Challenges of Decoding Data in Spectroscopy, Reflectometry, X-Ray and Electron Diffraction

Challenges of Decoding Data in Spectroscopy, Reflectometry, X-Ray and Electron Diffraction

By

Felix N. Chukhovskii, Petr V. Konarev and Vladimir V. Volkov

Cambridge
Scholars
Publishing

Challenges of Decoding Data in Spectroscopy, Reflectometry,
X-Ray and Electron Diffraction

By Felix N. Chukhovskii, Petr V. Konarev and Vladimir V. Volkov

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): 1-5275-8604-9
ISBN (13): 978-1-5275-8604-8

TABLE OF CONTENTS

CHAPTER ONE

THE INFRARED MULTICOMPONENT SPECTROSCOPY

1.1. Introduction

Spectroscopic analysis of complex objects is impossible without using the chemometric methods of analyzing experimental data. One of the main directions of chemometrics is using the mathematical, statistical, and other numerical methods to obtain the most relevant chemical information by analyzing spectroscopic data (Malinowski, 1980). Further, we will discuss the analysis of multicomponent spectroscopy data to demonstrate methods for determining the number of components.

Mathematical chemometric algorithms, actively developed after 1970 have been reviewed in the publications, the number of which exceeds dozens, or even hundreds of thousands. Multidimensional empiric relationships and structural features of the object.

In this field, the point is the development and application of mathematical algorithms designed to separate spectrum sets of multicomponent mixtures into items of individual compounds. At the same time, the well-known chromatographic and mass spectrometric methods are powerful, but they are not good enough to study equilibrium systems.

The field of the separation method of additive spectra data is large enough. By additive data can be meant the simultaneous measurement data of different characteristics of objects with the help of measuring instruments or multi-channel sensors, proportionally registering physical or any other parameters. The term “spectrum” will represent a set of such measurements in the form of dependence of “result of measurement – the ordinal number of the sensor channel”.

There are different parameters, which lead to changes in the spectra of the mixture components. In chemistry, the parameters are the following: a time and conditions of chromatography, *i.e.,* pH of the medium, temperature, concentration of initial substances, etc. when studying equilibrium chemical systems; conditions of vital activity of biological objects; time and other parameters when researching intermediate products of multistage reactions; excitation energy in luminescence analysis and photoelectronic experiments; ionic strength of a solution and many others.

Separation methods of the multicomponent spectra (MCS) into the individual component spectra can be divided into two groups. One group allows one to find a unique set of spectrum components that coincides with the reference true ones. As to the other methods, using some *a priori criteria*, it is possible to find approximate contours of the individual spectra from the MCS. The application of the different techniques of the MLS expansion into individual spectra is determined by the nature of the information used, which, in turn, can also be divided into two groups.

Further, we will refer to the spectroscopic data matrix as the information source. Any other information is the additional one, which may concern either the relative component concentrations in the mixtures, the shape of the individual spectra, or the presence of known components.

Even in the simple case of the single mixture spectrum, its decomposition over the known components requires using the solution quality criteria, which would not be confined only by calculation of the Chi 2-value, but contain more details of the decomposition result reliability. In practice, the spectroscopic characteristics of the individual components are often unknown. In this case, the MCS can be decomposed into several individual spectra. New possibilities open up in opposition to a sole MCS, a set of several MCS with different component compositions has to be used. In addition to the non-negativity of the individual spectra, the main requirement is that the individual spectra additivity holds. For example, in the case of IR spectra of organic compounds, the molecules of the components should not form chemical bonds between themselves, which would lead to a shift in the frequencies of the spectral bands and a change in their shape (Hyvarinen, 2001). As to additional information that concerns the mathematical operational processing of the additive signal decomposition, one can refer to the work (Lee, 1998).

1.2. Theoretical fundamentals of the multicomponent spectra decomposition

Mathematical methods of decomposition of the additive spectral contour sets are based on the construction of linear combinations, *i.e.,* addition and subtraction with some coefficients of the observed MCS.

Let the contours of the spectra be represented on some grid of N values of the abscissa $i = 1, ..., N$ (in the future, we will not be interested in the absolute values and units of measurement of the abscissa scale, the data are assumed to be dimensionless). We denote the vector-spectra of mixtures $\boldsymbol{d}_j$, $j = 1, 2, ..., M$ (M is the number of mixtures in the studied multivariate set), the spectra of components are $\boldsymbol{e}_k$, $k = 1, 2, ..., K$ (K is the number of components). The additivity condition means that each of the mixture spectra must be a linear combination of the component spectra (for simplicity, we will not consider the influence of noise here):

$$d_{ij} = \sum_{k=1}^{K} e_{ik} \cdot c_{kj}, \tag{1.1}$$

where e_{ik} is the relative spectral contribution of the kth component to the jth spectrum of the mixture with the weight (concentration) c_{kj}. The index i denotes the serial number of the point in the spectrum. In matrix form, the linearity relation is written as

$$D = E \cdot C, \tag{1.2}$$

where D is the N x M matrix of mixture spectra, E is the N by K matrix of component spectra, and C is the K by M matrix of relative concentrations.

The most common approach is self-modeling, first proposed by Lawton (1971) to decompose two-component mixtures. Self-modeling is based on the representation of the component spectra x_{ik} as a linear combination of the eigenvectors $\boldsymbol{u}$ of the matrix of second moments (covariance matrix)

$$D \cdot D^T = U \cdot \Lambda \cdot U^T, \tag{1.3}$$

Λ is a diagonal matrix of eigenvalues ordered in descending order:

$$x_{ik} = \sum_{j=1}^{M} u_{ij} \cdot y_{jk} \tag{1.4}$$

where y_{jk} are the unknown coefficients, the index i refers to the number of the spectrum point, k is the component index, and j is the eigenvector index.

The vector representation of spectroscopic information taking into account relations (1.1) and (1.4) makes it possible to demonstrate the peculiarities of the problem. If the number of spectral points N exceeds the number of components K, a set of the component spectra vectors is defined in the N-dimensional space, K – dimensional subspace, which contains all spectra of mixtures (see (1.1)). *Figure 1.1* shows the case of $N = 3$, $M = 4$, $K = 2$, i.e., the subspace of the mixture spectra represents a plane in 3-dimensional space (intensities are plotted along the coordinate axes), and the component spectra vectors form the basis of this plane (non-orthogonal in the general case).

When matrix D is only known, the problem of determination of the component spectra matrix E reduces to finding the unknown matrices E and C involved in (1.2) in the form of a system of nonlinear equations, the solution of which is not unique. Indeed, if X' and C' are solutions of (1.2), it is always possible to find a nondegenerate matrix A, so that

$$D = X'' \cdot C'', \tag{1.5}$$

where $X'' = X' \cdot A$ and $C'' = A^{-1} \cdot C'$. This problem of non-uniqueness is the main one in analyzing multicomponent mixtures by their spectra and is solved by involving additional information and restrictions on the type of solution.

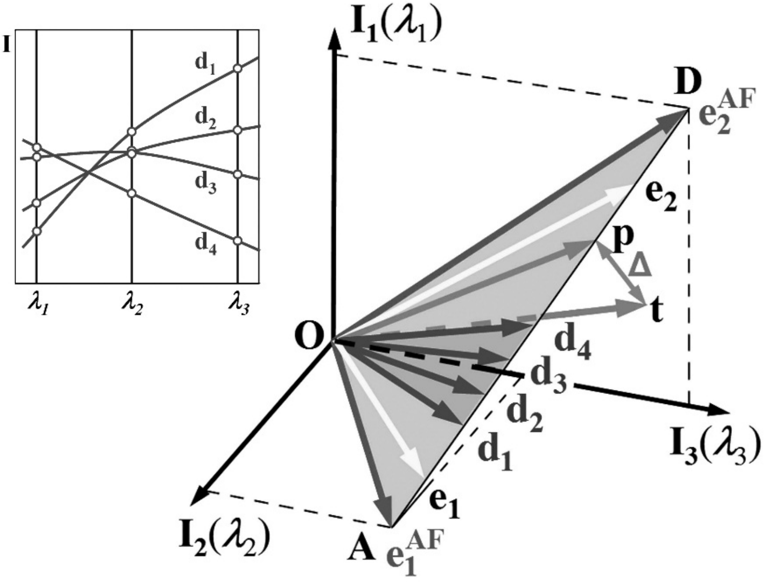

Figure 1.1. Vector representation of the decomposition problem for the case of 2-component spectra of 4 mixtures recorded at three wavelengths $\lambda_1 \div \lambda_3$. $\boldsymbol{d}_1 \div \boldsymbol{d}_4$ are the mixture spectra. $\boldsymbol{e}_1$ and $\boldsymbol{e}_2$ are the spectra of components. $\boldsymbol{e}_1^{AF}$ and $\boldsymbol{e}_2^{AF}$ are the non-overlapped component spectra. Along the axes, $I_1 \div I_3$, the spectrum intensities are deposited at three values of the abscissa $\lambda_1 \div \lambda_3$. $\boldsymbol{t}$ is a vector that is not a component of the mixture, $\boldsymbol{p}$ is its projection on the plane of mixtures AOD, Δ is the difference $\boldsymbol{t} - \boldsymbol{p}$. All vectors, except $\boldsymbol{t}$, are normalized for clarity so that their ends lie on the line AD. The areas of acceptable solutions are represented by sectors AO-$\boldsymbol{d}_1$ and DO-$\boldsymbol{d}_4$.

As is known, the common restrictions are the non-negativity of the spectra of components and their concentrations (Lawton, 1971; Sawal, 2014). Since the mixture spectra are non-negative linear combinations of the component spectra, regions of the solution existence are limited by the subspace regions defined by the mixture spectra vectors (2D-plane AOD in *Figure 1.1*). Accordingly, they are located in the 1st orthant inside a cone formed by the mixture spectra vectors (regions AO-d_1 and DO-d_4 in *Figure 1.1*).

It is shown in (Fok, 1973) that when the conditions of non-overlapping (called the Alentsev-Fok conditions) are met, the vectors of linear combinations that are most distant from each other correspond to the actual component spectra.

Consider some practical approaches for the decomposition problem.

Several methods of self-modeling curve resolution, such as RADICAL, SIMPLISMA (Rodionova, 2005), MCR-ALS (Mas, 2011), have become traditional. Recently, algorithms implementing the MILCA (Mutual Information Least Component Analysis) and SNICA (Stochastic Non-Negative Independent Component Analysis (Astakhov, 2006)) methods have been proposed. To date, the MILCA and SNICA algorithms have been successfully validated for multicomponent analysis of complex objects. It should be noted that the software packages implementing the MILCA and SNICA algorithms are available to users. Detailed Internet resources, including tutorials (URL http://www.ihes.fr/~zinovyev/#books) have been prepared on issues related to multivariate data analysis. In (Ohta, 1973), the self-modeling technique is modified for processing three-component mixtures. The authors (Meister, 1984) note that for four or more components, the algorithm is significantly complicated and has not been developed by them.

The non-negativity conditions are fundamental in the factor analysis, which is widely used under the decomposition of the mixture spectra (Malinowski, 1980; Martens, 1979). The latter is based on the possible linear combinations of eigenvectors. In (Martens, 1979), the main provisions of the factor analysis are formulated.

In several works, additional requirements to the shape of individual spectra are used to obtain a single solution, which ensures the successful operation of modern software packages, such as the already mentioned MILCA and SNICA. Thus, the authors (Kawata, 1985; Friedrich, 1987) suggest supplementing self-modeling with the requirement of the minimum information entropy of the component spectra, which is calculated under the assumption that the absolute value of the first derivative of the spectral contour p_{ik} means the probability density function, and then the entropy of the kth spectrum is equal to

$$H_k = -\sum_{i=1}^{N} p_{ik} \cdot \log_2 p_{ik} \cdot \qquad (1.6)$$

Authors (Friedrich & Yu, 1987) also use other complexity criteria. These are either the value of

$$\int_{\nu}\left[1+\left(\frac{\partial x(\nu)}{\partial \nu}\right)^{2}\right]^{1/2}, \tag{1.7}$$

where $\boldsymbol{x}(\nu)$ is the spectral contour.

The method provides the weakly overlapped areas in the spectra via similar ones in the contours of the eigenvectors *U* and the linear combinations of the latter to obtain the pure zero or resulting low-intensity spectra. This method is based on the Alentsev-Fok procedure (Fok, 1973) but applied to the eigenvectors *U* of the covariance matrix (1.3). The degree of the approach solution correctness of the component spectra depends on their non-overlap degree. The technique has been successfully applied to the 2- and 3-component mixture spectra decomposition.

The known decomposition methods turn out to be insufficient when the non-overlap zones are too short, or their number is not enough following to Alentsev-Fok (Fok, 1973).

The conditions of non-negativity of spectra and concentrations of components are successfully applied when separating small-angle X-ray scattering data in the **REGALS** package (Meisburger, 2021). The scattering intensity data is obtained, in particular, in the process of preparative chromatography of the sample in the stop-flow mode and then processed by a method called Evolving Factor Analysis (EFA) (Keller, 1992). From a set of experimental scattering intensity curves, sections with different numbers of components are selected that are estimated by analyzing graphs of singular numbers of data matrices, the size of which is increased by sequentially adding new measurements. After that, the individual scattering spectra are calculated using a variant of the ALS (Alternating Least Squares) method of non-negative least squares.

Interactive algorithms to analyze the spectra of mixtures are very effective in the case of a small (3-4) number of components (Bu, 2000; Leger, 2002).

There are several publications in which geometric analysis, like *Figure 1.1*, extended to three components, is effectively used to separate spectra (Akbari, 2013; Rajko, 2005, 2006). In what follows, we will consider approaches that are not constrained by the number of components.

1.3. Determination of the number of components

To determine the number of components *K* from the spectroscopic data matrix *D* can be regarded as a preliminary analysis. The known methods, as a rule, are based on knowledge of the magnitude of experimental errors and are essentially statistical (Lawley, 1971). In addition, several non-statistical methods have been proposed (Culler, 1984; Knorr, 1981; Shrager, 1982), which allow us to obtain plausible results. As to a real problem, it seems necessary to use several criteria since sole statistical ones may be incorrect due to the erroneous statistical hypotheses concerning experimental errors, and non-statistical estimates usually do not have sufficient justification. Most of the well-known methods of estimating K are based on a factor analysis of the matrix of initial data, which allows us to identify a set of "significant factors", the number of which is the number of components. Thus, in (Malinowski, 1980), more emphasis is placed on the practical use of factor analysis, and in (Lawley, 1971) – on the statistical aspect of the methods.

The estimation of the number of components can be done using the singular decomposition of the data matrix D:

$$D = U \cdot \Lambda \cdot V^T, \tag{1.8}$$

where U is an N x M matrix of eigenvectors of $D \cdot D^T$ (in the literature devoted to the factor analysis, it is a covariance matrix). V is the eigenvector matrix of the correlation matrix $D^T \cdot D$, and Λ is a diagonal matrix of singular numbers arranged in the non-increasing order (positive square roots from non-zero eigenvalues of the matrix or). Afterward, we will assume that the number of spectral points N, the number of mixtures M, and components K satisfy the inequality

$$N > M > K, \tag{1.9}$$

although the inequalities are sufficient

$$N \geq K \text{ and } M \geq N, \tag{1.10}$$

To facilitate the systematization of methods for estimating the number K, let us review the decomposition methods of (1.8).

Following to (Malinowski, 1980), we will present the experimental data matrix D in the form of the sum of the unperturbed matrix $D^*$and the matrix of experimental errors F:

$$D = D^* + F\ . \tag{1.11}$$

Under the condition (1.9), the rank of $D^*$is equal to K (in other words, the rank of the matrix D^* coincides with the rank of the relative concentration matrix C, this condition is assumed to be in further). But, due to the uncorrelated experimental errors, the rank of D is equal to the number of mixtures M.

In turn, the error matrix can be represented as

$$F = F^\# + F^0, \tag{1.12}$$

where $F^\#$ is formed by the part of errors that contribute to the first K singular vectors U, and F^0 is composed of errors that are responsible for non-zero singular numbers λ_i, $i = M$+1, ..., M and the corresponding vector columns u_i. Directly from (1.11) and (1.12), one obtains:

$$\begin{aligned} D^{\#} &= D^{*} + F^{\#} \\ D &= D^{\#} + F^{0} \end{aligned} \tag{1.13}$$

The matrix F^0 is usually called the residual error, which can be discarded by reconstructing the data matrix $D^{\#}$ from the first singular vectors of the original matrix D according to the relation (1.2). It should be noted that the initial errors are distributed over the matrices $D^{\#}$ and F^0, *i.e.*, they are formed only by some part of the errors.

The first group of methods for finding the number of components may be reduced to finding the splitting (1.13) (these methods are called residue analysis). At the first stage, the decomposition (1.8) is calculated, and then the singular vectors are divided into two groups, called primary and secondary axes. The primary axes are the first K singular vectors that form the basis of the column space (in the case of $\boldsymbol{u}$ – vectors) or rows (in the case of $\boldsymbol{v}$-vectors) of the matrix $D^{\#}$. If $Q = U \cdot S$, then

$$D = Q \cdot V^T = Q^{\#} \cdot V^{\#T} + Q^0 \cdot V^{0T}, \tag{1.14}$$

where $Q^{\#}$ is a matrix of primary axes of size N by M, whose columns $\{K$+1, ..., $M\}$ are zero, and Q^0 is a matrix of N by M secondary axes with zero columns 1,2, ..., K. Then, the first method for estimating the number of components (Residual Standard Deviation, RSD) will consist in the calculation of the expressions

$$\mathrm{RSD} = \sum_{i=1}^{N} \sum_{k=K+1}^{M} (q_{ik}^0)^2, \tag{1.15}$$

where q_{ik}^0 are the elements of the matrix Q^0;

$$\mathrm{RSD} = \left(\frac{\sum\limits_{k=K+1}^{M} \lambda_k^2}{N \cdot (M-K)} \right)^{1/2}, \tag{1.16}$$

(λ_k are singular numbers), which is equivalent to (1.9). Calculations are performed by changing the hypothesis about the number of components K starting from K = 1 and comparing the corresponding RSD value with the value of the experimental error estimate σ_{exp}. The value of K that leads to the RSD ≈ σ_{exp} is the number of components in the mixtures.

In practice, it is better to overestimate the number of components continuing the evaluation procedure until the RSD becomes strictly less than the experimental error. In (Bulmer, 1973), a method similar to the considered above is proposed, called the Residual Standard Deviation Method, based on the calculation

$$\sigma_K = \left(\frac{\mathrm{trace}\left(D^T \cdot D\right) - \sum_{k=1}^{K} \lambda_k^2}{N - K} \right)^{1/2}, \tag{1.17}$$

where K also takes increasing numbers until σ_K becomes compared with the value σ_{exp}. The main difficulty is caused by the correct estimation of the experimental errors that are not due to random noise with known dispersion, but may be systematic errors (instrument drift, intermolecular interactions, etc.). Moreover, the estimates considered are developed for the case of normally distributed noise.

It can be shown that the RSD estimate is related to the total quadratic discrepancy between the original D and $D^{\#}$ data matrices reconstructed by (1.13):

$$\mathrm{RSD} = \left(\frac{\sum_{i=1}^{N} \sum_{j=1}^{M} \left(d_{ij} - d_{ij}^{\#}\right)^2}{N \cdot (M - K)} \right)^{1/2}. \tag{1.18}$$

The RSD value is applied in the frequently used Factor Indicator Function (IND) (Culler, 1984; Knorr, 1981):

$$\mathrm{IND} = \frac{\mathrm{RSD}}{(M - K)^2}, \tag{1.19}$$

which is calculated by varying the estimate of K from 1 to M. The curve of the dependence of the IND value on M should reach a minimum value when estimating the number K is correct.

The natural method of estimating K seems to be based on comparing the variance of the elements of the residual matrix with the variance of the experimental errors. This method is implemented in some software packages. In practice, it can lead to incorrect results since the variance equality criterion of the two sets of random variables using the Fisher-Snedekor distribution is very sensitive to violations of the distribution normality (Johnson, 1977). In turn, as noted above, there is no reason to consider the distribution of errors in the data matrix normal. For the *F* criterion to be applied, it is first necessary to prove the hypothesis about the normality of the distribution law using the consent statistical criterion (Pearson's, Kolmogorov's, Smirnov's ones, see, *e.g.,* Johnson & Leone, 1977). Therefore, it is better to use statistical methods free from the distribution law knowledge of the error matrix elements or stable to deviations from distribution normality.

Here it is necessary to make an important remark. Quite widely used statistical analysis for the absence of correlations in elements of the residual matrix F^0, which corresponds to the difference between initial data matrix and matrix reconstructed by first singular vectors with indices j=1, ..., K:

$$F^0 = D - D^0, \qquad D^0 = U_{j=1,..K} \cdot \Lambda \cdot V^T_{j=1,..K}, \qquad (1.20)$$

which theoretically should contain only the noise component. In practice, it can give an underestimated number of components since the low-intensity component spectra are masked by noise and "captured" in F^0. Due to this circumstance, one will not discuss this method further.

A requirement for the residuals matrix is the absence of autocorrelation on the sequence of elements of its columns. The Durbin-Watson criterion (Durbin, 1971) can be used as the statistical test for residuals, which is governed by the formula:

$$\mathrm{DW} = \frac{\sum_{i=2}^{N\cdot M} \left(F_i^0 - F_{i-1}^0\right)^2}{\sum_{i=1}^{N\cdot M} \left(F_i^0\right)^2}, \qquad (1.21)$$

where F_i^0 represents a consecutive element of a matrix transformed into a one-dimensional array of size *M* by *N* by consecutive connection of vector-columns: i = (1,1); (1,2); ... (1,*M*); (2,1); (2,2); ... (2,*M*); ... (*N*,*M*).

The criterion value lies in the region of {0.0 – 4.0}, taking values of about 2.0 in the case of absence of correlation. If the hypothesis of the existence of a correlation between consecutive elements of the array has a significance level of 0.05, it should be rejected if the value of the criterion lies in the range {1.7 - 2.2}.

In practice, it is necessary to assume the presence of a weak residual correlation due to a violation of the additivity of the spectra during their interaction in mixtures or unaccounted weak systematic measurement errors. In most cases, the boundary needs to be slightly expanded, for example, to {1.5 – 2.5}, justifying this expansion empirically based on the experience of solving similar problems from this subject area.

Another criterion is based on calculating the variance of the eigenvalues of $D \cdot D^T$ or $D^T \cdot D$ matrices. According to the definition of the singular decomposition (1.8), these eigenvalues are the squares of the corresponding singular values.

The standard error in the *m*th eigenvalue is calculated as (Bulmer, 1973):

$$\sigma_m = \left(\sum_{j=1}^{M} \sum_{k=1}^{M} v_{mj}^2 \cdot v_{mk}^2 \cdot \sigma(Z)_{jk}^2 \right)^{1/2}, \tag{1.22}$$

where v_{mj}^2 и v_{mk}^2 are jth and *k*th elements of the *m*th singular vector, and

$$\sigma(Z)_{jk}^2 = \begin{pmatrix} \sum_{i=1}^{N} \left(d_{ij}^2 \cdot \sigma_{ik}^2 + d_{ik}^2 \cdot \sigma_{ij}^2 \right), & j \neq k \\ \sum_{i=1}^{N} \left(4 d_{ij}^2 \cdot \sigma_{ij}^2 \right), & j = k \end{pmatrix}^{1/2}. \tag{1.23}$$

The value σ_{ij}denotes the error in the experimental point d_{ij}. The number of components is the number of eigenvalues exceeding the estimate (1.22) at a given significance level. As seen from (1.23), this method allows us to take into account the error estimate in each of the spectral points of the data matrix *D* individually.

When the standard deviation varies from point to point in the spectrum and is not a constant for a given data matrix, one suggests applying the chi-square criterion (Bulmer, 1973). The disadvantage of this method is the necessity to have a fair error estimate at each experimental point. As are obtained, they are defined as

$$\chi_N^2 = \sum_{i=1}^{N} \sum_{j=1}^{M} \frac{\left(d_{ij} - d_{ij}^{\#} \right)^2}{\sigma_{ij}^2} \tag{1.24}$$

where $d_{ij}^{\#}$ are still the elements of the data matrix reconstructed from the first *K* singular vectors (1.2). The number of components *K* is determined by changing it, starting from 1, calculating χ_N^2 for each trial value *K*, and comparing it with the expected value χ_α^2

$$\chi_\alpha^2 = (N-K)(M-K), \tag{1.25}$$

where $(N-K)(M-K)$ is the number of degrees of freedom, α is the significance level. Choose such a value of *K*, which χ_N^2 is closest to χ_α^2 (1.25).

From a practical viewpoint, methods based on the spectrum of singular numbers λ_j of the matrix D analysis (or the eigenvalues of the covariance matrix (Knorr, 1981)) are of great interest. As is shown in (Ohta, 1973; Culler, 1984; Gillette, 1982), due to the normal distribution of errors, the singular numbers of the error matrix F from (1.11) lie about on a straight, when they are arranged in descending order on the graph in logarithmic magnitudes against ordinal numbers. The mean square deviation from this line proposed to be estimated by the formula (Ohta, 1973):

$$\xi_0 = 0.3 \cdot \sigma_D^2 \cdot \sqrt{N \cdot (M-2)}\,, \tag{1.26}$$

where σ_D^2 is the estimate of the experimental error variance. The deviation of the eigenvalues from the straight line on the graph becomes the more significant, the greater the dependence of σ on d_{ij} (for example, it may be in the case of Poisson noise), which sometimes limits the application of this method. The number of singular values falling out of the linear dependence is the number of components. This semi-empirical method is very convenient. In the REGALS package (Meisburger, 2021), for example, the index of the smallest singular number λ_{j} is taken as the number of components for which the condition is still met

$$\frac{\lambda_j - \sqrt{M}}{\sqrt{N}} > 1.0\,, \tag{1.27}$$

where M is the number of points in the spectrum, N is the number of the mixture spectra. However, this approach is valid for data in which the variance expectation does not depend on the abscissa, so the resulting estimates should be treated with caution.

Among the methods for estimating the number of components that do not require knowledge of the experimental error value, we should note the frequently used Embedded Error Function (IE) (Culler, 1984):

$$\mathrm{IE} = \left(\frac{K \cdot \sum\limits_{j=K+1}^{M} \lambda_j^0}{N \cdot M \cdot (M-K)} \right)^{1/2}, \tag{1.28}$$

which is calculated for the numbers K changed from 1 to M. The curve of the IE-dependence on K looks like the log (λ_j)-dependence on the number j. The number of components corresponds to the breaking point on the curve that corresponds to the sharp change in the slope angle of the curve relative to the abscissa axis).

The number of components can be found by a frequency analysis of the contours of the left singular vectors. As is noted in (Shrager, 1982), singular vectors corresponding to "significant" singular values contain mainly lower harmonics (in the Fourier spectrum). The "noise" vectors represent the high noise harmonics. The latter is valid only if the noise components have a higher frequency than the spectroscopic information data. Therefore, under the number of components estimated by this method, it is essential that the spectra sampling is carried out with a sufficiently small argument step (wavelength). In turn, the sampling noise will exceed its upper Fourier spectrum frequency.

The two other criteria for checking individual spectra included in the mixture spectra, namely, formulated in (Malinowski, 1981), are Reliability (RELI) and Spoil (SPOIL) ones. The criteria, developed for the library search of the spectrum components, show reliable results when estimating the number of components by projecting the library-known spectrum onto the data matrix $D^{\#}$ column space. For this purpose, the matrix $D^{\#}$ is reconstructed from the K-first singular vectors of the matrix D (1.8). These criteria use the "apparent" error value in the vector $\boldsymbol{t}$ of the trial spectrum of the component:

$$\mathrm{AET} = \left(\frac{\sum_{i=1}^{N} (p_i - t_i)^2}{N} \right)^{1/2}, \tag{1.29}$$

where $\boldsymbol{p}$ is the vector of the projection of $\boldsymbol{t}$ on D (or $D^{\#}$) (see *Figure 1.1*), which may be calculated using the SVD decomposition (1.2):

$$\boldsymbol{p} = U \cdot U^T \boldsymbol{t}, \quad \text{or} \quad \boldsymbol{p} = U^{\#} \cdot U^{\#T} \boldsymbol{t} \ , \tag{1.30}$$

where the sign # denotes the matrix composed of the first K columns of U, and K is the assumed (verifiable) number of components.

The value of AET is compared with the "real" error in the projection vector $\boldsymbol{p}$ (λ_{j} are the singular numbers from (1.2)):

$$\mathrm{REP} = \left(\frac{\sum_{j=1}^{M} \lambda_j^{\,2}}{\max(N,M) \cdot (M-K)} \right)^{1/2}, \tag{1.31}$$

$$\mathrm{RET} = \left(\mathrm{AET}^2 - \mathrm{REP}^2\right)^{1/2}.$$

The criteria used in deciding whether a test vector is a component are calculated using the formulae:

$$SPOIL = \frac{\mathrm{RET}}{\mathrm{REP}}, \qquad \mathrm{RELI} = \left(\frac{\mathrm{RET}^2 - \mathrm{RET}_{\mathrm{est}}^2}{\mathrm{AET}^2} \right)^{1/2}, \tag{1.32}$$

where the RET_{est}^2 is the estimated error of the data used for building the test vector.

According to (Lawton, 1971), the following threshold values are proposed: RELI > 0.5 means that ***t*** (or a set of the test spectra) is acceptable as the spectrum of the component, as well as SPOIL < 3.0. SPOIL = 3.0...6.0 indicates the possibility that ***t*** is the component spectrum; SPOIL > 6.0 means a negative answer.

Despite the apparent rigidity of the error theory, based on the "triangle ratio":

$$\mathrm{AET} = \mathrm{RET}^2 - \mathrm{REP}^2.$$

This equality is introduced artificially and not supported by any statistical conclusions. Notice we must consider the RET and REP errors as they would be independent, and their variances could be summed. Nevertheless, the considered criteria allow one to obtain good results.

Thus, there is a wide variety of methods for estimating the number of components, the simultaneous use of which makes it possible to increase the reliability of the answer by comparing the results obtained by different methods. The very fact that the estimates do not match can carry useful information. In opposition to the statistical techniques, the overestimated number of components obtained by the non-statistical ones can evidence that errors caused, *e.g.*, by the temporary instability of the device (wavelength drift, *etc.*) are predominant. In turn, intermolecular interactions in mixtures lead to changes in the spectra of components, *etc.*

In the next Section, several examples of estimating the number of components are presented.

1.3.1. Data preparation

Several mixtures of organic compounds were prepared to demonstrate the procedures for determining the number of components. Vacuum distillation and preparative liquid chromatography were used. The chromatography device was a 250x3.5 mm glass column filled with Silasorb 600 (LC) sorbent of 30 microns, equipped with a syringe sample inlet and pressure system (a nitrogen cylinder with a pressure reducer for 0-0.4 MPa overpressure). The flow rate was 0.75 ml/min. The device design is simple, and we do not give its scheme since it is not of fundamental importance. Perkin Elmer 580B IR spectrometer coupled to an Interdata 6/16 micro-computer was used as a detector.

The spectra of the mixtures-fractions were taken in a flowing cuvette of KBr with a thickness of 0.15 mm, stopping the solvent flow with a needle valve at approximately equal intervals. The level of spectroscopic noise was 0.4% T. The obtained spectra have been smoothed using standard programs (included in the spectrometer software). The spectrum of the solvent (carbon tetrachloride, which was a known component) was compensated using a cuvette of variable thickness placed in the reference spectrometer channel or by a program for spectra subtracting. The initial data was formed by encoding the most informative sections in the mixture spectra with a variable step. The relative component concentration matrix has been determined onto the chromatographic peaks obtained on a gas chromatograph LXM 8MD, namely: the carrier is 3%E-30 on Panchrome 3, column length 3m, temperature 120oC, with an accuracy up 3.5% rel. (Volkov, 1996).

The model mixtures of hexane, toluene, acetone, cyclohexanone, diethyl ether, and amyl acetate (a model of technical waste of polymer production) have been used. Chromatograms have been recorded at the frequency of 1450 cm^{-1} when absorption is approximately the same for all the components. The example of a typical chromatogram is shown in *Figure 1.2*. The spectra of the mixtures have been recorded in the range of 1750-1110 cm^{-1} with a resolution not worse than 5 cm^{-1}. After that, when preparing the spectroscopic data matrix *D*, the low-intensity areas were discarded, and the regions 1750-1690, 1610-1595, 1510-1300, and 1160-1110 cm^{-1} were left.

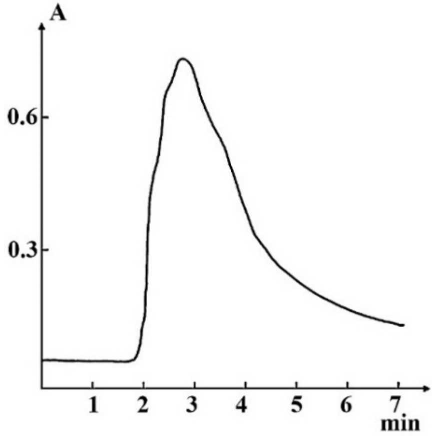

Figure 1.2. Typical chromatogram of a 5-component mixture **No. 1** of hexane, toluene, acetone, cyclohexanone, and diethyl ether at an absorption frequency of 1450 cm^{-1}. The monotony of the peak points to poor component separation under these chromatographic conditions.

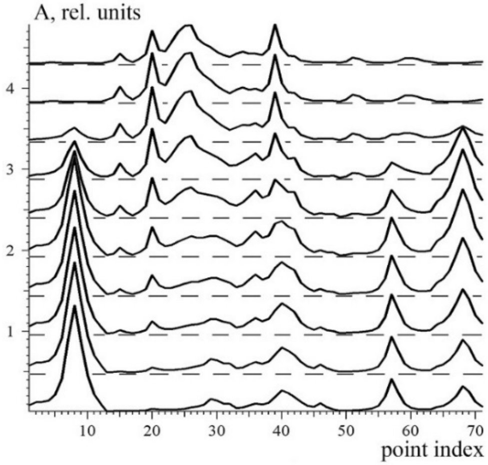

Figure 1.3. IR spectra of the 5-component mixture №1 after digitization and resampling. The spectral range is 1750-1110 cm^{-1}. The serial numbers of spectral points on the abscissa axis are given (since the wavelength value is not considered in the decomposition, the spectra are digitized with a variable step to exclude non-informative regions). The spectra are shifted vertically for better visualization, and the dashed lines indicate zero absorption levels.

The decoded spectra have been digitized with the fixed step of 5 cm^{-1}. *Figure 1.3* shows the typical mixture spectra **No. 1**, the five components: hexane, toluene, acetone, cyclohexanone, and diethyl ether. The 10-12 temporal spectra of fractions have been selected from each mixture. To verify the decomposition performance is desirable to get all the precise parameters asap. Based on equation (1.1), the concentration matrix *C* was calculated using the pure component spectra measured on the same spectrometer.

The second (**No. 2**) model was a mixture of the six (6) components: isooctane, toluene, benzene, diethyl ether, acetone, and amyl acetate in carbon tetrachloride as a nonpolar solvent. In this set, two pairs of components are relatively strongly overlapped: acetone – amyl acetate and benzene – toluene (see *Figure 1.13*). In addition, these substances are poorly separated under these chromatographic conditions. The set of absorption spectra is shown in *Figure 1.4*.

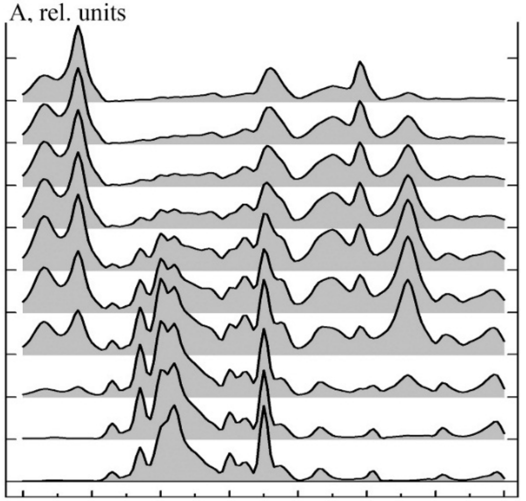

Figure 1.4. The IR spectra of fractions of the 6-component mixture (**№ 2**) of isooctane, toluene, benzene, diethyl ether, acetone, and amyl acetate. For other details, see the capture in *Figure 1.3*.

Model **No. 3** was composed of six (6) components: hexane, isooctane, benzene, dioxane, diethyl ether, and acetone. The mixture fractions (10 fractions) were obtained by distillation with a deflagrator. Determination of the number of components by matrix (*Figure 1.5d*) gave an estimate of 4 instead of 6.

Model **No. 4** was the 4-component mixture of acetone, diethyl ether, cyclohexanone, and toluene. The fractionation has been carried out using liquid chromatography. However, a worse component separation was obtained due to the larger column diameter (5 mm).

1.3.2. Analysis of singular values and residual matrices

As was described above, the preliminary estimate of the number of components is obtained by counting the number of singular numbers above the line drawn through the set of small $log(\lambda_k)$. By the break of the graph line in the transition from "noisy" $log(\lambda_k)$ to "significant" level, one can judge about nature of the spectroscopic information. Loosely speaking, one can perform the splitting into significant and noisy components more confidently since the decomposition procedure is better conditioned. *Figure 1.5* shows the singular values for several modeled and real data sets.

The second method, which can be efficient for estimating the number of components, is based on statistical criteria, which work not with the data values, but with their ranks, which make the estimates independent of the distribution law (Johnson, 1977). Rank criteria are invariant regarding any monotonic transformation of the measurement scale. One of the rank criteria was formulated by Mann-Whitney and Wilcoxon (Wilcoxon, 1945). Mann-Whitney's *U* test is a nonparametric statistical criterion to assess differences between two samples. The hypothesis stated that the distribution laws of the two independent random variable sets are identical can be used as the criterion of the number of components to test the residuals matrix $F_0(K)$. The latter is obtained by calculating the matrix $D^{\#}$ for different numbers K of the sample of the experimental error matrix (see equations (1.12)-(1.15)). As to the number K, it is assumed to be the minimum value when the hypothesis about the identity of the distribution laws is valid at the chosen level of significance. In other words, K is assumed to be equal to the minimum number of singular numbers, which corresponds to the residual matrix F_0 with elements having the same distribution law as the errors in D.

First, all n_1 elements of $F_0(K)$ and n_2 elements of the array of experimental errors are combined into one vector $\boldsymbol{x}$ in descending order. Then form a vector $\boldsymbol{y}$ from n_1 numbers representing order numbers of elements $F_0(K)$ in vector $\boldsymbol{x}$ and calculate the sum

$$S = \sum_{i=1}^{n_1} y_i \tag{1.33}$$

According to the tables (Johnson, 1977) for the given n1 and n2 and the selected significance level, we find the critical region (S_{min}, S_{max}). The value S belonging to the interval (S_{min}, S_{max}) means that the distribution laws are identical. For the IR spectra of 6-component mixtures shown in *Figure 1.4*, the S value for the 4-component hypothesis is 194, which is outside the interval (348, 472) at a 5% level of significance. For the 5-7 component hypotheses, the value S was equal to 429, 394, and 414, respectively. They may indicate the presence of errors associated with instrument drift and lead to the fact that the experimentally estimated noise vector contains a fraction of information about the spectra of the substances. Hence, the acceptable number of components is 5.

A slight modification of the Wilcoxon procedure makes it possible to create the criterion of equality of dispersions under the condition of equality of position characteristics for the two populations (Johnson, 1977). In our case, since one assumes the mean values of $F_0(K)$ and experimental errors to be zero, this criterion can be applied. The modification consists of changing the way of assigning ranks to elements of vector x. Instead of the sequence 1,2,3, ..., n_1+n_2, one constructs the following: 1,4,5,8,9, ..., 7,6,3,2 (this trick is to destroy autocorrelations in the sequence) and from this chain form the vector y of n1 ranks (ordinal numbers in x) for sample F0(K). Then, the number S is calculated, and the criterion is evaluated using statistical tables (Johnson, 1977). In the example under consideration, the values of S are 292, 362, and 440 for the 5, 6, and 7 component hypotheses. As is seen, the acceptable K is equal to 6, *i.e.*, the criterion of variance equality, in this case, was more effective.

It should be noted that it is better to apply both the above criteria together. The number K is chosen as a minimum estimation that is simultaneously valid for both approaches. Even the estimate of experimental errors is unknown, the number of components can be found from testing the hypothesis that supposes that the mean value of elements $F_0(K)$ is equal to zero.

The practice has shown that the analysis of the residuals matrix in the presence of autocorrelation gives less reliable results due to the influence of systematic measurement errors. The dependence of the autocorrelation criteria values on the estimated number of components under consideration has a significantly less pronounced kink than the plots of the logarithms of the singular numbers.

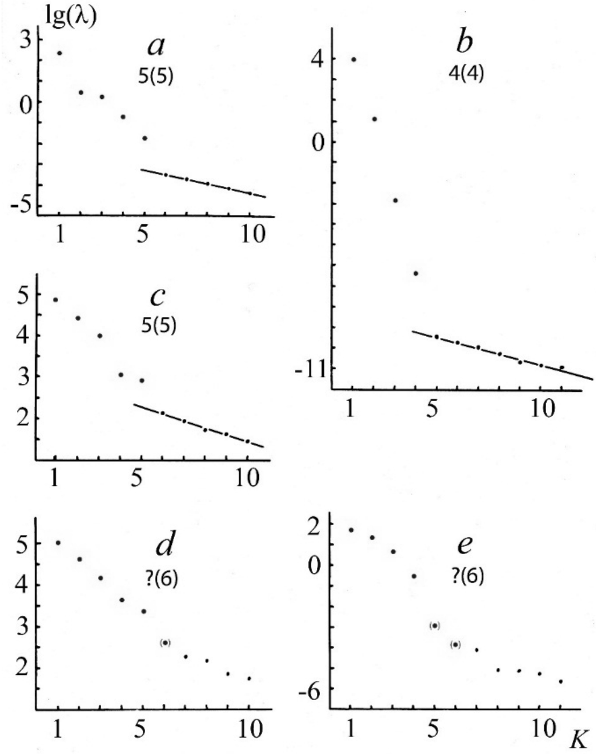

Figure 1.5. Logarithms of singular numbers of the IR spectra matrices of the test mixtures: a) the model data matrix at spectral noise level 0.2%T, b) mixture No.4, c) mixture No.1, d) mixture No.2, e) mixture No.3. Each plot shows the number of absorbing components found and the a priori known (in parentheses). The ordinal numbers of singular numbers are plotted on the abscissa. The question marks show that the nonlinear dependences occur at $R^2 > 0.7$, and the estimate is still a problem.

1.3.3. Analysis of singular vectors of the data matrix

The contributions of component spectra and spectroscopic noise to the singular vectors U of the data matrix are nonuniformly distributed over them. Let u_j be the jth column of the matrix U. u_1 represents an averaged mixture spectrum, which exhausts the maximum variance in the original matrix. By definition (1.8),

$$\boldsymbol{u}_j = \frac{1}{\lambda_j} D \cdot \boldsymbol{v}_j$$

$$\boldsymbol{v}_j = \frac{1}{\lambda_j} D^T \cdot \boldsymbol{u}_j \tag{1.34}$$

In addition, the DD^{T} and $D^{\mathrm{T}}D$ matrices are symmetric and non-negative (since $d_{ij} \geq 0$ – we consider the case of non-negative spectra). If there are no regions in the mixture spectra whose intensity is equal to zero for all the spectra, the DD^{T} and $D^{\mathrm{T}}D$ matrices are irreducible.

According to the Perron-Frobenius theorem (Lancaster, 1985), non-negative irreducible matrices have a positive eigenvalue λ_1

$$\lambda_1 = \sqrt{}\underset{x\neq 0}{max}\frac{|D\cdot D^T\cdot x|}{|x|} = \sqrt{}max\frac{|D^T\cdot D\cdot y|}{|y|}, \tag{1.35}$$

equal to the matrix spectral radius, and it refers to a positive eigenvector u_1 (and v_1, respectively).

As follows from (1.29), u_1 is the positive linear combination of the mixture spectra (matrix D columns), whereas v_1 is the positive linear combination of the matrix D rows.

If the matrix A_k of the smaller rank $k<M$ is obtained using the k first singular vectors

$$A_k = U \cdot \Lambda \cdot V^T, \tag{1.36}$$

the residual matrix $A - A_k$ has the minimal possible consistent norm for all $A - B_k$, where B_k is a matrix of rank k, i.e.

$$\left\| A - B_k \right\|_E = \min \qquad if \qquad B_k = A_k. \tag{1.37}$$

Consequently, when $k = 1$, the residual matrix after removing the 1-st singular vector will have the minimum possible element variance. It means that in u_1 or v_1 the signal-to-noise ratio will be maximum if the errors are uncorrelated. Similarly, u_2 and v_2 maximally exhaust the correlations in the residual matrix, and so on, until the vectors $\{u_K{+}1, ..., M\}$ and $\{v_K{+}1, ..., M\}$ are composed of elements whose non-zero values are due almost exclusively to the uncorrelated component in the raw data, in practice, to noise.

Figure 1.6 shows the left singular vectors calculated from a matrix of 10 mixture spectra (6-component mixture No.2, see 1.3.2) shown in *Figure 1.4*. It can be seen from this figure that the relative contribution of high-frequency components increases with increasing vector numbers. If the sampling interval in the data exceeds the maximum frequency in the spectra, the high-frequency part corresponds to the noise component.

An effective in-practice method for estimating the number of components is based on the statistical analysis of the sequence of elements in the left singular vectors. The condition for the applicability of this method is a sufficiently small step of spectra digitization. The latter ensures the correlation (monotonic change) of consecutive points in the non-noise spectra.

Consider now the analysis of autocorrelations in singular vectors, on the example of the set of 6-component IR mixture spectra (model **No. 2**, *Figure 1.4*). The left singular vectors are given in *Figure 1.6,* while *Figure 1.7* shows a plot of the logarithms of the singular numbers. As a result, one can see that the choice can be made between the 5 and 6 components. *Figure 1.6* is visually more informative: singular vectors 1 ÷ 6 describe the spectral data, while vectors 7 ÷ 10 have a noise origin.

Tables 1.1 and 1.2 show the statistical analysis results of the singular vectors using Wilcoxon and Durbin-Watson (1.21) tests. The results show that at the most commonly used significance level of 0.02, we should take the seven components instead of six ones as the number estimate.

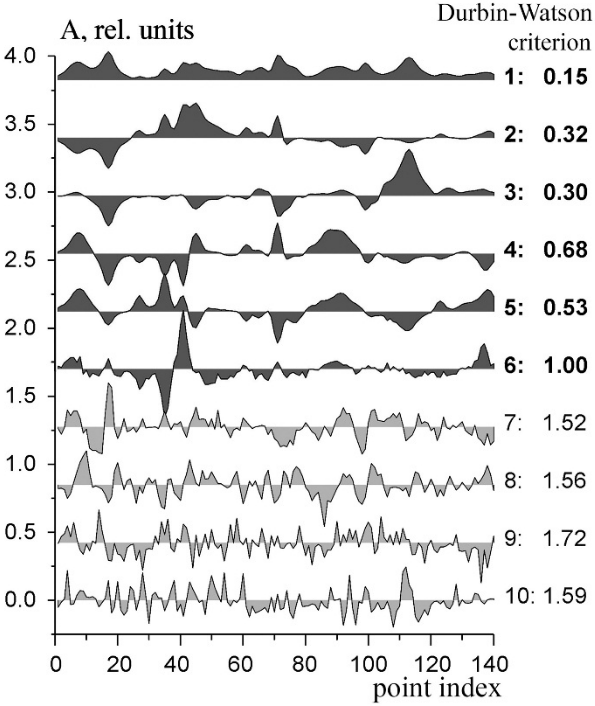

Figure 1.6. Left singular vectors u_j of the 6-component data matrix (*Figure 1.4*) presented as spectral contours. The contours are shifted vertically for better visualization, the zero-level lines are in the middle of the grey shading areas.

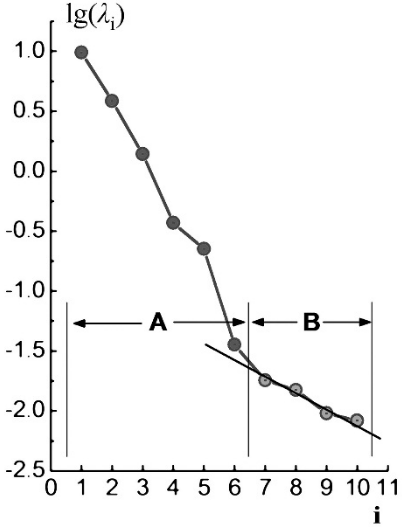

Figure 1.7. Plot of logarithms of singular numbers of the matrix of 6-component mixtures (model No. 2, see 1.3.2). The coefficient of determination R^2 for linear approximation of the right part of the graph is 0.89 > 0.7. Area A corresponds to significant numbers, B to the noise part.

Table 1.1. Values of the Wilcoxon and Durbin-Watson (1.21) criteria for autocorrelation of elements in the left singular vectors *U* calculated for the data set No.2.

Number of the singular vector	Wilcoxon test	The decision about the number of components (at significance level 0.05)	Durbin-Watson test (1.14)	The decision about the number of components (at significance level 0.05)
1	1.0	-	0.15	-
2	2.0	-	0.32	-
3	8.0	-	0.30	-
4	14.0	-	0.68	-
5	11.0	-	0.53	-
6	17.0	-	1.00	-
7	25.0	6 (0.002)	1.52	(6)
8	28.0	7 (0.02)	1.56	-
9	28.0	8 (0.02)	1.72	8
10	30.0	9 (0.05)	1.59	-

Table 1.2. Significance levels for the Wilcoxon test calculated for data set No.2.

Level of significance Q	0.05	0.02	0.01	0.002
Lower critical value	28.00	27.00	26.00	23.00
Upper critical value	43.00	45.00	46.00	48.00

The Durbin-Watson DW criterion (1.21) showed the overestimation of $K = 8$ due to weak correlations along the vectors due to molecular interactions of components. This criterion proved to be more sensitive to such violations of additivity. As stated earlier, the hypothesis of the presence of correlations is rejected if the DW value is outside the confidence interval {1.7-2.2} for the typical significance level $Q = 0.05$. However, if a confidence interval is enlarged to {1.5-2.4}, then the number 6 can be taken as an estimate of K.

The general conclusion for the example above is to estimate the number of components as 6 or 7. However, number 8 must also be considered when searching for individual spectra discussed later.

The Durbin-Watson criterion at $K = 8$ was rejected in this experiment because the contribution of the 8th singular vector in the data matrix can be neglected due to the relative smallness of the corresponding singular number (*Figure 1.7*).

In addition to the Durbin-Watson criterion, the nonparametric Wald-Wolfowitz series criterion (Wald, 1940) can be recommended. It allows one to obtain unbiased and consistent estimates for small sample sizes (parametric criteria usually require a size of 1000 at least) and *a priori* unknown distribution laws. The Wald-Wolfowitz statistic checks the randomness of alternation of positive (A) and negative (B) elements in the binary sequence AAABBBAAAAAABA (in our case, the signs of consecutive elements of singular vectors). Let N_A be the number of positive singular vector elements, N_B is the number of negative elements, and N_R is the number of groups of both.

The expectation and variance of the number of groups are equal to

$$M\{N_R\} = 1 + \frac{2 \cdot N_A \cdot N_B}{N_A + N_B}, \quad D\{N_R\} = \frac{2 \cdot N_A \cdot N_B \cdot (2 \cdot N_A \cdot N_B - N_A - N_B)}{(N_A + N_B)^2 \cdot (N_A + N_B - 1)}.$$

These values are used to test the null hypothesis that the sequence is not random.