Mathematical Processing of Spectral Data in Analytical Chemistry

Mathematical Processing of Spectral Data in Analytical Chemistry:

A Guide to Error Analysis

By

Joseph Dubrovkin

Cambridge
Scholars
Publishing

Mathematical Processing of Spectral Data in Analytical Chemistry:
A Guide to Error Analysis

By Joseph Dubrovkin

This book first published 2018

Cambridge Scholars Publishing

Lady Stephenson Library, Newcastle upon Tyne, NE6 2PA, UK

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

ISBN (10): 1-5275-1152-9
ISBN (13): 978-1-5275-1152-1

For my wife at the golden wedding anniversary

TABLE OF CONTENTS

PART III: MULTIVARIATE CALIBRATION AND PREDICTION

PART IV: DECOMPOSITION OF OVERLAPPING PEAKS

ACKNOWLEDGEMENTS

The author is grateful to prof. V. I. Tomin and dr. D. V. Ushakou (Institute of Physics, Pomeranian University, Słupsk, Poland) for providing experimental data and discussion of the results obtained by separation of overlapping peaks in fluorescence spectra.

PREFACE

The technological revolution in electronics and computers which has taken place during the recent decades has radically changed the outlook for analytical laboratories. Robotic techniques have replaced test tubes used for sample preparation. Compact computerized instruments have allowed practitioners to perform rapid and precise analyses. Commercially available software packages based on highly efficient algorithms of mathematical statistics and signal processing have opened new horizons for data processing. Significant progress has been achieved in the field of computer-enhanced spectrochemical analysis, mostly due to the emergence of a new scientific discipline called chemometrics.

An international journal called "Chemometrics and Intelligent Laboratory Systems" defines chemometrics as "the chemical discipline that uses mathematical and statistical methods to design or select optimal procedures and experiments, and to provide maximum chemical information by analyzing chemical data."

Chemometrics can be presented as a bridge between the obtaining of raw data and the validation of analytical methods. The main metrological characteristics that must be considered during such validation include accuracy and precision. There is a large set of the international standards which define the standard practice for conducting the studies aimed at determining these characteristics.

Chemometrics originated in the 1970s as a means of solving the multivariate calibration problems of the quantitative analytical measurements. Today, this field includes a diverse set of computerized tools for preprocessing of analytical data and for its further use in the qualitative and quantitative analysis for extraction the information necessary for practical and theoretical applications. The ultimate goal of chemometrics is to estimate hidden values of the system parameters under investigation.

Any measurement of these parameters is to be accompanied by an estimate of the error. That is, a measurement uncertainty must necessarily characterize the data obtained by analytical measurements. The study of the error propagation in the data collection, processing and analysis is one of the main tasks of chemometrics.

In solving this problem, chemometrics uses cumbersome mathematical methods based on statistics and the signal processing theory. Unfortunately,

scientific literature on this issue had to be drawn from diverse sources (extracts of manuals, monographs, scattered articles, and tutorials).

It is somewhat tricky to independently study all the aspects of the error propagation problem since the documents mentioned above usually do not contain numerical calculations. Moreover, such calculations cannot be checked by a practitioner who has a basic knowledge of mathematics and statistics. These estimates have to be taken "on faith."

Open source software in statistics is a "black box" for non-professional programmers. Also, there is always a danger of a blind use of the problem-solving recipes which may lead to unpredictable results.

Apparently, the only scholar who has so far given some valuable examples of signal processing ready for practical use is Prof. T. O'Haver, author of "A Pragmatic Introduction to Signal Processing: with applications in scientific measurement," 2017. Available from: https://terpconnect.umd.edu/~toh/spectrum/ (Accessed February, 1st, 2018).

The goal of this guidebook was to provide the readers with a comprehensive presentation of the problems in the error analysis in analytical spectrometry. We did not intend to scrutinize the published data but only to briefly point out the primary studies in each particular case and to redirect those who are eager to investigate the relevant sources.

Theoretical discussions on this issue are illustrated by various examples supplied by a simple programme code on MATLAB which can be easily modified by non-professional users. The readers who may wish to study the problem further can validate numerical data given in the guide by using computer calculations. Thus, they will be able to understand the details of the algorithm and, if necessary, modify corresponding computer programs.

This book is intended for a broad range of readers including practitioners and researchers of industrial and university analytical laboratories as well as for students specializing in analytical spectroscopy and chemometrics.

ABOUT THE STRUCTURE OF THE BOOK

The main topics are organized into four parts divided up into brief chapters according to the thematic areas they cover. Each chapter is supplied with suitable references to make it more convenient for readers.

In the first part, we introduce the readers to the main problems posed in the book, so that they understand what awaits them while reading the following chapters. We give general characteristics of the noise in spectroscopic measurements, define the accuracy and the precision in analytical chemistry and formulate the main tasks of preprocessing, calibration and prediction. The second and the third parts review the widespread methods of the univariate and multivariate calibration and prediction and give the accepted mathematical expressions of the noise propagation. The last part dedicated to the analysis of the uncertainty in determining parameters of the overlapping peaks.

Each chapter has a large number of examples focused on the subject matter and is supplied with exercises based on the computer programs.

The book closes with appendices included supplementary materials that are necessary to facilitate the readers' ability to understand more deeply the theoretical problems discussed in the main text.

Reading requires the knowledge of the secondary school courses on the differential calculus, linear algebra, and statistics. To perform the exercises, the readers must have programming skills for beginners in MATLAB.

All programs are open sources which can be downloaded from in the project "Computer-Based Tutorial on Chemometrics"
https://www.researchgate.net/profile/Joseph_Dubrovkin.

A significant part of the book is based on the original research carried out by the author.

For simplicity, the captions of figures, tables, exercises, and examples have the following structure: "part.chapter-current number."

The author would be very grateful for the criticisms, comments and proposals about this book and hopes take them into account in his future work.

ABBREVIATIONS

AC-Analytical Signal
BLS-Bivariate Least Squares
CI-Confidence Interval
BWF-Breit-Wigner-Fano
CR-Constrained Regression
D2-Second-Order Derivative
DSP-Digital Signal Processing
EGN-Exponential-Gaussian Hybrid
FSD-Fourier Self Deconvolution
FT-Fourier Transform
FTIR-Fourier Transform Infrared Spectroscopy
FWHM-Full Peak Width at Half-Maximum
GDH-Generalized Discrete Harmonics
GEMG-Generalized EMG
GUI-Graphic User Interface
IF-Instrumental Function
IFT-Inverse FT
IID- independently and identically normally distributed
ILS-Inverse Least Squares
IHM-Indirect Hard Modelling
LC/MS-Liquid-Chromatography-Mass Spectrometry
LEL-Lower Error Limit
LHO-Leave Half Out
LM-Levenberg-Marquardt
LOO-Leave One Out
LS-Least Squares
MCR(-ALS)-Multivariate Curve Resolution(-Alternating LS)
MCRS-Mean Centring RS
MhD-Mahalanobis Distance
NAS-Net Analyte Signal
NLDM-Natural Logarithm Derivative Method
NLICF-Non-Linear Iterative Curve Fitting
OLS-Ordinary Least Squares
OP-Orthogonal Polynomial
PAT-Process Analytical Technology
PC-Principal Component
PCA-PC Analysis
PCR-PC Regression
PDF- Probability Density Function
PMG-Polynomial Modified Gauss
PVMG-Parabolic-Variance Modified Gauss
PSD-Pseudo-Deconvolution
RBias-Relative Bias
RMSEC-Root Mean Squared Error of Calibration
RMSEP-Root Mean Squared Error of Prediction
RGK- Regularized Gaussian kernel
RS-Ratio Spectra
RSTD- Relative Standard Deviation
SCV-Segment Cross-Validation
SEL-Standard Error of the Laboratory
SG-Savitzky-Golay
SP-Sparrow
SVD-Singular Value Decomposition
TR-Tikhonov Regularization
UI-Undetectable Increment
UP-Undetectable Perturbation
UR-Unconstrained Regression
WT-Wavelet Transform

PART I:

INTRODUCTION TO ERROR ANALYSIS

CHAPTER ONE

ANALYSIS OF NOISE IN SPECTRAL MEASUREMENTS

Measurements of any physical value are influenced by random and systematic errors. The systematic errors may arise from incorrect measurement (a typical example is an improper preparation of the blank cuvette in the spectrophotometer) and from the imperfection of the instrument (e.g., spurious reflection and radiation). Often the systematic errors can be decreased to be negligible by improving the apparatus and the measurement process (e.g., by correct calibration). However, the random errors cannot be eliminated in principle; they only can be decreased by improving the measurement procedure (e.g., by an expansion of the measurement scale) and by analog or numerical processing of obtained results (e.g., by smoothing). It is important to emphasize that reduction of the random errors can cause significant unpredictable distortions of the actual value of the quantity to be measured (systematic errors).

The random errors vary randomly with time. They are due both to the errors in analog-to-digital conversion of the measured value and to the impact of different external factors on the measurement process (e.g., a variation of the sample temperature, vibrations). In spectroscopy, the sources of the random errors are the random noises arising in different parts of the spectrometer, mainly, in the radiation detector. The origins of these noises are very different and can be approximately described in every case by the particular mathematical model based on the probability theory. The following classification was given in the series of the articles which were summarized by Mark and Workman [1].

The sources of noise

Detector-dependent noise

The thermal noise (IR, NIR spectrometers) is the noise in the thermal detectors. This noise does not depend on the intensity of the electromagnetic

radiation that falls on a sensor. In the time domain, this noise has a Gaussian (normal) intensity distribution (We hope that the readers remember this from the university course in probability and statistics). This noise is often called "white noise" because it has the constant power density of Fourier spectrum (Appendix A2) in the vast range (Fig. 1.1-1). The noise power is proportional to the width of an interval of Fourier frequencies (bandwidth) of the recording device.

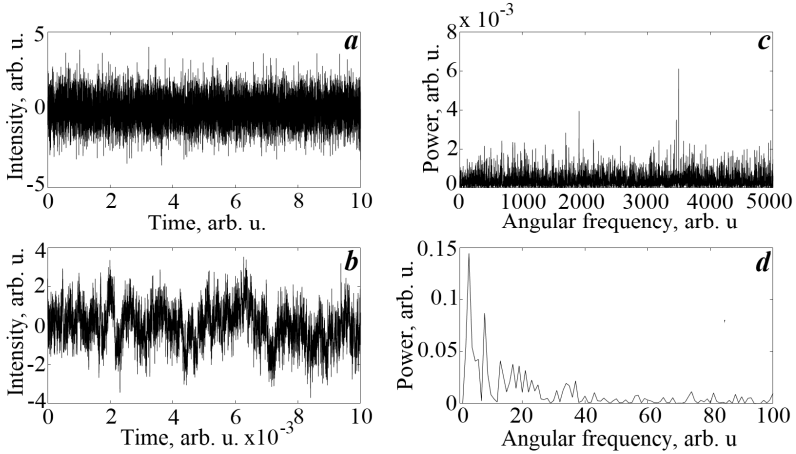

Fig. 1.1-1. White and pink noises in the time and the Fourier domain (panels a, b and c, d, respectively). The noise mean value and the standard deviation are zero and one, respectively.

The standard deviation of the absorbance (A) measurements distorted by a low-intensity thermal noise [2]:

$$\sigma_{Term} = \{k_1/ln(10)\}\sqrt{1 + 100^A}, \qquad (1.1-1)$$

where k_1 is a constant indicating precision of a spectrophotometer.

Plot $[\sigma_{Term}/A](A)$ (Fig. 1.1-3) shows that the absorbance region 0.2-1.2 is the most suitable for spectroscopic measurements since in this range the relative standard deviation is approximately constant.

The shot-noise (UV-VIS, X-ray, and gamma-ray spectrometry) in the photon counting detectors has a Poisson intensity distribution in the time domain. Its intensity increases with the square root of the signal. This noise is also white.

The standard deviation of the absorbance distorted by a low-intensity short-noise:

$$\sigma_{Shot} = \{k_2/(ln(10))\}\sqrt{1+10^A}, \qquad (1.1-2)$$

where k_2 is a constant similar to k_1 in Eq. (1.1-1).

Exercise 1.1

The readers are invited to represent noise data obtained by their instruments similar to Fig. 1.1-2.

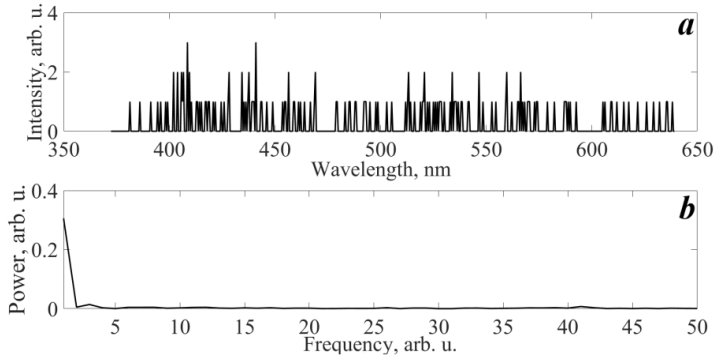

Fig. 1.1-2. Noise measured on Bruker Optics 2501S spectrograph by D. V. Ushakou (Pomeranian University in Slupsk, Poland) and the noise power spectrum (panels a and b, respectively).

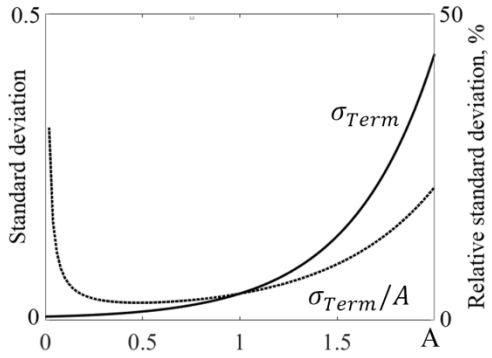

Fig. 1.1-3. Eq. (1.1-1). $k_1 = 0.01$.

Variations in energy fall on the detector due to vibrations of the source and the changing geometry of the radiation cause the flicker (pink) noise ($1/f$-noise). The noise intensity is proportional to the signal energy. The

noise power spectrum depends on frequency: $f^{-\alpha}$, where $\alpha \simeq 1$ (Fig. 1.1-1). If the noise is small then the standard deviation of absorbance σ_f is approximately constant.

Detector-independent noise

The noise sources include mechanical vibrations of the optical instruments and different kinds of instabilities:

- The variation of the pathlength in the absorption spectroscopy due to the changes in the sample position [2].
- The variability of the sample properties that cannot be measured. However, these properties influence on the measurement property (e.g., the changing of light reflectance in the transmittance measurements, inhomogeneous of the sample, the artefacts of the blood motion in the blood analysis using Functional NIR Spectroscopy [3, 4]).
- Random drifts of optical and electronic devices (the flicker noise) caused by slow changes of their parameters due to the temperature variations and other factors.

The mathematical analysis of these sources can be performed only in each case using appropriate assumptions which simplify the solution of the problem.

Computer modelling of correlated noise

Let us consider the noise intensity distribution in the time domain [5]. As was pointed out above, Gaussian (normal) and Poisson noise distributions are usually used to model noise models in spectroscopic measurements. In the time domain, noise is characterized mathematically by the error covariance matrix $\boldsymbol{COV}$ which diagonal elements

$$COV_{ii} = \sigma_i^2 \qquad (1.1-3)$$

represent the noise variances (dispersions) in the i^{th} point. If no two points of the noise in the time domain are correlated with each other, the non-diagonal elements $COV_{ij} = 0$. If all variances (Eq. (1.1-3)) are the same, the noise is called *homoscedastic*, otherwise-*heteroscedastic*. The source of the heteroscedastic noise, in particular, is a randomly varying baseline (background) [6, 7], due to both instrumental and physical-chemical factors. Total baseline compensation is practically impossible.

Suppose that the baseline is approximated by the second-order polynomial which coefficients are random numbers. This baseline is added to some spectrum. The procedure is repeated r times. Then the intensity of the *t*-spectrum the point *i*:

$$y_{it} = y_i + \eta_i + a_{0t} + a_{1t}i + a_{2t}i^2 , \qquad (1.1-4)$$

where y_i is the undistorted value, η_i is the normal noise with zero mean and dispersion σ_y^2, a_{qt} is the q coefficient of Eq. (1.1-4) which is constant for a given spectrum, but it changes randomly for each spectrum. The estimation of the covariance matrix element [8]:

$$COV(y_k, y_l) = [1/(r-1)] \sum_{t=1}^{r} (y_{kt} - \bar{y}_k)(y_{lt} - \bar{y}_l), \qquad (1.1-5)$$

where the bar is the average symbol. According to Eq. (1.1-4):

$$\bar{y}_i = y_i + \bar{a}_0 + \overline{(a_1)}i + \overline{(a_2)}i^2 \qquad (1.1-6)$$

Substituting Eqs. (1.1-4) and (1.1-6) into Eq. (1.1-5), we have

$$COV(y_k, y_l) = \sigma_y^2 \xi_{kl} + BS(y_k, y_l), \qquad (1.1-7)$$

where $BS(y_k, y_l) = [1/(r-1)] \sum_{t=1}^{r} \left(\delta_{a_{0t}} + \delta_{a_{1t}}k + \delta_{a_{2t}}k^2\right)\left(\delta_{a_{0t}} + \delta_{a_{1t}}l + \delta_{a_{2t}}l^2\right)$,

$\xi_{kl} = \begin{Bmatrix} 1, k = l \\ 0, k \neq l \end{Bmatrix}$ is the Kronecker symbol, $\delta_{a_{qt}} = a_{qt} - \bar{a}_q$.

Suppose that $\delta_{a_{qt}}$ is a normal random variable with zero mean and covariance matrix $\sigma_{a_q}^2 \boldsymbol{I}$ ($\boldsymbol{I}$ is the identity matrix). Then, by neglecting the small contributions to the sum of the cross members for sufficiently large r, we obtain from Eq. (1.1-7):

$$COV(y_k, y_l) = \sigma_y^2 \xi_{kl} + \sigma_{a_0}^2 + \sigma_{a_1}^2 kl + \sigma_{a_2}^2 k^2 l^2 . \qquad (1.1-8)$$

Correlations between analytical points may also be due to the pre-processing, e.g., digital smoothing [6]. Since a digital filter is usually much shorter than a spectrum, the covariance matrix will be a sparse matrix (populated primarily with zeros) and, therefore, singular (Appendix D2). If the matrix is near to singular, a correctly computed inverse is impossible. We found that stabilization of the inverting process with Tikhonov regularization (Appendix D3) is not useful. Therefore, in further, we will not consider inverting of this correlation matrix.

CHAPTER TWO

PREPROCESSING OF SPECTRAL DATA

Data preprocessing is one of the most critical steps for any data analysis problem in signal processing [1]. Data preprocessing involves data cleansing, analysis of missing values and data transformations for further modelling and extracting necessary information. Preprocessing of spectrochemical data includes the following steps:

- Smoothing of noisy spectra
- Baseline removal
- Decomposition of overlapping peaks
- Data transformation and compression

These steps can be combined with each other; for example, the first three items can be carried out together.

The historically first preprocessing methods were digital smoothing and differentiation of spectral data introducing in analytical chemistry by Savitzky and Golay (SG) [2] (Appendix A1.1). The primary goal of these methods was made the spectrum plot more visible for its interpretation by denoising and improving spectral resolution using the even-order derivatives of the graph [3]. These derivatives are very similar to raw spectra but significantly narrower, and they flatten the polynomial baseline. The exceptional success of using *SG* filters in 1960-1980 is explained by the simplicity of the algorithm design, possibility of the manual calculations and by the fact that the interpretation of the transformed and non-transformed spectra is similar. Unfortunately, the error analysis of the derivative spectrometry showed many drawbacks of this method [3-6]. As an example, consider Gaussian triplet (Fig. 1.2-1):

$$A(x) = \sum_{j=1}^{3} R_j \exp\left\{-4ln2\left[\left(x - x_{0j}\right)/r_j\right]^2\right\}, \qquad (1.2-1)$$

where parameters of each component $R, x_0,$ and r are the intensity, the position of the maximum and the width of the peak, respectively. The last two parameters are defined in the dimensionless abscissa x. The peak parameters: $[x_0, R, r]$= [-2, 1, 2; 0, 2, 2; 1.3, 0.5, 0.5];

Fig. 1.2-1 demonstrates:

- The shift of the peak maxima of the first peak observed in the second-order derivative from its correct position (The left-side shift is shown by the shift of the arrows).
- The erroneous structure appears due to the satellites (negative peaks in the Fig.).
- The wrong relative intensity in the derivative spectra of the non-equal peak widths (compare the third peaks in the spectrum and its second-order derivative).

Also, it is well-known that noise in the derivative spectrum significantly increases; therefore, a smoothing is needed.

These items will be discussed below in Chapters 4.6 and 4.7.

Exercise 1.2

The readers are invited to calculate a set of triplets by varying the peak parameters, the noise intensity, the step of x and the number of smoothing. You will be able to detect that the optimal differentiation is a very cumbersome task!

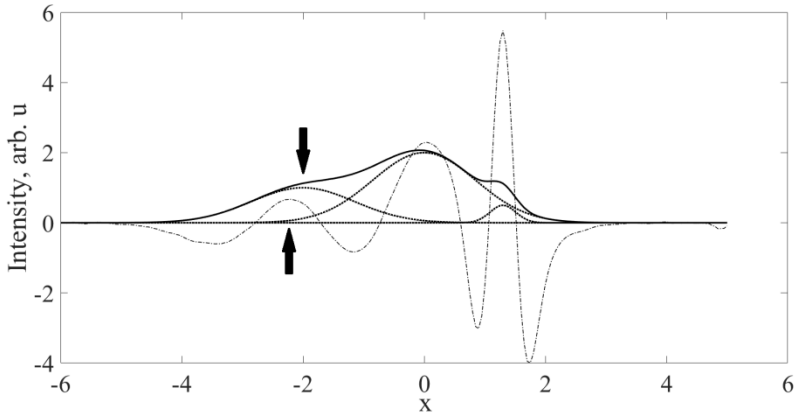

Fig. 1.2-1. Gaussian triplet, its components and the negative second-order derivative of the triplet (solid, dotted and dashed curves, respectively).

Parallel to SG filters, a variety of the smart preprocessing methods has been developed at the end of the past century [7]. Among these, the mathematically rigorous spectral decomposition (deconvolution)

algorithms are of the most importance [6]. This is because they are based on:

- Mathematical models of the measured spectrum which were predetermined theoretically and experimentally proven.
- Non-linear curve fitting procedures which are strict from a statistical point of view [8].

These issues will be discussed in Chapter 4.1.

Digital data filtration is one of the families of the *Linear Data Transformation methods* widely used for data compression [9]. For this goal, the conventional coordinate system of spectra (e.g., absorbance versus wavelength) is transformed to another coordinate frame. In this frame, Y and X axes represent the generalized Fourier series expansion coefficients of the spectrum (generalized discrete harmonics - *GDH*) [10] and their consecutive numbers, respectively.

Transformed data were used for qualitative and quantitative purposes without reconstruction [11]. The examples of *GDH* are Fourier, Walsh, and wavelet harmonics [12], statistical moments [13] and coefficients in the series expansions using classical orthogonal polynomials [14]. According to the information theory, the *GDH* method involves some loss of information. It cannot decrease the errors of unbiased estimates in determining mixture concentrations, but high compression ratios can be obtained by ignoring some low informative harmonics [11].

The *GDH* method belongs to the general group of the linear transformations of analytical signals. This group also includes the ordinate transformations, while the abscissa is not changed; e.g., derivative spectra [3] and the *Net Analyte Signals* [15].

CHAPTER THREE

CALIBRATION AND PREDICTION IN ANALYTICAL SPECTROMETRY

Calibration is one of the significant methodological problems in the spectrochemical analysis [1-10]. The multivariate calibration methods which have laid the foundation of chemometrics are wide-spread in modern analytical chemistry for laboratory studies and *Process Analytical Technology*.

There are various definitions of calibration which dependent on science [1]. In analytical chemistry, calibration is a process of setting relations between physical-chemical properties of the sample (e.g., the analyte concentration) and the analytical signal (e.g., spectral absorbance).

Danzer [1] distinguished between three types of calibration: absolute, definitive and experimental. He pointed out that the following conditions of correct calibration must be fulfilled:

- reliable and traceable standards,
- fixed calibration parameters,
- proper mathematical treatment of an appropriate calibration model.

These issues have been carefully discussed and are supplied with numerous references [1]. They are beyond the scope of this book.

The most straightforward procedure for measuring the content of a single mixture component (analyte) using one analytical point (wavelength) in the measured spectrum is called *Univariate Calibration* (Fig. 1.3-1a). It uses the direct calibration model: estimation of the relation of the measured value (absorbance, A) to the content (concentration, c) of a reference sample using predefined parameters of the linear model [6]:

$$A = k_1 c + k_0 + \epsilon, \qquad (1.3-1)$$

where the slope k_1 is the absorptivity constant, k_0 is the intercept, ϵ is a residual due to the imperfections in the model and to the random variations of the absorbance. The constants k_0 and k_1 are estimated using reference samples containing different amounts of the analyte under study. The

concentration of the unknown sample is calculated by substituting the measured absorbance into:

$$c = (A - k_0)/\, k_1 \qquad (1.3-2)$$

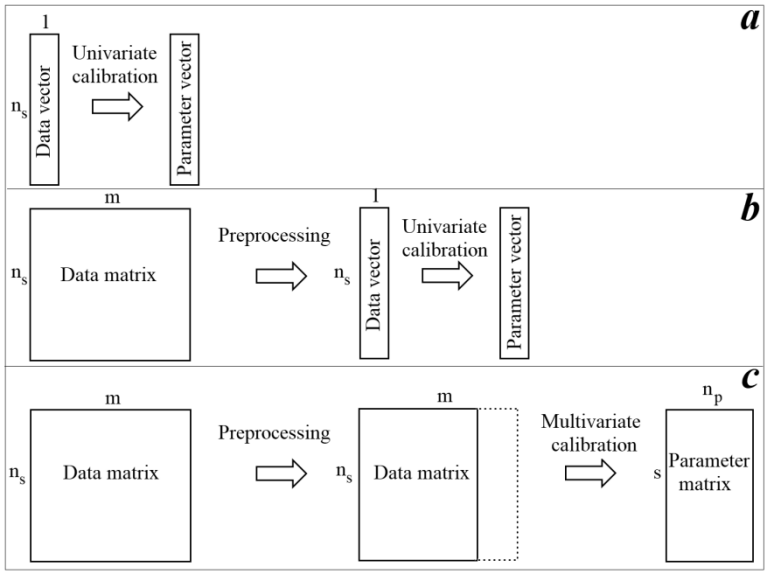

Fig. 1.3-1. Block diagrams of univariate (a, b) and multivariate (c) calibration. n_s, m and n_p are the number of samples, analytical points and the sample parameters, respectively.

To improve the metrological properties of the calibration procedure, the calibration matrix (set) is obtained by collecting spectra of different reference samples measured in several analytical points (Fig. 1.3-1b).

As noted in IUPAC Technical Report [2], for univariate calibration the spectrum must be "highly selective for the analyte of interest." In other words, there should not be interference from other components. Many technical procedures have been developed to eliminate the contribution of the interfering constituents including the linear transformation methods described in the previous section [5]. However, these methods are not suitable in many cases, especially, for the samples which contain complex mixtures of many different chemicals.

Now *the multivariate calibration methods* (Fig. 1.3-1c) have taken the dominant position in chemometrics. They have become a standard means

of scientific research [6-10]. The multivariate calibration is mostly based on the inverse calibration model:

$$\boldsymbol{C} = \boldsymbol{AB} + \boldsymbol{\epsilon}, \qquad (1.3-3)$$

where $\boldsymbol{C}$ is the matrix of the concentrations of the analytes in the calibration mixtures which spectra are in the data matrix $\boldsymbol{A}$, $\boldsymbol{B}$ is the regression matrix, $\boldsymbol{\epsilon}$ is the matrix of residuals. For predefined matrices $\boldsymbol{C}$ and $\boldsymbol{A}$, matrix $\boldsymbol{B}$ is estimated by solving Eq. (1.3-3). However, in practice, the solution is made complicated by the fact that the matrix $\boldsymbol{A}$ is usually poorly conditioned (Appendix D2) since the spectra of different mixtures may be very similar to each other [6]. To overcome this, so called, the collinearity problem [11], smart mathematical methods (the *Principal Component Regression*, the *Partial Least Squares Regression*, and *Tikhonov regularization*) were used [6-9, 12]. The review [13] discusses various modifications of the regularization procedures in chemometrics.

Here we first mentioned the term *'Regression'* which needs clarification since there is often confusion between the concepts of *'Calibration*' and *'Regression*.'

Consider a simple example [14]. Suppose that there is some relationship (F) between the response (dependent) variable (y) and other independent variables $x_1, x_2, \dots, x_t$ (called explanatory or regressor variables):

$$y \approx F(x_1, x_2, \dots, x_t). \qquad (1.3-4)$$

Some of regressors may have fixed values, others-random (measured) values. Also, we suppose that the mathematical expression (linear or nonlinear) of the function F has a *priory* predefined, based on the theoretical and experimental data. In other words, the form and the parameters of the model F are known. Then the goal of regression is to estimate the model parameters provided the best-fit of $F(x_1, x_2, \dots, x_t)$ to the response y. E.g., according to the *Least Squares Method* (*LS*), the best-fit model provided a minimum:

$$min\{\sum_{i=1}^{n}(y_i - F_i)^2\}, \qquad (1.3-5)$$

where $i = 1, 2, \dots, n$ is a number of a measured sample.

If "we have little or no idea about the form of the relationship" [14] (Eq. (1.3-4)) then it is necessary to select a family of the models and find the most appropriate model. Unfortunately, mathematics in principle cannot give an unambiguous solution without involving additional physicochemical information.

We conclude that *calibration* set relations between the physical-chemical properties of the sample and the analytical signal using a statistical (mathematical) method called *regression*. In principle, the calibration may be performed without a rigorous mathematical tool, e.g., using the plot: response over a property.

The preprocessing step is also wide-spread in the multivariate calibration [6-10]. This step is called "*Multivariate signal processing*" [4] similar common term "*signal processing*" being used in electronics and communication. Often the preprocessing step includes data compression.

There is a large number of software packets designed to solve calibration problems. Among them, the most popular are MATLAB [15] and UNSCRAMBLER [16].

As we pointed above, the calibration process is closely related to the choice of the optimal model involving preparation of the calibration (training) set, selection of the optimal analytical points and preprocessing procedures for a predefined mathematical method of calibration. These very complex problems are not reviewed in this book. The interested reader is encouraged to refer to the bibliographical sources [6-10].

The calibration process also closely related to the *prediction ability* of the developed model, that is, its validation [6]. The monograph [6] describes theoretical and practical aspects of the estimation of prediction errors in the linear regression models.

The rigorous discussion of the theoretical concepts of the calibration and prediction requires from the reader a solid mathematical background and extensive training in signal processing and computers. Consequently, in what follows, we will try to simplify the presentation of the material taken from the books [6-9] and IUPAC documentation [2, 3, 10], but not emasculating its essence. Without loss of generality, for convenience, we will consider the spectrophotometric absorption methods and use the appropriate terminology.

CHAPTER FOUR

ACCURACY AND PRECISION IN ANALYTICAL SPECTROMETRY

Before analysts start using a newly developed analytical method, they should validate it. The goal of the validation is to confirm that the analytical method is suitable for its intended use, that is, its metrological characteristics indicate that the method is reliable and consistent [1]. In other words, measurement results must be meet the tolerance limits of the material under study. However, there are always consumer and producer risks of the false decisions [2, 3]. The mathematical aspects of the false accept risk were carefully discussed by Castrup [2] which gave common guidelines for managing in-tolerance compliance decisions. Kuselman et al. [3] studied this issue in analytical chemistry. Generally speaking, the measured analyte content (the measurand) must be treated as a random variable and expressed in terms of a probability density function (*PDF*). The probability of a random variable to fall within a particular range is equal to the area under *PDF* and is mathematically given by the integral of its *PDF* over that range.

The *PDF*

"combines prior knowledge of the measurand and new information acquired during the chemical analysis/ measurement/testing" [3].

Since the population of the measurand is unknown, one can only estimate the population parameters such as mean and standard deviation using a set of random samples. These estimates (the best guesses) are called "point estimates" [4]. It is essential to underline that these estimates are not deterministic but have probabilistic properties.

The estimator is called unbiased if there is no difference between an estimator's expected value and the true value of the estimated parameter. For example, if the mean analyte concentration obtained in repeated measurements approaches to the correct value of the measurand while the number of the repetitions increases, then this mean concentration is unbiased.

To determine how accurate the point estimate of the population mean,