Chemometrics

Chemometrics is the science of extracting information from chemical systems by data-driven means. 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.

Contents

Introduction

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 very large and highly complex, involving hundreds to thousands of variables, and hundreds to thousands of cases or observations.

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.

Origins

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.

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), and partial least-squares (PLS). 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*,* Anal. Chim. 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,[2] *Sharaf, Illman and Kowalski's* Chemometrics,[3] *Massart et al.* Chemometrics: a textbook,[4] *and* Multivariate Calibration *by Martens and Naes.*[5]

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.[6][7]

Techniques

Multivariate calibration

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.[5][8] 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).[9] 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,[10][11][12] 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.

Classification, pattern recognition, clustering

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. [13] 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.

Multivariate curve resolution

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.[14][15] 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).[16][17]

Other techniques

Experimental design remains a core area of study in chemometrics and several monographs are specifically devoted to experimental design in chemical applications.[18][19] 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.[20] Signal pre-processing may affect the way in which outcomes of the final data processing can be interpreted. [21]

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.[22] 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.[23][24][25] 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,[26] or the newer term process analytical technology continues to draw heavily on chemometric methods and MSPC.

Multiway methods are heavily used in chemometric applications.[27][28] 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.

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 (https://doi.org/10.1016%2F0169-7439%2895%2900042-9).