International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

| ISSN: 2582-7219 | Monthly, Peer Reviewed & Referred Journal |

Volume 2, Issue 3, March 2019

Chemometrics - Science of Extracting Information from Chemical Systems by Data-Driven Means

Dr. Anil Kumar

Dept. of Chemistry, D.A.V. (PG) College, Dehradun, India

ABSTRACT: Chemometrics is inherently interdisciplinary, using methods frequently employed in core data-analytic disciplines such as multivariate statistics, applied mathematics, and computer science, in order to address problems in chemistry, biochemistry, medicine, biology and chemical engineering. In this way, it mirrors other interdisciplinary fields, such as psychometrics and econometrics. Chemometrics is applied to solve both descriptive and predictive problems in experimental natural sciences, especially in chemistry. In descriptive applications, properties of chemical systems are modeled with the intent of learning the underlying relationships and structure of the system (i.e., model understanding and identification). In predictive applications, properties of chemical systems are modeled with the intent of predicting new properties or behavior of interest. In both cases, the datasets can be small but are often large and complex, involving hundreds to thousands of variables, and hundreds to thousands of cases or observations.

KEYWORDS: chemometrics, multivariate, complex, observations, cases, learning, descriptive, predictive, analytic

I. INTRODUCTION

Chemometric techniques are particularly heavily used in analytical chemistry and metabolomics, and the development of improved chemometric methods of analysis also continues to advance the state of the art in analytical instrumentation and methodology. It is an application-driven discipline, and thus while the standard chemometric methodologies are very widely used industrially, academic groups are dedicated to the continued development of chemometric theory, method and application development. Although one could argue that even the earliest analytical experiments in chemistry involved a form of chemometrics, the field is generally recognized to have emerged in the 1970s as computers became increasingly exploited for scientific investigation. The term 'chemometrics' was coined by Svante Wold in a 1971 grant application,[1] and the International Chemometrics Society was formed shortly thereafter by Svante Wold and Bruce Kowalski, two pioneers in the field. Wold was a professor of organic chemistry at Umeå University, Sweden, and Kowalski was a professor of analytical chemistry at University of Washington, Seattle.[2]

Many early applications involved multivariate classification, numerous quantitative predictive applications followed, and by the late 1970s and early 1980s a wide variety of data- and computer-driven chemical analyses were occurring.

Multivariate analysis was a critical facet even in the earliest applications of chemometrics. Data from infrared and UV/visible spectroscopy are often counted in thousands of measurements per sample. Mass spectrometry, nuclear magnetic resonance, atomic emission/absorption and chromatography experiments are also all by nature highly multivariate. The structure of these data was found to be conducive to using techniques such as principal components analysis (PCA), partial least-squares (PLS), orthogonal partial least-squares (OPLS), and two-way orthogonal partial least squares (O2PLS).[3] This is primarily because, while the datasets may be highly multivariate there is strong and often linear low-rank structure present. PCA and PLS have been shown over time very effective at empirically modeling the more chemically interesting low-rank structure, exploiting the interrelationships or 'latent variables' in the data, and providing alternative compact coordinate systems for further numerical analysis such as regression, clustering, and pattern recognition. Partial least squares in particular was heavily used in chemometric applications for many years before it began to find regular use in other fields.

Through the 1980s three dedicated journals appeared in the field: Journal of Chemometrics, Chemometrics and Intelligent Laboratory Systems, and Journal of Chemical Information and Modeling. These journals continue to cover both fundamental and methodological research in chemometrics. At present, most routine applications of existing chemometric methods are commonly published in application-oriented journals (e.g., Applied Spectroscopy, Analytical Chemistry, Analytica Chimica Acta, Talanta). Several important books/monographs on chemometrics were also first published in the 1980s, including the first edition of Malinowski's Factor Analysis in Chemistry,[4] Sharaf, Illman and

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

| ISSN: 2582-7219 | Monthly, Peer Reviewed & Referred Journal |

Volume 2, Issue 3, March 2019

Kowalski's Chemometrics,[5] Massart et al. Chemometrics: a textbook,[6] and Multivariate Calibration by Martens and Naes.[7]

Some large chemometric application areas have gone on to represent new domains, such as molecular modeling and QSAR, cheminformatics, the '-omics' fields of genomics, proteomics, metabonomics and metabolomics, process modeling and process analytical technology.

An account of the early history of chemometrics was published as a series of interviews by Geladi and Esbensen.[8][9]

Many chemical problems and applications of chemometrics involve calibration. The objective is to develop models which can be used to predict properties of interest based on measured properties of the chemical system, such as pressure, flow, temperature, infrared, Raman, NMR spectra and mass spectra. Examples include the development of multivariate models relating 1) multi-wavelength spectral response to analyte concentration, 2) molecular descriptors to biological activity, 3) multivariate process conditions/states to final product attributes. The process requires a calibration or training data set, which includes reference values for the properties of interest for prediction, and the measured attributes believed to correspond to these properties. For case 1), for example, one can assemble data from a number of samples, including concentrations for an analyte of interest for each sample (the reference) and the corresponding infrared spectrum of that sample. Multivariate calibration techniques such as partial-least squares regression, or principal component regression (and near countless other methods) are then used to construct a mathematical model that relates the multivariate response (spectrum) to the concentration of the analyte of interest, and such a model can be used to efficiently predict the concentrations of new samples.

Techniques in multivariate calibration are often broadly categorized as classical or inverse methods.[7][10] The principal difference between these approaches is that in classical calibration the models are solved such that they are optimal in describing the measured analytical responses (e.g., spectra) and can therefore be considered optimal descriptors, whereas in inverse methods the models are solved to be optimal in predicting the properties of interest (e.g., concentrations, optimal predictors).[11] Inverse methods usually require less physical knowledge of the chemical system, and at least in theory provide superior predictions in the mean-squared error sense,[12][13][14] and hence inverse approaches tend to be more frequently applied in contemporary multivariate calibration.

The main advantages of the use of multivariate calibration techniques is that fast, cheap, or non-destructive analytical measurements (such as optical spectroscopy) can be used to estimate sample properties which would otherwise require time-consuming, expensive or destructive testing (such as LC-MS). Equally important is that multivariate calibration allows for accurate quantitative analysis in the presence of heavy interference by other analytes. The selectivity of the analytical method is provided as much by the mathematical calibration, as the analytical measurement modalities. For example, near-infrared spectra, which are extremely broad and non-selective compared to other analytical techniques (such as infrared or Raman spectra), can often be used successfully in conjunction with carefully developed multivariate calibration methods to predict concentrations of analytes in very complex matrices.

II. DISCUSSION

Supervised multivariate classification techniques are closely related to multivariate calibration techniques in that a calibration or training set is used to develop a mathematical model capable of classifying future samples. The techniques employed in chemometrics are similar to those used in other fields – multivariate discriminant analysis, logistic regression, neural networks, regression/classification trees. The use of rank reduction techniques in conjunction with these conventional classification methods is routine in chemometrics, for example discriminant analysis on principal components or partial least squares scores.

A family of techniques, referred to as class-modelling or one-class classifiers, are able to build models for an individual class of interest. [15] Such methods are particularly useful in the case of quality control and authenticity verification of products.

Unsupervised classification (also termed cluster analysis) is also commonly used to discover patterns in complex data sets, and again many of the core techniques used in chemometrics are common to other fields such as machine learning and statistical learning. In chemometric parlance, multivariate curve resolution seeks to deconstruct data sets with limited or absent reference information and system knowledge. Some of the earliest work on these techniques was done by Lawton and Sylvestre in the early 1970s.[16][17] These approaches are also called self-modeling mixture analysis, blind source/signal separation, and spectral unmixing. For example, from a data set comprising fluorescence spectra from a series of samples each containing multiple fluorophores, multivariate curve resolution methods can be used to extract the fluorescence spectra of the individual fluorophores, along with their relative concentrations in each of the samples, essentially unmixing the total fluorescence spectrum into the contributions from the individual

components. The problem is usually ill-determined due to rotational ambiguity (many possible solutions can equivalently represent the measured data), so the application of additional constraints is common, such as non-negativity, unimodality, or known interrelationships between the individual components (e.g., kinetic or mass-balance constraints).[18][19]

Experimental design remains a core area of study in chemometrics and several monographs are specifically devoted to experimental design in chemical applications.[20][21] Sound principles of experimental design have been widely adopted within the chemometrics community, although many complex experiments are purely observational, and there can be little control over the properties and interrelationships of the samples and sample properties.

Signal processing is also a critical component of almost all chemometric applications, particularly the use of signal pretreatments to condition data prior to calibration or classification. The techniques employed commonly in chemometrics are often closely related to those used in related fields.[22] Signal pre-processing may affect the way in which outcomes of the final data processing can be interpreted.[23]

Performance characterization, and figures of merit Like most arenas in the physical sciences, chemometrics is quantitatively oriented, so considerable emphasis is placed on performance characterization, model selection, verification & validation, and figures of merit. The performance of quantitative models is usually specified by root mean squared error in predicting the attribute of interest, and the performance of classifiers as a true-positive rate/false-positive rate pairs (or a full ROC curve). A recent report by Olivieri et al. provides a comprehensive overview of figures of merit and uncertainty estimation in multivariate calibration, including multivariate definitions of selectivity, sensitivity, SNR and prediction interval estimation.[24] Chemometric model selection usually involves the use of tools such as resampling (including bootstrap, permutation, cross-validation).

Multivariate statistical process control (MSPC), modeling and optimization accounts for a substantial amount of historical chemometric development.[25][26][27] Spectroscopy has been used successfully for online monitoring of manufacturing processes for 30–40 years, and this process data is highly amenable to chemometric modeling. Specifically in terms of MSPC, multiway modeling of batch and continuous processes is increasingly common in industry and remains an active area of research in chemometrics and chemical engineering. Process analytical chemistry as it was originally termed,[28] or the newer term process analytical technology continues to draw heavily on chemometric methods and MSPC.

Multiway methods are heavily used in chemometric applications.[29][30] These are higher-order extensions of more widely used methods. For example, while the analysis of a table (matrix, or second-order array) of data is routine in several fields, multiway methods are applied to data sets that involve 3rd, 4th, or higher-orders. Data of this type is very common in chemistry, for example a liquid-chromatography / mass spectrometry (LC-MS) system generates a large matrix of data (elution time versus m/z) for each sample analyzed. The data across multiple samples thus comprises a data cube. Batch process modeling involves data sets that have time vs. process variables vs. batch number. The multiway mathematical methods applied to these sorts of problems include PARAFAC, trilinear decomposition, and multiway PLS and PCA.

III. RESULTS

Principal component analysis (PCA) is a popular technique for analyzing large datasets containing a high number of dimensions/features per observation, increasing the interpretability of data while preserving the maximum amount of information, and enabling the visualization of multidimensional data. Formally, PCA is a statistical technique for reducing the dimensionality of a dataset. This is accomplished by linearly transforming the data into a new coordinate system where (most of) the variation in the data can be described with fewer dimensions than the initial data. Many studies use the first two principal components in order to plot the data in two dimensions and to visually identify clusters of closely related data points. Principal component analysis has applications in many fields such as population genetics, microbiome studies, and atmospheric science.[1]

PCA can be thought of as fitting a p-dimensional ellipsoid to the data, where each axis of the ellipsoid represents a principal component. If some axis of the ellipsoid is small, then the variance along that axis is also small.

To find the axes of the ellipsoid, we must first center the values of each variable in the dataset on 0 by subtracting the mean of the variable's observed values from each of those values. These transformed values are used instead of the original observed values for each of the variables. Then, we compute the covariance matrix of the data and calculate the eigenvalues and corresponding eigenvectors of this covariance matrix. Then we must normalize each of the orthogonal eigenvectors to turn them into unit vectors. Once this is done, each of the mutually-orthogonal unit eigenvectors can be interpreted as an axis of the ellipsoid fitted to the data. This choice of basis will transform the covariance matrix into a

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

| ISSN: 2582-7219 | Monthly, Peer Reviewed & Referred Journal |

Volume 2, Issue 3, March 2019

diagonalized form, in which the diagonal elements represent the variance of each axis. The proportion of the variance that each eigenvector represents can be calculated by dividing the eigenvalue corresponding to that eigenvector by the sum of all eigenvalues.

Biplots and scree plots (degree of explained variance) are used to explain findings of the PCA.

The singular values (in Σ) are the square roots of the eigenvalues of the matrix X^TX. Each eigenvalue is proportional to the portion of the "variance" (more correctly of the sum of the squared distances of the points from their multidimensional mean) that is associated with each eigenvector. The sum of all the eigen values is equal to the sum of the squared distances of the points from their multidimensional mean. PCA essentially rotates the set of points around their mean in order to align with the principal components. This moves as much of the variance as possible (using an orthogonal transformation) into the first few dimensions. The values in the remaining dimensions, therefore, tend to be small and may be dropped with minimal loss of information (see below). PCA is often used in this manner for dimensionality reduction. PCA has the distinction of being the optimal orthogonal transformation for keeping the subspace that has largest "variance" (as defined above). This advantage, however, comes at the price of greater computational requirements if compared, for example, and when applicable, to the discrete cosine transform, and in particular to the DCT-II which is simply known as the "DCT". Nonlinear dimensionality reduction techniques tend to be more computationally demanding than PCA.

PCA is sensitive to the scaling of the variables. If we have just two variables and they have the same sample variance and are completely correlated, then the PCA will entail a rotation by 45° and the "weights" (they are the cosines of rotation) for the two variables with respect to the principal component will be equal. But if we multiply all values of the first variable by 100, then the first principal component will be almost the same as that variable, with a small contribution from the other variable, whereas the second component will be almost aligned with the second original variable. This means that whenever the different variables have different units (like temperature and mass), PCA is a somewhat arbitrary method of analysis. (Different results would be obtained if one used Fahrenheit rather than Celsius for example.) Pearson's original paper was entitled "On Lines and Planes of Closest Fit to Systems of Points in Space" – "in space" implies physical Euclidean space where such concerns do not arise. One way of making the PCA less arbitrary is to use variables scaled so as to have unit variance, by standardizing the data and hence use the autocorrelation matrix instead of the autocovariance matrix as a basis for PCA. However, this compresses (or expands) the fluctuations in all dimensions of the signal space to unit variance.

Mean subtraction (a.k.a. "mean centering") is necessary for performing classical PCA to ensure that the first principal component describes the direction of maximum variance. If mean subtraction is not performed, the first principal component might instead correspond more or less to the mean of the data. A mean of zero is needed for finding a basis that minimizes the mean square error of the approximation of the data.[15]

Mean-centering is unnecessary if performing a principal components analysis on a correlation matrix, as the data are already centered after calculating correlations. Correlations are derived from the cross-product of two standard scores (Z-scores) or statistical moments (hence the name: Pearson Product-Moment Correlation). Also see the article by Kromrey & Foster-Johnson (1998) on "Mean-centering in Moderated Regression: Much Ado About Nothing". Since covariances are correlations of normalized variables (Z- or standard-scores) a PCA based on the correlation matrix of X is equal to a PCA based on the covariance matrix of Z, the standardized version of X.

PCA is a popular primary technique in pattern recognition. It is not, however, optimized for class separability.[16] However, it has been used to quantify the distance between two or more classes by calculating center of mass for each class in principal component space and reporting Euclidean distance between center of mass of two or more classes.[17] The linear discriminant analysis is an alternative which is optimized for class separability.

IV. CONCLUSIONS

In PCA, it is common that we want to introduce qualitative variables as supplementary elements. For example, many quantitative variables have been measured on plants. For these plants, some qualitative variables are available as, for example, the species to which the plant belongs. These data were subjected to PCA for quantitative variables. When analyzing the results, it is natural to connect the principal components to the qualitative variable species. For this, the following results are produced.

- Identification, on the factorial planes, of the different species, for example, using different colors.
- Representation, on the factorial planes, of the centers of gravity of plants belonging to the same species.

- For each center of gravity and each axis, p-value to judge the significance of the difference between the center of gravity and origin.

These results are what is called introducing a qualitative variable as supplementary element. This procedure is detailed in and Husson, Lê & Pagès 2009 and Pagès 2013. Few software offer this option in an "automatic" way. This is the case of SPAD that historically, following the work of Ludovic Lebart, was the first to propose this option, and the R package FactoMineR.

REFERENCES

1. As recounted in Wold, S. (1995). "Chemometrics; what do we mean with it, and what do we want from it?". Chemometrics and Intelligent Laboratory Systems. 30 (1): 109–115. doi:10.1016/0169-7439(95)00042-9.
2. ^ Kowalski, Bruce R. (1975). "Chemometrics: Views and Propositions". J. Chem. Inf. Comput. Sci. 15 (4): 201-203. doi:10.1021/ci60004a002.
3. ^ Cotrim, G. S.; Silva, D. M.; Graça, J. P.; Oliveira Junior, A.; Castro, C.; Zocolo, G. J.; Lannes, L. S.; Hoffmann-Campo, C. B. (2018). "Glycine max (L.) Merr. (Soybean) metabolome responses to potassium availability". Phytochemistry. 205: 113472. doi:10.1016/j.phytochem.2018.113472. ISSN 0031-9422. PMID 36270412. S2CID 253027906.
4. ^ Malinowski, E. R.; Howery, D. G. (1980). Factor Analysis in Chemistry. New York: Wiley. ISBN 978-0471058816. (other editions followed in 1989, 1991 and 2002).
5. ^ Sharaf, M. A.; Illman, D. L.; Kowalski, B. R., eds. (1986). Chemometrics. New York: Wiley. ISBN 978-0471831068.
6. ^ Massart, D. L.; Vandeginste, B. G. M.; Deming, S. M.; Michotte, Y.; Kaufman, L. (1988). Chemometrics: a textbook. Amsterdam: Elsevier. ISBN 978-0444426604.
7. ^ Martens, H.; Naes, T. (1989). Multivariate Calibration. New York: Wiley. ISBN 978-0471909798.
8. ^ Geladi, P.; Esbensen, K. (2005). "The Start and Early History of Chemometrics: Selected Interviews. Part 1". J. Chemometrics. 4 (5): 337–354. doi:10.1002/cem.1180040503. S2CID 120490459.
9. ^ Esbensen, K.; Geladi, P. (2005). "The Start and Early History of Chemometrics: Selected Interviews. Part 2". J. Chemometrics. 4 (6): 389–412. doi:10.1002/cem.1180040604. S2CID 221546473.
10. ^ Franke, J. (2002). "Inverse Least Squares and Classical Least Squares Methods for Quantitative Vibrational Spectroscopy". In Chalmers, John M (ed.). Handbook of Vibrational Spectroscopy. New York: Wiley. doi:10.1002/0470027320.s4603. ISBN 978-0471988472.
11. ^ Brown, C. D. (2004). "Discordance between Net Analyte Signal Theory and Practical Multivariate Calibration". Analytical Chemistry. 76 (15): 4364–4373. doi:10.1021/ac049953w. PMID 15283574.
12. ^ Krutchkoff, R. G. (1969). "Classical and inverse regression methods of calibration in extrapolation". Technometrics. 11 (3): 11–15. doi:10.1080/00401706.1969.10490714.
13. ^ Hunter, W. G. (1984). "Statistics and chemistry, and the linear calibration problem". In Kowalski, B. R. (ed.). Chemometrics: mathematics and statistics in chemistry. Boston: Riedel. ISBN 978-9027718464.
14. ^ Tellinghuisen, J. (2000). "Inverse vs. classical calibration for small data sets". Fresenius' J. Anal. Chem. 368 (6): 585–588. doi:10.1007/s002160000556. PMID 11228707. S2CID 21166415.
15. ^ Oliveri, Paolo (2017). "Class-modelling in food analytical chemistry: Development, sampling, optimisation and validation issues – A tutorial". Analytica Chimica Acta. 982: 9–19. doi:10.1016/j.aca.2017.05.013. hdl:11567/881059. PMID 28734370.
16. ^ Lawton, W. H.; Sylvestre, E. A. (1971). "Self Modeling Curve Resolution". Technometrics. 13 (3): 617–633. doi:10.1080/00401706.1971.10488823.
17. ^ Sylvestre, E. A.; Lawton, W. H.; Maggio, M. S. (1974). "Curve Resolution Using a Postulated Chemical Reaction". Technometrics. 16 (3): 353–368. doi:10.1080/00401706.1974.10489204.
18. ^ de Juan, A.; Tauler, R. (2003). "Chemometrics Applied to Unravel Multicomponent Processes and Mixtures. Revisiting Latest Trends in Multivariate Resolution". Analytica Chimica Acta. 500 (1–2): 195–210. doi:10.1016/S0003-2670(03)00724-4.
19. ^ de Juan, A.; Tauler, R. (2006). "Multivariate Curve Resolution (MCR) from 2000: Progress in Concepts and Applications". Critical Reviews in Analytical Chemistry. 36 (3–4): 163–176. doi:10.1080/10408340600970005. S2CID 95309963.
20. ^ Deming, S. N.; Morgan, S. L. (1987). Experimental design: a chemometric approach. Elsevier. ISBN 978-0444427342.

21. ^ Bruns, R. E.; Scarminio, I. S.; de Barros Neto, B. (2006). Statistical design – chemometrics. Amsterdam: Elsevier. ISBN 978-0444521811.
22. ^ Wentzell, P. D.; Brown, C. D. (2000). "Signal Processing in Analytical Chemistry". In Meyers, R. A. (ed.). Encyclopedia of Analytical Chemistry. Wiley. pp. 9764–9800.
23. ^ Oliveri, Paolo; Malegori, Cristina; Simonetti, Remo; Casale, Monica (2018). "The impact of signal pre-processing on the final interpretation of analytical outcomes – A tutorial". Analytica Chimica Acta. 1058: 9–17. doi:10.1016/j.aca.2018.10.055. PMID 30851858. S2CID 73727614.
24. ^ Olivieri, A. C.; Faber, N. M.; Ferre, J.; Boque, R.; Kalivas, J. H.; Mark, H. (2006). "Guidelines for calibration in analytical chemistry Part 3. Uncertainty estimation and figures of merit for multivariate calibration". Pure and Applied Chemistry. 78 (3): 633–650. doi:10.1351/pac200678030633. S2CID 50546210.
25. ^ Illman, D. L.; Callis, J. B.; Kowalski, B. R. (1986). "Process Analytical Chemistry: a new paradigm for analytical chemists". American Laboratory. 18: 8–10.
26. ^ MacGregor, J. F.; Kourti, T. (1995). "Statistical control of multivariate processes". Control Engineering Practice. 3 (3): 403–414. doi:10.1016/0967-0661(95)00014-L.
27. ^ Martin, E. B.; Morris, A. J. (1996). "An overview of multivariate statistical process control in continuous and batch process performance monitoring". Transactions of the Institute of Measurement & Control. 18 (1): 51–60. doi:10.1177/014233129601800107. S2CID 120516715.
28. ^ Hirschfeld, T.; Callis, J. B.; Kowalski, B. R. (1984). "Chemical sensing in process analysis". Science. 226 (4672): 312–318. Bibcode:1984Sci...226..312H. doi:10.1126/science.226.4672.312. PMID 17749872. S2CID 38093353.
29. ^ Smilde, A. K.; Bro, R.; Geladi, P. (2004). Multi-way analysis with applications in the chemical sciences. Wiley.
30. ^ Bro, R.; Workman, J. J.; Mobley, P. R.; Kowalski, B. R. (1997). "Overview of chemometrics applied to spectroscopy: 1985–95, Part 3—Multiway analysis". Applied Spectroscopy Reviews. 32 (3): 237–261. Bibcode:1997ApSRv..32..237B. doi:10.1080/05704929708003315