Journal of Analytical Chemistry, Vol. 60, No. 10, 2005, pp. 994–996. Translated from Zhurnal Analiticheskoi Khimii, Vol. 60, No. 10, 2005, pp. 1115–1118.

CRITIQUE AND BIBLIOGRAPHY

Brereton, R.G., *Chemometrics: Data Analysis for the Laboratory and Chemical Plant*, Chichester: Wiley, 2003, 489 pp.

The book considers present-day approaches to treating experimental data in analytical chemistry, primarily in spectrometry and chromatography, in sufficient detail. The author is guided by his many years of experience in teaching chemometrics to undergraduate and graduate students of the School of Chemistry at the University of Bristol (Great Britain). Professor Brereton is group supervisor at the Bristol Centre for Chemometrics. He has abundant experience in organizing short-term intensive courses for chemical engineers, and this fact has left its mark on the style of the monograph.

Chemometrics covers a diversity of methods and approaches that can hardly be described in one book. The readers are introduced to the basic ideas of chemometrics. Having a thorough grasp of the basic ideas, after solving certain work problems to understand these ideas in practice, the reader can go deeper into the processing of the data of multifactorial experiments. The book is written so that, depending on his or her own needs and background, the reader can either learn the material step-by-step or read only selected chapters. Much attention is given to the analysis of particular examples. Each chapter is supplemented with a set of problems, which illustrate the material presented and are aimed at developing skills used in mathematical processing of experimental data. It is natural that, to master the subject, one has to make certain efforts. However, this task is not very difficult, because all of the necessary basic ideas of mathematical statistics, matrix algebra, and the basic principles of software operation are considered in the Appendix. As noted above, the examples cover mainly spectrometry and chromatography; the data generated by these methods bear important information about the test sample. However, this information cannot be extracted without using chemometrics methods. Nevertheless, the approaches to solving analytical problems described in Brereton's book are universal and not restricted to particular instrumental methods.

The book consists of five is chapters and an Appendix, in addition to an Introduction. It is devoted to the use of chemometrics in chemistry, first of all, analytical chemistry, and to its relations to other branches of science, such as statistics and computational mathematics. The Introduction also describes the main mathematical software used for analyzing and processing experimental data. A short literature review on the main divisions of chemometrics will assist the reader in deepening his/her knowledge in this field. The first chapter is devoted experiment design. In the author's opinion, this is the first step of experimental work, in particular, in cases when experiments are expensive or the experimental data will be used for mathematical simulation. Different scenarios of designing an experiment are considered, from complete and fractional factorial experiments to central composite design and simplex optimization. This branch has been thoroughly studied and described in numerous papers by Russian researchers as well as in foreign monographs and handbooks translated into Russian. However, from the methodological viewpoint, this chapter seems necessary in the book under review.

The second chapter is devoted to the processing of sequential signals. The material of this chapter is also traditional and familiar to Russian scientists. The necessity of processing sequential signals arises when one works with instruments interfaced with computers. This is done, first of all, in infrared (IR) spectrometry, including near-infrared spectrometry; nuclear magnetic resonance (NMR); and high-performance liquid chromatography (HPLC). Chromatograms and spectra usually consist of sets of peaks or lines superimposing a baseline and noise. Therefore, the author focuses special attention on two main problems. The first is the description of the shape of peaks and their characterization with standard parameters, such as the position of the center, width, and area. The second problem is the separation of the useful signal from the background and noise. This is done with different digital filters, such as moving average, the Savitsky–Golay method, and filtering using Fourier transformation. These methods are described in sufficient detail; at the end of the chapter, the reader is given some problems to solve, involving the processing of different signals. For example, there are problems to calculate the parameters of a Gaussian peak, to smooth the results of measurements in UV-VIS spectrometry using the Savitsky–Golay method, and to process an NMR spectrum using Fourier transformation. More complex methods, such as Kalman filtering, wavelet analysis, and the entropy maximum method, are also briefly described in this chapter to give the reader a more comprehensive idea of methods for signal processing.

1061-9348/05/6010-0994 © 2005 Pleiades Publishing, Inc.

The description of the basic methods of present-day chemometrics begins in the third chapter. Pattern recognition belongs to the main problems that added popularity to chemometrics in the early years of its development. Many methods of pattern recognition are based on the Fisher method of principal component analysis (PCA). The central concept of this method is the idea of the principal component. Principal components point to hidden correlations typical of the initial data set. They can be ranked in order of the longest distance between the measured objects or, which is the same, in descending order of the corresponding eigenvalue. The presentation of the experimental data in a new space of principal components opens up wide possibilities for the visualization of both the experimental data and their mutual arrangement. The corresponding methods are named projection methods, because data presentation in the principal component space is essentially the projection of the initial data onto a subspace of a lower dimension. The projection approach significantly differs from the traditional one, when an experimentalist first selects a small number of experimental factors, for example, one or two spectral lines, and then analyses the relationship between the input and output signals. In the book under the review, different aspects of principal components analysis are considered using practical examples. The book also describes what types of useful information can be obtained by properly applying this approach. The main versions of principal components analysis are demonstrated for two problems. The first is analyzing the data of HPLC with a diode array detector (HPLC–DAD). In this case, principal components analysis was used to resolve the overlapping peaks. The second problem is the assessment of the efficiency (productivity) of eight industrial chromatography columns. Eight substances were passed through each column, and each chromatographic peak was characterized by four standard parameters. In this case, principal components analysis was used for revealing the similarity in the work of individual columns. Other methods of pattern recognition and discrimination are considered in the book. Among these methods are cluster analysis, factor analysis, and linear discrimination analysis. All these methods are intended for work with large data sets represented as matrices. The rows of these matrices are observations, for example, UV-VIS spectra recorded at certain retention times in HPCL–DAD; the columns of the matrix correspond to variables, spectral wavelength in our case.

In the last decade, the wide introduction of analytical methods similar to HPLC–DAD to gas and liquid chromatography–mass spectrometry (GC–MS, LC–MS) has stimulated interest in processing even more complex data sets, including three-dimensional ones. The initial data in this case are represented as a parallelepiped rather than a table. These data sets are called three-modal; they arise, for example, in monitoring chemical reactions in time using HPLC–DAD. Working with three-modal data involves processes using more complex methods, for example, parallel factor analysis (PARAFAC) or Tucker algorithms. However, these methods are also based on principal components analysis.

The fourth chapter is devoted to calibration problems. Multivariate calibration problems occupy a special place in chemometrics. Their main goal is to simultaneously study two data sets and describe a correlation between them, that is, to construct a regression model. One of the data sets is usually a set of measurements, for example, a spectrum or a chromatogram; the second set consists of data on one or several analytical properties, for example, concentrations of components in a mixture. Calibration is rarely used to describe a correlation between two data sets; more often, it is used for predicting, for example, unknown concentrations based on the data of newly obtained spectral measurements. A number of problems in chemistry can successfully be solved using multivariate calibration methods. The simplest problem is the determination of individual substance concentration from spectroscopic data; this is done using an entire spectral region rather than one or two spectral lines. A more complex problem is the determination of the concentrations of several components in the mixture; spectra of individual components can be either known or unknown. An even more complex problem arises when an integral property characterizing the quality of the test object must be determined instead of the concentration of one substance or another. Many difficulties arise in constructing calibration models. One of the main difficulties is the variability of the background signal (baseline) due to the instability of instrument operation (instrumental drift) or external factors, for example, scattering in the near-infrared reflection spectrum. The problem is considered using an analysis of a mixture of ten polynuclear aromatic hydrocarbons (PAHs) as an example. The input data (predictors) are electronic absorption spectra in the region 220–350 nm digitized in 5-nm steps. The output data (responses) are PAH concentrations in 25 samples. The problem is to construct a calibration model and then predict concentrations of individual PAHs from a single spectrum of a mixture. The beginning of the fourth chapter is devoted to direct and inverse calibration problems; these problems are solved using traditional regression analysis for both one variable and multiple regression. Then, a detailed description of principle component regression (PCR) and projection to latent structures (PLS) is given. The PCR method has been considered in sufficient detail in Russian publications and used for a long time. The PLS method is less known in Russia. In this method, principal components are selected differently than in PCR, so as to maximize the covariation between the input and output parameters. Since both methods occupy a special place in chemometrics, the book comprehensively describes all steps of their use, such as the selection of the number of principal components, estimation of the accuracy of the constructed model, and assessment of the predictive

properties of the model. Of the thirteen examples given at the end of the fourth chapter, the reader has the opportunity to solve a problem of monitoring a chemical reaction using flow-injection analysis. The most complicated practical task is the construction of a calibration model for a three-modal data set resulting from an HPCL–DAD analysis of ten samples. One should determine the concentration of 3-hydroxypyridine in 2-hydroxypyridine.

The fifth chapter is devoted to one more important application of chemometrics, the processing and analysis of evolutionary signals, that is, signals depending on time. Such data appear more and more often and usually include spectral measurements upon the variation of a system parameter, for example, time or pH. The spectrum evolves as a result of parameter variation. In present-day laboratories, such data are obtained, for example, using hybrid methods of analysis, such as HPLC–DAD, GC–MS, LC–MS, or LC–NMR. Multivariate methods of data analysis described in the preceding chapters can be used for processing similar data; however, these methods do not take into account the order of data arrival, that is, the time evolution of the signal. From a practical viewpoint, one can formulate three main problems that can be solved for such data using the chemometrics approach. In ascending order of complexity, these problems are as follows:

(1) How many peaks are present in the given signal? Is it possible to recognize small impurities and detect embedded peaks?

(2) Is it possible to characterize peaks of individual components, what are their spectra, and is it possible to obtain a mass spectrum or an NMR spectrum in the case when one chromatographic peak entirely superimposes another peak?

(3) Is it possible to quantitatively determine the components of the mixture and to determine small impurities? Is it possible to use chromatographic analysis for monitoring the kinetics of a chemical reaction?

The processing of an evolutionary signal is usually started from preliminary data analysis. At this step, the background signal is subtracted or suppressed, data are scaled and preliminarily processed, the graphs constructed using principal components analysis are analyzed, and variables are selected for further investigation. At the second step, it is proposed to examine chromatograms and distinguish selective sections, that is, sections corresponding to individual mixture components, in the chromatogram. Many methods are used to search for and describe these sections. Principle components analysis is often used, but different versions of factor analysis are also highly efficient. From the viewpoint of mathematics, these methods are based on the calculation of the eigenvalues of the data matrix. In this case, the number of nonzero eigenvalues (rank of matrix) characterizes the number of mixture components in the particular section. At the third step, the problem is deconvolution, that is, the separation of two-dimensional signals into constituents, in the ideal case into signals corresponding to individual mixture components. This is also done using projection methods. Among the most interesting examples are the problem of selecting essential variables in analyzing GC–MS data and the problem of resolving an IR spectrum of a ternary mixture of 1,2,3-trimethylbenzene, 1,3,5-trimethylbenzene, and toluene.

The Appendix describes the main rules of vector and matrix algebra, the basic notions of mathematical statistics, and the main algorithms of projection methods. The last two sections of the Appendix are devoted to Excel and Matlab software. The description of elementary reactions is given for beginners. The book is supplemented with electronic data available on the John Wiley & Sons website, where one can find special macros for work with projection methods in Excel and a corresponding Matlab code. The reader can also find there the initial conditions for all problems given in the book.

The book is intended for researchers engaged in treating and analyzing large data sets that are obtained in chemical analysis. The advantages of this book are its clarity and the description of the corresponding mathematical apparatus. Therefore, after reading this book, one can master not only the basic concepts of chemometrics, but also the corresponding software, and then use this knowledge for solving particular problems. The book can be used as a textbook for students in chemistry: its material is presented straightforwardly and clearly and is not overwhelmed with complex mathematics expressions. Examples at the end of each chapter can be used as laboratory exercises.

O. Ye. Rodionova