Egyptian Journal of
Forensic Sciences

REVIEW **Open Access**

Overview of chemometrics in forensic toxicology

Sukhwinder Singh[1*], Hanan Shakeel[1] and Rakesh Sharma[1]

Abstract

Background The beginning of chemometrics within pattern recognition of the 1960s and 1970s is defined. This article shows a comprehensive deliberation on application of the chemometric techniques to chemical data analysis.

Main body of the abstract Many review papers along with the usage of chemometrics in forensic chemistry have been available. The present article has been distributed into several parts which comprise chemometrics, its history, its function and chemometrics methods.

Conclusion It is advised that these new chemometrics methods should be applied in forensic chemistry to get accurate and fast results.

Keywords Chemometric, Multivariate analysis, Forensic chemistry

Background

Chemometrics is the science that extracts information from mathematical and statistical systems. It is an inherently multidisciplinary field whose relevance among the chemical disciplines in general and analytical chemistry has considerably grown over the years. This is shown in Fig. 1. These techniques are used in modern chromatographic and spectrometric techniques, which furnish digitalized data. Chemometrics approaches can be used for analysis protocols, process modelling, multivariate data collection, calibration, predictions, classification and pattern recognition, compression and graphical display and statistical process control. In particular, chemometric applications explain low-dimensional and high-dimensional data and are also useful in the forensic discipline. In forensic casework, it is important to classify and identify samples correctly. For this reason, the forensic field shows a clear trend towards increasing the use of chemometrics (Wold 2017). Chemometrics can provide information and enhance productivity in different fields.

Main text

Chemometric field was recognized in the 1970s when personal computers became gradually exploit for scientific purposes. The word 'chemometrics' was first used by Svante Wold in 1971. In 1974, Svante Wold created the International Chemometrics Society (ICS) together with Bruce Kowalski. He communicated his paper that is first along with the word chemometrics. This happened first time in 1980s in chemometrics field (Bovens et al. 2019).

1. First devoted journals *Chemometrics* journal and *Chemometrics and Intelligent Laboratory Systems*
2. First book titled Chemometrics
3. First devoted software UNSCRAMBLER, SIMCA and ARTHUR

The aim of chemometrics is to provide statistical and mathematical tools to convert raw data into information. The systematic diagram has been shown in Fig. 2.

The role of chemometrics is to intensify the use of a basic knowledge of statistics not simply as a tool for

*Correspondence:
Sukhwinder Singh
sukhwinder3956@gmail.com
[1] Saraswati Group of Colleges, Mohali, Punjab, India 140413

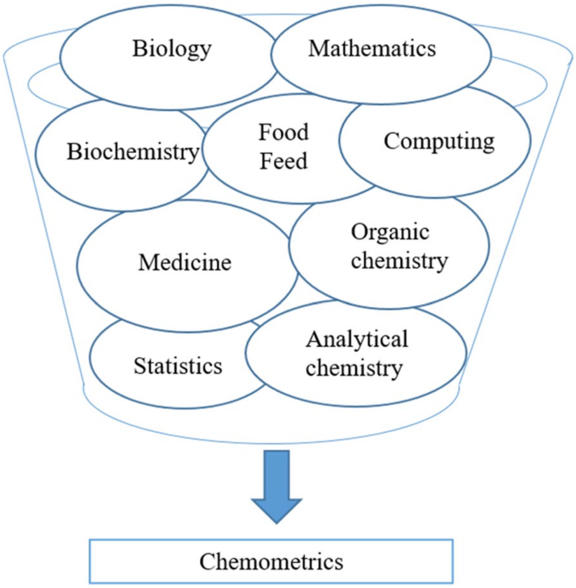

Fig. 1 Different areas of life science accompanying the application of chemometric method

interpreting data but rather as a whole way of approaching experimental procedures.

Role of chemometrics

There are three types of roles for chemometrics (Brereton 2015):

I.Analysis

Earlier, analysis was the most devoted part of chemometrics. Further, two more applications of chemometrics have been identified — factor analysis and calibration. Factor analysis is derived from other metric sciences like econometrics and psychometrics. The aim of factor analysis is to take an overview of the chemical process and detect the factors that are involved in collection of data. Calibration is used to give more attention to the chemical process. The aim of this application is to extract particular information from collection of data.

II. Facilitation

Lately, the role of chemometrics has been to accelerate the placement of instrumentation in every field. To develop a calibration model for each instrument, there is much expense and labour. So, chemometrics also develop standardized compact models. Therefore, the calibration model is adaptable among instruments.

III. Design

The most utilized role is design in process analysis. Chemometrics deeply provides insights on what information is necessary or unnecessary from an analysis.

The guiding principles and software tool will assist forensic workflow and enhance the productivity of samples. Based on the literature, chemometric methods are still used in analytical chemistry as these methods give notable and expedient results (Colina 1988). The guiding principles and software will not be limited to examine low-dimensional data (drug examination) but also applicable for other forensic disciplines.

Furthermore, the results of these analyses are used for identification and need to be communicated in an extensive manner in the forensic field. Apart from this use in chemical or physical casework, chemometric can be used to analyse large sets of data for intelligence tasks as well as crime analysis purposes (Frank and Friedman 2013).

In this review, we have been given a brief introduction of chemometric methods, difference between supervised and unsupervised pattern recognition, and also evaluate

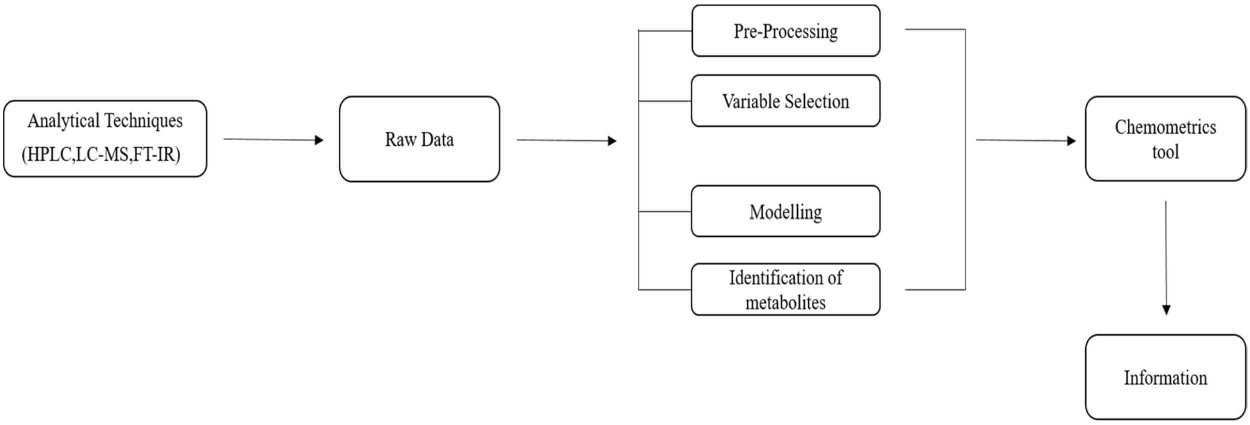

Fig. 2 Structural outline of converting data into information

the problems from forensic chemistry that can be solved by chemometrics tool.

Chemometrics algorithm

Modern computer technologies and spectroscopic analysis when combined with chemometrics algorithms provide methods of chemical analysis that are fast, inexpensive and simple to use (Booksh 2006). Flow chart diagram has been shown in Fig. 3.

Chemometrics algorithm are divided into two parts:

Regression algorithm

Regression algorithm on observational data creates a major part of chemometric studies (Marta 2014). This can be further divided into two parts:

(a) Linear regression algorithm

In a linear regression algorithm, the regression function is linear (Fig. 4).

$$Y = \mathrm{a} + \mathrm{bx} + \mathrm{u}$$

where x = independent variable, a = the intercept, Y = dependent variable, b = the slope and u = error term.

This is again classified into the following categories:

(b) Multiple linear regression (MLR)

This algorithm is the basic form of linear regression method and was established by Karl Norris in the 1970s. It is used for quantitative analysis by using modern methods (Singh et al. 2013). It is used to explain the relationship between two or more independent variables and one continuous dependent variable in a predictive analysis, and these independent variables can be continuous or categorical (Fig. 5).

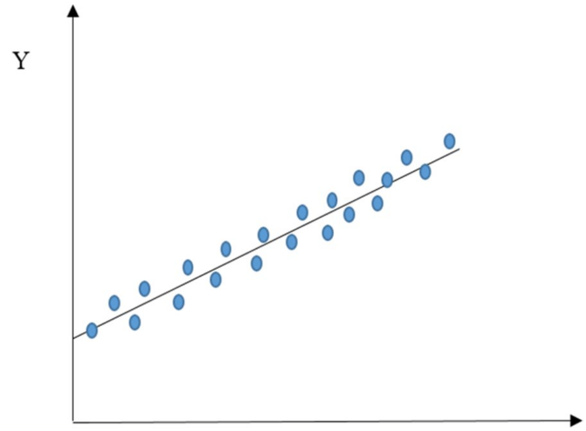

Fig. 4 Graphical representation of linear regression algorithm

$$Y = \mathrm{a} + \mathrm{b_1x_1} + \mathrm{b_2x_2} + \mathrm{b_3x_3} \ldots\ldots + \mathrm{b_tx_t} + \mathrm{u}$$

where x = independent variable, a = the intercept, b = the slope (coefficient of x_1), Y = dependent variable and u = error term.

(c) Principal component regression (PCR)

It is a combination of PCA and MLR. Primarily, the principal components are calculated, and the scores obtained

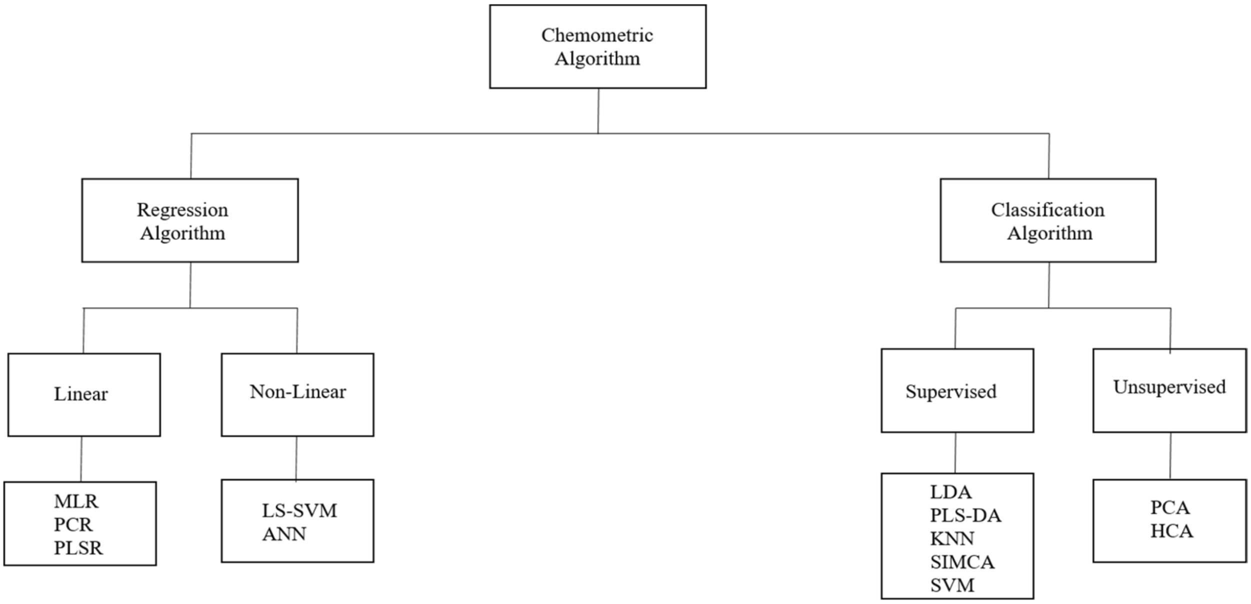

Fig. 3 Classification of chemometric methods

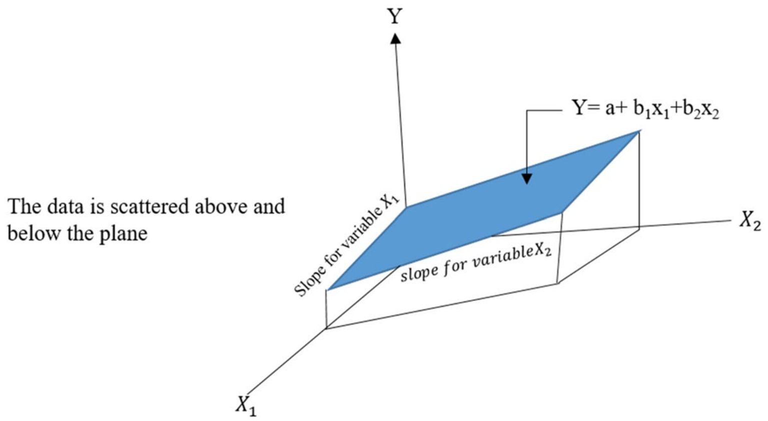

Fig. 5 Diagrammatic representation of multiple linear regression

are used as the basis for the MLR with the target data y. The graphical abstract has been shown in Fig. 6. Multiple regression data are obtained from multicollinearity when analysed by the PCR technique. When multicollinearity takes place, least squares evaluated are impartial; on the other hand, they can be far from the true value because their variances are huge. When adding a degree of bias to the regression estimates, the regression of the principal component reduces the standard errors (Siebert 2011).

(d) Partial least square regression (PLSR)

PLSR is a multivariate method which is used to create a regression model between various independent and dependent variables when variables have multicollinearity. Briefly, PLS evaluates the principal components. It was established for social sciences but became famous in chemometrics. PLSR is using NIPALS algorithm.

The linear regression model is defined as (Meglen 1988).

$$Y = \mathrm{a} + \mathrm{bx} + \mathrm{u}$$

where b = regression coefficients that are evaluated during calibration, u = error term and b is calculated by using the equation.

$$b = \mathrm{W}(\mathrm{P}^{\mathrm{T}}\mathrm{W})^{-1} - 1\mathrm{Q}^{\mathrm{T}}\mathrm{QT}$$

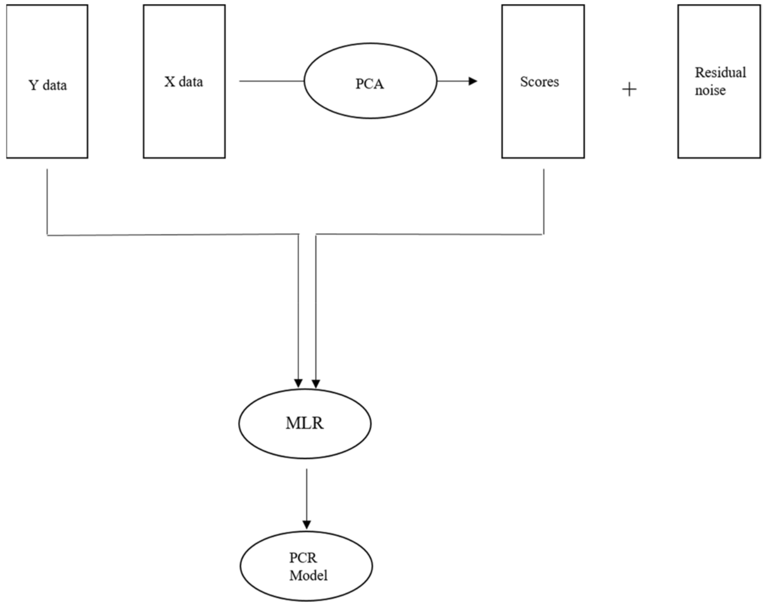

Fig. 6 Graphical representation of principal component regression

where P = two sets of x loading that are calculated in PLS calibration and explain the relation between the x data matrix and its scores W = loading weights, T = W relate the matrices Y and x through regression and Q = regression coefficients connect the Y-variables to T scores and also used to estimate b.

PLSR contains of outer relations and an inner relation connecting both blocks: Fig. 7.

Outer relation for the x-block is as follows:

$$x = \Sigma \mathrm{TP}^{\mathrm{T}} + \mathrm{E} = \Sigma \mathrm{t_h P_h}^{\mathrm{T}} + \mathrm{E}$$

where t_{h} = scores vector, $P_{\mathrm{h}}{}^{\mathrm{T}}$ = loadings vector for the x-block, T = loadings time and E = residuals.

Outer relation for the Y-block is as follows:

$$Y = \Sigma \mathrm{UQ}^{\mathrm{T}} + \mathrm{F} = \Sigma \mathrm{U_h q_h}^{\mathrm{T}} + \mathrm{F}$$

where U_{h} = scores for the Y block, $\mathrm{q_h}^{\mathrm{T}}$ = loadings for the Y-block and U = score vectors derived from the Y-matrixes are the starting points for the t-score vectors in decompositions of the X-matrix.

Plotting scores of Y-block (U) against x-block scores (T) calculate inner relation for each component.

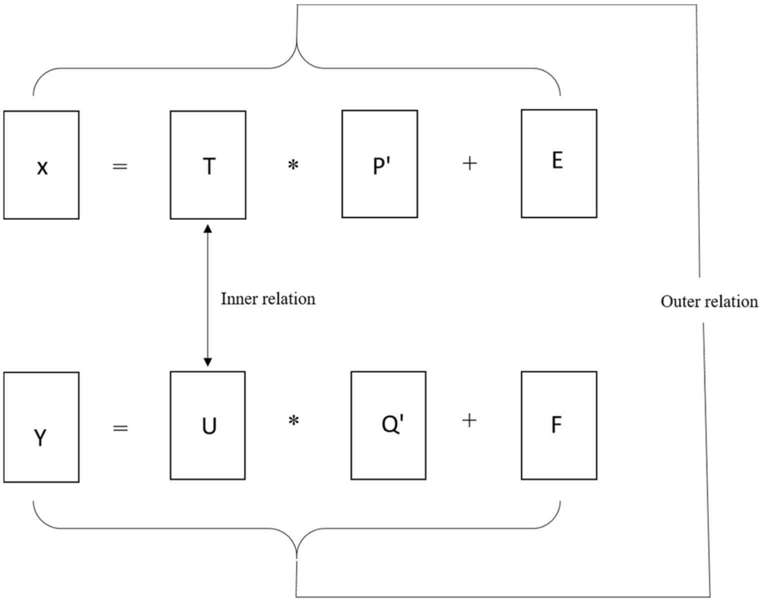

Fig. 7 Graphical representation of partial least square regression

Non-linear regression algorithm

In statistics, non-linear regression algorithm of the regression function is not linear (Fig. 8), whereas in linear regression algorithm it is as follows (Brereton et al. 2018):

$$Y = \mathrm{f}(\mathrm{X}, \beta) + \varepsilon$$

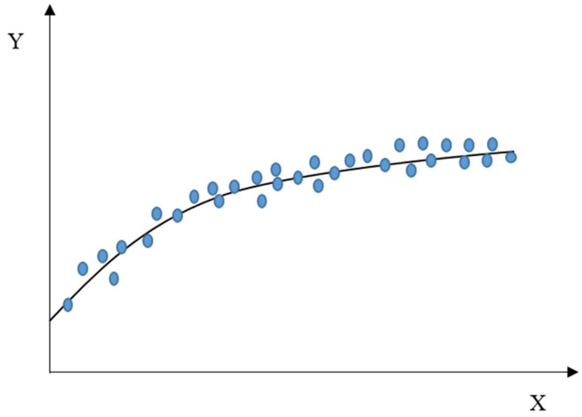

Fig. 8 Graphical representation of non-linear regression algorithm

where Y = dependent variable, X = independent variable, β = vector of parameter, ε = error term and f = regression function.

This can also be further divided into a few parts.

(a) Least squares support vector machine (LS-SVM)

This technique is the least square form of support vector machine. This method was proposed by Suykens. It is based on a kernel learning method. In this method, the objective function is the same as for SVM, but the loss function is changed by the classical squared loss function. Lack of sparseness is the advantage of this method. SVM implementation needs three parameters (σ, γ, ξ), while LS-SVM implementation needs only two parameters (σ, γ) (Chauchard et al. 2010).

(b) Artificial neural network (ANN)

This technique is a non-linear tool and appropriate for practical application due to adaptability and flexibility. In 1943, McCulloch communicated the first modelling and pits in the matter of computational model 'nervous activity'. This model explains about neurons as a computing unit with several inputs and single output either 0 or 1. This network has a unit called perceptron, which gives an output as 1 or −1 based on linear combinations of inputs. As a computational outlook, the simplest presentation of artificial neural network is in Fig. 9.

$$y = \mathrm{f}(\sum \mathrm{wixi} + \mathrm{w0})$$

where y = output, x_i = non-linear combinations of inputs, w_i = weights, w_0 = offset term called bias and f = function

ANNs can perform mapping, regression, modelling, classification and clustering, so they have a wide field of application. They are ideal for solving non-linear problems, where conventional statistical methods do not work. They deal with introduction into chemometric studies in different networks and algorithms (Marini et al. 2008). Neural networks are showing different types of techniques which are used in chemistry (Fig. 10).

Many studies have been conducted in analytical chemistry using ANN chemometric techniques which aim at obtaining multivariate calibration and analysis of spectroscopic data.

Classification algorithm

Many methods for classifying and identifying spectral data have been employed. According to the problem, they have been divided into two categories.

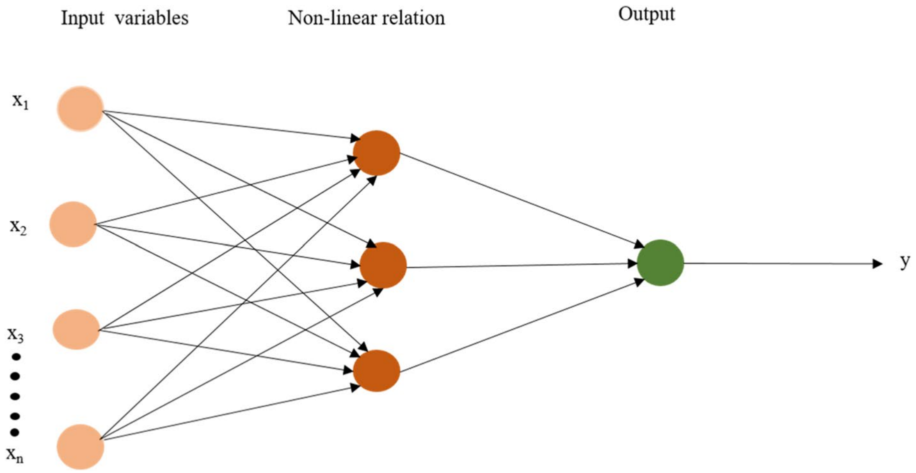

Fig. 9 Graphical representation of artificial neural network

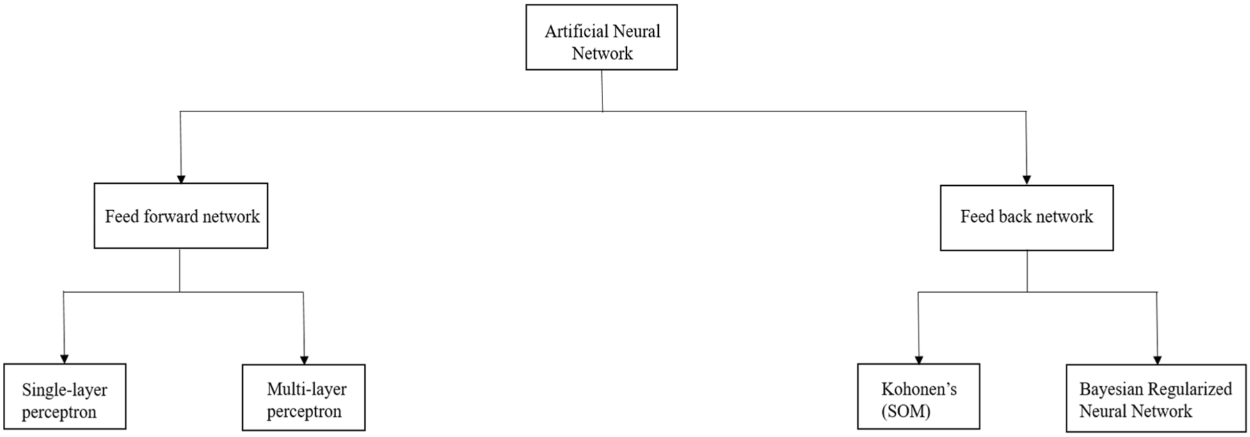

Fig. 10 Classification of artificial neural networks methods

Supervised pattern recognition

Supervised classification algorithms are enhanced with pairs of observations for which the class membership is already studied. Supervised algorithm is used on a large scale with different applications like individualization, classification and discrimination (Kumar and Sharma 2018). Supervised pattern recognition is divided into categories:

1. Discrimination between the classes
2. Modelling the individual classes

There are several methods for this.

(a) Linear discriminant analysis (LDA)

This relates to a statistical method and proposed by Ronald Fisher in 1936. He classified an observation as one of two possible groups based on many measurements. These methods are developed to predict the ratio of the variance between classes to the variance within the classes (Fig. 11).

$$J(W) = \frac{W^T S_B W}{W^T S_W W}$$

where W = transformation matrix of LDA, S_B = between class variance and S_W = within class variance

Application of these methods is used in biometrics, bioinformatics and chemistry. There are two types of classes in which LDA is involved (Tharwat et al. 2017).

- Class-dependent linear discriminant analysis — To project the data of each class on, it is calculated a separate class of lower dimension.

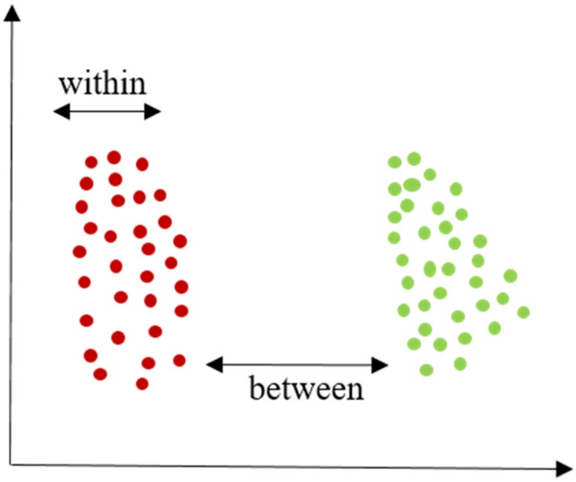

Fig. 11 Graphical abstract of linear discriminant analysis

- Class independent linear discriminant analysis — Each class against the other class will be treated as an individual class.

The aim of this method is to estimate the original data matrix in a lower dimensional disk. There are three steps to be completed for achieving this aim:

- Calculate the distinctiveness of class variance.
- Measure the variance within class.
- Create the lower dimensional space that maximizes the distance between the different classes of media and minimizes the distance between the mean and the sample of each class.

(b) Partial least square discriminant analysis (PLS-DA)

Primarily, a partial least square algorithm was used for a regression task and then used for categorization that is notable as PLS-DA. It has been used for descriptive, predictive modelling and also used in discriminative variable selection. It has many applications in different fields: food science, medical science, forensic science and soil science. PLS-DA modelling has eight significant functional aspects (Fig. 12).

Experiential work is essential for upgrading the PLS-DA modelling practice plans especially in multifaceted datasets, i.e. high-dimensional, multiclass and imbalanced which approach future real-world issues closely (Lee 2018).

(c) k-nearest neighbour (kNN)

The KNN permits a specimen or group to be used to detect other specimens or groups. A new instance (object) is classified by a widely held vote for its neighbour classes. The new instance (object) is allocated to its nearest neighbour to the most common class (Fig. 13). This method is better than LDA. These algorithms are used for evaluating continuous variables. It works as follows (Hall 2008):

i. Euclidean

$$D = \sqrt{\sum_{i=1}^{k} (x_i - y_i)^2}$$

ii. Manhattan

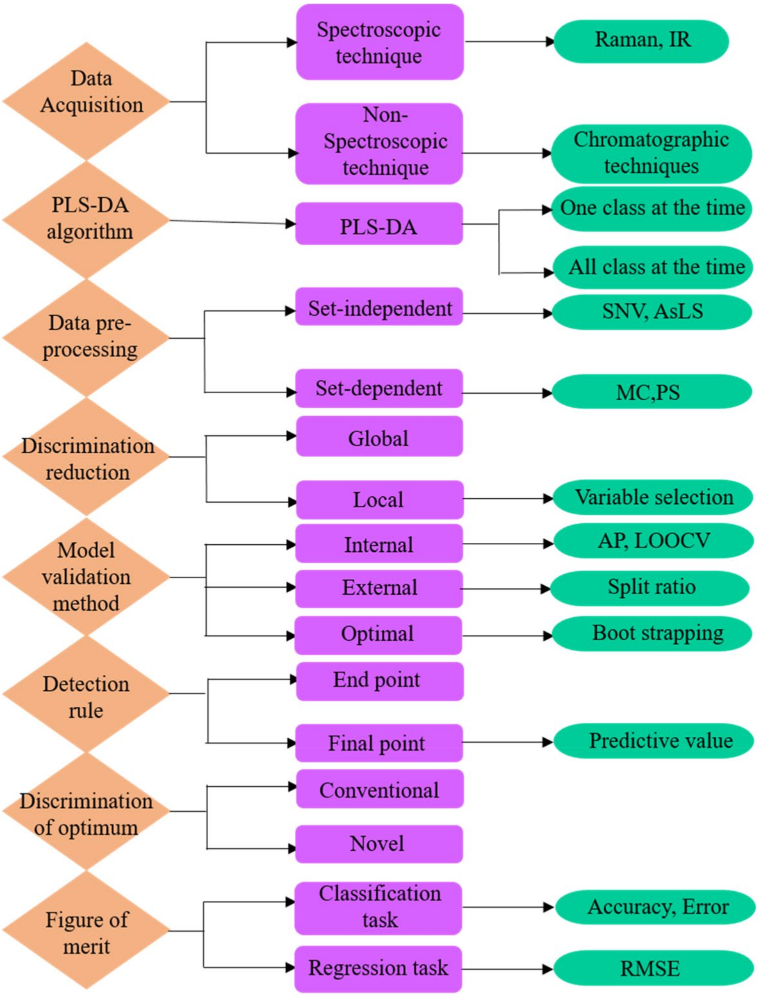

Fig. 12 Different areas of science and use of the chemometric method

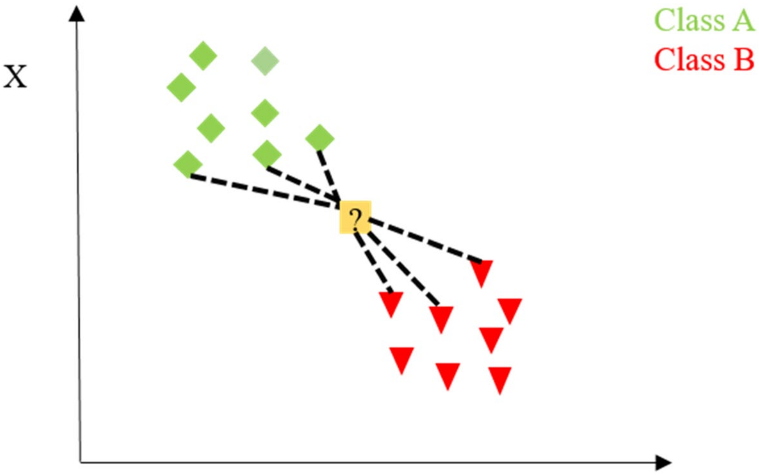

Fig. 13 k-NN method of classification of unknown object

$$D = \sum_{i=1}^{k} |x_i - y_i|$$

iii. Minkowski

$$D = \left\{ \sum_{i=1}^{k} (|x_i - y_i)|^q \right\} 1/q$$

(d) Soft independent modelling of class analogy (SIMCA)

SIMCA was first proposed by Wold. It is a class modelling technique and used to classify local models for probable groups and to predict a probable class membership for new observation. This technique runs on PCA. It implements a PCA on each of the pre-set classes from the data set (Fig. 14). The optimal number of principal components may be pre-determined and explained by cross-validation. SIMCA is a resilient method and gives additional information about the class membership. Many possibilities should be evaluated prior to modelling (Branden and Hubert 2005):

a) Scaling of the variable
b) Assess the figure of principal components
c) Assess the figure of principal components in enlarged or reduced range
d) Different values in the interior and the exterior space for the distances from the model
e) Variable scoring after class auto scaling

Main advantage of SIMCA permits classifying a sample into multiple classes.

(e) Support vector machine (SVM)

This is the latest method. It was proposed by Vapnik. This technique has become progressively accepted after their initiation in the late 1990s. This technique has so many applications in several fields like bioinformatics, biometrics and chemical science. It includes make an edge among clusters of specimens that present into different classes. This technique can be used to calculate kernel roles for such data by taking multidimensional data types beyond the feature vector (e.g. graphs, sequence) (Taylor et al. 2007). This method is explained in three parts:

a) Mathematical derivation of separable classes.
b) Extension to the non-linearly separable class by using kernel functions.
c) Incorporation of trade-off parameters to control complexity along with presentation of generalized solutions.

SVM was created as a nonlinear classifier by implementing the kernel trick to maximum hyperplane margin (Fig. 15). With a standardized dataset, hyperplane can be defined by these equations:

$$W \times X + b = \left(\frac{-10}{+1} \right)$$

where W = normal vector, X = real vector, (b = +1) = above the boundary and (b = −1) = below the boundary.

Unsupervised pattern recognition

These methods use the features of the whole data set to understand if samples relating to a class tend to jointly and segregate from other classes. The aim is to determine if there is any grouping in the data set without using the class about samples. The main unsupervised methods are as follows (Rácz et al. 2018):

a) Principal component analysis (PCA)

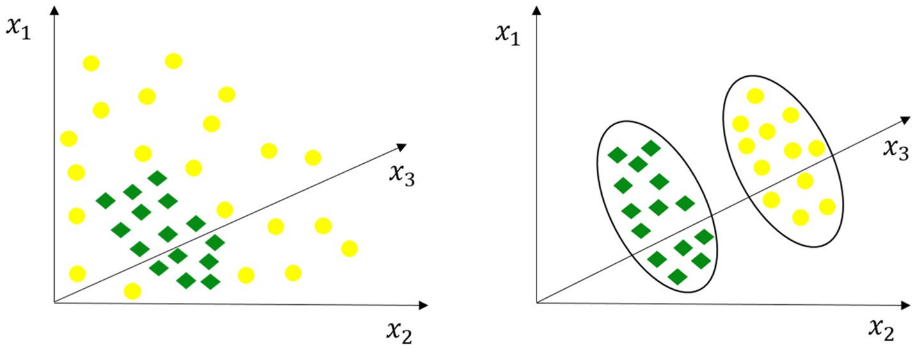

Fig. 14 SIMCA classification

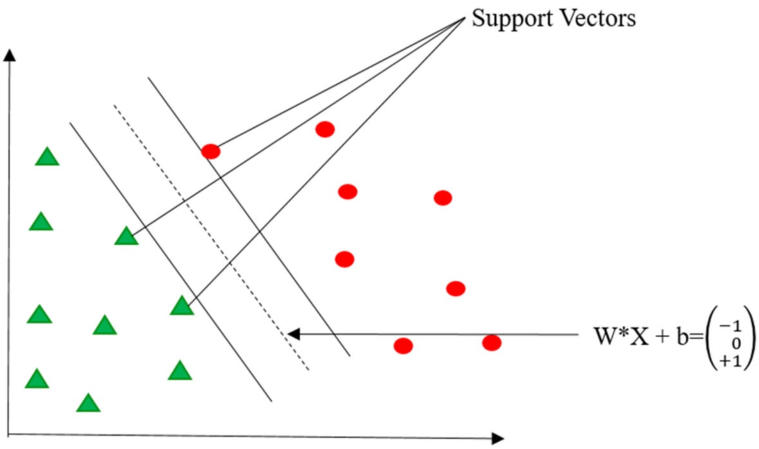

Fig. 15 Linear and nonlinear support vector machine hyperplanes

b) Hierarchical cluster analysis (HCA)

(a) Principal component analysis (PCA)

This method was proposed by Pearson in 1901. The aim of this method is to find differences and similarity between the samples. It decreases the data set into three new variables (Fig. 16).

1) Principal component
2) Scores
3) Loadings

These variables can be developed and examined latent variations. Each principal component captures as much of the difference within the data as feasible. This variation is replaced, and a new principal component is identified (Santos et al. 2019). In PCA, scores disclose information between sample variation, and loadings disclose variables

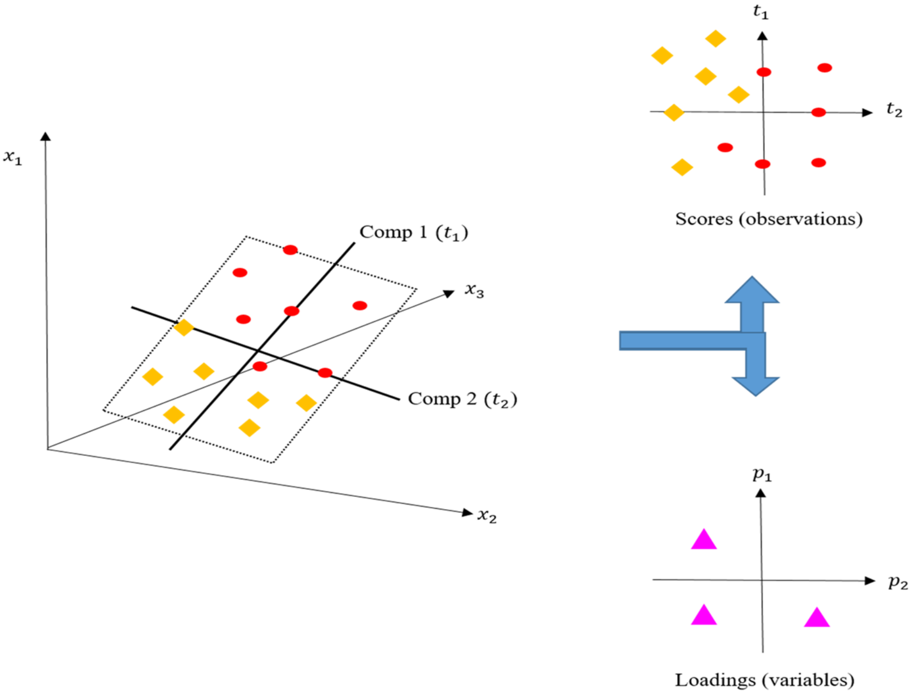

Fig. 16 Principal component analysis

which present in original data. This method can also be used to evaluate data to perform tasks such as outlier removal. It also helps in process understanding.

(b)Hierarchical cluster analysis (HCA)

The HCA is a category of chemometric method in which dendrogram (two-dimensional plot) is used to identify the distance between the samples and data set (Fig. 17). This distance can be evaluated with the help of various methods as Mahalanobis, Euclidean and Manhattan distance (Naeim et al. 2018).

Euclidean

$$D = \sqrt{(X_1 - Y_1)^2 + (X_2 + Y_2)^2 + \cdots + (X_n + Y_n)^2}$$

where $X_{\mathrm{n}=}$ coordinates of sample X in the nth dimension of row space, Y_{n} = coordinates of sample Y in the nth dimension of row space and D = distance.

Mahalanobis

$$D = \sqrt{(X_i - Y_j)^T C^{-1} + (X_i + Y_j)}$$

where X_{i} = column vectors for objects i, Y_{j} = column vectors for objects j, C = covariance matrix and D = distance.

Manhattan

$$D = \sum_{i=1}^{p} |x_i - y_i|$$

where X_{i} = vectors, Y_{i} = vectors and D = distance.

Applications of chemometric in forensic chemistry

The literature of forensic reveals a clear movement towards expanding the use of chemometrics as shown in Figs. 18–19. There is a growing body of research in which these methodologies are being used in subdisciplines of forensic chemistry. The amalgamation of spectroscopic techniques with chemometric methods yields rapid, cost-effective, and precise outcomes within the realm of chemical analysis (Smith and Siegel 2016). This part of the paper furnishes an overview of recent literature examples.

SVM-DA

Liu and his co-workers communicated two researches in which they used SERS instrument for quantification and detection of amphetamine drugs in human urine. They used SERS (laser excitation 785 nm) for detection of urine samples which mixed with various concentration of drug (0, 0.01, 0.1, 1, 10 and 100 ppm). In this instrument, gold nanorod medium was used, and specimens were evaluated on this medium when it dried but not entirely like typical SERS method. In order to identify the different concentrations of the two methamphetamines in human urine by using SVM. SERS spectra were used by calibration model for identification of 50 human urine in drugs with different concentrations (0.1, 2.5 and 50 ppm). The accuracy of the categorization was 96% for one drug, and 94% was the lowest concentration (0.1 ppm) as predicted. This procedure gave 11% better result as compare to traditional SERS method. Three actual drug addicts revealed 90% accurate classification with urine samples by using model validation. For on-site use, the analysts supported the procedure as fast (2-min analysis time) and practical (2-μL sample volume) (Han et al. 2015).

In second analysis, they formed SERS substrate from an oil in water emulsion which was made by surfactant and silver nanoparticles. Drugs were isolated from human urine in high alkaline media using cyclohexane. There

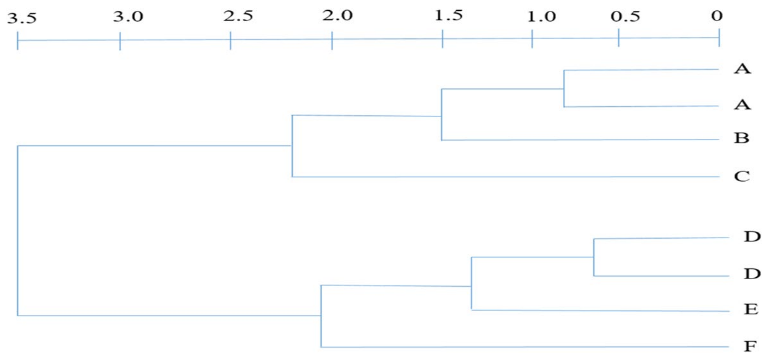

Fig. 17 A dendogram showing samples and distances towards right and upper side respectively

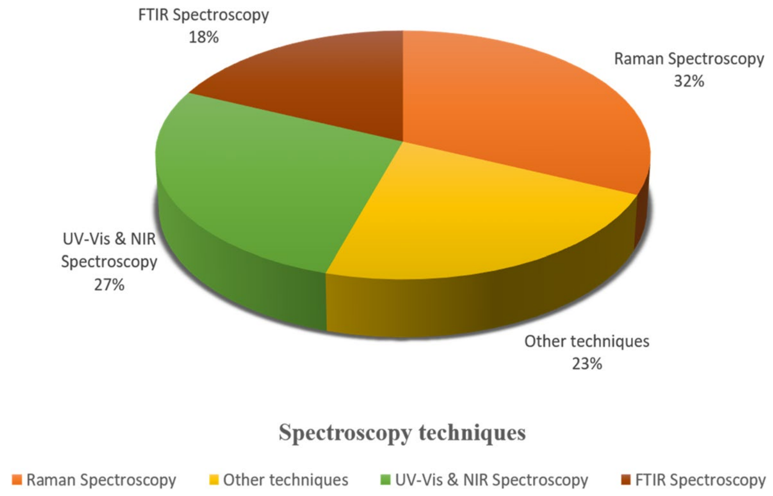

Fig. 18 Technique used in forensic science

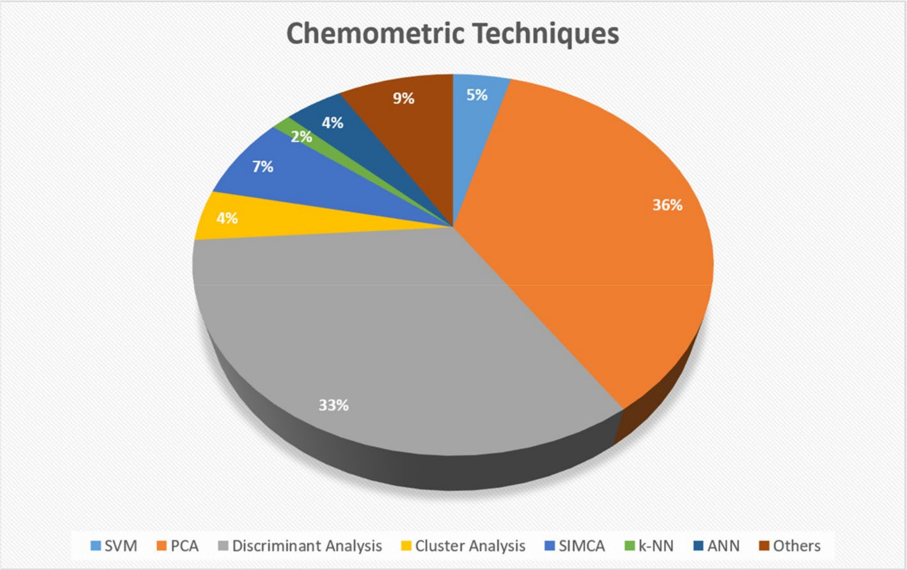

Fig. 19 Chemometrics used in forensic science

was detection limit (10 ppb) in SERS tool. In addition to ultimate detection, they proposed 98% accurate classification in different ratios of two methamphetamines from human urine by using SVM-DA and PCA (for data reduction) (Dong et al. 2015).

Dies et al. used confocal Raman (laser excitation 633 nm) for detection of various drugs (heroine, oxycodone, tetrahydrocannabinol and cocaine) from oral fluid. They formed silver nanoparticle substrates for SERS which were electrochemically immobilized. For samples analysis, the analyst used SVM method with leave-one-out cross-validation, but that model did not give more information. They found 98.3% accurate classification in different ratios of cocaine samples using this model and 100% accurate classification for drugs in solution (Dies et al. 2018).

SIMCA

Liu et al. communicated two researches for quantitative and qualitative classification and also used NIR instrument for identification of drugs (heroin, cocaine, ketamine and methamphetamine). In this method, 836 and

282 samples were used for prediction and calibration respectively. The stage of classification was used to determine the class to which the sample belonged (heroin, cocaine, ketamine and methamphetamine). The samples were guided to respective PLS model using these details. Due to greater efficiency for eliminating false-positive outcomes, SIMCA and SVM were assessed. This system was deemed precise for quantitative and qualitative examination of drugs samples (Liu et al. 2018).

E. Deconinck et al. analysed 26 synthetic and seized tablets of Viagra® using Raman micro-spectroscopic imaging (laser excitation 785 nm) and also used 8 standard tablets. MATLAB performs the data pretreatment, and SIMCA modelling is conducted when using PLS toolbox. They also used PCA for decreasing the multifaceted data set into 2D. There are no significant variations in the peak position of the standard and seized Viagra® spectrum. Even then, relative to the standard sample, the frequency of seized Viagra® is strong (Sacré et al. 2011).

HCA

Marcelo et al. evaluated 500 samples and used SVM, PLS-DA, PCA and HCA for classification of data according to their forms and patterns respectively. They used 413 samples for prediction and 100 samples for training sets. For both approaches, they found 100% vulnerability and selectivity (Marcelo et al. 2014).

PCA

Andreou et al. applied varimax rotation and PCA on microfluidic device for SERS trace identification of amphetamine from saliva. They identified the various SERS bands in every LV when rotation procedure was completed and LV was associated with illicit drug. According to this procedure, they evaluated methamphetamine in low concentration (10 nM). In this, PCA is used for automated classification of spectra (Andreou et al. 2013).

Risoluti et al. used DR-NIR tool for determination of phenethylamines and synthetic cannabinoids. They analysed standard and imitated (blotter papers and dried herbs) samples separately and examined spectroscopic spectra due to matrix influence. Originally, PCA was used to distinguish between drug forms. Based on examination of samples, the results of imitated and seizure samples revealed few clusters while scattered in the matrix (Risoluti et al. 2016).

PLS

There have been many researches communicated on portable MIR and NIR tools with multivariate calibration models (mainly PLS). These calibration data sets are used for quantitative analysis of illicit drugs which are available in powder or tablets forms. These drugs are methylamphetamine, (Hughes et al. 2013) heroin (Moros et al. 2008) and cocaine (Galipienso et al. 2014; Botelho 2015; Kahmann et al. 2018; Silva and a A. B. and M. F. P. 2019; Grobério et al. 2014).

Perez-Alfonso et al. analysed cocaine samples which were infused in seized samples and identify these samples by using PLS regression. Four various kinds of fabrics, paper and foam and black and white fabric, were distinguished accurately from the surface by diffuse reflectance DRNIR and ATR-FTIR. The samples were infused with cocaine in different concentrations such as 68.1% for foam, 38.1% and 54.1% for black and white fabric and 50.7% for paper. The fabric samples were classified using PLS model. The comparability of the validation set from new samples was tested using PCA. They resulted that if the drug was administered homogeneously, then in that case PLS was effective (approximately 4% RMSECV and RMSEP), while the matrix may be distinctive (Pérez-alfonso et al. 2018).

PLS-DA

Mabbott et al. created substrate on British coins using silver nanoparticles for SERS and analysed two amphetamines (3,4-methylenedioxymethamphetamine (MDMA), 5,6-methylenedioxy-2-aminoindane (MDAI)) and mephedrone. They used PCA and PLS-DA for measuring the reproducibility of SERS and classification respectively. Every drug is separated using three different models and also used bootstrap validation method (1000 iterations). Proportionally, the efficacy, reliability and validity were greater than 95% for these three drugs. Due to the high number of false negatives recognized, the models showed better performance for MDAI and mephedrone, however lower for MDMA. They stated that the loadings for each drug displayed distinguishing absorption spectra (Goodacre 2013).

Two studies linked to this were reported by Pereira et al. In the first study, they evaluated 21 similar papers without drugs and 73 seized samples comprising NPS on blotters. These samples were analysed using ATR-FTIR and hierarchical process through PLS-DA. In categorizing specimens into three classes (NBOMe, 2C-H and methallylescaline) and distinguishing the drug from the blank documents, these models were reasonably good. The average of CON, ACC and RLR was 86.1%, 91.1% and 88.9%, respectively (Neto et al. 2018; Pereira et al. 2017).

Massarini et al. created viable substrate which was made of gold inverted pyramids for SERS and also analysed 10 drugs in different concentrations (cocaine, methadone, amphetamine, diazepam, oxazepam, methylphenidate, morphine, tramadol, 6-monoacetylmorphine and buprenorphine). These drugs evaluated using portable Raman (laser excitation 785 nm) and confocal instruments. They detected limit of identification through spectra and used PLS-DA model for this. Since at least three spectra of 6-monoacetylmorphine, buprenorphine and morphine with an adequate signal-to-noise ratio could not be obtained, so this model could not analyse these drugs. The classification model contained 42 spectra, and validation data set contained 149 spectra from seven drugs in various concentrations with solvent that was not present in data set. The findings revealed that cocaine was not properly classified as just one amphetamine specimen. In case of limit detection, PLS-DA method worked better with exception of methylphenidate, and this method also gave better result with different concentration as compare to conventional method (Massarini et al. 2015).

Rodrigues et al. analysed 91 samples of seized cocaine powder using ATF-FTIR. The existence of caffeine, lidocaine and benzocaine along with details relevant to the chemical type of cocaine was seen in the PCA data. To differentiate between the diluted, concentrated, base and salt cocaine samples, two PLS-DA models have been constructed. They reported rates between 95 and 97% for true positives and 83 and 88% for true negatives (Federal et al. 2013).

D'Elia et al. identified cocaine in oral fluid using confocal Raman (laser excitation785 nm) and SERS which is substrate made of gold nanorods. They used OPLS-DA to differentiate between oral fluid samples at different cocaine concentrations. It had been obvious that the set of data was limited because there was not any information about model development. Although, the model was capable to identify the concentration of cocaine (1ng/ml) from oral fluid and also indicated strong distinction between the samples. The results indicated that with expanding the amount of drug doses, sample replication and the incorporation of oral fluids from various donors, this approach might be used for quantitative analysis (Elia et al. 2018).

Deconinck et al. used ATR-FTIR and DR-NIR for analysis of apprehended 267 ecstasy tablets and other party drugs. They used PLS and PLS-DA for quantitative analysis of MDMA tablets and classification for these drugs respectively. The DR-NIR showed better performance in the qualitative classification, with 96% accurate interpretation for the prediction set. Even so, the inconsistencies reported false positive. Despite it, the analysts cautioned about regular procedures because PLS quantification would be included in the follow-up review and incorrect judgments would be outcome. False negatives may also be slightly troublesome when specimens would be submitted for confirmatory examination to a laboratory. The studies indicated that periodically updating the model with new specimens seems to be the only solution to this issue, making the model quite versatile but also more accurate (Deconinck et al. 2018).

SVM

Y. Roggo et al. analysed 25 various pharmaceutical tablets using Raman instrument (laser excitation 785 nm) and chemometrics methods. They reported classification and detection of these tablets by using SVM method. In this study, the author proposed concentration of API (active pharmaceutical ingredients) in the tablets. The limit of identification found 0.59% API. SVM model gave accurate detection and classification of these 25 tablets (Roggo et al. 2010).

Challenges of chemometrics in forensic chemistry

Forensic chemistry is the core domain of forensic science, encompassing the application and advancement of chemical methodologies and techniques for examining materials pertinent to criminal investigations and legal disputes. This discipline poses numerous analytical challenges from the perspective of analytical chemistry. A significant issue frequently encountered in various applications pertains to substrate influence, particularly in the identification of body fluids and gunshot residue (GSR). The approach to mitigate this problem relies on various factors, including the choice of analytical technique. Chemometric approaches, specifically involving preprocessing techniques and weighted least squares, have proven to be particularly intriguing in this regard.

In real-world scenarios, the integration and everyday application of chemometric models continue to pose difficulties. This is primarily due to the requirement for a strong statistical foundation in model development, leading to some reluctance among forensic professionals to incorporate and create chemometric models in their daily work. In addition, it is important to emphasize that these professionals are not averse to their responsibilities; rather, they actively contribute to the criminal justice system, working towards justice for society. Furthermore, several other challenges should be taken into account when deploying these models for regular use. It is essential to continually improve and maintain a particular model, especially when dealing with complex sample analyses. Lastly, it is essential to recognize that the forthcoming prospects of forensic analysis, in combination with chemometrics, are evolving and garnering increased interest for improvement. The demand for resilient, rapid and dependable models remains a persistent challenge.

Conclusions

After wide literature research, the advances prepared when spectrographic and chromatographic techniques have been combined with chemometric tools. These techniques will always be used according to the evidence. This present review summarized the extensive discussion about unsupervised and supervised process of recognition of pattern together with their application in forensic chemistry.

About supervised techniques, PLS-DA, LDA and SVM play a role in the field of forensic biology and toxicology. These methods lead to complex data sets that need linear and nonlinear boundaries for discrimination.

About unsupervised techniques, PCA plays a major role in forensic investigation. This method has the ability to establish differences and similarities. It is also referred to as classification technique.

In present time, advanced modelling methods such as support vector machine and soft independent modelling of class analogy achieve admiration in the field of forensic chemistry. It is essential to know the future setup for chemical analysis when chemometrics will be attached with spectroscopic and chromatographic techniques. The need for a fast and dependable model is still challenging in the field of forensic chemistry.

Abbreviations

MLR	Multi-linear regression
PCR	Principal component regression
PLSR	Partial least square regression
LS-SVM	Least square support vector machine
ANN	Artificial neural network
LDA	Linear discriminant analysis
PLS-DA	Partial least square discriminant analysis
k-NN	k-nearest neighbour
SIMCA	Soft independent modelling of class analogy
SVM	Support vector machine
PCA	Principal component analysis
HCA	Hierarchical cluster analysis
IR	InfraRed
SNV	Standard normal variate
AsLS	Asymmetric least squares
MC	Mean centring
PS	Pareto scaling
AP	Auto prediction
LOOCV	Leave-one-out cross-validation
RMSE	Root-mean-squared error

Acknowledgements

Not applicable.

Authors' contributions

SS wrote the article, and HS and RS gave the structure of the article. The authors read and approved the final manuscript.

Funding

Not applicable.

Availability of data and materials

Not applicable.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Received: 19 September 2023 Accepted: 26 November 2023
Published online: 13 December 2023

References

Andreou C, Hoonejani MR, Barmi MR, Moskovits M, Meinhart CD (2013) Rapid detection of drugs of abuse in saliva using surface enhanced Raman spectroscopy and microfluidics. Am Chem Soc 7(8):7157–7164. https://doi.org/10.1021/nn402563f

Booksh KS (2006) Chemometric methods in process analysis. Encyclopedia of Analytical Chemistry: Applications, Theory and Instrumentation. pp 1–25. https://doi.org/10.1002/9780470027318.a2102

Botelho D (2015) Discrimination and quantification of cocaine and adulterants in seized drug samples by infrared spectroscopy and PLSR abstract. Forensic Sci Int 257:297–306. https://doi.org/10.1016/j.forsciint.2015.09.012

Bovens M, Ahrens B, Alberink I, Nordgaard A, Salonen T, Huhtala S (2019) Chemometrics in forensic chemistry — part i: implications to the forensic workflow. Forensic Sci Int 301:82–90. https://doi.org/10.1016/j.forsciint.2019.05.030

Branden K. Vanden, Hubert M (2005) Robust classification in high dimensions based on the SIMCA method. Chemom Intell Lab Syst 79:10–21. https://doi.org/10.1016/j.chemolab.2005.03.002

Brereton, R. G. (2015). Chemometrics and intelligent laboratory systems pattern recognition in chemometrics. Chemom Intell Lab Syst. 90–96. https://doi.org/10.1016/j.chemolab.2015.06.012

Brereton RG, Jansen J, Lopes J, Marini F, Pomerantsev A, Rodionova O (2018) Chemometrics in analytical chemistry — part II : modeling, validation, and applications. Anal Bioanal Chem 410:6691–6704. https://doi.org/10.1007/s00216-018-1283-4

Chauchard F, Cogdill R, Roussel S, Roger JM, Maurel VB, Chauchard F, Maurel VB (2010) Application of LS-SVM to non-linear phenomena in NIR spectroscopy : development of a robust and portable sensor for acidity prediction in grapes to cite this version : HAL Id : hal-00450846. J Chemolab 71:141–150. https://doi.org/10.1016/j.chemolab.2004.01.003

Colina A (1988) Chemometric methods in analytical research: a program of practical. J Chemom 3(S1):323–327. https://doi.org/10.1002/cem.1180030525

Deconinck E, Campenhout R. Van, Aouadi C, Canfyn M, Bothy JL, Gremeaux L, Courselle P (2018) Combining attenuated total reflectance- infrared spectroscopy and chemometrics for the identification and the dosage estimation of MDMA tablets. Talanta 195:142–151. https://doi.org/10.1016/j.talanta.2018.11.027

Dies H, Raveendran J, Escobedo C, Docoslis A (2018) Sensors and actuators B: chemical rapid identification and quantification of illicit drugs on nanodendritic surface-enhanced Raman scattering substrates. Sens Actuators B Chem 257:382–388. https://doi.org/10.1016/j.snb.2017.10.181

Dong R, Weng S, Yang L, Liu J (2015) Detection and direct readout of drugs in human urine using dynamic surface-enhanced Raman spectroscopy and support vector machines. Anal Chem 87(5):2937–2944. https://doi.org/10.1021/acs.analchem.5b00137

Elia VD, Rubio-retama J, Ortega-ojeda FE, García-ruiz C (2018) Gold nanorods as SERS substrate for the ultratrace detection of cocaine in non-pretreated oral fluid samples. Colloids Surfaces A 557(May):43–50. https://doi.org/10.1016/j.colsurfa.2018.05.068

Federal P, Regional S, Gerais DM (2013) Analysis of seized cocaine samples by using chemometric methods and FTIR spectroscopy. J Braz Chem Soc 24(3):507–517. https://doi.org/10.5935/0103-5053.20130066

Frank IE, Friedman JH (2013) A Statistical View of Some Chemometrics Regression Tools (Vol. 35) (Retrieved from http://www.jstor.org/stable/1269656)

Galipienso N, Garrigues S, De M, Pe C (2014) A green method for the determination of cocaine in illicit samples. Forensic Sci Int 237:70–77. https://doi.org/10.1016/j.forsciint.2014.01.015

Goodacre R (2013) 2p or not 2p: tuppence-based SERS for the detection of illicit materials†. Analyst 138(1):1–372. https://doi.org/10.1039/c2an35974j

Grobério TS, Zacca JJ, Braga JWB (2014) Quantification of cocaine hydrochloride in seized drug samples by infrared spectroscopy and PLSR. J Br Chem Soc 25(9):1696–1703. https://doi.org/10.5935/0103-5053.20140164J

Hall BP (2008) Choice of neighbor order in nearest-neighbor classification. The Annals OfStatistics 36(5):2135–2152. https://doi.org/10.1214/07-AOS537

Han Z, Liu H, Wang B, Weng S, Yang L, Liu J (2015) Three-dimensional SERS hotspots in spherical colloidal superstructure for rapid identification and detection of drugs in human urine three-dimensional SERS hotspots in spherical colloidal superstructure for rapid identification and detection of drugs in H. Anal Chem 87(9):4821–4828. https://doi.org/10.1021/acs.analchem.5b00176

Hughes, J., Ayoko, G., Collett, S., & Golding, G. (2013). Rapid quantification of methamphetamine : using attenuated total reflectance Fourier transform infrared spectroscopy (ATR-FTIR) and chemometrics. PLoS One. 8(7). https://doi.org/10.1371/journal.pone.0069609

Kahmann A, Anzanello MJ, Fogliatto FS, Marcelo MCA, Ferrão MF, Ortiz RS, Mariotti KC (2018) Journal of Pharmaceutical and Biomedical Analysis wavenumber selection method to determine the concentration of cocaine and adulterants in cocaine samples. J Pharm Biomed Analysis 152:120–127. https://doi.org/10.1016/j.jpba.2018.01.050

Kumar R, Sharma V (2018) Chemometrics in forensic science. Trends Anal Chem. https://doi.org/10.1016/j.trac.2018.05.010

Lee, L. C. (2018). Partial least squares-discriminant analysis (PLS-DA) for classification of high-dimensional (HD) data: a review of contemporary practice strategies and knowledge gaps. Rsc.Li/Analyst, (1–3), 1–14. https://doi.org/10.1039/C8AN00599K

Liu C, Han Y, Min S, Jia W, Meng X, Liu P (2018) Rapid qualitative and quantitative analysis of methamphetamine, ketamine, heroin, and cocaine by near-infrared spectroscopy. Forensic Sci Int 290:162–168. https://doi.org/10.1016/j.forsciint.2018.07.008

Marcelo MCA, Mariotti KC, Ferrão MF, Ortiz RS (2014) Profiling cocaine by ATR-FTIR. Forensic Sci Int 246:65–71. https://doi.org/10.1016/j.forsciint.2014.11.011

Marini F, Bucci R, Magrì AL, Magrì AD (2008) Artificial neural networks in chemometrics : history, examples and perspectives. Microchem J 88:178–185. https://doi.org/10.1016/j.microc.2007.11.008

Marta B (2014) Chemometric classification techniques as a tool for solving problems in analytical chemistry. J Aoac Int 97:19–29. https://doi.org/10.5740/jaoacint

Massarini E, Wästerby P, Landström L, Lejon C, Beck O, Ola P (2015) Sensors and actuators B : chemical methodologies for assessment of limit of detection and limit of identification using surface-enhanced Raman spectroscopy. SensActuators B Chem 207:437–446. https://doi.org/10.1016/j.snb.2014.09.116

Meglen R (1988) Chemometrics: its role in chemistry and measurement sciences. Chemom Intell Laboratov Syst 3(1–2):17–29. https://doi.org/10.1016/0169-7439(88)80062-5

Moros J, Galipienso N, Garrigues S, De. Guardia M (2008) Nondestructive direct determination of heroin in seized illicit street drugs by diffuse reflectance near-infrared spectroscopy. Anal Chem 80(19):7257–7265. https://doi.org/10.1021/ac800781c

Naeim M, Asri M, Nur W, Mat S (2018) Combined principal component analysis (PCA) and hierarchical cluster analysis (HCA): an efficient chemometric approach in aged gel inks discrimination combined principal component analysis (PCA) and. Aust J Forensic Sci 0618(May):1–22. https://doi.org/10.1080/00450618.2018.1466913

Neto C, Vallad FN, Sena MM (2018) Screening method for rapid classification of psychoactive substances in illicit tablets using mid infrared spectroscopy and PLS-DA6. Forensic Sci Int 288:227–235. https://doi.org/10.1016/j.forsciint.2018.05.001

Pereira LSA, Lisboa FLC, Coelho J, Valladão FN, Sena MM (2017) Direct classi fi cation of new psychoactive substances in seized blotter papers by ATR-FTIR and multivariate discriminant analysis ☆. Microchem J 133:96–103. https://doi.org/10.1016/j.microc.2017.03.032

Pérez-alfonso C, Galipienso N, Garrigues S, De M (2018) Preliminary results on direct quantitative determination of cocaine in impregnated materials by infrared spectroscopy. Microchem J 143(May):110–117. https://doi.org/10.1016/j.microc.2018.07.026

Rácz, A., Bajusz, D., & Héberger, K. (2018). Chemometrics in analytical chemistry. In T. E. and J. Gasteiger. (Ed.), Applied Chemoinformatics: Achievements and Future Opportunities (1st ed., pp. 471–499).

Risoluti R, Materazzi S, Gregori A, Ripani L (2016) Talanta early detection of emerging street drugs by near infrared spectroscopy and chemometrics. Talanta 153:407–413. https://doi.org/10.1016/j.talanta.2016.02.044

Roggo Y, Degardin K, Margot P (2010) Talanta identification of pharmaceutical tablets by Raman spectroscopy and chemometrics. Talanta 81(3):988–995. https://doi.org/10.1016/j.talanta.2010.01.046

Sacré P, Deconinck E, Saerens L, De Beer T, Courselle P, Vancauwenberghe R, De Beer JO (2011) Journal of Pharmaceutical and Biomedical Analysis detection of counterfeit Viagra ® by Raman microspectroscopy imaging and multivariate analysis. J Pharm Biomed Analysis 56:454–461. https://doi.org/10.1016/j.jpba.2011.05.042

Santos, M. C., Andrade, P., Nascimento, M., & Guedes, W. N. (2019). Chemometrics in analytical chemistry – an overview of applications from 2014 to 2018. Iq.Unesp.Br/Ecletica, 44(2), 11–25. https://doi.org/10.26850/1678-4618eqj.v44.2.11-25

Siebert KJ (2011) Using chemometrics to classify samples and detect misrepresentation. Progress in Authentication of Food and Wine. pp 39–65. https://doi.org/10.1021/bk-2011-1081.ch004

Carolina S. Silva, a A. B. and M. F. P. (2019). Vibrational spectroscopy and chemometrics in forensic chemistry: critical review, current trends and challenges. J Braz Chem Soc, 30(11). 2259–2290. https://doi.org/10.21577/0103-5053.20190140

Singh I, Juneja P, Kaur B, Kumar P (2013) Pharmaceutical applications of chemometric techniques. ISRN Anal Chem 2013:13. https://doi.org/10.1155/2013/795178

Smith R (2016) Chemometrics. In: Siegel JA (ed) Forensic Chemistry: Fundamentals and Applications, 1st edn. John Wiley & Sons, USA, pp 469–503

Taylor P, Xu Y, Zomer S, Brereton RG, Xu Y, Zomer S, Brereton RG (2007) Critical reviews in analytical chemistry support vector machines : a recent method for classification in chemometrics support vector machines : a recent method for classification in chemometrics. Crit Rev Anal Chem 36(2013):177–188. https://doi.org/10.1080/10408340600969486

Tharwat A, Gaber T, Ibrahim A, Ella A (2017) Linear discriminant analysis : a detailed tutorial. AI Commun 30:169–190. https://doi.org/10.3233/AIC-170729

Wold, S. S. (2017). What is chemometrics? In Chemometrics: Statistics and Computer Application in Analytical Chemistry (Third, pp. 1–13). John Wiley & Sons.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.