ELSEVIER

Analytica Chimica Acta 379 (1999) 143–158

ANALYTICA CHIMICA ACTA

An automated procedure to predict the number of components in spectroscopic data

Abdalla Elbergali*, Jan Nygren, Mikael Kubista

Department of Biochemistry and Biophysics, Chalmers University of Technology, SE-413 90 Göteborg, Sweden

Received 21 April 1998; received in revised form 6 August 1998; accepted 4 September 1998

Abstract

We have compared various statistical methods to estimate the number of components that contribute to a set of spectra. The methods are tested both on simulated and on experimental data. No assumptions are made about noise level, since this in most experimental situations is unknown. For tests that formally require such information we have devised novel criteria for their predictions. The criteria have been integrated with the NIPALS algorithm to create a routine that in an automated way predicts the number of components. We find that the methods almost always predict the correct number of components when the quality of data is high. Also for multi-component samples and at high-noise levels most of these methods make satisfactory predictions. Those that gave the overall best results were the factor indicator function (IND) and the imbedded error function (IE). The *F*-test also worked well, but it has the disadvantage that a significance level must be chosen rather arbitrarily. The residual standard deviation (RSD), the root mean square (RMS), the χ-squared and the residual percentage variance (RPV) tests also gave satisfactory results. Less good were the eigenvalue (EV) and the reduced eigenvalue (REV). The ability of all indicators to predict the number of components was significantly improved when the degree of digitalization of the spectra was increased.

Keywords: Chemometrics; Indicator functions; Spectroscopic data

1. Introduction

Chemometric methods are today commonly used in spectral analysis of test samples. In general, such an analysis is made in two steps: first, the number of components is determined, and then the spectral responses and concentrations of the components are calculated. Several methods have been devised for the second step, and they widely depend on auxiliary information and experimental design [1–3]. The first step, determining the number of components, is pertinent to all forms of spectral analysis [4,5].

There are several approaches to determine the number of components that contribute to a given set of spectra. Here we only consider those based on pure principal component analysis (PCA) and not those combined with cross-validation [6–10]. All real data contain experimental noise that may mask the true dimensionality of a data set, which is the number *r* of components that are present. Malinowski [11,12] showed that the error associated with a data set can be divided into imbedded error and extracted error. Extracted error is the error contained within the minor

*Corresponding author. Tel.: +46-31-773-1000; fax: +46-31-773-3910.

0003-2670/99/$ – see front matter
PII: S0003-2670(98)00640-0

PC dimensions (e.g. $r+1$, $r+2$, ..., m) and can be removed, or extracted, from the data by retaining only the first r PCs. Imbedded error is the error which mixes into the factor scheme and is contained within the first r PCs. It can never be completely removed from a data set [13].

Methods to identify the true dimensionality of a data set can be classified into two categories: precise methods that are based on comparison with the experimental error, and approximate methods that do not require such information. Many of the latter methods are empirical functions. During the last decade also some methods based on formal statistical tests for this estimation have been developed. Here we extensively compare these methods on both simulated and experimental data. For tests that formally require knowledge about the experimental error associated with the data set, we have devised novel criteria for their predictions.

2. Theory

2.1. Principal components analysis

Most measurements are not selective for only the constituents of interest; but the data also contain noise. In principal component analysis (PCA) the measured data are reduced to contain only the information that is relevant to the system [14–18]. Its systematic variations are extracted and the information in the many variables is concentrated into a few underlying (latent) variables called principal components.

The first step in PCA is to decompose the data matrix $\mathbf{A}$ into an orthonormal basis set:

$$\mathbf{A}_{n,m} = \mathbf{T}_{n,q}\mathbf{P}'_{q,m} = \sum_{i=1}^{q} \mathbf{t}_i \mathbf{p}'_i, \tag{1}$$

where $\mathbf{A}_{n,m}$ contains the n recorded spectra as rows, each digitized into m data points, $\mathbf{T}_{n,q}$ is the score matrix which relates to sample composition, $\mathbf{P}'_{q,m}$, where the prime ‘′’ denotes transpose, is the loading matrix which relates to spectra and q is the least of n and m, which in spectroscopy usually is n. Eq. (1) is exact. The complete set of score and loading vectors accounts for both the systematic variations in the data and the experimental noise.

The second step in PCA is to separate the eigenvectors that account for the systematic variations from those corresponding to noise:

$$\mathbf{A} = \mathbf{T}_{n,r}\mathbf{P}'_{r,m} + \mathbf{E}_{n,m} = \sum_{i=1}^{r} \mathbf{t}_i \mathbf{p}'_i + \mathbf{E}_{n,m} = \hat{\mathbf{A}} + \mathbf{E}_{n,m}, \tag{2}$$

where $\hat{\mathbf{A}}$ is the predicted data matrix, $\mathbf{E}_{n,m}$ is the residual matrix, and r is the number of significant components. It corresponds to the number of compounds that contribute significantly to the measured spectra.

2.2. Methods to estimate the number of significant components (r)

Several methods have been proposed for the determination of the number of significant factors in PC decomposition. The most common ones are utilized in the following indicators.

2.2.1. Eigenvalues

Eigenvalues (EV or g) are conventionally used as a measure of the size of a PC [19]. The sum of squared elements of the data matrix $\mathbf{A}$ is equal to the sum of the eigenvalues of $\mathbf{A}'\mathbf{A}$. Each eigenvalue is proportional to the variance in the data that the corresponding principal component accounts for. For principal components that span only random noise, the corresponding eigenvalues should be small and roughly equal.

Eigenvalues can be calculated as the sum of squares of the score vectors,

$$\mathrm{EV}_l = g_l = \sum_{i=1}^{n} t_{li}^2 \quad (l = 1, 2, \ldots, r, \ldots, q). \tag{3}$$

The first set of r eigenvalues that contain useful information are called significant eigenvalues or primary eigenvalues. These have contributions from the real components and should be considerably larger than those containing only noise. The second set of $(q-r)$ eigenvalues are referred to as non-significant eigenvalues or secondary eigenvalues.

2.2.2. Reduced eigenvalues

Reduced eigenvalues (REVs) are normalized eigenvalues. An REV is defined as the eigenvalue divided by the degree of freedom employed in its extraction [20]:

$$\mathrm{REV}_l = g_l/((n-l+1)(m-l+1)). \quad (4)$$

2.2.3. Residual standard deviation

The residual standard deviation (RSD) is a measure of the lack of fit of a PC model to a data set [12]. It is defined as

$$\mathrm{RSD}(l) = \sqrt{\left[\frac{\sum_{j=l+1}^{q} g_j}{n(q-1)}\right]}, \quad (5)$$

where g_j is the eigenvalue as defined above, and q is the least of n and m.

2.2.4. Root mean square error

The root mean square (RMS) error of a data matrix is defined as

$$\mathrm{RMS}(l) = \sqrt{\frac{\sum_{i=1}^{n}\sum_{j=1}^{m}(x_{ij}-\hat{x}(l)_{ij})^2}{nm}}, \quad (6)$$

where

$$\hat{x}(l)_{ij} = \sum_{k=1}^{l} t_{ik}p'_{kj}. \quad (7)$$

The term within parentheses on the right-hand side in Eq. (6) represents the difference between the experimental data and the reproduced data from the l most significant PCs [12].

2.2.5. χ-Square

For data sets where the standard deviation varies from one data point to another Bartlett proposed the χ-square criterion [21]:

$$\chi^2(l) = \sum_{i=1}^{n}\sum_{j=1}^{m}\frac{(x_{ij}-\hat{x}(l)_{ij})^2}{\sigma_{ij}^2}, \quad (8)$$

where σ_{ij} is the standard deviation associated with measurement x_{ij}. The method is particularly useful when such information is available. Here we assume constant variance throughout the data, which allows us to factorize σ^2. We also normalize χ^2 to obtain the reduced χ-square $\chi_r^2(l)$:

$$\chi_r^2(l) = \sum_{i=1}^{n}\sum_{j=1}^{m}\frac{(x_{ij}-\hat{x}(l)_{ij})^2}{\sigma^2(n-l)(m-l)}. \quad (9)$$

The χ_r^2 formula is similar to the square of the RMS indicator. It differs only in the normalization.

2.2.6. Scree test

The scree test was proposed by Cattell [22], and is based on the observation that the residual variance levels off before the dimensions containing random error are included. The residual variance expressed as percentage, associated with a reproduced data matrix, is defined as

$$\mathrm{RPV}(l) = 100\left[\frac{\sum_{i=1}^{n}\sum_{j=1}^{m}(x_{ij}-\hat{x}(l)_{ij})^2}{\sum_{i=1}^{n}\sum_{j=1}^{m}x_{ij}^2}\right]. \quad (10)$$

The residual percentage variance (RPV) can also be expressed in eigenvalues,

$$\mathrm{RPV}(l) = 100\left[\frac{\sum_{j=l+1}^{q} g_j}{\sum_{j=1}^{q} g_j}\right]. \quad (11)$$

2.2.7. Imbedded error

The imbedded error (IE) function is an empirical function of the non-significant eigenvalues [11,12]:

$$\mathrm{IE}(l) = \sqrt{\left[\frac{l\left(\sum_{j=l+1}^{q} g_j\right)}{nm(q-l)}\right]}, \quad (12)$$

which is equivalent to $\sqrt{l/m}$ RSD (l).

2.2.8. Factor indicator function

The factor indicator (IND) function is an empirical function that has been claimed to be more sensitive than the IE function [11,12]:

$$\mathrm{IND}(l) = \frac{\sqrt{\left[\frac{\left(\sum_{j=l+1}^{q} g_j\right)}{n(q-l)}\right]}}{(q-l)^2}. \quad (13)$$

It is equivalent to RSD (l)/$(q-l)^2$.

2.2.9. F-test

A statistically more rigorous procedure to estimate r of a data matrix is based on the Fisher variance ratio test (F-test) [23,24]. By considering the statistical distribution of the non-significant reduced eigenvalues (REV), it is possible to determine whether a given REV is significantly larger than the mean of all

subsequent (higher rank) REVs. The calculated *F*-ratio can then be compared to the expected value for a particular significance level.

$$F_l(1, q-l) = \frac{\sum_{j=l+1}^{q}(n-j+1)(m-j+1)}{(n-l+1)(m-l+1)} \times \frac{\mathrm{REV}_l}{\sum_{j=l+1}^{q}\mathrm{REV}_j}. \quad (14)$$

The degrees of freedom have been suggested to be $n_1=1$ and $n_2=q-l$, which is equal to the expressions in the numerator and denominator, respectively [23].

3. Materials and methods

3.1. Simulation of data

The indicator functions were tested on several simulated data sets. The spectra were generated with Gaussian shapes and represented by m data points each.

$$v_{jk} = A_k \mathrm{e}^{-(j-c_k)^2/w_k^2} \quad (j = 1, 2, \ldots, m \text{ and } k = 1, 2, \ldots, r), \quad (15)$$

where A_k is the maximum intensity of the spectrum of component k and is found at wavelength c_k. w_k is proportional to the width of the spectrum. The parameters used in the simulations are listed in Table 1. The component concentrations were generated by the MATLAB random number generator, RAND('uniform'), from a uniform distribution in the interval 0–1. Random noise was added to the spectra by generating random numbers with a Gaussian distribution with mean 0 and standard deviations of 0.5, 0.8 and 1.2, respectively. These correspond to signal to noise (*S*/*N*) ratios of about 22, 14 and 10 (defined as the ratio between maximum signal and maximum noise).

Table 1
Parameters used to generate component spectra for the simulations

k	c_k	w_k
1	0.25	0.60
2	0.45	0.30
3	0.60	0.45
4	0.75	0.70
5	0.80	0.15

The parameters are normalized in the interval 0–1. For a data set of size m the parameters should be multiplied by m.

3.2. Experimental data

3.2.1. Absorption spectra

Data sets A and B were collected on mixtures of thiazole orange (TO) and calf thymus DNA. Since only TO absorbed light in the studied wavelength interval, the contributing components were free TO and TO bound to calf thymus DNA. In set A spectra were collected in the wavelength interval 400–600 nm (m=1001), and the amount of bound TO was varied by changing the salt concentration in the interval 0.01–0.5 M NaCl (n=15). In set B the amount of bound TO was varied by changing the temperature in the interval 10–60°C (n=10), and spectra were collected in the wavelength interval 375–500 nm (m=626). Data set C was collected on a sample containing only TO under conditions where TO forms dimers. Spectra were collected in the interval 260–600 nm (m=1701), and the amount of dimer was varied by changing the temperature (15–70°C, n=23).

Data set D was absorption spectra measured on fluorescein samples in the pH range 1–9. In this range four protolytic forms are present (r=4), each with a characteristic response [25]. Twenty-four samples (n=24) containing 14 μM fluorescein (purchased from Sigma) in 1 M NaCl and 50 mM buffer (phosphate buffer at pH>5 and citrate buffer at pH<5) were prepared and data were collected between 250 and 550 nm (m=1501).

3.2.2. Fluorescence spectra

Data set E were 15 (n=15) fluorescence emission spectra collected on the protein ΔOBP in the presence of single stranded DNA (dT_{65}) at different mixing ratios. Only the protein, which was present in free and bound state, had fluorescence (r=2). Spectra were recorded in the interval 300–450 nm (m=301).

Data sets F–I were fluorescence spectra measured on solutions containing 1,4-bis[5-Phenyl-2-oxazolyl]-benzene-2,2′-*p*-Phenylene-bis[5-phenyloxazole] (POPOP), anthracene and 9,10-diphenyl anthracene (purchased from Sigma) in *n*-hexane (r=3). Two of the sets (F and G) contained emission spectra collected between 380 and 500 nm (m=601) at 21 excitation

wavelengths (n=21) ranging from 270 to 370 nm, and two sets (H and I) contained excitation spectra measured between 270–370 nm (m=501) at 25 emission wavelengths (n=25) ranging from 380 to 500 nm. Four additional data sets (J–M) were collected accordingly on samples containing also dimethyl POPOP (r=4).

3.2.3. HPLC chromatogram

Data set N was a chromatogram of degraded chlorophyll into four major components (r=4) [26,27]. The mixture was analyzed by HPLC using a Waters 990-diode array equipped with a 600E multi-solvent delivery system and a C_{18} reversed phase column. The flow rate was 1 ml min^{-1} and the chromatogram was recorded between 0 and 50 min after injection. The peaks of interest eluted between 39.3 and 42.3 min, and were sampled every 2 s (n=76). Absorption was measured at every fifth wavelength between 350 and 800 nm (m=91).

4. Results

The number of components can be predicted from the indicator values by comparing them with the experimental error, i.e. using the noise level as a threshold. This is the common criterion to determine r [12]. However, in most experimental situations sufficient information about the noise is not available and such comparison cannot be made. We therefore propose to determine the point where l=r from the dependence of the indicators on the number of principal components (l) used to calculate them. Fig. 1 shows the indicators as a function of the number of PCs (l) for one of the simulated data sets with r=5,

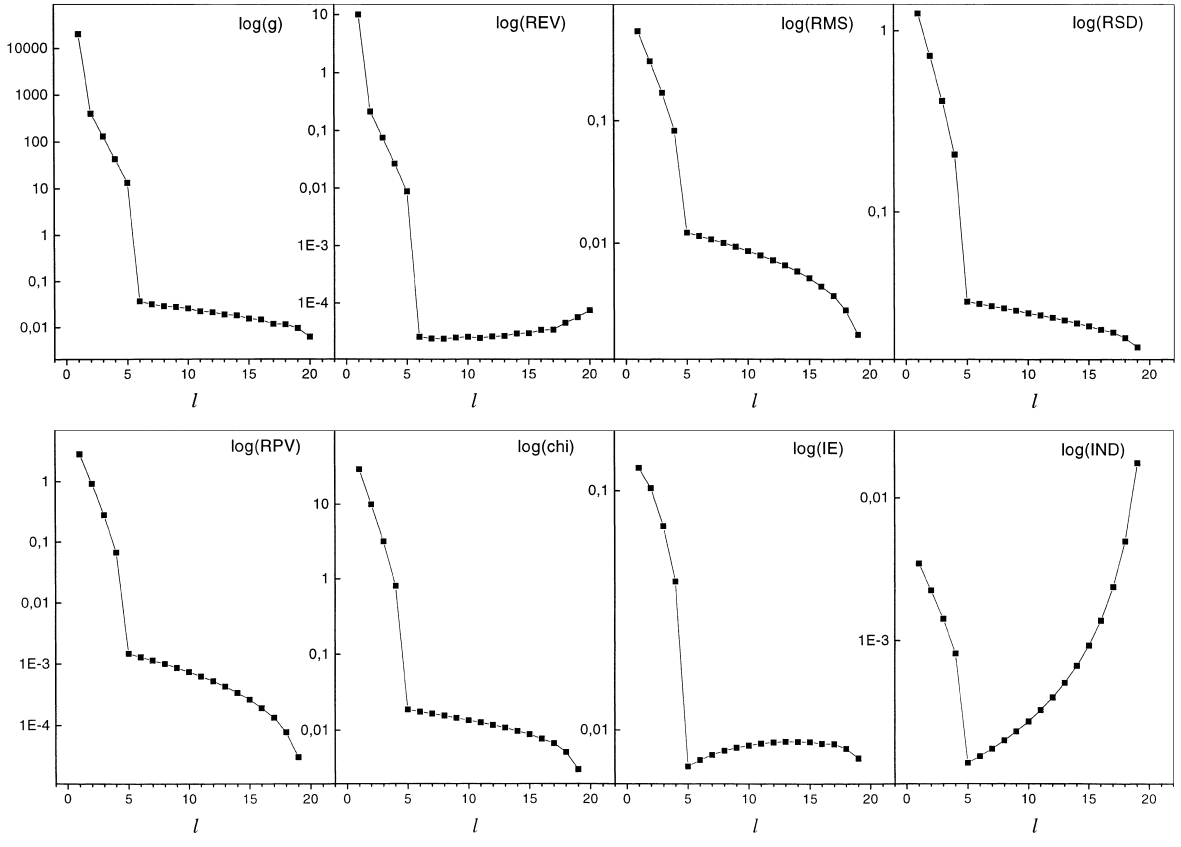

Fig. 1. Logarithm of the indicators as a function of the number of PCs (l) for a simulation with r=5; n=20; m=100.

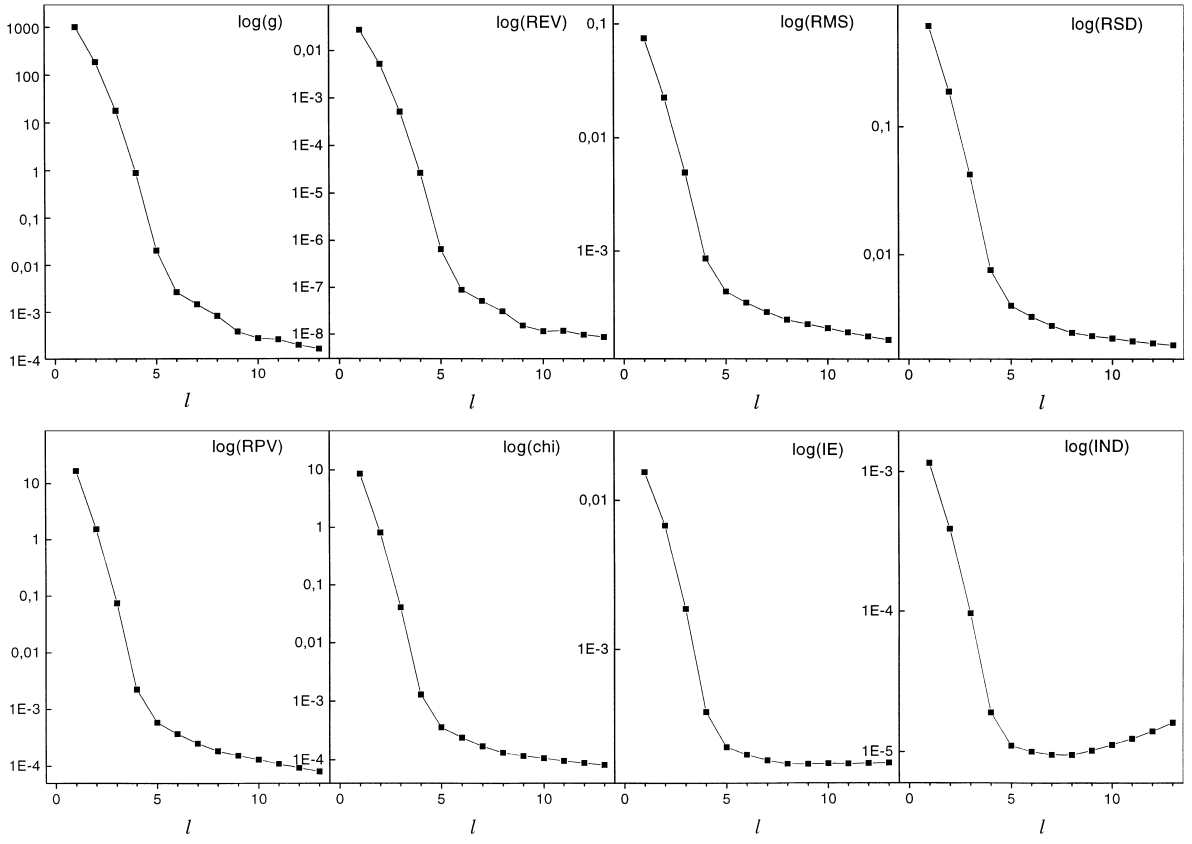

Fig. 2. Logarithm of the indicators as a function of the number of PCs (l) for the experimental absorption data set D.

n=20, m=100 and S/N=12, and Fig. 2 shows the same indicator functions for one of the absorption data sets (r=4, set D). Due to the large variations in the indicator values they are plotted on a logarithmic scale. We note that the indicators fall into two categories. The g_l and REV_l are functions of the lth PC, and should change substantially when $l+1=r$, while the other indicators reflect the cumulated effect of the first l PCs and should change when $l=r$. In these plots a change in slope can be seen around $l=r$ (r=5 for the simulated data and r=4 for the experimental data) for the RMS, RSD, RPV, χ-square, IE and IND indicators, and at $l=r+1$ for g and REV.

It is not trivial to identify where $l=r$ in the plots. Some indicators that are appropriately normalized may exhibit a minimum at this point [28,29]. This has previously been suggested to be used as criterion [11,28]. These indicators decrease as more primary eigenvectors are used in the data reproduction, and when the correct number of factors (r) is exhausted, and secondary eigenvectors are included, they start to increase. In our study we find that only the IE and IND indicators frequently exhibit a minimum at $l=r$. However, this seems to be the case only for simulated data, where the error is random and uniform throughout the entire data set. For experimental data, the indicators frequently exhibited minimum at a too high l resulting in an overestimation of the number of components. We therefore tested some other criteria to see if they worked better on experimental data. In general, they are all based on finding the point where the slope of the indicator function changes.

4.1. The second derivative criterion

The most obvious way to identify the point where the slope changes is to calculate the second derivative (SD) of the indicator function (INF). At this point SD

should have a maximum. Our simulations showed that the SD criterion made best predictions when using logarithmic scale, since the break in the slope was more easily detected when all indicators were of about the same magnitude. The SD was calculated as

$$\mathrm{SD}(l) = \log[\mathrm{INF}(l+1)] - 2 \times \log[\mathrm{INF}(l)] + \log[\mathrm{INF}(l-1)], \quad (16)$$

l=r should be at the first maximum of the SD (l) function. This criterion proved to be very accurate for simulated data. For experimental data, however, we found that the SD (l) function occasionally had about the same magnitude for two successive points around l=r. To determine which one was correct we found the following test reliable. The first maximum of the SD (l) function is calculated. The SD (l) function is then also evaluated at two successive points. If SD ($l_{\max}$+1) is closer to SD ($l_{\max}$) than to SD ($l_{\max}$+2), r=$l_{\max}$+1 is considered correct, while if SD ($l_{\max}$+1) is closer in magnitude to SD ($l_{\max}$+2), r=$l_{\max}$ is assumed. The approach is illustrated in Fig. 3(a). The SD $(l)^{\mathrm{IE}}$ has a maximum at l=4. Since (SD $(4)^{\mathrm{IE}}$−SD $(5)^{\mathrm{IE}}$)> ($\mathrm{SD}(5)^{\mathrm{IE}}$−$\mathrm{SD}(6)^{\mathrm{IE}}$), r should be 4.

4.2. The third derivative criterion

The abrupt change in the slope of the indicator function also leads to changes in the third derivative (TD). The TD value crosses zero and reaches a negative minimum, which can be used as a criterion. The TD is calculated as

$$\mathrm{TD}(l) = \log[\mathrm{INF}(l+2)] - 3 \times \log[\mathrm{INF}(l+1)] + 3 \times \log[\mathrm{INF}(l)] - \log[\mathrm{INF}(l-1)], \quad (17)$$

r should be equal to l where TD (l) has its first minimum. As seen in Fig. 3(b) TD has a sharp minimum at l=4.

4.3. The ratio of derivatives criterion

The change in slope can also be found by calculating the ratio of derivatives (ROD):

$$\mathrm{ROD}(l) = \frac{\mathrm{INF}(l-1) - \mathrm{INF}(l)}{\mathrm{INF}(l) - \mathrm{INF}(l+1)}. \quad (18)$$

Ideally ROD (l) should have a maximum at the point where l=r. It may have additional maxima at higher l,

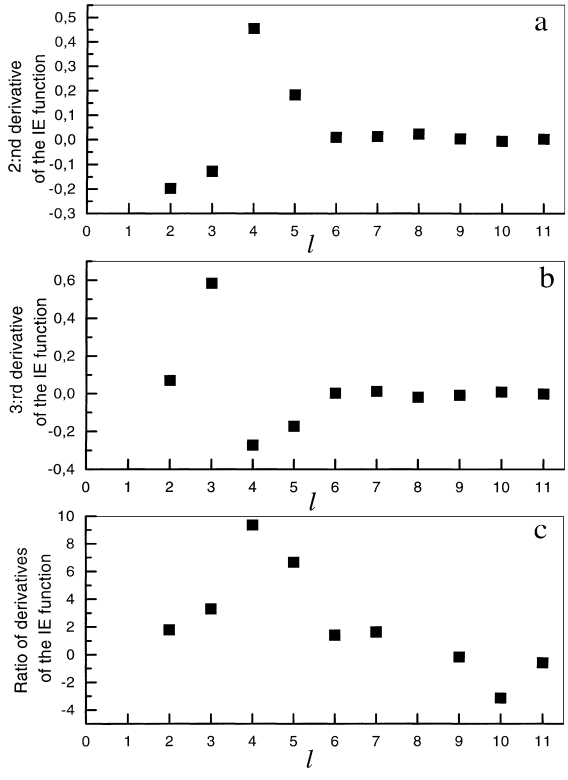

Fig. 3. The detection criteria for the IE function of data set D: (a) the second derivative (SD); (b) the third derivative (TD); (c) the ratio of derivatives (ROD).

caused by fluctuations in the indicator values when non-significant PCs are being included, but these are not relevant. Hence, the criterion is that the first ROD maximum determines r. As seen in the example in Fig. 3(c), ROD has a maximum at l=4.

The ROD criterion for the IE and IND indicators occasionally failed to predict r correctly. This happened when the indicator function itself exhibited a minimum, which resulted in negative ROD values. Combining the minimum and ROD criteria eliminated this problem. In this combined test, (RODM), r=l should be the point where either the indicator function has a minimum or ROD has a maximum.

4.4. F-test

The F-test is used differently from the other indicators. An F-value is calculated by Eq. (14) for a certain number of PCs and is compared with tabulated

values, which were generated using the MATLAB routine fdis.m from 'Numerical Recipes' [30]. The degree of freedom for the numerator (n_1) is taken as 1 and the degree of freedom for the denominator (n_2) is initially $q-1$ and is decreased by 1 for each REV [23]. One starts with the highest rank (q) and works down assuming that the REV will become insignificant at a particular significance level α. As long as the calculated F-value is larger than the tabulated one, significant components are included. Table 2 shows the result of the F-test for the same data as analyzed by the other indicators in Fig. 1. The calculated F-values for $1 \leq l \leq 5$ are much larger than the F-table values. For $l=6$ the calculated F-value is lower than that from the F-table. Hence, the F-test also predicts five components.

4.5. Comparison of the indicators on simulated data

The indicators were extensively tested on simulated data. The number of components (r), the number of samples (n), the S/N ratio, and the degree of digitalization (m) were varied. In initial tests we found five components to be suitable for rigorous comparison of the indicators, and chose $r=5$ for extensive studies.

Table 2
F-test results for the simulated data shown in Fig. 2

F-value	F-table (α=5%)	Degree of freedom n_2	Rank (l)
303.0251	4.38	19	1
17.7283	4.41	18	2
18.2197	4.45	17	3
21.9445	4.49	16	4
79.3660	**4.54**	**15**	**5**
0.9291	*4.60*	*14*	*6*

We also found $n=20$ to be a reasonable number of samples. At these conditions we performed simulations at different signal to noise ratios (S/N) using either $m=100$ or $m=1000$. For $S/N=22$ and 14 all indicators predicted correctly irrespective of criteria. Decreasing S/N to 10 provided very noisy spectra (Fig. 4(b)). As shown in Fig. 4(c), using $m=1000$ and $S/N=10$, all indicators predicted correctly in at least 81% of the cases, with the exception of the F-test which only predicted 64% of the cases correctly (Table 3). Best results were obtained when using the SD and TD criteria for the IE, IND, RSD, RMS, RPV and χ-square indicators which predicted correctly in at least 91% of the cases. The EV and REV were correct in less than 83%. When using the ROD criterion all indicators predicted correctly between 81 and 86% of the cases. Decreasing m to 100, IE, IND, RSD, RMS, RPV and χ-square still predicted correctly in over 70% of the cases, irrespectively of criterion, as shown in Table 3.

For the F-test the user must specify a significance level. In the tests above a significance level of 0.10 was used, which is the upper limit of what has been recommended [23]. This predicted correctly in 43% and 63% for $m=100$ and $m=1000$, respectively (Table 3). When we tested the effect of α, we found 100% success rate with $\alpha=0.30$ (Fig. 5). However, such a high significance level can hardly be motivated to use.

4.6. Analysis of experimental data

The indicators were used to analyze three types of experimental data, using the SD, TD and ROD criterion. For the IE and IND indicators also the minimum criterion (M) was tested [11,28].

Table 3
Simulated data results

Criterion	m	Indicators								
		EV (%)	REV (%)	IE (%)	IND (%)	RMS (%)	RSD (%)	RPV (%)	χ^2 (%)	F-test (%)
ROD	100	79	78	82	85	71	77	76	77	
	1000	86	84	85	86	81	83	84	84	
SD	100	74	77	91	92	87	88	87	86	43
	1000	82	82	95	96	92	93	91	91	64
TD	100	75	75	92	93	88	89	88	87	
	1000	83	82	95	96	92	93	91	91	

$S/N=10$; for the F-test $\alpha=0.10$.

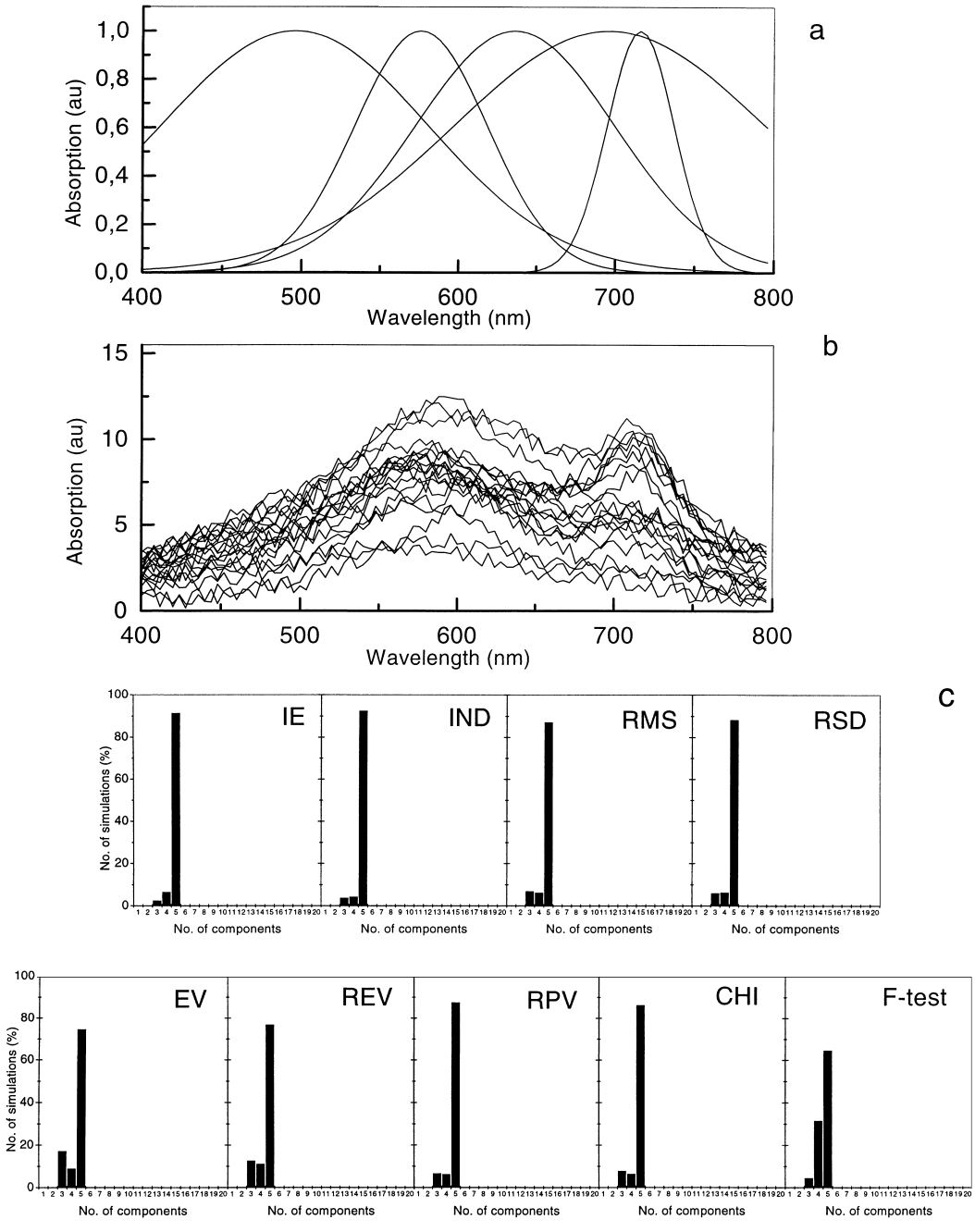

Fig. 4. (a) Simulated spectra of five components; (b) generated test spectra with S/N=10; (c) predicted number of components, expressed in percentage, by the nine indicators for 1000 simulations (m=1000, S/N=10 and SD criterion).

4.6.1. Absorption spectra

Four sets of absorption spectra (A–D) of a rather typical quality were analyzed. For the three data sets with two components (A–C), all indicators but the *F*-test estimated r correctly (Table 4). For set D, which has four components (Fig. 6), the IE, IND, RSD,

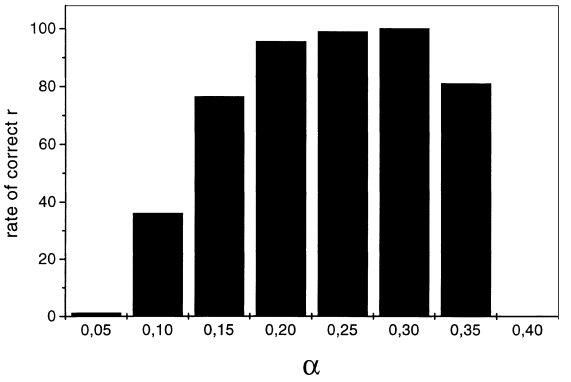

Fig. 5. Percentage successful predictions by the *F*-test at various significance levels for the simulated data with *S*/*N*=10.

RMS, RPV and χ-square indicators predicted correctly with any criterion. The EV and REV indicators predicted correctly when using the ROD criterion, but overestimated with the SD and TD criteria. The *F*-test overestimated in sets C and D with α=0.01, and overestimated substantially when α was set to 0.30. The minimum criterion worked only for the IND indicator in set B.

4.6.2. Fluorescence measurements

Nine fluorescence data sets (E–M) with 2, 3 and 4 components were analyzed. For the 2 and 3 component mixtures all indicators predicted the correct number of components. Even the minimum criterion worked for the 3 component mixtures with the IE indicator but failed with IND. The *F*-test worked with α=0.01 but failed with α=0.30. The data sets with four components (Fig. 7) are not very typical for fluorescence analysis but rather represent a very hard case. The EV and REV indicators did not predict correctly. The RMS, RSD, RPV and χ-square indicators predicted correctly for three of the data sets when using the SD and TD criteria and IE and IND predicted correctly also with the ROD criterion. For data set K, no indicator (except for the *F*-test) predicted correctly. This data set, however, has two components with very similar spectral shapes, out of which one is minor, and questionably significant. Only

Table 4
Experimental data results

Indicator	Criterion[a]	Absorption				Fluorescence									HPLC
		A	B	C	D	E	F	G	H	I	J	K	L	M	N
EV	SD	**2**	**2**	**2**	5	**2**	**3**	**3**	**3**	**3**	5	3	5	3	**4**
	ROD	**2**	**2**	**2**	**4**	**2**	**3**	**3**	**3**	**3**	3	3	**4**	3	**4**
REV	SD	**2**	**2**	**2**	5	**2**	**3**	**3**	**3**	**3**	5	3	5	3	**4**
	ROD	**2**	**2**	**2**	**4**	**2**	**3**	**3**	**3**	**3**	3	3	5	3	**4**
RMS	SD	**2**	**2**	**2**	**4**	**2**	**3**	**3**	**3**	**3**	**4**	3	**4**	**4**	**4**
	ROD	**2**	**2**	**2**	**4**	**2**	**3**	**3**	**3**	**3**	3	3	**4**	3	**4**
RSD	SD	2	**2**	**2**	**4**	**2**	**3**	**3**	**3**	**3**	**4**	3	**4**	**4**	**4**
	ROD	**2**	**2**	**2**	**4**	**2**	**3**	**3**	**3**	**3**	3	3	**4**	3	**4**
IE[b]	SD	**2**	**2**	**2**	**4**	**2**	**3**	**3**	**3**	**3**	**4**	3	**4**	**4**	**4**
	RODM	**2**	**2**	**2**	**4**	**2**	**3**	**3**	**3**	**3**	**4**	3	**4**	5	**4**
	M	4	4	6	9	**2**	**3**	**3**	**3**	**3**	5	5	5	6	–
IND[b]	SD	**2**	**2**	**2**	**4**	**2**	**3**	**3**	**3**	**3**	**4**	3	**4**	**4**	**4**
	RODM	**2**	**2**	**2**	**4**	**2**	**3**	**3**	**3**	**3**	**4**	3	**4**	**4**	**4**
	M	4	**2**	5	8	**2**	4	4	4	4	5	5	5	5	19
RPV	SD	**2**	**2**	**2**	**4**	**2**	**3**	**3**	**3**	**3**	**4**	3	**4**	**4**	**4**
	ROD	**2**	**2**	**2**	**4**	**2**	**3**	**3**	**3**	**3**	3	3	**4**	3	**4**
χ	SD	**2**	**2**	**2**	**4**	**2**	**3**	**3**	**3**	**3**	**4**	3	**4**	**4**	**4**
	ROD	**2**	**2**	**2**	**4**	**2**	**3**	**3**	**3**	**3**	3	3	**4**	3	**4**
f	α=0.01	**2**	**2**	4	5	**2**	**3**	**3**	**3**	**3**	**4**	**4**	**4**	**4**	9
	α=0.30	5	4	9	11	4	9	9	35	36	8	8	14	15	49

[a]The results using SD and TD criteria were identical for all experimental data sets.
[b]M is the minimum criterion and RODM is the combined ROD and M criterion.

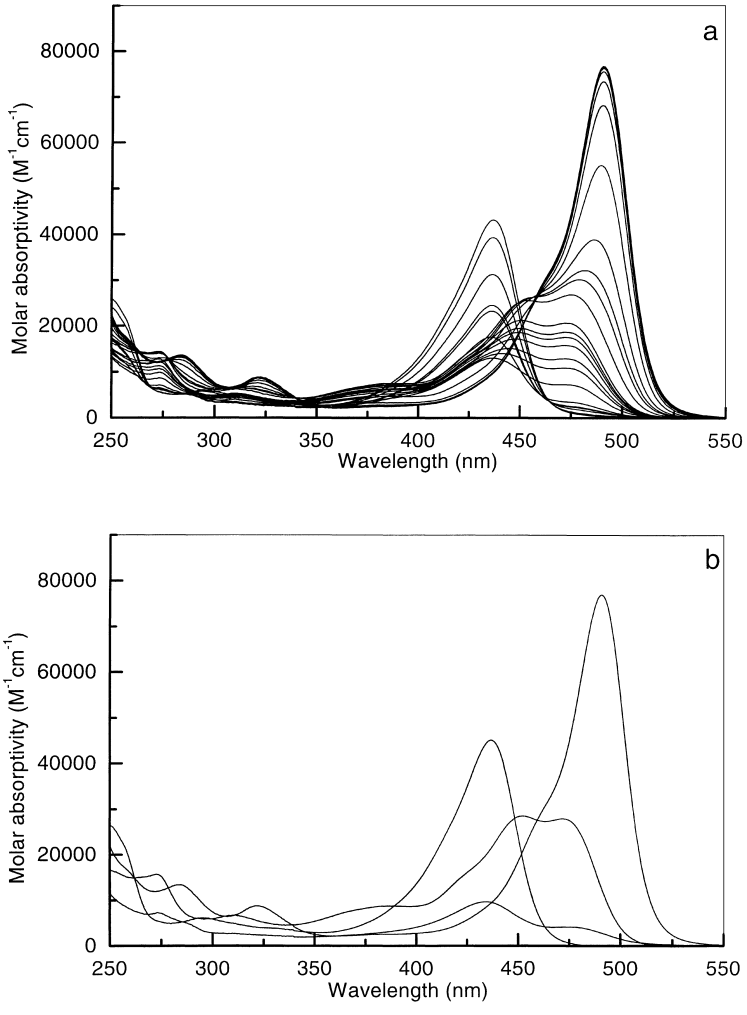

Fig. 6. (a) Absorption spectra of fluorescein at various pH (data set D); (b) absorption spectra of the four fluorescein species. The experimental data are taken from [27].

the *F*-test predicted correctly for all nine data sets when using a 1% significance level α=0.01). It, however, over-estimated all data sets when using α=0.30.

4.6.3. HPLC measurements

For the HPLC diode array data (set N, Fig. 8) all indicators but the *F*-test correctly predicted four components. For the IND and IE indicators the minimum criterion (M) did not work at all. IND displayed a minimum at l=19 and IE did not have any minimum at all (Table 4).

5. Discussion

Nine statistical indicators were tested for their ability to predict the number of components in spectral data. All are functions of the number of PCs (l) into which the data are decomposed, and from their dependence on l the number of components is determined. The indicators decrease steeply with increasing number of PCs as long as the PCs are significant. When these are exhausted the indicators fall off, some of them even display a minimum. At this point r=l for all indicators except g and REV for which r=l+1. The

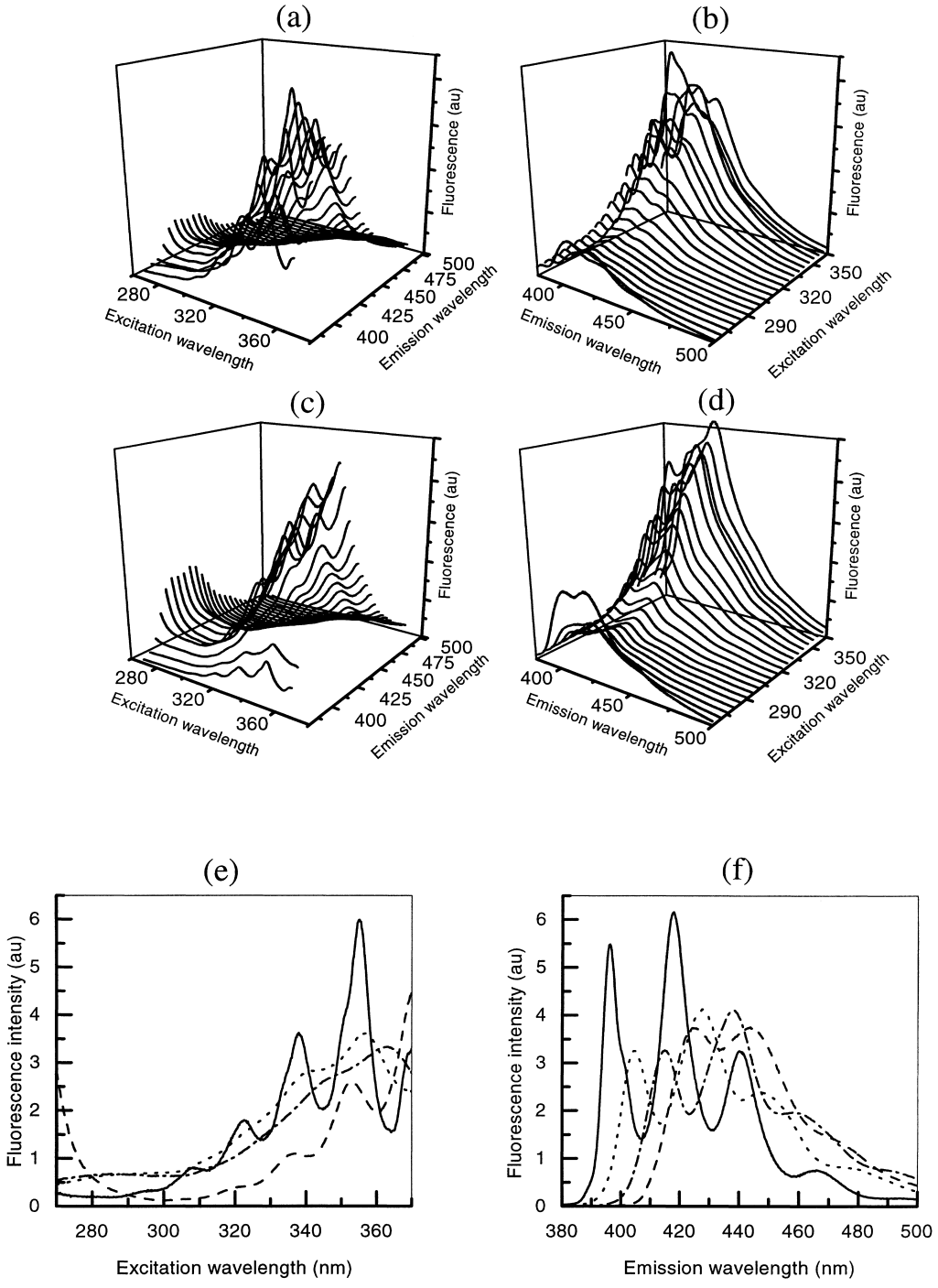

Fig. 7. Fluorescence spectra of samples containing anthracene, diphenyl-anthracene, POPOP and dimethyl-POPOP. (a), (c) and (e): Excitation recorded at different emission wavelengths; (b), (d) and (f): emission spectra recorded at different excitation wavelengths; (e) and (f): spectra of the pure components.

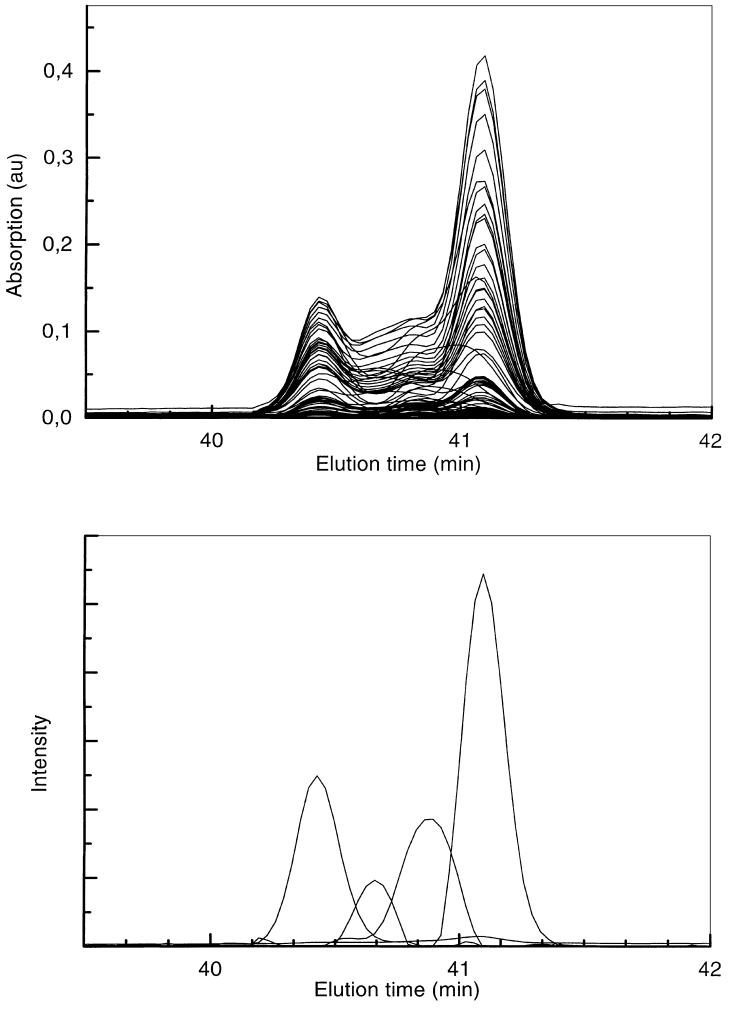

Fig. 8. (a) Chromatogram of degraded chlorophyll recorded by DAD-HPLC (data set N); (b) resolved chromatograms of the four components (normalized to the same peak height). The experimental data are taken from [25].

indicator values at this point can be predicted from the properties of the noise, which may be used as a criterion to determine r [12]. However, detailed information about noise is rarely available in experimental data, which urged us to introduce novel criteria. We tested the ratio of derivatives (ROD), the second derivative (SD) and the third derivative (TD) criteria, which all identify points where the indicator functions change slope. The number of significant components r should be equal to l at the first maximum of the ROD (l) and SD (l) functions, and at the first minimum of the TD (l) function. For the IE and IND indicators, these criteria were also compared with the minimum criterion to determine r, which was used both as an independent criterion and also in combination with the ROD criterion.

The ROD criterion was most effective for the EV and REV indicators that reflect the importance of the individual PCs. For the IE, IND, RMS, RSD, RPV and χ-square indicators, which reflect the cumulative effect of the first l PCs, the SD and TD criteria worked better. The testing occasionally turned out to be complicated by the first indicator value (at $l=1$) being much larger than the second, resulting in two components being predicted although the sample contained more. In these cases there was a second extreme

(minimum or maximum) point at $l=r$. These cases were readily recognized by visual inspection of either a plot of the indicator function or a plot of some of the indicator test functions (ROD (l), SD (l) or RD (l)). To avoid these rare mistakes we recommend backing up any automatic indicator test by giving the operator the option to plot and inspect the indicator functions.

5.1. The IND and IE indicators are the most reliable

Our extensive simulations and tests showed that most of the indicators accurately predict the number of components that contribute to a set of spectra. The indicators were applied to four absorption data, nine fluorescence data sets, one HPLC data set and a large number of simulated sets of data (Tables 3 and 4). For the experimental data sets having 2 or 3 components all indicators predicted correctly when using either of the SD, TD or ROD criteria that we have introduced. With the traditional minimum criterion, r was frequently overestimated. For the samples with four components (except data set K) all indicators but EV and REV predicted the correct number of components when using the SD or TD criterion. The IE and IND indicators predicted correctly also with the ROD criterion.

For the simulated spectra all indicators predicted highly accurate, all being absolutely correct for data sets with *S/N* ratios of at least 14. For data with higher noise the EV and REV indicators turned out least well. The RMS, RSD, RPV and χ-square indicators were somewhat better, and the IE and IND predicted best. The F-test predicted correctly for all sets of fluorescence data when using $\alpha=0.01$. This, however, is a completely different significance level than what worked best for the simulated data ($\alpha=0.30$). Since there is no simple way to tell a priori what significance level will work best for a particular set of data, this is a major problem. We therefore do not recommend the F-test unless the experimentalist has a way to choose optimum significance level. Hence, out of the indicators tested the IND and IE indicators make the most reliable predictions particularly in combination with the SD or TD criteria.

We note that from the IND and IE indicators the error (noise) in the data matrix can be estimated by calculating the residual standard deviation at the point where $l=r$:

$$\mathrm{RSD} = (q-r)^2\,\mathrm{IND}(r) = \sqrt{m/r}\,\mathrm{IE}(r). \tag{19}$$

5.2. Increasing the number of data points improves the indicator's predictability

Our simulations show that a higher number of data points collected in a given wavelength range may considerably improve the ability of the indicators to predict the number of contributing components (Table 3). Since computers today run most instruments, digitalization is rarely a problem. One should therefore take it as a habit to digitize the recorded spectra into the maximum feasible number of data points. This is particularly important for data sets with low *S/N* ratios and many components.

5.3. Combining the indicator tests with the NIPALS algorithm

There are several computational algorithms for PCA [12,31,32]. In spectroscopy one often uses the nonlinear iterative partial least squares (NIPALS), which is designed to extract the principal components (score vectors $\mathbf{t}_i$ and loading vectors $\mathbf{p}_i'$) directly from the data, in the order of decreasing significance [32]. The predicted data matrix $\hat{\mathbf{A}}(l)$, where l is the number of PCs used to calculate $\hat{\mathbf{A}}$, can be formed any time during the decomposition, and the process can be terminated when the desired accuracy is obtained. This considerably reduces the analysis time compared to non-iterative algorithms, and is particularly useful when the number of components is substantially lower than the number of samples ($r \ll n$). To know when to terminate the NIPALS iterations, NIPALS must be combined with the indicator test. This requires that the test can be performed without knowing all the PCs. From Eqs. (5), (11), (12) and (14) it follows that most indicators require evaluation of the sum of the remaining eigenvalues $\sum_{j=l+1}^{q} g_j$.

Since

$$\sum_{j=1}^{q} g_j = \sum_{i=1}^{n}\sum_{j=1}^{m} x_{ij}^2 \tag{20}$$

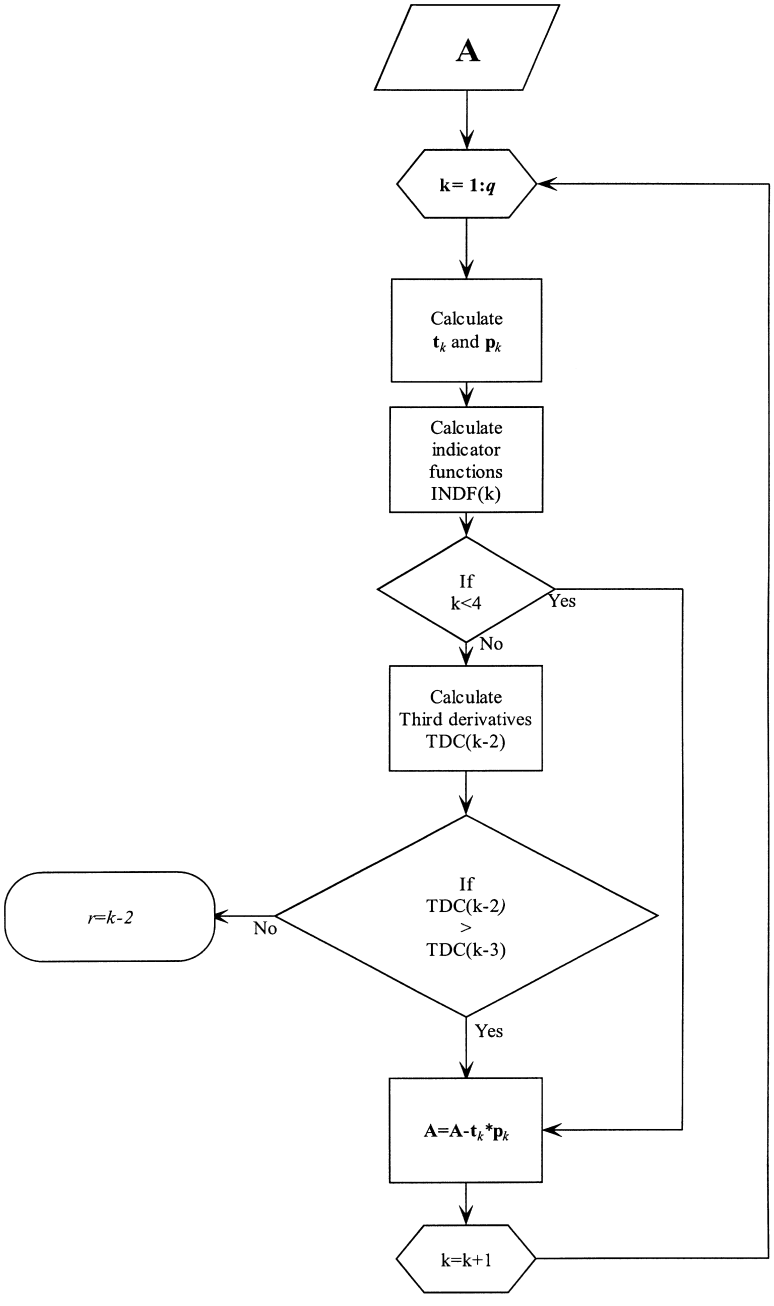

Fig. 9. Flow chart showing NIPALS combined with indicator test, using the third derivative criterion.

it follows that

$$\sum_{j=l+1}^{q} g_j = \sum_{i=1}^{n}\sum_{j=1}^{m} x_{ij}^2 - \sum_{j=1}^{l} g_j. \qquad (21)$$

The right-hand side expression can be calculated from the data matrix and the first l eigenvalues. Fig. 9 is a flow chart that shows how the indicator test, using the TD criterion, can be combined with NIPALS. After

each iteration, yielding a new pair of principal components, the indicator value is calculated and the TD(l) is evaluated. When the TD(l) function increases, it reflects a local minimum and NIPALS is terminated. The number of PCs that must be calculated for complete analysis is $r+3$.

6. Conclusion

We conclude that statistical methods can be used to accurately predict the number of components that contribute to spectral data sets even when information about experimental error is not available. The statistical methods can furthermore be combined with the NIPALS algorithm to constitute an automated prediction routine. The most reliable indicators under a wide range of conditions are the factor indicator function (IND) and the imbedded error function (IE). The number of components contributing to the spectral data is most accurately extracted from the indicator function as the first minimum of its third derivative. The degree of digitalization of the experimental data is important for successful analysis, and should always be as large as possible.

References

[1] M. Kubista, Chemom. Intell. Lab. Syst. 7 (1990) 273.
[2] M. Kubista, R. Sjöback, J. Nygren, Anal. Chim. Acta 302 (1995) 121.
[3] A.K. Elbergali, R.G. Brereton, Chemom. Intell. Lab. Syst. 23 (1994) 97.
[4] A.K. Elbergali, R.G. Brereton, Chemom. Intell. Lab. Syst. 27 (1995) 55.
[5] Y-Z. Liang, O. Kvalheim, A.M. Rahmani, R.G. Brereton, J. Chemom. 7 (1993) 15.
[6] M.J. Stone, J. Roy. Stat. Soc. B 36 (1974) 111.
[7] S. Wold, Technom. 20 (1978) 397.
[8] H.T. Eastment, W.J. Krzanowski, Technom. 24 (1982) 73.
[9] W.J. Krzanowski, Biometr. 43 (1987) 575.
[10] D.W. Stone, J. Chem. 2 (1988) 39.
[11] E.R. Malinowski, Anal. Chem. 49 (1977) 606.
[12] E.R. Malinowski, Factor Analysis in Chemistry, 2nd ed., Wiley, Chichester, 1991.
[13] S. Wold, C. Albano, W.J. Dunn, K. Esbensen, S. Hellberg, E. Johansson, M. Sjöström, Proceedings of the IUFOST Conference, Food Research and Data Analysis, Applied Science Publishers, London, 1983, p. 147.
[14] M.A. Sharaf, D.L. Illman, B.R. Kowalski, Chemometrics, Wiley, Chichester, 1986.
[15] H. Martens, T. Naes, Multivariate Calibration, Wiley, Chichester, 1989.
[16] D.R. Cox, D. Oakes, Analysis of Survival Data, Chapman and Hall, London, 1984.
[17] D.L. Massart, R.G. Brereton, R.E. Dessy, P.K. Hopke, C.H. Spiegelman, W. Wegscheider (Eds.), Chemometrics Tutorials, Elsevier, Amsterdam, 1990.
[18] D.L. Massart, W. Wegscheider, B.G. Vandeginste, S.N. Deming, Y. Michotte, L. Kaufman, Chemometrics: A Textbook, Elsevier, Amsterdam, 1990.
[19] T.M. Rossi, I.M. Warner, Anal. Chem. 54 (1986) 810.
[20] E.R. Malinowski, J. Chemom. 1 (1987) 33.
[21] M.S. Bartlett, Brit. J. Psych. Stat. Sec. 3 (1950) 77.
[22] R.D. Catell, Multi. Beha. Rese 1 (1966) 245.
[23] E.R. Malinowski, J. Chemom. 3 (1988) 49.
[24] K. Faber, B.R. Kowalski, Anal. Chim. Acta. 337 (1997) 57.
[25] R. Sjöback, J. Nygren, M. Kubista, Spectro. Acta Part A 51 (1995) L7.
[26] A.K. Elbergali, R.G. Brereton, A. Rahmani, Analyst 120 (1995) 2207.
[27] A.K. Elbergali, R.G. Brereton, A. Rahmani, Analyst 121 (1996) 585.
[28] R.E. Malinowski, Anal. Chem. 49 (1977) 612.
[29] I. Scarminio, M. Kubista, Anal. Chem. 65 (1993) 409.
[30] W.H. Press, B.P. Flannery, S.A. Teukolsky, W.T. Vetterling, Numerical Recipes in C, Cambridge University Press, Cambridge, 1988.
[31] SAS Institute Inc., SAS/STAT User's Guide, Version 6, 4th ed., vol. 2, 1989.
[32] S. Wold, K. Esbensen, P. Geladi, Chemom. Intell. Lab. Syst. 2 (1987) 37.