
31. Lu MY, Yang K, Song PF, et al. The study of classification modeling method for near infrared spectroscopy of tobacco leaves based on convolution neural network. *Spectrosc Spect Anal*. 2018;38(12):78–82.
32. Li LQ, Pan XP, Feng XC, et al. Deep convolution network application in identification of multi-variety and multi-manufacturer pharmaceutical. *Spectrosc Spect Anal*. 2019;39(11):3606–13.
33. Zhao Y, Rong K, Tan AL. Qualitative analysis method for raman spectroscopy of Estrogen based on one -dimensional convolutional neural network. *Spectrosc Spect Anal*. 2019;39(12):3755–60.
34. Meng SY, Huang YL, Zhao P, et al. Wood quality of chinese zither panels based on convolutional neural network and near-infrared spectroscopy. *Spectrosc Spect Anal*. 2020;40(1):284–9.
35. Dong XD, Guo PY, Xu P, et al. Fusing hyperspectral features and image deep features for classification and retrieval of meat. *Food Ind Sci Technol*. 2018;39(23):261–6.
36. Tuan LB, Xiao D, Mao YC, et al. Coal classification based on visible, near infrared spectroscopy and CNN-ELM algorithm. *Spectrosc Spect Anal*. 2018;38(7):2107–12.

37. Zong QQ, Ding XQ, Han F, et al. Study on near infrared spectroscopy model of tobacco leaves based on regression CNN. *Comput Digit Eng.* 2019;47(2):275–80.
38. Wang Z, Wu XH, Li LQ, et al. Convolutional neural network application in prediction of soil moisture content. *Spectrosc Spect Anal.* 2018;39(1):36–41.
39. Tsakiridis NL, Keramaris KD, Theocharis JB, et al. Simultaneous prediction of soil properties from VNIR-Swir spectra using a localized multi-channel 1-D convolutional neural network. *Geoderma.* 367:114208.
40. Shi Y, Wang RJ, Wang YB. Soil organic carbon prediction based on convolutional neural networks and near infrared spectroscopy. *Comput Appl Softw.* 2018;35(10):147–52.
41. Zhang XL, Lin T, Xu JF, et al. Deep spectra: an end-to-end deep learning approach for quantitative spectral analysis. *Anal Chim Acta.* 2019;1058:48–57.
42. Liu JC, Osadchy M, Ashton L, et al. Deep convolutional neural networks for Raman spectrum recognition: a unified solution. *Analyst.* 2017;142:4067–74.
43. Fan XQ, Ming W, Zeng HT, et al. Deep learning-based component identification for the Raman spectra of mixtures. *Analyst.* 2019;144:1789–98.
44. Nie PC, Zhang JN, Feng XP, et al. Classification of hybrid seeds using near-infrared hyperspectral imaging technology combined with deep learning. *Sens Actuat: B. Chem.* 2019;296:126630.
45. Zhang L, Ding XQ, Hou RC. classification modeling method for near-infrared spectroscopy of tobacco based on multimodal convolution neural networks. *J Anal Methods Chem.* 2020;2020(22):1–13.
46. Weng SZ, Yuan HC, Zhang XY, et al. Deep learning networks for the recognition and quantitation of surface-enhanced raman spectroscopy. *Analyst.* 2020;145(14):4827–35.
47. Tan AL, Wang XS, Chu ZY, et al. Research on quantitative modeling method of maize composition based on near infrared spectrum fusion and deep learning. *Food Ferment Ind.* 2020.
48. Shi Y, Ma DH, Lv J, et al. Hyperspectral image classification based on manifold spectral dimensionality reduction and deep learning method. *Trans Chinese Soc Agricult Eng.* 2020;36(6):151–60.
49. Zhang LH, Zhang L, Wu ZC, et al. Quantitative modeling for earth sample's LIBSS spectra of curiosity rover based on inception network. *Acta Photon Sin.* 2020;49(6):0630002.
50. Lai WH, Zhou MR, Wang Y, et al. Application of counterfeit liquor recognition of counterfeit liquor recognition based on deep learning and laser induced fluorescence. *Laser Optoelectron Progr.* 2018;55(4):388–94.
51. Yang JC, Wang XL, Wang RH, et al. Combination of convolutional neural networks and recurrent neural networks for predicting soil properties using Vis-NIR spectroscopy. *Geoderma.* 2020;380:114616.
52. Fang MM, Liu J. Evaluation method of apple chips quality by near infrared spectroscopy based on regressive convolutional neural network. *Food Sci Technol.* 2020;45(7):303–8.
53. Weng SZ, Tang PP, Zhang XY, et al. Non-destructive identification method of famous rice based on image and spectral features of hyperspectral imaging with convolutional neural network. *Spectrosc Spect Anal.* 2020;40(9):2826–33.
54. Assadzadeh S, Walker CK, McDonald LS, et al. Multi-task deep learning of near infrared spectra for improved grain quality trait predictions. *J Near Infrared Spectrosc.* 2020;28(5–6):275–86.
55. Yang SY, Kwon O, Park Y, et al. Application of neural networks for classifying softwood species using near infrared spectroscopy. *J Near Infrared Spectrosc.* 2020;28(5–6):298–307.
56. Hu WF, Tang RN, Li C, et al. Fractional order modeling and recognition of nitrogen content level of rubber tree foliage. *J Near Infrared Spectrosc.* 2020;28.
57. Chen YY, Wang ZB. Feature selection based convolutional neural network pruning and its application in calibration modeling for NIR spectroscopy. *Chemom Intell Lab Syst.* 2019;191:103–8.
58. Chen YY, Wang ZB. Quantitative analysis modeling of infrared spectroscopy based on ensemble convolutional neural networks. *Chemom Intell Lab Syst.* 2018;181:1–10.

59. Yi L, Lu J, Ding JL, et al. soft sensor modeling for fraction yield of crude oil based on ensemble deep learning. *Chemom Intell Labor Syst.* 2020;204:104087.
60. Fu WF, Yang HH, Liu ZB, et al. Medicine identifying method based on deep belief network and random forests. *Comput Integr Manuf Syst.* 2018;35(4):325–30.
61. Zhang R, Ding XQ, Gao ZX, et al. Study on near infrared spectral model of tobacco leaves based on dropout deep belief network. *Comput Digit Eng.* 2019;47(2):383–7.
62. Wu XP, Guan YP, Li WD, et al. Visible-near infrared spectroscopy based chronological classification and identification of ancient ceramic. *Spectr Spect Anal.* 2019;39(3):756–64.
63. Huang H, Lan HY, Huang YB. A detection method of SO₂ concentration based on DBN and ELM. *J Atmos Environ Opt.* 2020;15(3):207–16.
64. Zhang M, Zhao ZG. Near infrared spectral analysis modeling method based on deep belief network. *Spectr Spect Anal.* 2020;40(8):2512–7.
65. Liu XB, Yin X, Liu HB, et al. Classification of hyperspectral remote sensing image based on deep transfer learning: a review. *J Qingdao Univ Sci Technol (Nat Sci Ed).* 2019;40(3):1–11.
66. Yue XJ, Ling KJ, Wang LH, et al. Inversion of potassium content for citrus leaves based on hyperspectral and deep transfer learning. *Trans Chinese Soc Agricult Mach.* 2019;50(3):193–202.
67. Mozaffari MH, Tay LL. A review of 1D convolutional neural networks toward unknown substance identification in portable Raman spectrometer. 2020. arXiv:2006.10575.
68. Tan K, Wang X, Du PJ. Research progress of the remote sensing classification combining deep learning and semi-supervised learning. *J Image Graph.* 2019;24(11):1823–41.
69. Lian XQ, Cheng KY, An F, et al. Fruit image classification based on deep learning and transfer learning. *Meas Control Technol.* 2019;38(6):15–8.
70. Wang LW, Li JM, Zhou GM, et al. Application of deep transfer learning in hyperspectral image classification. *Comput Eng Appl.* 2019;55(5):187–92.
71. Liu L, Ji M, Buchroithner M. Transfer learning for soil spectroscopy based on convolutional neural networks and its application in soil clay content mapping using hyperspectral imagery. *Sensors.* 2018; 3168–70.
72. Padarian J, Minasny B, Mcbratney AB. Transfer learning to localise a continental soil vis-NIR calibration model. *Geoderma.* 2019;340:279–88.
73. Kraub SD, Roy R, Yosef HK, et al. Hierarchical deep convolutional neural networks combine spectral and spatial information for highly accurate raman-microscopy-based cytopathology. *J Biophotonics.* 2018;11(10):1–12.
74. Sun XT, Yuan HF, Song CF. Study on pattern recognition method using ‘dynamic’ NIR spectroscopy with deep learning-based image identification and transfer learning. *J Instrum Anal.* 2020;38(10):1247–53.

Chapter 19

Chemometrics Software and Toolkits



19.1 Introduction

So far, there are many types of chemometric methods used in spectral analysis. For spectral analysts, it is relatively easy to master the basic principles of these methods, but turning these algorithms into applications requires proficiency in mathematics, statistics, and advanced programming skills. The development of chemometrics software and toolkits plays a very crucial role in the popularization and application of analysis techniques such as spectroscopy combined with chemometrics. Mastering this software can solve most of the problems in practical applications. Spectrometer hardware and software (mainly including spectrum acquisition software and chemometrics software) constitute the technical platform of modern spectroscopic analysis. The above chapters of this book have given a detailed introduction of the common chemometrics involved in modern spectroscopy techniques and their latest developments. The following in this chapter mainly introduces the basic structure, functions, and commercial software and toolkits of chemometrics software.

19.2 Basic Structure and Functions of Software

The chemometric software used for spectral analysis is mainly to establish calibration models and predict unknown samples. As shown in Fig. 19.1, in terms of structure, this type of software usually consists of three parts: sample set managing, calibration, and blind sample prediction. Sample set managing is to stack the spectral data and reference data into a matrix to form a sample set file that can be used for model establishment and validation. Calibration refers to the establishment of a quantitative or qualitative calibration model. Commonly used chemometric algorithms such as

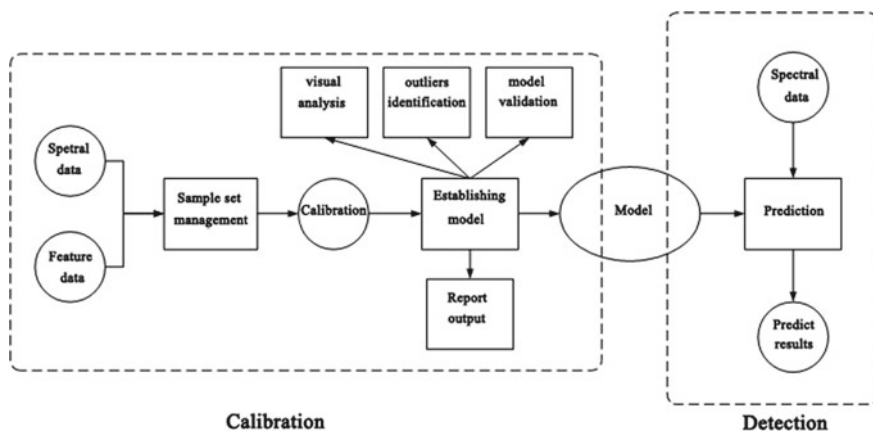


Fig. 19.1 Scheme of chemometrics software for spectral analysis

spectral preprocessing algorithms, multivariate calibration, and classification algorithms are all concentrated in this module. Blind sample prediction is to use the built model to calculate the concentration or property data of the unknown sample.

(1) Samples Managing

The main function of calibration set managing is to stack the spectra of a group of samples and reference data into a matrix to form a database. Thus, the sample set managing should be able to identify and call common spectral file formats, and input reference data in different ways. Calibration set managing usually also has the function of selecting samples to form a representative calibration set and validation set. Moreover, the real-time spectra and spatial distribution diagram of the sample can be displayed on this interface to determine extremely outlier spectra, and the concentration value of the sample can be statistically analyzed. Calibration set managing is supposed to be an open interface, and be easy to add and delete samples.

(2) Calibration Establishment

The function of establishing a calibration model is the core function of chemometrics software, which is divided into two types: establishing the qualitative and quantitative model. Both types include three steps: spectral preprocessing, spectral range selection, and method selection. After establishment, the model should be evaluated and optimized by visual operation.

Commonly used spectral preprocessing algorithms include baseline correction (first and second derivatives, subtraction), smoothing, multiplicative scatter correction, standard normalization of vector, standardization, centralization, etc. Commonly used quantitative calibration algorithms usually include MLR, PCR, PLS, SVR, ANN, etc. Qualitative algorithms mainly include cluster analysis, KNN,

SIMCA, etc. Spectral range or interval selection generally adopts a visual interactive mode, which can be directly conducted on the spectra with the mouse or can be automatically selected by parameters such as correlation coefficients.

View analysis after modeling is very important for judging whether the model is acceptable or not and removing outliers, generally including PRESS diagram, regression curve, spectral residual distribution, score and loading diagram, etc. At the same time, the evaluation results such as SEC, SECV, and R^2 should be observable. According to ASTM E1655, three types of outliers in the calibration set, such as Mahalanobis distance outliers, property residual outliers, and spectral residual outliers, should be eliminated during modeling. Therefore, the software needs to provide corresponding view analysis functions.

External validation is the main way to test whether the model is reasonable. Model validation can provide multiple statistical parameters (such as RMSEP, RPD, t -test, etc.), as well as the comparison of measured and predicted values so as to evaluate the pros and cons of the model.

Some software has the function of the automatic output of modeling parameters, such as spectral preprocessing parameters, PLS main factors, spectral interval, etc. Generally, this function is only for reference, the final model parameters still need to be determined by the users based on the necessary chemical knowledge.

(3) Prediction

The main function of the predictive module is to perform predictive analysis on the unknown samples. As shown in Fig. 19.2, when calculating, the spectra of the unknown sample are first preprocessed by the saved preprocessing parameters, and

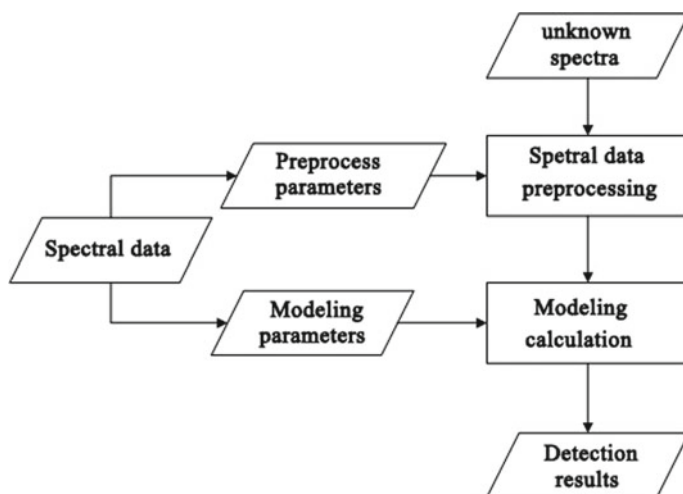


Fig. 19.2 Basic steps of predictive analysis of unknown samples

then the calibration method and setup parameters are run for calculation. Quantitative models generally need to determine whether unknown samples are within the model range, such as Mahalanobis distance, spectral residuals, and the nearest-neighbor distances. Prediction results are usually displayed directly or output to the corresponding file in the form of a report.

19.3 Common Software and Toolkits

Nowadays, almost all large-scale spectrometer manufacturers, especially near-infrared spectroscopy suppliers, have developed dedicated chemometric software, such as FOSS WinISI, Thermo TQ Analyst, Bruker OPUS, Metrohm Vision, Buchi NIRCcal, etc.

Some chemometric calculation software includes the Unscrambler of Norway Camo, Solo of Eigenvector Research of the U.S., and the PLS_Toolbox developed based on Matlab, Pirouette of InfoMatrix of the U.S., and the SIMCA MVDA of Sartorius of Germany, etc. There is also chemometrics software developed by some universities, such as the ParLeS software of the University of Sydney, Australia [1], Caunir of China Agricultural University, RIPP software of SINOPEC Research Institute of Petroleum Processing, etc.

Commercial chemometrics software can solve most of the problems encountered in daily analysis, and plays an important role in the popularization and application of modern spectroscopic technology. However, the updates of commercial software would be relatively slow. As well, the improvement of new algorithms or classic algorithms sometimes requires users' programming. The commercialization of MATLAB, R, and Python significantly provides great convenience for the program implementation of chemometric algorithms. There have been many commercial or open access chemometrics software and toolkits, such as the PLS Toolbox based on MATLAB, the mdatools based on R language [2], the scikit-learn toolkit based on Python [3], etc.

MATLAB software comes with many toolboxes that can be directly or slightly modified for spectral analysis, such as statistics and machine learning toolbox, wavelet toolbox, neural network toolbox, deep learning toolbox, global optimization toolbox, optimization toolbox, etc.

Table 19.1 is some MATLAB toolbox and open source code of certain algorithms written by chemometrics researchers [4–13]. The emergence of these toolboxes has greatly promoted the application research of new algorithms in chemometrics [21, 22].

Table 19.1 Some MATLAB toolboxes that can be used for chemometrics

Names	Resources	Directions
SAISIR	http://www.chimietrie.fr/sai-sirdownload.html	Complete chemometrics toolbox [4]
ChemoAC	http://minf.vub.ac.be/~fabi/research/chemoac	Complete chemometrics toolbox [5]
Pre-screen	https://www.cpact.com/	Data preprocessing and multivariable process control toolbox [6]
TOMCAT	http://www.chemometria.us.edu.pl/RobustToolbox/	Robust multivariate correction algorithm toolbox [7]
SPA toolbox	http://www.ele.ita.br/~kawakami/spa	Successive projection algorithm selection feature variable toolbox [8]
Multiblock_toolbox	https://github.com/puneetmisra2/Multi-block	Multi-block data analysis toolbox [9]
PO/SO-PLS	https://nofimamodeling.org/software-downloads-list/	Sequential orthogonal PLS and parallel orthogonal PLS toolbox for multi-block analysis [10–12]
VSN	https://www.chem.uniroma1.it/romechemometrics/research/algorithms/	Weighted normal variable transformation toolbox
PLS-genetic algorithm toolbox	http://models.life.ku.dk/algorithms	Genetic algorithm PLS method toolbox
N-way Toolbox	http://models.life.ku.dk/algorithms	Multi-dimensional data processing method toolbox
iToolbox	http://models.life.ku.dk/algorithms	PLS-based feature variable selection toolbox
MCR-ALS toolbox	https://mcrals.wordpress.com/download/mcr-als-toolbox/	Multivariate curve resolution-alternating least squares toolbox [13–15]
FastICA	http://research.ics.aalto.fi/ica/fastica/	Independent component analysis (ICA) toolbox
ELM	https://personal.ntu.edu.sg/egbhuang/elm_kernel.html	Extreme learning machine (ELM) toolbox
libPLS	http://www.libpls.net/	Variable selection (CARS, MWPLS, IRIV, etc.) toolbox [16]
Gaussian processes	http://gaussianprocess.org/gpml/code/matlab/doc/index.html	Gaussian process regression toolbox
MATLAB toolbox for dimensionality reduction	http://homepage.tudelft.nl/19j49/Matlab_Toolbox_for_Dimensionality_Reduction.html	Data dimensionality reduction method toolbox

(continued)

Table 19.1 (continued)

Names	Resources	Directions
LibSVM	https://www.csie.ntu.edu.tw/~cjlin/libsvm/	Support vector machine toolbox
Pattern recognition and machine learning in MATLAB	https://github.com/covartech/PRT	Pattern recognition and machine learning toolbox
Data-driven SIMCA tool	https://github.com/yzontov/dd-simca	Data-driven SIMCA toolbox [17]
IRootLab toolbox	http://trevisanj.github.io/irootlab/	Vibration biological spectroscopy data analysis toolbox [18]
LS-SVM	https://www.esat.kuleuven.be/sista/lssvmlab/	Least squares support vector machine toolbox
Classification toolbox	https://micchem.unimib.it/download/matlab-toolboxes/	Supervised pattern recognition toolbox
FRUITNIR	https://github.com/puneetmisra2/FRUITNIR	Migration component analysis toolbox [19]
MEDA-toolbox	https://github.com/josecamacho/MEDA-Toolbox	Big data chemometrics toolbox [20]
Cluster toolbox	https://github.com/Biospec/cluster-toolbox-v2.0	Latent structure orthogonal projection (OPLS), multi-level simultaneous component analysis (MSCA) toolbox
Sparse projection pursuit analysis	https://github.com/S-Driscoll/SparseProjectionPursuit	Projection pursuit analysis toolbox
Peak fit toolbox	https://github.com/heriantolim/PeakFit	Spectral peak fitting toolbox
MVC3 graphical interface	http://www.iquir-conicet.gov.ar/descargas/mvc3.rar	Multi-dimensional data processing method toolbox

References

1. Rossel RAV. ParLeS: software for chemometric analysis of spectroscopic data. *Chemom Intell Lab Syst.* 2008;90(1):72–83.
2. Kucheryavskiy S. mdatools – R package for chemometrics. *Chemom Intell Labor Syst.* 2020;198:103937.
3. Torniaainen J, Afara IO, Prakash M, et al. Open-source python module for automated preprocessing of near infrared spectroscopic data. *Anal Chim Acta.* 2020;1108:1–9.
4. Cordella C, Bertrand D. SAISIR: a new general chemometric toolbox. *Trends Anal Chem.* 2014;54:75–82.
5. Vandeginste B, Smeyers-Verbeke J. ChemoAC: its contribution to the advancement of chemometrics. *J Chemom.* 2007;21:257–62.
6. Yi G, Herdsman C, Morris J. A MATLAB toolbox for data pre-processing and multivariate statistical process control. *Chemom Intell Laborat Syst.* 2019;194:103863
7. Daszykowski M, Serneels S, Kaczmarek K, et al. TOMCAT: a MATLAB toolbox for multivariate calibration techniques. *Chemom Intell Lab Syst.* 2007;85:269–77.

8. Paiva HM, Soares SF, Galvao RK, et al. A graphical user interface for variable selection employing the successive projections algorithm. *Chemom Intell Lab Syst.* 2012;116:260–6.
9. Mishra P, Roger JM, Rutledge DN, et al. MBA-GUI: a chemometric graphical user interface for multi-block data visualisation, regression, classification, variable selection and automated pre-processing. *Chemom Intell Labor Syst.* 2020;205:104139.
10. Næs T, Mage I, Segtnan V. Incorporating interactions in multi-block sequential and orthogonalised partial least squares regression. *J Chemom.* 2011;25(11):601–9.
11. Mage I, Menichelli E, Næs T. Preference mapping by PO-PLS: separating common and unique information in several data blocks. *Food Qual Prefer.* 2012;24(1):8–16.
12. Biancolillo A, Mage I, Næs T. Combining SO-PLS and linear discriminant analysis for multi-block classification. *Chemom Intell Lab Syst.* 2015;141:58–67.
13. Jaumot J, Gargallo R, Juan A, et al. A graphical user-friendly interface for MCR-ALS: a new tool for multivariate curve resolution in MATLAB. *Chemom Intell Lab Syst.* 2005;76:101–10.
14. Jaumot J, Juan AD, Tauler R. MCR-ALS GUI 2.0: new features and applications. *Chemom Intell Labor Syst.* 2015;140:1–12.
15. Juan AD, Tauler R. Multivariate curve resolution: 50 years addressing the mixture analysis problem - a review. *Anal Chim Acta.* 2021;1145:59–78.
16. Li HD, Xu QS, Liang YZ. libPLS: an integrated library for partial least squares regression and linear discriminant analysis. *Chemom Intell Lab Syst.* 2018;176:34–43.
17. Zontov YV, Rodionova OY, Kucheryavskiy SV, et al. DD-SIMCA-A MATLAB GUI tool for data driven SIMCA approach. *Chemom Intell Lab Syst.* 2017;167:23–8.
18. Trevisan J, Angelov PP, Scott AD, et al. IRootLab: a free and open-source MATLAB toolbox for vibrational biospectroscopy data analysis. *Bioinformatics.* 2013;29(8):1095–7.
19. Mishra P, Roger JM, Marini F, et al. FRUITNIR-GUI: a graphical user interface for correcting external influences in multi-batch near infrared experiments related to fruit quality prediction. *Postharv Biol Technol.* 2020;174:111414
20. Tortorella S, Servili M, Toschi TG, et al. Subspace discriminant index to expedite exploration of multi-class omics data. *Chemom Intell Labor Syst.* 2020;206:104160
21. Morais CLM, Lima KMG, Singh M, et al. Tutorial: multivariate classification for vibrational spectroscopy in biological samples. *Nat Protoc.* 2020;15:2143–62.
22. Yang QX, Zhang LX, Wang LX, et al. MultiDA: chemometric software for multivariate data analysis based on Matlab. *Chemom Intell Lab Syst.* 2012;116:1–8.

Chapter 20

Discussion of Some Issues



20.1 Comparison of Different Spectroscopic Analysis

In terms of chemical information, different spectra contain the same or similar information about molecular functional groups. For example, although the NIR, MIR, Raman, and Terahertz spectra are produced by different mechanisms, they are all generated by the interaction of molecular vibrations with electromagnetic radiation and mainly reflect information about the vibrational energy transition of chemical bonds in molecules [1, 2], known as the “four sisters of vibrational spectroscopy”. At present, these four spectroscopic methods combined with chemometrics are practiced in various fields. Each of the four analytical techniques has its own characteristics, and if investment costs are not taken into account, they can be substituted for each other in many cases, but in some cases, one technique seems to be the only choice [3].

Thus, a brief summary of their technical characteristics is given below:

- (1) Compared to IR and Raman spectroscopy, NIRS is less finger-printable and less sensitive. It is weakly selective for molecular structure, and it is difficult to find independent characteristic absorption peaks of functional groups. Therefore, NIRS is rarely used for structural identification of molecules in the laboratory. However, the NIRS of different substances also differs significantly, with strong fingerprinting, and these differences can be used in combination with chemometric methods for rapid identification and analysis of substances, for example, for the rapid screening of incoming raw materials for pharmaceutical plants [4]. NIRS is also widely used for in situ forensic analysis due to its strong advantages such as high signal-to-noise ratio (strong light source energy, high detector sensitivity), low cost (common optical materials), robustness, adaptability to the environment, and simplicity and flexibility of measurement methods.

The current status and acceptance of NIRS for rapid laboratory and in situ online analysis of complex substances (e.g., oils and agricultural products), especially in large process industries such as petrochemicals, pharmaceuticals, and foodstuffs, are

difficult to replace by the other two techniques, reflecting the strong practicality of NIRS.

However, NIRS is not suitable for the analysis of trace substances. Further, due to the heavy reliance on calibration models and the lack of distinctive spectral features, NIRS is not as advantageous in areas of research such as reaction dynamics. The exception is functional NIRS for cutting-edge research in brain science.

- (2) Currently, of the three types of vibrational spectra, MIRS has the highest popularity, the most complete library of MIRS, and its fingerprint is relatively strong, so it has been playing an important role in laboratory structural identification, especially in the characterization of polar functional groups (e.g., carbonyl groups). In combination with chemometric methods and ATR measurements, MIRS has also made great progress in recent years in the rapid qualitative and quantitative analysis of complex samples, such as the authentication of food products, the determination of biodiesel blending ratios, and the monitoring and analysis of the quality of in-service lubricants [5].

However, for online process analysis, the long-distance transmission of MIR light is limited by optical fiber materials (usually <10 m) and is currently mostly used for laboratory studies of reaction processes, and less often for actual industrial production. However, MIRS has certain advantages over near-infrared spectroscopy for gas detection [6].

- (3) A significant advantage of Raman spectroscopy over MIRS is that it is not afraid of water, as Raman spectroscopy measures information about the fundamental frequency vibrations of molecules in the ultraviolet, visible, and near-infrared spectral regions. Among the three types of vibrational spectroscopy, Raman spectroscopy has the strongest experimental methods, such as resonance Raman, surface-enhanced Raman, confocal Raman, Raman imaging, etc. Depending on the application object, a suitable experimental method can be selected [7], which makes Raman spectroscopy irreplaceable for MIRS or near-infrared spectroscopy, such as quantitative and qualitative analysis of trace substances.

With its strong fingerprint and wide wavenumber measurement range (4000–50 cm^{-1}), Raman spectroscopy has a wide range of applications in many fields, especially in the testing of inorganic materials and biological samples, etc. It is widely used in scientific research. Due to the remarkable fingerprint nature of Raman spectroscopy, there are applications where a wealth of qualitative or quantitative information can often be obtained without the need for complex chemometric methods.

Unfortunately, the Raman scattering signal is very weak and is obtained from absolute measurements, which are susceptible to instrument variations and external environments, and the signal-to-noise ratio and repeatability of the spectra are relatively poor, which is a disadvantage for analytical methods combined with chemometrics, especially for the quantitative analysis of complex mixtures (e.g., oils and pharmaceuticals). In addition, satisfactory Raman spectra cannot be obtained for

some samples (e.g., heavy oils) due to interference from fluorescence, which limits the use of Raman spectroscopy for some applications.

The terahertz spectrum (or far-infrared spectrum), located between infrared and microwave, is in a special region of the transition from macroscopic classical theory to microscopic quantum theory and from electronics to photonics. Weak interactions between molecules (e.g., hydrogen bonding), skeletal vibrations of macromolecules (configuration bending), rotational and vibrational jumps of dipoles, and low-frequency vibrational absorption of lattices in crystals have a wealth of information in terahertz spectroscopy, which is of great scientific significance and practical application for detecting and understanding the structure and properties of matter and intermolecular interactions [8].

There is an extensive literature comparing NIR, mid-IR, Raman, and terahertz spectroscopy combined with chemometric methods for the measurement of complex sample systems. Examples include adulteration or origin identification of edible oils and honey, composition determination of food and feed, blending ratios of biodiesel, quantitative analysis of physical and chemical parameters of oils and polymers, monitoring of reaction processes, quantitative analysis and authenticity identification of pharmaceuticals, and diagnosis of clinical diseases [9–16]. The results obtained vary according to the object under study, the purpose of the study, and the chemometric methods used. However, some basic principles for the selection of spectroscopic analysis techniques can be summarized from the above literature:

- (1) The amount of spectral information should be considered first and foremost, the spectroscopic technique chosen should contain sufficiently rich chemical and/or physical information about the substance to be measured, which is one of the most critical aspects and is a prerequisite and basis for all qualitative and quantitative analysis.
- (2) Simplicity, timeliness, and effectiveness of the experimental method, i.e., convenient sample measurement, less preparation before and after measurement, fast measurement speed, good spectral repeatability, high signal-to-noise ratio, easy standardized operation, etc.
- (3) Easy to maintain and popularize, that is, comprehensively considering the characteristics of the instrument (specifications, stability, and consistency of spectral instrument, etc.) and cost, the difficulty of working curve or calibration model, and the cognition of industry personnel on the technology.

These basic principles mentioned above also apply to the selection of analytical techniques for atomic spectroscopy such as LIBS as well as spectral imaging [17, 18]. For example, Fig. 20.1 shows a Raman spectral imaging and NIR spectral imaging of a three-component tablet, and it can be seen that Raman spectral imaging provides much richer information on the spatial distribution of the chemical components. However, the measurement speed of Raman spectroscopy is still very slow (about 3.5 h), which makes it difficult to be widely used in the pharmaceutical industry, while the speed of NIRS is relatively fast (about 13 min) and the amount of information can meet the needs of many practical applications. Therefore, the choice of spectral

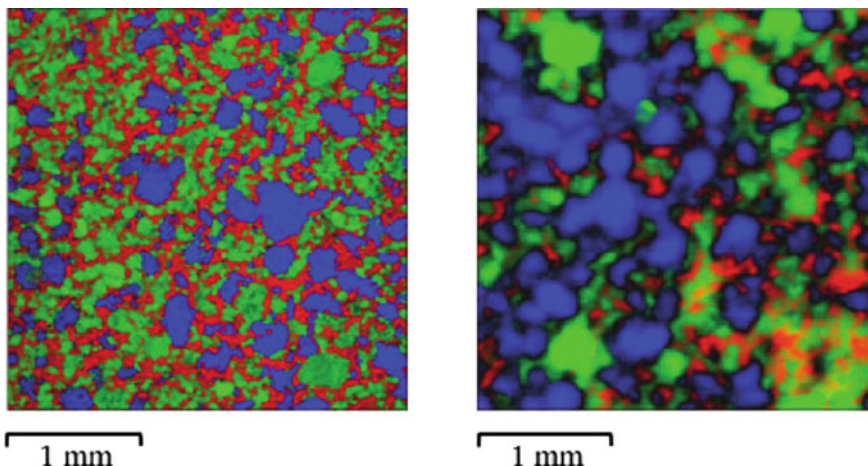


Fig. 20.1 Raman image (Left) and NIR image (Right) of ternary tablet. Blue: Microcrystalline cellulose; Green: Saccharin; Red: Eletriptan HBr

technology should be made by combining various factors such as the amount of spectral information, testing speed, and convenience [19].

In order to obtain more comprehensive and richer information about the samples, there has been an increasing emphasis in recent years on spectral fusion techniques, which include fusion between molecular vibrational spectra, but also between molecular and atomic spectra, and between spectroscopy and imaging techniques [20, 21]. For an introduction to spectral fusion methods and their algorithms please attend to Chap. 15 of this book.

20.2 Selection of Chemometric Methods

Chemometric methods are playing an increasingly important role in the quantitative and qualitative analysis of spectra and therefore the selection of the appropriate chemometric method for the problem to be solved is a very critical aspect [22, 23]. For example, Fig. 20.2 shows a framework of commonly used multivariate calibration and pattern recognition methods. In practice, the choice should be based on the specific problem, deciding whether to use a pattern recognition method or a multivariate calibration method, and then deciding whether to use a supervised pattern recognition method or an unsupervised pattern recognition method, a linear calibration method (PLS) or a non-linear calibration method (ANN), etc.

Spectral analysis is a highly practical technique and, from a practical point of view, not the more complex the algorithm, the better. Obtaining satisfactory results in the most concise way is one of the main principles followed in the selection of

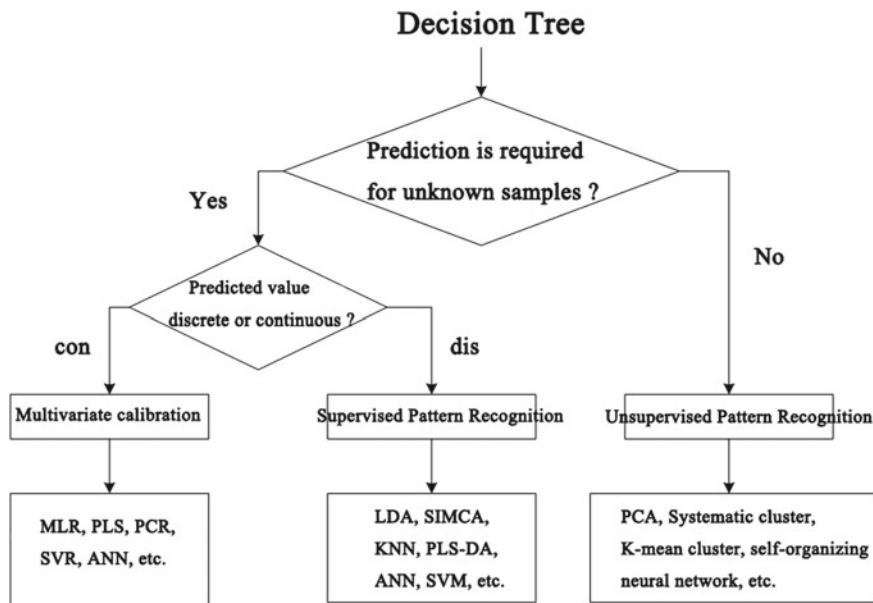


Fig. 20.2 Decision tree selection for pattern recognition and multivariate calibration methods

chemometric methods, which must be based on familiarity with chemometric algorithms and a good understanding of the technical problem to be solved. A typical example is the use of the concept of spectral standard deviation to determine the homogeneity of a powdered mixture. This application does not involve complex chemometric methods but makes full use of the relationship between spectral variation and standard deviation during the mixing process to determine the homogeneity of the mixture. With the simplicity of computing, the requirements on the hardware of the spectroscopy instrument are also significantly reduced, as there is no calibration model, less demanding long-term stability of the instrument, and no cumbersome model maintenance issues at a later stage.

20.2.1 Selection of Multivariate Calibration Methods

In the multivariate quantitative calibration of spectra, PLS can usually solve most of the problems, which is dictated by the characteristics of the algorithm itself, as it is based on multiple linear regression and principal component regression, overcoming the problem of multicollinearity between variables and allowing the strongest correlation between latent variables (principal components) and concentrations. However, PLS can only be used for linear or weakly non-linear analytical systems, and if severe

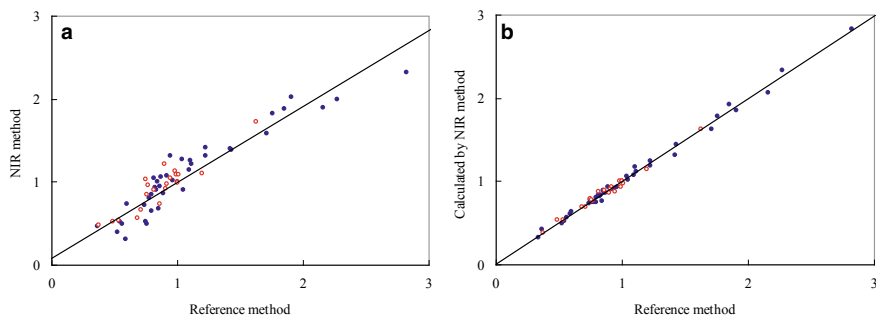


Fig. 20.3 Calibration and validation results of the NIR determination of the alcohol to alkene ratio using different PLS (a) and SVR (b) methods respectively (●—calibration sample; ○—validation sample)

non-linear systems are encountered, non-linear calibration methods such as ANN or SVR are required [24–26].

For example, when NIR was used to determine the alcohol to alkene ratio of methyl tert-butyl ether (MTBE) feed, there was a severe non-linear relationship between this ratio and the NIR spectrum as the alcohol to alkene ratio is the ratio of moles of methanol to moles of isobutylene in the feed to the reactor mixture (methanol and mixed carbon tetra fraction), which is different from the concentration of the pure component. In this case, the PLS method will not yield accurate calibration and prediction results (as shown in Fig. 20.3a). To directly build the calibration model for the determination of the alcohol to alkene ratio, a non-linear calibration method such as SVR must be used (as shown in Fig. 20.3b) [27]. Of course, the PLS method was also used to develop separate calibration models for the determination of methanol moles and isobutene moles, and then the alcohol-to-alkene ratio was calculated from their PLS predictions. Linear regression methods include Lasso methods and elastic networks in addition to PLS methods; non-linear calibration methods include extreme learning machines, Gaussian process regression, and convolutional neural networks in addition to SVR and BP-ANN [28, 29], which are described in Chaps. 7, 8, and 18, respectively. Strategies based on local modeling can also be used to solve non-linear calibration problems [30] and to improve the robustness of the predictions of quantitative calibration models, an ensemble or consensus strategy can be used [31], which can be found in Chap. 14.

20.2.2 Selection of Pattern Recognition Methods

In spectral pattern recognition, traditionally, the most applied method is principal component analysis (PCA). In solving most problems, clustering or identification analysis using principal component scores as features can give satisfactory results.

However, PCA decomposes the spectral array along the direction of variance maximization, so that the resulting principal component scores are not necessarily the most relevant to the category. In particular, PCA often does not give satisfactory results when the feature information associated with the category is not significant in the spectrum. In this case, supervised algorithms such as canonical variance analysis (CVA) or PLS-DA can be used. For example, Yuan H et al. classified the IR spectra of 454 residual oils (105 atmospheric residual oils, 98 vacuum residual oils, and 269 hydrogenated residual oils). Since hydrogenated residual oils differ significantly in composition from atmospheric and vacuum residual oils, while atmospheric and vacuum residual oils differ relatively little in composition, it is easy to separate hydrogenated residual oils from atmospheric and vacuum residual oils by the PCA score as a characteristic variable, but not from the PCA score can easily separate hydrogenated residue from atmospheric and vacuum residual oils, but not atmospheric residual oils from vacuum residual oils (see Fig. 20.4). However, using the PLS-DA method, which assigns a value of -1 to atmospheric vacuum residual oils, 0 to vacuum residual oils, and 1 to hydrogenated residual oils, it is easy to classify and identify these three types of residual oils (see Fig. 20.5) [32].

In pattern recognition, there is actually another class of recognition method, named by similarity analysis, which is used to determine the degree of similarity between two samples, and the traditional spectral library search method falls into this category. Similarity analysis mostly uses correlation coefficients or distances as evaluation indicators and is commonly used for IR and Raman spectral library searches of pure compounds, but it is difficult to perform library searches of samples from a particular class of complex mixed systems (e.g., crude oil species). Because the chemical composition of the subjects in a class of samples is extremely similar

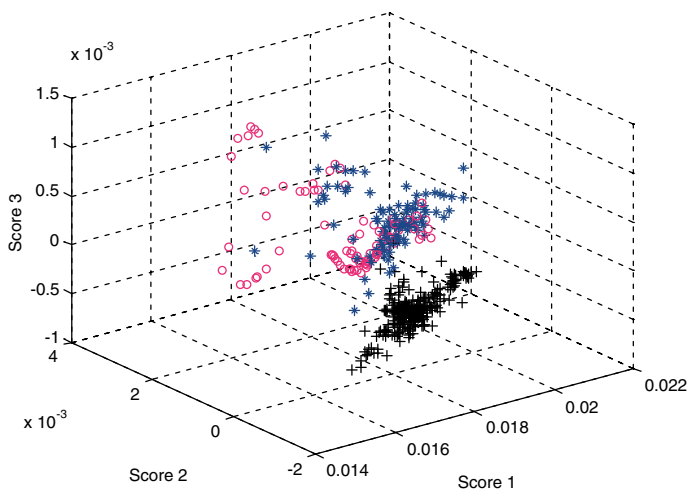


Fig. 20.4 Distribution of the first three scores obtained by PCA for the IR spectra of the three types of residual oils

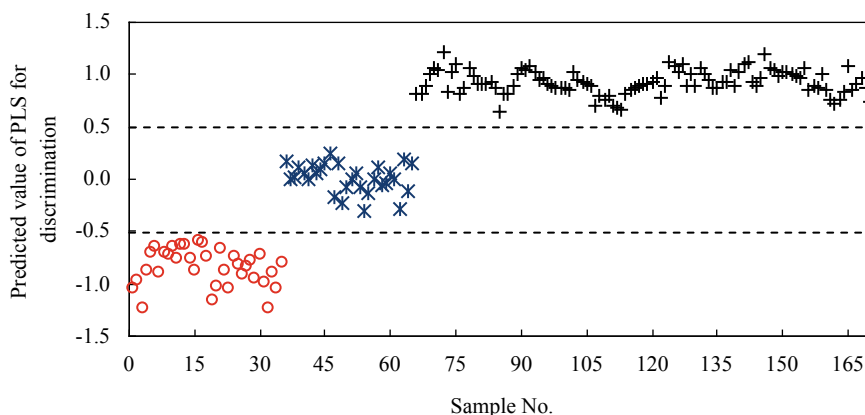


Fig. 20.5 Infrared spectra of three types of residue oils identified by PLS-DA; O—atmospheric residue; *—reduced pressure residue; +—hydrogenated residue

and the correlation coefficients between the spectra are mostly above 0.98, or even close to 1 in some cases, but there are some differences in the composition and properties between the samples, and it is not possible to distinguish precisely between these samples with very similar spectra by traditional methods such as correlation coefficients.

Various methods have been tried to improve the accuracy of similarity analysis, and Blanco et al. proposed to improve the accuracy of traditional correlation coefficient identification by building a sub-spectral library for the identification of NIRS of drug ingredients [33]. Loudermilk et al. used a consensus strategy to integrate the results of multiple identification methods for the retrieval of a library of infrared spectra of cotton contaminants [34]. Xu et al. used the segmented correlation coefficient method (array correlation coefficient) for the identification and analysis of infrared spectra of Chinese herbs, dividing the whole spectral range into several regions and calculating the correlation coefficient for each region separately, which can improve the difference between spectra to a certain extent [35].

To identify the fast identification of crude oil species, Chu et al. combined the moving window concept with the correlation coefficient method to propose a new similarity calculation method—the moving window correlation coefficient method. For two spectra that the similarity calculations were to be performed, the correlation coefficient values were calculated in each moving window sub-wavelength region, and then the obtained correlation coefficient values were plotted against the starting position of the corresponding window to obtain a moving correlation coefficient plot [36]. As shown in Fig. 20.6, the moving correlation coefficient method yields a vector from which the degree of similarity between the two spectra can be easily seen, and if the two spectra are identical, the moving correlation coefficient value is 1 over the entire spectral range. If the two spectra differ only in one interval, the value of the moving window correlation coefficient for that interval will be significantly lower.

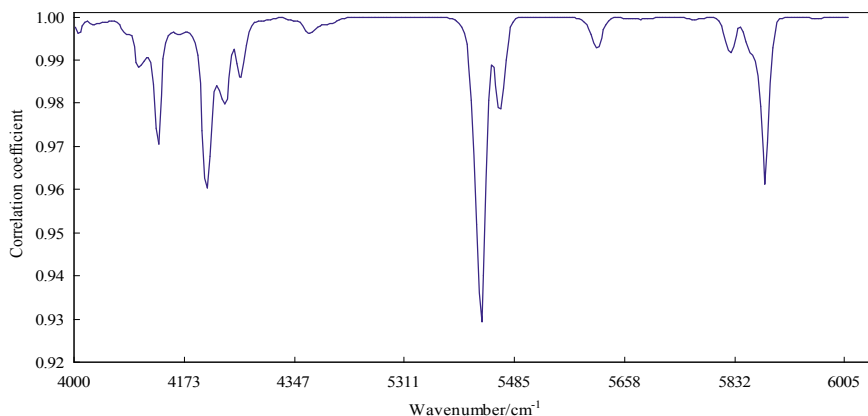


Fig. 20.6 Figure of moving window correlation coefficient for two similar crude oils

This method allows small differences between spectra to be discerned, facilitating spectral resolution and information extraction. Asemani et al. used this method for the rapid identification of bitumen by extracting features from the absorbance ratio of infrared spectra and obtained satisfactory results [37]. In recent years, methods such as random forest and convolutional neural networks have been increasingly used for pattern recognition of complex sample spectra [38, 39], please refer to Chaps. 12 and 18 for details.

20.2.3 Selection of Spectral Preprocessing Methods and Spectral Variables

The choice of spectral pre-treatment method and spectral range is also very important. The most commonly used spectral preprocessing methods are first-order derivative, second-order derivative, MSC, and SNV, with liquid transmission measurements mostly using preprocessing methods such as derivative and solid diffuse reflectance measurements mostly using preprocessing methods such as MSC. For Raman spectroscopy, a baseline correction method is often also required to eliminate the effect of fluorescence on the spectrum. A number of papers have compared different pre-treatment methods, and the optimal pre-treatment method varies for different measurement systems [40]. When comparing different preprocessing methods, especially some of the more complex ones, it is important to consider whether the method is a substantial improvement in predictive ability or just an “improvement” within the margin of error (so-called numbers game). If the latter is the case, it is recommended that the classical conventional pre-treatment method is still used. If multiple pre-treatment methods are chosen at the same time, attention should also be paid to the order in which they treat the spectra.

The most commonly used method of wavelength selection is the correlation coefficient method, which intuitively gives the most informative spectral interval. Other methods, such as genetic algorithms, often yield better calibration and prediction results, but their parameters are more complex to choose and more computationally intensive, and need to be chosen with care when building commercial calibration models. In recent years, variable selection methods based on model population analysis, represented by the competitive adaptive re-weighting algorithm (CARS), have received the most widespread attention and use [41]. When selecting spectral intervals, attention should also be paid to the use of chemical knowledge, especially spectroscopic knowledge. Some chemometric software has an automatic wavelength screening function, but it is often necessary to adjust its screening results according to spectroscopic knowledge, because the characteristic bands most relevant to the component or property to be measured may not be selected in the automatic screening process.

In addition, for some training sets, combining spectral variables and performing mathematical operations, such as ratios of certain wavelength variables or ratios after differences, etc., instead of the original spectral variables to build a calibration model can improve the predictive ability of the model.

The order of precedence between spectral preprocessing and wavelength interval selection should also be noted. For methods such as derivatives and mean-center, whether preprocessing or selection of wavelength intervals is performed first has no effect on the final calibration and prediction results. However, for methods such as MSC and SNV, there is a certain influence and the wavelength interval needs to be selected before the preprocessing operation.

The incorporation of spectral preprocessing and wavelength selection methods into the multivariate calibration step is an important development. Examples include the sequential preprocessing method based on orthogonal operations (SPORT) proposed by Roger et al. [42] and the parallel preprocessing method based on orthogonal operations (PORTO) proposed by Mishra et al. [43] that are detailed in Chaps. 4 and 5.

20.3 Influencing Factors of Model Prediction Ability

Spectroscopy combined with chemometrics is an indirect measurement technique, and the establishment of robust, reliable, and accurate calibration models is key to the successful application of such analytical methods. All aspects involved in the modeling process affect the reproducibility and accuracy of the analytical results. The main influencing factors include the representativeness of the calibration sample, the accuracy of the reference data, the method and conditions of spectral acquisition, the chemometric methods, and the performance of the spectroscopic instrument. The selection of chemometric methods has been discussed in the previous section, and the following discussion focuses on the effects of calibration samples, reference data, spectral acquisition methods and conditions, and instrument performance.

20.3.1 Effect of Calibration Samples

The impact of calibration samples on the analytical model relates to the representativeness, number, range, and distribution of calibration samples, the homogeneity of calibration samples (such as grain size, budding rate, shriveled rate, water content, color, and impurities of agricultural product samples) and the pre-treatment of calibration samples (grating, slicing, and extraction) among many other things. The following section focuses on the problem of uneven distribution of calibration samples, which is often encountered in practical applications (Fig. 20.7).

As shown in Fig. 20.8 [44], during the establishment of the NIRS for the determination of density of naphtha, the main body of calibration samples collected had a density distribution between 0.66 and 0.72 g/cm³ but contained a class of samples with a higher density (around 0.75 g/cm³). It is also clear from the PCA score space that these small samples are also significantly different from those of the

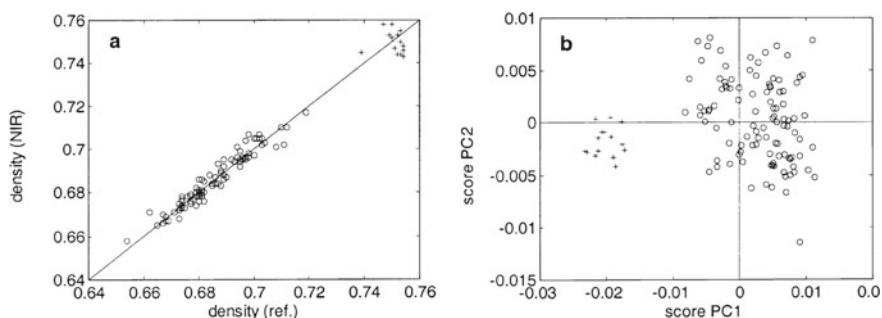


Fig. 20.7 Example plots of uneven distribution of calibration set samples. **a** Results of PLS cross validation; **b** Distribution of calibration set samples in PCA score space; o—main calibration sample, +—a small number of calibration samples)

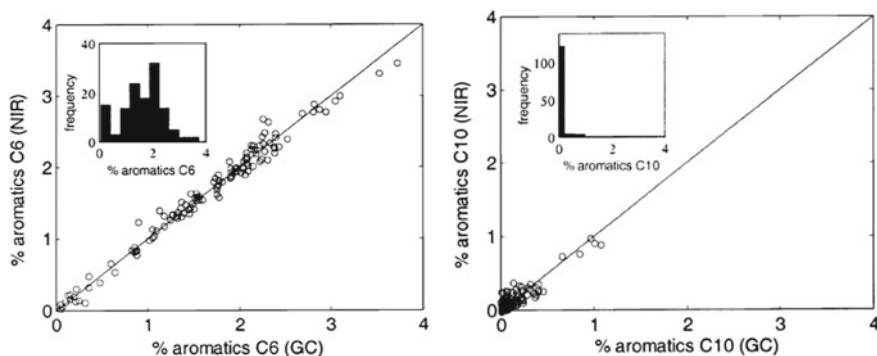
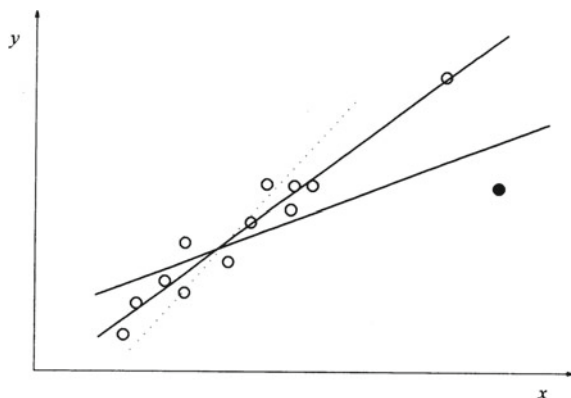


Fig. 20.8 Cross-validation results for different components to be predicted in the same calibration set samples

Fig. 20.9 Schematic representation of the effect of outlier samples on the model



main distribution. If these samples are involved in the modeling, the SECV of the constructed model becomes significantly worse, with the SECV increasing from 0.028 to 0.038 g/cm³, although the predicted trend remains. In practice, the decision to involve this part of samples in modeling needs to be made on a case-by-case basis. Because this type of sample is different from anomalous (out-of-bounds) samples, as shown in Fig. 20.9, anomalous samples are necessary to be removed before building a calibration model, otherwise the predictive ability of the model will be seriously affected, especially when the number of calibration samples is small.

It is also important to note that a set of calibration samples may have a uniform concentration distribution for some components to be measured, but not for others. As shown in Fig. 20.8, a calibration set of 132 naphtha samples had an essentially homogeneous distribution of C₆ aromatics. For C₁₀ aromatics, however, the distribution was extremely heterogeneous, with most of the calibration samples containing <0.4% C₁₀ aromatics and only five samples at around 1.0%. If this calibration set is used to build a calibration model for the determination of C₁₀ aromatics it is clearly inappropriate and additional samples will need to be collected.

For samples with poor homogeneity, the predictive accuracy of the model can be improved if the necessary processing is carried out prior to measurement. For example, for NIR spectroscopic diffuse reflectance measurements, samples grated to a fine powder are usually better than predicted for intact particles. For instance, the protein and starch content of flour is better predicted than that of wheat seeds, and the nicotine and total sugar content of tobacco powder are better predicted than that of tobacco slices, but this comes at the expense of convenience and speed of analysis. The choice of sample pre-treatment for a specific application needs to be decided on a case-by-case basis. Other factors involved in influencing the sample are sample temperature, water content and residual solvent, sample thickness, loading tightness, optical properties, polymorphism, and the actual storage time of the sample [45–48].

The number of representative samples in a calibration set has an important impact on the quality of the model, and the number of samples required for modeling is

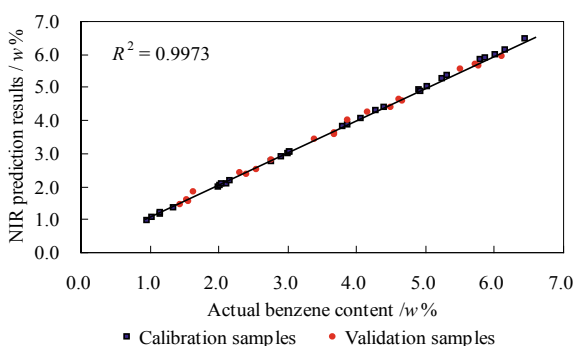
closely related to factors such as the object of application and the quantitative calibration algorithm used [49–54]. In general, the larger the number of calibration samples, the more variable information they contain about themselves and the outside world, the more significant the non-linear relationship between the spectrum and the physical properties to be measured may be, the more non-linear calibration methods such as neural networks may be used. In contrast, for non-linear methods such as neural networks, the greater the number of representative samples used in modeling, the better the robustness of the model built. The number of calibration samples also involves issues such as modeling strategy and model maintenance, which can be found in Chaps. 11 and 14.

20.3.2 Effect of Reference Data

The accuracy of the reference data has a large impact on the predictive ability of the calibration model. To examine the influence of the reference data, Chu et al. carried out NIR spectral simulations. Fifty samples of a four-component mixture of benzene, toluene, xylene, and isooctane were prepared and analyzed for benzene content. The reference data for benzene concentration were obtained by weighing during the preparation process and the distribution of benzene content ranged from 1.0 to 6.5%. Thirty samples were selected by the K-S method to form the calibration set, and the validation set consisted of the remaining 20 samples. A quantitative calibration model was developed using PLS method, and the spectra were processed by first-order derivatives and the chosen spectral interval was 750–1050 nm. Figure 20.10 gives the actual-predicted correlation plots for the calibration (cross validation) and for the benzene content during validation, with a SECV of 0.06% for the cross validation and a SEP of 0.09% for the validation set [55].

To examine the effect of the accuracy of the reference data on the model and its predictive ability, errors were added to the benzene content of the calibration set samples artificially, the calibration model was rebuilt, and the benzene content of the 20 validation set samples was predicted. Increase the error in the reference data

Fig. 20.10 Actual–predicted correlation plot for the determination of benzene content in a four-component mixture of benzene, toluene, xylene, and isooctane by NIRS



in three ways: (1) Add the absolute error Δy to the original reference data y_i of the calibration set samples, i.e., $y_i \pm \Delta y$, where Δy takes a positive value, using a plus sign for samples with even serial numbers and a minus sign for samples with odd serial numbers, and the serial numbers of the calibration set samples are arranged randomly. (2) Relative error $y_i \times r\%$ is added to the original reference data y_i of the calibration set samples, i.e., $y_i \pm y_i \times r\%$, where plus signs are used for samples with even serial numbers and minus signs are used for samples with odd serial numbers. (3) Add a normally distributed random error Δy_i to the original reference data y_i of the calibration set samples, i.e., $y_i \pm \Delta y_i$ where the random error Δy_i is automatically generated with the standard deviation as an indicator of the error.

Absolute errors of 0.1–1.0% (w%) were sequentially added or subtracted to the benzene content of the calibration set samples according to the reference data error addition method (1) to examine the effect of variation in the accuracy of the reference data on the model and its prediction results. The variation curves of SEC and SEP with absolute error Δy are given in Fig. 20.11, while Fig. 20.12 shows the cross-validation results of the calibration set and the prediction results of the validation set for an absolute error of ± 0.7 (w%). It can be seen that as the accuracy of the reference data becomes worse, both SEC and SEP become correspondingly larger, but the increase in SEP is much smaller than that of SEC, indicating that the NIR calibration model built with it can still obtain more accurate prediction results despite the large absolute errors in the reference data within a certain range. Similar results were obtained by adding the other two types of error to the reference data of the calibration sample, as shown in Figs. 20.13 and 20.14.

The same results are obtained for the testing of complex mixtures such as oils. Chung et al. used two different sets of reference data of lubricant pour points with different accuracies to build an analytical model for NIRS using PLS [56]. The accuracy of one reference method for determining pour points (1 °C reading interval) was

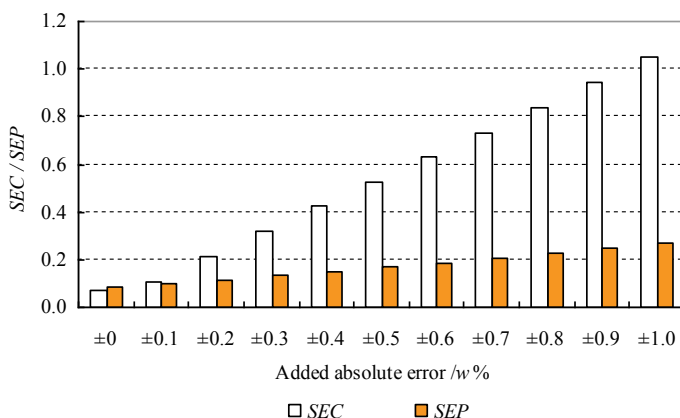


Fig. 20.11 Effect of adding absolute error to the reference data of the calibration sample on calibration and prediction

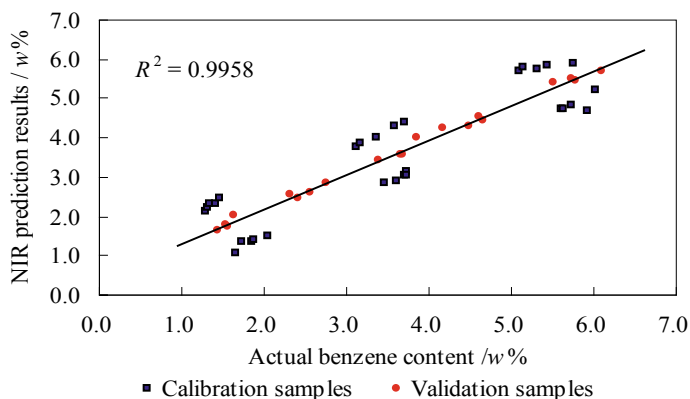


Fig. 20.12 Effect of adding an absolute error of 0.7% to the reference data of the calibration sample on calibration and prediction

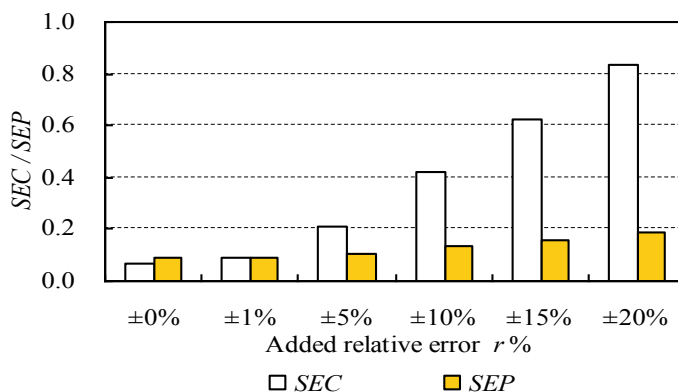


Fig. 20.13 Effect of adding relative error to the reference data of the calibration sample on calibration and prediction

significantly better than the other method (3 °C reading interval), and the results are shown in Table 20.1. Under the same parameter conditions, the SECV obtained when modeling with high accuracy pour point data was better than that of the low-accuracy data. Figure 20.15 shows a plot of the first two principal factor scores obtained from the PLS regression, which reveals that despite the difference in accuracy between the two sets of reference data, their scores match almost exactly. It can be suggested that the difference in SECV is mainly caused by the quality of the reference data.

From the above simulations, practical examples, and relevant references [57–61], the following conclusions can be drawn.

- (1) The accuracy of the reference data has an impact on both the calibration model and its prediction results. The more accurate the reference data, the higher

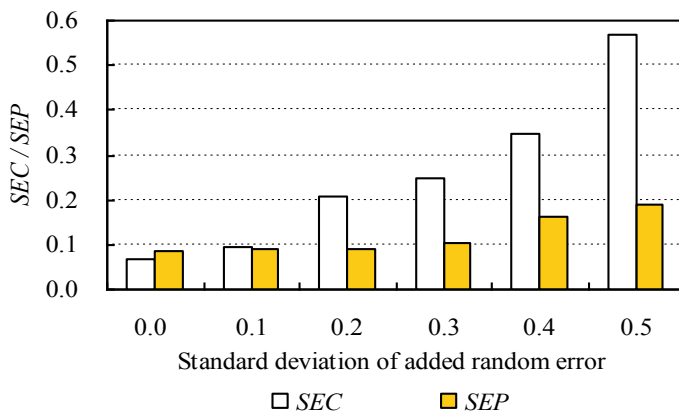
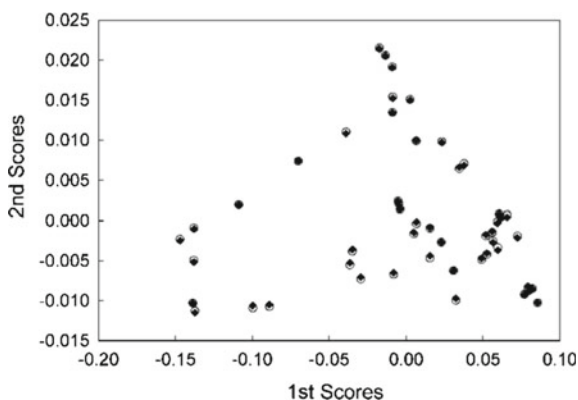


Fig. 20.14 Effect of adding random error to the reference data of the calibration sample on calibration and prediction

Table 20.1 Results obtained from separate PLS regressions for two sets of reference data with different accuracies

Spectral range (nm)	Low accuracy reference data		High accuracy reference data	
	PLS factor	SECV/°C	PLS factor	SECV/°C
1100–1580	3	1.70	3	1.17
1100–1580 and 1870–2140	3	1.66	3	1.14

Fig. 20.15 Plots of the first two factor scores obtained from separate PLS regressions for two sets of reference data with different precision



the accuracy of the model built and the more accurate its prediction results for unknown samples. To obtain reliable reference data, reference methods with high accuracy and repeatability should be used, sometimes averaged over several measurements, and where possible, the same instrument and skilled operators should be used to measure the reference data of calibration samples.

If necessary, the accuracy and repeatability of the conventional methods used to obtain the reference data need to be assessed. Samples used for reference data measurements must be the same as those used for spectral acquisition and, where possible, the reference data and spectra must be measured in time after sampling to avoid changes in sample composition affecting the accuracy of the calibration model. The units of the reference data should also be considered, e.g., the linearity between volumetric concentration units and spectral absorbance is better than that of weight concentration units [62, 63].

- (2) Although the calibration model is obtained from the regression between reference data and the corresponding spectra, the spectral approach has the potential to yield predictions that are closer to the true value. Particularly for the reference data provided by relatively poor accurate test methods, it will be possible to obtain more accurate prediction results by processing a large number of samples through statistical analysis. This does not mean, however, that the accuracy and repeatability of the spectroscopic method are necessarily better than that of the reference method.
- (3) When actually building the calibration model, samples with relatively large deviations in the cross validation can be retained in the calibration set (deviations should generally not exceed 1.5–2.0 times the reproducibility requirements of the reference test method), which can increase the robustness and applicability of the model without basically affecting its prediction accuracy.

20.3.3 Effect of Spectral Measurement Methods

The measurement method of the spectra is one of the important factors that determine the quality of the spectra (signal-to-noise ratio, repeatability, spectral information, etc.), and the quality of the spectrum will significantly affect the predictive ability of the calibration model. Therefore, it is very important to choose an appropriate spectrum measurement method. A suitable measurement method should meet the following conditions: (1) the repeatability and reproducibility of the spectrum are excellent; (2) the test is convenient and fast; (3) the signal-to-noise ratio of the spectrum is high; (4) the sample physicochemical information contained in the spectrum is rich and complete.

Each spectral analysis technique has a variety of measurement methods. For example, the commonly used measurement methods for mid-infrared spectroscopy include transmission, ATR, and diffuse reflection; the commonly used methods for NIRS include diffuse reflection, transmission, and diffuse transmission; Raman spectroscopy has more measurement methods, such as backscattering, SERS, SORS, and transmission. Spectral measurement methods are closely related to the selection of measurement accessories. For the same sample type (such as transparent liquids, viscous bodies, solid particles, powders, etc.), a variety of measurement methods and accessories can be used for spectral measurement. In the process of feasibility study before practical application, the advantages and disadvantages of all feasible

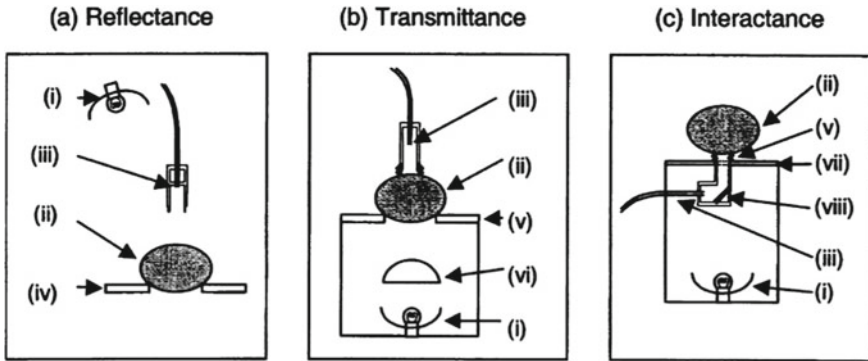


Fig. 20.16 Three ways to measure the near-infrared spectrum of fruit. (i) light source; (ii) fruit; (iii) fiber optic probe; (iv) black foam holder; (v) light sealing ring; (vi) condenser lens; (vii) glass cover; (viii) plane mirror

measurement methods and their potential accessories should be compared to select a suitable and convenient spectral measurement method.

For example, when NIRS is used to non-destructively determine fruit quality, a variety of measurement methods can be used. Figure 20.16 shows three common spectral measurement methods: Transmission, Reflectance, and Interactance. A study compared the advantages and disadvantages of these three methods for determining fruit soluble solids content, density, and flesh color [64]. The results showed that relatively accurate results can be obtained by the method of interactance measurement. The interactance method is actually a deformation of the reflectance method. Since there is a certain distance between the incident light irradiation area and the light collection area, this method collects not the diffuse reflection light on the surface of the fruit, but the light that penetrates into the fruit and comes out, so it carries more physical and chemical information of the internal components of the fruit, but avoids the influence of the fruit kernel on the spectrum caused by the diffuse transmission measurement method. At present, some commercialized portable near-infrared fruit analyzers use ring optical fiber light source, which essentially adopt this internal reflection measurement method.

When Raman spectroscopy is used to determine the content of active ingredients in tablets, in order to obtain a spectrum with good repeatability and can reflect the bulk information, measurement methods such as large illumination spot, rotating sample, and transmission can be used. Compared with the traditional backscattering measurement method, the transmission Raman can often obtain more accurate measurement results. Using the same spectral range ($700\text{--}1700\text{ cm}^{-1}$) and calibration methods, the predictive ability of the transmission Raman is better than that of the traditional backscattering measurement method [65, 66].

Similarly, when NIRS is used to determine the content of active ingredients in tablets or capsules, the transmission is often better than the diffuse reflection measurement method, which is because the transmission usually contains more information

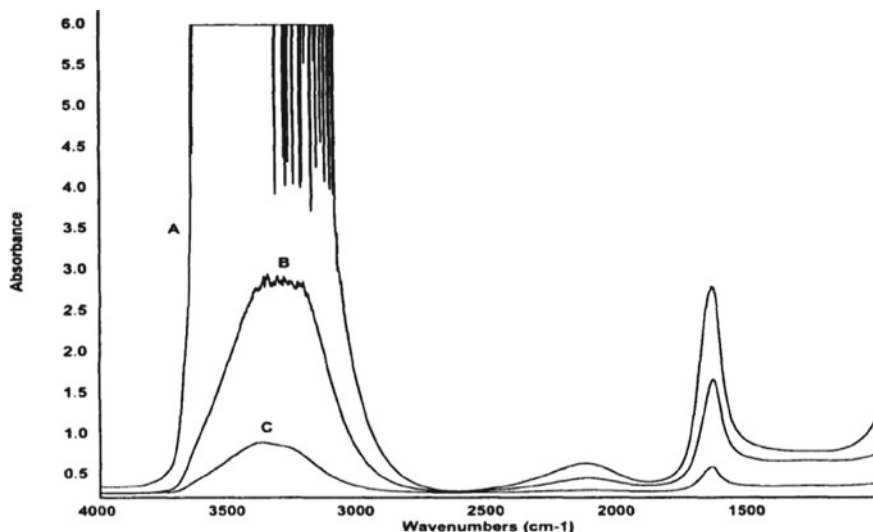


Fig. 20.17 Mid-infrared spectra of water measured in different ways. (A) 25 μm transmission cell; (B) 45° ZnSe horizontal ATR (12 reflections); (C) 45° Ge horizontal ATR (12 reflections)

about the chemical composition of the sample than the reflection spectrum, so the transmission method is more conducive to the analysis of the composition or properties closely related to the composition [67–69]. Mid-infrared spectroscopy also has transmission, ATR, diffuse reflectance, and photoacoustic measurement methods, and Fig. 20.17 shows the infrared spectra of water obtained by different measurement methods. In the actual applications, it is necessary to evaluate the convenience of testing and the amount of spectral information to select [70–73].

As shown in Fig. 20.18, for the liquid transmission measurement method, the selection of optical path should seek the best advantage among the factors such as spectral range, absorbance linearity, and spectral signal-to-noise ratio [74–77]. The choice of background reference material is also very important for the diffuse reflectance measurement of solid substances.

20.3.4 Effect of Spectral Acquisition Conditions

Spectral acquisition conditions include spectral range, resolution, number of spectral accumulation measurements, and uniformity and consistency of sample loading. The influence of different spectral acquisition conditions on the calibration model is described below by taking NIR spectral analysis as an example.

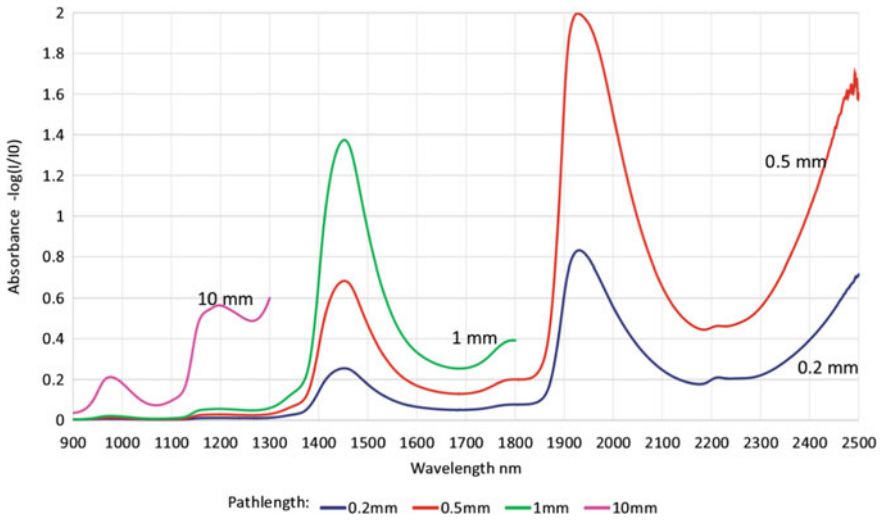


Fig. 20.18 Near-infrared spectra of water under different optical path [78]

(1) Effect of spectral range

The wavelength range of the NIRS is usually divided into two sections, the short-wave region of 700–1100 nm and the long-wave region of 1100–2500 nm. Among them, 1100–2500 nm can be divided into three sections, 1100–1540 nm ($9090\text{--}6500\text{ cm}^{-1}$), 1540–2000 nm ($6500\text{--}5000\text{ cm}^{-1}$), and 2000–2500 nm ($5000\text{--}4000\text{ cm}^{-1}$). As shown in Fig. 20.19, the NIRS of different wavelengths and their spectral characteristics are quite different.

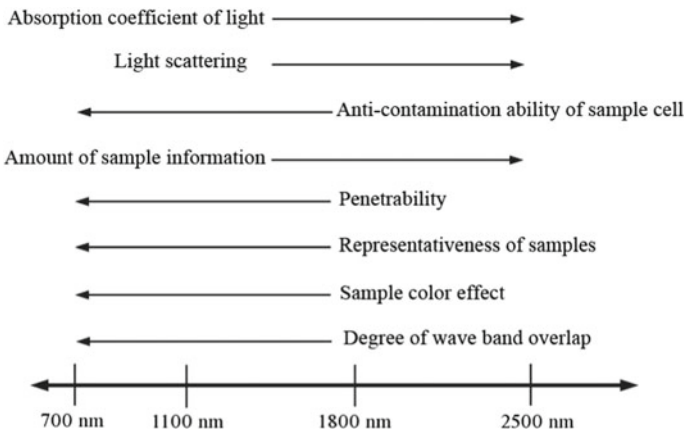


Fig. 20.19 Near-infrared light and its characteristics as a function of wavelength

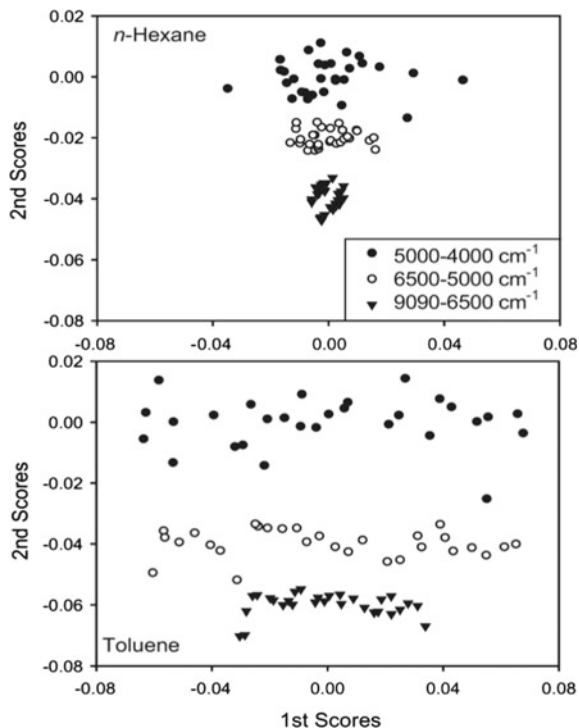
The short-wave NIR spectral region is mainly the absorption of third-order and fourth-order overtone and combination frequency, and the long-wave region is mainly the absorption of first-order and second-order overtone and combination frequency. The light transmittance in the short-wavelength region is strong, and the absorption coefficient is small. Long optical paths such as 30–50 mm are often used. The representativeness of samples and the anti-contamination ability of sample cell are relatively strong. The long-wavelength NIR spectral region is richer in information than the short-wavelength region, especially in the combination region of 2000–2500 nm, the band overlap is not as serious as that in the short-wave region. However, the required optical path is shorter in the long-wavelength NIR spectral region, usually 0.5 mm.

Cho et al. [79] investigated the influence of different spectral ranges on the NIRS calibration model by using a prepared three-component (*n*-hexane, *n*-heptane, and toluene) system with *n*-heptane as the solvent and *n*-hexane and toluene as the analytical objects. A total of 30 samples were prepared in which the concentrations of *n*-hexane and toluene ranged from 0.05 to 3.0% (w/w). In order to obtain three spectra with similar absorbance intensities, different optical path lengths were selected, the optical path in the range of 9090–6500 cm^{-1} is 10 mm, the optical path in the range of 6500–5000 cm^{-1} is 2 mm, and the optical path in the range of 5000–4000 cm^{-1} is 0.5 mm. The SECV obtained by PLS cross validation in different spectral ranges can be concluded that the results in the range of 5000–4000 cm^{-1} are the best, and the results of toluene are obviously better than those of *n*-hexane. This shows that there are differences in the amount of information in different spectral ranges. Compared with the other two bands, the combined frequency region of 5000–4000 cm^{-1} has more information, and the overlap of the spectral bands is relatively small. Because the spectra of *n*-hexane and *n*-heptane are very similar, the results for *n*-hexane are relatively poor relative to toluene. This can also be explained from the PLS factor score plots. As shown in Fig. 20.20, the scores of toluene in the range of 5000–4000 cm^{-1} change the most, indicating that they have the most information and the corresponding SECV is the smallest. The spectra of *n*-hexane in the range of 9090–6500 cm^{-1} have the smallest change in scores, indicating that its information content is the least, and the corresponding SECV is the largest.

The above experiments were obtained based on the transmission measurement method. In practical applications, the problem of cross selection between different spectral ranges and different measurement methods is often encountered. For example, for granular samples, whether the calibration model established by the diffuse reflection in the long wavelength region or the diffuse transmission in the short and medium wavelength region is better, which needs to be determined through feasibility experiments for different measurement modes.

In addition, the choice of spectral range is also limited by many conditions such as spectrometer type, accessory type, and measurement method. For example, when using optical fiber measurement accessories, due to the self-absorption of optical fiber materials, if the optical fiber distance is too long, the spectral range above 2200 nm will be unusable. For another example, when using the diffuse transmission method to measure the NIRS of a tablet, only the spectral region between 800 and

Fig. 20.20 The first two-factor scores obtained by PLS regression of n-hexane and toluene in the three spectral regions



1600 nm is available due to the weakened penetrating ability of NIR light in the long wavelength region.

(2) Effect of resolution

Spectral resolution is a measure of an instrument's ability to distinguish two adjacent absorption peaks and is usually characterized by spectral bandwidth, which is the width at half the maximum intensity of the monochromatic spectral band emitted by a monochromator. The spectral resolution mainly depends on the monochromator of the spectroscopic instrument. The resolution of the grating spectroscopic instrument is related to the design of the slit. The narrower the slit, the higher the resolution, but the optical throughput will decrease, so that the signal-to-noise ratio of the spectrum will decrease. The resolution of the array detector is also related to the pixels of the detector. The resolution of the Fourier spectrometer is determined by the moving distance of the moving mirror. The higher the resolution, the further the moving distance of the moving mirror, the slower the scanning speed, and the lower the signal-to-noise ratio per unit time. In addition, high-resolution spectral files can also affect the speed of math operations. Therefore, in the practical application of spectroscopy combined with chemometric methods, high resolution is generally not pursued, usually not exceeding 4 cm^{-1} .

It is especially true for the NIR spectral region, because the absorption bands are mostly broad and overlapping, and high instrument resolution is usually not required for quantitative or qualitative analysis. A resolution of 16 cm^{-1} or 10 nm (at 2500 nm) is sufficient for most analytical applications. For example, in the determination of the octane number of gasoline by NIRS, a resolution of 40 nm can generally meet the requirements of routine analysis accuracy. However, for complex samples with very similar structural characteristics, if accurate analysis results are to be obtained, certain requirements must be placed on the resolution of the instrument [80–83].

To examine the effect of resolution on the NIR calibration model, Chung et al. [84] designed an experiment in which they prepared 55 samples of mixtures using 25 pure hydrocarbons to simulate the composition of naphtha. The NIRS of these 55 samples were measured at different resolutions (4 , 8 , 16 , and 32 cm^{-1}), with a spectral range of $4000\text{--}4500\text{ cm}^{-1}$ and an optical path length of 0.5 mm . The calibration model was established by the PLS method. Table 20.2 shows the SECV of group compositions (paraffins, *n*-paraffins, isoparaffins, naphthenes, and aromatics) and some pure compounds. It can be seen that the 4 , 8 , and 16 cm^{-1} resolutions have little effect on the group composition, while the 32 cm^{-1} resolution has a greater impact, the number of PLS factors required increases, and the SECV becomes worse. For pure compounds, due to the small difference in the NIRS of *n*-alkanes, the resolution has a great influence on the SECV of *n*-heptane and *n*-hexane, and satisfactory results cannot be obtained when the resolution is 32 cm^{-1} . However, for benzene and toluene, due to their relatively strong characteristics, under the same resolution, they use a small number of PLS factors, and the SECV value is small, and the impact of the resolution on benzene and toluene is also smaller than that on *n*-heptane and *n*-hexane.

Table 20.2 Effect of resolution on the SECV of group compositions and some pure compounds; the numbers in the inner brackets are the number of PLS factor

Composition	SECV (4 cm^{-1})	SECV (8 cm^{-1})	SECV (16 cm^{-1})	SECV (32 cm^{-1})
Total alkanes	0.54 (12)	0.51 (12)	0.52 (13)	1.25 (12)
Total <i>n</i> -alkanes	0.44 (15)	0.37 (15)	0.43 (16)	0.76 (22)
Total isoparaffins	0.36 (15)	0.37 (15)	0.40 (15)	0.96 (21)
Total naphthene	0.78 (14)	0.81 (14)	0.89 (14)	
Total aromatics	0.81 (12)	0.81 (12)	0.84 (12)	1.60 (10)
<i>n</i> -Hexane	0.29 (16)	0.29 (17)	0.30 (17)	1.64 (17)
<i>n</i> -Heptane	0.29 (20)	0.28 (21)	0.42 (22)	
2, 2-dimethylbutane	0.24 (17)	0.23 (16)	0.23 (17)	0.78 (22)
Cyclohexane	0.16 (16)	0.17 (17)	0.17 (19)	0.37 (24)
Benzene	0.17 (6)	0.17 (6)	0.17 (9)	0.23 (20)
Toluene	0.21 (16)	0.21 (17)	0.26 (16)	0.32 (19)

Table 20.3 The influence of measurement times on spectral noise and SECV

Times of the measurement	RMS noise of the spectra $\times 10^{-5}$	Mean standard deviation of the spectra $\times 10^{-5}$	SECV/V%		
			<i>n</i> -Hexane	Cyclohexane	Toluene
4	8.29	0.35	0.280(3)	0.045(3)	0.031(4)
8	5.88	0.35	0.135(3)	0.037(3)	0.039(4)
16	4.12	0.14	0.083(3)	0.025(3)	0.021(5)
32	2.98	0.10	0.073(3)	0.023(3)	0.017(4)
64	2.14	0.12	0.053(4)	0.020(3)	0.015(4)

* The numbers in parentheses in the table are the number of main factors of PLS.

(3) Impact of spectral scans

Increasing the number of spectral scans of the sample, that is, averaging through multiple measurements, is a common method to improve the spectral signal-to-noise ratio.

The signal-to-noise ratio of the spectrum affects the predictive ability of the model. Cho and Chung [85] investigated the influence of the number of spectral scans on the NIR calibration model based on a set of artificially prepared samples. They mixed *n*-hexane, *n*-heptane, cyclohexane, and toluene at different concentrations, and prepared dozens of samples with *n*-heptane as the solvent. The concentrations of *n*-hexane, cyclohexane, and toluene ranged from 0.05 to 2.0% (v/v). The spectrum collection range was 4000–4500 cm^{-1} , the resolution was 4 cm^{-1} , and the optical path was 0.5 mm. The calibration model was established by PLS method. Table 20.3 showed the root mean square noise (RMS), the average standard deviation of the measured spectra, and the effect on the SECV of *n*-hexane, cyclohexane, and toluene under different measurement times. It could be seen that with the increase in the number of scans, the spectral noise decreased, and the SECV of *n*-hexane, cyclohexane, and toluene also decreased to varying degrees, and the influence of *n*-hexane was the most significant. This was because the spectral difference between *n*-hexane and the solvent *n*-heptane is small, and a high signal-to-noise ratio of the spectra is required to obtain satisfactory prediction results.

In practical application, it is not that the more the number of scans, the better. The increase in the number of scans will prolong the measurement time of the spectrum. In addition, when the number of scans increases to a certain value, the attenuation of the noise will no longer be obvious. Therefore, in the process of spectral acquisition, the setting of the number of scans should be chosen as a compromise between the spectral measurement time and the spectral signal-to-noise ratio.

20.3.5 *Effect of Instrument Performance*

Spectroscopic instrument performance includes the effective wavelength range, resolution, signal-to-noise ratio, baseline stability, wavelength accuracy and repeatability, absorbance accuracy and repeatability, temperature applicable range and resistance to voltage fluctuations, and many other aspects. The influence of spectral range, signal-to-noise ratio, and resolution on the calibration model has been introduced above. The influence of long-term stability and consistency of the instrument is briefly discussed below.

The stability and consistency of spectrometer is one of the restrictive factors for the wide popularization of spectrum combined with chemometrics. Spectrometer stability refers to the long-term repeatability of wavelength and absorbance, and spectrometer consistency refers to the accuracy of wavelength and absorbance. The calibration model is based on a large number of actual samples, and the determination of reference data requires considerable human and financial resources. Therefore, the built model must be able to be used for a long time and can be used on multiple instruments. This requires an instrument to have long-term stability and consistency between different instruments.

For different spectroscopic instruments, different measurement objects, and analysis parameters, the required instrument performance is also different. There is no unified or fixed standard, and it needs to be determined through experiments in the feasibility study.

On the basis of the long-term stability of the instrument, the calibration transfer methods in chemometrics can be used to solve the problem of model incompatibility caused by differences between instruments to a certain extent. This part of the content can refer to Chap. 17 of this book.

20.4 Outlook

Spectroscopy combined with chemometrics has been widely used in practice, especially playing an increasingly important role in on-site fast and industrial scenes. As the subject of chemometrics gradually enters the classroom of college students, this analysis technology will inevitably become more and more popular and common for chemical analysts and process analysts. However, as an emerging technology, if it wants to play its due role, there is still considerable work to be carried out, which mainly includes the following aspects.

(1) **Spectrometer hardware**

The hardware level is a key factor restricting the rapid development of this technology. Whether it is a laboratory, portable, or online spectrometer, the overall performance needs to be further improved, especially in the long-term stability and consistency of

the instrument, and a higher level of technical specifications needs to be formulated to realize the long-term effectiveness and universality of the calibration models. On the measurement accessories, it is necessary to develop more efficient, adaptable and targeted special accessories according to the specific application objects. On this basis, the miniaturization of the spectrometer as well as its supporting components is always an eternal goal.

The miniaturization of spectroscopic instruments and imaging instruments will bring important changes in terms of cost, performance, and application scenarios. Due to reasons such as data storage and computing speed, the application scenarios of these spectrometers and spectral imaging instruments will benefit from the development of technologies such as 5G communication, deep learning, and cloud computing platforms in the future, and become key components and important nodes in the construction of the Internet of Things.

Combination and fusion of multi-spectral instruments have been other significant research hotspots in recent years. For example, the combination of Raman and MIRS, LIBS and Raman, MIRS and NIRS, and various imaging combination of instruments, etc. can obtain more and richer chemical composition information [86, 87].

(2) **Experimental technology**

The experimental technology (spectral acquisition method and sample preprocessing, etc.) is an important link to determine the repeatability and accuracy of the analysis results. For different application objects, the experimental technology needs to be deeply and carefully studied in order to obtain high-quality spectra. Experimental techniques and measurement accessories are inseparable, and the two complement each other. The emergence of new experimental techniques will promote the improvement and development of measurement accessories. The commercialization of measurement accessories will improve the overall level of experimental technology and will also promote the development of spectroscopic instruments. SORS and transmission experimental techniques in Raman spectroscopy are examples, and the research and development of similar experimental techniques will remain one of the important development directions in the future.

(3) **Chemometric methods and software**

Chemometric methods and software are an important part of this analytical technique. Although the existing methods have been able to solve most of the problems, the research boom in this direction has not diminished. Although some of the research is only published as a paper, “practical application-driven” is still a strong driving force for its development. For example, the steps of establishing a quantitative analysis model in the multivariate calibration method are cumbersome and require more professional personnel to maintain, which limits its application scope to a large extent. Therefore, developing new algorithms to fundamentally solve the workload problem of modeling and its maintenance should be a key research direction.

In recent years, deep learning algorithms represented by convolutional neural networks (CNNs) have been used in the establishment of spectral quantitative and qualitative models [88]. Compared with traditional machine learning methods, CNNs

can gradually extract microscopic and macroscopic features hidden in spectral data through multiple convolutional layers and pooling layers. To a certain extent, the preprocessing of the spectrum and the selection of variables before modeling are reduced, and the workload of modeling is significantly reduced. Application of deep learning algorithms in spectral analysis has just begun, and issues such as network size, optimal selection of parameters, overfitting, and model interpretability are still worthy of further study. Strategies such as transfer learning, domain adaptation, and multi-task learning in deep learning are expected to provide new ideas for calibration transfer, and to a certain extent, solve the problem of the universality of quantitative and qualitative models in different instruments.

For some specific application requirements, multivariate calibration-free model methods based on strategies such as spectral fitting calculation are also attracting attentions. This kind of method can avoid the traditional complicated modeling and maintenance process and has certain advantages in oil products, drugs, and other fields [89–91].

It is worth mentioning that although new and effective chemometrics algorithms are constantly emerging, the functions of the computing software supporting the instruments are often not upgraded in time. This problem is expected to be solved through the wider application of cloud computing platforms.

(4) Maintenance of calibration models and data mining

The calibration models are built on the software and hardware platform based on the spectra of a large number of representative samples and their reference data. On the one hand, we need to build an official and commercial network model maintenance and sharing platform through different ways, so as to continuously expand and improve the established model database and make it play its due role in practical application. On the other hand, it is necessary to make full use of these data resources, especially the petrochemical, tobacco, grain spectral model libraries, etc., to further dig more and more useful information. In addition, it is also a very meaningful research work to dig out the information of process affecting product quality and reaction mechanism from a large number of process analysis spectra obtained in laboratories and industrial devices.

The rise of cloud computing platforms has improved the conditions for the processing and application of spectral big data. Through cloud computing platforms, NIR spectral data from different sources such as raw material production, online products, and laboratory research can be managed and stored. At the same time, the big data analysis method is used to analyze and mine the collected spectral big data, and then output the analysis results in a visual way, which can effectively provide data for the production process and product quality control in real time and provide a reliable basis for the management and storage of raw materials, the sales of products, and the monitoring and enforcement of relevant departments at higher levels.

Model maintenance is inseparable from calibration transfer, especially with the continuous popularity of portable and pocket-sized spectrometers on the consumer side, how to transfer the model on the laboratory host to the consumer-side instruments, and how to maintain the model using a large amount of spectral data without

standard reference data (or no label) on consumer instruments, new model maintenance, and model transfer methods will become more and more important, especially the standard-free model transfer method will play an extremely important role [92].

In recent years, some semi-supervised calibration transfer methods (it is not necessary to obtain the spectra of a set of standard samples on the master and the slave at the same time, but only need to obtain the spectra of a set of samples and their reference data on the slave) have been verified and applied, such as Dynamic Orthogonal Projection (DOP) and Semi-supervised parameter-free framework for calibration enhancement (SS-PFCE) [93, 94]. There are even some well-performing unsupervised (requiring only the spectra of a set of samples on the slave) calibration transfer method, such as Domain Invariant Partial Least Squares (di-PLS), Transfer Component Analysis (TCA), and Non-supervised parameter-free framework for calibration enhancement (NS-PFCE), etc. [94, 95].

(5) Spectral imaging technology

Spectral imaging techniques (NIRS, MIRS, Raman, Terahertz, LIBS, etc.) are an important branch of spectroscopy application combined with chemometrics [80]. Due to the inherent advantages of spectral imaging technology, the supporting hardware and algorithms will be the research hotspots in the future and will become an important supplement to traditional spectral analysis. In recent years, the miniaturization and portability of spectral imaging instruments have also been developed rapidly, and it has wide application potential in the fields of environment, geology, food, biomedicine, medicine, archaeology, cultural relics, forensic, etc.

(6) Expansion and deepening of practical applications

At present, although the applications of this kind of analysis technology have been studied and implemented in almost every field, their application breadth, depth, and role are still in the process of superfast development. According to different practical requirements, it is necessary to improve the whole set of platform technology to obtain the best application effect. At the same time, with the extensive and in-depth application of this type of analysis technology, it will also have a profound positive impact on the production process and production management and will play an important role in optimizing the production process and ensuring the quality of finished products.

At present, the industry is in a period of transition from the traditional production pattern to digital and intelligent pattern. Depth of information “Self-perception”, intelligent optimization “Self-decision”, and precise control “Self-execution” are the three key characteristics of smart factories. Among them, the depth of information “Self-perception” is the basis of intelligent industries. Molecular composition and physical property data of raw materials, intermediate materials and products are an important part of information perception. Modern process analytical technology with spectroscopy as one of the cores provides a very effective means for chemical information perception. Application of spectroscopic technology, especially online

technology in the fields of food, pharmacy, and chemical industry, has just begun with the general trend of refined management and intelligent processing and will bring changes to the process industries.

References

1. Diem M. Modern vibrational spectroscopy and micro-spectroscopy theory, instrumentation and biomedical applications. New Jersey: Wiley; 2015.
2. Bunaciu AA, Aboul-Enein HY, Hoang VD. Vibrational spectroscopy applications in biomedical, pharmaceutical and food sciences. Amsterdam: Elsevier; 2020.
3. Andrews J, Dallin P. Choosing your approach. *Spectrosc Eur.* 2003;15(3):27–9.
4. Caporaso N, Whitworth MB, Fisk ID. Near-infrared spectroscopy and hyperspectral imaging for non-destructive quality assessment of cereal grains. *Appl Spectrosc Rev.* 2018;53(8):667–87.
5. Kaavya R, Pandiselvam R, Mohammed M, et al. Application of infrared spectroscopy techniques for the assessment of quality and safety in spices: a review. *Appl Spectrosc Rev.* 2020;55(7):593–611.
6. Vodopyanov KL. Laser-based mid-infrared sources and applications. New Jersey: Wiley; 2020.
7. Lohumi S, Kim MS, Qin J, et al. Raman imaging from microscopy to macroscopy: quality and safety control of biological materials. *Trends Anal Chem.* 2017;93:183–98.
8. Wu JZ, Liu CL. Terahertz technology and its application in the detection of agricultural products. Beijing: Chemical Industry Press; 2020.
9. Baranska M, Schutze W, Schulz H. Determination of lycopene and β -carotene content in tomato fruits and related products comparison of FT-Raman, ATR-IR, and NIR spectroscopy. *Anal Chem.* 2006;78(24):8456–61.
10. McGill CA, Nordon A, Littlejohn D. Comparison of in-line NIR, Raman and UV-visible spectrometries, and at-line NMR spectrometry for the monitoring of an esterification reaction. *Analyst.* 2002;127(2):287–92.
11. Kim M, Noh J, Chung H. Comparison of near-infrared and Raman spectroscopy for the determination of the density of polyethylene pellets. *Anal Chim Acta.* 2009;632(1):122–7.
12. Salomonsen T, Jensen HM, Stenbak D, et al. Chemometric prediction of alginate monomer composition: a comparative spectroscopic study using IR, Raman NIR and NMR. *Carbohydr Polym.* 2008;72(4):730–9.
13. Sacre PY, Deconinck E, Beer TD, et al. Comparison and combination of spectroscopic techniques for the detection of counterfeit medicines. *J Pharm Biomed Anal.* 2010;53(3):445–53.
14. Yu XL, Sun DW, He Y. Emerging techniques for determining the quality and safety of tea products: a review. *Compr Rev Food Sci Food Saf.* 2020;19(5):2613–38.
15. Fakayode SO, Baker GA, Bwambok DK, et al. Molecular (Raman, NIR, and FTIR) spectroscopy and multivariate analysis in consumable products analysis. *Appl Spectrosc Rev.* 2020;55(8):647–723.
16. Nowak MR, Zdunek R, Plinski E, et al. Recognition of pharmacological bi-heterocyclic compounds by using Terahertz time domain spectroscopy and chemometrics. *Sensors.* 2019;19:3349.
17. Chen TT, Zhang TL, Li H. Applications of laser-induced breakdown spectroscopy (libs) combined with machine learning in geochemical and environmental resources exploration. *Trends Anal Chem.* 2020;133:116113.
18. Wang Q, Xie L, Ying Y. Overview of imaging methods based on Terahertz time-domain spectroscopy. *Appl Spectrosc Rev.* 2021;56:1–16.
19. Carruthers HL, Clark D, Clarke F, et al. Comparison of Raman and near-infrared chemical mapping for the analysis of pharmaceutical tablets. *Appl Spectrosc.* 2020;75(4):000370282095244.

20. Zhang H, Liu Z, Zhang J, et al. Identification of edible gelatin origins by data fusion of NIRS, fluorescence spectroscopy, and LIBS. *Food Anal Methods*. 2021;14(3):1–12.
21. Ballabio D, Robotti E, Grisoni F, et al. Chemical profiling and multivariate data fusion methods for the identification of the botanical origin of honey. *Food Chem*. 2018;266:79–89.
22. Ramirez CAM, Greenop M, Ashton L, et al. Applications of machine learning in spectroscopy. *Appl Spectrosc Rev*. 2020.
23. Biancolillo A, Marini F. Chemometric methods for spectroscopy-based pharmaceutical analysis. *Front Chem*. 2018;6:576.
24. Kuang B, Tekin Y, Mouazen AM. Comparison between artificial neural network and partial least squares for on-line visible and near infrared spectroscopy measurement of soil organic carbon, pH and clay content. *Soil Tillage Res*. 2015;146:243–52.
25. Buchmann NB, Josefsson H, Cowe IA. Performance of European Artificial Neural Network (ANN) calibrations for moisture and protein in cereals using the Danish Near-Infrared Transmission (NIT) network. *Cereal Chem*. 2001;78(5):572–7.
26. Clavaud M, Roggo Y, Degardin K, et al. Global regression model for moisture content determination using near-infrared spectroscopy. *Eur J Pharm Biopharm*. 2017;119:343–52.
27. Chu XL, Yuan HF, Luo XH, et al. Developing near infrared spectroscopy calibration model of molar ratio between methanol and isobutylene by support vector regression. *Spectrosc Spect Anal*. 2008;28(6):1227–31.
28. Le B. Application of deep learning and near infrared spectroscopy in cereal analysis. *Vib Spectrosc*. 2020;106:103007–9.
29. Nawar S, Mouazen AM. Comparison between Random Forests, Artificial Neural Networks and gradient boosted machines methods of on-line Vis-NIR spectroscopy measurements of soil total nitrogen and total carbon. *Sensors*. 2017;17:2428.
30. Wan XH, Li G, Zhang MQ, et al. A review on the strategies for reducing the non-linearity caused by scattering on spectrochemical quantitative analysis of complex solutions. *Appl Spectrosc Rev*. 2020;55(5):351–77.
31. Cernuda C, Lughofer E, Klein H, et al. Improved quantification of important beer quality parameters based on nonlinear calibration methods applied to FT-MIR spectra. *Anal Bioanal Chem*. 2017;409(3):841–57.
32. Yuan HF, Chu XL, Li HR, et al. Determination of multi-properties of residual oils using mid-infrared attenuated total reflection spectroscopy. *Fuel*. 2006;80(12–13):1720–8.
33. Blanco M, Romero MA. Near-infrared libraries in the pharmaceutical industry: a solution for identity confirmation. *Analyst*. 2001;126(12):2212–7.
34. Loudermilk JB, Himmelsbach DS, Barton FE, et al. Novel search algorithms for a mid-infrared spectral library of cotton contaminants. *Appl Spectrosc*. 2008;62(6):661–70.
35. Xu YQ, Zhu J, Qin Z, et al. Design and test of program for comparing FTIR spectra of chinese herbs by array of correlation coefficient. *Comput Appl Chem*. 2002;19(3):223–6.
36. Chu XL, Xu YP, Tian SB, et al. Rapid identification and assay of crude oils based on moving-window correlation coefficient and near infrared spectral library. *Chemom Intell Lab Syst*. 2011;107(1):44–9.
37. Asemani M, Rabbani AR, Sarafdokht H. Evaluation of oil fingerprints similarity by a novel technique based on FTIR spectroscopy of asphaltenes: modified moving window correlation coefficient technique. *Marine Petrol Geol*. 2020;104542.
38. Liang J, Li MG, Du Y, et al. Data fFusion of Laser Induced Breakdown Spectroscopy (LIBS) and Infrared Spectroscopy (IR) coupled with Random Forest (RF) for the classification and discrimination of compound *Salvia Miltiorrhiza*. *Chemomet Intell Lab Syst*. 2020;207:104179.
39. Lee W, Lenferink ATM, Otto C, et al. Classifying Raman spectra of extracellular vesicles based on convolutional neural networks for prostate cancer detection. *J Raman Spectrosc*. 2020;51:293–300.
40. Moghimi A, Aghkhani MH, Sazgarnia A, et al. Vis/NIR spectroscopy and chemometrics for the prediction of soluble solids content and acidity (pH) of Kiwi fruit. *Biosys Eng*. 2010;106(3):295–302.

41. Yun YH, Li HD, Deng BC, et al. An overview of variable selection methods in multivariate analysis of near-infrared spectra. *Trends Anal Chem.* 2019;113:102–15.
42. Roger JM, Biancolillo A, Marini F. Sequential preprocessing through orthogonalization (sport) and its application to near infrared spectroscopy. *Chemomet Intell Lab Syst.* 2020;199:103975.
43. Mishra P, Nordon A, Roger JM. Improved prediction of tablet properties with near-infrared spectroscopy by a fusion of scatter correction techniques. *J Pharm Biomed Anal.* 2021;192:113684.
44. Macho S, Larrechi MS. Near-infrared spectroscopy and multivariate calibration for the quantitative determination of certain properties in the petrochemical industry. *Trends Anal Chem.* 2002;21(12):799–806.
45. Duan YQ, Yang T, Kong XY, et al. Effects of sample granularity and spectral resolution on tobacco nicotine NIR predictive model. *J Yunnan Univ.* 2006;28(4):340–4.
46. Li JH, Qin XY, Zhang WJ, et al. Influence of sample loading and test conditions on NIR veracity and study of analysis error source. *Spectrosc Spect Anal.* 2007;27(9):1751–3.
47. Roudier P, Hedley CB, Lobsey CR, et al. Evaluation of two methods to eliminate the effect of water from soil Vis-NIR spectra for predictions of organic carbon. *Geoderma.* 2017;296:98–107.
48. Nawar S, Munnaf MA, Mouazen AM. Machine learning based on-line prediction of soil organic carbon after removal of soil moisture effect. *Remote Sens.* 2020;12(8):1308.
49. Debaene G, Niedzwiecki J, Pecio A, et al. Effect of the number of calibration samples on the prediction of several soil properties at the farm-scale. *Geoderma.* 2014;214–215:114–25.
50. Rossel RAV, Behrens T, Ben-Dor E, et al. A global spectral library to characterize the world's soil. *Earth-Sci Rev.* 2016;155:198–230.
51. Kuang B, Mouazen AM. Influence of the number of samples on prediction error of visible and near infrared spectroscopy of selected soil properties at the farm scale. *Eur J Soil Sci.* 2011;63(3):421–9.
52. Guerrero C, Wetterlind J, Stenberg B, et al. Do we really need large spectral libraries for local scale SOC assessment with NIR spectroscopy? *Soil Tillage Res.* 2015;155:501–9.
53. Luca F, Conforti M, Castrignano A, et al. Effect of calibration set size on prediction at local scale of soil carbon by Vis-NIR spectroscopy. *Geoderma.* 2017;288:175–83.
54. Goge F, Gomez C, Jolivet C, et al. Which strategy is best to predict soil properties of a local site from a national Vis-NIR database? *Geoderma.* 2014;213:1–9.
55. Chu XL, Yuan HF, Lu WZ. Effects of the accuracy of reference data on NIR prediction results. *Spectrosc Spect Anal.* 2005;25(6):886–9.
56. Chung H, Ku MS. Near-infrared spectroscopy for on-line monitoring of lube base oil processes. *Appl Spectrosc.* 2003;57(5):545–50.
57. Sorensen LK. True accuracy of near infrared spectroscopy and its dependence on precision of reference data. *J Near Infrared Spectrosc.* 2002;10:15–25.
58. Cayuela JA. Assessing Olive oil peroxide value by NIRS, and on reference methods. *NIR News.* 2017;28(3):12–6.
59. Bazar G, Kovacs Z. Checking the laboratory reference values with NIR calibrations. *NIR News.* 2017;28(3):17–20.
60. Coates DB. Is near infrared spectroscopy only as good as the laboratory reference values? An empirical approach. *Spectrosc Eur.* 2002;14(4):24–6.
61. Isengard HD, Merkh G, Schreib K, et al. The Influence of the reference method on the results of the secondary method via calibration. *Food Chem.* 2010;122(2):429–35.
62. Mark H, Workman J Jr. Units of measure in spectroscopy, Part III: summary of our findings. *Spectroscopy.* 2015;30:24–33.
63. Mark H. Effect of measurement units on NIR calibrations. *NIR News.* 2017;28(3):7–11.
64. Schaare PN, Fraser DG. Comparison of reflectance, interactance and transmission modes of visible-near infrared spectroscopy for measuring internal properties of Kiwifruit (*Actinidia chinensis*). *Postharvest Biol Technol.* 2000;20(2):175–84.
65. Johansson J, Sparen A, Svensson O, et al. Quantitative transmission Raman spectroscopy of pharmaceutical tablets and capsules. *Appl Spectrosc.* 2007;61(11):1211–8.

66. Aina A, Hargreaves MD, Matousek P, et al. Transmission Raman spectroscopy as a tool for quantifying polymorphic content of pharmaceutical formulations. *Analyst*. 2010;135(9):2328–33.
67. Schneider RC, Kovar KA. Analysis of ecstasy tablets: comparison of reflectance and transmittance near infrared spectroscopy. *Forensic Sci Int*. 2003;134(2–3):187–95.
68. Ito M, Suzuki T, Yada S, et al. Development of a method for the determination of Caffeine Anhydrate in various designed intact tablets by near-infrared spectroscopy: a comparison between reflectance and transmittance technique. *J Pharm Biomed Anal*. 2008;47(4–5):819–27.
69. Dowell FE, Pearson TC, Maghirang EB, et al. Reflectance and transmittance spectroscopy applied to detecting fumonisin in single Corn Kernels infected with *Fusarium verticillioides*. *Cereal Chem*. 2002;79(2):222–6.
70. Gishen M, Damberg RG, Cozzolino D. Grape and wine analysis—enhancing the power of spectroscopy with chemometrics. A review of some applications in the Australian Wine Industry. *Austr J Grape Wine Rese*. 2005;11(3):296–305.
71. Yang J, Tsai FP. Comparison of SPME/transmission IR and SPME/ATR-IR spectroscopic methods in detection of Chloroanilines in Aqueous Solutions. *Appl Spectrosc*. 2001;55(7):919–26.
72. Moros J, Garrigues S, de la Guardia M. Comparison of two partial least squares infrared spectrometric methods for the quality control of pediculosis lotions. *Analytica Chimica Acta*. 2007;582(1):174–80.
73. Koulis CV, Reffner JA, Bibby AM. Comparison of transmission and internal reflection infrared spectra of Cocaine. *J Forensic Sci*. 2001;46(4):822–9.
74. Jensen PS, Bak J. Near-infrared transmission spectroscopy of aqueous solutions: influence of optical pathlength on signal-to-noise ratio. *Appl Spectrosc*. 2002;56(12):1600–6.
75. Francisco J, Martin G. Optical path length and wavelength selection using Vis/NIR spectroscopy for Olive oil's free acidity determination. *Int J Food Sci Technol*. 2015;50:1461–7.
76. Manley M, Eberle K. Comparison of Fourier Transform near infrared spectroscopy partial least square regression models for South African extra virgin Olive oil using spectra collected on two spectrophotometers at different resolutions and path lengths. *J Near Infrared Spectrosc*. 2006;14(1):111–26.
77. Wang SP, Gong ZM, He YJ, et al. Effects of background and optical path length on predicting the contents of tea polyphenol with NIR models. *J Huazhong Agric Univ*. 2015;34(2):120–4.
78. Ozaki Y, Huck C, Tsuchikawa S, et al. Near-infrared spectroscopy: theory, spectral analysis, instrumentation, and applications. Singapore: Springer; 2020.
79. Cho S, Kwon K, Chung H. Varied Performance of PLS calibration using different overtone and combination bands in a near-infrared region. *Chemom Intell Lab Syst*. 2006;82(1–2):104–8.
80. Yang D, Liu X, Liu HG, et al. Effect of the near infrared spectrum resolution on the nitrogen content model in green tea. *Spectrosc Spect Anal*. 2013;33(7):1786–90.
81. Zhang Y, Tan LH, He Y. Study on brand discrimination of differential oil using near-infrared spectroscopy with different resolutions. *Spectrosc Spect Anal*. 2015;35(7):1889–93.
82. Liu XY, Tang XY, Sun BZ, et al. Comparative study on the prediction of beef nutrients by near infrared spectroscopy under two resolutions. *Sci Technol Food Indus*. 2013;34(3):302–5.
83. Dong GM, Yang RJ, Wu HY, et al. Effect of experimental parameters on quantitative model of soil moisture content by NIRS. *Spectrosc Spect Anal*. 2020;40(S1):91–2.
84. Chung H, Choi SY, Choo J, et al. Investigation of Partial Least Squares (PLS) calibration performance based on different resolutions of near infrared spectra. *Bull Korean Chem Soc*. 2004;25(5):647–51.
85. Cho S, Chung H. Investigation of chemometric calibration performance based on different chemical matrix and signal-to-noise ratio. *Anal Sci*. 2003;19(9):1327–9.
86. Watari M, Nagamoto A, Genkawa T, et al. Use of near-infrared–mid-infrared dual-wavelength spectrometry to obtain two-dimensional difference spectra of Sesame oil as inactive drug ingredient. *Appl Spectrosc*. 2020;1:000370282096919.
87. Muller-Maatsch J, Alewijn M, Wijten M, et al. Detecting fraudulent additions in skimmed milk powder using a portable, hyphenated, optical multi-sensor approach in combination with one-class classification. *Food Control*. 2021;121:107744.

88. Zhang XL, Yang J, Lin T, et al. Food and agro-product quality evaluation based on spectroscopy and deep learning: a review. *Trends Food Sci Technol.* 2021;112:431–41.
89. Shi ZQ, Hermiller J, Munoz SG. Estimation of mass-based composition in powder mixtures using Extended Iterative Optimization Technology (EIOT). *AIChE J.* 2019;65(1):87–98.
90. Li J, Chu X. Rapid determination of physical and chemical parameters of reformed gasoline by NIR combined with Monte Carlo virtual spectrum identification method. *Energy Fuels.* 2018;32(12):12013–20.
91. Sun X, Yuan H, Song C, et al. Rapid and simultaneous determination of physical and chemical properties of asphalt by ATR-FTIR spectroscopy combined with a novel calibration-free method. *Construct Build Mater.* 2020;230:116950.
92. P Mishra, R Nikzad-Langerodi, F Marini, et al. Are standard sample measurements still needed to transfer multivariate calibration models between near-infrared spectrometers? The answer is not always. *Trends Anal Chem.* 2021;143:116331.
93. Mishra P, Roger JM, Rutledge DN, et al. Two standard-free approaches to correct for external influences on near-infrared spectra to make models widely applicable. *Postharvest Biol Technol.* 2020;170:111326.
94. Zhang J, Li BY, Hu Y, et al. A parameter-free framework for calibration enhancement of near-infrared spectroscopy based on correlation constraint. *Analytica Chimica Acta.* 2020;1142:169–78.
95. Mishra P, Nikzad-Langerodi R. A brief note on application of domain-invariant PLS for adapting near-infrared spectroscopy calibrations between different physical forms of samples. *Talanta.* 2021;232:122461.