International Journal of Life Sciences
International Peer Reviewed Open Access Journal

Special Issue A18, May 2022
ISSN:2320-7817(p) 2320-964X(0)

Original Article Open Access

Chemometrics in Forensic Science

Raheena Bin Mohammed

Department of chemical engineering, National Institute 0f Technology Agartala
Email: raheenabinmohd@gmail.com

Manuscript details:

Available online on http://www.ijlsci.in
ISSN: 2320-964X (Online)
ISSN: 2320-7817 (Print)

Cite this article as:
Raheena Bin Mohammed (2022) Chemometrics in forensic science, *Int. J. of. Life Sciences*, Special Issue, A18: 31-36.

Article published in Special issue of 1st National Conference on Forensic Science & Digital Forensics 2022 organised by Applied Forensic Research Science From 18th to 20th March 2022.

ABSTRACT

The last decade has seen the application of the chemometric methods combined with analytical techniques for characterization and discrimination of samples, which leads to the informative and representative examinations of the samples. Many research articles with reference to the use of chemometrics in forensic science have been published. This review has been divided into various sections which include chemometrics, its history, multivariate methods, and the application of chemometrics in various disciplines of forensic science. Chemometric methods are expedient due to their ease of interpreting results, reliability, and speed. Advanced modeling methods such as SIMCA and SVM are gaining popularity. It is suggested that these new techniques and mathematical/statistical methods should be utilized in forensic science casework to get statistical confidence in the results.

Keywords: Multivariate, methods, SIMCS, SVM, Forensic.

INTRODUCTION

A primary aim in many forensic investigations is to establish links between people, places, or objects in order to reconstruct events surrounding a crime (Coyle, 2010). This is typically done through the recovery, analysis and interpretation of physical evidence. Many items of physical evidence are macroscopic items such as clothing or firearms. However, this category also includes 'trace evidence' (such as soil, glass, paint, hair, fibres, or explosive particulates) that can be cross-transferred between surfaces through physical contact or proximity.

Assuming that an item of physical evidence is successfully recovered and analysed, significant challenges arise in its interpretation. Many forensic disciplines rely on visual comparisons of complex images or multivariate chemical data in the form of spectra (Roux and Robertson, 2013), chromatograms or other analytical output 4,5 These comparisons require substantial time and expertise on the part of the examiner, and the visual complexity of the data may veil potentially useful information.

A widely accepted way of approaching this is through statistical inference, which allows the strength of evidence to be described in probabilistic terms.

- "What is the probability that a randomly selected individual other than the suspect would exhibit an indistinguishable DNA profile?"
- "What is the probability of the profile being observed if the DNA originated from the suspect, compared to if it originated from another random individual?"

However, this does not address potential error or bias in the interpretation of the evidence itself. For this reason, an increasing volume of literature has emerged investigating chemo metric techniques for the analysis and interpretation of physical evidence. The use of statistical techniques allows more objective and quantitative measures of data to be made compared to visual inspections.

Chemo metrics

Chemo metrics is the chemical discipline that uses mathematical, statistical, and other methods employing formal logic to design or select optimal measurement procedures and experiments, and to provide maximum relevant chemical information by analyzing chemical data.

Chemo metrics in forensic

Chemo metrics has been recognized as a powerful tool within forensic science for interpretation and optimization of analytical procedures. The chemo metric methods give better resolution or separation quality of the samples which define its incorporation with analytical (spectroscopic/chromatographic) techniques in recent times.

Unsupervised pattern recognition

The purpose of unsupervised learning is to detect patterns in datasets without setting any prior labels or outcomes. The algorithm is thus left to infer patterns (Morgan and Bartick, 2013) with minimal human intervention. Whilst these methods cannot be directly applied to classification or regression problems, they are ideal for probing the underlying structure of data. Additionally, some unsupervised methods allow new samples to be projected onto a pre-existing dataset. This is useful for comparative purposes but should not strictly be considered classification, as there is no assumption of specific classes existing.

Hierarchical cluster analysis

Cluster analysis refers to a group of algorithms (Saferstein, Forensic Science, 2009) in which objects (samples) are grouped according to their relative similarity. The choice of algorithm and its parameters depends on the properties of the dataset and the purpose of the analysis. This process is most often agglomerative; starting with single objects and progressively grouping them into larger clusters.

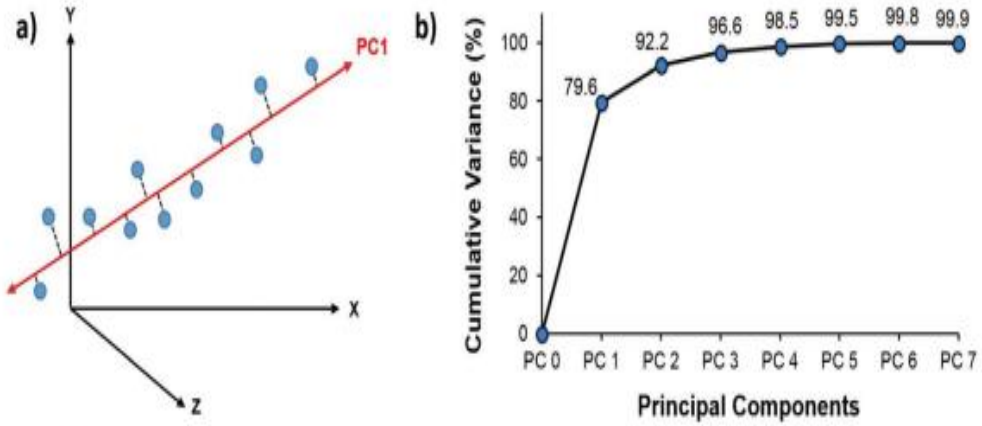

Fig. 2 (a) Diagram illustrating generation of the first principal component (PC) for a simplified dataset described by variables *X*, *Y* and *Z*. Dashed lines indicate the projection of each sample onto the PC; (b) example scree plot showing cumulative variance accounted for by each successive PC.

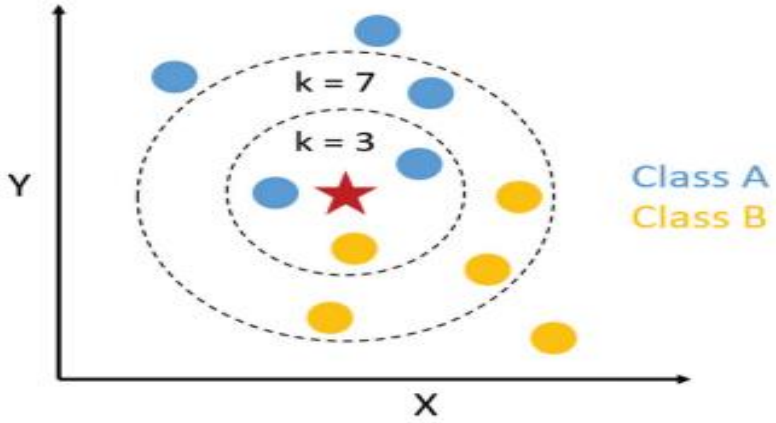

Fig. 4 Representation of a kNN analysis where an unknown sample (red star) must be assigned to Class A or Class B, determined by the mode class value of the '*k*' closest training samples. The resulting classification may vary depending on the selection of '*k*'.

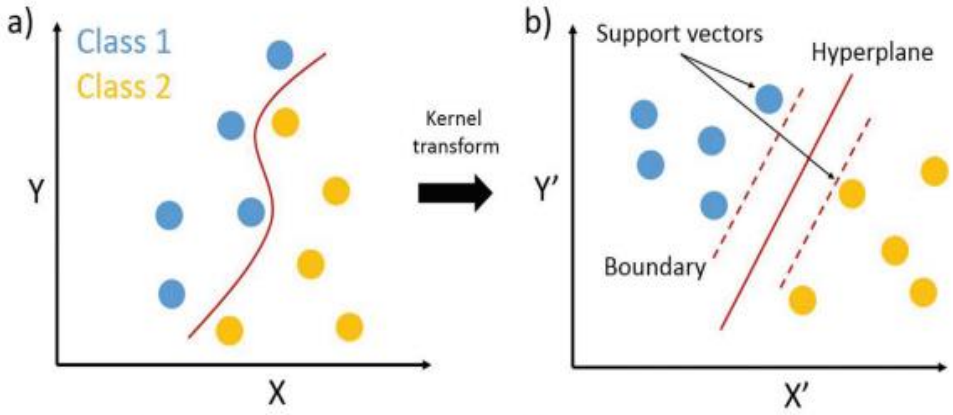

Fig. 6 (a) Two hypothetical sample classes that cannot be linearly separated, (b) samples mapped to a higher dimensional plane in which a linear hyperplane can be constructed. When mapped back into the lower dimensional plane, this hyperplane becomes a non-linear classifier.

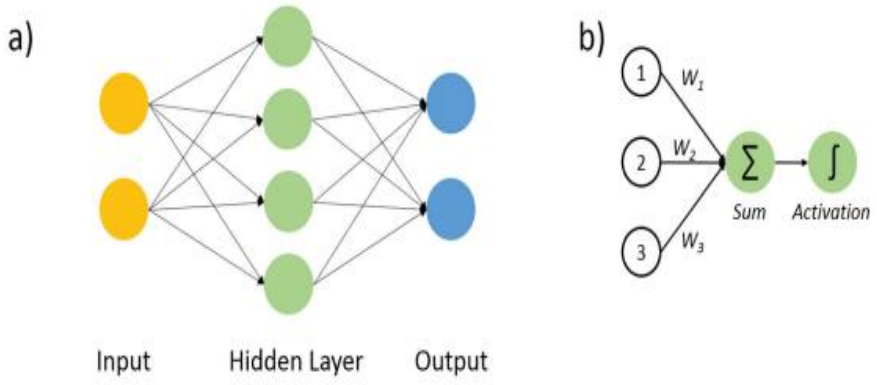

Fig. 8 (a) General schematic showing the architecture of a fully connected ANN with a single hidden layer, (b) diagrammatic representation of the processing functions within a neuron.

Principal component analysis

Identifying relationships within a dataset can be challenging when examining complex multivariate data such as spectra or chromatograms. Additionally, it may be of interest to identify the features responsible for similarity or dissimilarity between objects. This information can only be extracted from cluster analysis by inferring from visual comparisons of the original data which may not be practical for large datasets. PCA reduces data dimensionality by extracting the orthogonal (Adam et al., 2016) sources of variation; known as 'principal components' or PCs. These PCs are a linear combination of the original variables, each multiplied by a loading. Each successive PC is calculated to describe the maximum proportion of remaining variation in the data. This process is represented in Fig. 2a for a simplified dataset.

Here, samples are initially represented as data points in a feature space described by the original variables; X, Y and Z. Each PC is determined by finding the direction along which the remaining dispersion of the data poinis the greatest. Projection of the samples onto the components then allows them to be alternatively described using the PC coordinates, or scores. A screen plot, which shows the cumulative variance accounted for with each successive PC, can be used to indicate those containing 'useful' information.

Supervised pattern recognition

The purpose of supervised learning is to map labelled (Brereton, 2003) inputs to an expected output. Input variables are paired with the desired output or classification, with the algorithm being tasked to develop a function that correlates the two. In other words, supervised algorithms are designed to create functions based on known data that can then draw inferences about new samples. These methods are ideal for classification and regression problems, as the model can be 'trained' to detect and accurately model specific patterns. The model must then be validated to assess how it will generalise these patterns to independent datasets. The importance of rigorous validation cannot be overstated, as failure to establish accurate error rates has significant implications for criminal justice (Smith, 2016).

K – NEAREST NIEGHBOUR (knn)

The k-nearest neighbor (kNN) method is based on the distance between known and unknown samples. The training set is divided into known classes, and the distance metric between the unknown and each training sample determined. The unknown is subsequently assigned to the pre-determined class that is most common among its 'k' nearest neighbor's (Fig. 4). 'k' is generally a small integer, with the optimal value often determined through cross-validation.

Linear discriminant analysis

Discriminant analysis (DA) aims to classify objects (Kassin et al., 2013) into pre-defined, mutually exclusive classes based on scores derived from a discriminant function. This function is a combination of input variables, calculated to maximise the ratio of between-class to within-class variance (the Fisher ratio). If it is assumed that the separating function is linear, this method is referred to as linear discriminant analysis (LDA). This approach is similar to PCA, in that they each look for linear combinations of variables to best describe trends in complex data. LDA is hence a valuable tool for data visualisation and classification. However, LDA requires the number of samples to exceed the number of descriptor variables. For this reason, LDA is often carried out following a preliminary data reduction method such as PCA.

SIMCA

In some scenarios, classification needs to be more flexible than kNN or LDA would permit. Soft independent modelling of class analogy (SIMCA) is a disjoint technique, meaning that it constructs separate models (in this case derived using PCA) describing a boundary around each class (Jackson and Jackson, 2011). The number of PCs retained for each model is highly influential, as too few will result in a loss of information whilst too many will introduce noise. Cross-validation is typically used for determining the optimal number of PCs describing each class, although other measures such as the Malinowski indicator function can also be used. Classification of a new object is based on its sample-tomodel distance for each class as determined by two limits; residual variance and leverage (Forensic Science Regulator, 2017). The residual variance of an object, calculated as a residual sum of squares, is that which remains unexplained after projection onto a known class. The critical limit for

residuals distances is commonly determined using a chi-square distribution or Jackson-Mudholkar approximation.

Leverage is the Mahalanobis distance between an object and the centroid of a given class, with a high leverage indicating outlying objects in the model space. An object fallswithin the above limits if the residual variance and leverage do not exceed the selected cut-off values. Objects will be assigned to any class for which they meet both of the above limits. Unlike other supervised techniques, SIMCA thus permits classification into one, multiple, or none of the known classes. 'Soft' classifications allow easier identification of atypical samples compared to kNN or LDA. On the other hand, the disjoint modelling of each class means that there is limited measure of between-class to within-class variation. Consequently, SIMCA is highly sensitive to sample leverage and variance, which may result in incorrect rejections (false negatives). This may be compensated for by reducing the significance level (α), or the percentage of training samples deemed acceptable as outliers.

SVM

In situations where known classes cannot be linearly (Christensen et al., 2014) separated, support vector machines (SVMs) are considered by many as an ideal classifier. In SVM analysis, samples are considered as points in a multidimensional space. Boundaries between different classes are established in the form of a separating gap or hyperplane, which is mapped to be as wide as possible (Fig. 6a). The hyperplane (classifier) can be expressed in terms of the samples lying closest to its boundaries, which are known as support vectors. Classification then occurs by mapping new samples into the same space as the training set, and assigning them toa known class based upon which side of the boundary they fall on. In order to establish the classifier between non-linearly separable classes, the data is transformed through a mathematical function (referred to as a kernel). This transforms the data into a higher-dimensional space, in which linear separation is then achievable (Fig. 6b). The mathematical basis behind such kernel transformations (the most common being linear, polynomial or Gaussian radial basis function) or computation of the classifier function are beyond the scope of this review, but have been covered in other literature

Artificial neural networks

Artificial neural networks (ANNs) are computational models based on the assembly and functions of biological neural structures. ANNs consist of interconnected nodes or 'neurons' that are typically aggregated into an input layer, one or more hidden layers, and an output layer, as shown in Fig. 8a. This figure shows a fully connected network, where each neuron in a given layer connects to every neuron in the next layer. Other architectures include pooled (feed-forward) networks, (Forensic Science Regulator, 2015) in which neurons of one layer connect to a single neuron in the subsequent layer; or recurrent networks that allow links to be formed to the same or previous layers.56 The more complex the network, the more powerful it becomes. However, they will also require a greater degree of training in order to establish relationships. Inputs received by a neuron are multiplied by a weighting through various mathematical operations (Fig. 8b) that can be considered to reflect the strength of the neural connection. The resulting products of the inputs are summed, then processed through a non-linear activation function to generate an output.22 This output must meet a minimum threshold in order to be passed on to subsequent neurons. When the information reaches the output layer, an appropriate response is generated.

Multivariate regression

Multivariate regression methods are used to establish quantitative relationships between multiple predictor variables and a dependent response. The resulting model can then predict the response of an unknown sample based on measured predictor data. Such approaches can also be used for binary classification by assigning arbitrary responses to each class.

Multiple linear regression (MLR)

Principal component regression (PCR)

Partial least squares regression (PLSR)

Conflicts of interest: The authors stated that no conflicts of interest.

REFERENCES

Coyle T (2010) in Crime Scene to Court: The Essentials of Forensic Science, ed. P. White, RSC Publishing, Cambridge, 2010, pp. 106–126

Roux C and Robertson J (2013) in Encyclopedia of Forensic Sciences, ed. J. A. Siegel and P. J. Saukko, Academic Press, Waltham, 2nd edn, pp. 279–285.

Saferstein R, Forensic Science: From the Crime Scene to the Crime Lab, Pearson Education, Upper Saddle River, NJ, 2009.

Jackson ARW and Jackson JM (2011) in Forensic Science, Pearson Education, Harlow, England, 3rd edn, 2011, pp. 1–14.

Smith R, in Forensic Chemistry: Fundamentals and Applications, ed. J. A. Siegel, John Wiley & Sons, Chichester, UK, 2016, pp. 469–503.

R. G. Brereton, Chemometrics: Data Analysis for the Laboratory and Chemical Plant, John Wiley & Sons, London, 2003.

S. L. Morgan and E. G. Bartick, in Forensic Analysis on the Cutting Edge: New Methods for Trace Evidence Analysis, ed. R. D. Blackledge, John Wiley & Sons, New Jersey, 2007, pp. 333–374.

A. M. Christensen, C. M. Crowder, S. D. Ousley and M. M. Houck, J. Forensic Sci., 2014, 59, 123–126.

I. E. Dror, D. Charlton and A. E. Péron, Forensic Sci. Int., 2006, 156, 74–78.

S. M. Kassin, I. E. Dror and J. Kukucka, J. Appl. Res. Mem. Cogn., 2013, 2, 42–52

European Network for Forensic Science Institutes, 2015.

Forensic Science Regulator, Annual Report November 2014–November 2015, 2015.

Forensic Science Regulator, Annual Report November 2015–November 2016, 2017.

Forensic Science Regulator, Annual Report November 2016–November 2017, 2018.

Forensic Science Regulator, Annual Report November 2017–November 2018, 2019.

President's Council of Advisors on Science and Technology, Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods, Washington DC, 2016.

National Academy of Sciences and T. N. A. Press, Strengthening Forensic Science in the United States: A Path Forward, Committee on Identifying the Needs of the Forensic Sciences Community, National Research Council, Washington DC, 2009.