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

X. Chu et al., *Chemometric Methods in Analytical Spectroscopy Technology*,
https://doi.org/10.1007/978-981-19-1625-0_19

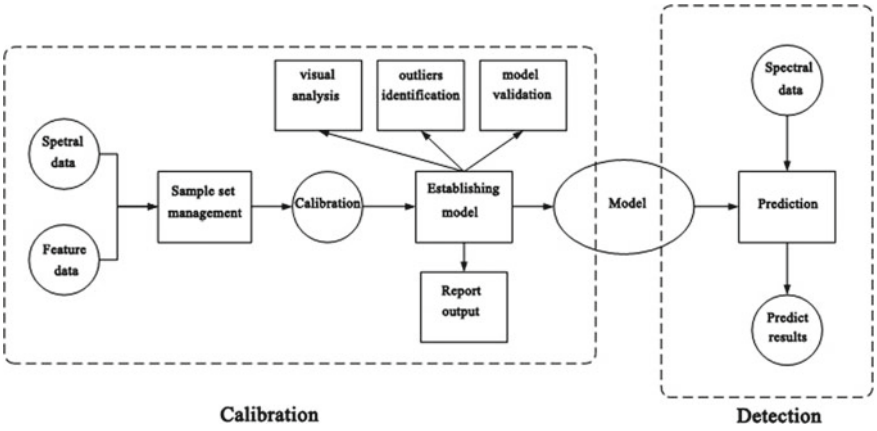

Fig. 19.1 Scheme of chemometrics software for spectral analysis

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

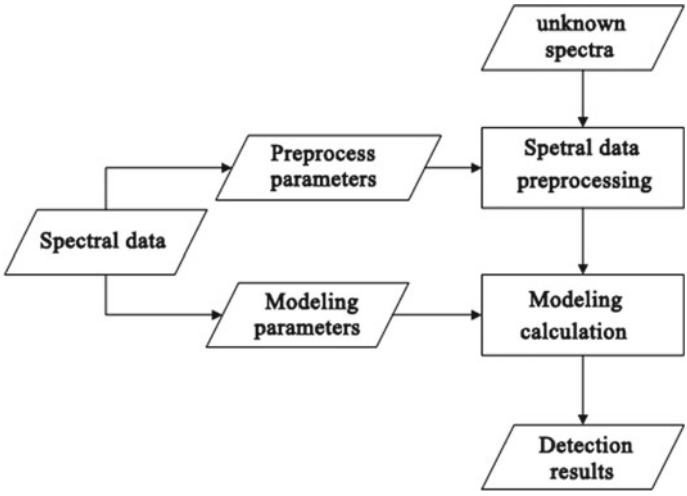

Fig. 19.2 Basic steps of predictive analysis of unknown samples

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 NIRCal, 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 assess 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.chimiometrie.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/puneetmishra2/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://michem.unimib.it/download/matlab-toolboxes/	Supervised pattern recognition toolbox
FRUITNIR	https://github.com/puneetmishra2/FRUITNIR	Migration component analysis toolbox [19]
MEDA-toolbox	https://github.com/josecamachop/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. Torniainen 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.