Chemometrics in Nondestructive Quality Evaluation

Md. Nahidul Islam

Abstract Because of the widespread use of nondestructive measurement techniques, such as spectroscopy, spectral imaging, which allow scientists to swiftly obtain a complete spectrum for a single sample. The datasets nowadays tend to have a smaller number of samples and a larger number of variables. In order to extract information from these high-dimensional and high-volume data, traditional univariate analysis is widely considered to be inadequate. Effective data preprocessing and applying the appropriate chemometric methods are required to gain insights and obtain essential information from these datasets. In this way, chemometrics has made substantial advancements and recognition in nondestructive quality assessment of foods. The purpose of this chapter is to introduce an understanding of the latest ideas, methodologies, techniques, and fundamental processes used during nondestructive analysis of fruits and vegetables, where chemometrics and/or multivariate analytical approaches were performed.

Keywords Principal component analysis (PCA) · Partial least squares regression (PLSR) · Partial least squares discriminant analysis (PLS-DA)

1 Introduction

Analytical techniques used in laboratories are frequently insufficient since they necessitate a large number of samples, a longer time to receive results, and highly technical personnel (Zou & Zhao, 2015). In an environment where speed is critical, engineering advances must require fewer samples or, at the very least, no one (nondestructive techniques): a) they must provide prompt, if not immediate, responses in order for the operator to make an informed decision on the next steps to regulate or release the product to the market; b) they must be simple to use in order to encourage their use across the manufacturing chain, where analytical laboratories

M. N. Islam (✉)
Department of Agro-Processing, Bangabandhu Sheikh Mujibur Rahman Agricultural University, Gazipur, Bangladesh
e-mail: nahidul.islam@bsmrau.edu.bd

P. B. Pathare, M. S. Rahman (eds.), *Nondestructive Quality Assessment Techniques for Fresh Fruits and Vegetables*, https://doi.org/10.1007/978-981-19-5422-1_14

are not always available. As a result, technology needs to be adjusted to a new strategy of production: the use of sensors and the necessarily associated information extraction system, which allows “measurement,” to meet the demands of the agri-food stakeholders. Moreover, the manufacturers of technologies often provide devices that require calibration phases not always easy to perform but that are often the subject of actual researches. These are particularly complex when similar processes need to run repeatedly (Eriksson et al., 2013).

Hence, chemometrics techniques in nondestructive quality evaluation aim to produce an empirical or semi-empirical model from data that may be used to predict one or more chemical properties of a system from observations (Cocchi, 2017). Chemometrics employs mathematical and statistical methodologies to optimize experimental processes through the scientific design of experiments, to treat the experimental data and to extract as much relevant chemical information as possible from generated data (Guidetti et al., 2012). Chemical systems are often multivariate, which means that numerous information are obtained at the same time. As a result, the majority of chemometric procedures fall within the category of analytical methods known as multivariate statistical analysis, which contains many measurements on a number of individuals, objects, or data samples. Therefore, multiple measurements and analyses of the variable dependence are central to chemometrics (Marini, 2013). This chapter aims to present an overview and to provide a clear understanding of the chemometric methods with their advantages and disadvantages used in the nondestructive quality evaluation of fruits and vegetables.

2 Major Chemometric Tools in Food Analysis

Chemometric approaches are used to optimize the experimental process and extract relevant chemical information from massive quantities of data, identify hidden relationships, and provide a visual approach. There are several chemometric approaches: design of experiment (DoE), preprocessing, explorative analysis, classification, regression, validation, feature selection, multiway analysis, etc. These methods are utilized for nondestructive quality analysis of fruits and vegetables, as well as in other areas of food science and technology. The chosen method is determined by the challenge, type of experimental data, as well as by considering the pros and cons of that particular chemometric approach (Martens & Martens, 2001).

2.1 Design of Experiment

DoE technique ensures representativeness of the sample, allows for the evaluation of the primary sources of variability, and is the most effective way to optimize analytical measurement processes (Lawson, 2014). Experimental designs are

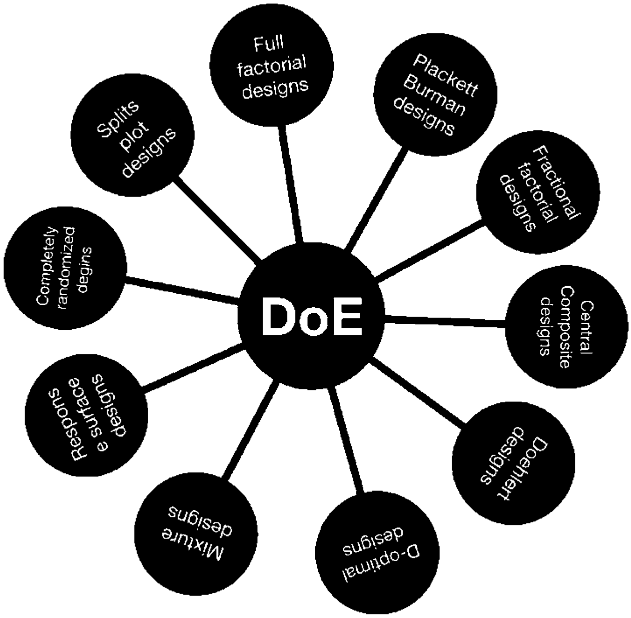

Fig. 1 Common experimental design methods in chemometrics

frequently neglected or undervalued; however, in order to illustrate the need for variable optimization as well as the development of adequate methods for carrying out the tests, a correct experimental design must be established in advance (Granato and de Araújo Calado, 2013; Leardi, 2009; Wold et al., 2004). It is essentially decided how to carry out scientific research. When developing new detection systems, the optimization protocol is especially important. A well-defined DoE not only allows scientists to investigate different factors and their interactions, but it also saves money (Granato and de Araújo Calado, 2013; Leardi, 2006, 2009; Wold et al., 2004). Figure 1 illustrates various methods of experimental designs used in chemometrics.

The use of a specific design depends on the specific problem statement. For example, some methods are used in optimization while some methods are used in screening experiments. The advantages and disadvantages of some common experimental designs are presented in Table 1.

2.2 Preprocessing of Data

After collecting data, data preprocessing is frequently the deciding factor between excellent and poor chemometric models (Rinnan, 2014). Preprocessing is used to reduce variation that is not connected to the topic of interest, allowing the variation

Table 1 Advantages and disadvantages of some common experimental designs

Experimental designs	Advantages	Disadvantages	References
Full factorial designs	Allows scientists to look at the effect of treatment	Difficult to ensure the elimination of in-group variations	(Ebrahimi-Najafabadi et al. (2014)
	Good for experiments with more than five factors	Need to report experiments in the standard order to avoid systematic error.	
		Bad for more than five factors	
Randomized block designs	Opportunity to block nuisance factors inside the block.	Need to eliminate the contribution of nuisance factors.	Huynh and Feldt (1976)
Plackett Burman designs	Good for more factors.	These designs have run numbers that are a multiple of four.	Vanaja and Shobha Rani (2007)
Fractional factorial design	Possible to give a first interpretation just looking at the highly informative data.	Require many runs.	Rakić et al. (2014)
	Use low-resolution designs for screening among main effects and use higher resolution designs when interaction effects and response surfaces need to be investigated.		
Response surface methodology	Quadratic models are almost always sufficient for industrial applications.	Response surface models may involve just main effects and interactions, or they may also have quadratic and possibly cubic terms to account for the curvature.	Bezerra et al. (2008)
Central composite design	Start with factorial or fractional factorial design (with center points) and add "star" points to estimate curvature.	The position of the star points is important.	Asghar et al. (2014)

of interest to stand out more and be more easily modeled (Islam et al., 2018b). There are some situations where enhancing spectral features is essential; for example, in a situation where intriguing spectral characteristics differ slightly from the global intensity, where small peaks are difficult to see in the presence of a large one and due to overlapped peaks. According to Roger et al. (2020), there are several types of systematic variations that are not related to topics of interest, for example, the shift of the baseline due to light scattering effects as a result of various particle sizes, offsets of baseline due to differences in instrumentation, and variations in the signal intensities due to size, shape, and volume of the sample. Figure 2 illustrates the available data preprocessing methods in chemometrics.

Among all the preprocessing methods, mean centering, standard normal variate (SNV) normalization, baseline correction, orthogonal signal correction (OSC),

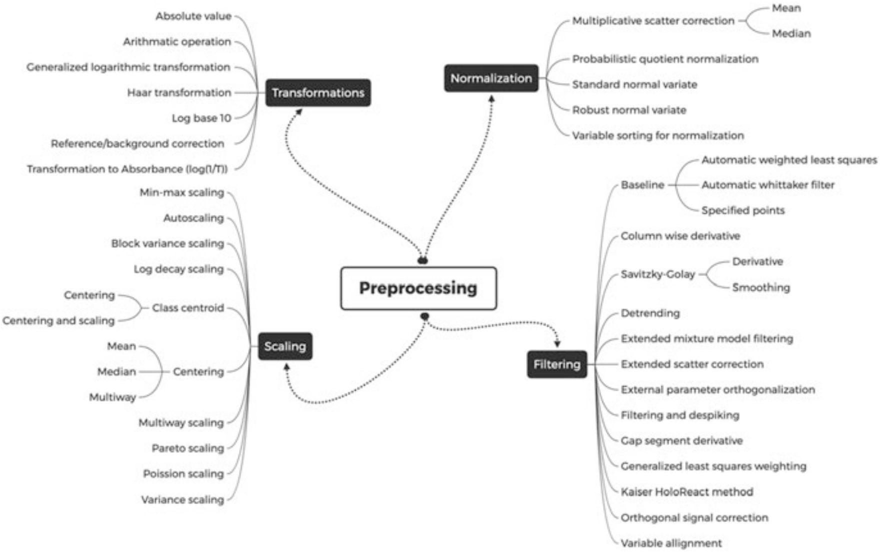

Fig. 2 Available preprocessing methods in chemometrics

Savitzky–Golay (SAV-GOL) smoothing and derivatization, and multiplicative signal correction (MSC) are the most common data preprocessing methods used in chemometrics (Vidal and Amigo 2012). The visualization of the data and the removal of severe bands that are driven by noise are the first steps in near-infrared (NIR) spectroscopic preprocessing. Then, to reduce any high-frequency noise, window-based smoothing techniques can be performed. The SAV-GOL algorithm is a widely used approach for reducing high-frequency noise. It includes fitting a polynomial of chosen order into a band of the defined size that is moved throughout the entire spectrum (Rinnan et al., 2009). In an ideal circumstance, the smoothed spectra should be ready for regression or classification modeling in the presence of just absorption features. The smoothing stage is frequently followed by scattering correction methods due to the prominence of scattering effects. The estimation of the second derivative of the spectra is the most frequent method since it can quickly remove first-order additive (baseline shift) effects and also show underlying peaks that would otherwise be invisible (Rinnan et al., 2009). Another widely used method is SNV, which includes subtracting each spectrum’s mean spectral intensity from each intensity response and then dividing by its spectral-domain standard deviation (Barnes et al., 1989). SNV can be used to eliminate additive and multiplicative effects. In NIR modeling, both the second derivative and SNV are quite useful and usually increase model prediction performance. Another prominent method is MSC, which assumes the spectrum has a multiplicative, additive, and residual component (Isaksson & Næs, 1988). In order to describe these impacts, the Extended MSC (EMSC) model incorporates higher-order complex relations (Martens et al., 2003).

Table 2 Some common preprocessing methods with their pros and cons

Preprocessing methods	Advantages	Disadvantages	References
Mean centering	It may reduce model complexity by reducing factors. Provide better calibration model.	Risk of uncertainty. Poor prediction model outside of calibration space.	Enders and Tofighi (2007)
Autoscaling	All variables become equally important.	Inflation of the measurement errors.	van den Berg et al. (2006)
Variance scaling	A weak signal may provide information.	Noisy parts get weight. Difficult to interpret loading and beta coefficient.	Noda (2008)
Pareto scaling	Isolated weak signal provides information. Does not give much weights to noisy part as variance scaling.	Difficult to interpret loading and beta coefficient.	Kasprzak and Lewis (2001)
Logarithmic transformation	Transform multiplicative effect to additive effects.	Not applicable to data containing negative or zero values.	Bartlett and Kendall (1946)
SAV-GOL	Derivation may enhance less apparent spectral features.	A wide window may remove important information, while a narrow window keeps lots of noises.	Press and Teukolsky (1990)
MSC	Good at NIR reflectance measurement of fruits and vegetables.	Not good at predicting physical properties. Depends on the reference spectral data.	Isaksson and Næs (1988)
SNV	Effective in correcting systematic effects, and NIR scattering effects.	Contains negative value in the processed data.	Barnes et al. (1989)
Detrending	Can deal with data collected over time where baseline/background may have drift.	Processed spectra have a negative value.	Tanabe et al. (2002)

MSC multiplicative scatter correction, *SNV* standard normal variate, *SAV-GOL* Savitzky–Golay

In many cases, the scattering effect is also important to perfectly describe the quality of fresh fruits and vegetables. Therefore, the removal of the scattering effect in those cases may lead to the wrong chemometric model (Mishra et al., 2021). Robust normal variate (RNV) (Guo et al., 1999), probabilistic quotient normalization (PQN) (Dieterle et al., 2006), and variable sorting for normalization (VSN) (Rabatel et al., 2020) have all been offered as improvements and alternatives to SNV. To summarize, there are numerous chemometric preprocessing approaches for removing/reducing scattering effects from spectral data. Table 2 summarizes the advantages and disadvantages of the common data preprocessing methods used in chemometrics.

3 Principal Component Analysis (PCA)

Principal component analysis (PCA) is one of the most important and powerful methods in chemometrics (Bro & Smilde, 2014). PCA is a bilinear reduction approach that may condense enormous amounts of data into a few parameters known as principal components (PCs) or latent variables, which reflect the levels, differences, and similarities among the samples and variables that make up the modeled data. A linear transformation is used to accomplish this goal, with the constraints of conserving data variance and imposing orthogonality on the latent variables (Smilde et al., 2005).

3.1 PCA Data Analysis

PCA can be used to visualize the *X* data matrix in the multivariate space, cluster identification and detection of outliers, reducing the dimensionality of the data and removing the noise. The starting point for PCA is a matrix of data with *N* rows (observations) and *M* columns (variables), here denoted by *X*. Technically, PCA seeks lines, planes, and hyperplanes in *K*-dimensional space that best approximate the data in terms of least squares. It is obvious that a line or plane that is the least squares approximation of a set of data points minimizes the variance of the coordinates on the line or plane (Wold et al., 1987).

The first PC is the line in *K*-dimensional space that best approximates the data in terms of least squares. The line intersects the mean point. Hence, each observation can be projected onto this line to obtain a coordinate value along the PC line. This new coordinate value is referred to as a score. A second PC is a line in *K*-dimensional variable space that is orthogonal to the first PC. This line likewise crosses through the average point and enhances the *X*-data approximation as much as feasible. If *X* is a data matrix with *N* rows and *M* columns, and with each variable being a column and each sample a row, PCA decomposes *X* as the sum of *r* t_i and p_i, and where *r* is the rank of the matrix *X* (Eq. 1).

$$X = t_1 pT_1 + t_2 pT_2 + \ldots\ldots + t_m pT_m + \ldots\ldots + t_r pT_r \qquad (1)$$

$$r \leq min\ \{M, N\}$$

$$X = t_1 pT_1 + t_2 pT_2 + \ldots\ldots + t_m pT_m + E \qquad (2)$$

The amount of variance captured by t_i, p_i pairs are ordered. The scores are vectors that include information about how the samples relate to one another. The vectors are called loadings and they provide information on how the variables interact. In general, after *m* components, the PCA model is usually truncated, and the small variance factors are consolidated into a residual matrix *E* (Eq. 2).

The basic premise is that the investigated systems are "indirectly observable," meaning that the relevant phenomena that cause data variation/patterns are concealed and not directly measurable/observable. This is where the phrase “latent variables” comes from. Latent variables (PCs) can be expressed as scatter plots in the Euclidean plane once they have been discovered. A loading plot can be discussed in conjunction with the associated score plot, which is generated for the same pair of PCs, or it can be directly shown in the same figure, which is called a biplot. It becomes easier to explain the groups or patterns observed in the PC space in terms of the original variables in this way. Although the biplot format for spectral data is difficult to visualize, specific spectral regions that are responsible for the separation of process phases can be highlighted.

3.2 Outlier Detection

Q residuals are the sum of squared residuals for each sample. In other words, *Q* is a measure of the distance of a sample from the PCA model. Therefore, a higher *Q* value means a lower model fit. Hotelling’s T^2 is the sum of normalized squared scores (Hotelling, 1947). T^2 is a measure of the variation in each sample within the PCA model (Fig. 3). Figure 4 presents *Q* residuals versus Hotelling’s T^2 plot, which is very useful to determine the outlier sample.

The region of extreme samples (bottom right) exhibits unusual behavior since they adhere to the variable correlation structure recorded by the PCA model while achieving high scores in the scores space. Because they pull the PC axes toward them, these samples with high Hotelling’s T^2 values are said to have strong leverage. The region far from model samples (top left): these samples, with high Q residuals values, appear to be “well behaving” when projected onto model space because they

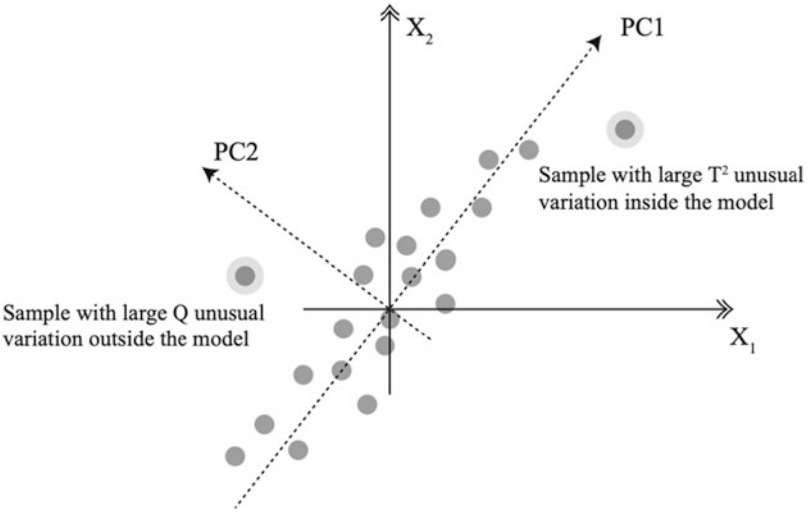
X_2
PC1
PC2
Sample with large T^2 unusual variation inside the model
Sample with large Q unusual variation outside the model
X_1

Fig. 3 Graphical representation of the principal components space for a two-component model

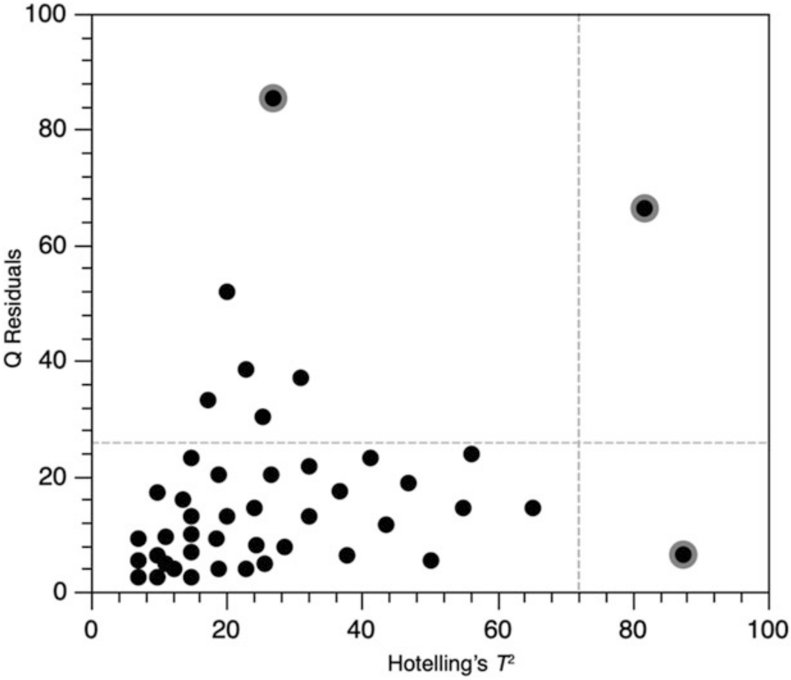

Fig. 4 Q residuals versus Hotelling's T^2 plot

share some characteristics with the modeled category, but they are not well modeled because part of their variation is not accounted for by the model. The anomalous, extreme, and non-modeled samples that have both T^2 and Q high values belong in the outliers region (top right) (Westerhuis et al., 2000).

4 Partial Least Squares Regression

Partial least squares regression (PLSR) is a regression extension of PCA, which is used to connect the information in two blocks of variables, *X* and *Y*, to each other (Wold et al., 2001). PLSR is a method of relating two data matrices, *X* and *Y*, to each other by a linear multivariate model. PLSR stands for projections to latent structures by means of partial least squares. It derives its usefulness from its ability to analyze data with many noisy, collinear, and even incomplete variables in both *X* and *Y*. For parameters related to the observations (samples, compounds, objects, items), the precision of a PLSR model improves with the increasing number of relevant *X*-variables. This corresponds to the intuition of most chemists, technicians, and engineers that many variables provide more information about the observations than just a few variables do (Martens & Naes, 1991).

PLSR can be seen as a particular regression technique for modeling the association between X and Y, but it can be seen as a philosophy of how to deal with complicated and approximate relationships (Geladi & Kowalski, 1986). Because PLSR considers not just the correlation between two variables but also the amount of variation in each, the criterion for defining the PLS latent variables is formulated using covariance, which is a good metric of interrelation, component-based criterion because converting it to a global loss function is quite challenging. As a result, PLSR is a sequential algorithm: PLS latent variables are computed in such a way that the first PLS component is the dependent variables' direction of maximum covariance. The second PLS component, for example, is orthogonal to the first and has the highest residual covariance, and so on (Wold et al., 1983).

Outlier samples that are far from the center within the space given by the PLS model can be detected using plots of leverage or Hotelling's T^2. The critical limit for Hotelling's T^2 statistics is based on an F-test (Hotelling, 1992), while the critical limit for Leverage is based on ad hoc knowledge (Martens & Naes, 1991). A predicted versus measured plot should, in a good PLS model, display a straight-line relationship between predicted and measured values, ideally with a slope of one and a correlation close to one. A residual plot may be plotted against the value of the y-variable to check that the residuals are not depending on the value of Y. Outliers of various types, such as samples with significant residuals and influential samples, are commonly detected using F residuals versus Hotelling's T^2 plot. Outliers are samples with high residual variance or those that lie at the top of the plot. Influential samples are those that have high leverage, that is, those that lie to the right of the plot (Rousseeuw & Leroy, 1987). This indicates that they are attracting the model in order for it to better describe them. Influential samples are not always risky if the variables follow the same pattern as the more "average" samples. A sample with significant residual variance and leverage is referred to as a "potential outlier," and in the presence of these outliers, the model focuses on the differences between that sample and the outlier rather than defining more general traits common to all samples.

5 Classification

Datasets are frequently made up of samples from various groups or "classes." Groups may differ for a variety of reasons, including variations in sample preparation, chemical constituent types such as aromatic, aliphatic, etc., or process conditions. A number of approaches for classifying samples based on measured responses have been developed, as shown in Fig. 5. Cluster analysis and unsupervised pattern recognition are methods for attempting to find groups or classes without the use of prior knowledge regarding class memberships. On the other hand, classification or supervised pattern recognition are terms used to describe methods that leverage known class memberships (Ballabio & Consonni, 2013).

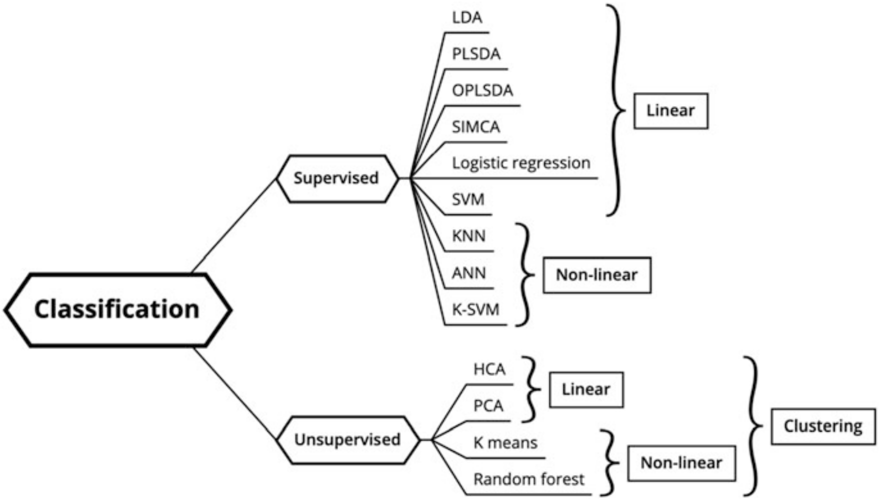

Fig. 5 Overview of the classification techniques in chemometrics

Most cluster analysis approaches are based on the concept that samples that are close together in the measurement space are comparable and so likely to belong to the same class. However, there are other ways to define the distance between samples. The most popular is the simple Euclidean distance. A Mahalanobis distance accounts for the fact that variance in some directions is substantially greater in some datasets than in others. As a result, distance in some directions is more relevant than the distance in others (De Maesschalck et al., 2000).

Soft Independent Modeling of Class Analogy (SIMCA) makes use of the model features and incorporates information about the calibration data types. A SIMCA model is made up of a set of PCA models, one for each class in the dataset (Wold, 1976). The number of major components in each class can vary. The number is determined by the data in the class. Each PCA sub-model includes all of the standard components of a PCA model, such as the mean vector, scaling information, preprocessing such as smoothing and derivatizing, and so on. The oldest and most studied supervised pattern recognition approach is linear discriminant analysis (LDA) (Fisher, 1936). It is a linear approach in the sense that the decision boundaries dividing the classes of variables in their multidimensional space are linear surfaces (hyperplanes). The purpose of LDA is to identify the ideal linear surface in a multidimensional space that corresponds to the best two-dimensional straight line. Partial least squares discriminant analysis (PLS-DA) is quite similar to LDA, another common discriminating approach. Indeed, Barker and Rayens (2003) demonstrated that PLS-DA is simply the inverse-least squares method to LDA, producing essentially the same result but with the noise reduction and variable selection benefits of PLS. PLS is used in PLS-DA to create a model that predicts the class number for each sample (Næs et al., 2002). Table 3 summarizes the advantages and

Table 3 Advantages and disadvantages of some common chemometric methods

Chemometric tools	Advantages	Disadvantages
PCA	PCA shows the similarities and differences between the samples as well as the relationships between the variables quickly and easily.	It does not allow for the classification of samples and the assignment of a class to each one.
PLSR	The ability to handle more descriptor variables than compounds, nonorthogonal descriptors, and multiple biological outcomes with more predictability and a considerably reduced probability of chance correlation.	Increased risk of missing "real" correlations and sensitivity to the relative scaling of the descriptor variables.
PLS-DA	The categorization model may be created quickly and easily, and the results are usually extremely good.	Classification errors might occur if the distinction between the areas of the various classes is not clear enough.
SIMCA	It can train a binary classification model exclusively with the target class since it defines an acceptance region that contains all of the target class's objects/samples.	There may be overlapping between acceptance areas that contain samples from distinct classes in models trained with two or more classes. As a result, some samples may be assigned to one or more classes.
OPLS-DA	To improve classification, increase the difference in means as well as the difference in within-class variance between the two classes.	It is impossible to see different treatment effects among the participants in the community.
KNN	Method of application that is simple to use.	Because the dominant class affects the classification, if there are more samples of one class than the other (skewed distribution of classes), the samples may be incorrectly classified.
SVM	When the demarcation between the regions of the different classes of samples is not sufficiently obvious, this can be used to get around the technical challenge.	Alternative kernel functions must be utilized for non-linear SVM models. As a result, the model's development is complex, and a lot of informatics resources are required.
Random forest	Because the variance is reduced, this is an excellent choice for unstable models or class imbalance issues. Overfitting is avoided as much as possible.	Because the classification is not displayed as a graphical tree, understanding the results is difficult.

PCA principal component analysis, *PLSR* partial least squares regression, *PLS-DA* partial least squares discriminant analysis, *SIMCA* soft independent modeling of class analogy, *OPLS-DA* orthogonal partial least squares discriminant analysis, *KNN* K-nearest neighbors, *SVM* support vector machines

disadvantages of the most common chemometric methods used in nondestructive quality evaluation.

6 Model Validation

The most conservative validation method is to run the model on a sufficiently large representative independent test set. Several methodologies can be used to quantify sources of variation that are in principle unknown for future objects in order to make a model more robust to changes in the sample matrix, raw materials, chemical reagents, and so on (Westad & Marini, 2015). Though the goal is to have enough items to set aside a decent amount as a test set, this is not always practicable due to factors such as sample costs or reference testing. Cross-validation is the best alternative to using an independent test set for validation (Westad & Kermit, 2003).

Cross-validation is a practical and reliable way to test the significance of a PLS model. This procedure has become standard in chemometric analysis and is incorporated in one form or another in most commercial software. With CV, the basic idea is to keep a portion of the data out of the model development, develop a number of parallel models from the reduced data, predict the omitted data by different models, and finally compare the predicted values with the actual ones. The square differences between predicted and observed values are summed to form the predictive residual sum of squares, which is a measure of the predictive power of the tested model (Stone, 1974). Various ways of cross-validation is available, for example, full cross-validation, segmented cross-validation, systematic segmented cross-validation, and validating across categorical information about the objects (Kos et al., 2003).

7 Model Performances

The number of latent variables in a PLSR model is determined by minimizing the root mean square error of cross-validation (RMSECV). Given the data and number of latent variables, overfitting is a possibility, but the purely data-driven strategy is the best option. The root mean square error of prediction (RMSEP) is a direct estimate of the model's prediction error in PLSR modeling. The RMSEP can be calculated using Eq. (3). Alternatively, the PLSR model's accuracy and precision are represented by the bias and standard error of performance (SEP), respectively. Equations (4) and (5) can be used to compute the SEP and bias, respectively, where, and are the predicted and measured values of the ith observation in the test set and n is the size of the validation set (Amigo, 2021).

The accuracy, precision, and linearity of the models can be used to assess their performance. The root mean square error of calibration (RMSEC), RMSECV, RMSEP, and bias can all be used to express the model's correctness. The SEP can be used to examine the PLSR model's precision, and R^2 can be used to assess linearity using a linear fit of predicted versus measured values. Low RMSEC, RMSECV, RMSEP, and SEP values, as well as a high R^2 value, indicate a good model (Islam et al., 2018a).

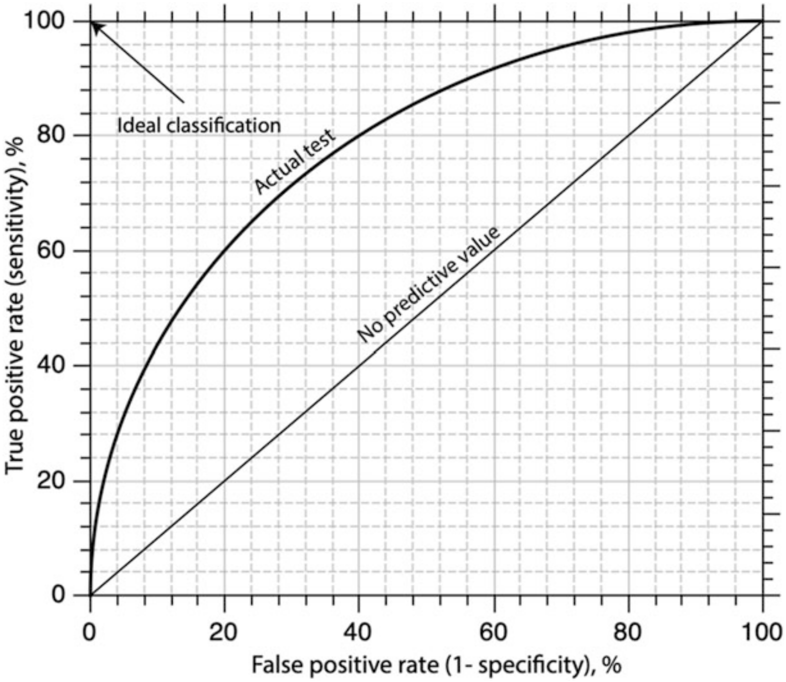

Fig. 6 Receiver operating characteristic (ROC) curve

$$RMSEP = \sqrt{\frac{\sum\left(\hat{y}_{cal} - y_{val}\right)^{2}}{n}} \qquad (3)$$

$$SEP = \sqrt{\frac{\sum_{i=1}^{n}\left(\hat{y}_{i} - y_{i} - bias\right)^{2}}{n-1}} \qquad (4)$$

$$bisa = \sum_{i=1}^{n}\frac{\left(\hat{y}_{i} - y\right)}{n} \qquad (5)$$

The receiver operating characteristic (ROC) curve demonstrates the trade-off between sensitivity (or TPR) and specificity (1 − FPR). Classifiers that produce curves closer to the top-left corner perform better. A random classifier is expected to give points along the diagonal as a baseline (FPR = TPR). The test becomes less accurate when the curve approaches the ROC space's 45-degree diagonal. The class distribution has no bearing on the ROC. This makes it ideal for testing classifiers that anticipate infrequent events like rotten products. Using accuracy (*TP* + *TN*)/(*TP* + *TN* + *FN* + *FP*) to evaluate performance, on the other hand, would favor classifiers that always predict a negative outcome for uncommon events (Fig. 6).

The confusion matrix and ROC curves can also be used to measure a classifier's error rate. To understand the confusion matrix, consider a classification problem where there are two classes, *X* is the negative class, and *Y* is the positive class. And four possible outcomes: sample from class *X* is assigned to class *X*, the sample from class *Y* is assigned to class *Y*, the sample from class *X* is assigned to class *Y* (false positive), and sample from class *Y* is assigned to class *X* (false negative).

To keep track of these various outcomes, a confusion matrix (or contingency table) is utilized. The confusion matrix's columns correspond to the samples' actual classes, while the rows correspond to the assigned classes. The main diagonal of the matrix shows the number of correctly categorized samples in each class, while the off-diagonal elements show the number of wrongly classified samples. The off-diagonal members of the matrix are zero if the data is perfectly classified. The accuracy, sensitivity (also known as precision, recall, hit rate, or true-positive rate), false-positive rate (also known as false alarm rate), and specificity of the classification can all be determined using the confusion matrix.

$$\begin{array}{lcc} (\mathit{Confusion\ matrix}) & \underline{\mathit{True\ Group}\ \boldsymbol{X}} & \underline{\mathit{True\ Group}\ \boldsymbol{Y}} \\ \mathit{Classified\ as\ Group}\ \boldsymbol{X} & m & n \\ \mathit{Classified\ as\ Group}\ \boldsymbol{Y} & o & p \end{array}$$

Equation (6) gives the accuracy of the classification, where *m* is the number of samples from class ***X*** that are assigned to class ***X*** by the classifier, *p* is the number of samples from class ***Y*** that are assigned to class ***Y*** by the classifier, *n* is the number of samples from class ***Y*** assigned to class ***X*** by the classifier, and *o* is the number of samples from class ***X*** assigned to class ***Y*** by the classifier. The sensitivity of the classification is given by Eq. (7), and the false-positive rate is given by Eq. (8). The specificity of the classification is given by Eq. (9) (Islam et al., 2018b).

$$\mathit{Accuracy} = \frac{m+p}{m+n+o+p} \tag{6}$$

$$\mathit{Sensitivity} = \frac{p}{n+p} \tag{7}$$

$$\mathit{False\ positive} = \frac{o}{o+m} \tag{8}$$

$$\mathit{Specificity} = \frac{m}{m+o} \tag{9}$$

8 Variable Selection Methods

A model that uses the entire spectral range may be at threat of overfitting, resulting in decreased predictive performance. Furthermore, a spectrum contains a significant quantity of data, most of which is unnecessary. Given the large redundancy in spectroscopic data, variable selection can typically improve chemometric models (Fig. 7).

The advantages of variable selection have been concluded in the following three aspects: (a) improve the prediction accuracy of the model because of the elimination of uninformative variables that must lead to less precision as proved theoretically; (b) selecting wavelengths probably responsible for the property of interest makes the model more interpretative; and (c) enhance the computational efficiency for modeling with a small number of variables. The advantages and disadvantages of the most common variable selection methods are presented in Table 4.

9 Multiway Analysis

In some cases, data structures are more complex than typical. Most multivariate methods are designed to work with matrices, which can be thought of as data tables. If, on the other hand, the measurements for each sample are stored in a matrix, the structure of the data is then more effectively stored in a data "box." Such data is referred to as multiway data. If each sample produces a matrix of size $M \times N$ and there are L samples, then an $L \times M \times N$ three-way array is produced. There are several methods available to deal with these kinds of three-way data (Bro, 1998).

The Generalized Rank Annihilation Method (GRAM) is a simple method with various applications; many second-order analytical procedures, such as GC-MS, are bilinear, meaning that the data may be described as the outer product of concentration profiles and pure component spectra. The main issue with GRAM is that the concentration profiles in many systems alter due to drift in the analytical apparatus

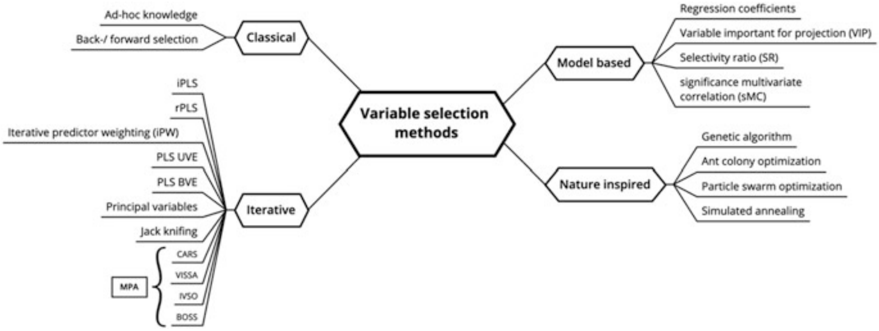

Fig. 7 Common variable selection methods in chemometrics

Table 4 Advantages and disadvantages of available variable selection methods

	Advantages	Disadvantages	References
Ad hoc	Models of causation.	Expert knowledge is required, and there are no indirect measurements (firmness, digestibility, obesity).	Mehmood et al. (2012)
Back/forward selection	Straightforward and simple.	Defining the limit is challenging.	Marcucci (1997)
*i*PLS	Simple procedure. Slow (particularly backward *i*PLS).	Autocorrelated data is required (i.e., spectra, chromatograms, process variables)	Nørgaard et al. (2000)
rPLS	Quick. Converges to a small number of variables, which is very useful for interpretation.		Rinnan et al. (2014)
VIP	Select essential variables in the PLSR model quickly.	Combines information that is associated with the parameter of interest with information that is orthogonal to it.	Wold et al. (1998)
SR	Select important variables for predicting Y quickly.	A bit too sensitive to small changes in Y pred.	Kvalheim (2010)
sMC	The sMC approach highlights variables that have little bias (in terms of parameter estimation) and are statistically significant in the model. The F-values of the variables are used to rank them.	For noisy data, it is dependable.	Tran et al. (2014)
Jack-knifing	Fast-paced. It focuses on the ambiguity in the X/Y relationship.	Data must be resampled (normally cross-validation, but Monte Carlo and bootstrap also works).	Martens and Martens (2000)
CARS	Takes into account the interactions between factors.	It is possible that critical variables will be removed forcibly. The calibration model's stability may be harmed by very collinear variables.	Li et al. (2009)
VISSA	VISSA theoretically optimizes the variable space at each phase. The variable space's statistical data is highlighted. It decreases the variable space smoothly, lowering the danger of omitting informative variables and ignoring variable combination effects. It results are indifferent to parameters like sample number		Deng et al. (2014)

(continued)

Table 4 (continued)

	Advantages	Disadvantages	References
	and model ratio, and the software stops immediately without further requirements.		
IVSO	Eliminate uninformative variables gradually, considering their interactions.		Wang et al. (2015)
BOSS	Reduces the chance of missing vital variables. Compensates for collinearity's influence.	Ignores high correlation between variables. Pick fewer options.	Al-Kaf et al. (2020)
PLS-UVE	Quickly (fast compared to backward *i*PLS, which does the same). Removes irrelevant variables (useful for large datasets).	It is not a variable selection method but rather a variable elimination method.	Centner et al. (1996)
Genetic algorithm	It looks into variable combinations that none of the "traditional" approaches looks into.	There are numerous parameters. Prone to overfitting. Best to combine with intervals or dimension reduction.	Leardi (2000)
Ant Colony optimization	Investigates variable combinations that none of the "traditional" approaches has looked into.	Prevent overfitting by combining intervals or reducing dimensions.	Shamsipur et al. (2006)
Particle swarm optimization	Investigates variable combinations which do not seek "standard" techniques	Numerous parameters. Prone to overfitting, preferable to be combined with intervals or reduced dimensions.	Lin et al. (2008)
Simulated annealing	Looks into variable combinations that none of the "normal" methods look into.	Lots of variables. Risk of overfitting, hence intervals or dimension reduction should be used in conjunction.	Meiri and Zahavi (2006)

iPLS interval partial least squares, *rPLS* recursive weighted partial least squares, *VIP* variable importance projection, *SR* selectivity ratio, *sMC* significance multivariate correlation, *CARS* competitive adaptive reweighted sampling, *VISSA* variable iterative space shrinkage approach, *IVSO* iteratively variable subset optimization, *BOSS* bootstrapping soft shrinkage, *PLS-UVE* partial least squares uninformative variable elimination

(changes in the GC column in a GC-MS, for example). These changes have the potential to rapidly damage GRAM solutions. GRAM's early implementations resulted in nonsensical fictitious solutions.

The most remarkable difference between Parallel Factor Analysis (PARAFAC) and PCA is that PARAFAC is unique in terms of scaling and permutation. Scaling ambiguity means that a column of A can be scaled by any value α as long as the corresponding column of B or C is scaled inversely i.e., by $1/\alpha$. Component one, component two, and vice versa can be called as results of permutation ambiguity. Aside from these minor uncertainties, the PARAFAC model is special in that it has only one solution. When compared to the non-uniqueness of a bilinear model, this uniqueness is the direct cause of much of PARAFAC's popularity (Bro, 1997). If the measured data fit a PARAFAC model, the model's underlying parameters can be calculated without rotational ambiguity.

The Tucker3 model, also known as the three-way PCA model, is among the most fundamental three-way models used in chemometrics (Tucker, 1966). The number three in the name Tucker3 refers to the fact that all three modes have been reduced. If the Tucker model is applied to a four-way dataset and all modes are decreased, the model will be called a Tucker4 model.

The so-called PARAFAC2 model, designed by Harshman (1972), is a more exotic yet extremely useful model. However, a workable method was not developed until 1999 (Kiers et al., 1999). A dataset may be ideally trilinear but not correspond to the PARAFAC model in some instances. It could be due to sampling issues or physical artifacts. Another issue arises when the array's slabs do not have the same row (or column) dimension. It turns out that the PARAFAC2 model can be used to solve both the problem of axis shifts and the problem of shifting axis diameters in some circumstances. (Amigo et al., 2008). One of the essential features of the PARAFAC2 model is that, like PARAFAC, it is unique in some situations. The PARAFAC2 model conditions for uniqueness have received far less attention than the PARAFAC model.

10 Tools for Chemometric Analysis

To perform chemometric analyses, several MATLAB toolboxes, R packages and software's are available. Different tools offer different functionalities. Common chemometric tools with their functionalities are listed in Table 5.

11 Conclusion

The growth in instrumentation is causing a data overload, and as a result, a large portion of the data is "wasted," meaning that no usable information is collected from it. The issue occurs with data compression as well as extraction. In general, laboratory and process measurements contain a lot of correlated or redundant data. This data must be gathered in such a way that keeps the relevant information while making it easier to show than each variable individually. Furthermore, crucial information is frequently found not in any particular element but in how the parameters vary in relation to each other; that is, the manner in which they co-vary. The information must be taken from the data in this scenario. Furthermore, in the presence of a lot of noise, it is always preferable to use some type of data processing. Therefore, a proper chemometric tool is essential for data cleaning,

Table 5 Available tools for Chemometrics

Software package	Available methods	Website
MATLAB environment		
PLS_toolbox	PCA, PLS, PLS-DA, SIMCA, SVM, HCA, PARAFAC, MLR, CLS, ANN, MCR DoE, preprocessing	https://eigenvector.com/software/pls-toolbox/
iToolbox	Interval PLS (*i*PLS), backward interval PLS (*bi*PLS), moving window PLS (*mw*PLS), synergy interval PLS (*si*PLS), and interval PCA (*i*PCA)	http://www.models.life.ku.dk/itoolbox
GAPLS toolbox	Genetic algorithm PLS	http://www.models.kvl.dk/GAPLS
SPA toolbox	Successive projections algorithm	http://www.ele.ita.br/~kawakami/spa/
LS-SVM lab	Kernel PCA, kernel CCA, and kernel PLS	https://www.esat.kuleuven.be/sista/lssvmlab/
Hypertools	Multispectral and hyperspectral image analysis	https://www.hypertools.org
A good number of essential source code/toolbox for chemometric analysis are freely available at http://www.models.life.ku.dk/algorithms		
R environment		
mdatools	Preprocessing, exploring data	https://mdatools.com
Chemometrics	PCA, PLSR	https://rdrr.io/cran/chemometrics/
ChemoSpec	Chemometrics for spectroscopy	https://cran.r-project.org/web/packages/ChemoSpec/index.html
prospectr	Preprocessing of data	https://cran.r-project.org/web/packages/prospectr/index.html
The Unscrambler	Mostly cover all chemometrics methods	https://www.aspentech.com/en/products/msc/aspen-unscrambler
SIMCA	Multivariate tools, data visualizations, and process intelligence	https://www.sartorius.com/en/products/process-analytical-technology/data-analytics-software/mvda-software/simca
Latentix	Preprocessing	https://www.latentix.com
Pirouette	Multivariate calibration and prediction	https://infometrix.com/pirouette/
PerClass	Spectral image analysis	https://www.perclass.com
OriginLab	PCA, DoE, logistic regression	https://www.originlab.com
DesignExpert	Experimental design	https://www.statease.com/software/design-expert/

preprocessing, and extracting the most relevant chemical information from the experimental data.

Acknowledgement The author acknowledges the support from the research management wing, Bangabandhu Sheikh Mujibur Rahman Agricultural University, Gazipur-1706, Bangladesh. The author declares no conflict of interest.

References

Al-Kaf, H. A. G., Alduais, N. A. M., Saad, A. M. H. Y., Chia, K. S., Mohsen, A. M., Alhussian, H., Mahdi, A. A. M. H., & Salam, W. S. I. W. (2020). A bootstrapping soft shrinkage approach and interval random variables selection hybrid model for variable selection in near-infrared spectroscopy. *IEEE Access, 8*, 168036–168052. https://doi.org/10.1109/ACCESS.2020.3023681

Amigo, J. M. (2021). Data mining, machine learning, deep learning, chemometrics definitions, common points and trends (spoiler alert: VALIDATE your models!). *Brazilian Journal of Analytical Chemistry, 8*, 22–38. https://doi.org/10.30744/brjac.2179-3425.ar-38-2021

Amigo, J. M., Skov, T., Bro, R., Coello, J., & Maspoch, S. (2008). Solving GC-MS problems with PARAFAC2. *TrAC Trends in Analytical Chemistry, 27*, 714–725. https://doi.org/10.1016/j.trac.2008.05.011

Asghar, A., Abdul Raman, A. A., & Daud, W. M. A. W. (2014). A comparison of central composite design and Taguchi method for optimizing Fenton process. *The Scientific World Journal, 2014*, 869120. https://doi.org/10.1155/2014/869120

Ballabio, D., & Consonni, V. (2013). Classification tools in chemistry. Part 1: Linear models. PLS-DA. *Analytical Methods, 5*, 3790–3798. https://doi.org/10.1039/c3ay40582f

Barker, M., & Rayens, W. (2003). Partial least squares for discrimination. *Journal of Chemometrics, 17*, 166–173. https://doi.org/10.1002/cem.785

Barnes, R., Dhanoa, M. S., & Lister, S. J. (1989). Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra. *Applied Spectroscopy, 43*, 772–777. https://doi.org/10.1366/0003702894202201

Bartlett, M. S., & Kendall, D. (1946). The statistical analysis of variance-heterogeneity and the logarithmic transformation. *Supplement to the Journal of the Royal Statistical Society, 8*, 128–138. https://doi.org/10.2307/2983618

Bezerra, M. A., Santelli, R. E., Oliveira, E. P., Villar, L. S., & Escaleira, L. A. (2008). Response surface methodology (RSM) as a tool for optimization in analytical chemistry. *Talanta, 76*, 965–977. https://doi.org/10.1016/j.talanta.2008.05.019

Bro, R. (1997). PARAFAC. Tutorial and applications. *Chemometrics and Intelligent Laboratory Systems, 38*, 149–171. https://doi.org/10.1016/s0169-7439(97)00032-4

Bro, R. 1998. *Multi-way analysis in the food industry-models, algorithms, and applications* (Doctoral dissertation). Royal Veterinary and Agricultural University.

Bro, R., & Smilde, A. K. (2014). Principal component analysis. *Analytical Methods, 6*, 2812–2831. https://doi.org/10.1039/c3ay41907j

Centner, V., Massart, D.-L., de Noord, O. E., de Jong, S., Vandeginste, B. M., & Sterna, C. (1996). Elimination of uninformative variables for multivariate calibration. *Analytical Chemistry, 68*, 3851–3858. https://doi.org/10.1021/ac960321m

Cocchi, M. (2017). Chemometrics for food quality control and authentication. In *Encyclopedia of analytical chemistry* (pp. 1–29). Wiley. https://doi.org/10.1002/9780470027318.a9579.

De Maesschalck, R., Jouan-Rimbaud, D., & Massart, D. L. (2000). The Mahalanobis distance. *Chemometrics and Intelligent Laboratory Systems, 50*, 1–18. https://doi.org/10.1016/s0169-7439(99)00047-7

Deng, B.-C., Yun, Y.-H., Liang, Y.-Z., & Yi, L.-Z. (2014). A novel variable selection approach that iteratively optimizes variable space using weighted binary matrix sampling. *The Analyst, 139*, 4836. https://doi.org/10.1039/c4an00730a

Dieterle, F., Ross, A., Schlotterbeck, G., & Senn, H. (2006). Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. Application in 1H NMR Metabonomics. *Analytical Chemistry, 78*, 4281–4290. https://doi.org/10.1021/ac051632c

Ebrahimi-Najafabadi, H., Leardi, R., & Jalali-Heravi, M. (2014). Experimental design in analytical chemistry—Part I: Theory. *Journal of AOAC International, 97*, 3–11. https://doi.org/10.5740/jaoacint.sgeebrahimi1

Enders, C. K., & Tofighi, D. (2007). Centering predictor variables in cross-sectional multilevel models: A new look at an old issue. *Psychological Methods, 12*, 121. https://doi.org/10.1037/1082-989x.12.2.121

Eriksson, L., Byrne, T., Johansson, E., Trygg, J., & Vikström, C. (2013). *Multi-and megavariate data analysis basic principles and applications*. Umetrics Academy. https://doi.org/10.1002/cem.713

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics, 7*, 179–188. https://doi.org/10.1111/j.1469-1809.1936.tb02137.x

Geladi, P., & Kowalski, B. R. (1986). Partial least-squares regression: A tutorial. *Analytica Chimica Acta, 185*, 1–17. https://doi.org/10.1016/0003-2670(86)80028-9

Granato, D., & de Araújo Calado, V. M. (2013). The use and importance of design of experiments (DOE) in process modelling in food science and technology. In *Mathematical and statistical methods in food science and technology* (pp. 1–18). Wiley. https://doi.org/10.1002/9781118434635.ch01.

Guidetti, R., Beghi, R., & Giovenz, V. (2012). Chemometrics in food technology. In *Chemometrics in Practical Applications*. InTech. https://doi.org/10.5772/34148.

Guo, Q., Wu, W., & Massart, D. L. (1999). The robust normal variate transform for pattern recognition with near-infrared data. *Analytica Chimica Acta, 382*, 87–103. https://doi.org/10.1016/S0003-2670(98)00737-5

Harshman, R. (1972). PARAFAC2: Extensions of a procedure for “explanatory” factor-analysis and multidimensional scaling. *The Journal of the Acoustical Society of America, 51*, 111–111. https://doi.org/10.1121/1.1981298

Hotelling, H. (1947). Multivariate quality control-illustrated by the air testing of sample bombsights. In C. Eisenhart, M. Hastay, & W. Wallis (Eds.), *Techniques of statistical analysis* (pp. 111–184). McGraw-Hill.

Hotelling, H. (1992). The generalization of Student’s ratio. In S. Kotz & N. L. Johnson (Eds.), *Breakthroughs in statistics* (pp. 54–65). Springer. https://doi.org/10.1007/978-1-4612-0919-5_4

Huynh, H., & Feldt, L. S. (1976). Estimation of the box correction for degrees of freedom from sample data in randomized block and split-plot designs. *Journal of Educational Statistics, 1*, 69–82. https://doi.org/10.3102/10769986001001069

Isaksson, T., & Næs, T. (1988). The effect of multiplicative scatter correction (MSC) and linearity improvement in NIR spectroscopy. *Applied Spectroscopy, 42*, 1273–1284. https://doi.org/10.1366/0003702884429869

Islam, M. N., Nielsen, G., Stærke, S., Kjær, A., Jørgensen, B., & Edelenbos, M. (2018a). Noninvasive determination of firmness and dry matter content of stored onion bulbs using shortwave infrared imaging with whole spectra and selected wavelengths. *Applied Spectroscopy, 72*, 1467–1478. https://doi.org/10.1177/0003702818792282

Islam, M. N., Nielsen, G., Stærke, S., Kjær, A., Jørgensen, B., & Edelenbos, M. (2018b). Novel non-destructive quality assessment techniques of onion bulbs: A comparative study. *Journal of Food Science and Technology, 55*, 3314–3324. https://doi.org/10.1007/s13197-018-3268-x

Kasprzak, E. M., & Lewis, K. E. (2001). Pareto analysis in multiobjective optimization using the collinearity theorem and scaling method. *Structural and Multidisciplinary Optimization, 22*, 208–218. https://doi.org/10.1007/s001580100138

Kiers, H. A., Ten Berge, J. M., & Bro, R. (1999). PARAFAC2—Part I. A direct fitting algorithm for the PARAFAC2 model. *Journal of Chemometrics: A Journal of the Chemometrics Society, 13*, 275–294. https://doi.org/10.1002/(sici)1099-128x(199905/08)13:3/4<275::aid-cem543>3.0.co;2-b

Kos, G., Lohninger, H., & Krska, R. (2003). Validation of chemometric models for the determination of deoxynivalenol on maize by mid-infrared spectroscopy. *Mycotoxin Research, 19*, 149–153. https://doi.org/10.1007/bf02942955

Kvalheim, O. M. (2010). Interpretation of partial least squares regression models by means of target projection and selectivity ratio plots. *Journal of Chemometrics, 24*, 496–504. https://doi.org/10.1002/cem.1289

Lawson, J. (2014). *Design and analysis of experiments with R*. Taylor & Francis. https://doi.org/10.1201/b17883

Leardi, R. (2000). Application of genetic algorithm–PLS for feature selection in spectral data sets. *Journal of Chemometrics, 14*, 643–655. https://doi.org/10.1002/1099-128X(200009/12)14:5/6<643::AID-CEM621>3.0.CO;2-E

Leardi, R. (2006). *D-optimal designs Encyclopedia of analytical chemistry: Applications, theory and instrumentation* (pp. 1–11). Wiley. https://doi.org/10.1002/9780470027318.a9646

Leardi, R. (2009). Experimental design in chemistry: A tutorial. *Analytica Chimica Acta, 652*, 161–172. https://doi.org/10.1016/j.aca.2009.06.015

Li, H., Liang, Y., Xu, Q., & Cao, D. (2009). Key wavelengths screening using competitive adaptive reweighted sampling method for multivariate calibration. *Analytica Chimica Acta, 648*, 77–84. https://doi.org/10.1016/j.aca.2009.06.046

Lin, S.-W., Ying, K.-C., Chen, S.-C., & Lee, Z.-J. (2008). Particle swarm optimization for parameter determination and feature selection of support vector machines. *Expert Systems with Applications, 35*, 1817–1824. https://doi.org/10.1016/j.eswa.2007.08.088

Marcucci, M. (1997). *Applied multivariate techniques*. Taylor & Francis. https://doi.org/10.2307/1270777

Marini, F. (2013). *Chemometrics in food chemistry*. Newnes. https://doi.org/10.1016/c2011-0-08492-2

Martens, H., & Martens, M. (2000). Modified Jack-knife estimation of parameter uncertainty in bilinear modelling by partial least squares regression (PLSR). *Food Quality and Preference, 11*, 5–16. https://doi.org/10.1016/s0950-3293(99)00039-7

Martens, H., & Martens, M. (2001). *Multivariate analysis of quality: An introduction*. Wiley. https://doi.org/10.1088/0957-0233/12/10/708

Martens, H., & Naes, T. (1991). *Multivariate calibration*. Wiley. https://doi.org/10.2307/2532682

Martens, H., Nielsen, J. P., & Engelsen, S. B. (2003). Light scattering and light absorbance separated by extended multiplicative signal correction. Application to near-infrared transmission analysis of powder mixtures. *Analytical Chemistry, 75*, 394–404. https://doi.org/10.1021/ac020194w

Mehmood, T., Liland, K. H., Snipen, L., & Sæbø, S. (2012). A review of variable selection methods in partial least squares regression. *Chemometrics and Intelligent Laboratory Systems, 118*, 62–69. https://doi.org/10.1016/j.chemolab.2012.07.010

Meiri, R., & Zahavi, J. (2006). Using simulated annealing to optimize the feature selection problem in marketing applications. *European Journal of Operational Research, 171*, 842–858. https://doi.org/10.1016/j.ejor.2004.09.010

Mishra, P., Rutledge, D. N., Roger, J.-M., Wali, K., & Khan, H. A. (2021). Chemometric pre-processing can negatively affect the performance of near-infrared spectroscopy models for fruit quality prediction. *Talanta, 229*, 122303. https://doi.org/10.1016/j.talanta.2021.122303

Næs, T., Isaksson, T., Fearn, T., & Davies, T. (2002). *A user-friendly guide to multivariate calibration and classification*. NIR. https://doi.org/10.1255/978-1-906715-25-0

Noda, I. (2008). Scaling techniques to enhance two-dimensional correlation spectra. *Journal of Molecular Structure, 883*, 216–227. https://doi.org/10.1016/j.molstruc.2007.12.026

Nørgaard, L., Saudland, A., Wagner, J., Nielsen, J. P., Munck, L., & Engelsen, S. B. (2000). Interval partial least-squares regression (iPLS): A comparative chemometric study with an example from near-infrared spectroscopy. *Applied Spectroscopy, 54*, 413–419. https://doi.org/10.1366/0003702001949500

Press, W. H., & Teukolsky, S. A. (1990). Savitzky-Golay smoothing filters. *Computers in Physics, 4*, 669–672. https://doi.org/10.1063/1.4822961

Rabatel, G., Marini, F., Walczak, B., & Roger, J.-M. (2020). VSN: Variable sorting for normalization. *Journal of Chemometrics, 34*, e3164. https://doi.org/10.1002/cem.3164

Rakić, T., Kasagić-Vujanović, I., Jovanović, M., Jančić-Stojanović, B., & Ivanović, D. (2014). Comparison of full factorial design, central composite design, and Box-Behnken design in chromatographic method development for the determination of fluconazole and its impurities. *Analytical Letters, 47*, 1334–1347. https://doi.org/10.1080/00032719.2013.867503

Rinnan, Å. (2014). Pre-processing in vibrational spectroscopy—When, why and how. *Analytical Methods, 6*, 7124–7129. https://doi.org/10.1039/c3ay42270d

Rinnan, Å., Andersson, M., Ridder, C., & Engelsen, S. B. (2014). Recursive weighted partial least squares (rPLS): An efficient variable selection method using PLS. *Journal of Chemometrics, 28*, 439–447. https://doi.org/10.1002/cem.2582

Rinnan, Å., Nørgaard, L., Berg, F. V. D., Thygesen, J., Bro, R., & Engelsen, S. B. (2009). Data pre-processing. In D.-W. Sun (Ed.), *Infrared spectroscopy for food quality analysis and control* (pp. 29–50). Academic Press. https://doi.org/10.1016/B978-0-12-374136-3.00002-X

Roger, J.-M., Boulet, J.-C., Zeaiter, M., & Rutledge, D. N. (2020). Pre-processing methods. In S. Brown, R. Tauler, & B. Walczak (Eds.), *Comprehensive chemometrics* (2nd ed., pp. 1–75). Elsevier. https://doi.org/10.1016/b978-0-12-409547-2.14878-4

Rousseeuw, P. J., & Leroy, A. M. (1987). *Robust regression and outlier detection*. Wiley. https://doi.org/10.1002/0471725382

Shamsipur, M., Zare-Shahabadi, V., Hemmateenejad, B., & Akhond, M. (2006). Ant colony optimisation: A powerful tool for wavelength selection. *Journal of Chemometrics: A Journal of the Chemometrics Society, 20*, 146–157. https://doi.org/10.1002/cem.1002

Smilde, A., Bro, R., & Geladi, P. (2005). Two-way component and regression models. In *Multi-way analysis with applications in the chemical sciences* (pp. 35–45). Wiley. https://doi.org/10.1002/0470012110.ch3.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society: Series B: Methodological, 36*, 111–133. https://doi.org/10.1111/j.2517-6161.1974.tb00994.x

Tanabe, J., Miller, D., Tregellas, J., Freedman, R., & Meyer, F. G. (2002). Comparison of detrending methods for optimal fMRI preprocessing. *NeuroImage, 15*, 902–907. https://doi.org/10.1006/nimg.2002.1053

Tran, T. N., Afanador, N. L., Buydens, L. M. C., & Blanchet, L. (2014). Interpretation of variable importance in partial least squares with significance multivariate correlation (sMC). *Chemometrics and Intelligent Laboratory Systems, 138*, 153–160. https://doi.org/10.1016/j.chemolab.2014.08.005

Tucker, L. R. (1966). Some mathematical notes on three-mode factor analysis. *Psychometrika, 31*, 279–311. https://doi.org/10.1007/BF02289464

van den Berg, R. A., Hoefsloot, H. C., Westerhuis, J. A., Smilde, A. K., & van der Werf, M. J. (2006). Centering, scaling, and transformations: Improving the biological information content of metabolomics data. *BMC Genomics, 7*, 1–15. https://doi.org/10.1186/1471-2164-7-142

Vanaja, K., & Shobha Rani, R. (2007). Design of experiments: Concept and applications of Plackett Burman design. *Clinical Research and Regulatory Affairs, 24*, 1–23. https://doi.org/10.1080/10601330701220520

Vidal, M., & Amigo, J. M. (2012). Pre-processing of hyperspectral images. Essential steps before image analysis. *Chemometrics and Intelligent Laboratory Systems, 117*, 138–148. https://doi.org/10.1016/j.chemolab.2012.05.009

Wang, W., Yun, Y., Deng, B., Fan, W., & Liang, Y. (2015). Iteratively variable subset optimization for multivariate calibration. *RSC Advances, 5*, 95771–95780. https://doi.org/10.1039/c5ra08455e

Westad, F., & Kermit, M. (2003). Cross validation and uncertainty estimates in independent component analysis. *Analytica Chimica Acta, 490*, 341–354. https://doi.org/10.1016/s0003-2670(03)00090-4

Westad, F., & Marini, F. (2015). Validation of chemometric models—A tutorial. *Analytica Chimica Acta, 893*, 14–24. https://doi.org/10.1016/j.aca.2015.06.056

Westerhuis, J. A., Gurden, S. P., & Smilde, A. K. (2000). Generalized contribution plots in multivariate statistical process monitoring. *Chemometrics and Intelligent Laboratory Systems, 51*, 95–114. https://doi.org/10.1016/s0169-7439(00)00062-9

Wold, S. (1976). Pattern recognition by means of disjoint principal components models. *Pattern Recognition, 8*, 127–139. https://doi.org/10.1016/0031-3203(76)90014-5

Wold, S., Esbensen, K., & Geladi, P. (1987). Principal component analysis. *Chemometrics and Intelligent Laboratory Systems, 2*, 37–52. https://doi.org/10.1016/0169-7439(87)80084-9

Wold, S., Josefson, M., Gottfries, J., & Linusson, A. (2004). The utility of multivariate design in PLS modeling. *Journal of Chemometrics, 18*, 156–165. https://doi.org/10.1002/cem.861

Wold, S., Martens, H., & Wold, H. (1983). The multivariate calibration problem in chemistry solved by the PLS method. In B. Kågström & A. Ruhe (Eds.), *Matrix pencils* (pp. 286–293). Springer. https://doi.org/10.1007/bfb0062108

Wold, S., Sjöström, M., & Eriksson, L. (1998). Partial least squares projections to latent structures (PLS) in chemistry. In P. von Ragué Schleyer (Ed.), *Encyclopedia of computational chemistry* (pp. 2006–2021). Wiley. https://doi.org/10.1002/0470845015.cpa012

Wold, S., Sjöström, M., & Eriksson, L. (2001). PLS-regression: A basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems, 58*, 109–130. https://doi.org/10.1016/s0169-7439(01)00155-1

Zou, X., & Zhao, J. (2015). NIR spectroscopy detection. In Z. Xiaobo & J. Zhao (Eds.), *Nondestructive measurement in food and agro-products* (pp. 57–126). Springer. https://doi.org/10.1007/978-94-017-9676-7_3