ELSEVIER

Analytica Chimica Acta 296 (1994) 1–20

ANALYTICA CHIMICA ACTA

Aspects of pseudorank estimation methods based on an estimate of the size of the measurement error

N.M. Faber *, L.M.C. Buydens, G. Kateman

Department of Analytical Chemistry, University of Nijmegen, Toernooiveld 1, 6525 ED Nijmegen, Netherlands

Received 2nd March 1994

Abstract

The estimation of the pseudorank of a matrix, i.e., the rank of a matrix in the absence of measurement error, is a major problem in multivariate data analysis. In the practice of analytical chemistry it is often even the only problem. An important example is the determination of the purity of a chromatographic peak. In this paper we discuss three pseudorank estimation methods that make use of prior knowledge about the size of the measurement error. The first method (Method A) is based on the standard errors in the diagonal elements of the row-echelon form of the matrix, the second method (Method B) is based on the eigenvalues of principal component analysis (PCA) and the third method (a *t*-test) is based on the singular values. Methods A and B are modifications of methods that are well known in analytical chemistry. However, these methods cannot provide significance levels for the estimated pseudorank. This holds for the original methods as well as the present modifications. The main reason for introducing these modifications is that in this way relationships are established between the *t*-test and methods that are already known. The aspects that are covered in this paper include the sampling distribution of the test statistic, the number of degrees of freedom to be used in the test, the adequacy of theoretical predictions and the bias that results from random measurement noise. The object of this paper is to demonstrate that using prior knowledge about the size of the measurement error may yield powerful pseudorank estimation methods. This is illustrated by comparing the significance levels obtained by the *t*-test and Malinowski's *F*-test. The *t*-test yields sharper significance levels for experimental data obtained from the literature as well as simulated data. This can be satisfactorily explained by the larger number of degrees of freedom that is employed in this test. The viability of the new *t*-test is supported by a thorough evaluation of the test data by a large number of conventional methods. As a remarkable by-product of the present investigation we find that a plot of the singular values yields a promising graphical pseudorank estimation method. Graphical methods have proved their use in the past in the case that the size of the measurement error is unknown. This new graphical method therefore provides a natural complement to the *t*-test.

Keywords: Principal component analysis; Pseudorank estimation

* Corresponding author.

SSDI 0003-2670(94)00201-V

1. Introduction

The estimation of the pseudorank of a matrix, i.e., the rank of a matrix in the absence of measurement error, is a major problem in multivariate data analysis. It typically arises when highly redundant data produced by modern analytical instruments are to be compressed to a more relevant format. This problem is not likely to be solved in the near future by a better instrumentation since for many applications the main interest is focused on very fast data acquisition, e.g., for the on-line monitoring of an industrial process. Therefore many methods have been proposed over the years to tackle this problem mathematically and the field of analytical chemistry has been especially fertile in producing such methods [1].

A useful and general classification of pseudorank estimation methods is based on the required prior knowledge about size and/or distribution of the measurement error. Methods that require such input are called *parametric* whereas methods that work without this prior knowledge are called *non-parametric*. Parametric methods are only reliable if it is safe to make the necessary assumptions concerning the noise. Otherwise they will not work. Interest in these methods has greatly faded with the introduction of Malinowski's methods based on error functions [2] and Wold's cross-validation [3]. Both methods are non-parametric and make use of a very popular method for the compression and subsequent analysis of multivariate data, i.e., principal component analysis (PCA).

Characteristic for these and many other methods in extensive use today is their inability to establish a *significance level* for the estimated pseudorank. A notable exception is Malinowski's *F*-test [4,5] which is recently developed from the concept of reduced eigenvalues [6]. It is a parametric method that however works without knowledge about the size of the measurement error. The only assumption being implicitly made is that the residuals are Gaussian distributed. (However, from classical analysis of variance it is well known that even this assumption is not necessarily restrictive [7].) This method is therefore essentially different from a number of very recently introduced methods for which significance levels can be obtained from simulation studies or resampling methods. Examples are canonical correlation [8] which may also be applied if only one data matrix is available [9], an algebraic method based on the Wronskian determinant [10], a non-parametric method based on resampling the zero-component region in the data matrix [11] and a non-parametric method based on the residuals of consecutive PC models [12].

In this paper we will discuss three parametric pseudorank estimation methods that explicitly require knowledge about the size of the measurement error. The first method (Method A) is based on the standard errors in the *diagonal elements of the row-echelon form* of a matrix [13]. The original formulation of this method [13] allows for a complication that critically depends on the nature of the test data matrix, i.e., the data matrix under consideration. Consequently we will propose a possible solution to this problem. The second method (Method B) constitutes a modification of the method of Hugus and El-Awady [14,1]. In the method of Hugus and El-Awady the *eigenvalues* of PCA of the test data matrix are compared to their standard error. In the current modification the eigenvalues of the test data matrix are corrected before they are compared to their standard error. The correction emulates the value that should be expected if an eigenvalue were caused by measurement error and is obtained as the dominant eigenvalue of a 'reference' matrix. In a previous paper we showed that an appropriately sized random matrix can be used as a suitable reference matrix [15]. Furthermore, we make use of the previously derived standard errors in the eigenvalues of PCA [16]. The third method (a *t*-test) is based on results obtained by Goodman and Haberman [17] for the *singular values* of a matrix. This method comes down to comparing the singular values of the test data matrix to the associated reference value in a similar way as Method B. It will be shown that in all these methods the data matrix is reduced to a 'canonical' form that is suitable for revealing the pseudorank. However, only the *t*-test is able to give *significance levels* for the estimated pseudorank.

The main reason for also introducing and discussing methods A and B is that in this way a relationship is established between the *t*-test and methods that are already introduced in analytical chemistry. This should lead to an improved understanding of the working of the proposed *t*-test and parametric methods in general. The aspects that are covered in this paper include the sampling distribution of the test statistic, the number of degrees of freedom to be used in the test, the adequacy of theoretical predictions and the bias that results from random measurement noise.

It is to be expected that prior knowledge about the size of the measurement error should yield a method that gives sharper significance levels than a method that does not use this extra knowledge. This will be illustrated by comparing the significance levels obtained by the *t*-test and Malinowski's *F*-test for literature data as well as simulated data. It is found that Malinowski's *F*-test gives rather conservative confidence levels. This can be explained by the small number of degrees of freedom employed in this test. The number of degrees of freedom associated with PCA will be further discussed in the Appendix. Support for the viability of the new *t*-test may also come from a thorough evaluation of the test data by a large number of methods that are currently in use in analytical chemistry. Thus we will also pay considerable attention to the background of these conventional methods.

In the remaining part of this paper we will assume that no data preprocessing has taken place and that the data matrix under consideration is open. Data preprocessing is absolutely necessary if the measurement error is heteroscedastic [18]. The consequences of closure and mean centering for the estimated rank have recently been discussed by Pell et al. [19]. Finally we will assume that pure data for the individual contributing sources are not available. Otherwise the Kalman filter (KF) approach developed at our laboratory [20] provides an excellent parametric pseudorank estimation method.

The following notation will be adapted throughout this paper. Bold upper-case letters will denote matrices, e.g., **M**. Bold lower-case letters will denote column vectors, e.g., **u**. Matrix and vector transposition are indicated by a superior 'T', e.g., $\mathbf{M}^T$ and $\mathbf{u}^T$. Italic letters (upper-case as well as lower-case) will denote scalars, e.g., M_{ij} is the element in row i and column j of M. The elements of diagonal matrices, e.g., Θ_{nn} and Λ_{nn}, are denoted by lower case letters, e.g., θ_n and λ_n, where the index indicates the position on the diagonal.

2. Theory

All parametric methods discussed in this section except Malinowski's *F*-test have in common that they are based on standard errors derived by the method of error propagation. Thus before introducing the parametric methods we will outline the principle behind the derivations. Since the parametric methods to be discussed primarily rely on a dependable estimate of the measurement error in the data matrix **M**, $\sigma(\mathbf{M})$, we will also briefly discuss how it could be estimated in practice.

2.1. The method of error propagation

The method of error propagation deals with the way in which uncertainties are carried over or propagated from the data points to the estimated parameters. The parameters are written as a function of the data and this function is approximated by a truncated Taylor expansion. The function is expanded around the errorless values and truncation usually proceeds after the linear or quadratic term. It follows that the function should be differentiable in a sufficiently small neighbourhood of the errorless values. The method works well if the measured data points are unbiased estimates of the true data points and the errors are small. The characteristics and limitations of this method are discussed in detail by Moran and Kowalski [21]. We emphasize that for the methods described in this paper the error propagation is carried out to first-order. As a result the derived standard errors can only be expected to be accurate if the standard deviation of the measurement error is small.

2.2. Estimation of the standard deviation of the measurement error in a data matrix

It is evident that in general one should use all the information available in order to ensure that the estimate of the size of the error is accurate. For example, Bubert and Jenett [22] recommend to extend the data matrix obtained from Auger electron spectrometry with sputter cycles in the lower and higher energy regions where no lines can be detected. Essentially the same approach can be followed for chromatographic data. Here the zero-component regions should provide the necessary information [11].

2.3. Pseudorank estimation method based on the standard errors in the diagonal elements of the row-echelon form of the data matrix (Method A)

The oldest methods developed in (analytical) chemistry for pseudorank estimation are based on mathematical definitions of matrix rank in terms of the largest non-zero submatrix [23–25] or the number of non-zero rows in the row-echelon form of the matrix [13]. In this paper we will only consider the method of Wallace and Katz [13] although it should be clear that the main disadvantage connected with the other submatrix methods (excessive computing time) is greatly alleviated by the use of modern high-speed computers.

The method of Wallace and Katz – in the sequel referred to as method A – consists of setting up, in addition to the data matrix **M**, a companion matrix **E**, whose elements E_{ij} are the estimated error of M_{ij}. **M** is reduced to row-echelon form by Gaussian elimination [26] with *complete* pivoting:

$$M'_{ij} = M_{ij} - \frac{M_{i1}}{M_{11}} M_{1j} \tag{1}$$

The elements of **E** are transformed during the reduction of **M** by computing new values based on the propagation of errors in **M**:

$$E'_{ij} = \left[E_{ij}^2 + E_{1j}^2 \left(\frac{M_{i1}}{M_{11}} \right)^2 + E_{i1}^2 \left(\frac{M_{ij}}{M_{11}} \right)^2 + E_{11}^2 \left(\frac{M_{i1} M_{1j}}{M_{11}^2} \right)^2 \right]^{1/2} \tag{2}$$

Complete pivoting is used in order to minimize the rate of propagation of errors. The pseudorank is now determined by investigating for each dimension the following ratio

$$\rho_n(\varepsilon) = \left| \frac{M'_{nn}}{E'_{nn}} \right| \tag{3}$$

Thus given a good estimate of the amount of measurement error it can be established whether a row is zero in the statistical sense by examining the diagonal elements of the transformed matrices **M'** and **E'**. Two decision rules have been found in the literature. Wallace and Katz consider a diagonal element of the reduced data matrix to be significant if it is three times its estimated error whereas Halket [27] proposes a ratio of one.

There is a complication not accommodated for by the simultaneous transformation of the companion matrix. This complication is best illustrated by the following example given by Golub and Van Loan [26]

$$\begin{pmatrix} 1 & -1 & -1 & -1 & . & . \\ 0 & 1 & -1 & -1 & . & . \\ 0 & 0 & 1 & -1 & . & . \\ 0 & 0 & 0 & 1 & . & . \\ . & . & . & . & . & . \end{pmatrix}$$

Suppose for the sake of simplicity that this matrix is the data matrix we start with. Then application of the decision rule that the ratio should exceed, say, k would immediately lead to the conclusion that **M** is full rank if the estimated standard deviation of the measurement error is less than $1/k$. However applying the matrix to the vector whose elements are 1, 1/2, 1/4, 1/8,... shows that the columns of the matrix are nearly dependant because this weighted sum of the

columns is nearly zero. The matrix is very ill-conditioned and a much smaller perturbation then expected may cause the resulting matrix to be singular. As a result we need an independent test for invertibility. Such a test is easily constructed by using a well known result from numerical analysis that says that the smallest singular value of **M** is the L_2-norm distance of **M** to the set of all rank-deficient matrices [26]. Therefore we propose to compute, in addition to the ratios of Eq. 3, the following index τ:

$$\tau_n = \mathrm{cond}_2(\mathbf{M}'_n)\frac{\|\mathbf{E}'_n\|_2}{\|\mathbf{M}'_n\|_2} \tag{4}$$

where $\mathrm{cond}_2(\bullet)$ is the L_2-condition number, $\|\bullet\|_2$ is the L_2-norm and $\mathbf{M}'_n$ and $\mathbf{E}'_n$ denote the $n \times n$ leading principal submatrix of $\mathbf{M}'$ and $\mathbf{E}'$ respectively. The pseudo rank of $\mathbf{M}'$ should be at least n if $\tau_n < 1$.

2.4. *Pseudorank estimation method based on the standard errors in the eigenvalues of PCA (Method B)*

PCA is a method that finds new (orthogonal) base vectors that span the space of the matrix in an optimal way. The new base vectors are constructed in such a way as to explain successively the maximum amount of variation in the data. According to Malinowski [1] the data matrix can be reconstructed using only the significant dimensions found by PCA. The remaining dimensions will only contain measurement error. In the terminology of Malinowski the significant dimensions are denoted as *primary* and the remaining ones as *secondary*.

PCA is directly related to the singular value decomposition (SVD) of **M**. Let s be equal to r or c whichever is smaller. The SVD decomposes **M** into a product of three matrices:

$$\mathbf{M} = \mathbf{U}\boldsymbol{\Theta}\mathbf{V}^{\mathrm{T}} \tag{5}$$

where **U** is an $r \times s$ orthogonal matrix whose columns $\mathbf{u}_n$ are the left singular vectors, **V** is an $s \times c$ orthogonal matrix whose columns $\mathbf{v}_n$ are the right singular vectors and $\boldsymbol{\Theta}$ is an $s \times s$ diagonal matrix with elements $\theta_1 \geq \theta_2 \geq \ldots \geq \theta_s$. These elements, the singular values, are the (positive) square roots of the eigenvalues λ_n of the cross-product matrices $\mathbf{MM}^{\mathrm{T}}$ and $\mathbf{M}^{\mathrm{T}}\mathbf{M}$:

$$\lambda_n = \mathbf{u}_n^{\mathrm{T}}(\mathbf{MM}^{\mathrm{T}})\mathbf{u}_n = \mathbf{v}_n^{\mathrm{T}}(\mathbf{M}^{\mathrm{T}}\mathbf{M})\mathbf{v}_n \tag{6}$$

The singular vectors $\mathbf{u}_n$ and $\mathbf{v}_n$ are seen to be eigenvectors of the cross-product matrices. The eigenvalue decompositions of $\mathbf{MM}^{\mathrm{T}}$ and $\mathbf{M}^{\mathrm{T}}\mathbf{M}$ are often referred to as *Q*-mode and *R*-mode PCA respectively.

Hugus and El-Awady [14] have developed a test based on the following expression for the standard errors (*R*-mode analysis):

$$\sigma(\lambda_n) = \left(\sum_{k=1}^{c} V_{kn}^2 \sum_{l=1}^{c} V_{ln}^2 \sum_{i=1}^{r} \left(M_{ik}^2 \sigma(M_{il})^2 + M_{il}^2 \sigma(M_{ik})^2 \right)(1 + \delta_{kl}) \right)^{1/2} \tag{7}$$

Here, δ_{kl} is the well-known Kronecker delta. In their test a PC is considered to be significant if the associated eigenvalue is larger than its standard error. (It should be noted that Bubert and Jenett [22] employ a critical ratio of three.) In a previous paper [16] we showed that the expression of Hugus and El-Awady is incorrect and should be replaced by

$$\sigma(\lambda_n) = 2\lambda_n^{1/2}\sigma(\mathbf{M}) \tag{8}$$

Furthermore, in the test of Hugus and El-Awady the eigenvalues are directly compared to their standard error. The underlying assumption is that an eigenvalue resulting from measurement error should be zero. However, measurement error also contributes to the variation in the data and one should try to take this fact into account. Thus we propose to test the eigenvalues for significance after correcting them for the value that could be the result of chance effect alone. In a previous publication [15] we showed that the secondary eigenvalues of the test data matrix can very well be estimated by the eigenvalues of a pure random matrix. This random matrix should preferably have the same *number of degrees of freedom* as the test data matrix. It is well known that the number of degrees of freedom left *after* extracting the nth PC from the test data is $(r - n)$ $(c - n)$. Since we need the number of degrees of

freedom *before* extracting the nth PC, the number of degrees of freedom of the reference matrix is therefore found to be $(r-n+1)\ (c-n+1)$. Since no parameters can be estimated from a random matrix, the number of degrees of freedom automatically equals the total number of data points. Thus the *size* of the reference matrix should be $(r-n+1) \times (c-n+1)$. This leads to the following correction procedure before testing the significance of the eigenvalues. The first eigenvalue of the test data matrix is corrected by subtracting the first eigenvalue of an $r \times c$ random matrix, the second eigenvalue of the test data matrix is corrected by subtracting the first eigenvalue of an $(r-1) \times (c-1)$ random matrix, and so on. In general, the correction for the nth eigenvalue under scrutiny is found as the dominant eigenvalue of an $(r-n+1) \times (c-n+1)$ random matrix. Accurate reference values are obtained by averaging the eigenvalues of a large number of random matrices. Tables with accurate reference eigenvalues have, for example, been published by Mandel [28] and are easily extended by Monte Carlo (MC) simulations [15]. As a modification to the method of Hugus and El-Awady – in the sequel referred to as method B – we therefore propose to examine the following ratio

$$\rho_n(\lambda) = \frac{\lambda_n - \lambda_{n,\mathrm{ref}}}{\sigma(\lambda_n)} \tag{9}$$

where $\lambda_{n,\mathrm{ref}}$ denotes the nth reference eigenvalue. It is to be expected that only the ratio associated with the last significant PC (complete model) should be consistent with the ratio found by equation (3) since here the data reduction proceeds in an entirely different way. Furthermore these values should only agree as long as the reference values $\lambda_{n,\mathrm{ref}}$ are negligible since we do not apply a correction in Method A.

In a previous publication we showed that the standard errors predicted by Eq. 8 overestimate the true standard errors [16]. The overestimate is negligible for the large primary eigenvalues but may be notable for the small ones. (The true standard errors were estimated by MC simulations.) Thus especially for the high-numbered primary eigenvalues the ratios calculated by Eq. 9 are an underestimate of the true ratios and consequently method B is expected to give conservative estimates of the pseudorank. This will, however, only constitute a problem if for the specific application at hand a false negative declaration, i.e., a primary PC is deemed non-significant, causes more harm than a false positive declaration, i.e., a secondary PC is deemed significant. It should be noticed that for many PCA based methods e.g., iterative target testing factor analysis (ITTFA) the incomplete model will lead to erroneous results. In that case the conservative estimate could still provide a useful lower bound.

There is still one point we want to discuss with respect to Eq. 9. Goodman and Haberman [17] have shown that the eigenvalues of PCA are biased as a result of random measurement noise. Although they only give the relevant expression for a one-dimensional PC model their result is easily generalized to an n^*-dimensional PC model by invoking Malinowski's error functions. We will show in a future publication that the bias in eigenvalue λ_n can be predicted as $\mathrm{bias}(\lambda_n) = (r + c - n^*)\sigma(\mathbf{M})^2$. The adequacy of this *theoretical* prediction, however, depends on the signal-to-noise ratio (*SNR*). For low *SNR* this prediction is not accurate enough in order to construct a confidence interval for the eigenvalue in absence of noise. It will be shown in this paper that an accurate *empirical* 'bias correction' is always provided by the reference eigenvalue $\lambda_{n,\mathrm{ref}}$.

2.5. *Pseudorank estimation method based on the standard errors in the singular values of SVD (t-test)*

Methods A and B are both characterized by comparing a ratio with a *fixed critical value* (one or three). This procedure is not in the spirit of hypothesis testing in statistics. In statistics a hypothesis is formulated about a test statistic and the validity of the hypothesis is derived from the sampling distribution of the test statistic, the appropriate number of degrees of freedom and a certain (predetermined) significance level. The number of degrees of freedom and thus also the critical value of the test statistic should depend on the test data at hand. It is extremely difficult

to devise such a procedure for the statistics given by Methods A and B because their sampling distribution is unknown. For Method B this is caused by the fact that the numerator and denominator in Eq. 9 are not independent. In general it is possible to infer the sampling distribution of a statistic from MC simulations. However, we will not pursue this line because it is possible to derive a significance test [1] from Method B in a straightforward manner without resorting to (additional) simulations.

In the case that the variance in an estimated parameter depends on the parameter itself, the standard procedure in statistics consists of 'stabilizing' the variance by transforming the parameter in such a way that the transformed parameter is independent of the associated variance [29]. It immediately follows from Eq. 8 that the standard error in the singular values is constant and equal to $\sigma(\mathbf{M})$ (see also [17,16]). Thus the stabilized parameters are simply given by the singular values. [2] Furthermore, the singular values are linear functions of the data. Thus given Gaussian distributed measurement errors, the sampling distribution of the singular values is also given by the Gaussian distribution [17]. The assumption of Gaussian distributed noise is often not justified in practice. However, it is well known that deviations from normality can be neglected if the number of observations (i.e., in our case the number of matrix elements) is sufficiently large. As a general guide, a number of at least 50 can be considered to be large enough [30,31]. If we have, in addition, some prior knowledge which suggests that the distribution of the matrix elements resembles the Gaussian in some way, e.g., symmetry, then this would allow us to regard a smaller number as large enough. It is interesting to note that very recently Booksh and Kowalski [32] have demonstrated a considerable 'normalizing' effect for the generalized rank annihilation method (GRAM), a calibration method that is based on PCA.

It is possible to examine the statistic that is obtained by simply replacing the eigenvalues in Eq. 9 by the corresponding singular values. It is easily shown that the resulting statistic is always larger. However, the Eq. 9 statistic does not take into account that the eigenvalues of the reference matrix vary in a similar way as the eigenvalues of the test data matrix. Thus an additional modification is necessary before the test statistic is complete. Approximating the standard errors in both singular values by $\sigma(\mathbf{M})$ yields as a possible test statistic

$$\rho_n(\theta) = \frac{\theta_n - \theta_{n,\mathrm{ref}}}{\sqrt{2}\,\sigma(\mathbf{M})} \tag{10}$$

If the test data matrix is large enough the ratio given by Eq. 10 is approximately distributed as Student's t independent of the distribution of the measurement error (large sample assumption). The number of degrees of freedom associated with this test statistic is determined by the size of the reference matrix, i.e., $\nu = (r - n + 1)(c - n + 1)$. The ratio of Eq. 10 is designed to test the null-hypothesis

$$\mathrm{H}_0: \theta_n = \theta_{n,\mathrm{ref}}$$

against the alternative hypothesis (a one-sided test)

$$\mathrm{H}_1: \theta_n > \theta_{n,\mathrm{ref}}$$

Hence we reject H_0 at the α level of significance if the realization of $\rho_n(\theta)$ is greater than or equal to the tabulated $t_\nu(1-\alpha)$. Analogous to the F-test of Malinowski (that is to be briefly discussed next) the proposed significance test starts at the high-numbered PCs. First, the singular value with $n = s$ is tested against the singular value of an $(r - s + 1) \times 1$ matrix. If the calculated ρ is less than the tabulated t, the singular value under test is added to the secondary set. Next, we examine the ratio for $n = s - 1$, and so

[1] We make a distinction here between pseudorank estimation methods: only methods that are able to provide a significance level are denoted as significance tests.

[2] It is important to note that Lawson and Hanson [30] give deterministic error bounds for the singular values of a perturbed matrix. If M is perturbed by a matrix E, an upper bound for the error in a singular value is given as the largest singular value of E. It should be clear that these error bounds are not statistical in nature (no assumptions are made about the elements in E) and therefore not accurate enough for the purpose of this paper.

on. The process of testing and adding to the null set is repeated until the calculated ρ exceeds the tabulated t.

It is seen that the variability of both singular values is taken into account in Eq. 10 in a pessimistic fashion since the standard error in the singular values of the reference matrix is smaller than $\sigma(\mathbf{M})$. (An overestimate by a factor of two should be expected [16].) The conservative character of the proposed t-test should guard the user against the consequences of, for example, violating the large sample assumption in practice. However, a thorough evaluation should demonstrate whether the test still has enough discriminating power or that it is useless.

2.6. Reduced eigenvalues and Malinowski's F-test

Malinowski discovered that the following function of the eigenvalues of PCA

$$REV_n = \frac{\lambda_n}{(r-n+1)(c-n+1)} \tag{11}$$

is constant for the secondary PCs [6]. Using these 'reduced' eigenvalues (*REV*s) an F-test was developed [4,5]:

$$F(\nu_1,\nu_2) = \frac{\sum_{j=n+1}^{s}(r-j+1)(c-j+1)}{(r-n+1)(c-n+1)} \times \frac{\lambda_n}{\sum_{j=n+1}^{s}\lambda_j} \tag{12}$$

with degrees of freedom $\nu_1 = 1$ and $\nu_2 = s-n$. The procedure consists of testing λ_n against the pool of $(s-n)$ remaining eigenvalues. One starts with eigenvalue λ_{s-1} and works backwards through the list of eigenvalues. If an eigenvalue is found to be insignificant, it is pooled with the remaining error eigenvalues, the counter n is lowered by one and the next eigenvalue is considered. It is seen that the number of degrees of freedom is taken as the number of eigenvalues involved in the test. The number of degrees of freedom associated with PCA is further discussed in the Appendix. It was found that in general testing on the 5% level tends to underestimate whereas testing on the 10% level tends to overestimate the pseudorank of the matrix [4].

3. Experimental

The methods discussed in the preceding section are evaluated by analyzing data obtained from the literature as well as simulated data.

3.1. Literature data

Investigating data from the literature in order to test a new method is useful since these data should be readily available for other researchers thus making the present results reproducible. Some of the data sets considered in this section are based on computer simulations. They should

Table 1
Characterization of literature test data

Data	FIAL80 [a]	GUTM68 [b]	HAVE85 [c]	MALI82 [d]	RITT76 [e]	WEIN70 [f]	WEIN71 [g]
Size	20×5	9×5	10×8	20×10	17×7	14×9	22×6
Pseudorank	3	2	3	5	2	3	2
$\hat{\sigma}(\mathbf{M})^h$	0.45	0.049	0.013	0.59	0.14	0.65	0.56

[a] Computer simulated powder diffraction intensities [33].
[b] Half-wave potentials of metal ions in various solvents [34]. Two PCs are deemed significant by the tests discussed below while Howery reports that three PCs are needed to adequately reproduce the data [35].
[c] Potentiometric data [36].
[d] Simulated mass spectra [37].
[e] Mass spectra for which the row corresponding to contaminating nitrogen is deleted [38].
[f] Chemical shifts in various solvents [39].
[g] Chemical shifts in various solvents [40]. Two significant PCs are found by the tests discussed below while Weiner et al. report that three PCs are needed for the reproduction of the data within the experimental error (0.5Hz).
[h] Estimated as $\hat{\sigma}(\mathbf{M}) = \sum_{j=n^*+1}^{s}\lambda_j/(r-n^*)(c-n^*)$ where n* denotes the pseudorank.

be particularly useful for the purpose of this paper since simulated data can be expected to be well-behaved with respect to the 'measurement' error. The selection of these data sets is primarily based on the following consideration: if we want to discover whether a procedure has failed to indicate the correct pseudorank, we should apply it to data for which a reliable estimate for the pseudorank is available. In Table 1 we have described the literature test data that meet this requirement. The selected data sets cover a wide range of experimental techniques. The first row gives the data matrix under investigation. In the second row the size of the matrices ($r \times c$) is given. It is seen that in general the number of data points is rather small. The obvious reason is that it is not practical to publish large data matrices in journal articles. The number of data points ranges between 45 for data set GUTM68 and 200 for data set MALI82. The unfavourable size of some of the data sets may lead to a small number of degrees of freedom and critically influence the outcome of the proposed t-test. In the third row we have listed the estimated pseudorank n^*. The quoted estimate is found by a large number of methods from which the following are the most widely used: cross-validation [2], reduced eigenvalues [6], imbedded error function [3], indicator function [3] and the eigenvalue ratio [41–43]. With two exceptions, i.e., GUTM68 and WEIN70, the determined pseudorank agrees with the value reported earlier. (These exceptions show that data reproduction – formerly a popular method – only provides a reliable pseudorank if $\sigma(\mathbf{M})$ can be estimated accurately.) In the last row we give the estimated standard deviation of the measurement error that is based on the residuals of the correct PC model. These values will be used as input for the parametric methods since for many of these data sets a reliable estimate independent of the data is not available. Exceptions are formed, for example, by data sets WEIN70 and WEIN71 for which a measurement error of 0.5 Hz is reported. One of these datasets, i.e., WEIN70, will be investigated using both values, i.e., 0.5 and 0.65, in order to evaluate the robustness of the parametric methods with respect to an inaccurate estimate of $\sigma(\mathbf{M})$.

3.2. Simulated data

In a previous investigation [16] we constructed a dilution series by simulating a number of multicomponent systems for which the signal of one component was systematically lowered. This way the usefulness of theoretical results like Eq. 8 was tested. A three-component LC-UV data matrix was simulated by multiplying Gaussian functions and the (normalized) UV spectra of adenine, cytidine and guanine taken from the work of Zscheile et al. [44]. The size of the resulting matrices was 36×36. Artificial Gaussian noise with standard deviation 0.5 mAU was added. In this paper we will restrict the discussion to dilutions where theoretical predictions should be expected to start to break down. For these dilutions the peakheights of adenine and guanine are 1000 mAU while the peakheight of cytidine is only 6, 5 and 3 mAU respectively. The resulting data sets are denoted as EXP1, EXP2 and EXP3. Details about the simulations are summarized in Table 2. A plot of data matrix EXP1 is shown in Fig. 1. The unfavourable ratio of peakheights and high overlap of the pure component responses (in both instrumental modes) are apparent. These data

Table 2
Characterization of simulated test data [a]

	Adenine	Cytidine	Guanine
Peak positions, μ	9	18	27
Standard deviation peaks, σ	5	5	5
Peakheights h for EXP1 in mAU	1000	6	1000
Peakheights h for EXP2 in mAU	1000	5	1000
Peakheights h for EXP3 in mAU	1000	3	1000
Number of spectra		36	
Number of wavelengths		36	
$\sigma(\mathbf{M})$ in mAU		0.5	

[a] The elements of the data matrices are generated as $M_{ij} = \sum_{k=1}^{K} C_{ik} S_{jk} + N(0,\sigma(\mathbf{M}))$ where K is the number of components (i.e., 3 in our case), C_{ik} is the value of the elution profile of component k at time i, S_{jk} denotes the absorbance of component k at wavelength j and $N(0,\sigma(\mathbf{M}))$ is a normally distributed number with zero mean and standard deviation $\sigma(\mathbf{M})$. The elements of the elution profiles are calculated as $C_{ik} = h_k \cdot \exp\,[-1/2(i-\mu_k)^2/\sigma_k^2]$ where the symbols have the meaning as indicated above.

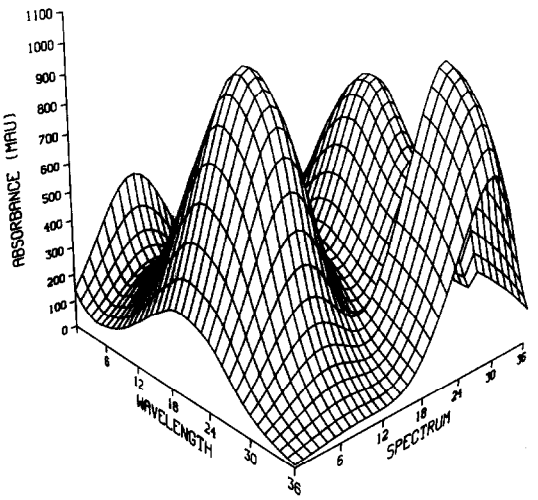

Fig. 1. Simulated three-component LC-UV data matrix EXP1.

matrices should therefore constitute an interesting test case for the methods discussed in this paper. Only data matrices EXP2 and EXP3 have been analyzed before in [16]. For data set EXP2 it was found that the prediction of the standard error in the eigenvalue was excellent for the two main components but wrong (too high) by 20% for the dilute component. However, given the low value of the relevant eigenvalue ratio ($\lambda_3/\lambda_4 = 1.77$) this result was seen as very promising. For data set EXP3 the true standard error for the dilute component was overestimated by 85%. Additional results showed that the third dimension for this data set primarily consists of noise.

3.3. Calculations

The computer program is written in Fortran77 and all calculations are performed in double precision arithmetic on a HDS-EX60 mainframe computer. Built-in subroutines and functions of the IMSL library [45] are used. The SVD of the data sets is performed by subroutine DLSVRR. Significance levels for the F-test are calculated from the output of function DFDF as $\%\alpha = 100 \times (1 - \mathrm{DFDF}(F,\nu_1,\nu_2))$. Occasionally very large F-values may cause a floating point underflow in the evaluation of DFDF. This problem is solved by setting $\%\alpha$ to zero if $F > 100$. Significance levels for the t-test are calculated from the output of function DTDF as $\%\alpha = 100 \times (1 - \mathrm{DTDF}(t,\nu))$.

4. Results and discussion

4.1. Literature data

Before presenting the results for the parametric methods introduced in the theoretical section, we want to discuss the performance of three conventional pseudorank estimation methods. These methods are based on (functions of) the eigenvalues and are often used graphically. Additionally, we will show that the singular values are a promising alternative for these conventional methods.

Table 3
Eigenvalues of PCA for literature test data. The numbers marked in bold indicate the estimated pseudorank

n	FIAL80	GUTM68	HAVE85	MALI82	RITT76	WEIN70	WEIN71
1	2.60×10^5	1.89×10^2	9.67×10	2.56×10^5	2.43×10^3	1.02×10^7	4.90×10^5
2	1.87×10^4	$\mathbf{2.38 \times 10^{-1}}$	4.61×10^{-1}	2.12×10^4	$\mathbf{3.34 \times 10^2}$	4.77×10^2	$\mathbf{5.54 \times 10^3}$
3	$\mathbf{2.53 \times 10^3}$	3.17×10^{-2}	$\mathbf{4.10 \times 10^{-2}}$	1.77×10^4	7.60×10^{-1}	$\mathbf{9.55 \times 10}$	1.13×10
4	3.53	1.51×10^{-2}	3.30×10^{-3}	1.02×10^4	3.35×10^{-1}	1.16×10	8.52
5	3.43	3.05×10^{-3}	1.87×10^{-3}	$\mathbf{2.42 \times 10^3}$	2.56×10^{-1}	8.79	3.73
6			3.40×10^{-4}	1.00×10	1.20×10^{-1}	3.94	1.54
7			2.98×10^{-4}	5.87	5.76×10^{-2}	1.56	
8			1.15×10^{-6}	5.20		1.36	
9				3.58		0.70	
10				1.33			

Table 4
Indicator function for literature test data. The numbers marked in bold indicate the estimated pseudorank

n	FIAL80	GUTM68 ($\times 10^{-2}$)	HAVE85 ($\times 10^{-3}$)	MALI82	RITT76 ($\times 10^{-2}$)	WEIN70 ($\times 10^{-1}$)	WEIN71
1	1.02	0.56	1.74	0.209	5.04	0.36	0.284
2	0.72	**0.48**	0.78	0.215	**0.54**	0.23	**0.033**
3	**0.10**	0.80	**0.43**	0.194	0.67	**0.16**	0.051
4	0.41	1.84	0.50	0.126	1.03	0.19	0.086
5	–	–	0.51	**0.020**	1.81	0.23	0.264
6			0.97	0.028	5.82	0.33	–
7			0.34	0.046	–	0.68	
8			–	0.088		2.24	
9				0.258		–	
10				–			

In Table 3 the eigenvalues of PCA are listed. Using the simple argument that primary eigenvalues should be relatively large one easily arrives at the values for the pseudorank given in Table 1. (It is interesting to note that support for the two-dimensional model for WEIN71 comes from comparing the eigenvalue pattern with that for WEIN70.) However, in general the jump between primary and secondary eigenvalues is not so prominent and several methods have been introduced to aid in the decision making process.

The results for two of these methods, the indicator function [2] and the reduced eigenvalues [6], are shown in Tables 4 and 5. Malinowski has postulated that the indicator function should exhibit a minimum for the true dimension of the data matrix [3]. Thus we arrive at the same pseudorank estimates using the indicator function. It is often stated as a disadvantage of this method that the minimum is shallow. This is also the case here for data set WEIN70 but the point is that we can not quantify such a statement without knowing the standard errors in the indicator function. It is possible to derive these standard errors from Eq. 8. This could lead to theoretical evidence for the postulated minimum. However, there is already considerable experimental evidence in the literature that indicates that the minimum is significant (the method is very successful) and we will not pursue this line. According to Malinowski the reduced eigenvalues should be constant for the secondary PCs while the values for the primary PCs should be larger. In another publication [15] we have shown by simulations of random matrices that it depends on the ratio of the rows and the columns of the matrix, the so-called divergence coefficient, whether the reduced eigenvalues are (approximately) constant. This numerical result is explained by a theoretical result of multivariate statistics for the joint probability density function (pdf) of the eigenvalues of a random matrix. (The joint pdf gives the probability of finding *any* eigenvalue in a certain range.) The shape of the joint pdf depends primarily on the divergence coefficient of the matrix. The consequences for the reduced eigenvalues are that different patterns should be expected depending on the value of the divergence coefficient. We have found, for example, that a divergence coefficient of (approximately) one leads to a large probability of finding a very small eigenvalue. This is confirmed here for data set HAVE85. (This very small eigenvalue is believed to be the reason for the dip in the indicator function.) For the other data sets the divergence coefficient ranges between 1.5 and 4. This is the range where the reduced eigenvalues of random matrices were shown to be approximately constant [15]. Thus for the data sets under investigation the method

[3] Occasionally a local minimum is observed as already noted by Malinowski. For the present data sets we find a local minimum for HAVE85 and MALI82. The local minimum for HAVE85 is discussed later. The local minimum for MALI82 is caused by the small ratio between the second and third eigenvalue.

Table 5
Reduced eigenvalues for literature test data. The numbers marked in bold indicate the estimated pseudorank

n	FIAL80	GUTM68	HAVE85	MALI82	RITT76	WEIN70	WEIN71
1	2.60×10^{3}	4.20	1.21	1.28×10^{3}	2.05×10	8.13×10^{4}	3.71×10^{3}
2	2.45×10^{2}	$\mathbf{7.43 \times 10^{-3}}$	7.32×10^{-3}	1.24×10^{2}	**3.48**	4.59	$\mathbf{5.27 \times 10}$
3	$\mathbf{4.69 \times 10}$	1.51×10^{-3}	$\mathbf{8.54 \times 10^{-4}}$	1.23×10^{2}	1.01×10^{-2}	**1.14**	0.14
4	0.10	1.26×10^{-3}	9.42×10^{-5}	8.60×10	0.60×10^{-2}	0.18	0.15
5	0.21	0.61×10^{-3}	7.80×10^{-5}	$\mathbf{2.52 \times 10}$	0.66×10^{-2}	0.18	0.10
6			2.27×10^{-5}	0.13	0.50×10^{-2}	0.11	0.09
7			3.72×10^{-5}	0.11	0.52×10^{-2}	0.07	
8			0.04×10^{-5}	0.13		0.10	
9				0.15		0.12	
10				0.12			

should work and this is confirmed by the results given in Table 5.

In Table 6 we have listed the singular values of the literature test data. The reason for showing the singular values is as follows. In the theoretical section it is argued that the error in the singular values is constant (and equal to the original measurement error). Contrary to the primary singular values the secondary singular values only consist of measurement error. Thus it is to be expected that the distance between the secondary singular values is bounded by the size of the measurement error whereas the distance between the primary singular values is also affected by the size of the systematic variation. Since the size of the systematic variation should be larger than the size of the measurement noise for the data to be analyzable at all, it seems logical to simply inspect the distance between the singular values. It is seen that for all data sets the distance between the secondary singular values is of the same order of magnitude. Furthermore, it is easily verified that the order of magnitude is given by the standard deviations in Table 1. This leads to the conclusion that plotting the singular values yields a promising graphical pseudorank estimation method: the singular values should (approximately) lie on a straight line for the secondary PCs whereas for the primary PCs the curve deviates upwards. It is worth mentioning that the logarithm of the eigenvalues is reported to yield a straight line for the secondary eigenvalues [46]. However, a systematic evaluation [15] showed that a straight line should only be expected for the low-numbered secondary PCs. This numerical result has now been explained since the logarithm and the square root may transform the eigenvalues in a similar way over a restricted range (they are both weak transformations) [4]. It should be kept in mind that we are using qualitative arguments here which we will try to quantify by means of the proposed t-test on the singular values. It is evident that such a quantification should be based on an estimate of the size of the measurement error.

The working and use of the parametric methods is illustrated by discussing the results in detail for data set WEIN70. This data set has a large number of degrees of freedom, i.e., $(r - n^*)(c - n^*) = (14 - 3)(9 - 3) = 66$, and has already been treated extensively in the literature (see e.g., [4] and [43]). Additional results (not shown here) obtained for the χ^2-test as well as the number of 3σ-misfits [14] are further evidence for the suitability of this data set for the evaluation of pseudorank estimation methods. The results of Methods A and B are summarized in Table 7. It should be noted that all calculations are performed with the reported value for the standard deviation of the measurement error (0.5 Hz) which

[4] In multivariate statistics standard errors in the eigenvalues as a result of sampling errors have been derived (see [16] for a detailed discussion). These standard errors also depend on the size of the eigenvalues but now Equation (8) is no longer appropriate. Now the stabilizing transform is given by the logarithm. This motivates inspecting the logarithm of the eigenvalues in the case that sampling errors play a role.

is 30% smaller that the standard deviation estimated from the residuals of the 3-dimensional PC model (0.65 Hz). The first column gives the PC under investigation. The second and third column give the diagonal elements of the reduced matrices, M′ and E′. The resulting ratio calculated from Eq. 3 is given in the next column. Four PCs are deemed significant if we use the (fixed) critical value of three. It is clear that three significant PCs would have been found if the input standard

Table 6
Singular values for literature test data. The numbers marked in bold indicate the estimated pseudorank

n	FIAL80	GUTM68	HAVE85	MALI82	RITT76	WEIN70	WEIN71
1	5.10×10^2	1.374×10	9.835	5.06×10^2	4.93×10	3.20×10^3	7.00×10^2
2	1.37×10^2	**0.488**	0.679	1.46×10^2	**1.83×10**	2.18×10	**7.44×10**
3	**5.03×10**	0.178	**0.202**	1.33×10^2	0.87	**9.77**	3.35
4	1.88	0.123	0.057	1.01×10^2	0.58	3.41	2.92
5	1.85	0.055	0.043	**4.92×10**	0.51	2.97	1.93
6			0.018	3.16	0.35	1.98	1.24
7			0.017	2.42	0.24	1.25	
8			0.001	2.28		1.17	
9				1.89		0.84	
10				1.15			

Table 7
Results of method A and B for literature data matrix WEIN70. The numbers marked in bold indicate the estimated pseudorank

n	Method A			Method B			
	diag(M′)	diag(E′) [a]	$\rho(\varepsilon)$	λ	λ_{ref} [a]	$\sigma(\lambda)$	$\rho(\lambda)$
1	4.56×10^2	0.50	9.11×10^2	1.02×10^7	9.28	3.20×10^3	3.20×10^3
2	1.49×10	0.55	2.71×10	4.77×10^3	8.34	2.18×10	2.15×10
3	7.11	0.83	8.61	9.55×10	7.38	9.77	**9.02**
4	−2.99	0.98	**3.04**	1.16×10	6.42	3.41	1.53
5	−2.22	1.24	1.79	8.79	5.47	2.97	1.12
6	−2.01	0.90	2.23	3.94	4.50	1.98	−0.28
7	1.55	1.74	0.89	1.56	3.55	1.25	−1.59
8	1.55	2.21	0.70	1.36	2.56	1.17	−1.03
9	−0.15	1.82	0.08	0.70	1.49	0.84	−0.94

[a] Calculated with $\hat{\sigma}(\mathbf{M}) = 0.50$.

Table 8
Results of t-test and F-test for literature data matrix WEIN70. The numbers marked in bold indicate the estimated pseudorank

n	t-test						F-test				
	Θ	Θ_{ref} [a]	$\sigma(\Theta_{ref})$	t	ν	%α	REV	F	ν_1	ν_2	%α
1	3.20×10^3	3.04	0.26	4.52×10^3	126	0.0	8.13×10^4	5.20×10^4	1	8	0.0
2	2.18×10	2.88	0.27	2.68×10	104	0.0	4.59	1.04×10	1	7	1.4
3	9.77	2.70	0.27	1.00×10	84	**0.0**	1.14	7.96	1	6	**3.0**
4	3.41	2.52	0.28	1.26	66	10.6	0.18	1.40	1	5	29.0
5	2.97	2.32	0.29	0.92	50	18.1	0.18	1.86	1	4	24.4
6	1.98	2.10	0.29	−0.17	36	56.7	0.11	1.33	1	3	33.2
7	1.25	1.86	0.31	−0.86	24	80.1	0.07	0.63	1	2	51.0
8	1.17	1.57	0.32	−0.57	14	71.1	0.10	0.83	1	1	53.0
9	0.84	1.17	0.35	−0.47	6	67.3	0.12	–	–	–	–

[a] Calculated with $\hat{\sigma}(\mathbf{M}) = 0.50$.

deviation would have been larger by only 2%. In the next two columns we give the eigenvalues of PCA and the appropriate reference eigenvalues. The standard error in the eigenvalues predicted from Eq. 8 is given in the next column. The resulting ratio calculated from Eq. 9 is given in the last column. Due to the considerable correction by the reference eigenvalue we now find only three significant PCs. It is clear that the outcome would only change if we would underestimate $\sigma(\mathbf{M})$ by a factor of two. It is seen that Method B is more robust with respect to errors in the input value of $\sigma(\mathbf{M})$ than Method A.

The results for the t-test and Malinowski's F-test are given in Table 8. The first three columns give the PC under consideration, the test singular value and the reference singular value respectively. In the next column we have listed the standard error in the reference singular values. Over the whole range this value is (much) smaller than $\sigma(\mathbf{M}) = 0.50$. This means that the t-values listed in the next column are conservative as indicated before. The degrees of freedom to be used in the test are given in the next column. The resulting significance levels clearly indicate the presence of three significant PCs and nearly a fourth one on the 10% level. However, the input value of $\sigma(\mathbf{M})$ is rather small compared to the value estimated from the residuals of the correct PC model and these results are therefore promising. (It follows that the robustness found earlier for Method B is misleading.) The results for the F-test are shown next. They have already been discussed in detail by Malinowski [4]. Three PCs are deemed significant at the 5% level of significance. Thus both tests agree about the true dimension of the data. However, there is a sharp contrast with respect to the significance levels. The significance level supplied by the t-test is essentially zero while the F-test attaches an uncertainty of 3% to the model. The reason for the discrepancy is that the t-test uses much more degrees of freedom than the F-test (see Appendix). Although the t-test should also give conservative estimates of the significance of a PC model it is seems to be less conservative than the F-test. Using the extra knowledge about the measurement error has indeed led to a sharper significance level. In the case that an important decision has to be made based on the result of a significance test the improvement obtained by the t-test may be appreciable.

In Table 9 we have summarized the results of the various parametric methods for the last significant PCs. It should be noted that for all data sets the value of $\sigma(\mathbf{M})$ taken from Table 1 is inserted in the relevant expressions. (As a result the first secondary PC is deemed non-significant in all cases.) The first column lists the data set under consideration. The next two columns gives the ratios calculated from Eqs. 3 and 9, respectively. It is seen that the agreement is very well in all cases. This result is remarkable since the evaluation of Eq. 9 involves the correction by a reference value obtained from random matrices. The next three columns list the results for the t-test. The t-values are extremely high in all cases leading to very small significance levels. The last three columns summarize the results for the F-test. In all cases the F-values are larger than the

Table 9
Results of various pseudorank estimation methods [a] for last significant PC of literature test data

Data	Method A	Method B	t-test			F-test		
	$\rho(\varepsilon)$	$\rho(\lambda)$	t	ν	$\%\alpha$	F	ν_2	$\%\alpha$
FIAL80	4.03×10	5.54×10	7.49×10	54	0.0	3.37×10^2	2	0.3
GUTM68	4.37	4.18	4.24	32	8.9×10^{-2}	5.66	3	9.8
HAVE85	7.19	7.19	7.89	48	0.0	1.25×10	5	1.7
MALI82	4.16×10	4.16×10	5.51×10	96	0.0	1.99×10^2	5	3.2×10^{-3}
RITT76	6.01×10	6.38×10	1.22×10^2	96	0.0	4.66×10^2	5	4.0×10^{-4}
WEIN70	6.58	6.52	6.79	84	0.0	7.96	6	3.0
WEIN71	6.26×10	6.64×10	8.98×10	105	0.0	4.00×10^2	4	3.7×10^{-3}

[a] Evaluated with $\hat{\sigma}(\mathbf{M})$ given in Table 1.

Table 10
Reference singular values [a] for literature test data. The numbers marked in bold indicate the estimated pseudorank for the test data. Reference values that are higher than the test values (see Table 3) are underlined

n	FIAL80	GUTM68	HAVE85	MALI82	RITT76	WEIN70	WEIN71
1	2.72	2.180	0.068	4.12	0.87	3.95	3.61
2	2.54	**0.196**	0.063	3.94	**0.82**	3.74	**3.41**
3	**2.34**	0.171	**0.058**	3.77	0.77	**3.52**	3.20
4	2.11	0.142	0.053	3.58	0.71	3.28	2.96
5	1.79	0.104	0.047	**3.37**	0.65	3.02	2.67
6			0.040	3.17	0.57	2.74	2.28
7			0.032	2.92	0.47	2.42	
8			0.021	2.66		2.04	
9				2.34		1.52	
10				1.91			

[a] Calculated with $\hat{\sigma}(\mathbf{M})$ given in Table 1.

t-values (this is not a general rule). However, as a result of the much smaller number of degrees of freedom the resulting significance levels are much larger than those found by the t-test. This is without consequence for the estimated pseudorank except for the (extremely small) data set GUTM68. For this data set Malinowski's F-test gives one significant PC at the 5% level and two significant PCs at the 10% level. This result is in agreement with the conclusion of Malinowski about the tendency to underestimate or overestimate the pseudorank at the 5 and 10% level respectively.

In Table 10 we give the reference singular values for the literature data matrices. Using these numbers it is possible to reproduce the results for the t-test given in Table 9. In another publication [15] we show that the secondary eigenvalues of the test data matrix are approached from above by the eigenvalues of the reference matrix. As a result we find that the test singular values tend to be smaller than their reference values. The differences are small, however, especially for the first secondary PC. The tendency of the reference values to be too large contributes to the conservative character of the t-test.

For the literature data sets we have found an excellent agreement between the results of the F-test and the t-test with respect to the estimated pseudorank. The reason is that these data sets are not selected for their discriminating ability. It is possible to investigate the difference in sensitivity in more detail by performing the following 'Gedankenexperiment' on the data. A hypothetical matrix is constructed from the test matrix by lowering the last significant singular value while keeping all other things fixed. The size of the hypothetical singular value is determined by the significance level it would give for a certain test. It is to be expected that in order to find a predetermined significance level, say 1%, the size

Table 11
Comparison of actual and critical values for the eigenvalue ratio (ER) at different confidence levels

Data	ER (actual value)	t-test			F-test		
		ER (1%)	ER (5%)	ER (10%)	ER (1%)	ER (5%)	ER (10%)
FIAL80	7.17×10^2	4.22	3.29	2.84	2.10×10^2	3.95×10	1.82×10
GUTM68	7.51	4.20	3.09	2.59	4.52×10	1.34×10	7.35
HAVE85	1.24×10	3.19	2.41	2.05	1.62×10	6.57	4.04
MALI82	2.42×10^2	2.85	2.26	1.98	1.98×10	8.04	4.94
RITT76	4.39×10^2	2.23	1.76	1.54	1.53×10	6.23	3.82
WEIN70	8.21	2.80	2.19	1.91	1.42×10	6.18	3.90
WEIN71	4.92×10^2	2.48	1.99	1.75	2.61×10	9.48	5.58

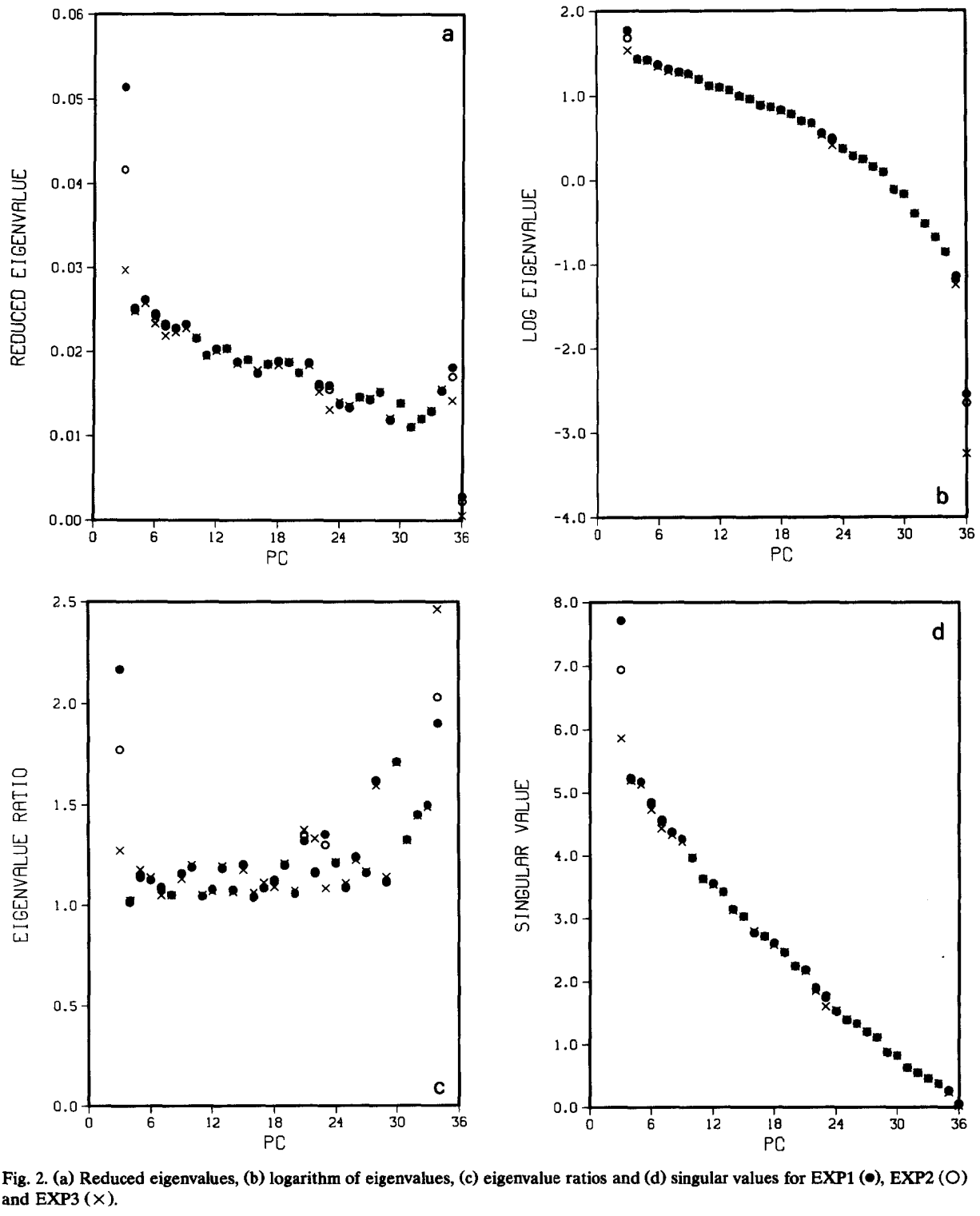

Fig. 2. (a) Reduced eigenvalues, (b) logarithm of eigenvalues, (c) eigenvalue ratios and (d) singular values for EXP1 (●), EXP2 (○) and EXP3 (×).

of this singular value is smaller for the *t*-test than for the *F*-test. Since there is an arbitrary difference in scale between the different data sets, we have listed in Table 11 the ratio of the last significant (hypothetical) and the first non-significant (actual) eigenvalue that would result in significance levels of 1, 5 and 10% respectively for both tests. The first column contains the dataset under consideration. The second column gives the eigenvalue ratio that is actually found. The next three columns give the eigenvalue ratios that would enable testing at the 1, 5 and 10% respectively by the *t*-test. The last three columns give the same results for the *F*-test. From these numbers it is easily discerned that the *t*-test is more sensitive than the *F*-test. The situation is especially favourable for the *t*-test if testing at the 1% level is required. The difference in sensitivity decreases rapidly with increasing significance level. It is interesting to compare the values found for WEIN70 to the critical region found by Hirsch et al. [43] for this data set from extensive simulations: 'If one wishes to ensure the detection of all significant factors and is not concerned that too many factors might be accepted, one should use an ER (eigenvalue ratio) test, probably with a critical value in the range 2.0 to 2.5.' It is remarkable that (approximately) the same critical region is found here by a test that is designed to be conservative.

4.2. Simulated data

In this section we will restrict ourselves to the comparison of the *t*-test and Malinowski's *F*-test. But before presenting these results we discuss the outcome of a large number of conventional methods. In this way we hope to discover what we can reasonably expect from the two significance tests.

In Fig. 2 we show the reduced eigenvalues, the eigenvalue ratios, the logarithm of the eigenvalues and the singular values for the simulated three-component systems. (The values for the first two PCs are not included for visual clarity.) The reduced eigenvalues slowly decrease for the secondary PCs [5]. The logarithm of the eigenvalues lie (approximately) on a straight line for the low-numbered non-significant PCs. This is further illustrated by the eigenvalue ratios being (approximately) constant in that region. (Plotting the eigenvalue ratios instead of the logarithm of the eigenvalues has the advantage of leading to a more practical scale.) These trends are in agreement with the results found earlier [15] for random matrices with an equal number of rows and columns. It is seen that the singular values lie (approximately) on a straight line for all non-significant PCs. This lends credit to the use of the singular values for visual inspection. The graphical methods strongly indicate the presence of three significant PCs for EXP1 and EXP2. For EXP3 no evidence for the presence of the dilute component can be deduced from these plots. The results for EXP2 should be contrasted to the minimum found at the second PC for both imbedded error function and indicator function (not shown here). It should be emphasized that finding a minimum for the imbedded error function can be satisfactorily explained. It is simply the point where less error (random *and* systematic) is introduced in the model by *not* including a PC that in fact contains systematic variation. Cross-validation [3] confirms the choice of a two-dimensional model by giving the ratios 0.04, 0.02 and 1.02 for the first three PCs (cut-off value is one). Since much structure is present in the pure component responses used to construct the data we also investigated the eigenvectors. The first-order autocorrelation function has proved to be a very sensitive method for this kind of data [47]. For the third PC the time constants found are 4.63 and 3.24 for the left and right singular vector respectively while the time constants are 0.47 and 0.58 for the corresponding vectors of the first secondary PC. This result is in excellent agreement with the cut-off value of 0.60 proposed by Shrager [47]. However, for data matrix EXP3 the values 0.75 and 0.69 are found for the third PC and basing a decision on this method becomes difficult. From all the conventional methods ap-

[5] The fact that the reduced eigenvalues are not constant is of little consequence for the application of the *F*-test. The *F*-test guards against violations of assumptions by the small number of degrees of freedom.

Table 12
Results of t-test and F-test for simulated data matrix EXP1. The numbers marked in bold indicate the estimated pseudorank

n	t-test						F-test				
	Θ	Θ_{ref} [a]	$\sigma(\Theta_{ref})$	t	ν	$\%\alpha$	REV	F	$\nu 1$	$\nu 2$	$\%\alpha$
1	4.23×10^3	5.75	0.22	5.97×10^3	1296	0.0	1.38×10^4	4.62×10^2	1	35	0.0
2	6.68×10^2	5.66	0.22	9.36×10^2	1225	0.0	3.64×10^2	1.54×10^4	1	34	**0.0**
3	7.71	5.58	0.22	3.01	1156	**0.1**	5.14×10^{-2}	2.44	1	33	12.8
4	5.24	5.49	0.22	−0.35	1089	63.7	2.52×10^{-2}	1.22	1	32	27.9
5	5.18	5.40	0.22	−0.31	1024	62.2	2.62×10^{-2}	1.30	1	31	26.4
6	4.85	5.31	0.23	−0.65	961	74.2	2.45×10^{-2}	1.24	1	30	27.4

[a] Calculated with $\hat{\sigma}(\mathbf{M}) = 0.50$.

plied to this matrix only a plot of the singular vectors indicates the presence of the third dimension (see Fig. 7 in [16]). The human eye is seen to be an excellent pattern recognizer. The application of this method is, however, restricted to ordered data [48]. In the light of the above results it is reasonable to demand that the significance tests yield certainly three significant PCs for EXP1 and EXP2 and possibly two for EXP3.

In Table 12 we have summarized the relevant results for data matrix EXP1 (for explanation of the symbols see Table 8). It is seen that the t-test correctly yields three significant PCs whereas the F-test fails. The failure is caused by the combined effect of the relatively small F-value and the small number of degrees of freedom. The results for data matrix EXP2 (not shown here) are very similar. Only the numbers for the third PC, which corresponds to the dilute component, change significantly. The t-value for the third PC is decreased to 1.92, yielding a significance level of 2.8%. Inspecting the plots in [16] may lead to the conclusion that this is somewhat pessimistic. Finally, for data matrix EXP3 we find a t-value of 0.40 which is not significant ($\%\alpha = 34.5$). The results for the simulated data sets support the results obtained for the literature data. The t-test makes effective use of the extra knowledge. The estimated pseudoranks are summarized in Table 13.

5. Conclusions

In this paper three methods are discussed with respect to their usefulness as a parametric pseudorank estimation method. These methods explicitly need an estimate of the size of the measurement error. Thus they are restrictive in their use but should be expected to be more sensitive than methods that cannot employ this extra knowledge. Method A (based on the row-echelon form of the matrix) and Method B (based on the eigenvalues of PCA) have the disadvantage of not supplying significance levels for the estimated pseudorank. (Furthermore, Method A is shown to be sensitive for an incorrect estimate of the size of the measurement error.) Thus they offer little advantage over a large number of methods that is already available and does not need the extra knowledge. The third method (a t-test) makes good use of the extra knowledge and does supply a significance level. As a result of the assumptions made in order to construct the test statistic, the t-test should be expected to give conservative estimates of the pseudorank. In or-

Table 13
Results of various pseudorank estimation methods for simulated test data

Data	Cross-validation	Imbedded error function	F-test	Indicator function	t-test	Reduced eigenvalue	Logarithm eigenvalue	Eigenvalue ratio	Singular value
EXP1	2	2	2	3	3	3	3	3	3
EXP2	2	2	2	2	3	3	3	3	3
EXP3	2	2	2	2	2	2	2	2	2

der to gauge the sensitivity of this t-test, a comparison is carried out with Malinowski's F-test for data obtained from the literature and simulated data. For the data matrices obtained from the literature the estimated pseudorank agrees very well for both significance tests. However, the t-test gives sharper confidence levels as a result of the larger number of degrees of freedom involved in the test. For the simulated data matrices Malinowski's F-test fails to indicate the correct dimension in cases where the t-test still yields sharp confidence levels. It is concluded that prior knowledge of the size of the measurement is put to effective use by the currently developed t-test. Additional support for the viability of the new t-test comes from a thorough analysis of the test data by a large number of conventional methods. Finally, as a remarkable by-product of the current research we have found that a plot of the singular values yields a promising graphical pseudorank estimation method. (This is only remarkable, since two modern textbooks on PCA do not mention this possibility [49,50].) Graphical methods have proved their use in the past in cases where the size of the measurement error is unknown. This new graphical method therefore provides a natural complement to the t-test.

Appendix 1

In essence Malinowski's theory deals with the number of degrees of freedom associated with secondary PCs. It is interesting to compare his results, viz. Eq. 11, with the theory developed by Mandel [28]. Mandel argues that an eigenvalue explains a portion of the sum of squares associated to the data. In order to arrive at the portion variance explained by an eigenvalue, the eigenvalue should be divided by an appropriate number of degrees of freedom. The consequence of this reasoning is as follows: for a secondary PC the expectation of the eigenvalue divided by the appropriate number of degrees of freedom should be an unbiased estimator of the variance of the measurement error $\sigma(\mathbf{M})^2$. Alternatively, dividing the expectation of the eigenvalue λ_n by the variance of the measurement error should yield an unbiased estimator of the appropriate number of degrees of freedom, ν_n:

$$\nu_n = \frac{E[\lambda_n]}{\sigma(\mathbf{M})^2} \tag{13}$$

where $E[\cdot]$ denotes expected value. In fact these expected values are the reference values discussed earlier. Since the variance accounted for by the secondary PCs should be constant the denominator in Eq. 11 should be proportional to the number of degrees of freedom associated with the PC under scrutiny. The proportionality constant N (normalization) is found by observing that the number of degrees of freedom summed over the secondary PCs should add up to the total number of degrees of freedom left after extracting n^* components, i.e., building the correct model:

$$N = \frac{(r-n^*)(c-n^*)}{\sum_{j=n^*+1}^{s}(r-j+1)(c-j+1)} \tag{14}$$

Hence $\nu_n = N(r-n+1)(c-n+1)$, found this way, should equal ν_n found from evaluating Eq. 13. We will return to this question in more detail in another publication [15]. It is tempting to evaluate Eq. 12 using the number of degrees of freedom associated with the sources of variance that are tested instead of using the number of sources as the number of degrees of freedom, i.e., take $\nu_1 = N(r-n+1)(c-n+1)$ and $\nu_2 = \Sigma_{n+1}^{s} N(r-j+1)(c-j+1)$ instead of $\nu_1 = 1$ and $\nu_2 = s-n$. In general this should lead to larger numbers of degrees of freedom and consequently the resulting F-test should yield an increased discriminating ability. Some obvious disadvantages are connected with this 'alternative F-test'. In general the resulting numbers will not be integral as already indicated by Mandel [28]. This problem is easily solved by rounding the numbers to the nearest integer. The confidence levels do not change very much by this operation, especially for large data matrices. A second disadvantage is the presence of the correct dimension of the PC model in Eq. 14. This problem cannot be solved since it leads to circular reasoning: the

pseudorank needs to be known in order to estimate it.

References

[1] E.R. Malinowski, Factor Analysis in Chemistry, Wiley, New York, 1991.
[2] E.R. Malinowski, Anal. Chem., 49 (1977) 612.
[3] S. Wold, Technometrics, 20 (1978) 397.
[4] E.R. Malinowski, J. Chemom., 3 (1988) 49.
[5] E.R. Malinowski, J. Chemom., 4 (1990) 102.
[6] E.R. Malinowski, J. Chemom., 1 (1987) 33.
[7] L. Ståhle and S. Wold, Chemom. Intell. Lab. Syst., 6 (1989) 259.
[8] X.M. Tu, D.S. Burdick, D.W. Millican and L.B. McGown, Anal. Chem., 61 (1989) 2219.
[9] X.M. Tu, J. Chemom., 5 (1991) 333.
[10] S.P. Koinis, A.T. Tsatsas and D.F. Katakis, J. Chemom., 5 (1991) 21.
[11] Y.-Z. Liang, O.M. Kvalheim and A. Höskuldsson, J. Chemom., 7 (1993) 277.
[12] V. Tomišić and V. Simeon, J. Chemom., 7 (1993) 381.
[13] R.M. Wallace and S.M. Katz, J. Phys. Chem., 68 (1964) 3890.
[14] Z.Z. Hugus and A.A. El-Awady, J. Phys. Chem., 75 (1971) 2954.
[15] N.M. Faber, L.M.C. Buydens and G. Kateman, Chemom. Intell. Lab. Syst., in press.
[16] N.M. Faber, L.M.C. Buydens and G. Kateman, J. Chemom., 7(1993) 495.
[17] L.A. Goodman and S.J. Haberman, JASA, 85 (1990) 139.
[18] R.N. Cochrane and F.H. Horne, Anal. Chem., 49 (1977) 846.
[19] R.J. Pell, M.B. Seasholtz and B.R. Kowalski, J. Chemom., 6 (1992) 57.
[20] C.B.M. Didden and H.N.J. Poulisse, Anal. Lett., 13 (1980) 921.
[21] M.G. Moran and B.R. Kowalski, Anal. Chem., 56 (1984) 562.
[22] H. Bubert and H. Jenett, Z. Anal. Chem., 335 (1989) 643.
[23] R.M. Wallace, J. Phys. Chem., 64 (1960) 899.
[24] G. Weber, Nature, 190 (1961) 27.
[25] S. Ainsworth, J. Phys. Chem., 65 (1961) 1968.
[26] G.H. Golub and C.F. Van Loan, Matrix Computations, Hopkins University Press, Baltimore, 1983.
[27] J.M. Halket, J. Chromatogr., 175 (1979) 229.
[28] J. Mandel, Technometrics, 13 (1971) 1.
[29] P.J. Bickel and K.A. Doksum, Mathematical Statistics, Holden-Day, San Francisco, CA, 1977.
[30] C.L. Lawson and R.J. Hanson, Solving Least Squares Problems, Prentice-Hall, Englewood Cliffs, 1974.
[31] J.R. Green and D. Margerison, Statistical Treatment of Experimental Data, Elsevier, Amsterdam, 1978.
[32] K. Booksh and B.R. Kowalski, J. Chemom., 8 (1994) 45.
[33] J. Fiala, Anal. Chem., 52 (1980) 1300.
[34] V. Gutmann, Co-ordination Chemistry in Non-aqeous Solutions, Springer Verlag, Vienna, 1968, p. 33
[35] D.G. Howery, Bull. Chem. Soc. Jpn, 45 (1972) 2643.
[36] J. Havel and M. Meloun, Talanta, 32 (1985) 171.
[37] E.R. Malinowski, Anal. Chim. Acta, 134 (1982) 129.
[38] G.L. Ritter, S.R. Lowry, T.L. Isenhour and C.L. Wilkins, Anal. Chem., 48 (1976) 591.
[39] P.H. Weiner, E.R. Malinowski and A.R. Levinstone, J. Phys. Chem. 74 (1970) 4537.
[40] P.H. Weiner and E.R. Malinowski, J. Phys. Chem., 75 (1971) 1207.
[41] H.B. Woodruff, P.C. Tway and L.J. Cline Love, Anal. Chem., 53 (1981) 81.
[42] R.A. Hearmon, J.H. Scrivens, K.R. Jennings and M.J. Farncombe, Chemom. Intell. Lab. Syst., 1 (1987) 167.
[43] R.F. Hirsch, G. Lam Wu and P.C. Tway, Chemom. Intell. Lab. Syst., 1 (1987) 265.
[44] F.P. Zscheile, H.C. Murray, G.A. Baker and R.G. Peddicord, Anal. Chem., 34 (1962) 1776.
[45] IMSL MATH/LIBRARY User's Manual, version 1.1, IMSL, Inc., Houston, TX, 1989
[46] N. Ohta, Anal. Chem., 45 (1973) 553.
[47] R.I. Shrager, SIAM J. Alg. Disc. Meth., 5 (1984) 351.
[48] C. Shen, T.J. Vickers and C.K. Mann, J. Chemom., 5 (1991) 417.
[49] I.T. Joliffe, Principal Component Analysis, Springer Verlag, New York, 1986.
[50] J.E. Jackson, A User's Guide to Principal Components, Wiley, New York, 1991.