ELSEVIER

Chemometrics and Intelligent Laboratory Systems 25 (1994) 203–226

Chemometrics and intelligent laboratory systems

Aspects of pseudorank estimation methods based on the eigenvalues of principal component analysis of random matrices

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

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

Received 4 February 1994; accepted 25 May 1994

Abstract

Nowadays, analytical instruments that produce a data matrix for one chemical sample enjoy a widespread popularity. However, for a successful analysis of these data an accurate estimate of the pseudorank of the matrix is often a crucial prerequisite. A large number of methods for estimating the pseudorank are based on the eigenvalues obtained from principal component analysis (PCA). In this paper methods are discussed that exploit the essential similarity between the residuals of PCA of the test data matrix and the elements of a random matrix. In the literature of PCA these methods are commonly denoted as parallel analysis. Attention is paid to several aspects that have to be considered when applying such methods. For some of these aspects asymptotic results can be found in the statistical literature. In this study Monte Carlo simulations are used to investigate the practical implications of these theoretical results. It is shown that for sufficiently large matrices the distribution of the measurement error does not significantly influence the results. Down to a very small signal-to-noise ratio the ratio of the number of rows and the number of columns constitutes the major influence on the expected value of the eigenvalues associated with the residuals. The consequences are illustrated for two functions of the eigenvalues, i.e. the logarithm of the eigenvalues and Malinowski's reduced eigenvalues. Both methods are graphical and have been applied in the past with considerable success for a variety of data. Malinowski's reduced eigenvalues are of special interest since they have been used to construct an F-test. Finally, a modification is proposed for pseudorank estimation methods that are based on the principle of parallel analysis.

1. Introduction

It is becoming common practice that modern analytical instruments produce a large amount of data for one chemical sample. This development has inspired chemometricians to introduce new multivariate techniques and extend existing ones, especially for the purpose of calibration [1]. In this area it is also proposed to classify the techniques according to the order of the data that are analyzed: zero-order data are scalars, first-order data are vectors and second-order data are matrices. The use of first-order data enables the quantitation of an analyte in the presence of an interfering signal that is accounted for in the model. This is the so-called first-order advantage. The use of second-order data enables the quantitation of an analyte in the presence of an interfering signal that is *not* accounted for in the model [1]. This is the so-called second-order advantage. The importance of the second-order advantage cannot be overstrained, especially since the techniques developed for exploiting this advantage only need one calibration sample for the analysis of the test sample.

For many techniques that handle second-order data the concept of pseudorank is of pivotal importance. The pseudorank is defined as the mathematical rank of the matrix in the absence of noise. Finding a good estimate

* Corresponding author.

SSDI 0169-7439(94)00051-J

for the pseudorank is often critical for the overall success of the data analysis. In practice, this may lead to problems that are far from trivial, since the analytical instrument is usually not optimized for the measurement of a specific sample. Instead, a large data matrix is produced in order to determine only a few parameters, e.g. the concentrations and the physical descriptors (elution profile, spectrum) of the analytes. It is not uncommon that the data are overdetermined by orders of magnitude. Stated otherwise: the information contained by the resulting data matrix is highly redundant. Or equivalently, there is a large number of linear relations constraining the data within the level of the noise. Redundancy leads to ill-conditioned problems which e.g. in the field of calibration invariably accumulate to the inversion of a nearly singular matrix. The inversion should be carried out in a space of lower dimension in order to avoid excessive error propagation. This dimension is preferably equal to the pseudorank of the data matrix. An overestimate leads to unnecessary error propagation whereas an underestimate leads to loss of information, especially information concerning the minor substituents. Thus it seems appropriate to consider the problem of pseudorank estimation as a serious *second-order disadvantage*. It is important to note that the technique of rank annihilation factor analysis (RAFA) can accommodate for a small *overestimate* of the pseudorank [2]. This empirical result has very recently been given a theoretical basis by the derivation of the appropriate variance and bias expressions [3].

An important class of pseudorank estimation methods is based on principal component analysis (PCA). PCA is a multivariate technique that finds new axes that span the space of the data matrix in an optimal way. The projection of the data swarm onto the first principal axis gives the best least-squares reproduction of the data in a one-dimensional space. The projection of the data swarm onto the plane spanned by the first two principal axes gives the best least-squares reproduction of the data in a two-dimensional space. In general, each axis successively accounts for a maximum amount of variation in the data by minimizing the residuals. Using PCA it is possible to retain the systematic part of the variation in the first axes, the so-called primary principal components (PCs), while most of the noise is described by the remaining axes, the so-called secondary PCs. This is the essential result of Malinowski's theory of errors for PCA [4]. PCA is also often referred to as abstract factor analysis (AFA) since it leads to an abstract decomposition of the data matrix.

It is often overlooked that determining the pseudorank is not necessarily a difficult task. In another paper [5] it is shown that parametric methods may be put to effective use if a dependable estimate of the standard deviation of the noise is available. Parametric methods have the advantage of replacing subjective decision rules by a formal significance test. This is illustrated for data matrices presented in the literature for testing the adequacy of a new pseudorank estimation method. If some considerations about the measurement error are met (i.e. uncorrelated and homoscedastic) it is possible to obtain accurate confidence levels for the primary PCs using a parametric method. Thus it may even be concluded that in favorable cases the correct procedure constitutes of applying a parametric method.

In the past several methods have been proposed that are based on the eigenvalues of PCA. In this paper the focus will be on a certain class of methods that does not depend on prior knowledge about the standard deviation of the noise and may therefore be classified as non-parametric. These methods try to exploit the essential similarity between the residuals of PCA of the test data matrix, i.e. the data matrix under consideration, and the elements of a random matrix. This similarity is assumed to be carried over to the corresponding eigenvalues of PCA. Methods based on comparison of (functions of) the eigenvalues of random matrices are commonly denoted as *parallel analysis* [6].

Several aspects have to be considered when applying such methods. For some of these aspects asymptotic results can be found in the statistical literature. These results usually concern the matrix size or the signal-to-noise ratio. However, asymptotic results derived for infinitely large matrices with infinitely large signal-to-noise ratio are not very useful for the analytical chemist who wishes to analyze a matrix of a specific size with a finite signal-to-noise ratio. Thus in order to investigate the practical implications of these theoretical results one will have to perform an evaluation for a variety of matrix sizes and signal-to-noise ratios. This evaluation is carried out by performing Monte Carlo simulations. Deviations from 'ideal behavior' resulting from finite sized data matrices with finite signal-to-noise ratio will be compared with the inherent variability of the eigenvalues given by the standard error.

Deviations from ideal behavior can be tolerated if they are (sufficiently) smaller than the standard error.

Furthermore, in practice it is not justified to make strong assumptions about the distribution of the noise. (A vast majority of the theory in multivariate statistics is developed around the assumption of normally distributed errors.) Thus it is necessary to test the influence of the distribution. The distributions that are investigated only have in common that they are symmetric around the mean. It is emphasized that homoscedastic noise is simulated. Otherwise, weighted PCA should be used [6]. Moreover, the effect of outliers will not be investigated. If outliers are expected to be important, robust estimation of PCs becomes mandatory [6]. Neglecting heteroscedasticity and outlying data simplifies reality without immediately leading to trivial or meaningless results. Many of the conclusions based on these simulations are e.g. also relevant for the multivariate detection limit very recently developed by Liang et al. [7] for chromatographic data. This detection limit is based on the resampling of so-called zero-component regions. In this way random matrices are constructed by sampling an 'experimental' distribution. Thus the simulations described in this paper can be attributed to hold a place between the restrictive normal assumption and the method of Liang et al. which is *completely* free of assumptions. Discrepancies due to the use of different distributions will also be compared to the standard error in the eigenvalues.

By combining the theoretical and numerical results, two functions of the eigenvalues that have been proposed in the past as pseudorank estimation method will be evaluated. These functions are the logarithm of the eigenvalues [8] and Malinowski's reduced eigenvalues [9]. The logarithm of the eigenvalues are reported to yield a straight line for the secondary PCs whereas the reduced eigenvalues should be constant in that region. This important property of the reduced eigenvalues has recently led to the construction of an F-test [10,11]. In this paper the assumption is tested that the reduced eigenvalues are constant for the secondary PCs. In another paper [5] the number of degrees of freedom that can be used for the F-test is discussed.

Finally, a modification is proposed for pseudorank estimation methods that are based on the principle of parallel analysis. In this modification the size of the random matrices is varied in order to account for the loss of degrees of freedom due to the systematic contribution to the data. The modification is therefore iterative in nature in contrast to the old methods where random matrices are generated that have the same size as the test data matrix.

It should be noted that throughout this paper it is assumed that the elements of the data matrix are unknown constants contaminated with measurement error [12]. This is the case for second-order data, e.g. high-performance liquid chromatography with a diode array–UV/Visible spectrophotometer as a detector: the data matrix is obtained for one chemical sample. The situation may be different if the data matrix is constructed from first-order data. In that case the row index usually corresponds to an object whereas the column index corresponds to a variable and the elements denote the observations made. Since the objects are randomly drawn from a population, an additional error is present in the resulting data matrix, the so-called sampling error (selecting other objects by chance leads to a different data matrix). The relative importance of the sampling error depends on the number of objects and the standard deviation of the measurement noise [13]. Examples of this kind of data are abundant, especially in the field of pattern recognition [14]. The work of Duewer and Kowalski [15] is one of the very few studies that involve both sampling error and measurement error. In this paper we will confine ourselves to the effect of the measurement error. For second-order data like the popular spectrochromatograms mentioned above data preprocessing other than background subtraction and selection of a time and spectral window is not customary. Thus it is assumed that the test data matrix is open and no mean centering has taken place ('covariance about the origin'). The consequences of closure and mean centering for the estimated pseudorank have recently been discussed by Pell et al. [16].

The following notation will be adopted throughout this paper. Bold upper-case letters will denote matrices, e.g. $\mathbf{M}$. Bold lowercase italic letters will denote column vectors, e.g. $\boldsymbol{v}$. Matrix and vector transposition are indicated by a superior 'T', e.g. $\mathbf{M}^{\mathrm{T}}$ and $\boldsymbol{v}^{\mathrm{T}}$. Italic letters (uppercase as well as lowercase) will denote scalars, e.g. M_{ij} is the element in row i and column j of $\mathbf{M}$. The elements of diagonal matrices, e.g. Λ_{aa} and Θ_{aa}, are denoted by lower case letters with one index indicating the position on the diagonal, e.g. λ_a and θ_a.

2. Theory

This section is organized so that first it is discussed how PCA and the related singular value decomposition (SVD) can be used to reveal the pseudorank, i.e. the essential dimension of the data matrix. Next, the difficult problem of the number of degrees of freedom in PCA is treated. In parallel analysis it is assumed that the secondary eigenvalues of the test data matrix can be approximated by the eigenvalues of a random matrix with the same size [6]. However, Mandel [17] has shown that ideally the (random) reference matrices should have the same number of degrees of freedom instead of the same size. (As a matter of fact this result will be used here to develop a modification of parallel analysis.) In the following part attention is paid to the fact that the application of such methods is always complicated by the systematic variation in the data because it affects the distribution of the secondary eigenvalues. This in turn will be of immediate consequence for theoretical predictions about the primary eigenvalues. (This insight can be seen as a useful byproduct of the current investigation.) Next, theoretical results from multivariate statistics are shown that indicate the importance of the ratio of the number of rows and columns for the distribution of the eigenvalues of a random matrix. This ratio will be denoted as the divergence coefficient d. The expected influence of the distribution of the noise is also shortly discussed. Next, three methods will be discussed that are based on the expected behavior of the eigenvalues of random matrices: the logarithm of the eigenvalues [8], Malinowski's reduced eigenvalues [9] and the F-test based on Malinowski's reduced eigenvalues [10,11]. At the end of this section, Mandel's 'reduced eigenvalues' are briefly introduced followed by an outline of the proposed modification of parallel analysis.

2.1. Principal component analysis (PCA) and singular value decomposition (SVD)

Algebraically, PCA comes down to performing an eigenvalue decomposition (EVD) of one of the cross-products of the data matrix $\mathbf{M}$, i.e. $\mathbf{M}^T\mathbf{M}$ or $\mathbf{M}\mathbf{M}^T$. If the objective of PCA is the estimation of the pseudorank, it is customary to analyze the smallest of the two matrices. Let the data points be arranged in such a way that the number of rows I is larger than the number of columns J, then PCA calculates the following decomposition:

$$\mathbf{M}^T\mathbf{M} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^T \tag{1}$$

Since $\mathbf{M}^T\mathbf{M}$ is symmetric, the columns of $\mathbf{V}$ are orthogonal eigenvectors in R^J and the diagonal elements of $\mathbf{\Lambda}$, the eigenvalues λ_a, are real numbers arranged in non-increasing order. Furthermore, in the presence of noise, $\mathbf{M}^T\mathbf{M}$ is a positive definite matrix so that the eigenvalues are all positive.

The following identities show that PCA leads to an apportionment of the total sum of squares of the data matrix to the eigenvalues:

$$\sum_{i=1}^{I}\sum_{j=1}^{J} M_{ij}^2 = \mathrm{tr}[\mathbf{M}^T\mathbf{M}] = \mathrm{tr}[\mathbf{V}^T\mathbf{M}^T\mathbf{M}\mathbf{V}] = \sum_{a=1}^{J}\lambda_a \tag{2}$$

where tr[ ] denotes the trace of a matrix. According to Malinowski [4] only the largest eigenvalues represent systematic variation whereas the remaining eigenvalues represent noise. If the signal-to-noise ratio is large, simple inspection of the size of the eigenvalues leads to a reliable estimate of the pseudorank of $\mathbf{M}$. In the past several functions of the eigenvalues have been proposed in order to facilitate this task in cases where the signal-to-noise ratio is intermediate or low.

Another group of methods has been developed that tries to exploit the characteristics of the eigenvectors v_a. Important examples are the frequency distribution of the Fourier transformed eigenvectors [18] and canonical correlation analysis [19]. The primary argument is that the eigenvectors contain more information, since the eigenvalues are merely single numbers. This argument is, however, not sufficient to unconditionally prefer the eigenvector-based methods, since the precision of an eigenvalue is better than that of the associated eigenvector. This is immediately clear if we consider e.g. the way an eigenvalue–eigenvector pair is calculated by the power method. At convergence the following holds:

$$\|\mathbf{M}^T\mathbf{M}v_a\|_E = \|z_a\|_E = \lambda_a \tag{3}$$

where $\|\ \|_E$ represents the Euclidean vector norm and z_a is the converged iterate. Since the eigenvalue λ_a is found from the Euclidean vector norm of z_a, some noise averaging will take place. Thus λ_a is expected to be more precise than the individual elements of v_a, the

normalized converged iterate. The effect should be particularly notable if $J \gg 1$ (a common situation for data in analytical chemistry) and we merely show this qualitative argument to justify why only eigenvalue-based methods are considered in this paper.

A computationally very stable alternative to the EVD of a cross-product matrix is given by the SVD of the original data matrix:

$$\mathbf{M} = \mathbf{U}\boldsymbol{\Theta}\mathbf{V}^{\mathrm{T}} = \mathbf{U}\boldsymbol{\Lambda}^{1/2}\mathbf{V}^{\mathrm{T}} \tag{4}$$

where the columns of $\mathbf{U}$ are orthonormal vectors in $\mathbb{R}^I$. (They are in fact normalized eigenvectors of $\mathbf{M}\mathbf{M}^{\mathrm{T}}$.) The diagonal elements of $\boldsymbol{\Theta}$, the singular values θ_a, are (by convention) the positive square roots of the eigenvalues λ_a. Using the SVD, the element-wise representation of $\mathbf{M}$ becomes:

$$M_{ij} = \sum_{a=1}^{J} U_{ia}\theta_a V_{ja} \tag{5}$$

Every term in the expansion of Eq. (5) successively improves the reproduction of the data according to the least-squares criterion. The SVD is useful for detecting near-dependencies (constraints) among the columns of $\mathbf{M}$. Near-dependencies are indicated by the elements of the eigenvector $\boldsymbol{v}_a$ associated with a small singular value θ_a. This is immediate from

$$\|\mathbf{M}\boldsymbol{v}_a\|_{\mathrm{E}} = \|\theta_a \boldsymbol{u}_a\|_{\mathrm{E}} = \theta_a \tag{6}$$

The vector $\mathbf{M}\boldsymbol{v}_a$ is close to the null-vector if θ_a is 'small'. For the practical worker, the key question is: how to relate the 'smallness' of θ_a to the variation in the data contributed by the measurement error? It follows that pseudorank estimation can also be interpreted as finding the number of near-dependencies in the data. The subsequent analysis of the data should preferably be executed in a space from which all near-dependencies are removed.

2.2. Number of degrees of freedom

According to Eq. (2) the eigenvalues of a cross-product matrix represent a partitioning of the total sum of squares of the data matrix. Thus according to Mandel [17] it is more appropriate to speak about the 'portion trace explained by an eigenvalue' than the 'portion variance explained by an eigenvalue' which is common practice now.

In order to test the portion variance explained by each successive PC one needs to assess the number of degrees of freedom associated to a single eigenvalue. The question of the correct number of degrees of freedom amounts to one of the most intriguing problems of multivariate statistics [6]. We will summarize the essential results from the literature and hereby use the terminology that is common in analytical chemistry. Two numbers of degrees of freedom are considered separately. Let A denote the pseudorank of $\mathbf{M}$ we wish to determine.

Total number of degrees of freedom for the residuals

First, there is the total number of degrees of freedom for the residuals. This number is $(I-A)(J-A)$ if A significant PCs are extracted from the data. (In the case row, column or grand average are subtracted from the data, modifications of this number given by Mandel [17] should be used.) Dividing the sum of squares of the residuals by this number of degrees of freedom should give an accurate estimate of the variance of the noise, σ_{M}^2:

$$\hat{\sigma}_M^2 = \frac{\sum_{a=A+1}^{J} \lambda_a}{(I-A)(J-A)} \tag{7}$$

where the hat indicates that the variance is estimated. This is confirmed by different authors [12,17,20–22]. (Strictly speaking this is an asymptotic result: see below and the Results and Discussion section.) Note that this number is used in cross-validation for the primary as well as the secondary PCs [23] [1]. (We will return to this point in the Results and Discussion section.) It is emphasized that it differs from the number that is used to evaluate the real error function [4], i.e. $I(J-A)$. Several derivations can be found in the literature. A simple proof is given by Paatero and Tapper [12]: the $I \times J$ pseudorank A data matrix is reproduced by the product of the $I \times A$ score matrix $\mathbf{S} = \mathbf{U}\boldsymbol{\Theta}$, and

[1] In this paper the focus is on the residual variation and therefore only the degrees of freedom pertinent to the residual variation are discussed. Interestingly, Wold and Sjöström introduced an empirical function in their cross-validation procedure in order to account for the decreasing number of degrees of freedom due to the extraction of *primary* PCs [24]. This perfectly makes sense, since in cross-validation the PCs are examined in decreasing order of importance, while the methods discussed in this paper proceed *backwards* through the list of PCs.

the $A \times J$ loading matrix $\mathbf{L} = \mathbf{V}^T$. This reproduction is fixed up to an $A \times A$ transformation matrix. As a result one finds for the number of free variables for the A-dimensional PC model: $I \times J - I \times A - A \times J + A \times A = (I-A)(J-A)$. A formal proof based on projection matrices is given by Mandel [25]. It follows that the real error function gives estimates for the standard deviation of the noise that are biased low compared to the estimates obtained from Eq. (7).

Number of degrees of freedom of an individual secondary PC

Second, there is the number of degrees of freedom that is associated with an individual secondary PC. Mandel states that for the portion trace explained by an eigenvalue no equivalent exists for the number of degrees of freedom well known from the additive analysis of variance (ANOVA) model [17]. However, a number of degrees of freedom can be *defined* by recognizing that the expected value of a secondary eigenvalue λ_a (as it represents a sum of squares), divided by an appropriate 'number of degrees of freedom' ν_a, should be an unbiased estimate of the error variance σ_M^2. Thus ν_a should be related to σ_M^2 and λ_a as

$$\hat{\sigma}_M^2 = \frac{E[\lambda_a]}{\nu_a} \tag{8}$$

where $E[\ ]$ denotes taking expectation. (As noted by Mandel the degrees of freedom for secondary PCs are generally not integral numbers.) Mandel's degrees of freedom are determined by simulating a large number of random matrices of appropriate size for which the individual elements are drawn from some distribution with variance $\sigma_M^2 = 1$. The eigenvalues for these matrices are averaged and the average constitutes an estimate for the expected value in Eq. (8). The precision of this estimate depends on the number of matrices that has been generated. Since $\sigma_M^2 = 1$, the average eigenvalue automatically yields an estimate for the desired number of degrees of freedom. The degrees of freedom for the leading PCs of a variety of matrix sizes have been tabulated by Mandel. These numbers were obtained by simulating normally distributed noise and averaging the eigenvalues of 625 random matrices. Inserting these degrees of freedom in Eq. (8) gives an estimate of σ_M^2 for each secondary eigenvalue. Evidently, this estimate can be improved by 'pooling' the individual estimates [17].

A correct number of degrees of freedom has many applications apart from the cross-validation mentioned above, e.g. the evaluation of the Exner function [26] and the construction of fitting criteria in curve resolution [27]. It is one of the purposes of this paper to compare the number of degrees of freedom defined by Eq. (8) and the number of degrees of freedom implied by Malinowski's reduced eigenvalues. (The discussion of Malinowski's reduced eigenvalues is deferred to a later stage.)

2.3. Influence of the systematic variation in the data (signal-to-noise ratio) on the distribution of the secondary eigenvalues and the consequences for the validity of theoretical expressions for the primary eigenvalues

Let τ denote the minimum for the following ratio of successive PCs: $(\theta_a - \theta_{a+1})/\sigma_M$ for $1 \le a \le A$ where (by definition) $\theta_a = 0$ if $a > A$. Goodman and Haberman [22] have proved that after extracting the A primary PCs the residual variation approaches a central χ^2 distribution with $(I-A)(J-A)$ degrees of freedom if τ approaches ∞. Thus in the limiting case the number of degrees of freedom for the residuals previously given as $(I-A)(J-A)$ becomes essentially correct.

Immediately the question arises how this result can be exploited in practice. By performing Monte Carlo simulations Mandel [17] has found that the distribution of the secondary eigenvalues depends only little on the value of the primary eigenvalues. This means that the secondary eigenvalues of the $I \times J$ pseudorank A test data matrix can adequately be approximated by the eigenvalues of an $(I-A)(J-A)$ random matrix. Johnson and Graybill [21] have confirmed Mandel's numerical results. In this study we will relate the adequacy of the approximate number of degrees of freedom to the value of the signal-to-noise ratio which is — contrary to τ — a typical figure of merit in analytical chemistry.

The fact that a theoretical prediction for the secondary eigenvalues such as Eq. (7) is an asymptotic result is of direct consequence for the theoretical prediction of the influence of the measurement error on the primary eigenvalues. Random measurement errors lead to a standard error and bias in the primary eigenvalues that can be predicted if an estimate of σ_M is available [22,13]. It is to be expected that these expressions will

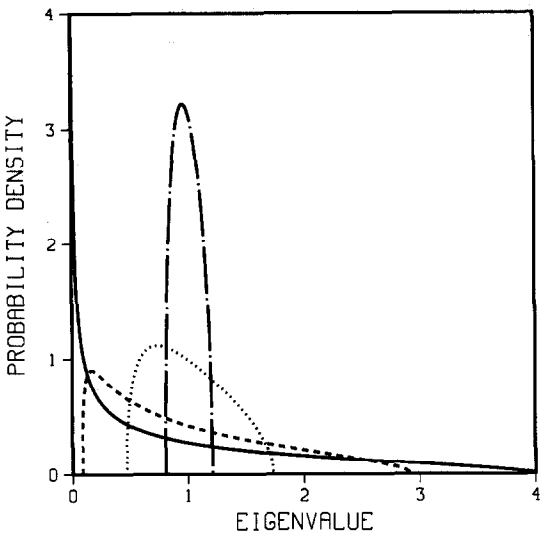

Fig. 1. Distribution of eigenvalues for divergence coefficient $d=1$ (——), 2 (– – –), 10 (· · ·) and 100 (· – · – ·).

not be valid if the distribution of the secondary eigenvalues is markedly different from the asymptotic distribution. This conjecture will be tested in the Results and Discussion section.

2.4. *Influence of the divergence coefficient d of the matrix and the distribution of the noise*

In a theoretical study of Grenander and Silverstein [28] the elements of the data matrix were allowed to take the values -1 and $+1$ exclusively (i.e. there is no systematic variation). This is the so-called random sign distribution. The cross-product matrix $\mathbf{M}^T\mathbf{M}$ was standardized by dividing all elements by the average eigenvalue. The probability density of finding *any* eigenvalue with value λ was derived for the standardized cross-product matrix of an infinitely large matrix (i.e. $I \to \infty, J \to \infty$ and $d = I/J =$ constant). It was shown that the probability density function, $f(\lambda)$, only depends on the divergence coefficient d:

$$f(\lambda) = \frac{d\sqrt{(\lambda - b_1)(b_2 - \lambda)}}{2\pi\lambda} \quad b_1 < \lambda < b_2$$
$$= 0 \quad \text{otherwise} \qquad (9)$$

The borders of the existence region of the eigenvalues are given by $b_1 = (d + 1 - 2\sqrt{d})/d$ and $b_2 = (d + 1 + 2\sqrt{d})/d$.

Plots of $f(\lambda)$ are shown in Fig. 1 for various values of d. For square matrices ($d=1$) there is a relatively large probability of finding very small eigenvalues while for very large values of d the eigenvalues start to cluster around one. This is an indication that the standardized cross-product matrix approaches the unity matrix. For intermediate values of d the effect of chance correlations is clearly visible in this plot.

The results obtained for the random sign distribution might not seem very useful for practical applications. However, from the central limit theorem it is expected that if the number of rows I is 'large enough', the elements of $\mathbf{M}^T\mathbf{M}$ (standardized or not) approach the same distribution, irrespective of the distribution of the elements of $\mathbf{M}$. In that case one would only have to assume that the elements of $\mathbf{M}$ are independently distributed with some known mean and variance. (Thus only pathological distributions for which these parameters do not exist, e.g. the Cauchy distribution, are excluded from this discussion.)

Clearly, simulations are needed to assess how large the number of rows I should be before the results become sufficiently independent of the distribution of the noise. It should be noted that the probability distribution for the eigenvalues of a matrix with normally distributed elements has received more attention in the statistics literature [29]. However, we prefer to evaluate the practical usefulness of Eq. (9) because this expression is much simpler to interpret than the expression obtained from the normal assumption.

It is e.g. straightforward to obtain an expression for the expected spacing of the eigenvalues, since it should be inversely proportional to the probability density function given above. (If the density is large, the expected spacing should be correspondingly small and vice versa.) In nuclear physics the eigenvalues of random matrices have been studied in order to estimate the energies associated to the state of a system [30]. For this kind of applications it is obvious that the spacing is an important property. However, in the case of PCA the eigenvalues and the associated spacing do not have a clear interpretation [2]. Still, it is interesting to introduce this concept because the postulated properties of the logarithm of the eigenvalues and reduced

[2] The spacing of the eigenvalues does, however, play an important part in the error analysis of PCA [13,22]. An example is the asymptotic result for the distribution of the residuals mentioned earlier.

eigenvalues (see below) can directly be interpreted as a statement about the spacing of the original eigenvalues. While the properties for logarithm of the eigenvalues and reduced eigenvalues are derived from numerical experiments there is a well-established theoretical result for the spacing. This knowledge should be used when constructing the appropriate data for the simulations. In analytical practice one expects to find considerable differences in the shape of the probability density function of the eigenvalues because d varies over a broad range for the data obtained by different techniques. For fluorescence excitation emission data one typically encounters a value around 1, while for the analysis of so-called spectrochromatograms small windows in the chromatographic mode and a large number of sensors in the spectral mode may lead to values (much) larger than 10. It follows that systematic simulations should include a large range for the expected dominating factor, i.e. the divergence coefficient d.

2.5. Logarithm of eigenvalues

Farmer [8] has found that a plot of the logarithm of the eigenvalue versus PC number may show three different regions: a straight line part of the so-called log-eigenvalue diagram which was attributed to random noise in the data, an upward deviation for the low-numbered PCs which was attributed to large scale patterns (systematic variation) and a downward deviation for the high-numbered PCs which was attributed to intercorrelations within the data. These conclusions were based on simulations of truly random data and serially correlated data (matrix size 200×30). The log-eigenvalue diagram has been introduced to analytical chemistry by Ohta [31]. The method proved to be successful in finding the correct dimension for simulated data (matrix size 64×31). Kormos and Waugh [32] applied the method to simulated data with varying signal-to-noise ratio (matrix size 520×10 and 600×10) and real data (matrix size 195×9). The results agreed with those obtained for Malinowski's indicator function [33]. It can be seen that the divergence coefficient d varies over an extreme range in these examples ($2<d<60$).

It should be noted that the straight line part in the log-eigenvalue diagram is equivalent to a constant value for the eigenvalue ratio. The eigenvalue ratio has already been thoroughly investigated by Hirsch et al. [34]. In this paper we prefer to discuss the logarithm of the eigenvalues for the following reason. In the Results and Discussion section we compile eigenvalues and standard errors for random matrices. These eigenvalues are useful for the evaluation of functions of the eigenvalues. The associated standard errors can, however, only be used to evaluate the standard error in the function of a *single* eigenvalue because the eigenvalues are not independent. In a previous paper it was shown that the eigenvalues are uncorrelated to first-order approximation [13]. The amount of correlation is described by the higher-order contributions which primarily depend on the spacing between the eigenvalues. Monte Carlo simulations showed that the correlation between the primary eigenvalues is negligible if the signal-to-noise ratio is high [13]. The same results indicate that it is certainly not negligible for the secondary eigenvalues. (This reasoning automatically applies to the eigenvalues of random matrices.) The eigenvalue ratio, however, depends on two successive eigenvalues which should be anti-correlated: if one eigenvalue rises the adjacent eigenvalues tend to decrease and vice versa. Thus it is to be expected that the eigenvalue ratio method is less stable than might be deduced from the standard errors for the individual eigenvalues. Correlations can easily be estimated by simulations but it is very bothersome to present the resulting tables in a journal article.

2.6. Malinowski's reduced eigenvalues

Recently, Malinowski presented his theory of the distribution of the secondary eigenvalues resulting from PCA [9]. Numerical experiments showed that the expected value of the eigenvalues of random matrices is proportional to $(I-a+1)\ (J-a+1)$, where a is the number of the extracted PC. A pseudorank estimation method was developed by constructing reduced eigenvalues as

$$REV_a = \frac{\lambda_a}{(I-a+1)(J-a+1)} \tag{10}$$

and comparing their relative size. The reduced eigenvalues for the secondary PCs should be constant and the primary PCs are easily distinguished, since their associated reduced eigenvalues are larger. However, it was found that the ideal behavior is only obeyed by

uniform distributed noise. For normally distributed noise one large reduced eigenvalue may be present that does not follow the proposed distribution. (This observation is not confirmed by the present study: see Results and Discussion section.) The results for spectroscopic data were in accordance with earlier work.

The relationship between Malinowski's reduced eigenvalues and the degrees of freedom defined by Eq. (8) is established as follows. Since the reduced eigenvalues should be constant for the secondary PCs, the denominator in Eq. (10) is proportional to the number of degrees of freedom of the eigenvalue. The proportionality constant N is found by recognizing that summing these numbers of degrees of freedom should give $(I-A)(J-A)$, the total number of degrees of freedom for the secondary PCs, so

$$N=\frac{(I-A)(J-A)}{\sum_{a=A+1}^{J}(I-a+1)(J-a+1)} \tag{11}$$

It follows that N can only be determined if A is known. Evidently, this is a problem in practice. However, in this paper simulations are used for the comparison of the different expressions for the degrees of freedom and, consequently, this problem does not exist.

The concept of reduced eigenvalues has proved to be useful for pseudorank estimation but it has much wider application. There are many expressions where an estimation of the magnitude of the secondary eigenvalues must be made in order to obtain a final result in closed form. The parametrization of Malinowski could be used without modification by the factor N in instances where the ratio of eigenvalues is important. An example of this kind is the expression for the rate of convergence of the non-linear iterative partial least squares (NIPALS) algorithm [35]. The NIPALS algorithm is a popular method for the calculation of a preselected number of PCs. Usually this preselected number is relatively small and the rate of convergence is a property of interest if the NIPALS algorithm has to be compared with another method with respect to the number of floating point operations to be expected.

2.7. *F-test based on Malinowski's reduced eigenvalues*

Malinowski [10,11] noticed that the reduced eigenvalues can be compared in an F-test because the associated eigenvectors are independent. It was found that the 5% level tends to underestimate, whereas the 10% level tends to overestimate the number of primary PCs. In another paper the number of degrees of freedom that can be used for this F-test is discussed [5].

2.8. *Mandel's 'reduced eigenvalues'*

It is interesting to note that Mandel [17] constructed 'reduced eigenvalues' by dividing the experimental eigenvalues by the eigenvalues obtained for random matrices. As mentioned above, this procedure, although essentially correct, is afflicted with a fundamental problem. The random matrices should have the same number of degrees of freedom as the test data matrix but the number of degrees of freedom depends on the true dimension of the test data matrix, which is just the parameter we want to determine. This calls for an iterative approach that comes down to a modification of parallel analysis (see below). It is worth mentioning that Mandel [17] also considered the use of an F-test. However the behavior of his 'reduced eigenvalues' was found to leave little doubt about the essential dimension in most practical applications.

2.9. *Modification of parallel analysis*

Mandel's 'reduced eigenvalues' have a sound theoretical basis but they should only be trusted if they are obtained from reference matrices with the same number of degrees of freedom. Thus a correct procedure for pseudorank estimation seems to be as follows. Given the $I\times J$ test data matrix one should generate reference matrices of size $(I-a)\times(J-a)$ where a takes all values that are compatible with the experiment. In order to reduce the amount of work, one could use an initial guess from another method. It is e.g. well known that the indicator function [33] often exhibits a minimum that is shallow and one could take as initial guess all dimensions that can hardly be distinguished from this minimum. A good initial guess may also be obtained from the original procedure of Mandel, i.e. examine the 'reduced eigenvalues' obtained from random matrices with the same size. Next, A is taken to be the value of a for which the eigenvalues obtained from the reference matrices yield the best match with the smallest $J-a$ eigenvalues of the test data matrix.

Additional support for the final choice may be obtained from the observation that Eqs. (7) and (8) yield two independent estimates for the standard deviation of the measurement error. It is seen that the estimate provided by Eq. (7) is based entirely on the test data while Eq. (8) uses external information, since the degrees of freedom ν_a are the average eigenvalues of random matrices. It seems therefore appropriate to base the final choice on consistency between the two estimates. It should, however, always be kept in mind that Eq. (7) gives estimates that are biased low. This will especially constitute a problem in the most interesting case, i.e. the situation where the signal-to-noise ratio is small.

3. Experimental

Random matrices are simulated in order to test the conjectures about the various influences on the value of the secondary eigenvalues of a test data matrix. The outcome of these simulations can, however, also be used to derive confidence levels for the primary PCs of a test data matrix as illustrated by the resampling method of Liang et al. [7]. Furthermore, we analyze one data matrix taken from the literature that is characterized by a large loss of degrees of freedom. This data matrix should therefore be ideally suited for demonstrating the consequences of using the eigenvalues of a random matrix with the same size instead of the same number of degrees of freedom for pseudorank estimation.

3.1. Random matrices

The divergence coefficient d is varied over a range of 1 to 10 by varying the number of rows I from 10 to 100 while keeping the number of columns J fixed to 10. The influence of the matrix size is investigated by constructing 20×10, 20×20, 40×20 and 40×40 matrices. Experimental noise is simulated according to the normal, the uniform and the random sign distribution. These distributions are expected to cover a large number of experimental distributions because some fundamental properties are varied. The normal and the uniform distribution are continuous while the random sign distribution is not. Furthermore, the normal distribution has an infinite range in contrast to the other two. All distributions have in common that they are symmetric around the mean, i.e. they all can be appropriately described with one standard deviation σ_M. σ_M will be 1 for all simulations. This means that the uniform distribution has a range of $\sqrt{3}\approx1.732$. The elements of a matrix generated according to the random sign distribution are restricted to have the values -1 and $+1$. Expected values for the eigenvalues and their standard error are estimated by the average obtained for Monte Carlo samples composed of 10 000 matrices. It should be noted that in most cases the standard error in the eigenvalues of a single matrix is reported. The standard error in the average eigenvalue of 10 000 matrices is a factor 100 smaller.

3.2. Literature data matrix

The data matrix taken from the literature consists of simulated mass spectra [36]. The size of the matrix is 20×10 and it corresponds exactly to the size of the random matrices used to sample the eigenvalue distribution for $d=2$. The true dimension of this data set is known to be 5, i.e. one has 200 data points and 125 parameters to be estimated by PCA. The 'measurement noise' is introduced by rounding the errorless data points to the nearest integer. As a result one would expect the noise to be uniformly distributed with range 0.5 and standard deviation $\sigma_M=1/6\sqrt{3}\approx0.289$. However, the residuals left after extracting five PCs lead to an estimated value $\hat{\sigma}_M=0.588$. (This value is obtained by dividing the total sum of squares of the residuals by $(I-A)(J-A)=75$ according to Eq. (7).) We have no explanation for the discrepancy between expected and estimated standard deviation but it is assumed to be of minor importance for the purpose of this research.

4. Results and discussion

In this section the systematic deviation from ideal behavior is always compared to the standard errors in the eigenvalues. Since these standard errors are inevitable in practice, the necessary playground is provided where asymptotic results may be assumed to be true. It should be evident that the discussion becomes academic if e.g. the effect of using the wrong dimension

for the reference matrix is small compared to the standard error in the eigenvalues resulting from noise [3].

4.1. Random matrices

The key assumption investigated in this study is that the secondary eigenvalues of the test data matrix can be approximated by the eigenvalues of an appropriately sized random matrix. At this point the influence of the distribution of the noise is not important because it can be assumed to be adequately simulated. A more disturbing factor is the systematic variation in the data because it is different for each test data matrix. Thus from the point of view of this research it is more justified to consider the stochastic contribution to the total signal, i.e. the measurement noise, as 'systematic' because the expected value of this contribution is (approximately) the same for each test data matrix.

Influence of the systematic variation in the data (signal-to-noise ratio) on the distribution of the secondary eigenvalues and the consequences for the validity of theoretical expressions for the primary eigenvalues

First, the influence of the systematic variation in the data on the expected value of the secondary eigenvalues is investigated. Furthermore attention will be paid to the consequences for the predicted bias and standard error in the primary eigenvalues. This is achieved by constructing a simple one-component system. To the elements of a 20×10 random matrix we add a constant systematic contribution.

The results of PCA are given in Table 1. In the first column the size of the elements of the error-less data is given. Since $\sigma_M = 1$, this number in fact constitutes the signal-to-noise ratio, in the sequel denoted by ρ.

The second column lists the value for $\hat{\sigma}_M$ estimated from the residuals of the correct PC model according to Eq. (7). For $\rho = 0$ the estimate is based on a zero-dimensional model while for the other values of ρ, it is based on a one-dimensional model. The figure in parentheses denotes the standard error in the Monte Carlo sample average. It is seen that σ_M is correctly reproduced for the random matrices ($\rho = 0$) while for the other levels of ρ the input value ($\sigma_M = 1$) is systematically underestimated. As predicted by Goodman and Haberman [22] the residual variation approaches the limiting distribution with increasing signal-to-noise ratio. In practice one should compare this 'bias' to the precision to which these numbers can be obtained for a *single* test data matrix. The standard errors for one matrix are larger than the standard errors for the sample average by a factor 100 and it is easily verified that for this specific example the improvement going from $\rho = 0.5$ to $\rho = 1$ is negligible, i.e. the estimated standard deviation $\hat{\sigma}_M$ may be considered to be constant [4]. (It is worth mentioning that the real error function underestimates $\hat{\sigma}_M$ by a factor $\sqrt{19}/20 \approx 0.975$.)

Table 1
Eigenvalues of a 20×10 matrix with constant elements ρ and normally distributed noise added. The figure in parentheses denotes the standard error in the average (expressed in units of the last reported digit)

ρ	$\hat{\sigma}_M$	λ_1	λ_2	λ_3
0.0	1.0005(5)	49.2(1)	37.74(5)	29.85(4)
0.5	0.9863(5)	82.9(2)	43.97(7)	33.26(5)
1.0	0.9965(5)	230.6(3)	45.05(7)	34.01(5)
2.0	0.9976(5)	829.7(6)	45.24(7)	34.14(5)
3.0	0.9984(5)	1828.3(9)	45.27(7)	34.07(5)

The third column gives the first eigenvalue obtained from PCA. In the absence of noise it should be equal to the total systematic variation in the data. Thus in the absence of noise one would expect to find $\lambda_1 = 0$, 50, 200, 800 and 1800, respectively. It can be seen that for the data matrices containing systematic variation the eigenvalues are biased upwards. Goodman and Haberman [22] have derived a theoretical expression for the expected bias in the first eigenvalue of a data matrix that has been corrected for row, column and grand average. It will be shown in a future publication that

[3] Very recently Liang et al. [7] have proposed a non-parametric multivariate limit of detection based on the analysis of reference matrices with the same size. The method is especially constructed to work in the presence of correlated noise. Correlated noise tends to increase these standard errors [7]. Thus it may turn out that for uncorrelated noise a significant effect would be predicted from using the wrong dimensions while Liang's detection limit is still correct.

[4] It is clear that this point defines the detection limit for this data matrix. For univariate calibration the detection limit is usually defined as the concentration of the analyte for which the signal-to-noise ratio is *three*. The results obtained in this study for second-order data show that extension of this definition to higher-order data is not so straightforward as implied by Wang et al. [1]. It should be noted that our (limited) results are confirmed by similar results obtained by Stewart [37] for an index which is inversely proportional to Lorber's multivariate signal-to-noise ratio [38].

without this data preprocessing the bias in the first eigenvalue is given by

$$b_\lambda = \lambda - \lambda_{\text{true}} = (I+J-1)\sigma_M^2 \qquad (12)$$

According to Goodman and Haberman the adequacy of this prediction depends on the value σ_M/θ_1. Thus the predicted bias is independent of the eigenvalue itself and in this specific example equal to 29. The error in the predicted bias ($=29-32.9=-3.9$) is much smaller than the standard error ($=20$) for $\rho=0.5$, i.e. the prediction is already useful. Furthermore, the predicted value is seen to be approached from above. Again, the improvement going from $\rho=0.5$ to $\rho=1$ should be compared to the expected experimental precision of the eigenvalues for a single test matrix. Consequently the same conclusion is arrived at as for the estimated σ_M: the signal-to-noise ratio should be approximately 0.5 before the asymtotic limit is virtually reached. For $\rho=0.5$, the level for which the approximation starts to work well, one finds $\sigma_M/\theta_1=0.11$, which is a reasonable value since $0.11 \ll 1$. Although this quantity is important for estimating the size of approximation errors we think that it is more convenient for the analytical chemist to express the level for which the theory predicts well in terms of the signal-to-noise ratio.

The fourth and fifth column list the second and third eigenvalue respectively. It can be seen that these eigenvalues also start to approach their limiting values for $\rho=0.5$. (The limit is approached even faster for the higher-numbered eigenvalues not shown here.) A note can be made about the general pattern that is displayed by the eigenvalues. Golub [39] has derived a formula for the updated eigenvalues after a so-called rank-one modification of a matrix. This formula has led to the development of fast updating algorithms that have found their use in e.g. the cross-validation procedure of Eastment and Krzanowski [24]. According to Golub's result the new eigenvalues interleave the old ones. Thus after adding systematic variation to the random elements, the second eigenvalue of the 'updated' matrix is bracketed by the first and second eigenvalue of the original random matrix, and so on. It immediately follows that the first secondary eigenvalue of the test data matrix will not equal the first eigenvalue of a random matrix of the same size. However, in practice only the following question is relevant: in how far is it justified to approximate the first secondary eigenvalue by the first eigenvalue of a random matrix with the same *size* instead of *number of degrees of freedom*? From simulations (analogous to the ones previously described) the expectation of the first eigenvalue of a 19×9 random matrix is found to be 45.4. We note that in our example the difference between the 'correct' and the simple 'substitute' value (taken from Table 1) is $49.2-45.4=3.8$. Again, this difference should be compared with the precision of the eigenvalue that is to be scrutinized for significance.

Table 2
Standard errors in the first eigenvalue of a 20×10 matrix with constant elements ρ and normally distributed noise added

ρ	Predicted	Monte Carlo	Relative error (%)
0.0	14.0	7.2	94
0.5	18.2	15.4	18
1.0	30.4	29.5	3.1
2.0	57.6	56.2	2.5
3.0	85.5	85.1	0.5

In Table 2 the standard errors in the first eigenvalue are displayed in more detail. In the second and third column we compare the predicted and Monte Carlo value. The theoretical prediction is based on the following expression [22,13]:

$$\sigma_\lambda = 2\lambda^{1/2}\sigma_M \qquad (13)$$

In contrast to the predicted bias, the standard error depends on the size of the eigenvalue itself. From the relative error in the fourth column it is seen that also this prediction works well for $\rho \geq 0.5$. (For $\rho<0.5$ the expression derived by a first-order approximation constitutes a gross overestimate.)

It is immediate that in this specific example the standard error in the eigenvalues obtained for a single matrix is much larger than the difference between the correct and the substitute estimate of the reference eigenvalue. Here, the simple substitution is certainly justified and pseudorank estimation by parallel analysis should work without the proposed modification. It should, however, be noted that for this example the loss of degrees of freedom is relatively small. The advantage of large intrinsic standard errors is not automatically copied to other situations. This example was primarily constructed in order to investigate the influence of the signal-to-noise ratio on the distribution of the secondary eigenvalues.

Influence of the divergence coefficient d of the matrix

From the preceding (lengthy) discussion it has become clear that down to a very small signal-to-noise ratio the secondary eigenvalues of a test matrix are virtually equal to reference values obtained from appropriately sized random matrices. Thus we have simulated random matrices and compiled the eigenvalues and their standard errors for random matrices for a wide range of d in Tables 3 and 4. The matrix elements are generated according to the normal distribution. These numbers can be conveniently used to evaluate functions of the eigenvalues and their standard error. (The tables given by Mandel [17] are restricted to the largest eigenvalues.) Of the few observations worth mentioning the presence of one very small and instable eigenvalue for the 10×10 matrix is particularly striking. Furthermore, the relative standard error is seen to decrease with increasing matrix size and increase with increasing PC number.

Number of degrees of freedom for a secondary principal component

Using the eigenvalues from Table 3 it is possible to investigate the different numbers of degrees of freedom proposed in the past for the individual secondary PCs. Fig. 2 shows the relevant numbers for a random matrix of varying size and divergence coefficient. The first two numbers, I and $I+J-2a+1$, are based on the numbers for the total residual variation, $I(J-A)$ and $(I-A)(J-A)$. The other two numbers, $N\times(I-a+1)(J-a+1)$ and $\lambda_a/\sigma_M{}^2$, are related to Malinowski's and Mandel's reduced eigenvalues, respectively. The 'reduced eigenvalues' of Mandel are based on a sound statistical argument and can be seen as a canonical limit. The other three numbers should be interpreted as approximations. It is found that among the three approximations, Malinowski's reduced eigenvalues generally perform best. For the 20×10, 20×20, 40×20 and 40×40 matrices the loss of degrees of freedom is underestimated for most PCs but there is a cross-over point after which the loss of degrees of freedom is overestimated (see Fig. 2a–d). For the 50×10 and 100×10 matrices the loss of degrees of freedom is overestimated for all PCs (see Fig. 2e,f). There is clearly a systematic deviation but the difference with the target values is usually much smaller than the difference with the competing values. Furthermore, a comparison of the plots for matrix size 20×10 and 20×20 (Fig. 2a,b) with the plots obtained after doubling the matrix size (Fig. 2c,d) shows that the agreement does not seem to improve by going to larger matrix sizes.

Table 3
Eigenvalues of a random matrix with normally distributed elements

Size	PC_1	PC_2	PC_3	PC_4	PC_5	PC_6	PC_7	PC_8	PC_9	PC_{10}
10×10	32.14	22.47	16.18	11.47	7.83	5.06	2.94	1.45	0.50	0.07
11×10	33.86	24.07	17.62	12.78	8.97	5.97	3.68	1.99	0.84	0.20
12×10	35.66	25.66	19.00	13.96	10.05	6.90	4.42	2.57	1.22	0.37
13×10	37.48	27.31	20.51	15.27	11.17	7.87	5.22	3.16	1.63	0.58
14×10	39.28	28.86	21.89	16.53	12.21	8.75	5.96	3.76	2.06	0.84
15×10	41.02	30.39	23.24	17.72	13.32	9.70	6.77	4.38	2.52	1.10
16×10	42.67	31.91	24.57	18.98	14.39	10.61	7.55	5.02	2.99	1.41
17×10	44.26	33.37	25.88	20.12	15.42	11.54	8.34	5.68	3.50	1.72
18×10	46.11	34.93	27.24	21.37	16.51	12.53	9.17	6.33	4.01	2.07
19×10	47.68	36.33	28.62	22.55	17.56	13.42	9.91	6.97	4.51	2.41
20×10	49.22	37.74	29.85	23.71	18.64	14.38	10.77	7.72	5.09	2.82
30×10	64.62	51.73	42.69	35.40	29.24	23.89	19.19	14.96	11.09	7.37
40×10	79.27	65.01	54.99	46.77	39.65	33.43	27.81	22.65	17.74	12.81
50×10	93.25	78.12	67.06	58.03	50.11	43.00	36.60	30.57	24.80	18.77
60×10	106.8	90.47	78.70	68.90	60.38	52.60	45.46	38.77	32.16	25.16
70×10	120.3	103.1	90.64	80.09	70.86	62.41	54.67	47.18	39.77	31.85
80×10	133.1	115.1	101.9	90.82	81.00	71.98	63.59	55.51	47.46	38.67
90×10	146.3	127.3	113.4	101.7	91.31	81.76	72.74	64.10	55.42	45.82
100×10	158.9	139.2	124.8	112.6	101.5	91.48	81.95	72.81	63.49	53.11

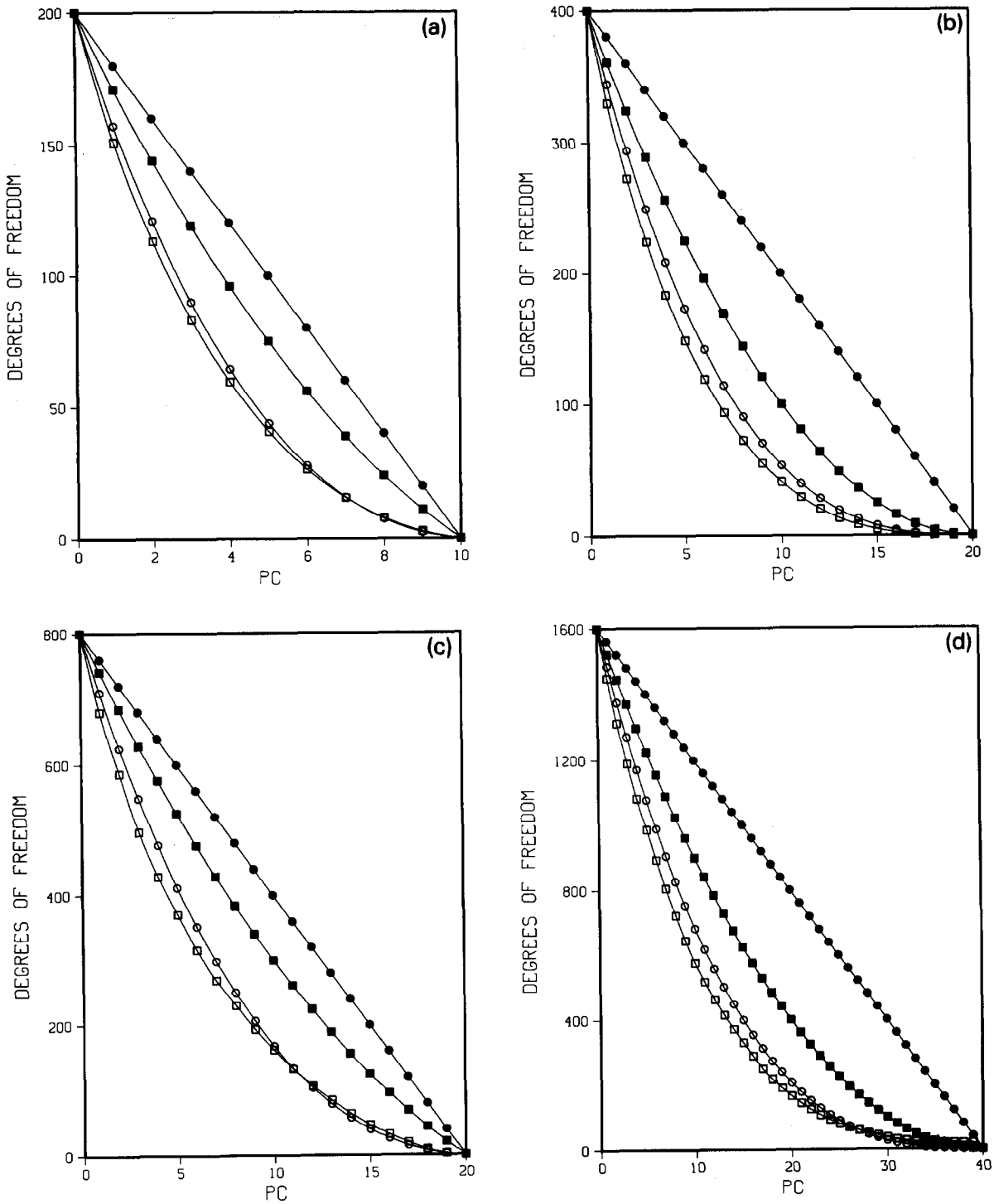

Fig. 2. Number of degrees of freedom left after extracting a PCs for (a) 20×10, (b) 20×20, (c) 40×20, (d) 40×40, (e) 50×10 and (f) 100×10 matrix. The number of degrees of freedom for the ath PC is estimated by I (●), $I+J-2a+1$ (■), $N(I-a+1)(J-a+1)$ (○) and λ_a/σ_M^2 (□).

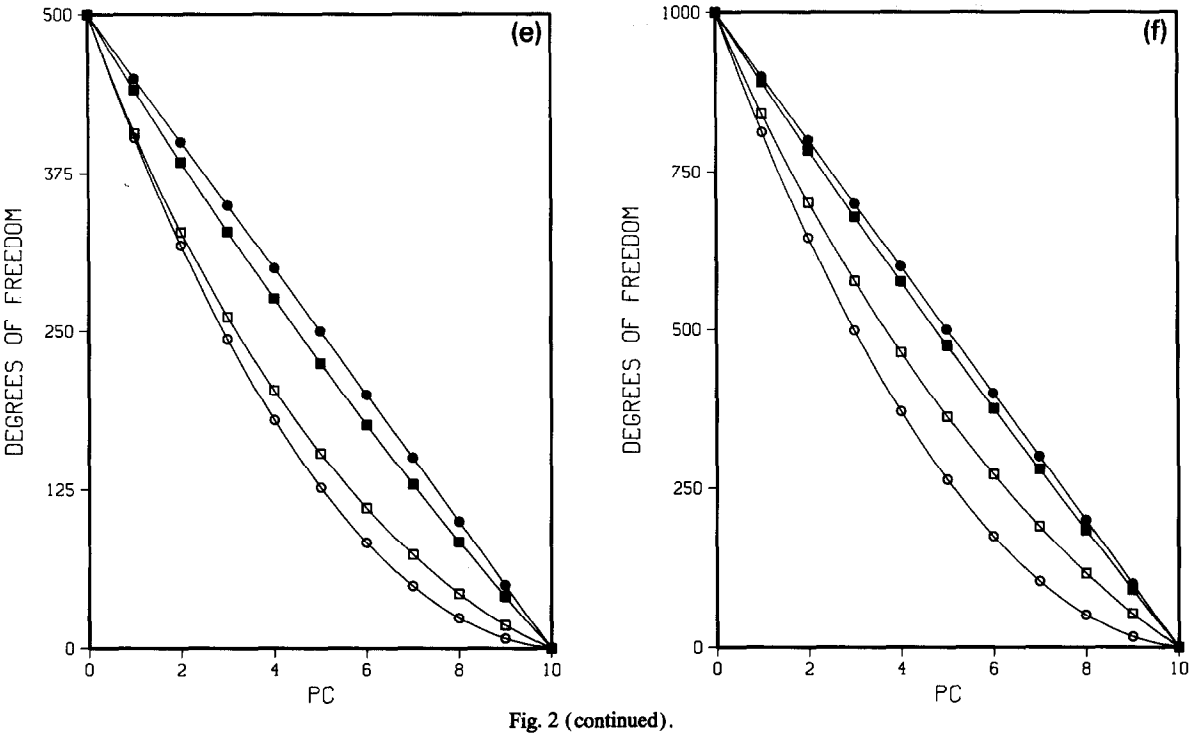

Fig. 2 (continued).

It is emphasized that for some important applications, e.g. cross-validation, the relevant quantity is a ratio where a number of degrees of freedom is substituted in the numerator as well as in the denominator. As a result errors of any kind will tend to cancel out in the final result and therefore the accuracy of the inserted number is not necessarily critical. (Application of Malinowski's reduced eigenvalues to cross-validation has not yet been reported to the authors' knowledge.)

Influence of the distribution of the noise

The influence of the distribution of the noise is shown in Figs. 3 and 4. In Figs. 3 and 4 the logarithm of the eigenvalues and Malinowski's reduced eigenvalues are plotted for matrices with normal, uniform and random sign distribution. The matrix sizes are equal to the ones just discussed. It is seen that the differences due to a different distribution are small, especially for the large matrices.

The logarithm of the eigenvalues is found to be on a straight line for the low-numbered PCs. A downward deviation for the high-numbered PCs occurs in all cases, although for the matrices with largest divergence coefficient (see Fig. 3e,f) the deviation is relatively small [5]. It follows that Farmer's result [8] (downward deviation is due to intercorrelations within the data) is not confirmed by these simulations. However, since the log-eigenvalue diagram is exclusively used to extrapolate towards the low-numbered PCs, this part of the plot is not particularly interesting anyway. Thus the log-eigenvalue diagram seems to provide a valid pseudorank estimation method over the range of matrix sizes considered in Fig. 3 if the number of primary PCs is not too large.

The reduced eigenvalues displayed in Fig. 4 show very different patterns. In all cases we find a systematic deviation from the ideal behavior. For $d=1$ the low-numbered PCs have reduced eigenvalues that are too high whereas the high-numbered PCs are characterized by reduced eigenvalues that are too low (see Fig. 4b,d). In the worst case the reduced eigenvalues differ by a factor of five. The situation is rather different for the other cases where the low-numbered PCs have reduced eigenvalues that are rather constant while the reduced eigenvalues for the high-numbered PCs are too high

[5] Different behavior is to be expected from the varying shapes of the distribution functions displayed in Fig. 1.

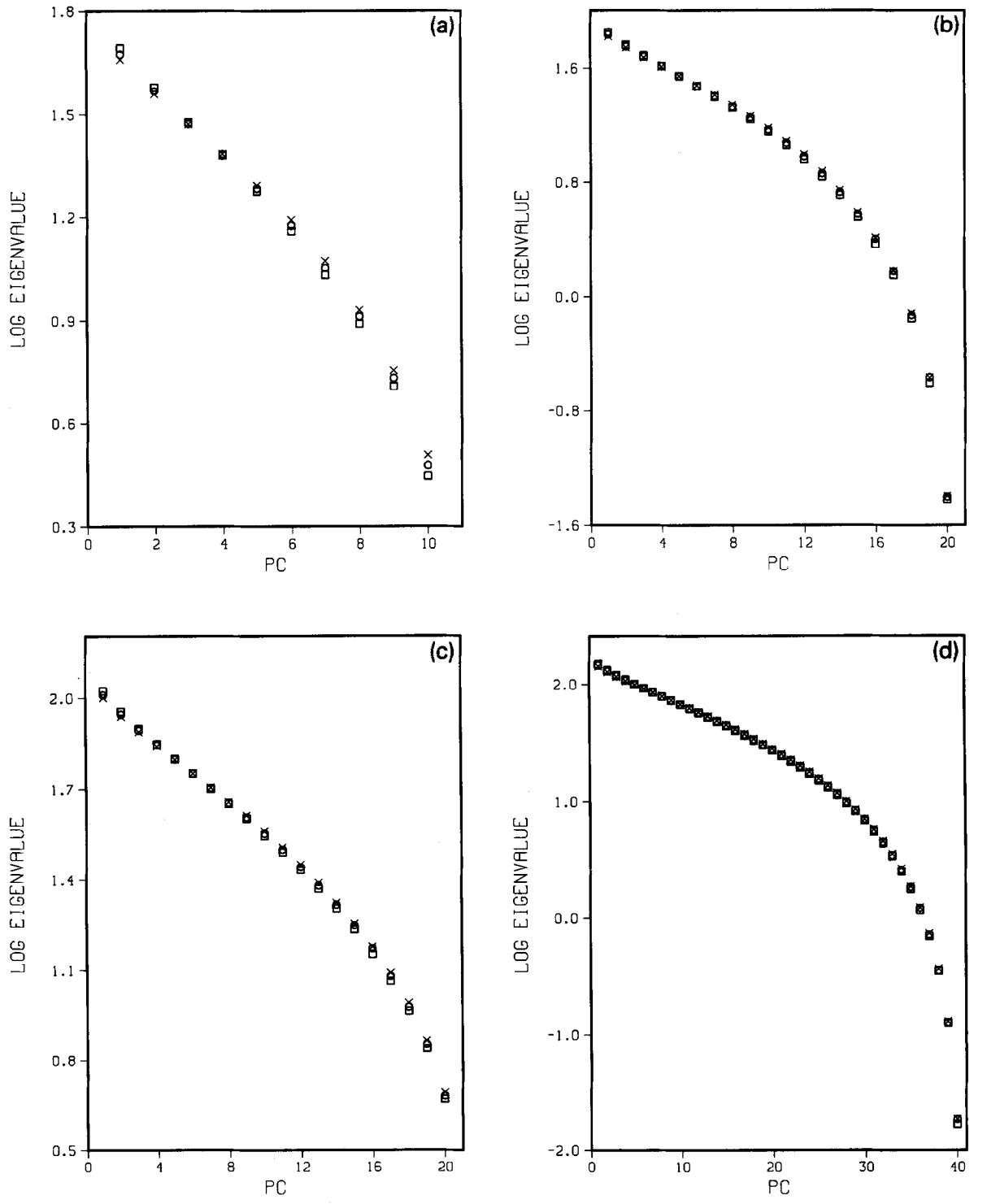

Fig. 3. Logarithm of eigenvalues for (a) 20×10, (b) 20×20, (c) 40×20, (d) 40×40, (e) 50×10 and (f) 100×10 random matrix with normal (□), uniform (○) and random sign (×) distribution.

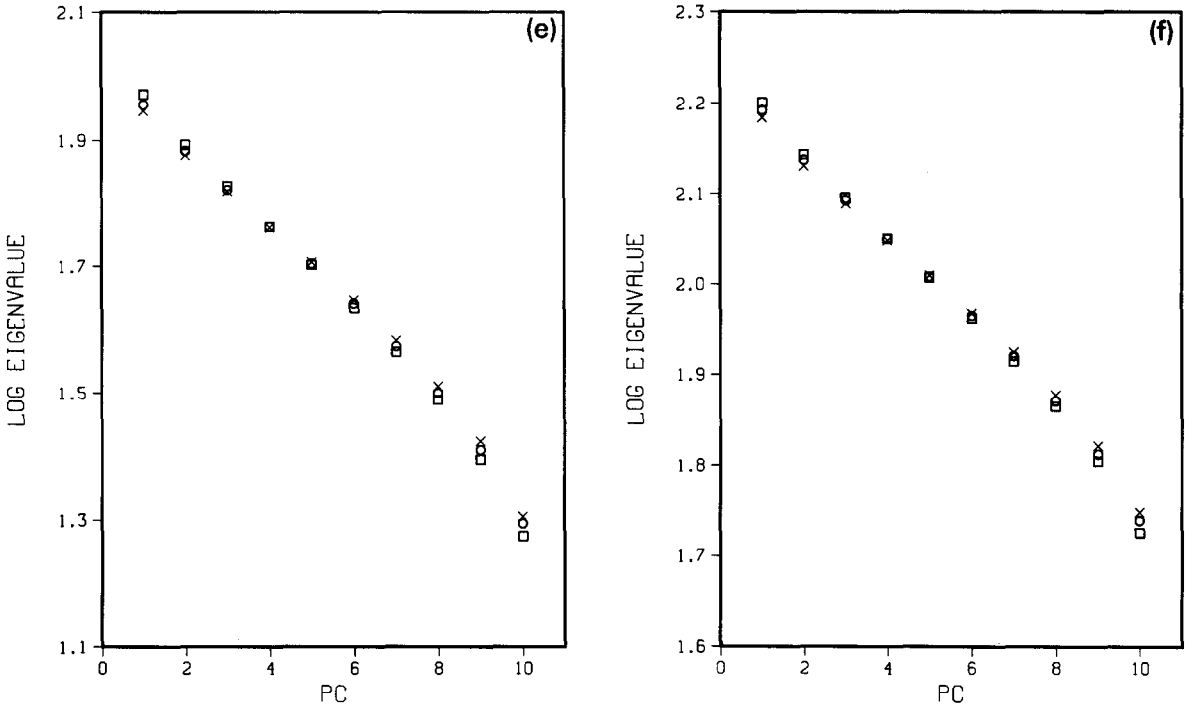

Fig. 3 (continued).

(see Fig. 4a,c,e,f). It should, however, be noted that from a numerical point of view even the worst case displayed here may be very acceptable, i.e. one might still obtain useful results after introducing Malinowski's reduced eigenvalues in theoretical equations. Furthermore, contrary to the results reported by Malinowski the same behavior is observed for random matrices generated according to the uniform and normal distribution. This finding extends the applicability of Malinowski's parametrization.

Using the standard errors from Table 4 it is possible to quantify the preceding statements obtained from plots. Thus it is easily verified that within the associated standard error the logarithm of the eigenvalues is lying on a straight line for a substantial part of the plot (we have not plotted the corresponding error bars because it does not improve the visibility). The situation is more complicated for Malinowski's reduced eigenvalues.

Table 5 gives the difference of the individual reduced eigenvalues with respect to the average reduced eigenvalue in units of the standard error of the particular reduced eigenvalue. Deviations from the 'ideal' behavior are easily tolerated if they are much smaller than the standard error. In that case one would not notice the difference in practice. The relative deviations tend to be large for the low-numbered PCs, since the standard error in the eigenvalue is relatively small then. This is an unfavorable situation since it concerns the interesting region. The results in Table 5 clearly show the range of d where Malinowski's reduced eigenvalues will provide a valid pseudorank estimation method. It is immediate that the deviations are too large for $d \geq 3$. Furthermore, it turns out that 20×10 is the only matrix size in this investigation that gives a systematic deviation smaller than the standard error for *all* PCs. From these results it seems dangerous to use the reduced eigenvalues (and an associated F-test) as a pseudorank estimation method for a particular matrix size without performing simulations in the direct neighborhood. At the same time these results show that it is likely that many other matrix sizes can indeed be found that follow the desired pattern.

4.2. Literature data matrix

From the results of the simulations it has become clear that the size of the literature data matrix (20×10)

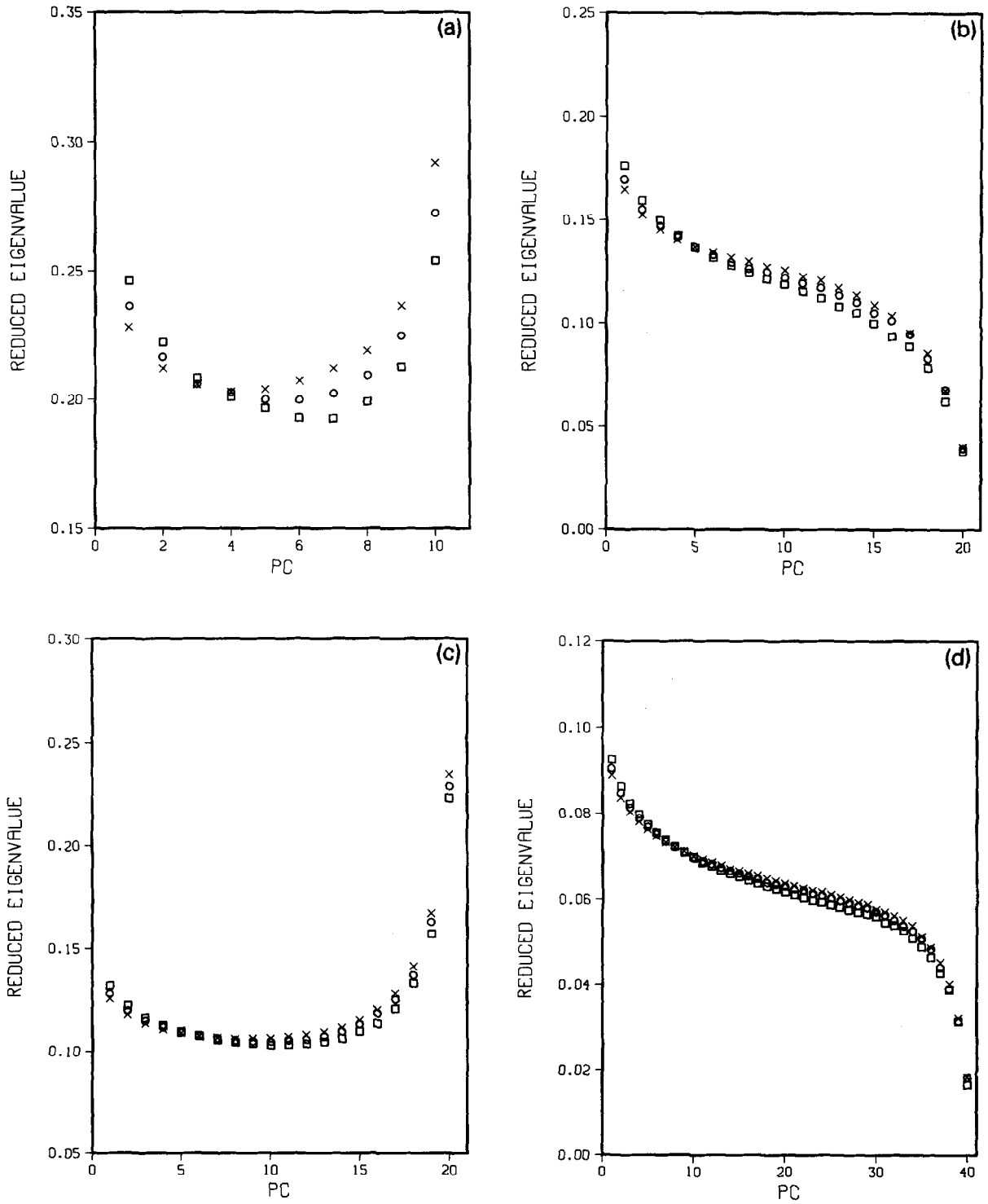

Fig. 4. Reduced eigenvalues for (a) 20×10, (b) 20×20, (c) 40×20, (d) 40×40, (e) 50×10 and (f) 100×10 random matrix with normal (□), uniform (○) and random sign (×) distribution.

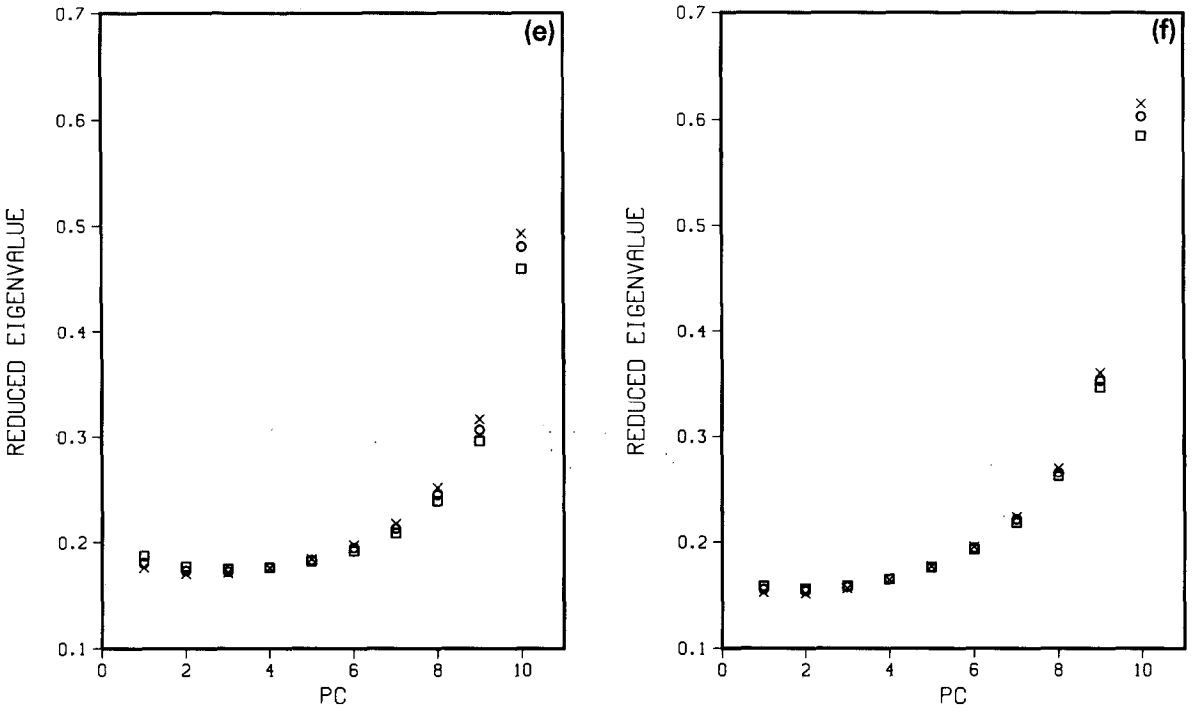

Fig. 4 (continued).

Table 4
Standard errors in the eigenvalues of a random matrix with normally distributed elements

Size	PC_1	PC_2	PC_3	PC_4	PC_5	PC_6	PC_7	PC_8	PC_9	PC_{10}
10×10	6.14	4.11	3.12	2.44	1.86	1.43	1.02	0.68	0.35	0.11
11×10	6.22	4.23	3.19	2.53	2.00	1.54	1.16	0.79	0.47	0.20
12×10	6.40	4.34	3.30	2.63	2.10	1.66	1.26	0.90	0.59	0.30
13×10	6.53	4.48	3.44	2.72	2.20	1.76	1.38	1.01	0.70	0.39
14×10	6.56	4.55	3.52	2.85	2.33	1.88	1.47	1.12	0.80	0.49
15×10	6.90	4.66	3.70	2.98	2.44	1.98	1.58	1.22	0.90	0.59
16×10	6.88	4.71	3.76	3.07	2.52	2.03	1.66	1.31	0.99	0.67
17×10	6.83	4.82	3.81	3.11	2.58	2.14	1.74	1.41	1.09	0.77
18×10	7.10	4.94	3.90	3.18	2.67	2.23	1.84	1.50	1.18	0.86
19×10	7.11	5.00	3.99	3.32	2.76	2.30	1.91	1.56	1.26	0.95
20×10	7.17	5.10	4.07	3.33	2.82	2.38	1.99	1.66	1.35	1.05
30×10	8.08	5.88	4.79	4.08	3.50	3.05	2.71	2.36	2.08	1.85
40×10	9.01	6.51	5.43	4.65	4.10	3.69	3.29	2.98	2.71	2.57
50×10	9.50	7.14	5.92	5.17	4.60	4.13	3.81	3.48	3.26	3.19
60×10	10.2	7.50	6.32	5.53	5.02	4.57	4.23	3.94	3.70	3.77
70×10	10.5	8.05	6.81	6.01	5.41	4.93	4.62	4.41	4.25	4.31
80×10	11.0	8.29	7.12	6.24	5.80	5.41	4.99	4.74	4.63	4.77
90×10	11.7	8.79	7.54	6.80	6.17	5.68	5.34	5.18	5.04	5.30
100×10	11.9	9.13	7.93	7.09	6.42	6.14	5.72	5.53	5.38	5.68

makes it particularly suited for the analysis with both functions of the eigenvalues previously considered, i.e. the logarithm of the eigenvalues and the reduced eigenvalues.

Functions of the eigenvalues

The results of PCA for the literature data matrix are given in Table 6. The first column lists the number of the PC under consideration. The large jump in the eigenvalues in the second column clearly points in the direction of a five-dimensional PC model. According to Eq. (13) the estimated standard error for the last primary PC is $\sigma_\lambda = 2 \times \sqrt{2422} \times 0.588 = 57.9$. This number should be compared to the gap with the first secondary eigenvalue. Thus the extreme significance of this model is established.

The log-eigenvalue diagram is shown in Fig. 5. It is important to note that the choice of a five-dimensional model from the log-eigenvalue diagram cannot be based on a straight part in the plot since there are too many primary PCs. Instead, this conclusion should now

Table 5
Relative deviation with respect to the mean of the reduced eigenvalues of a random matrix with normally distributed elements. Relative deviations smaller in absolute value than 1 are marked in bold

Size	PC_1	PC_2	PC_3	PC_4	PC_5	PC_6	PC_7	PC_8	PC_9	PC_{10}
10×10	1.90	1.43	**0.99**	**0.59**	**0.25**	**−0.04**	**−0.33**	**−0.58**	**−0.90**	−1.26
11×10	1.82	1.33	**0.90**	**0.51**	**0.18**	**−0.12**	**−0.36**	**−0.60**	**−0.82**	−1.06
12×10	1.72	1.23	**0.78**	**0.39**	**0.09**	**−0.17**	**−0.41**	**−0.57**	**−0.72**	**−0.82**
13×10	1.63	1.12	**0.68**	**0.30**	**0.01**	**−0.23**	**−0.41**	**−0.55**	**−0.62**	**−0.63**
14×10	1.57	1.02	**0.57**	**0.20**	**−0.10**	**−0.31**	**−0.46**	**−0.52**	**−0.54**	**−0.41**
15×10	1.43	**0.90**	**0.44**	**0.09**	**−0.17**	**−0.35**	**−0.45**	**−0.50**	**−0.44**	**−0.24**
16×10	1.36	**0.79**	**0.32**	**−0.00**	**−0.25**	**−0.43**	**−0.48**	**−0.47**	**−0.35**	**−0.08**
17×10	1.27	**0.67**	**0.20**	**−0.12**	**−0.35**	**−0.47**	**−0.50**	**−0.42**	**−0.25**	**0.07**
18×10	1.16	**0.56**	**0.08**	**−0.23**	**−0.43**	**−0.51**	**−0.50**	**−0.41**	**−0.17**	**0.20**
19×10	1.08	**0.45**	**0.00**	**−0.31**	**−0.50**	**−0.57**	**−0.54**	**−0.39**	**−0.10**	**0.32**
20×10	**0.96**	**0.30**	**−0.16**	**−0.45**	**−0.60**	**−0.63**	**−0.55**	**−0.32**	**0.00**	**0.47**
30×10	**−0.13**	**−0.92**	−1.33	−1.47	−1.41	−1.14	**−0.68**	**−0.06**	**0.70**	1.50
40×10	−1.16	−2.11	−2.43	−2.44	−2.15	−1.58	**−0.82**	**0.15**	1.25	2.28
50×10	−2.21	−3.18	−3.49	−3.32	−2.82	−2.04	**−0.95**	**0.31**	1.72	2.95
60×10	−3.16	−4.35	−4.57	−4.27	−3.49	−2.44	−1.09	**0.48**	2.18	3.53
70×10	−4.23	−5.34	−5.48	−5.03	−4.12	−2.85	−1.20	**0.62**	2.50	4.06
80×10	−5.14	−6.46	−6.48	−5.94	−4.69	−3.14	−1.33	**0.75**	2.87	4.58
90×10	−5.96	−7.36	−7.33	−6.51	−5.24	−3.54	−1.46	**0.86**	3.20	4.99
100×10	−6.93	−8.34	−8.16	−7.27	−5.87	−3.80	−1.56	**0.97**	3.53	5.48

be based on a jump in the logarithms which is extraordinarily large for this specific data matrix. However, in absence of a jump a justified extrapolation will not be possible and the fact that success or failure merely depends on the number of primary PCs is certainly a weakness of the log-eigenvalue diagram.

The next two columns in Table 6 give the reduced eigenvalues calculated according to Malinowski [9]

Table 6
Eigenvalues and reduced eigenvalues for literature data matrix

PC	*EV*	*REV*[a]	*REV*[b] (20×10)[c]	*REV*[b] (15×5)[c]
1	2.562×10^5	1.281×10^3	5.199×10^3	–
2	2.119×10^4	1.239×10^2	5.606×10^2	–
3	1.767×10^4	1.227×10^2	5.920×10^2	–
4	1.023×10^4	8.598×10	4.307×10^2	–
5	2.422×10^3	2.523×10	1.298×10^2	–
6	9.999	0.133	0.694	0.343
7	5.866	0.105	0.544	0.300
8	5.199	0.133	0.676	0.393
9	3.575	0.149	0.705	0.425
10	1.330	0.121	0.472	0.292

[a] Calculated according to Malinowski [9].
[b] Calculated according to Mandel [17].
[c] Size of random matrices.

and Mandel [17], respectively. The pattern in Malinowski's reduced eigenvalues is also characterized by a large jump. Moreover, there is no visible trend for the last five PCs. The same goes for Mandel's 'reduced eigenvalues' calculated from the eigenvalues of 20×10 random matrices. The five-dimensional model is easily discerned and it seems that parallel analysis works satisfactorily without modification. However, close examination shows that there is a problem. Pooling the 'reduced eigenvalues' should provide an efficient estimate of σ_M according to Eq. (8). The value found is $\hat{\sigma}_M = 0.786$. This value is not close to the value estimated from the residuals using Eq. (7), i.e. $\hat{\sigma}_M = 0.588$. Thus the behavior of Mandel's 'reduced eigenvalues' is partly misleading in this case. (Other examples show that constant 'reduced eigenvalues' cannot be expected in general if the size of the reference matrix is not correct.) It is, however, clear that an excellent initial guess is supplied by the results in the last two columns for the appropriate size of the reference matrix, i.e. $(20-5)\times(10-5)$. The 'reduced eigenvalues' calculated from the eigenvalues of 15×5 random matrices are also perfectly constant but now the pooled 'reduced eigenvalues' yield the estimate $\hat{\sigma}_M = 0.592$ which is very close to the correct value

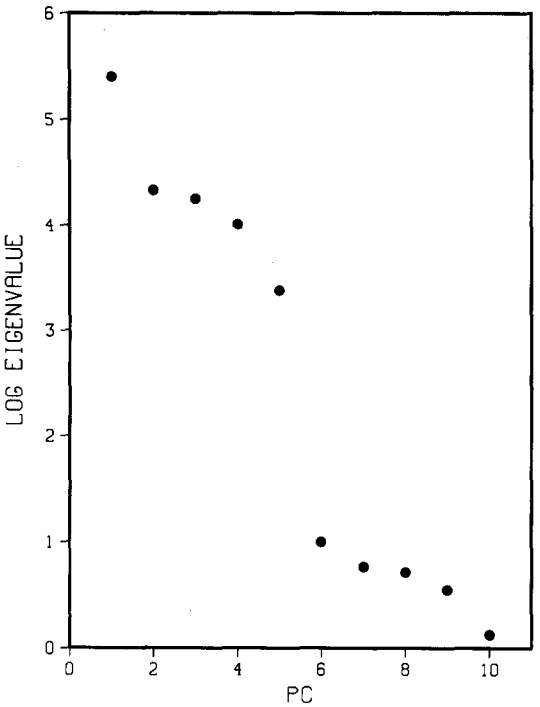

Fig. 5. Logarithm of eigenvalues for literature data matrix.

calculated from Eq. (7), i.e. $\hat{\sigma}_M = 0.588$. This lends credit to the followed approach. The only disadvantage connected to the method seems to be the trial-and-error character. Malinowski's reduced eigenvalues do not have this disadvantage but their applicability should be properly evaluated for the relevant matrix sizes.

Finally, it is emphasized that the example worked out here is selected in order to demonstrate the use of the modification rather than to put the method to the test. (In general the initial and final choice are not necessarily the same.) Work is currently in progress to evaluate the performance of the method on literature data that is generally accepted to be difficult.

Distribution of the residuals

In the preceding part it was assumed that the eigenvalues of random matrices with *normally* distributed elements could be used for the parallel analysis. It is therefore interesting to examine the distribution of the residuals of PCA. The distribution of the residuals of the five-dimensional PC model is shown in Fig. 6. It can be seen that the distribution is very close to the normal distribution with the same mean and standard deviation although the input round-off errors should be *uniformly* distributed. The tendency of the residuals to be more normally distributed then the original noise has been observed for other data as well. A 'theoretical' explanation could be as follows. Each principal axis is constructed with an error distributed according to some distribution. The size of the residuals, however, depends on the errors in all principal axes included in the model. Then according to the central limit theorem

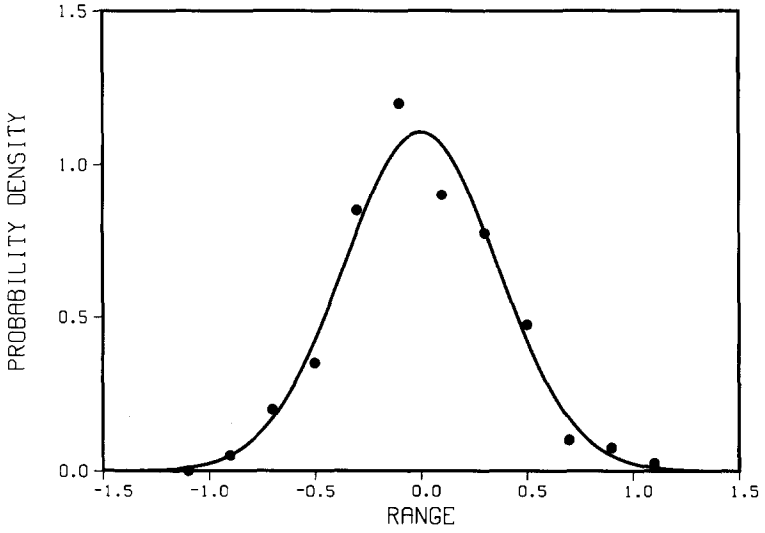

Fig. 6. Distribution of residuals after extracting five PCs for literature data matrix. The dots represent normalized frequency counts based on a binsize of 0.2. The normal distribution function with the same mean and standard deviation is drawn as guide to the eye.

the distribution of the residuals should converge to a normal distribution if enough PCs are extracted. (Note that this example has been selected because of the relatively large number of primary PCs.) This result should increase the value of Tables 3 and 4 and, perhaps more importantly, the applicability of certain significance tests that specifically demand normally distributed residuals, e.g. the χ^2 test and the number of 3σ misfits [40].

5. Conclusions

In the past, several pseudorank estimation methods have been proposed that are based on the similarity between the secondary eigenvalues of the test data matrix and the eigenvalues obtained from random matrices. These methods are commonly denoted as parallel analysis. In this study theoretical considerations have lead to a number of aspects that are expected to influence the applicability of such methods. The aspects thoroughly evaluated are the systematic contribution to the data, the divergence coefficient of the matrix and the distribution of the noise. The effect of possible approximations is always compared to the inherent variability of the eigenvalues, i.e. the standard error.

In this way the results of Monte Carlo simulations have shown that the size of the secondary eigenvalues depends only little on the value of the primary eigenvalues (systematic contribution) down to a very low signal-to-noise ratio ($=0.5$–1.0). Thus the secondary eigenvalues of a test data matrix can very well be approximated by the eigenvalues of an appropriately sized random matrix. As a useful byproduct of the current investigation it was found that theoretical predictions for the influence of the measurement errors on the primary eigenvalues start to work well for the same critical signal-to-noise ratio.

In the same way it is shown that the influence of the distribution of the noise becomes negligible if the data matrix is large enough. Differences between the eigenvalues that can be attributed to the distribution used (we have considered the normal, uniform and random sign distribution) are not significant with respect to the standard error in the eigenvalues for matrices as small as 20×10.

Thus if the data matrix is large enough, the distribution of the eigenvalues primarily depends on the divergence coefficient d. For square matrices ($d=1$) this distribution is characterized by a relatively large probability of finding a very small eigenvalue. For 'skinny' matrices ($d \gg 1$) the distribution approaches a spike, indicating that chance correlations vanish (the cross-product matrix becomes diagonal). Thus different values of d may lead to a completely different behavior for functions of the eigenvalues.

This has been illustrated for the logarithm of the eigenvalues and Malinowski's reduced eigenvalues. The logarithm of the eigenvalues is seen to yield a straight line for the low-numbered PCs. The logarithm of the eigenvalues for the high-numbered PCs may show a downward deviation. Since the use of the log-eigenvalue diagram is based on extrapolation towards the low-numbered PCs, it directly depends on the number of primary PCs which is an unfavorable situation. It is shown that, depending on the value of d, Malinowski's reduced eigenvalues are constant within the associated standard error. Thus they may be used for pseudorank estimation if the relevant matrix sizes have been properly investigated. The limited simulations described in this paper indicate that Malinowski's reduced eigenvalues will not work if $d \geq 3$.

A modification of parallel analysis is proposed that comes down to a trial-and-error procedure. The final estimate of the pseudorank is based on a consistent estimate of the variance of the measurement noise according to Eqs. (7) and (8). The procedure is inspired by the early result of Mandel [17] that ideally, one should simulate random matrices with the same number of degrees of freedom as the test data matrix in order to obtain the desired reference eigenvalues. It is emphasized that depending on the loss of degrees of freedom good results may still be expected if the reference eigenvalues are obtained from random matrices with the same size. In fact, simulating random matrices with the same size is useful in order to obtain an initial guess for the optimal size of the reference matrix.

Finally, numerical results show that the residual variance tends to become normally distributed if 'enough' PCs are extracted, independent of the distribution of the measurement error. As a consequence (parametric) methods that assume normally distributed residuals should have a wider range of applicability than previously assumed. This result is important for the applicability of the currently advertized method as well, since the choice of generating random matrices from

the normal distribution can now be motivated. The method may therefore especially hold a promise for the analysis of so-called high-rank data. It is emphasized that estimating the pseudorank for this kind of data constitutes a difficult problem in practice [41].

References

[1] Y. Wang, O.S. Borgen, B.R. Kowalski, M. Gu and F. Turecek, Advances in second-order calibration, *Journal of Chemometrics*, 7 (1993) 117–130.

[2] C.-N. Ho, G.D. Christian and E.R. Davidson, Application of the method of rank annihilation to fluorescent multicomponent mixtures of polynuclear aromatic hydrocarbons, *Analytical Chemistry*, 52 (1980) 1071–1079.

[3] N.M. Faber, L.M.C. Buydens, G. Kateman, Generalized rank annihilation method. II. Bias and variance in the estimated eigenvalues, *Journal of Chemometrics*, 8 (1994) 181–203.

[4] E.R. Malinowski, Theory of error in factor analysis, *Analytical Chemistry*, 49 (1977) 606–612.

[5] N.M. Faber, L.M.C. Buydens and G. Kateman, Aspects of pseudorank estimation methods based on an estimate of the size of the measurement error, *Analytica Chimica Acta*, in press.

[6] J.E. Jackson, *A User's Guide to Principal Components*, Wiley, New York, 1991.

[7] Y.-Z. Liang, O.M. Kvalheim and A. Höskuldsson, Determination of a multivariate detection limit and local chemical rank by designing a non-parametric test from the zero-component regions, *Journal of Chemometrics*, 7 (1993) 277–290.

[8] S.A. Farmer, An investigation into the results of principal component analysis of data derived from random numbers, *The Statistician*, 20 (1971) 63–72.

[9] E.R. Malinowski, Theory of the distribution of error eigenvalues resulting from principal component analysis with applications to spectroscopic data, *Journal of Chemometrics*, 1 (1987) 33–40.

[10] E.R. Malinowski, Statistical *F*-tests for abstract factor analysis and target testing, *Journal of Chemometrics*, 3 (1988) 49–60.

[11] Erratum in *Journal of Chemometrics*, 4 (1990) 102.

[12] P. Paatero and U. Tapper, Analysis of different modes of factor analysis as least squares fit problems, *Chemometrics and Intelligent Laboratory Systems*, 18 (1993) 183–194.

[13] N.M. Faber, L.M.C. Buydens and G. Kateman, Standard errors in the eigenvalues of a cross-product matrix: theory and applications, *Journal of Chemometrics*, 7 (1993) 495–526.

[14] J.M. Andrade, D. Prada, S. Muniategui, B. Gomez and M. Pan, Multivariate selection of variables in industrial quality control: optimizing aviation fuel final control, *Journal of Chemometrics*, 7 (1993) 427–438.

[15] D.L. Duewer, B.R. Kowalski and J.L. Fasching, Improving the reliability of factor analysis of chemical data by utilyzing the measured analytical uncertainty, *Analytical Chemistry*, 48 (1976) 2002–2010.

[16] R.J. Pell, M.B. Seasholtz and B.R. Kowalski, The relationship of closure, mean centering and matrix rank interpretation, *Journal of Chemometrics*, 6 (1992) 57–62.

[17] J. Mandel, A new analysis of variance model for non-additive data, *Technometrics*, 13 (1971) 1–18.

[18] T.M. Rossi and I.M. Warner, Rank estimation of excitation–emission matrices using frequency analysis of eigenvectors, *Analytical Chemistry*, 58 (1986) 810–815.

[19] X.M. Tu, D.S. Burdick, D.W. Millican and L.B. McGown, Canonical correlation technique for rank estimation of excitation–emission matrices, *Analytical Chemistry*, 61 (1989) 2219–2224.

[20] D.E. Johnson and F.A. Graybill, An analysis of a two-way model with interaction and no replication, *Journal of the American Statistical Association*, 67 (1972) 862–868.

[21] E.A. Sylvestre, W.H. Lawton and M.S. Maggio, Curve resolution using a postulated chemical reaction, *Technometrics*, 16 (1974) 353–368.

[22] L.A. Goodman and S. Haberman, The analysis of nonadditivity in two-way analysis of variance, *Journal of the American Statistical Association*, 85 (1990) 139–145.

[23] H.T. Eastment and W.J. Krzanowski, Cross-validatory choice of the number of components from a principal component analysis, *Technometrics*, 24 (1982) 73–77.

[24] S. Wold and M. Sjöström, SIMCA: a method for analyzing chemical data in terms of similarity and analogy, in B.R. Kowalski, (Editor), *Chemometrics, Theory and Application* (American Chemistry Society Symposium Series No. 52), American Chemical Society, Washington, DC, 1977, pp. 243–282.

[25] J. Mandel, The distribution of eigenvalues of covariance matrices of residuals in analysis of variance, *Journal of Research of the National Bureau of Standards*, 74B (1970) 149–154.

[26] O. Exner, Additive physical properties. I. General relationships and problems of statistical nature, *Collection of Czechoslovakian Chemical Communications*, 31 (1966) 3222–3251.

[27] F.J. Knorr, H.R. Thorsheim and J.M. Harris, Multichannel detection and numerical resolution of overlapping chromatographic peaks, *Analytical Chemistry*, 53 (1981) 821–825.

[28] U. Grenander and J.W. Silverstein, Spectral analysis of networks with random topologies, *Journal of Applied Mathematics*, 32 (1977) 499–519.

[29] S.S. Wilks, *Mathematical Statistics*, Wiley, New York, 1962.

[30] E.P. Wigner, On the distribution of the roots of certain symmetric matrices, *Annals of Mathematics*, 67 (1958) 325–327.

[31] N. Ohta, Estimating absorption bands of component dyes by means of principal component analysis, *Analytical Chemistry*, 45 (1973) 553–557.

[32] D.W. Kormos and J.S. Waugh, Abstract factor analysis of solid-state nuclear magnetic resonance spectra, *Analytical Chemistry*, 55 (1983) 633–638.

[33] E.R. Malinowski, Determination of the number of factors and the experimental error in a data matrix, *Analytical Chemistry*, 49 (1977) 612–617.

[34] R.F. Hirsch, G.L. Wu and P.C. Tway, Reliability of factor analysis in the presence of random noise or outlying data, *Chemometrics and Intelligent Laboratory Systems*, 1 (1987) 265–272.

[35] M.B. Seasholtz, R.J. Pell and K.E. Gates, Comments on the power method, *Journal of Chemometrics*, 4 (1990) 331–334.

[36] E.R. Malinowski, Obtaining the key set of typical vectors by factor analysis and subsequent isolation of component spectra, *Analytica Chimica Acta*, 134 (1982) 129–137.

[37] G.W. Stewart, Perturbation theory and least squares with errors in variables, *Contemporary Mathematics*, 112 (1990) 171–181.

[38] A. Lorber, Error propagation and figures of merit for quantification by solving matrix equations, *Analytical Chemistry*, 58 (1986) 1167–1172.

[39] G. Golub, Some modified matrix eigenvalue problems, *SIAM Review*, 15 (1973) 318–334.

[40] Z.Z. Hugus and A.A. El-Awady, The determination of the number of species present in a system: a new matrix rank treatment of spectrophotometric data, *Journal of Physical Chemistry*, 75 (1971) 2954–2957.

[41] A.K. Smilde, Y. Wang and B.R. Kowalski, Theory of medium-rank second-order calibration with restricted Tucker models, *Journal of Chemometrics*, 8 (1994) 21–36.