Data Compression in Spectroscopy

Data Compression in Spectroscopy

By

Joseph Dubrovkin

Cambridge
Scholars
Publishing

Data Compression in Spectroscopy

By Joseph Dubrovkin

This book first published 2022

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): 1-5275-8620-0
ISBN (13): 978-1-5275-8620-8

TABLE OF CONTENTS

PART III: LINEAR TRANSFORMS-BASED ONE DIMENSIONAL ANALYTICAL SPECTROMETRY

PART IV: LINEAR TRANSFORMS-BASED MULTIDIMENSIONAL ANALYTICAL SPECTROMETRY

PREFACE

In recent decades, the huge library cabinets filled with books and magazines have been successfully replaced by autonomous and cloud-based electronic digital data storage. High-speed information transmission channels and high-speed computers have made it possible to transfer, accumulate, and process unattainable arrays of numbers, symbols and pictures. Suffice it to mention the Google search engine, which in the middle of the last century could be the subject of fantastic fiction.

An inexperienced reader might think that the competition between the amount of information accumulated by mankind and the ability to efficiently archive it is won by the latest technical data storage facilities. Unfortunately, this statement is incorrect; the information ocean is threatened with a new flood but in the intellectual sphere. Where is the exit? To separate the wheat from the chaff [Matthew 13:24-30]? But who can decide what information should be filtered out and what information should be kept? A reasonable solution to this problem is to compress digital data by mathematical processing.

Data compression is based on reducing their redundancy, in other words, attenuating information noise (not relevant information) by compromising the fidelity of the useful signal and the level of residual noise. For example, a compressed image requires much less memory to store it than the original, and the deterioration of its quality is not noticeable to the naked eye.

For the first time, spectroscopic data, compressed by linear transforms (Fourier, Walsh-Hadamard) has been used to improve the efficiency of quantitative analysis [1] and to simplify the multivariate calibration in NIR spectroscopy [2 and ref. within]. The singular value decomposition of matrix data and use of the most informative principal components is the basic idea of chemometrics [3, 4]. In all these methods, the transformed data was not restored to the standard coordinate system. The main idea of compressed sensing spectral imaging is to perform optical transforms during data acquisition [5].

Now, compression of spectroscopy data is becoming an actual task since this processing allows effective storing and transferring huge datasets of multicomponent mixtures. The reconstructed matrix was successfully used for multivariate calibration and segment cross-validation, clearly demonstrating the potential of the proposed method for future applications to chemometrics-enhanced spectrometric analysis with limited options of memory size and data transfer rate [6]. Also, it was demonstrated that compression is a perspective method for the economical storage of spectral databases.

The present study is an attempt to give a detailed explanation of the reached theoretical and numerical results of data compression in spectroscopy. Since the reconstructed signal contains less noise than the original one, denoising is considered in parallel with compression.

While preparing the book, the author faced serious problems associated with its versatility, including the theory and technique of processing one and multidimensional signals, a description of optical devices, and applied physical-chemical tasks of atomic and molecular spectroscopy. The presentation of this material required the application of rather complex mathematical methods, usually little familiar to specialists in analytical spectroscopy.

The goal (as in our previous books [7, 8]) was to avoid, where possible, readers' blind faith in the validity of conclusions and recommendations.

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

References

1. Dubrovkin, J. (1985). Theory and use of the method of linear transformation of spectral coordinates in physicochemical studies. Izvestia Severo-Kavkazskogo Nauchnogo Centra Vysšey Školy, Yestestvennye Nauki, 2, 51-57 [Russian].

2. Dubrovkin, J. (2017). Linear transformations of multivariate calibration models in near infrared spectroscopy: A Comparative Study. Journal of Near Infrared Spectroscopy, 25, 223-230.

3. Workman Jr., J. J., Mobley, P. R., Kowalski, B. R., Bro, R. (1996). Review of Chemometrics Applied to Spectroscopy: 1985-95, Part 1. Applied Spectroscopy Reviews, 31, 73-124.

4. Mobley, P. R., Kowalski, B. R., Workman Jr., J. J., Bro, R. (1996). Review of Chemometrics Applied to Spectroscopy: 1985-95, Part 2. Applied Spectroscopy Reviews, 31, 347-368.

5. Gamez, G. (2016). Compressed sensing in spectroscopy for chemical analysis. Journal of Analytical Atomic Spectrometry, 31, 2165-2174.

6. Dubrovkin, J. (2022). A novel compression method of spectral data matrix based on the low-rank approximation and the fast Fourier transform of the singular vectors. Applied Spectroscopy, 76, 3, 369-378.
7. Dubrovkin, J. *Mathematical processing of spectral data in analytical chemistry: A guide to error analysis*. Cambridge Scholars Publishing. 2018.
8. Dubrovkin, J. *Derivative Spectroscopy*. Cambridge Scholars Publishing. 2021.

About the Structure of the Book

The book is organized into four parts. In the first part, the properties of the typical spectroscopic signals are discussed together with the information-theoretic aspects of their linear transforms. The last paragraph introduces the reader to the Big Data approach to analytical spectroscopy.

We assume that the material given in this part and Appendix 'Spectroscopy' will help the reader refresh and replenish the knowledge gained in the relevant university courses.

The second part introduces the common approaches to data compression, including the entropy-based methods and some image denoising and compression algorithms. The last paragraphs describe spectral imaging and compressed sensing that are little known to the typical spectroscopic audience.

The subjects of the two following parts are Linear Transforms-Based One Dimensional and Multidimensional Analytical Spectrometry in the framework of Chemometrics. Fourier (including interferogram-based), wavelet, and Walsh-Hadamard transformed methods are discussed. Using of classical orthogonal polynomials and splines for compression of spectra is also considered.

Compression of multidimensional data involves correlation spectroscopy and instrumental compressive sensing in 2D spectroscopy.

The bibliography tables briefly describe hundreds of applications of the compression methods and related topics in the industrial and research laboratories.

We sincerely apologize to all those researchers whose outstanding works are not cited because the book does not have enough free space to include a complete bibliography. The project "Data Compression in Spectroscopy" (https://www. researchgate.net/profile/Joseph_Dubrovkin) provides a bibliographical supplement, updated as new information becomes available.

The author wanted to give each chapter status of a self-contained article whose reading is independent of other sections. This material presentation allows readers to avoid cramming a previous text before moving on to another topic. MATLAB-based examples, which are focused on the subject matter, illustrate each chapter.

The book closes with appendices, which include supplementary materials necessary to facilitate the readers' understanding of the theoretical problems discussed in the main text more deeply. For example, a brief introduction to the mathematical method such as Fourier transform, spline approximation, and Tikhonov regularization is given. Reading requires the knowledge of the secondary school courses on differential calculus, linear algebra, and statistics. To perform the exercises, readers must have programming skills for beginners in MATLAB.

Appendix H includes some supplementary programs needed for the exercise.

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

The author would be very grateful for the criticisms, comments, and proposals about this book, which he hopes to consider in his future work.

ABBREVIATIONS

AS-Analytical Signal
BM3D-Block-Matching-3D Filtering
BDA-Big Data Approach
BPDN-Basis Pursuit Denoising Procedure
CCD- Charge-Coupled Device
CFT- Continuous FT
CMOS-Complementary Metal-Oxide Semiconductor
CNN-Convolutional Neural Network
CARS-Competitive Adaptive Reweighted Sampling
CPM-Combined Polynomial Method
CPSD-Cross PSD
CS-Compressed Sensing
CWT- Continuous WT
DCP-Dichlorophenol
DCT-Discrete Cosine Transform
DFT-Discrete FT
DIP-Digital Image Processing
DMD-Digital Micro-Mirror Device
FBP-Fourier Basic Patterns
FC-Fractal Compression
FCOS-Fluorescence Correlation Spectroscopy
FDA-Functional Data Analysis
FFT-Fast FT
FIR-Finite Impulse Response
FLC-Fixed-Length Code
FPN-Fixed-Pattern Noise
FT-Fourier Transform
FTIR-Fourier Transform IR
FWHM-Full Peak Width at Half-Maximum
GC-Gas Chromatography
GDH-Generalized Discrete Harmonics
GFT- Generalized FT
GRNN-Generalized Regression Neural Network
HA-Harmonic Analysis
HIS-Hyperspectral Imaging
HPLC- High-Performance LC
HT-Hadamard Transform
ICA-Independent Component Analysis
IFT-Inverse FT
KAP-Key Analytical Point
IT- Information Theory
KLT-Karhunen-Loève Transform
IR- Infrared
LC-Liquid Chromatography
LRA-Low Rank Approximation
LS-Least Squares
LTAS-Linear Transforms-based Analytical Spectrometry
LZ-Lempel-Ziv
MB-Multiblock
MDA-Multiscale Data Analysis
MDB-Multidimensional Database
MLR-Multiple Linear Regression
MM-Method of Moments
MMin- Matrix Minimization
MNF-Minimum Noise Fraction
MRSD-Multiresolution Signal Decomposition
MTF-Move-To-Front
NAS-Net Analyte Signal

NIR-Near IR
NMR-Nuclear Magnetic Resonance
MS-Mass Spectroscopy
OLAP-Sophisticated On-Line Analytical Processing
OLS-Ordinary Least Squares
OMS-Optical Multisensor Systems
OP-Orthogonal Polynomial
OSC-Orthogonal Signal Correction
DOSC-Direct OSC
OTM-Orthogonal Transformation Method
QMF-Quadrature Mirror-Filter
OWAVEC-Orthogonal Wavelet Correction
PC-Principal Component
PCA-PC Analysis
PCR-PC Regression
PLS-Partial Least Squares
PMG-Polynomial Modified Gaussian Function
PPM-Prediction by Partial Matching
PSD-Power Spectra Density
RGB-Red, Green, and Blue
RLC-Run-Length Coding
RS-Raman Spectrum
RSEP-Relative Standard Error of Prediction
SC-Streak Camera
SCI-Snapshot Compressive Imaging
SG-Savitzky-Golay
SFS-Synchronous Fluorescence Spectroscopy
S/N-Signal-to-Noise Ratio
SPD-Single-Pixel Detector
SPI-Single-Pixel Imaging
SR-Sparse Representation
SRS-Stimulated Raman Scattering
SVD-Singular Value Decomposition
SVM-Support Vector Machine
TD-Tucker Decomposition
TR-Tikhonov Regularization
TRS-Time-Resolved Spectroscopy
TVD-Total Variation Denoising
UV-Ultraviolet
VIS-Visible
VLC-Variable-Length Code
VQ-Vector Quantization
WHT-Walsh-HT Transform
WT-Wavelet Transform
XML-Extensible Markup Language

PART I:

SIGNALS AND NOISE IN SPECTROSCOPY

INTRODUCTION

Information obtained by modern spectral instruments is represented in the form of one or multidimensional sets of numerical data. The development of efficient methods for processing and, in particular, compressing this data requires knowledge of its internal structure. In other words, knowledge of the properties of spectral signals will be useful in designing algorithms for compressing spectra.

Since a significant part of the book is devoted to the technique of spectroscopy measurements for analytical purposes, we will use the concept of "analytical signal" for further presentation. This concept can also be successfully applied to the description of spectroscopic measurements in the general case.

According to Danzer [1], *analytical signal* (AS) in analytical chemistry is a response of the measurement system (analytical instrument) to the object under study [1]. Usually, the system's output is a linear or nonlinear mixture of the responses to different analytes, including noise and background.

In terms of signal processing theory, AS "refers to either a continuous or discrete measurement sequence which consists of a pure or undistorted signal corrupted by noise" [2]. In spectroscopy, this measurement sequence is usually a function of the frequency or wavelength. These arguments are, in turn, also time functions. In chromatography, the x-axis (abscissa) of the AS is often taken to represent time. AS processing is carried out in the time scale of coordinate x (the time domain) or the transformed x-scale, e.g., using the Fourier transform (FT) (the frequency domain). In the last case, the y-coordinates of the AS are the intensity of the Fourier harmonics. Each harmonic is a linear combination of the AS ordinates.

The y-axis of the AS may be the derivatives of the AS relative to the x-argument. Therefore, signal processing in the time domain involves extracting useful information from the transformed y-coordinate system. This method, e.g., derivative spectroscopy or chromatography, is one of the Linear Transform Coordinates Methods [3].

In the following chapters, simulated spectra are often used as models to study processing methods. Building these models requires knowledge of the shapes of the AS components and characteristics of the measurement noise. For simplicity, the AS structural elements, which have a bell-shaped form (symmetrical and asymmetrical), are named peaks.

CHAPTER ONE

PEAK SHAPES IN SPECTROSCOPY*

In spectroscopy [1] (Appendix F) and chromatography [2], a conventional model of AS is a sum of elementary peaks and a baseline. The peaks have symmetrical (Gaussian, Lorentzian, and Voigt) and asymmetrical [3-4] bell-shaped forms. These shapes are due to the impact of physical and instrumental factors. The last effects are essential in chromatography.

Spectral lines have "a natural width" due to the expansion of energy levels according to the Heisenberg uncertainty principle [5]. However, this width is negligible. The following effects cause the lines to become broader: the Zeeman Effect, thermal Doppler broadening, collisional broadening, and velocity broadening [5].

A significant contribution to spectral contour formation is made by the thermal motion and interactions between the particles of the object under study. The first factor causes Doppler broadening due to changes in the frequency of the radiation emitted or absorbed by the moving particles.

The instrumental distortions include:

- The limited maximum delay time in the interferogram, which is obtained using Fourier-transform infrared (FTIR) spectrometers, and the non-zero width of the instrumental function of monochromators.
- Inter- and/or intra-atomic and molecular interactions in the samples under study (e.g., Stark broadening in dense plasma [6]; spin-spin interactions in NMR spectroscopy [7]; and inter- and intra-molecular associations in IR spectroscopy of liquids [8]).

Doppler broadening forms the Gaussian contour [5] (Fig. 1.1-1):

$$F_G(\lambda) = F_0 \exp\{-4ln2[(\lambda - \lambda_0)/w]^2\}, \qquad (1.1 - 1)$$

where the abscissa argument λ is some physical quantity (e.g., a wavenumber); F_0 is the amplitude of the peak maximum whose position is λ_0; and w is the full peak width on the half-maximum (FWHM). The frequently used peak model in chromatography is the Gaussian type.

[*] Major parts of this chapter were taken from the following book: Dubrovkin, J. Derivative Spectroscopy. Cambridge Scholars Publishing. 2021.

Interactions between particles form the Lorentzian (or Cauchy) contour [5] (Fig. 1.1-1):

$$F_L(\lambda) = F_0\,\{1/[1+4[(\lambda-\lambda_0)/w]^2\}. \qquad (1.1-2)$$

In physics, the functions $F_G(\lambda)$ and $F_L(\lambda)$ are usually normalized to the unit area:

$$M_G\int_{-\infty}^{\infty} F_G(\lambda)d\,\lambda = 1.$$

Since $\int_{-\infty}^{\infty} F_G(\lambda)d\,\lambda = F_0\sqrt{\pi}\frac{w}{2\sqrt{ln2}}$ [6], $M_G = 2\sqrt{ln2/\pi}\,/(wF_0)$.

$$M_L\int_{-\infty}^{\infty} F_L(\lambda)d\,\lambda = 1.$$

$\int_{-\infty}^{\infty} F_L(\lambda)d\,\lambda = F_0\pi\frac{w}{2}$ [9], therefore $M_L = (2/\pi)/(wF_0)$.

Each spectrum, measured by a spectral instrument, is disturbed by this device. From a mathematical point of view, the measured ($F_m(\lambda)$) spectrum is the convolution of the undistorted ("true") spectrum (F_T) with the instrumental function (I) [10]:

$$F_m(\lambda) = \int_{-\infty}^{\infty} I(\lambda+\lambda')F_T(\lambda')d\lambda' + \eta(\lambda), \qquad (1.1-3)$$

where $\eta(\lambda)$ is the additive noise of the measurements.

Formally, the generalized function (I_g) includes instrumental factors and involves the physical-chemical interactions that lead to the distortions of the spectrum under study. According to [11], the undisturbed spectrum equals the measured one corrected by its weighted derivatives:

$$F_T(\lambda) = F_m(\lambda) + \sum\nolimits_{n=1}^{\infty}(b_n/j^n)d^nF_m(\lambda)/d\lambda^n, \qquad (1.1-4)$$

where constants b_n are defined by I_g; $j = \sqrt{-1}$.

Shapes of the spectral peaks, measured in practice, often differ from the "pure" functions (Eqs. (1.1-1), (1.1-2)) due to the combined impact of broadening factors and instrumental distortions. Therefore, the peaks are approximated by the Voigt profile, which is the convolution of Gaussian and Lorentzian shapes:

$$F_V(\lambda) = \int_{-\infty}^{\infty} F_G(\lambda')F_L(\lambda-\lambda')d\lambda'. \qquad (1.1-5)$$

The precise approximation of the Voigt FWHM with an accuracy of 0.02% [12]:

$$w_V = 0.5346w_L + \sqrt{0.2166w_L^2 + w_G^2}. \qquad (1.1-6)$$

Non-integral analytical expressions of the Voigt functions are cumbersome (for example, the pseudo-Voigt normalized profile [13] and its improved version [14]). Approximation of Voigt function combined two expressions, one of which was asymptotic [15]. Software products used many rough but simplified combinations of Gaussian and Lorentzian profiles (e.g., [16]).

The Voigt function and its derivatives were represented by series in Hermite polynomials [17].

Fig. 1.1-1 shows the intermediate position of the Voigt peak between the Lorentzian and Gaussian profiles.

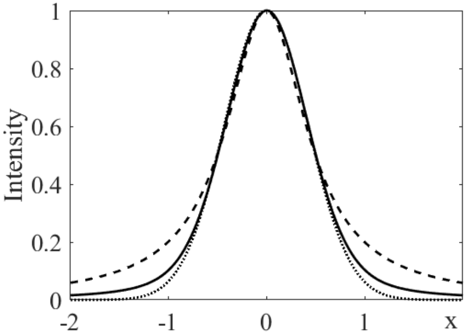


Figure 1.1-1. Lorentz, Gauss, and Voigt peaks (dashed, dotted, and solid curves, respectively). $F_0 = 1, x = (\lambda - \lambda_0)/w, w = 1$.

For the convenience of mathematical operations, Eqs. (1.1-1) and (1.1-2) are replaced by their Fourier transforms (FT) (Appendix A1):

$$\tilde{F}_G(p) = F_0 w_G\left(\sqrt{\pi}\,/2\sqrt{ln2}\right) exp(-p^2/16ln2), \qquad (1.1-7)$$

$$\tilde{F}_L(p) = F_0 w_L(\pi\,/2) exp\,(-|p|\,/\,2), \qquad (1.1-8)$$

where $p = \omega w$, and ω is the angular (Fourier) frequency.

One assumes that $\lambda_0 = 0$. If $\lambda_0 \neq 0$, then the FT is modified:

$$F(\lambda - \lambda_0) \rightarrow \tilde{F}(\omega)\exp(-j\omega\lambda_0). \qquad (1.1-9)$$

The FT of the Voigt peak is the product of the FT-convolution components normalized to the maximum intensity or area:

$$\tilde{F}_V = \tilde{F}_G * \tilde{F}_L. \qquad (1.1-10)$$

Asymmetrical peaks

There are a lot of mathematical models of asymmetrical peaks usually used in chromatography [18]. As examples, consider the polynomial modified Gaussian (PMG) and Lorentzian (Dobosz) functions:

$$F_{PMG} = exp\left(-y^2/B_y^2\right), \qquad (1.1-11)$$

where $y = 2\sqrt{ln2}(\lambda - \lambda_0)/w$;

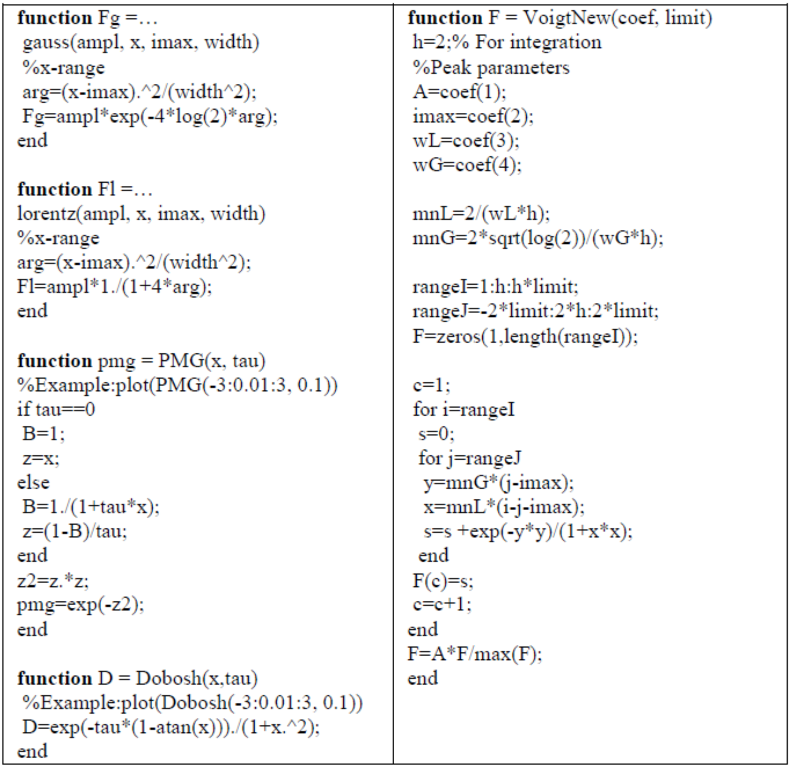
$B_y = 1 + \tau y$; τ is the asymmetry parameter.

$$F_D = exp\left(-\tau(1 - tan^{-1}y)\right)F_L, \qquad (1.1-12)$$

where $\tau \neq 0$; $F_L = 1/(1 + y^2)$.

Exercise 1.1-1

The reader is invited to plot all profiles, discussed in this chapter, using the following Matlab functions.



```
function Fg =...
 gauss(ampl, x, imax, width)
 %x-range
 arg=(x-imax).^2/(width^2);
 Fg=ampl*exp(-4*log(2)*arg);
end

function Fl =...
lorentz(ampl, x, imax, width)
%x-range
arg=(x-imax).^2/(width^2);
Fl=ampl*1./(1+4*arg);
end

function pmg = PMG(x, tau)
%Example:plot(PMG(-3:0.01:3, 0.1))
if tau==0
 B=1;
 z=x;
else
 B=1./(1+tau*x);
 z=(1-B)/tau;
end
z2=z.*z;
pmg=exp(-z2);
end

function D = Dobosh(x,tau)
 %Example:plot(Dobosh(-3:0.01:3, 0.1))
 D=exp(-tau*(1-atan(x)))./(1+x.^2);
end
```

```
function F = VoigtNew(coef, limit)
 h=2;% For integration
 %Peak parameters
 A=coef(1);
 imax=coef(2);
 wL=coef(3);
 wG=coef(4);

 mnL=2/(wL*h);
 mnG=2*sqrt(log(2))/(wG*h);

 rangeI=1:h:h*limit;
 rangeJ=-2*limit:2*h:2*limit;
 F=zeros(1,length(rangeI));

 c=1;
 for i=rangeI
  s=0;
  for j=rangeJ
   y=mnG*(j-imax);
   x=mnL*(i-j-imax);
   s=s +exp(-y*y)/(1+x*x);
  end
 F(c)=s;
 c=c+1;
end
F=A*F/max(F);
end
```

CHAPTER TWO

ANALYSIS OF NOISE IN SPECTRAL MEASUREMENTS*

Measurements of any physical value are influenced by random and systematic errors. Systematic errors may arise from incorrect measurements (a typical example is the improper preparation of the blank cuvette in the spectrophotometer) and from the imperfection of the instrument (e.g., spurious reflection and radiation). Often, systematic errors can be decreased until they are negligible by improving the apparatus and the measurement process (e.g., by correct calibration). However, 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 the analog or numerical processing of obtained results (e.g., by smoothing). It is important to emphasize that reducing the random errors can cause unpredictable, significant distortions of the actual value of the quantity to be measured (systematic errors).

Random errors vary randomly with time. They are due both to the errors in the 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 probability theory. The following classification was given in a series of articles summarized by Mark and Workman [1].

Since increasing the redundancy of the measurement process (number of measurements) can reduce the effect of measurement errors [2], data compression, its distortion, and denoising are interconnecting topics.

* Major parts of this chapter were taken from the book: Dubrovkin, J. Mathematical Processing of Spectral Data in Analytical Chemistry: A Guide to Error Analysis. Cambridge Scholars Publishing. 2018.

The sources of noise

Detector-dependent noise

The thermal noise (IR, NIR spectrometers) is the noise in the thermal detectors. This noise is independent of the intensity of the electromagnetic radiation that falls on the sensor; it has a Gaussian (normal) intensity distribution in the time domain. (We hope the readers remember the Gaussian distribution from their university courses in probability and statistics). This noise is often called "white noise" because it has the constant power density of the Fourier spectrum (Appendix A1) in the vast range (Fig. 1.2-1). Noise power is proportional to the width of an interval of Fourier frequencies (bandwidth) of the recording device. For more details on noise, see Appendix C.

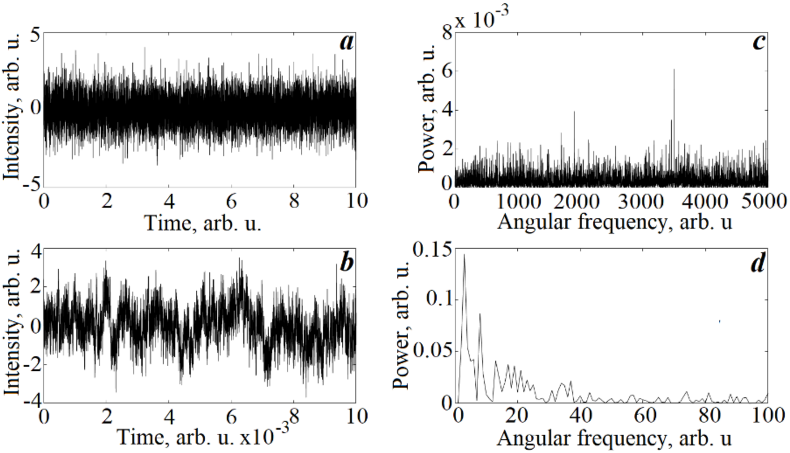


Figure 1.2-1. White and pink noises in the time (a, b) and Fourier (c, d) domains. The noise has a zero mean value; the standard deviation is one.

For numerical experiments, analytical signals are usually simulated by the composition of the peaks of known shapes. This model includes noise generated by a computer program. For example, in the MATLAB package, the function 'randn' generates pseudo-normal noise with zero mean and unit standard deviations. Numerical experiments show that these noise parameters are only accurate if the noise array contains hundreds of points. Figure 1.2-2 demonstrates that the noise parameters are different in short intervals. Moreover, the noise is well approximated by a parabola and not by a zero line as expected. Therefore, the research may observe a false structure in the analytical signal under processing.

However, the mean noisy data, obtained in a large number of repetitions, lack this flaw. We will discuss this issue below.

The standard deviation of the absorbance (A) measurements, distorted by the low-intensity thermal noise [3], $\sigma_{Term} = \{k_1/ln(10)\}\sqrt{1+100^A}$, $(1.2-1)$ where k_1 is the constant indicating precision of the spectrophotometer.

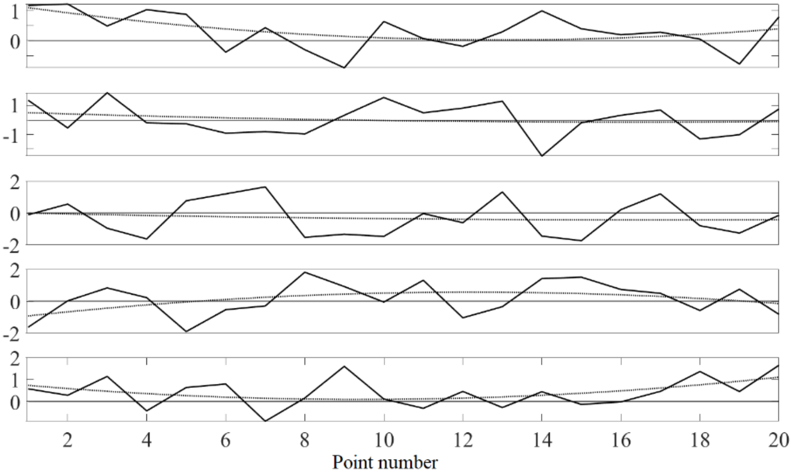


Figure 1.2-2. Parabola (dotted line) approximates the normal noise with zero mean and unit standard deviation (solid line) generated by the function randn. The estimated means and the standard deviations from the top to the bottom subplots: [0.3168, 0.6111; 0.0649, 1.1256; -0.3137, 1.1218; 0.1367 1.0445; 0.4027 0.6751].

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

The shot-noise (UV-VIS, X-ray, and gamma-ray spectrometry) in the photon-counting detectors (mostly, charge-coupled device-CCD) is due to the statistical nature of photon production [4].

The time domain probability distribution for p photons during the interval Δt obeys Poisson's law:

$$P_{\rho,\Delta t}(p) = n^p \exp(-n)/p!, \qquad (1.2-2)$$

where $n = \rho\Delta t$ is the number of photons; ρ is the photon rate.

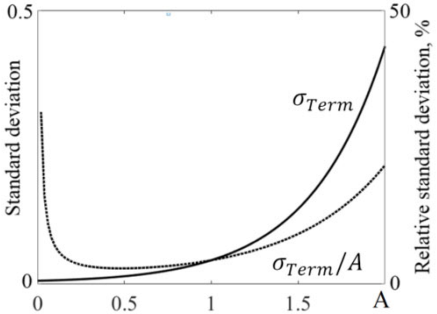


Figure 1.2-3. Eq. (1.2-1). $k_1 = 0.01$.

The mean value of the Poisson distribution and its standard deviation are

$$mean_{\Delta t}\left(P_{\rho,\Delta t}(p)\right) = n, \qquad (1.2-3)$$

$$\sigma_{\Delta t}\left(P_{\rho,\Delta t}(p)\right) = \sqrt{n}. \qquad (1.2-4)$$

Exercise 1.2-1

1.Check Eqs. (1.2-3) and (1.2-4) using the following MATLAB code:

```
L=4;
N=1e5;
 MV=mean(poissrnd(L*ones(1, N)))
 STD=std(poissrnd(L*ones(1, N)))
```

How do MV and STD depend on N?

2. Show that the photon noise depend on the signal?

3. Prove that this noise is not additive ($\eta(a+b) \neq \eta(a)+\eta(b)$)

4. Prove that the photon noise is white using the code:

```
plot(abs (fft(poissrnd(ones(1, 1e3)))))
```

Eq. (1.2-4) shows that the noise intensity increases with the square root of the mean value of the signal.

Poisson distribution can be approximated by the Gaussian function if $p > 20$.

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

$$\sigma_{shot} = \{k_2/ln(10)\}\sqrt{1+10^A}, \qquad (1.2-5)$$

where k_2 is a constant similar to k_1 in Eq. (1.2-1).

The dark current of the CCD is the origin of the thermal noise produced by "thermal electrons," whose number is independent of photon-induced signal. The dark noise is a form of shot noise and follows the Poisson relationship.

Readout (on-chip) noise appears in the process of converting CCD charge carriers into a voltage signal in the on-chip preamplifier. This noise is added uniformly to every image pixel. The noise has $1/f$ character at the high serial conversion time required to digitize a single pixel. It may significantly decrease the signal-to-noise ratio at very low signal levels.

Exercise 1.2-2

The readers are invited to represent the noise data obtained by their laboratory instruments, like in Figure 1.2-4.

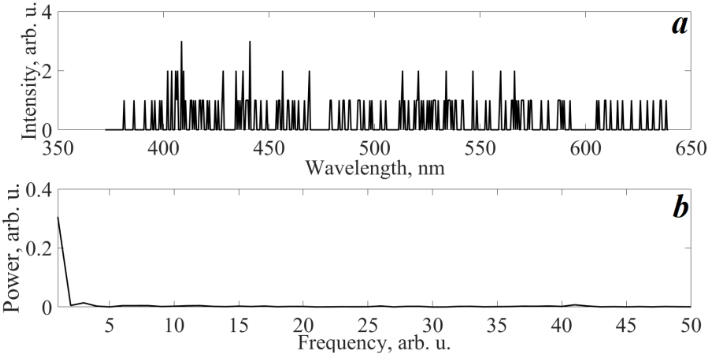


Figure 1.2-4. Noise measured on Bruker Optics 2501S spectrograph by D. V. Ushakou (Pomeranian University in Slupsk, Poland) and the noise power spectrum (panels 'a' and 'b', respectively).

Variations in energy, which are incident on the detector, are due to the vibrations of the source and the changing geometry of the radiation. They cause the flicker (pink) noise ($1/f$-noise). The noise intensity is proportional to the signal energy. The noise power spectrum depends on the frequency: $f^{-\alpha}$, where $\alpha \simeq 1$ (Fig. 1.2-1) (Appendix C). If the noise is small, then the standard deviation of absorbance σ_f is approximately constant.

Detector-independent noise

The noise sources include the 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 [3].
- The variability of the sample properties that cannot be measured. However, these properties influence the measurement property (e.g., the

changing of light reflectance in the transmittance measurements, the inhomogeneous of the sample, and the artefacts of the blood motion in the blood analysis using Functional NIR Spectroscopy (Optical Topography) [5, 6]).

- 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 only be performed in particular cases using appropriate assumptions that simplify the solution of the problem.

Image detectors [6]

An image detector (sensor) converts a 2D array of pixels of the light incident at its surface into an array of electrical signals (photocurrents), measured by the readout system. The RGB filters (Red, Green, and Blue) divide a colour image into three independent matrices.

The widely used image sensors are CCDs and complementary metal-oxide semiconductor (CMOS) devices. The range of illumination that the sensor can detect is determined by a temporal and fixed-pattern noise (FPN).

The temporal noise is due to photodetector shot noise, pixel reset circuit noise, readout circuit thermal and flicker noise, and quantization noise.

The FPN is a variation of the pixel-to-pixel output signals despite the uniform illumination of the sensor. The FPN contains two components: independent of pixel signal and increasing with its level (offset and gain FPN).

All technical details are given in the tutorial [7].

Computer modelling of correlated noise

Let us consider the noise intensity distribution in the time domain [8]. As was pointed out above, the Gaussian (normal) and Poisson noise distributions are usually used to model noise in spectroscopic measurements. In the time domain, noise is characterized mathematically by the error covariance matrix $\boldsymbol{COV}$, in which a diagonal element

$$COV_{ii} = \sigma_i^2 \qquad (1.2-6)$$

represents the noise variance (dispersion) in the ith 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 σ_i^2 (Eq. (1.2-6)) are the same, the noise is called *homoscedastic*; otherwise, it is *heteroscedastic*.

The heteroscedastic noise source, in particular, is a randomly varying baseline (background) [9, 10], 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, whose coefficients are random numbers. This baseline is added to some spectrum. The procedure is repeated r times. Then the intensity of the tth spectrum in the point i is

$$y_{it} = y_i + \eta_i + a_{0t} + a_{1t}i + a_{2t}i^2, \qquad (1.2-7)$$

where y_i is the undistorted value; η_i is the normal noise with zero mean and dispersion σ_y^2; a_{qt} ($q = 0, 1,$ and 2) are the constant coefficients that change randomly for each spectrum. The estimation of the covariance matrix element is [11]

$$COV(y_k, y_l) = [1/(r-1)]\ \sum_{t=1}^{r}(y_{kt} - \bar{y}_k)\,(y_{lt} - \bar{y}_l), \qquad (1.2-8)$$

where the bar is the average symbol. According to Eq. (1.2-7):

$$\bar{y}_i = y_i + \bar{a}_0 + \overline{(a_1)}i + \overline{(a_2)}i^2. \qquad (1.2-9)$$

Substituting Eqs. (1.2-7) and (1.2-9) into Eq. (1.2-8), we have

$$COV(y_k, y_l) = \sigma_y^2\xi_{kl} + BS(y_k, y_l), \qquad (1.2-10)$$

where $BS(y_k, y_l) = [1/(r-1)]\sum_{t=1}^{r}\Phi(k)\,\Phi(l)$;

$\Phi(v) = \delta_{a_{0t}} + \delta_{a_{1t}}v + \delta_{a_{2t}}v^2$;

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

Suppose that $\delta_{a_{qt}}$ is a normal random variable with a 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. (1.2-10):

$$COV(y_k, y_l) = \sigma_y^2\xi_{kl} + \sigma_{a_0}^2 + \sigma_{a_1}^2kl + \sigma_{a_2}^2k^2l^2\,. \qquad (1.2-11)$$

Correlations between analytical points may also be due to the pre-processing, e.g., digital smoothing [9]. Since a digital filter is usually much shorter than a spectrum, the covariance matrix will be populated primarily with zeros (sparse) and, therefore, singular. A correctly computed inverse is impossible if the matrix is near to singular.

CHAPTER THREE

INFORMATION-THEORETIC ASPECTS OF THE LINEARLY TRANSFORMED SPECTRA*

Using information theory (IT) in analytical chemistry and, notably, in analytical spectrometry allowed researchers to reveal new peculiarities of the objects under study [1]. A sample of the object may be characterized by some unknown quantity of intrinsic analytical information (I_{intr}) [2] (Fig. 1.3-1). The analytical instrument extracts from I_{intr} only a small part I_{instr} contained in the analytical signal (AS). Finally, the information content of output data $I_{data} \leq I_{instr}$ due to data processing.

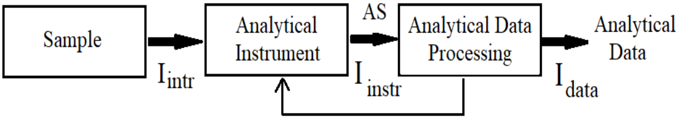


Figure 1.3-1. Information flow in analytical system (adapted from [2]).

Two right rectangles (Fig. 1.3-1) represent the information flow of the analytical data processing.

To illustrate some information-theoretical problems of the linear transformed spectra, let us consider the derivative spectrometry methods. In some publications, their success is associated, in a veiled form, with gains achieved by the differentiation. However, according to the general principles of information theory, any linear transformation cannot increase the data information content [3]. The theoretical study [4] discussed this problem in spectroscopy.

We will use the Savitzky-Golay (SG) digital filters (Appendix D) to demonstrate the information changes occurring in derivatives. This less formal method does not require specific knowledge in spectroscopy instrumentation [4].

* The first part of this chapter was taken from the book: Dubrovkin, J. Derivative Spectroscopy. Cambridge Scholars Publishing. 2021.