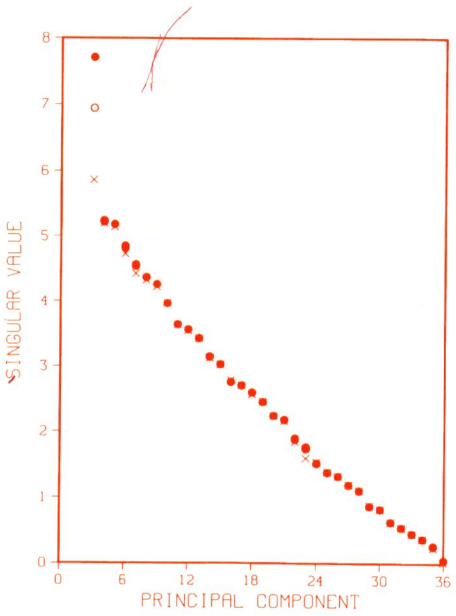
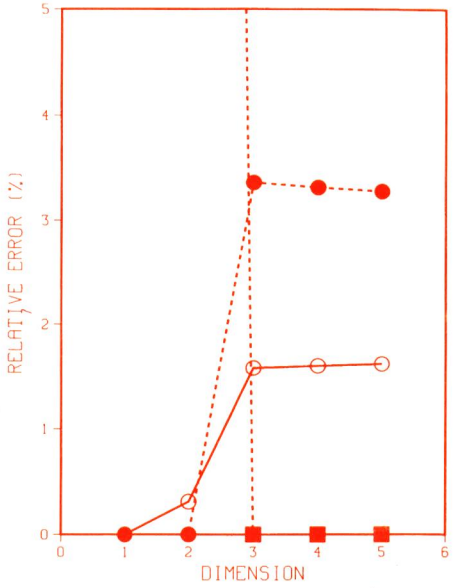
Contributions to multivariate data analysis in chemometrics





Klaas Faber

Contributions to multivariate data analysis in chemometrics

Manuscript commission: Dr. P. Geladi and Prof. Dr. A.K. Smilde

Contributions to multivariate data analysis in chemometrics

Een wetenschappelijke proeve op het gebied van de
NATUURWETENSCHAPPEN

Proefschrift

ter verkrijging van de graad van doctor
aan de Katholieke Universiteit Nijmegen,
volgens besluit van het College van Decanen
in het openbaar te verdedigen op
dinsdag 4 oktober 1994,
des namiddags te 1.30 uur precies

door

Nicolaas Maria Faber
geboren op 25 september 1957 te Aarle-Rixtel

Promotor: Prof. Drs. G. Kateman

Co-promotor: Dr. L.M.C. Buydens

Aan mijn ouders

Aan Marjan

CONTENTS

---

1 N.M. Faber, L.M.C. Buydens and G. Kateman, *J. Chemometrics*, **7**, 495 (1993).
2 N.M. Faber, L.M.C. Buydens and G. Kateman, *Chemometrics Intell. Lab. Syst.* in press.
3 N.M. Faber, L.M.C. Buydens and G. Kateman, *Anal. Chim. Acta*, in press.
4 N.M. Faber, L.M.C. Buydens and G. Kateman, *J. Chemometrics*, **8**, 147 (1994).
5 N.M. Faber, L.M.C. Buydens and G. Kateman, *J. Chemometrics*, **8**, 181 (1994).
6 N.M. Faber, L.M.C. Buydens and G. Kateman, *J. Chemometrics*, in press.

INTRODUCTION

This thesis is a reflection of the work which is carried out in the period of September 1990 to March 1994. The research has covered several topics within the field of multivariate data analysis (MDA) in chemometrics. Emphasis has been on the error analysis of multivariate methods (Part I and III) and pseudorank estimation (Part II). It has led to a number of papers and the present thesis consists mainly out of these articles.

Characterization of the current work

If only two characterizations of research were allowed, i.e. the research is either experimental or theoretical, the current work would have to be qualified as being theoretical in nature. However, pure mathematicians and statisticians would be equally irritated by the lack of rigor displayed in this thesis. Thus at all time it should be kept in mind that this research has been performed by a chemist and in many instances, the ground has been prepared by the work of other chemists. Mathematical or statistical rigor will not be discussed in this introduction and in most of the remaining sections. It will, however, return in the 'Conclusions and future research' section. The usefulness to practitioners in the field of chemometrics has always played a dominant role during the handling of mathematical and statistical subjects.* The aim of this introduction is to clarify the motivation of the current research within the established practical data analysis in chemometrics.

Background to the current work

Some background is needed in order to understand the motivation for the research currently presented. This background is indirectly reflected in the title of this thesis. It is seen that the title consists of two distinct parts, i.e. 'Contributions to multivariate data analysis' and 'in chemometrics'. The two separate parts can be conveniently described as follows. The first part covers the subject of the research: the analysis of multivariate data, i.e. the situation is considered in which several numbers are measured in only one experiment. The second part covers the environment where the research has taken place: chemometrics, i.e a (relatively young) discipline that embraces all topics associated with the 'non-instrumental' part of analytical chemistry. These topics are very diverse and range all the way from data analysis to laboratory organization. In the remaining part of this introduction I want to explain how this particular combination, i.e. performing research on MDA in the environment of chemometrics may lead to very specific difficulties. However, trying to solve some of them provided me with much of the motivation that was necessary for carrying out this work.

Two major difficulties resulting from the treated subject

MDA is a very broad subject which has received much attention in other disciplines whereas in chemometrics it is just one of many topics under investigation. It follows that usually the pioneering work is done outside the field of chemometrics. Thus one question immediately arises: in how far are the relevant results from other disciplines disseminated within chemometrics? When reading the different sections in chronological order one will find out that often important contributions have been missed. Some key references from applied statistics were only 'discovered' after submission of a substantial part of this thesis. (The present author is to blame for not knowing his literature and credit to this work is given whenever possible.) This problem is typical for any young science. Let us therefore focus on another important question: in how far are theoretical results obtained within other disciplines (or chemometrics as well) applicable to the particular kind of data we encounter in analytical chemistry? This question directly amounts to the practical usefulness of most of the present work.

Multivariate data in analytical chemistry

Many multivariate data nowadays generated in the laboratory result from so-called hyphenated methods. The most widely used of these methods typically consist of a combination of a chromatographic and a

---

*Kvalheim[1] has described this controversy in a striking way in connection with the formal criticism of early psychometric work: 'Perhaps more surprising for the psychometricians, who were struggling to establish a quantitative basis for their work, were the strong attacks from orthodox statisticians. The frustration felt by this unexpected and as he felt unfounded criticism made Thurstone pour out the following sarcastic statement: Let us also not forget a simple principle that every scientist takes for granted, namely, that he would rather measure something significant without any sampling distribution than to measure something trivial or irrelevant because its sampling distribution is known.'

spectroscopic technique: during chromatographic separation of a sample (by e.g. GC, LC, TLC or HPLC) a spectrum is recorded at regular time intervals (e.g. IR, MS, UV-vis, OES). The data thus obtained can be casted into a matrix where the rows (or columns) correspond to elution profiles and the columns (or rows) correspond to spectra. An example of such a 'spectro-chromatogram' is given in Figure 1 in Section 3. (The individual elution profiles and spectra are given in Figures 1 and 2 in Section 5.) Measuring a data matrix for one chemical sample has a tremendous advantage over just measuring a single number or a vector. In the field of calibration this advantage is now denoted as the second-order advantage following the terminology introduced by Sánchez and Kowalski.[2,3] The second-order advantage implies that a component can be quantified without calibrating for the interferents. Calibration methods that employ the second-order advantage have been developed. However, in order to correctly use these methods some considerations about the data should be met. These considerations are given by the mathematical model that is assumed to describe the data. For the generalized rank annihilation method (GRAM),[4] a method which receives much attention in this thesis, one of the requirements is that the signal be linear and additive. For spectroscopic data this amounts to the requirement that the law of Lambert-Beer holds.

Model errors in multivariate data

It is, however, well-known for most of the spectroscopic detection methods mentioned above that this law does not hold, since the necessary condition, the detected light should be monochromatic, is not fulfilled. (Analytic expressions for the resulting bias have been derived by Dose and Guiochon.[5]) This means that for a certain class of data model errors are inevitable in practice. (Model errors as a result of chemical interactions can often be eliminated by a suitable sample preparation.[6]) In essence, model errors are only problematic because it is difficult to develop a theory around a complicated model that adequately describes the resulting data. One may still obtain useful results from data that in fact violate the assumed simplified model. This should become clear from the following example.

Practical example of multivariate data analysis

If one is aware of the presence of model errors (and most data analysts in chemometrics are), how then should the MDA proceed in practice? During graduation much experience was obtained with practical data and the following problem had to be solved. A complex mixture containing four analytes and an unknown number of interferents was given a treatment with the purpose of extracting these analytes. A large data matrix (HPLC-UV-vis) was recorded before and after the treatment. The size of the data matrix (380x150) made data preprocessing necessary. First, the chromatogram was divided into time windows. Each window (A, B, C, D) contained one analyte of interest and several overlapped uncalibrated interferents. Next, the original spectrum (190-400nm) was reduced to bracket the region between 220 and 290nm. In this way the relevant fine structure was preserved while most of the background absorbance was removed. The calibration was performed using the 'old' method, univariate calibration at selective wavelengths, GRAM and iterative target-testing factor analysis (ITTFA).[7]

The first step of MDA is the determination of the dimension of the solution space, i.e. the so-called pseudorank. In practice, a number of procedures is available and an estimate is usually obtained by comparing the result for a number of suitable methods. For data of this kind, the second-derivative (2ND) method gives a solid lower bound.[8] Target testing is also a possible method and in this work the recommendation of Rasmussen *et al.*,[9] bilinear target testing (BTT), was employed. Furthermore, the 'extracted error' (XE) of Strasters[10] was applied. A summary of the results is given in Table 1. It is seen that the results for 2ND, BTT and XE are inconsistent. Therefore two additional methods, denoted by SIM and COC, were developed that exploit the advantage that similar mixtures (before and after treatment) are available. It should be kept in mind that is not important to know what information the different methods try to extract from the data and which (subjective) decision criteria have been used in order to arrive at these results. The example is merely worked out because tables like Table 1 which contain inconsistent results are frequently encountered in practice. It should illustrate the fact that determining the pseudorank may really constitute a bottle-neck in MDA. The results for XE, SIM and COC are seen to be 'fairly' consistent and have been combined to give the final estimates, N_{est}, in the last column.

Subsequent calibration, which is straightforward in this case once the pseudorank has been established, gives the quantities reported in Table 2. From these numbers it can be concluded that the three methods (GRAM, ITTFA and univariate calibration) agree very well. However, there is an advantage in using a multivariate method because it provides extra information in the form of the qualitative solution (pure spectra and elution profiles) for the analytes as well as the (eight!) interferents. For example, one could try to discover the identity of these interferents by means of a library search.

Table 1. Results of pseudorank estimation

Window	Treatment	2ND	BTT	XE	SIM	COC	N_{est}
A	-	2	1	2	2	2	2
	+	2	1	2	2	2	2
B	_	1	2	3	3	3	3
	+	2	2	3	3	3	3
C	_	1	2	4	4	4	4
	+	2	2	5	4	4	4
D	-	1	1	3	4	3	3
	+	1	1	4	4	3	3

Table 2. Results of calibration (mg)

Window	Treatment	GRAM	ITTFA	Univariate
A	-	56.1	62.1	60.6
	+	25.3	28.0	27.7
B	-	9.3	10.3	9.1
	+	3.3	3.8	3.3
C	-	27.5	30.0	28.8
	+	11.4	10.2	9.4
D	-	24.8	28.1	24.9
	+	10.3	9.3	8.1

Discussion of multivariate data analysis in practice

In the above example I have made a quantitative statement about the determined concentrations: the results for the different methods agree very well. I have not cared to mention that repeating the analysis some time later with somewhat different preprocessing of the data gave different results! Fortunately, it 'seems' that the differences are too small in order to affect any of the important conclusions. However, this example clearly shows that validation is mandatory in order to back up any quantitative statement about the obtained solution.

From basic statistics we know that a determined concentration should be accompanied by a confidence interval. Such a confidence interval is estimated from the data using the model that has also led to the concentration estimate. In the presence of model errors this procedure becomes impossible. The reason for the current mischief is that the concentration estimate is primarily determined by the adequacy of the model description of the data whereas the estimate of the confidence interval is primarily determined by the adequacy of the model description of the noise. In the presence of model errors the description of the data may be quite good while the description of the noise is completely insufficient. Thus useful concentration estimates may still be obtained while the theoretical estimate of the confidence interval is completely ridiculous.

A step in the direction of validation of results of multivariate methods has been set by the development of realistic simulations.[11] (Part of the graduation work was devoted to these simulations.) Very recently these simulations have been used to evaluate the effects of resolution, peak ratio and sampling frequency in diode-array fluorescence detection in liquid chromatography.[12] In the presence of model errors this kind of

simulations is a very sensible approach. The same can be concluded for the non-parametric multivariate detection limit proposed by Liang *et al.*[13] However, one must keep in mind that analyzing data with model errors leads to a very uncomfortable situation: one is allowed to use a model that is violated in order to predict concentrations but at the same time one is not allowed to use the same model in order to predict the reliability of these concentrations.

Motivation for the current work

From the point of view of the current thesis I will consider data analysis as described above as a data dependent necessity. It is to be expected that in future experimentation more attention will be paid to the elimination of model errors. As a result theoretically predicted confidence intervals will become increasingly important. Thus motivation for most of the current work is the possible usefulness for 'ideal' data in chemometrics (which may be just normal data in other disciplines). It is important to note that McCue and Malinowski[14] have reported the development of a liquid chromatograph that is especially designed for rank annihilation factor analysis (RAFA), the precursor of GRAM. I have analyzed their data and found no indication of model errors. Thus ideal data already exist in practice!

Motivation for the current work is also to be found in the better understanding of multivariate methods. For example, much effort has been devoted to the error analysis of GRAM: if second-order calibration (matrix data) is the natural extension of first-order calibration (vector data), it is just as natural to try to extend the corpus of error theory already available (but scattered in the literature) for first-order data to the analysis of second-order data. Much is to be learned about the relative importance of different sources of errors which is not directly apparent from the expression of the estimators themselves.

REFERENCES

1 O.M. Kvalheim, *Chemometrics Intell. Lab. Syst.* **14**, 1 (1992).
2 E. Sánchez and B.R. Kowalski, *J. Chemometrics*, **2**, 247 (1988).
3 E. Sánchez and B.R. Kowalski, *J. Chemometrics*, **2**, 265 (1988).
4 E. Sánchez and B.R. Kowalski, *Anal. Chem.* **58**, 496 (1986).
5 E.V. Dose and G. Guiochon, *Anal. Chem.* **61**, 2571 (1989).
6 J.D. Ingle and S.R. Crouch, *Spectrochemical Analysis*, Prentice-Hall, Englewood Cliffs, NJ (1988).
7 B.A. Roscoe and P.K. Hopke, *Comp. Chem.* **5**, 1 (1981).
8 B.G.M. Vandeginste, G. Kateman, J.K. Strasters, H.A.H. Billiet and L. de Galan, *Chromatographia*, **24**, 127 (1987).
9 G.T. Rasmussen, B.A. Hohne, R.C. Wieboldt and T.L. Isenhour, *Anal. Chim. Acta*, **112**, 151 (1979).
10 J.K. Strasters, PhD Dissertation, Delft (1989).
11 M.J.P. Gerritsen, N.M. Faber, M. van Rijn, B.G.M. Vandeginste and G. Kateman, *Chemometrics Intell. Lab. Syst.* **12**, 257 (1992).
12 R.B. Poe and S. Rutan, *Anal. Chim. Acta*, **283**, 845 (1993).
13 Y.-Z. Liang, O.M. Kvalheim and A. Höskuldsson, *J. Chemometrics*, **7**, 277 (1993).
14 M. McCue and E.R. Malinowski, *J. Chromatogr. Sci.* **21**, 229 (1983).

PART I PRINCIPAL COMPONENT ANALYSIS

Part I consists of a very large paper. Rather than deriving new results, I have aimed at discussing some seemingly unconnected topics from a central point of view while making use of statistical literature that is relatively unknown in the field of chemometrics. The central point of view is provided by the standard errors in the eigenvalues of a cross-product matrix. These standard errors are very important because cross-product matrices frequently arise in multivariate data analysis, especially in principal component analysis (PCA). PCA is a technique that is often used to discover the rank of the signal contribution to a data matrix. Very often this so-called pseudorank, i.e. the rank of the data matrix in the absence of noise, is much smaller than the number of rows or columns. Thus in subsequent data analysis the data matrix is replaced by its best least squares fit in this more appropriate dimension. This usually leads to a substantial reduction of data and removal of most of the noise. The derived standard errors account for the variability in the data as a result of measurement errors. The presented derivation (p.12-13) closely follows the work of Hugus and El-Awady (Reference 1 in the paper). However, the final expressions are entirely different. It is important to note that Goodman and Haberman[1] have recently given the same expressions as derived here. The title of their paper 'The analysis of nonadditivity in two-way analysis of variance', however, does not readily show a connection to PCA. This is one of the excuses for not finding this key reference in time. Goodman and Haberman also discuss bias in the eigenvalues of PCA. This paper is therefore highly recommended for anyone who is interested in the error analysis of PCA. For example, the current paper does *not* discuss bias. It should, however, be noted that Goodman and Haberman give expressions for bias in the eigenvalues for a one-dimensional principal component model. It will be shown in a future publication[2] how their results can be generalized to an n-dimensional model by making use of Malinowski's error functions.

REFERENCES

1 L.A. Goodman and S.J. Haberman, *JASA*, **85**, 139 (1990).
2 N.M. Faber, M.J. Meinders, P. Geladi, M. Sjöström, L.M.C. Buydens and G. Kateman, *Anal. Chim. Acta*, submitted.

STANDARD ERRORS IN THE EIGENVALUES OF A CROSS-PRODUCT MATRIX: THEORY AND APPLICATIONS

N. M. FABER, L. M. C. BUYDENS AND G. KATEMAN

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

SUMMARY

New expressions are derived for the standard errors in the eigenvalues of a cross-product matrix by the method of error propagation. Cross-product matrices frequently arise in multivariate data analysis, especially in principal component analysis (PCA). The derived standard errors account for the variability in the data as a result of measurement noise and are therefore essentially different from the standard errors developed in multivariate statistics. Those standard errors were derived in order to account for the finite number of observations on a fixed number of variables, the so-called sampling error. They can be used for making inferences about the population eigenvalues. Making inferences about the population eigenvalues is often not the purposes of PCA in physical sciences. This is particularly true if the measurements are performed on an analytical instrument that produces two-dimensional arrays for one chemical sample: the rows and columns of such a data matrix cannot be identified with observations on variables at all. However, PCA can still be used as a general data reduction technique, but now the effect of measurement noise on the standard errors in the eigenvalues has to be considered. The consequences for significance testing of the eigenvalues as well as the usefulness for error estimates for scores and loadings of PCA, multiple linear regression (MLR) and the generalized rank annihilation method (GRAM) are discussed. The adequacy of the derived expressions is tested by Monte Carlo simulations.

KEY WORDS Standard errors Eigenvalues PCA MLR GRAM Rank estimation

INTRODUCTION

The singular value decomposition (SVD) of a matrix $\mathbf{M}$ $(r \times c)$ is given by

$$\mathbf{M} = \mathbf{U}\Theta\mathbf{V}^{\mathrm{T}} \tag{1}$$

where $\mathbf{U}(r \times c)$ and $\mathbf{V}(c \times c)$ are matrices that satisfy $\mathbf{U}^{\mathrm{T}}\mathbf{U} = \mathbf{V}^{\mathrm{T}}\mathbf{V} = \mathbf{V}\mathbf{V}^{\mathrm{T}} = \mathbf{I}_c$ and $\Theta(c \times c)$ is a diagonal matrix that contains the singular values. $\mathbf{I}_c$ denotes the $c \times c$ identity matrix. It is assumed throughout this paper that $r \geqslant c$. The SVD of $\mathbf{M}$ is equivalent to the PCA or eigenvalue decomposition (EVD) of the cross-product matrices* of $\mathbf{M}$, $\mathbf{M}^{\mathrm{T}}\mathbf{M}$ and $\mathbf{M}\mathbf{M}^{\mathrm{T}}$, since the right singular vectors contained by $\mathbf{V}$ and the left singular vectors contained by $\mathbf{U}$ are

* The cross-product matrix $\mathbf{M}^{\mathrm{T}}\mathbf{M}$ itself is referred to as the covariance matrix (corrected for the number of degrees of freedom and the mean) in the statistical literature on PCA. To avoid confusion, we will only use the term covariance matrix to denote the matrix of covariances of the eigenvalues.

eigenvectors found by the so-called R-mode and Q-mode analysis respectively:

$$\mathbf{D}_R = \mathbf{M}^T\mathbf{M} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^T \tag{2a}$$

$$\mathbf{D}_Q = \mathbf{M}\mathbf{M}^T = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T \tag{2b}$$

(Actually, the opposite is not always true: in SVD, once the signs of **U** are fixed, then those of **V** are automatically defined. In EVD the signs of **U** and **V** are independent.) The singular values are the non-negative square roots of the eigenvalues contained by the diagonal matrix **Λ**. There is a numerical difference between the two procedures, since it takes double the precision to represent the eigenvalues. This is without consequence for data analysis in practice as long as the measurement noise is much larger than the machine precision.

Measurement noise in the data points results in approximate eigenvalues for the cross-product matrices. Hugus and El-Awady[1] have derived an expression for the standard errors in the eigenvalues. Assuming that error eigenvalues should be zero within the allowed statistical fluctuations, the significance of an eigenvalue is established by direct comparison with its associated error. Successful application of the criterion has been reported for infrared spectra[2] and more recently for Auger electron spectrometry depth profiles.[3] However, since a cross-product matrix is positive definite in the presence of noise, the eigenvalues are all non-zero. Therefore one should not test the eigenvalues to be equal to zero but equal to the expectation value of the eigenvalues of a random matrix with the same number of degrees of freedom.[4-7] (These 'reference' values can easily be obtained by simulating random matrices.) It follows that the derived equation must be re-examined.

Expressions for standard errors in the eigenvalues of PCA have also been known for a long time in the field of multivariate statistics.[8] There, however, a data matrix is seen as a limited sample* of observations on a population of random vectors. Assuming that the components of these vectors are Gaussian-distributed, expressions for the standard errors have been derived that primarily depend on the sample size.[9,10] These standard errors indicate how well the sample estimate is suited for making inferences about the population eigenvalues. Since the variability within the population is usually much larger than the measurement error, the contribution of the measurement error to the total variability is neglected, i.e. only sampling errors are considered. It follows that the use of these expressions is limited to data matrices where the elements correspond, in a general sense, to observations on variables. They can certainly not be used if the data matrix is measured on an analytical instrument that produces two-dimensional arrays for *one* chemical sample: the notion of observations on variables is not useful here. However, PCA can still be used as a general data reduction technique, but now the measurement error has to be considered in a derivation of standard errors in the eigenvalues. This brings us back to the work of Hugus and El-Awady. Because standard errors in the eigenvalues have a number of possible applications, one of the objectives of this paper will be to contrast the standard errors resulting from sampling errors with the standard errors resulting from measurement noise.

The remainder of this paper is organized as follows. First the different levels of variability in PCA will be outlined. Next, after introducing the method of error propagation, new expressions for the standard errors in the eigenvalues of a cross-product matrix will be derived. They will be compared with the result of Hugus and El-Awady and it will be shown that the expressions are essentially different. Two examples from the literature will be worked out to show the difference between standard errors resulting from measurement noise and standard

* It should be clear from the context whether a statistical sample or a chemical sample is meant.

errors resulting from the variability in the population. Next we will treat three important applications for the new standard errors. These applications comprise error estimates for the scores and loadings of PCA[11–13] and prediction error estimates for MLR and GRAM. In the classification of Sánchez and Kowalski, MLR and GRAM are examples of first-order[14] and second-order[15] tensorial calibration respectively. Here the terms first-order and second-order refer to the format of the data. For first-order tensorial calibration the data for a (chemical) sample consist of a vector; for second-order tensorial calibration the data for a sample consist of a matrix. (First-order and second-order are also often used to refer to the relation between concentration and response: linear and quadratic respectively.) Finally we will evaluate the adequacy of the derived expressions by Monte Carlo (MC) simulations of ideal bilinear data. For bilinear data the response is separable into two independent response functions. In this paper we will take HPLC–UV data as an example. In the ideal case the HPLC elution profile is independent of wavelength and the UV absorbance independent of time.

THEORY

Notation

The following notation is used with respect to errors and estimators: the errorless quantity (true value), the total error in the measured quantity and the first-order estimate of the total error are denoted by adding a 'tilde', a 'δ' and a 'd' respectively to the symbol for the measured quantity. For the element of data matrix $\mathbf{M}$ in row i and column k this gives $M_{ik} = \tilde{M}_{ik} + \delta M_{ik} \approx \tilde{M}_{ik} + dM_{ik}$. Taking expectation will lead to statistical errors. In order to deal with the covariance between the measurement error in different matrix elements, it is convenient to regard the statistical errors as vectors.[1] The covariance of the error in element M_{ik} and the error in element M_{jl} is given by the inner product $\mathrm{COV}(M_{ik}, M_{jl}) = E[dM_{ik}dM_{jl}] = |\vec{\sigma}_{M_{ik}} \cdot \vec{\sigma}_{M_{jl}}|$. Estimators will be indicated by a 'hat' unless estimation is performed by replacing the true values by the measured values. The same notation applies to the errors in the estimated parameters. The result of the derivation will be the covariance matrix for the parameters from which the standard errors will follow as the square roots of the diagonal elements.

Levels of variability in PCA

In textbooks on multivariate analysis the data collection process is usually presented as follows: a $p \times 1$ random vector $\mathbf{x}$ with population mean $\boldsymbol{\mu} = E[\mathbf{x}]$ and population covariance $\boldsymbol{\Sigma} = E[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}}]$ is observed by randomly drawing a sample $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$. The sample mean $\mathbf{m} = \Sigma_i\ \mathbf{x}_i$ and sample covariance $\mathbf{S} = [1/(N-1)]\ \Sigma_i\ (\mathbf{x}_i - \mathbf{m})(\mathbf{x}_i - \mathbf{m})^{\mathrm{T}}$ are efficient estimators of the population parameters. The eigenvectors of $\boldsymbol{\Sigma}$ and $\mathbf{S}$ are the principal components, while the corresponding eigenvalues are the variances for the population and the sample respectively. If the sample is large enough, an almost certain correspondence can be set up between the two sets of eigenvalues and eigenvectors.[8]

In order to assess the influence of the data collection process on the resulting sample estimates, theoretical expressions have been derived for the standard errors.[9,10] In this way it can be established how well the sample estimates resemble the unknown population parameters. A disadvantage of these expressions is that the components of $\mathbf{x}$ must be Gaussian-distributed. Furthermore, the results are asymptotic, i.e. they contain the values of the

population parameters. Nevertheless, useful error estimates are obtained if these values are replaced by their sample estimates.[16]

Deriving theoretical expressions will not be possible in general, i.e. for small samples, for different distributions and, more importantly, for statistics that depend in a more complicated way on the elements of the data matrix. This means that the uncertainties have to be estimated in an entirely different way, e.g. by performing simulations on the data. With respect to simulations Krzanowski has pointed out that two levels of variability are relevant in PCA.[17] The situation is outlined schematically in Figure 1. Starting from a population with known mean μ and covariance Σ, independent samples can be generated. Given a sample with known mean and covariance, say $\mathbf{m}_k$ and $\mathbf{S}_k$, different realizations for the data matrices are possible. This process is called *conditional* sampling, since the sample values are fixed. Krzanowski argues that traditional eigenvalue-based methods for rank determination will give identical results for data matrices that correspond to the same sample. For the assessment of the robustness of these methods, different samples must be generated. This process is called *unconditional* sampling, since the sample values are no longer fixed. For the assessment of the robustness of methods that depend on the values of all elements of the data matrix, the second level becomes relevant, since the outcome of these methods may be different for matrices corresponding to the same sample. Important examples are cross-validation[18,19] and matrix rank analysis.[20]

This scheme only accounts for sampling errors and becomes more complicated if measurement errors are also considered. As a result of measurement errors, a data matrix, say $\mathbf{M}_{k1}$, can be partitioned into an errorless part and the measurement error as $\mathbf{M}_{k1} = \tilde{\mathbf{M}}_{k1} + \delta\mathbf{M}_{k1}$. Now the eigenvalues are no longer fixed and the results of eigenvalue-based methods may no longer be identical for all data matrices derived from the same sample. Two simulation methods apply to the scheme of Figure 1 if measurement noise cannot be neglected. The effect of sampling error is estimated by resampling methods such as jack-knife and bootstrap.[21] These methods are distribution-free, because new samples are generated using the original data. The basic assumption is that the true (unknown) distribution is supported on the measured data points. The effect of measurement noise is estimated by the MC method. New data matrices are generated by perturbing the errorless data matrix with noise taken from some

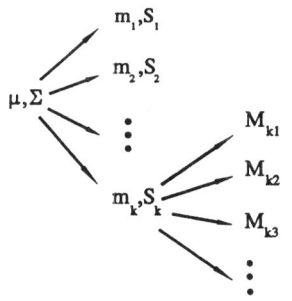


Figure 1. Schematic representation of different levels of variability for PCA. The standard errors in the population parameters μ and Σ are assessed by generating samples from the population (i.e. unconditional sampling). Conditional to the sample estimates $\mathbf{m}_k$ and $\mathbf{S}_k$ for the population values, data matrices $\mathbf{M}_{k1}$, $\mathbf{M}_{k2}$, ... can be generated in order to assess the variability within a sample (i.e. conditional sampling)

distribution. This method is therefore not distribution-free. We will use both simulation methods in order to compare the effect of the different sources of error. New realizations by the MC method will be called *trials* and new realizations by the bootstrap or jack-knife method will be called *replications*.[21]

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.[22] 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 true values and truncation usually proceeds after the linear or quadratic term. The method works well if the measured data points are unbiased estimates of the true data points and the errors are small. The method has been successfully applied in the context of matrix rank analysis[20] and multivariate calibration.[23,34] In order to derive tractable expressions, some assumptions must be made. The first assumption usually made is that the errors in the matrix elements are uncorrelated, i.e. with respect to the errors in the other matrix elements as well as the matrix element itself. This will be a valid assumption in many practical applications. The second assumption concerns homoscedasticity. This assumption may be realistic for HPLC–UV data (see the residual plot in Reference 25) but is not so for inductively coupled plasma–optical emission spectroscopy (ICP–OES) data.[24] The advantage of homoscedastic noise lies primarily in the easy interpretation of the resulting expressions.

If the distribution function of the errors is known, it is sometimes possible to derive the exact distributions of the parameters. Moran and Kowalski[23] have succeeded in deriving the distribution for the estimated calibration matrix $\hat{\mathbf{K}}$ but not for the matrix of estimated initial concentrations, $\hat{\mathbf{N}}_0 = (\hat{\mathbf{K}}^T)^{-1}\mathbf{Q}$, for the generalized standard addition method (GSAM), because the elements of $\hat{\mathbf{N}}_0$ are a complicated function of the elements of $\hat{\mathbf{K}}^T$. The distribution of the estimated parameters is needed if *exact* confidence limits are to be constructed. Even if the distribution function of the errors is known, the derived standard errors in the parameters do not automatically lead to confidence limits. Figure 2 shows a non-linear transformation of a symmetrically distributed random variable x. As a result of the non-linear transformation, y is not symmetrically distributed. Distribution and possible bias of the parameters are conveniently investigated by MC methods, giving *semiquantitative* results.

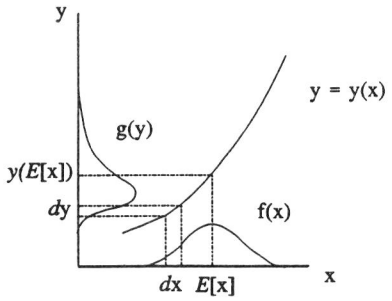


Figure 2. Propagation of errors under a non-linear transformation. The mode in the distribution function $g(y)$ is shifted downwards because in regions where $dy < dx$ the probability piles up faster for y than for x

In the following derivations no assumptions are made with respect to the distribution function of the errors, for two reasons. In the first place it is usually not realistic to assume e.g. Gaussian-distributed errors in practice. In the second place, for one of the applications, i.e. the standard errors in the eigenvalues for GRAM, the relation between the original data points and the eigenvalues is so complicated that it will be very difficult to obtain the distribution function for the eigenvalues in closed form.

Standard errors in the eigenvalues of PCA resulting from measurement noise

We start with the EVD of the $c \times c$ cross-product matrix $\mathbf{D}_R = \mathbf{M}^T\mathbf{M}$:

$$\mathbf{D}_R = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^T \tag{3}$$

$$\mathbf{\Lambda} = \mathbf{V}^T\mathbf{D}_R\mathbf{V} \tag{4}$$

Expressing the experimental quantities in terms of the errorless quantities and their respective total error gives

$$\tilde{\mathbf{\Lambda}} + \delta\mathbf{\Lambda} = (\tilde{\mathbf{V}} + \delta\mathbf{V})^T(\tilde{\mathbf{D}}_R + \delta\mathbf{D}_R)(\tilde{\mathbf{V}} + \delta\mathbf{V}) \tag{5}$$

which can be multiplied out to yield

$$\delta\mathbf{\Lambda} = \delta(\mathbf{V}^T)\tilde{\mathbf{D}}_R\tilde{\mathbf{V}} + \tilde{\mathbf{V}}^T\delta\mathbf{D}_R\tilde{\mathbf{V}} + \tilde{\mathbf{V}}^T\tilde{\mathbf{D}}_R\delta\mathbf{V} \tag{6}$$

Only the linear contribution to the error in the eigenvalues is considered at this point (first-order perturbation):

$$d\mathbf{\Lambda} = d(\mathbf{V}^T)\tilde{\mathbf{D}}_R\tilde{\mathbf{V}} + \tilde{\mathbf{V}}^T d\mathbf{D}_R\tilde{\mathbf{V}} + \tilde{\mathbf{V}}^T\tilde{\mathbf{D}}_R d\mathbf{V} \tag{7}$$

It is shown in Appendix I that the terms originating from errors in the eigenvectors cancel:

$$d\mathbf{\Lambda} = \tilde{\mathbf{V}}^T d\mathbf{D}_R\tilde{\mathbf{V}} \tag{8}$$

Introducing the definition of $\mathbf{D}_R$ gives

$$d\mathbf{\Lambda} = \tilde{\mathbf{V}}^T(d(\mathbf{M}^T)\tilde{\mathbf{M}} + \tilde{\mathbf{M}} + \tilde{\mathbf{M}}^T d\mathbf{M})\tilde{\mathbf{V}} \tag{9}$$

Hugus and El-Awady,[1] using indexed operations from the start, find that for the case of uncorrelated noise the standard errors can be estimated by (R-mode analysis)

$$\hat{\sigma}_{\lambda_n} = \left(\sum_{k=1}^{c} V_{kn}^2 \sum_{l=1}^{c} V_{ln}^2 \sum_{i=1}^{r} (M_{ik}^2\hat{\sigma}_{M_{il}}^2 + M_{il}^2\hat{\sigma}_{M_{ik}}^2)(1 + \delta_{kl}) \right)^{1/2} \tag{10}$$

where $\lambda_n = \Lambda_{nn}$ and δ_{kl} is the well-known Kronecker delta. The indices specify the locations (row and column) of the elements in their respective matrices. It can be seen that all errorless quantities in (9) have been replaced by the experimentally obtained values. Only the errors in the data points have to be estimated additionally.

However, considerable simplification results from recognizing that the two product matrices in parentheses in (9) are symmetrical and therefore identical, so that (9) reduces to

$$d\mathbf{\Lambda} = 2\tilde{\mathbf{V}}^T\tilde{\mathbf{M}}^T d\mathbf{M}\tilde{\mathbf{V}} \tag{11}$$

A more comprehensive form is obtained by using the singular equation $\tilde{\mathbf{M}}\tilde{\mathbf{V}} = \tilde{\mathbf{U}}\tilde{\mathbf{\Theta}}$:

$$d\mathbf{\Lambda} = 2\tilde{\mathbf{\Theta}}\tilde{\mathbf{U}}^T d\mathbf{M}\tilde{\mathbf{V}} \tag{12}$$

This result can also be derived in a more direct way by starting with the SVD of $\mathbf{M}$ (see

Appendix II). Equation (12) results in terms of standard deviations as vector quantities in

$$\vec{\sigma}_{\lambda_n} = 2\tilde{\theta}_n \sum_{i=1}^{r} \tilde{U}_{in} \sum_{k=1}^{c} \tilde{V}_{kn} \vec{\sigma}_{M_{ik}} \tag{13}$$

Here $\tilde{\theta}_n = \tilde{\Theta}_{nn}$. We change the subscripts n, i, k to m, j, l and take the inner product of the two expressions in order to obtain the covariance matrix of the eigenvalues:

$$\mathrm{COV}(\lambda_n, \lambda_m) = | \vec{\sigma}_{\lambda_n} \cdot \vec{\sigma}_{\lambda_m} | = 4\tilde{\theta}_n \tilde{\theta}_m \left| \left(\sum_{i=1}^{r} \tilde{U}_{in} \sum_{k=1}^{c} \tilde{V}_{kn} \vec{\sigma}_{m_{ik}} \right) \cdot \left(\sum_{j=1}^{r} \tilde{U}_{jm} \sum_{l=1}^{c} \tilde{V}_{lm} \vec{\sigma}_{M_{jl}} \right) \right| \tag{14}$$

Separation of the variance ($j = i$, $l = k$) and covariance ($j \neq i$, $l \neq k$) contributions gives

$$\mathrm{COV}(\lambda_n, \lambda_m)$$

$$= 4\tilde{\theta}_n \tilde{\theta}_m \left(\sum_{i=1}^{r} \tilde{U}_{in} \tilde{U}_{im} \sum_{k=1}^{c} \tilde{V}_{kn} \tilde{V}_{km} \sigma^2_{M_{ik}} + \sum_{i=1}^{r} \sum_{k=1}^{c} \sum_{j \neq i}^{r} \sum_{l \neq k}^{c} \tilde{U}_{in} \tilde{U}_{jm} \tilde{V}_{kn} \tilde{V}_{lm} \rho_{ik,jl} \sigma_{M_{ik}} \sigma_{m_{jl}} \right) \tag{15}$$

Here $\rho_{ik,jl}$ signifies the linear correlation coefficient between data points M_{ik} and M_{jl}.

In the case of uncorrelated noise (15) reduces to

$$\mathrm{COV}(\lambda_n, \lambda_m) = 4\tilde{\theta}_n \tilde{\theta}_m \sum_{i=1}^{r} \tilde{U}_{in} \tilde{U}_{im} \sum_{k=1}^{c} \tilde{V}_{kn} \tilde{V}_{km} \sigma^2_{M_{ik}} \tag{16}$$

and the standard errors of the eigenvalues are given as the square roots of the diagonal elements of the covariance matrix:

$$\sigma_{\lambda_n} = 2\tilde{\theta}_n \left(\sum_{i=1}^{r} \tilde{U}^2_{in} \sum_{k=1}^{c} \tilde{V}^2_{kn} \sigma^2_{M_{ik}} \right)^{1/2} = 2\tilde{\lambda}_n^{1/2} \left(\sum_{i=1}^{r} \tilde{U}^2_{in} \sum_{k=1}^{c} \tilde{V}^2_{kn} \sigma^2_{M_{ik}} \right)^{1/2} \tag{17}$$

If the error has a constant standard deviation σ_M, we may take it outside the summations. Since the singular vectors are orthonormal, the indices i and k are summed out and (16) becomes

$$\mathrm{COV}(\lambda_n, \lambda_m) = 4\tilde{\theta}_n \tilde{\theta}_m \sigma^2_M \, \delta_{nm} \tag{18}$$

The standard errors of the eigenvalues are given as

$$\sigma_{\lambda_n} = 2\tilde{\theta}_n \sigma_M = 2\tilde{\lambda}_n^{1/2} \sigma_M \tag{19}$$

The standard error of an eigenvalue is thus a weak function (square root) of the modulus of that particular eigenvalue.

Equations (17) and (19) are evaluated in practice as

$$\hat{\sigma}_{\lambda_n} = 2\lambda_n^{1/2} \left(\sum_{i=1}^{r} U^2_{in} \sum_{k=1}^{c} V^2_{kn} \hat{\sigma}^2_{M_{ik}} \right)^{1/2} \tag{20}$$

$$\hat{\sigma}_{\lambda_n} = 2\lambda_n^{1/2} \hat{\sigma}_M \tag{21}$$

The essential difference between equations (10) and (20) is the presence of the eigenvalue in the latter. Finally, it follows from (19) that the standard error in the singular values is (to first order) a constant and equal to the standard deviation of the error in the data:

$$\sigma_\theta \equiv \sigma_{\theta_n} = \sigma_M \tag{22}$$

Accordingly, the standard error in the singular values is estimated by

$$\hat{\sigma}_\theta = \hat{\sigma}_M \tag{23}$$

Estimation of measurement noise

The accuracy of the estimated standard errors will depend primarily on the accuracy of the estimate of the measurement error.[24] Different methods for estimating the magnitude of the noise can be found in the chemometrical literature. The methods that do not depend on the pseudorank of the data matrix, i.e. the mathematical rank in the absence of noise, are to be preferred, since determining the pseudorank is sometimes a difficult problem itself. Brown and Brown[26] have successfully estimated the magnitude of the noise in voltammograms from the high-frequency components of the Fourier power spectrum. Hirsch *et al.*[27] propose to add random noise to the experimental data matrix and extrapolate the resulting reproduction error after successively including principal components (PCs).

If a reliable choice for the pseudorank is available, the variance of the measurement noise can be estimated using

$$\hat{\sigma}_{\mathrm{M}} = \left(\frac{\sum_{p=K+1}^{c} \lambda_p}{(r-K)(c-K)} \right)^{1/2} \tag{24}$$

Here K denotes the pseudorank. Equation (24) is equivalent to the real error (RE) function of Malinowski.[28] The number of degrees of freedom in the denominator is, however, found by different authors[5,29] to give a better estimate than the number given by Malinowski, i.e. $r(c-K)$. It should be noted that the number of degrees of freedom is not derived in Malinowski's theory of error. A pragmatic solution to the dimensionality problem is to analyse the solution for different dimensions and choose the dimension that gives the best result according to some criterion.[30]

Standard errors in the eigenvalues of PCA resulting from sampling errors

The derived standard errors contrast strongly with the standard errors well-known in the statistical literature.[8] These standard errors are derived under the assumption that the variables are Gaussian-distributed in the population:[9]

$$\hat{\sigma}_{\lambda_n} = \sqrt{\left(\frac{2}{N-1}\right)} \lambda_n \tag{25}$$

Here the population eigenvalue $\tilde{\lambda}_n$ is replaced by the sample value and the result depends only on the sample size N. Lawley has extended this result with the second-order contribution.[10] Because of the restrictive assumption of Gaussian-distributed variables and large samples, these expressions have had little practical evaluation.[8]

Two numerical examples from the literature

The difference between the two standard errors is best illustrated with numerical examples. At all times it must be kept in mind that the standard errors apply to completely different error sources. The examples treated here are data matrices that can be seen as resulting from a genuine sample of observations on variables. Here the standard errors according to (25) should be used. However, 'interpreting' the residuals as measurement error makes it possible to compare the effects of both errors.

The first example is a marketing problems taken from Reference 8. Four attributes are

observed for 101 smokers and the resulting sample covariance matrix **S** is given by

$$\mathbf{S} = \begin{bmatrix} 2{\cdot}53 & 3{\cdot}50 & 2{\cdot}06 & 1{\cdot}45 \\ 3{\cdot}50 & 5{\cdot}05 & 2{\cdot}86 & 2{\cdot}02 \\ 2{\cdot}06 & 2{\cdot}86 & 1{\cdot}68 & 1{\cdot}19 \\ 1{\cdot}45 & 2{\cdot}02 & 1{\cdot}19 & 0{\cdot}86 \end{bmatrix}$$

The eigenvalues and standard errors according to equations (21) and (25) are given in Table 1. The estimate of the measurement error inserted in (21) is obtained by evaluating (24) for a one-dimensional model. Thus if the residuals are interpreted as measurement error, the standard error in the first eigenvalue equals about the sum of the secondary eigenvalues. However, if the residuals are interpreted as sampling error, the standard error for the first eigenvalue becomes quite large, although the ratio of observations and variables seems very favourable. The difference of a factor of ten between the two standard errors shows that the sampling error is much larger for this problem than the measurement error. Furthermore, the standard errors are about the same for principal components 2, 3 and 4. This confirms that the correct dimension of the problem is one.

The second example is the McReynolds' retention index matrix[31] (ten solutes on 226 liquid phases). This data matrix has been studied by a large number of investigators[18,19,32] and some of the results have been reviewed by Joliffe.[33] Of the 226 liquid phases (LPs), 13 are identified as outliers by Wold and Andersson[32] while one LP has a missing value. Wold and Andersson have deliberately restricted the number of PCs to three in order to obtain results that lead to a practical classification of the 213 'normal' LPs. Such a classification could help to reduce the number of LPs used, which is the original intention of the publication of McReynolds. In a later study, using cross-validation, Wold[18] comes to the conclusion that five PCs are significant for the 213×10 matrix. Eastment and Krzanowski[19] find four significant PCs using their modified cross-validation technique for the 212×10 matrix. Furthermore, Joliffe[33] gives the log eigenvalue (LEV) diagram and comments: 'There is, however, an indication of a straight line, starting at $m = 4$, in the LEV plot'. This reasoning would lead to the conclusion of three significant PCs.

In Table 2 we summarize the results of PCA for the retention index data matrix. Apart from the eigenvalues for the data measured about the mean, the results are divided into different estimates of the standard errors in the eigenvalues and a number of frequently used significance tests. We will start with a discussion of the results of the significance tests, because some of these results are a direct extension of the research mentioned above. (Note that the use of a particular significance test does not depend on the assumed source of error.) The cumulative percentage (CUM) criterion is only useful if a measure of the precision of the data is available. In the next column we give the value of $\hat{\sigma}_M$ in equation (24). This also provides a parametric

Table 1. Eigenvalues of PCA and standard errors for marketing problem

n	λ	σ_λ (equation (21))	σ_λ (equation (25))
1	10	0·132	1·414
2	0·1	0·013	0·014
3	0·02	0·006	0·003
4	0·01	0·004	0·001

Table 2. Results of PCA for gas chromatography retention index data

		Standard errors				Significance tests						
n	λ	σ_λ (equation (21))	σ_λ (equation (25))	σ_λ (bootstrap)	σ_λ (jack-knife)	CUM (%)	$\hat{\sigma}_M$	R^a	R^b	W^c	IND	ER
1	$6{\cdot}30 \times 10^7$	$0{\cdot}18 \times 10^6$	$6{\cdot}14 \times 10^6$	$6{\cdot}11 \times 10^6$	$6{\cdot}17 \times 10^6$	98·73	21·7	0·02	0·05(1)	494·98	0·255	$1{\cdot}35 \times 10^2$
2	$4{\cdot}68 \times 10^5$	$1{\cdot}52 \times 10^4$	$4{\cdot}56 \times 10^4$	$5{\cdot}41 \times 10^4$	$5{\cdot}38 \times 10^4$	99·47	15·3	0·43	0·63(2)	4·95	0·222	2·91
3	$1{\cdot}61 \times 10^5$	$0{\cdot}89 \times 10^4$	$1{\cdot}57 \times 10^4$	$1{\cdot}79 \times 10^4$	$1{\cdot}90 \times 10^4$	99·72	11·8	0·60	0·79(2)	1·90	0·226	2·26
4	$7{\cdot}13 \times 10^4$	$5{\cdot}94 \times 10^3$	$6{\cdot}94 \times 10^3$	$8{\cdot}41 \times 10^3$	$8{\cdot}73 \times 10^3$	99·83	9·4	0·70	0·88(3)	0·92	0·258	1·75
5	$4{\cdot}07 \times 10^4$	$4{\cdot}49 \times 10^3$	$3{\cdot}97 \times 10^3$	$6{\cdot}35 \times 10^3$	$6{\cdot}76 \times 10^3$	99·89	7·4	0·83	0·96(2)	0·41	0·322	1·48
6	$2{\cdot}76 \times 10^4$	$3{\cdot}69 \times 10^3$	$2{\cdot}68 \times 10^3$	$4{\cdot}21 \times 10^3$	$4{\cdot}73 \times 10^3$	99·94	6·5	0·99	0·90(2)	0·54	0·434	1·65
7	$1{\cdot}67 \times 10^4$	$2{\cdot}88 \times 10^3$	$1{\cdot}63 \times 10^3$	$1{\cdot}89 \times 10^3$	$2{\cdot}01 \times 10^3$	99·96	5·6	—	1·06(3)	0·13	0·684	1·46
8	$1{\cdot}15 \times 10^4$	$2{\cdot}38 \times 10^3$	$1{\cdot}12 \times 10^3$	$2{\cdot}26 \times 10^3$	$2{\cdot}86 \times 10^3$	99·98	4·8	—	0·98(8)	0·11	1·360	1·55
9	$7{\cdot}38 \times 10^3$	$1{\cdot}91 \times 10^3$	$0{\cdot}72 \times 10^3$	$1{\cdot}53 \times 10^3$	$2{\cdot}18 \times 10^3$	99·99	4·1	—	1·11(6)	0·03	4·916	1·51
10	$4{\cdot}88 \times 10^3$	$1{\cdot}55 \times 10^3$	$0{\cdot}48 \times 10^3$	$1{\cdot}04 \times 10^3$	$1{\cdot}86 \times 10^3$	100·00	—	—	—	—	—	—

[a] Reference 18 (213 × 10). [b] Average of ten cross-validations (212 × 10). [c] Reference 19 (212 × 10).

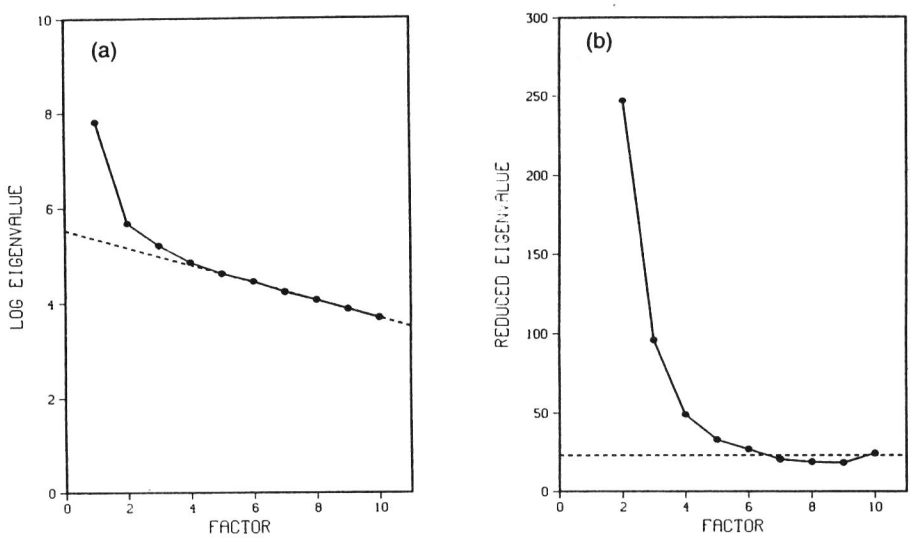


Figure 3. Functions of the eigenvalues for gas chromatography retention index data: (a) logarithm of eigenvalues, the dashed line (---) is fitted through the values for the last six PCs, (b) reduced eigenvalues, the dashed line (---) represents the average value for the last six PCs

significance test. Next we show the cross-validation ratio R obtained by Wold.[18] R should be smaller than unity for a significant PC, but it is clear that the PC with $R = 0{\cdot}99$ cannot have much predictive value. The next column gives the ratios obtained by randomly drawing subgroups ten times from the 212×10 matrix. In this way also the variability caused by the procedure itself can be estimated. The value of R is considerably closer to unity for the fifth PC but smaller for the sixth PC. The fourth PC must be considered significant if unity is taken as a hard limit. The next column gives the value of the W-statistic for the modified cross-validation technique.[19] Strictly speaking, PCs with $W > 1$ should be retained. However, the value of 0·92 for the fourth PC was considered to be large enough by the authors. This conjecture is supported by a later simulation study.[17] In the next column we show the values for the indicator (IND) function.[34] This function should give a minimum for the correct number of PCs. The minimum is not well-defined (as usual), but since the values for the first and fourth PCs are very close, we conclude that the correct number is indicated to be two or three. The last column shows the eigenvalue ratio (ER). For the secondary PCs the ratio should become constant.[33] Use of this criterion would lead to a choice between three and four PCs. This is supported by two graphical methods. Figure 3(a) shows the LEV diagram already discussed by Joliffe.[33] A straight line in the LEV diagram is equivalent to a constant ratio between successive eigenvalues, but the graphical method has the advantage that patterns in the ordered eigenvalues can be more easily detected. Figure 3(b) shows the reduced eigenvalues (REVs) of Malinowski.[7] The REVs should become constant for the secondary PCs:

$$\mathrm{REV}_p = \frac{\lambda_p}{(r-p+1)(c-p+1)} = \text{constant} \tag{26}$$

Again there is an indication that three or perhaps four PCs are significant. Additional support for a three-dimensional model comes from the correlation matrix. Taking solutes 1, 2 and 10

gives the values $\rho_{1,2} = 0{\cdot}988$, $\rho_{1,10} = 0{\cdot}988$ and $\rho_{2,10} = 0{\cdot}965$. Any other solute has a correlation of at least 0·995 with this 'key set'. This result gives credence to the original choice of Wold and Andersson to use the first three PCs for the classification of the LPs.

It follows that there is only a problem with Wold's cross-validation ratio R. The discrepancy can be explained via the number of degrees of freedom involved in the calculation. Mandel[5] has shown that for the secondary PCs a different number holds than for the primary PCs. Correct application of degrees of freedom would therefore inevitably lead to circular reasoning. However, this problem is not unique for cross-validation but applies equally well to every method that in some way makes use of degrees of freedom.

Finally, we show in Figure 4 the reproduction error that remains after successively fitting PCs when random noise is added to the data. If the standard deviation of the added noise is large enough (errors are added in quadrature), the root mean square error (RMS) left after fitting the significant PCs should extrapolate to the standard deviation of the noise present in the original data.[27] Although straight lines are obtained for a standard deviation larger then ten, the extrapolated values are not in the admissible range (11·8–9·4 for three to four PCs). This method should probably only be used for confirmation or to investigate whether there are abnormal data points.

We have chosen to use a three-dimensional model in order to estimate σ_M in (24). A value of $\hat{\sigma}_M = 11{\cdot}8$ is obtained in this way. This leads to the standard errors in the eigenvalues of the third column in Table 2. The fourth column contains the standard errors determined with (25), whereas columns 5 and 6 contain the bootstrap and jack-knife estimates respectively. It is clear that the theoretical standard errors in columns 3 and 4 should be different for the significant PCs. The standard errors in column 3 are smaller than the standard errors in column 4 for the first four PCs. The difference for the fourth PC is, however, rather small.

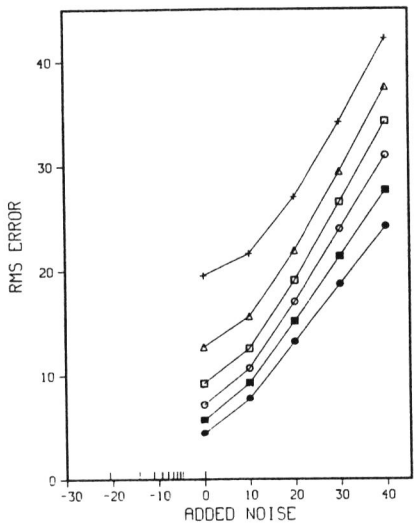


Figure 4. RMS error for gas chromatography retention index data after addition of uncorrelated Gaussian noise for the model with one PC (+), two PCs (Δ), three PCs (□), four PCs (○), five PCs (■) and six PCs (•). The tick marks on the abcissa represent the estimated measurement error for the different PC models

For the last six PCs the situation is reversed. Simulations, presented in a later section, show that the estimate for the secondary PCs is conservative, i.e. biased upwards, and we conclude that the last six or seven PCs must constitute noise. The bootstrap and jack-knife estimates correspond well with each other. The difference between the empirical and theoretical estimates becomes larger starting from the third PC. This all leads to a fairly consistent view of the 212×10 data matrix: take three PCs if maximizing the variance is the prime goal of the data reduction.

Higher-order contributions

Under the assumptions made to derive (18), the covariance matrix for the eigenvalues is diagonal, since the singular vectors associated with different factors are orthogonal. This is, however, in disagreement with Malinowski's theory of error for PCA:[28] the secondary eigenvalues must add up to the residual sum of squares from which the error in the data can be estimated. This is expressed by equation (24). Although the residual sum of squares is also a random variable, higher-order contributions may be important. A similar argument holds for the primary eigenvalues.

Expressions for the higher-order contributions are given in the excellent monograph of Wilkinson[35] on the algebraic eigenvalue problem. For the special case of linear elementary divisors (i.e. a complete set of eigenvectors exists) efficient error bounds can be derived. We will only summarize the relevant expressions and refer to the original text for a detailed proof.

If (i) $\mathbf{A}$ is a *general* $c \times c$ matrix with linear elementary divisors and (ii) $\mathbf{A}$ and $\mathbf{B}$ are matrices that satisfy $|A_{ij}| \leqslant 1$, $|B_{ij}| \leqslant 1$ and λ_n is a simple eigenvalue of $\mathbf{A}$ (i.e. λ_n has multiplicity one, a condition usually met in PCA), then $\lambda_n(\varepsilon)$ is an eigenvalue of $\mathbf{A} + \varepsilon\mathbf{B}$ and

$$\lambda_n(\varepsilon) = \lambda_n + k_n^{(1)}\varepsilon + k_n^{(2)}\varepsilon^2 + \cdots \tag{27}$$

is a convergent power series independent of the multiplicities of the other eigenvalues. Substituting $\mathbf{D}_R = \mathbf{A}$ and $d\mathbf{D}_R = \varepsilon\mathbf{B}$ leads to the following expressions for the first two error terms in the expansion:

$$k_n^{(1)}\varepsilon = 2\boldsymbol{\Theta}_{n\text{-row}} d\boldsymbol{\Theta}_{n\text{-col}} = d\Lambda_{nn} \tag{28a}$$

$$k_n^{(2)}\varepsilon^2 = \sum_{p \neq n}^{c} \frac{d\Lambda_{pn} d\Lambda_{np}^{\mathrm{T}}}{\lambda_n - \lambda_p} \tag{28b}$$

Here the indices 'n-row' and 'n-col' denote the nth row and nth column vector of the corresponding matrix respectively. It is obvious that the second-order term depends greatly on the spacing $\lambda_n - \lambda_p$ of the eigenvalues. The reduced eigenvalues of Malinowski[7] may be useful to estimate it for the secondary set. It follows that one must treat the error propagation as acting differently on the two subsets of eigenvalues rather than acting differently on the individual eigenvalues. Because of the smaller spacing, the contribution of the second-order term will be relatively large for the secondary eigenvalues. Furthermore, many of the terms in the summation become negative ($\lambda_n < \lambda_p$). Consequently, one may construct a significance test on the standard errors, but since it relies in a complicated way upon the magnitude of the eigenvalues, simpler tests are to be preferred, e.g. the eigenvalue ratio test.[27,33] As we have seen for the retention index data, only a very limited choice remains after applying several independent tests.

APPLICATIONS

In the preceding part we have shown that it is possible but not practical to determine the pseudorank of a matrix by estimating the standard errors in the eigenvalues of the cross-product matrix. In the current section we will discuss three topics in data analysis where our derived expressions may find useful application. We will focus on the applications and refer to the literature for a detailed introduction to these subjects.

Scores and loadings of PCA

Often the $r \times c$ data matrix $\mathbf{M}$ is decomposed as

$$\mathbf{M} = \mathbf{AF} \tag{29}$$

The matrices $\mathbf{A}(r \times c)$ and $\mathbf{F}(c \times c)$ are called scores and loadings; the loadings represent the principal axes of $\mathbf{M}$ and the scores represent the co-ordinates on this rotated system respectively. The relation with the SVD of $\mathbf{M}$ is given by the identities $\mathbf{A} = \mathbf{U\Theta}$ and $\mathbf{F} = \mathbf{V}^{\mathrm{T}}$. Expressions for the errors in the scores and loadings are derived by Malinowski[11] and Roscoe and Hopke.[12,13] The difference between the two methods is that the first leads to estimates of the error of the vectors as a whole whereas the second yields estimates of the error in the individual elements of the vectors. A drawback of both methods is that they depend explicitly on the number of PCs retained in the model. Using our derived standard errors, it is possible to find expressions for the errors in the loadings and scores without this restriction. We essentially follow the reasoning of Roscoe and Hopke.[12,13]

The error in $\mathbf{M}$ is expressed in the contributions from $\mathbf{A}$ and $\mathbf{F}$:

$$d\mathbf{M} = \tilde{\mathbf{A}}\,d\mathbf{F} + d\mathbf{A}\tilde{\mathbf{F}} \tag{30a}$$

Assuming that all measurement error resides in $\mathbf{A}$ gives

$$d\mathbf{M} = d\mathbf{A}\tilde{\mathbf{F}} \tag{30b}$$

Assuming that all measurement error resides in $\mathbf{F}$ gives

$$d\mathbf{M} = \tilde{\mathbf{A}}\,d\mathbf{F} \tag{30c}$$

Using equation (66) and substituting the identities for $\mathbf{F}$ and $\mathbf{A}$ leads after rearrangement (diagonal matrices commute) to

$$d\mathbf{A} = \tilde{\mathbf{A}}(d\mathbf{\Theta}\tilde{\mathbf{\Theta}}^{-1}) \tag{31a}$$

$$d\mathbf{F} = (d\mathbf{\Theta}\tilde{\mathbf{\Theta}}^{-1})\tilde{\mathbf{F}} \tag{31b}$$

$d\mathbf{A}$ is found by multiplying the *columns* of $\mathbf{A}$ by the relative error in $\mathbf{\Theta}$, whereas $d\mathbf{F}$ is found by multiplying the *rows* of $\mathbf{F}$ by the relative error in $\mathbf{\Theta}$. This result is consistent with the mixed use of row and column vectors in (29). Introducing the derived expression for $d\mathbf{\Theta}$ will finally give error estimates for the scores and loadings. For heteroscedastic noise (uncorrelated) one finds

$$\hat{\sigma}_{A_{in}} = \left(\sum_{k=1}^{c} (V_{kn})^2 \hat{\sigma}^2_{M_{ik}} \right)^{1/2} \tag{32a}$$

$$\hat{\sigma}_{F_{nk}} = \lambda_n^{-1/2} \left(\sum_{i=1}^{r} (U_{in})^2 \hat{\sigma}^2_{M_{ik}} \right)^{1/2} \tag{32b}$$

and for homoscedastic noise the indices k and i are summed out to give

$$\hat{\sigma}_A \equiv \hat{\sigma}_{A_{n\text{-col.}}} = \hat{\sigma}_{\mathbf{M}} \tag{33a}$$

$$\hat{\sigma}_{F_{n\text{-row}}} = \lambda_n^{-1/2} \hat{\sigma}_{\mathbf{M}} \tag{33b}$$

These estimates will be overestimates because of the assumptions underlying (30b) and (30c). Furthermore, they are internally inconsistent, since the relation between the errors in scores and loadings expressed by equation (65) no longer holds.

Comparison with the theory of Malinowski shows that (33b) is identical to equation (8) in Reference 11. It follows that the assumptions used by Roscoe and Hopke to arrive at their error estimates essentially come down to the assumption of 'small' errors in the derivation of Malinowski. The only remaining difference seems to be the additional assumption of homoscedastic noise by Malinowski. The agreement between the two theories for the homoscedastic case is a good indication of the reliability of the error estimates represented by (32) and (33).

Because the eigenvalues λ_n are monotonously decreasing, it is expected that the instability of the eigenvectors of $\mathbf{M}^T\mathbf{M}$ increases with factor number. Furthermore, since the secondary eigenvalues are usually much smaller than the primary eigenvalues, the associated secondary eigenvectors should be characterized by a much larger instability. Equation (33b) seems to be in contradiction with (21). However, after extracting the primary factors, the distribution of the residuals becomes almost spherical. Therefore the *length* of the principal axes must and will become more and more fixed, because a sphere is completely defined by the length of the radius, whereas the *direction* of the axes becomes more and more undeterminate. This relative instability of the secondary eigenvectors is e.g. employed by Duewer and Kowalski[36] for rank estimation.

First-order tensorial calibration for linear, additive model

If the response for a sample can be cast in a vector, first-order tensorial or multivariate calibration techniques can be used.[14] Quantitation always proceeds in two steps. In the first step the instrument responses are recorded from a set of calibration samples:

$$\mathbf{R} = \mathbf{SC} + \mathbf{E} \tag{34}$$

where $\mathbf{R}$ is the matrix of calibration data responses, $\mathbf{S}$ is the matrix of sensitivities, $\mathbf{C}$ is the matrix of concentrations in the calibration set and $\mathbf{E}$ is the matrix of response residuals. $\mathbf{R}$ and $\mathbf{E}$ are dimensioned $J \times I$, $\mathbf{S}$ is $J \times K$ and $\mathbf{C}$ is $K \times I$, with J the number of responses, I the number of calibration samples and K the number of chemical species. The matrix of sensitivities (pure component responses) is estimated from a least squares fit as $\hat{\mathbf{S}} = \mathbf{RC}^+$. In the second step the instrument response is recorded for the unknown sample and fitted to the model

$$\mathbf{r} = \hat{\mathbf{S}}\mathbf{c} + \mathbf{e} \tag{35}$$

where $\mathbf{r}$ denotes the $J \times 1$ response vector for the unknown sample and $\mathbf{c}$ is the $K \times 1$ vector of unknown concentrations. The unknown concentrations are estimated by

$$\hat{\mathbf{c}} = \hat{\mathbf{S}}^+\mathbf{r} = \mathbf{CR}^+\mathbf{r} \tag{36a}$$

The advantage of substituting for $\hat{\mathbf{S}}^+$ becomes clear from the expression for the individual

components:

$$\hat{c}_n = \mathbf{C}_{n-\text{row}}\mathbf{R}^+\mathbf{r} \tag{36b}$$

Only $\mathbf{C}_{n-\text{row}}$, i.e. the row in $\mathbf{C}$ corresponding to the analyte of interest, is needed for the prediction. This situation is usually referred to as *partial* calibration. It is sufficient that the interferents vary independently in the calibration set, so that $\mathbf{CC}^+$ equals the $K \times K$ identity matrix $\mathbf{I}_K$. For interferents that cannot be accounted for in this way, an explicit background correction has to be made.[24] This leads to a model where $\mathbf{R}$ and $\mathbf{r}$ are replaced by $\mathbf{R} - \mathbf{B}$ and $\mathbf{r} - \mathbf{b}$ respectively. A model for which the complete concentration matrix $\mathbf{C}$ has to be known is called a *total* calibration model.

Lorber and Kowalski[37] have derived estimates for the prediction error for the partial calibration approach starting from the equation (our notation)

$$\tilde{c}_n + \delta c_n = (\tilde{\mathbf{C}}_{n-\text{row}} + \delta\mathbf{C}_{n-\text{row}})(\tilde{\mathbf{R}} + \delta\mathbf{R})^+(\tilde{\mathbf{r}} + \delta\mathbf{r}) \tag{37}$$

Depending on how the pseudoinverse of $\tilde{\mathbf{R}} + \delta\mathbf{R}$ is estimated, the results apply to e.g. multiple linear regression (MLR), principal component regression (PCR) or a modified version of partial least squares (PLS).[38] It should be noted that for PCR and PLS, apart from the variance (spread around the mean), a bias term (deviation of the mean from the true value) also contributes to the mean square error (MSE). Error propagation does not account for bias.

The crucial step in their derivation is to separate the contributions of the true response and the corresponding error by decomposing $(\tilde{\mathbf{R}} + \delta\mathbf{R})^+$ according to the SVD

$$(\tilde{\mathbf{R}} + \delta\mathbf{R})^+ = \tilde{\mathbf{V}}(\tilde{\Theta} + \delta\Theta)^+\tilde{\mathbf{U}}^T = \tilde{\mathbf{V}}(\tilde{\Theta} + \delta\Theta)^{-1}\tilde{\mathbf{U}}^T \tag{38}$$

Consistent with the original significance test of Hugus and El-Awady,[1] the error in the singular values is taken to be equal to the first non-significant singular value θ_{K+1}. The results of this approach are disappointing and an alternative method is proposed that gives good results. However, the physical background of the second approach is unclear. This may explain the fact that a review[39] of the method shows that it has not been extensively used.

Recently, Bauer *et al.*[24] have developed prediction errors using first-order error propagation. Including a background term in the model results in

$$d\mathbf{c} = [-\mathbf{S}^+(d\mathbf{R} - d\mathbf{B}) + d\mathbf{C}]\mathbf{C}^+\mathbf{c} + \mathbf{S}^+(d\mathbf{r} - d\mathbf{b}) \tag{39}$$

and

$$d\mathbf{c} = [-\mathbf{C}(\mathbf{R} - \mathbf{B})^+(d\mathbf{R} - d\mathbf{B}) + d\mathbf{C}]\mathbf{C}^+\mathbf{C}(\mathbf{R} - \mathbf{B})^+(\mathbf{r} - \mathbf{b}) + \mathbf{C}(\mathbf{R} - \mathbf{B})^+(d\mathbf{r} - d\mathbf{b}) \tag{40a}$$

for the total and partial calibration approaches respectively. Since (40a) contains the matrix $\mathbf{C}^+$, it can only be evaluated if the complete matrix $\mathbf{C}$ is known. The authors conclude that for the calculation of the prediction error there is no advantage in using the partial calibration approach. They have tested the adequacy of (39) on experimental ICP–OES data and the errors in the concentrations are predicted satisfactorily, depending on the quality of the estimates of the errors of the measured signals.

However, using a consistent notation with errorless quantities,

$$d\mathbf{c} = [-\tilde{\mathbf{C}}(\tilde{\mathbf{R}} - \tilde{\mathbf{B}})^+(d\mathbf{R} - d\mathbf{B}) + d\mathbf{C}]\tilde{\mathbf{C}}^+\tilde{\mathbf{C}}(\tilde{\mathbf{R}} - \tilde{\mathbf{B}})^+(\tilde{\mathbf{r}} - \tilde{\mathbf{b}}) + \tilde{\mathbf{C}}(\tilde{\mathbf{R}} - \tilde{\mathbf{B}})^+(d\mathbf{r} - d\mathbf{b}) \tag{40b}$$

the pseudoinverse of the concentration matrix can be worked out, since $\tilde{\mathbf{C}}^+\tilde{\mathbf{C}}(\tilde{\mathbf{R}} - \tilde{\mathbf{B}})^+ = \tilde{\mathbf{C}}^+\tilde{\mathbf{S}}^+ = (\tilde{\mathbf{R}} - \tilde{\mathbf{B}})^+$. This simplification is not possible for the experimental quantities in (40a). Therefore we conclude that by approximating the right hand side of (38) by

$$\tilde{\mathbf{V}}(\tilde{\Theta} + \delta\Theta)^{-1}\tilde{\mathbf{U}}^T \approx \tilde{\mathbf{V}}(\tilde{\Theta} + d\Theta)^{-1}\tilde{\mathbf{U}}^T \approx \tilde{\mathbf{V}}(\tilde{\Theta}^{-1} - d\Theta\tilde{\Theta}^{-2})\tilde{\mathbf{U}}^T \tag{41}$$

and including the appropriate expression for the standard error in the singular values, the results should become essentially the same as for the expression of Bauer *et al.*[24] The only remaining difference would consist of the inclusion of the background term in the latter.

This is confirmed by the very recent publication of Karstang *et al.*,[40] where the procedure of Lorber and Kowalski is modified by multiplying out (37) using $(\tilde{\mathbf{R}} + \delta\mathbf{R})^+ = \tilde{\mathbf{R}}^+ + \delta(\mathbf{R}^+)$. The expressions of Karstang *et al.*[40] differ from those of Bauer *et al.*[24] by the substitution of the matrix of sensitivities by the score matrix. The use of scores enables the identification of the position of a sample in the predictor space. Karstang *et al.* show that the estimated prediction errors are close to the actual prediction errors for samples within the calibration range. For samples containing uncalibrated interferents a background correction technique must be applied before estimation of prediction errors.

Second-order tensorial calibration for linear, additive model

If the response for a sample can be cast in a matrix, second-order tensorial or bilinear calibration techniques can be used.[15] We will restrict ourselves to the discussion of a method developed for bilinear data. For non-bilinear data, non-bilinear rank annihilation (NBRA)[41] and residual bilinearization (RBL)[30] should be more appropriate. (RBL assumes that the residuals rather than the responses are bilinear.)

Ho *et al.*[42] have developed an iterative procedure for the analysis of a single analyte in the presence of an uncalibrated background. This method, called rank annihilation factor analysis (RAFA), is modified by Lorber[43] to a direct method which is generalized by Sánchez and Kowalski[44] to the simultaneous quantitation of several analytes in the presence of unknown interferents. Their method, called the generalized rank annihilation method (GRAM), makes use of two data matrices, **M** for the unknown and **N** for the calibration sample:

$$\mathbf{M} = \mathbf{X}\mathbf{C}_\mathbf{M}\mathbf{Y}^\mathrm{T} = \tilde{\mathbf{X}}\tilde{\mathbf{C}}_\mathbf{M}\tilde{\mathbf{Y}}^\mathrm{T} + \mathbf{E}_{\mathbf{M},\mathrm{real}} \tag{42a}$$

$$\mathbf{N} = \mathbf{X}\mathbf{C}_\mathbf{N}\mathbf{Y}^\mathrm{T} = \tilde{\mathbf{X}}\tilde{\mathbf{C}}_\mathbf{N}\tilde{\mathbf{Y}}^\mathrm{T} + \mathbf{E}_{\mathbf{N},\mathrm{real}} \tag{42b}$$

Suppose without loss of generality that the data matrices are collected from an HPLC–UV experiment and K is the total number of different components for the two samples. Then S spectra are obtained at W wavelengths, so that **M** ($S \times W$) and **N** ($S \times W$) contain the mixture spectra, **X** ($S \times K$) contains the pure component elution profiles, **Y** ($W \times K$) contains the pure component spectra and $\mathbf{C}_\mathbf{M}$ and $\mathbf{C}_\mathbf{N}$ are $K \times K$ diagonal matrices with calibration factors. The $S \times W$ error matrices $\mathbf{E}_{\mathbf{M},\mathrm{real}}$ and $\mathbf{E}_{\mathbf{N},\mathrm{real}}$ denote the difference between the experimental and the errorless data, i.e. the real error.[28]

In the most general case (i.e. both samples contain unique components) the matrices $\tilde{\mathbf{C}}_\mathbf{M}$ and $\tilde{\mathbf{C}}_\mathbf{N}$ contain zeros at different positions. Therefore a data matrix **Q** has to be constructed that spans a space describing all components. In the following we assume that **Q** is constructed as the sum of **M** and **N**, so that the matrix $\tilde{\mathbf{C}}_\mathbf{M} + \tilde{\mathbf{C}}_\mathbf{N}$ contains no zeros:

$$\mathbf{Q} = \mathbf{M} + \mathbf{N} = \tilde{\mathbf{X}}(\tilde{\mathbf{C}}_\mathbf{M} + \tilde{\mathbf{C}}_\mathbf{N})\tilde{\mathbf{Y}}^\mathrm{T} + \mathbf{E}_{\mathbf{Q},\mathrm{real}} \tag{43}$$

This operation can be interpreted as a 'simulated' standard addition. Next, **Q** is reproduced from the first F ($\geqslant K$) principal components by the truncated SVD

$$\mathbf{Q} = \mathbf{U}_\mathbf{Q}\boldsymbol{\Theta}_\mathbf{Q}\mathbf{V}_\mathbf{Q}^\mathrm{T} + \mathbf{E}_{\mathbf{Q},\mathrm{extr}} \tag{44}$$

Here the error matrix $\mathbf{E}_{\mathbf{Q},\mathrm{extr}}$ denotes the difference between the experimental and the reproduced sum matrix, i.e. the extracted error.[28] The resulting prediction equation is an $F \times F$

standard eigenvalue problem

$$(\mathbf{U}_Q^T\mathbf{M}\mathbf{V}_Q\boldsymbol{\Theta}_Q^{-1})\mathbf{Z} = \mathbf{Z}\boldsymbol{\Pi} \tag{45}$$

where $\boldsymbol{\Pi}$ is a diagonal matrix of eigenvalues, related to the calibration factors as $\boldsymbol{\Pi} = \mathbf{C}_M(\mathbf{C}_M + \mathbf{C}_N)^{-1}$, and $\mathbf{Z}$ is the matrix of right eigenvectors. Evidently, this eigenvalue problem is different from the eigenvalue problem of PCA. Relevant for the current discussion is the distinct possibility of degeneracy. Here degeneracy is the result of a combination of (nearly) identical concentration ratios and noise. The pure component responses and the concentrations for the desired analytes in the unknown sample can be derived using[44]

$$\mathbf{X}(\mathbf{C}_M + \mathbf{C}_N) = \mathbf{U}_Q\mathbf{Z} \tag{46a}$$

$$\mathbf{Y}^T = \mathbf{Z}^{-1}\boldsymbol{\Theta}_Q\mathbf{V}_Q^T \tag{46b}$$

$$\mathbf{C}_M = \mathbf{C}_N\boldsymbol{\Pi}(\mathbf{I}_F - \boldsymbol{\Pi})^{-1} \tag{46c}$$

where $\mathbf{I}_F$ denotes the $F \times F$ identity matrix. Only for simple eigenvalues will the solution for the pure component responses be unique (up to a constant), since the direction of eigenvectors corresponding to degenerate eigenvalues is not fixed.

Error estimates have been reported for the eigenvalues obtained by the iterative procedure[45] as well as for the eigenvalues obtained by GRAM.[46] Inspection of these error estimates shows that only the error in the decomposed data matrix $\mathbf{Q}$ is considered. Comparison with our error estimates will therefore be difficult.

Assuming uncorrelated errors with a constant and equal standard deviation in the data matrices $\mathbf{M}$ and $\mathbf{N}$, i.e. $\sigma_M = \sigma_N$, leads to the following expression for the estimated standard errors in the eigenvalues of the GRAM equation (see Appendix III):

$$\hat{\sigma}_{\pi_n} = \hat{\sigma}_M\left((1 - 2\pi_n + 2\pi_n^2)\sum_{p=1}^{F}\frac{(Z_{pn})^2}{\lambda_{Q,p}}\right)^{1/2} \tag{47}$$

Here $\pi_n = \Pi_{nn}$ and $\lambda_{Q,p} = \Theta_{Q,pp}^2$. (This expression also gives a unique result only for simple eigenvalues.) If the matrix $\mathbf{Q}$ is obtained by performing 'real' standard additions in the unknown sample, the standard errors are given by (see Appendix III)

$$\hat{\sigma}_{\pi_n} = \hat{\sigma}_M\left((1 + \pi_n^2)\sum_{p=1}^{F}\frac{(Z_{pn})^2}{\lambda_{Q,p}}\right)^{1/2} \tag{48}$$

Equations (47) and (48) clearly show the contribution of the noise factors ($F > K$) to the efficiency of the concentration estimates. We are currently investigating the merits of our error estimate, which will be discussed in a future publication. In this paper, we will confine ourselves to the illustration of the reduction of variance that can be achieved by 'simulating' standard addition.

EXPERIMENTAL

We have constructed three-component systems by convoluting the RNA spectra of Zscheile *et al.*[47] with Gaussian elution profiles. The spectra are normalized to unit length in order to make the contribution to the total variance proportional to the square of the peak height. Measures for the overlap are the inner product and the linear correlation coefficient. These are given in Table 3. It can be inferred from these numbers that there is only a moderate overlap between the spectra of adenine and guanine. This is balanced by the large chromatographic separation of these components. The signals of adenine and guanine are held constant at

Table 3. Inner product (right upper corner) and linear correlation coeffient (left lower corner) for UV spectra and elution profiles. Spectra and elution profiles are normalized

	UV spectra			Elution profiles		
Component	A	C	G	A	C	G
Adenine	1	0·70	0·95	1	0·45	0·04
Cytidine	−0·24	1	0·84	−0·05	1	0·45
Guanine	0·79	0·12	1	−0·78	−0·06	1

Table 4. Peak heights (in mAU) of elution profiles of adenine, cytidine and guanine for SNR = 2000, 10 and 6. For definition of SNR see text

Experiment	SNR	Adenine	Cytidine	Guanine
1	2000	1000	1000	1000
2	10	1000	5	1000
3	6	1000	3	1000

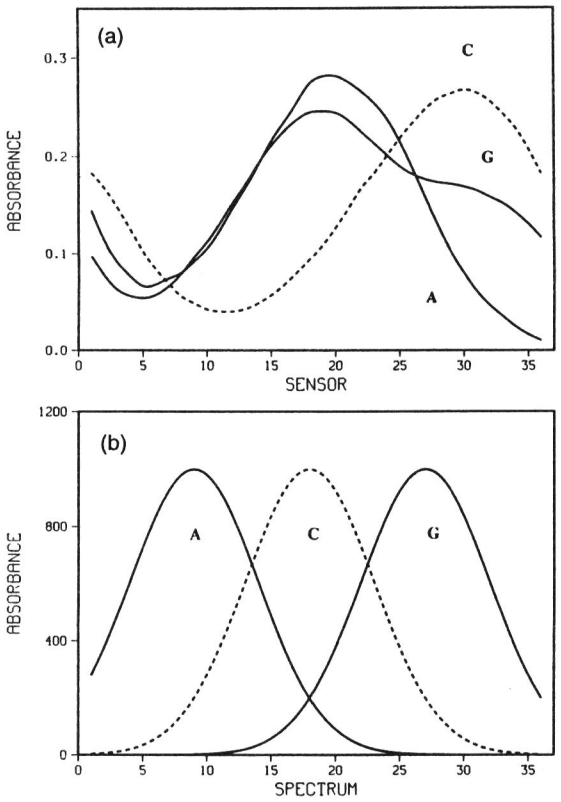


Figure 5. (a) Normalized UV spectra and (b) denormalized elution profiles of adenine (A), guanine (G) and cytidine (C). The dashed line (---) is the curve of the dilute component

1000 mAU while the signal of cytidine is lowered. Artificial Gaussian noise with standard deviation 0·5 mAU (absolute value) is added. The experiments show resemblance to the simulations executed by Tu *et al.*[48] The signal-to-noise ratio (SNR) is defined as the ratio of the peak height of the dilute component to the standard deviation of the noise. The degree of difficulty of an ideal bilinear data set is a combination of the factors overlap and SNR. The chosen levels are typical for HPLC–UV data in practice. A summary of the dilutions is given in Table 4. The elution profiles have a standard deviation of ten spectra and are located at positions 9, 18 and 27. Pictures of the spectra and elution profiles are shown in Figure 5. The resulting data matrices have dimensions 36×36.

RESULTS AND DISCUSSION

We will only give simulation results for the standard errors in the eigenvalues of PCA, the standard errors in the scores and loadings and the standard errors in the eigenvalues of GRAM. The approach has already proved to be successful in multivariate calibration by the work of Bauer *et al.*[24] and Karstang *et al.*[40]

Standard errors in the eigenvalues of PCA

We have tested the adequacy of our derived equation by generating large MC samples of ideal HPLC–UV data matrices. The MC method gives very precise estimates if sufficient simulations are performed. We have estimated the true variance of the eigenvalues from samples of 100 and 10 000 trials. In this way an idea of the precision of the MC estimate is obtained. Since we simulate homoscedastic noise, we compare (10) directly with (21) instead of (20). The evaluation of (21) does not involve MC simulations.

In the first experiment all components have peak heights of 1000 mAU. The resulting SNR is therefore 2000. The first six eigenvalues and estimated standard errors are given in Table 5. We notice that the estimates according to (10) are almost constant for the primary and secondary factors. This is the same behaviour as reported in the literature, and using the criterion of Hugus and El-Awady,[1] three factors are classified as significant. The estimates according to (21) are seen to be very good for the primary factors. Moreover, there is no visible trend. These estimates could have been 'improved' by using the average eigenvalue instead of the eigenvalue for one MC trial, but this would remove one of the approximations in an artificial way. The results for the secondary factors are a gross overestimate (by a factor of two). This is a consequence of the much smaller spacing of the secondary eigenvalues. The

Table 5. First six eigenvalues and estimated standard errors for SNR = 2000

n	λ	σ_λ[a] (equation (10))	σ_λ (equation (21))	σ_λ[a] (MC)	σ_λ[b] (MC)
1	$3{\cdot}79 \times 10^7$	$8{\cdot}71 \times 10^2$	$6{\cdot}15 \times 10^3$	$6{\cdot}11 \times 10^3$	$6{\cdot}51 \times 10^3$
2	$1{\cdot}25 \times 10^6$	$8{\cdot}13 \times 10^2$	$1{\cdot}12 \times 10^3$	$1{\cdot}12 \times 10^3$	$1{\cdot}05 \times 10^3$
3	$1{\cdot}84 \times 10^5$	$7{\cdot}10 \times 10^2$	$4{\cdot}29 \times 10^2$	$4{\cdot}25 \times 10^2$	$4{\cdot}52 \times 10^2$
4	$2{\cdot}84 \times 10$	$7{\cdot}60 \times 10^2$	5·32	2·51	2·01
5	$2{\cdot}66 \times 10$	$7{\cdot}59 \times 10^2$	5·15	1·90	1·68
6	$2{\cdot}39 \times 10$	$7{\cdot}62 \times 10^2$	4·88	1·59	1·44

[a] Monte Carlo estimate from 10^4 trials. [b] Monte Carlo estimate from 10^2 trials.

large overestimate for the secondary eigenvalues can be attributed to the extremely slow convergence of the expansion in (27). The factor of two indicates that including the second-order term will not really improve the situation. This is, however, an artificial problem, since in practice one is usually interested in the estimates for the significant factors. Furthermore, estimates for the standard errors in the secondary eigenvalues can be obtained from MC simulations of random matrices.[5] It is interesting to see that a very large MC sample is needed to accurately estimate the variance of the primary eigenvalues. This is no problem if the data are generated by computer, but clearly makes it difficult to test the theory if data matrices have to be collected in practice. For real data one could apply the method of bootstrapping[21] by drawing residuals with replacement if the dimensionality of the model is known. In that case one may also insert the average eigenvalue in (21) to obtain an improved error estimate. However, this method is not completely safe, because part of the error remains imbedded[28] in the model, so that abnormal data points may cause a bias.

In the second experiment cytidine is diluted to a peak height of 5 mAU. The resulting SNR is therefore ten. The first six eigenvalues and estimated standard errors are given in Table 6. Again the criterion of Hugus and El-Awady indicates the presence of three components. The first-order estimate according to (21) is excellent for the first two eigenvalues, but overestimates the true standard error for the third eigenvalue by 20%. The eigenvalue that corresponds to cytidine has become so close to the noise eigenvalues that higher-order contributions are no longer negligible. However, the crude first-order result can still be inserted in other expressions if conservative estimates are needed.

In the third experiment cytidine is diluted to a peak height of 3 mAU. The resulting SNR is therefore six. The first six eigenvalues and estimated standard errors are given in Table 7.

Table 6. First six eigenvalues and estimated standard errors for SNR = 10

n	λ	σ_λ[a] (equation (10))	σ_λ (equation (21))	σ_λ[a] (MC)	σ_λ[b] (MC)
1	$1{\cdot}79 \times 10^7$	$6{\cdot}67 \times 10^2$	$4{\cdot}23 \times 10^3$	$4{\cdot}19 \times 10^3$	$4{\cdot}35 \times 10^3$
2	$4{\cdot}46 \times 10^5$	$4{\cdot}09 \times 10^2$	$6{\cdot}68 \times 10^2$	$6{\cdot}65 \times 10^2$	$6{\cdot}53 \times 10^2$
3	$4{\cdot}82 \times 10$	$5{\cdot}16 \times 10^2$	6·94	5·67	5·49
4	$2{\cdot}72 \times 10$	$5{\cdot}20 \times 10^2$	5·21	2·40	2·03
5	$2{\cdot}68 \times 10$	$5{\cdot}19 \times 10^2$	5·18	1·81	1·73
6	$2{\cdot}33 \times 10$	$5{\cdot}19 \times 10^2$	4·83	1·52	1·40

[a] Monte Carlo estimate from 10^4 trials. [b] Monte Carlo estimate from 10^2 trials.

Table 7. First six eigenvalues and estimated standard errors for SNR = 6

n	λ	σ_λ[a] (equation (10))	σ_λ (equation (21))	σ_λ[a] (MC)	σ_λ[b] (MC)
1	$1{\cdot}79 \times 10^7$	$6{\cdot}67 \times 10^2$	$4{\cdot}23 \times 10^3$	$4{\cdot}19 \times 10^3$	$4{\cdot}35 \times 10^3$
2	$4{\cdot}46 \times 10^5$	$4{\cdot}09 \times 10^2$	$6{\cdot}68 \times 10^2$	$6{\cdot}65 \times 10^2$	$6{\cdot}53 \times 10^2$
3	$3{\cdot}44 \times 10$	$5{\cdot}19 \times 10^2$	5·86	3·16	3·07
4	$2{\cdot}70 \times 10$	$5{\cdot}20 \times 10^2$	5·20	2·05	1·80
5	$2{\cdot}64 \times 10$	$5{\cdot}19 \times 10^2$	5·14	1·67	1·48
6	$2{\cdot}24 \times 10$	$5{\cdot}19 \times 10^2$	4·73	1·45	1·37

[a] Monte Carlo estimate from 10^4 trials. [b] Monte Carlo estimate from 10^2 trials.

According to the test of Hugus and El-Awady, only two factors are significant. Again the first-order estimate is very good for the first two eigenvalues only. The standard error for the third eigenvalue is overestimated by 85%. However, considering the extreme SNR for this dilution, the results with (21) can be seen as promising.

To illustrate the fact that the last experiment is close to a general breakdown point, we have plotted the first three scores (i.e. abstract elution profiles) and loadings (i.e. abstract spectra) in Figures 6 and 7 for all dilutions. At the two highest values of SNR one clearly recognizes three significant factors. At the lowest value of SNR it becomes difficult to identify structure for the third eigenvector because of the large contribution of the noise. Many significance tests will fail to indicate the presence of the minor component in the last case. Visual inspection of the eigenvectors proves to be a very sensitive method for determining the number of significant factors for ordered data.[49]

We have also executed these simulations with uniform noise. The results are summarized in Table 8. The MC estimates are based on 10^4 trials. As expected, the results are identical to those obtained for Gaussian noise within the statistical uncertainty of the procedure.

Standard errors in the scores and loadings of PCA

The standard errors in the factor scores and loadings are given in Table 9. The simulation results are based on 10^4 MC trials with Gaussian noise. In order to obtain the reported MC values, the sign of the vectors has to be fixed. We have fixed the sign by calculating the correlation of the vectors of the noise-perturbed matrices with those of the errorless data and reversing the sign if the correlation is negative. The uncertainty in the standard errors is estimated by averaging over the components of the corresponding vector. This procedure does not apply to the secondary PCs, since these are missing for the errorless data. Therefore we only make the comparison for the primary PCs. It can be seen that the first-order estimates according to (33) are very accurate for the experiment with SNR = 2000. The empirical MC estimates are slightly overestimated by the theoretical values. For low values of SNR the MC values are much higher for the dilute component. For SNR = 10 the difference is already 28% for the scores and 30% for the loadings, compared with a value of 20% found for the eigenvalues. These results indicate the limitations of the assumption of 'small' errors.

Standard errors in the eigenvalues of GRAM

Using the spectra and elution profiles from the preceding parts, we have constructed standard addition data suitable for analysis with GRAM. The peak heights are chosen in such a way that degeneracy for the eigenvalues is unlikely to occur (see Table 10). Furthermore, the expected eigenvalues are close to the ideal value of 0·5, obtained only if the amount of the added standard is equal to the amount initially present. It is seen in Table 10 that the theoretical standard errors predict the MC values very well. A variance reduction by a factor of 1·6 is obtained if the standard addition is performed by adding the data matrices ('simulated') instead of adding the samples ('real'). However, the real virtue of this variance reduction must not be overestimated, since in practice the error in the concentrations of the calibration sample is often much larger than the error in the responses. The impact on the error in the unknown concentrations may in fact be negligible. The real gain could lie in the improvement of the qualitative solution, summarized in Table 11. The qualitative solution is important for the recognition of the reconstructed components. Here a gain by a factor of 1·6 would not be spoiled by other errors.

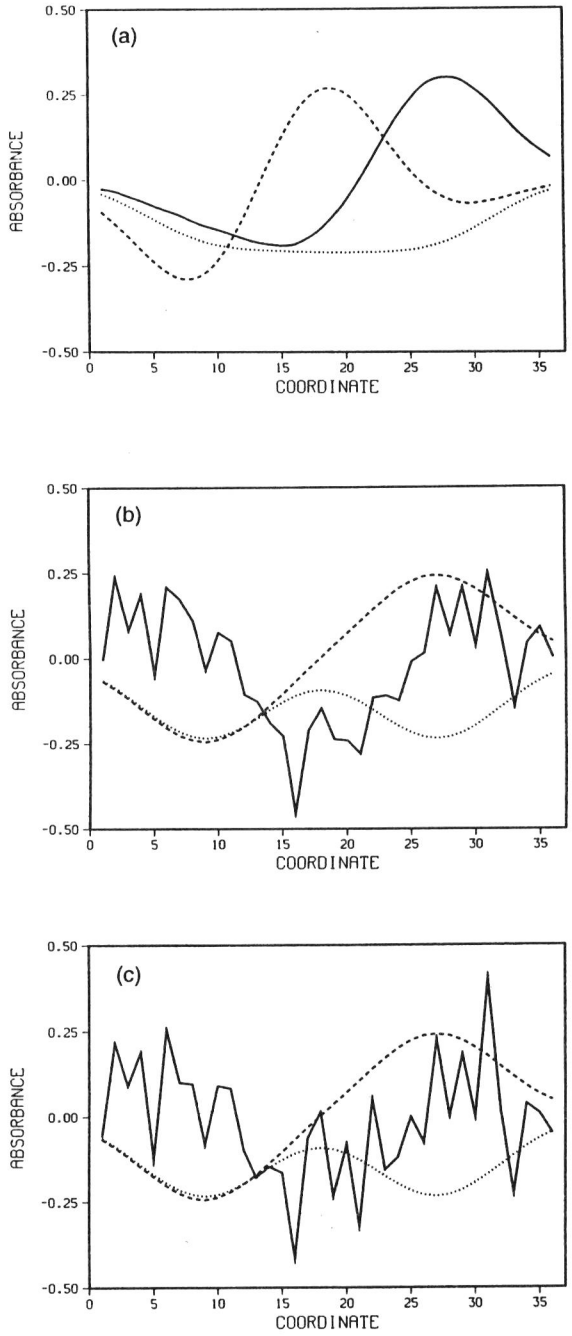


Figure 6. Normalized scores 1 (...), 2 (---) and 3 (——) for SNR = (a) 2000, (b) 10 and (c) 6

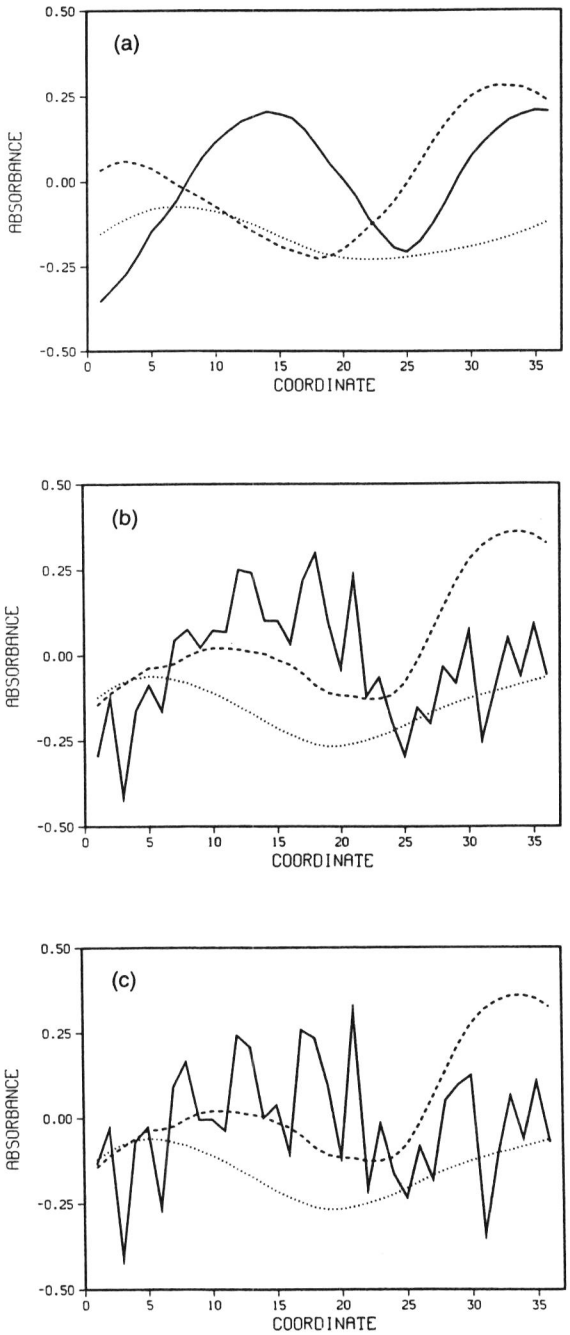


Figure 7. Loadings 1 (...), 2 (---) and 3 (——) for SNR = (a) 2000, (b) 10 and (c) 6

Table 8. Eigenvalues and standard errors within samples for uniform distributed noise

	SNR = 2000			SNR = 10			SNR = 6		
n	λ	σ_λ (equation (21))	σ_λ (MC)	λ	σ_λ (equation (21))	σ_λ (MC)	λ	σ_λ (equation (21))	σ_λ (MC)
1	$3{\cdot}79 \times 10^{7}$	$6{\cdot}15 \times 10^{3}$	$6{\cdot}10 \times 10^{3}$	$1{\cdot}79 \times 10^{7}$	$4{\cdot}23 \times 10^{3}$	$4{\cdot}19 \times 10^{3}$	$1{\cdot}79 \times 10^{7}$	$4{\cdot}23 \times 10^{3}$	$4{\cdot}19 \times 10^{3}$
2	$1{\cdot}25 \times 10^{6}$	$1{\cdot}12 \times 10^{3}$	$1{\cdot}13 \times 10^{3}$	$4{\cdot}46 \times 10^{5}$	$6{\cdot}68 \times 10^{2}$	$6{\cdot}72 \times 10^{2}$	$4{\cdot}46 \times 10^{5}$	$6{\cdot}68 \times 10^{2}$	$6{\cdot}72 \times 10^{2}$
3	$1{\cdot}84 \times 10^{5}$	$4{\cdot}29 \times 10^{2}$	$4{\cdot}27 \times 10^{2}$	$5{\cdot}60 \times 10^{2}$	$7{\cdot}48$	$5{\cdot}60$	$3{\cdot}82 \times 10^{2}$	$6{\cdot}13$	$2{\cdot}95$
4	$3{\cdot}08 \times 10$	$5{\cdot}55$	$2{\cdot}15$	$3{\cdot}07 \times 10$	$5{\cdot}54$	$2{\cdot}10$	$3{\cdot}06 \times 10$	$5{\cdot}53$	$1{\cdot}82$
5	$2{\cdot}73 \times 10$	$5{\cdot}23$	$1{\cdot}58$	$2{\cdot}56 \times 10$	$5{\cdot}05$	$1{\cdot}55$	$2{\cdot}41 \times 10$	$4{\cdot}91$	$1{\cdot}44$
6	$2{\cdot}29 \times 10$	$4{\cdot}79$	$1{\cdot}32$	$2{\cdot}30 \times 10$	$4{\cdot}79$	$1{\cdot}30$	$2{\cdot}29 \times 10$	$4{\cdot}79$	$1{\cdot}24$

Table 9. Standard errors in factor scores and loadings

	SNR = 2000			SNR = 10			SNR = 6		
n	σ_A (MC)	σ_F (equation (33b))	σ_F (MC)	σ_A (MC)	σ_F (equation (33b))	σ_F (MC)	σ_A (MC)	σ_F (equation (33b))	σ_F (MC)
1	0·5010(7)	$8{\cdot}127 \times 10^{-5}$	$8{\cdot}027(17) \times 10^{-5}$	0·5009(7)	$1{\cdot}182 \times 10^{-4}$	$1{\cdot}167(3) \times 10^{-4}$	0·5009(7)	$1{\cdot}834 \times 10^{-4}$	$1{\cdot}168(3) \times 10^{-4}$
2	0·4986(10)	$4{\cdot}465 \times 10^{-4}$	$4{\cdot}378(17) \times 10^{-4}$	0·4940(10)	$7{\cdot}489 \times 10^{-4}$	$7{\cdot}274(28) \times 10^{-4}$	0·4936(12)	$7{\cdot}489 \times 10^{-4}$	$7{\cdot}275(21) \times 10^{-4}$
3	0·4869(15)	$1{\cdot}167 \times 10^{-3}$	$1{\cdot}120(4) \times 10^{-3}$	0·6375(22)	$7{\cdot}203 \times 10^{-2}$	$9{\cdot}339(33) \times 10^{-2}$	0·8606(48)	$8{\cdot}529 \times 10^{-2}$	$1{\cdot}486(46) \times 10^{-1}$

Table 10. Summary of quantitative solution of GRAM analysis: eigenvalues and their standard errors

	Quantity			'Simulated' standard addition (all errors $\times 10^{-4}$)				'Real' standard addition (all errors $\times 10^{-4}$)			
n	Unknown sample	Standard addition	Expected eigenvalue	π	σ_π (equation (47))	σ_π[a] (MC)	σ_π[b] (MC)	π	σ_π (equation (48))	σ_π[a] (MC)	σ_π[b] (MC)
1	1100	1000	0·52381	0·52376	2·48	2·46	2·44	0·52372	4·01	3·93	3·81
2	1000	1000	0·50000	0·49991	1·71	1·69	1·62	0·49986	2·73	2·69	2·59
3	900	1000	0·47368	0·47445	3·55	3·52	3·51	0·47484	5·54	5·51	5·40

[a] Monte Carlo estimate from 10^4 trials. [b] Monte Carlo estimate from 10^2 trials.

Table 11. Summary of qualitative solution of GRAM analysis: normalized inner products with input spectra and elution profiles

		'Simulated' standard addition		'Real' standard addition	
n	Identity	Spectrum	Elution profile	Spectrum	Elution profile
1	Adenine	0·99999	0·99995	0·99999	0·99986
2	Cytidine	0·99995	0·99997	0·99983	0·99994
3	Guanine	0·99999	0·99998	0·99998	0·99994

Preliminary results show that the derived standard errors are also accurate for the secondary PCs. This is a result of cancellation, because the eigenvalues constitute ratio estimates. For the applications discussed, only the expressions for GRAM seem to be useful for the secondary PCs.

CONCLUSIONS

From the observations in the preceding sections we draw the following conclusions. The first-order estimate of the standard error in the eigenvalue of a cross-product matrix is proportional to the square root of the modulus of that eigenvalue if only measurement noise is relevant. This estimate is very precise if the eigenvalues are well-separated, as is often the case for the significant PCs. Higher-order contributions describe the covariance between the eigenvalues. This leads to an underestimation of the variance of the eigenvalues of the non-significant PCs. It follows that the derived standard errors are only discriminative in a significance test if second-order error propagation is considered or if the standard errors for the non-significant factors are estimated by simulating random matrices. The proposed standard errors are useful for deriving standard errors for other multivariate problems. This has been demonstrated for the standard errors in the scores and loadings of PCA and the prediction errors for MLR and GRAM. Previously derived expressions for the standard errors in scores and loadings are shown to be equivalent. The same conclusion can be made with respect to recently derived prediction errors for MLR. Prediction errors have been derived for GRAM that explicitly show the contributions of all possible sources of random error. Furthermore, they indicate that adding the unknown and calibration data matrix ('simulated standard addition') leads to a substantial reduction of variance in comparison with real standard addition.

ACKNOWLEDGEMENTS

Acknowledgement is due to the referees for comments and criticisms which have led to a number of improvements to the material.

APPENDIX I: DERIVATION OF THE CONTRIBUTIONS OF THE EIGENVECTORS TO THE ERROR IN THE EIGENVALUES

The $c \times c$ identity matrix $\mathbf{I}_c$ is a constant matrix, so

$$d\mathbf{I}_c = \mathbf{0}_c \tag{49}$$

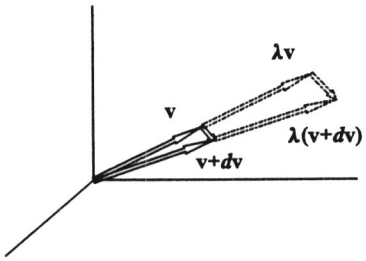


Figure 8. Transformation of approximate eigenvector $\mathbf{v} + d\mathbf{v}$. The error vector $d\mathbf{v}$ is (to first order) stretched by the same amount λ as the exact eigenvector $\mathbf{v}$

where $\mathbf{0}_c$ denotes the $c \times c$ null matrix. The eigenvectors are orthonormal:

$$\mathbf{V}^T\mathbf{V} = \mathbf{I}_c \tag{50}$$

Taking derivatives gives

$$d(\mathbf{V}^T\mathbf{V}) = d(\mathbf{V}^T)\tilde{\mathbf{V}} + \tilde{\mathbf{V}}^T d\mathbf{V} \tag{51}$$

Combining (49)–(51) gives

$$d(\mathbf{V}^T)\tilde{\mathbf{V}} + \tilde{\mathbf{V}}^T d\mathbf{V} = \mathbf{0}_c \tag{52}$$

So far, the derivation follows closely the reasoning in Appendices A and B of Reference 24, where the error in the pseudoinverse of an experimental matrix is expressed in terms of the error in the original matrix. Postmultiplication of (52) by the matrix of eigenvalues $\tilde{\mathbf{\Lambda}}$ yields

$$d(\mathbf{V}^T)\tilde{\mathbf{V}}\tilde{\mathbf{\Lambda}} + \tilde{\mathbf{V}}^T d\mathbf{V}\tilde{\mathbf{\Lambda}} = \mathbf{0}_c \tag{53}$$

When a small perturbation is applied to the eigenvectors, the following holds:

$$\tilde{\mathbf{D}}_R(\tilde{\mathbf{V}} + d\mathbf{V}) = (\tilde{\mathbf{V}} + d\mathbf{V})\tilde{\mathbf{\Lambda}} \tag{54}$$

This can easily be seen by interpreting matrix multiplication of a vector as a linear transformation. In the special case where a vector is an exact eigenvector, the transformation comes down to a scalar multiplication. The reasoning is still valid for a vector that constitutes a good approximation of the exact eigenvector. This situation is depicted in Figure 8, where the size of the error vector is exaggerated for visual clarity. By working out (54), one finds that

$$\tilde{\mathbf{D}}_R d\mathbf{V} = d\mathbf{V}\tilde{\mathbf{\Lambda}} \tag{55}$$

Inserting (55) in (53) finally results in

$$d(\mathbf{V}^T)\tilde{\mathbf{D}}_R\tilde{\mathbf{V}} + \tilde{\mathbf{V}}^T\tilde{\mathbf{D}}_R d\mathbf{V} = \mathbf{0}_c \tag{56}$$

APPENDIX II: DERIVATION OF EQUATION (10) FROM THE ERROR CONTRIBUTIONS OF THE SINGULAR VECTORS

Premultiplication of the SVD of $\mathbf{M}$ by $\mathbf{U}^T$ and postmultiplication by $\mathbf{V}$ gives

$$\mathbf{\Theta} = \mathbf{U}^T\mathbf{M}\mathbf{V} \tag{57}$$

The error matrix $d\boldsymbol{\Theta}$ can therefore be expanded as

$$d\boldsymbol{\Theta} = d(\mathbf{U}^{\mathrm{T}})\tilde{\mathbf{M}}\tilde{\mathbf{V}} + \tilde{\mathbf{U}}^{\mathrm{T}}d\mathbf{M}\tilde{\mathbf{V}} + \tilde{\mathbf{U}}^{\mathrm{T}}\tilde{\mathbf{M}}d\mathbf{V} \tag{58}$$

Again it is easily shown that the terms originating from errors in the singular vectors cancel. Equations (59)–(65) are similar to (50)–(56):

$$\mathbf{U}^{\mathrm{T}}\mathbf{U} = \mathbf{I}_c \tag{59}$$

$$d(\mathbf{U}^{\mathrm{T}}\mathbf{U}) = d(\mathbf{U}^{\mathrm{T}})\tilde{\mathbf{U}} + \tilde{\mathbf{U}}^{\mathrm{T}}d\mathbf{U} \tag{60}$$

$$d(\mathbf{U}^{\mathrm{T}})\tilde{\mathbf{U}} + \tilde{\mathbf{U}}^{\mathrm{T}}d\mathbf{U} = \mathbf{0}_c \tag{61}$$

$$d(\mathbf{U}^{\mathrm{T}})\tilde{\mathbf{U}}\tilde{\boldsymbol{\Theta}} + \tilde{\mathbf{U}}^{\mathrm{T}}d\mathbf{U}\tilde{\boldsymbol{\Theta}} = \mathbf{0}_c \tag{62}$$

$$\tilde{\mathbf{M}}(\tilde{\mathbf{V}} + d\mathbf{V}) = (\tilde{\mathbf{U}} + d\mathbf{U})\tilde{\boldsymbol{\Theta}} \tag{63}$$

$$\tilde{\mathbf{M}}d\mathbf{V} = d\mathbf{U}\tilde{\boldsymbol{\Theta}} \tag{64}$$

$$d(\mathbf{U}^{\mathrm{T}})\tilde{\mathbf{M}}\tilde{\mathbf{V}} + \tilde{\mathbf{U}}^{\mathrm{T}}\tilde{\mathbf{M}}d\mathbf{V} = \mathbf{0}_c \tag{65}$$

However, equation (65) is different from (56) because it gives a relation between the errors in the left and right singular vectors. The result is

$$d\boldsymbol{\Theta} = \tilde{\mathbf{U}}^{\mathrm{T}}d\mathbf{M}\tilde{\mathbf{V}} \tag{66}$$

The eigenvalues are the squares of the singular values, so

$$d\boldsymbol{\Lambda} = 2\tilde{\boldsymbol{\Theta}}d\boldsymbol{\Theta} \tag{67}$$

Inserting (66) in (67) gives equation (10).

Using the SVD of $\mathbf{M}$ rather than the EVD of $\mathbf{M}^{\mathrm{T}}\mathbf{M}$, the factor of two and the dependency of the error in the eigenvalues on the singular values of the pure data matrix arise in a very straightforward manner.

APPENDIX III: DERIVATION OF THE STANDARD ERRORS IN THE EIGENVALUES OF GRAM

Substituting $\boldsymbol{\Pi} = \boldsymbol{\Lambda}$, $\mathbf{W}^{\mathrm{T}} = \mathbf{V}^{\mathrm{T}}$ ($\mathbf{W}$ is the matrix of left eigenvectors), $\mathbf{U}_{\mathrm{Q}}^{\mathrm{T}}\mathbf{M}\mathbf{V}_{\mathrm{Q}}\boldsymbol{\Theta}_{\mathrm{Q}}^{-1} = \mathbf{D}_R$ and $\mathbf{Z} = \mathbf{V}$ in equation (7) and dropping the subscript Q for simplicity gives

$$d\boldsymbol{\Pi} = d(\mathbf{W}^{\mathrm{T}})\tilde{\mathbf{U}}^{\mathrm{T}}\tilde{\mathbf{M}}\tilde{\mathbf{V}}\tilde{\boldsymbol{\Theta}}^{-1}\tilde{\mathbf{Z}} + \tilde{\mathbf{W}}^{\mathrm{T}}d(\mathbf{U}^{\mathrm{T}}\mathbf{M}\mathbf{V}\boldsymbol{\Theta}^{-1})\tilde{\mathbf{Z}} + \tilde{\mathbf{W}}^{\mathrm{T}}\tilde{\mathbf{U}}^{\mathrm{T}}\tilde{\mathbf{M}}\tilde{\mathbf{V}}\tilde{\boldsymbol{\Theta}}^{-1}d\mathbf{Z} = \tilde{\mathbf{W}}^{\mathrm{T}}d(\mathbf{U}^{\mathrm{T}}\mathbf{M}\mathbf{V}\boldsymbol{\Theta}^{-1})\tilde{\mathbf{Z}} \tag{68}$$

Working out the remaining error term on the right-hand side of (68) in a straightforward manner as

$$d(\mathbf{U}^{\mathrm{T}}\mathbf{M}\mathbf{V}\boldsymbol{\Theta}^{-1}) = d(\mathbf{U}^{\mathrm{T}})\tilde{\mathbf{M}}\tilde{\mathbf{V}}\tilde{\boldsymbol{\Theta}}^{-1} + \tilde{\mathbf{U}}^{\mathrm{T}}d\mathbf{M}\tilde{\mathbf{V}}\tilde{\boldsymbol{\Theta}}^{-1} + \tilde{\mathbf{U}}^{\mathrm{T}}\tilde{\mathbf{M}}d\mathbf{V}\tilde{\boldsymbol{\Theta}}^{-1} + \tilde{\mathbf{U}}^{\mathrm{T}}\tilde{\mathbf{M}}\tilde{\mathbf{V}}d(\boldsymbol{\Theta}^{-1}) \tag{69}$$

and inserting the results obtained for $\mathbf{A} = \mathbf{U}\boldsymbol{\Theta}$ and $\mathbf{F} = \mathbf{V}^{\mathrm{T}}$ (see equation (31)) eventually leads to expressions for the standard errors in the eigenvalues. Although the results are close to the MC values, these expressions are not satisfying because of their inconsistent nature. The most notable inconsistency is the fact that the covariance matrix $\mathrm{COV}(\pi_n, \pi_m)$ is not exactly symmetrical and sometimes even leads to correlations slightly larger than unity. This means that the expressions for the standard errors in the loadings and scores are not efficient enough to be safely used in subsequent derivations. Since the error in the eigenvalues must come from the measurement errors in the data matrices $\mathbf{M}$ and $\mathbf{Q}$ – the error in the 'known' concentrations

comes in by applying error propagation to (46c) – it is natural to express the error in the eigenvalues as a sum of two independent contributions. This can be done by recognizing that the projection of **M** on to the row and column space of **Q** must leave **M** unchanged, since **Q** spans the space of **M** in an unbiased way. This procedure is known in the literature of rank annihilation as bilinear target testing:[44]

$$\mathbf{M} = \mathbf{U}\mathbf{U}^{\mathrm{T}}\mathbf{M}\mathbf{V}\mathbf{V}^{\mathrm{T}} = \mathbf{U}\mathbf{M}_{\mathrm{UV}}\mathbf{V}^{\mathrm{T}} \tag{70}$$

Here the substitution $\mathbf{M}_{\mathrm{UV}} = \mathbf{U}^{\mathrm{T}}\mathbf{M}\mathbf{V}$ has been made. (Wilson *et al.*[50] have used similar projections to derive an alternative algorithm for GRAM.) It follows that the error matrix on the left-hand side of (69) can be written as

$$d(\mathbf{U}^{\mathrm{T}}\mathbf{M}\mathbf{V}\boldsymbol{\Theta}^{-1}) = d(\mathbf{M}_{\mathrm{UV}}\boldsymbol{\Theta}^{-1}) = d(\mathbf{M}_{\mathrm{UV}})\tilde{\boldsymbol{\Theta}}^{-1} + \tilde{\mathbf{M}}_{\mathrm{UV}}d(\boldsymbol{\Theta}^{-1}) \tag{71}$$

Although, in contrast with the matrix $\mathbf{Q}_{\mathrm{UV}} = \mathbf{U}^{\mathrm{T}}\mathbf{Q}\mathbf{V} = \boldsymbol{\Theta}$, the matrix $\mathbf{M}_{\mathrm{UV}}$ and its errorless counterpart) is not diagonal, the arguments from the preceding appendices can still be used to show that

$$d(\mathbf{M}_{\mathrm{UV}}) = d(\mathbf{U}^{\mathrm{T}})\tilde{\mathbf{M}}\tilde{\mathbf{V}} + \tilde{\mathbf{U}}^{\mathrm{T}}d\mathbf{M}\tilde{\mathbf{V}} + \tilde{\mathbf{U}}^{\mathrm{T}}\tilde{\mathbf{M}}d\mathbf{V} = \tilde{\mathbf{U}}^{\mathrm{T}}d\mathbf{M}\tilde{\mathbf{V}} \tag{72}$$

Furthermore,

$$d(\boldsymbol{\Theta}^{-1}) = -d\boldsymbol{\Theta}\tilde{\boldsymbol{\Theta}}^{-2} = -\tilde{\boldsymbol{\Theta}}^{-1}d\boldsymbol{\Theta}\tilde{\boldsymbol{\Theta}}^{-1} \tag{73}$$

Combining (68) and (71)–(73) gives

$$d\boldsymbol{\Pi} = \tilde{\mathbf{W}}^{\mathrm{T}}(\tilde{\mathbf{U}}^{\mathrm{T}}d\mathbf{M}\tilde{\mathbf{V}}\tilde{\boldsymbol{\Theta}}^{-1} - \tilde{\mathbf{U}}^{\mathrm{T}}\tilde{\mathbf{M}}\tilde{\mathbf{V}}\tilde{\boldsymbol{\Theta}}^{-1}d\boldsymbol{\Theta}\tilde{\boldsymbol{\Theta}}^{-1})\tilde{\mathbf{Z}} \tag{74}$$

Working out (74) by inserting $\tilde{\mathbf{W}}^{\mathrm{T}}(\tilde{\mathbf{U}}^{\mathrm{T}}\tilde{\mathbf{M}}\tilde{\mathbf{V}}\tilde{\boldsymbol{\Theta}}^{-1}) = \tilde{\boldsymbol{\Pi}}\tilde{\mathbf{W}}^{\mathrm{T}}$ and $d\boldsymbol{\Theta} = \tilde{\mathbf{U}}^{\mathrm{T}}d\mathbf{Q}\tilde{\mathbf{V}}$, i.e. equation (66) applied to **Q**, leads to

$$d\boldsymbol{\Pi} = \tilde{\mathbf{W}}^{\mathrm{T}}\tilde{\mathbf{U}}^{\mathrm{T}}d\mathbf{M}\tilde{\mathbf{V}}\tilde{\boldsymbol{\Theta}}^{-1}\tilde{\mathbf{Z}} - \tilde{\boldsymbol{\Pi}}\tilde{\mathbf{W}}^{\mathrm{T}}\tilde{\mathbf{U}}^{\mathrm{T}}d\mathbf{Q}\tilde{\mathbf{V}}\tilde{\boldsymbol{\Theta}}^{-1}\tilde{\mathbf{Z}} \tag{75}$$

Now it is possible to substitute $\tilde{\mathbf{W}}^{\mathrm{T}}\tilde{\mathbf{U}}^{\mathrm{T}} = (\tilde{\mathbf{C}}_{\mathrm{M}} + \tilde{\mathbf{C}}_{\mathrm{N}})^{-1}\tilde{\mathbf{X}}^{+}$ and $\tilde{\mathbf{V}}\tilde{\boldsymbol{\Theta}}^{-1}\tilde{\mathbf{Z}} = (\tilde{\mathbf{Y}}^{\mathrm{T}})^{+}$ (see equation (46)). This will eventually lead to expressions that contain the physical decomposition of **Q**, i.e. equation (43). Expressing the error estimates in the responses of the individual components will help to identify which components contribute most to the error propagation and is therefore useful for optimization purposes. This line will be pursued in another publication, since here we want to concentrate on error estimates expressed in the abstract decomposition of **Q**. These expressions will show how the variance is built up by the subsequent addition of factors in (44) and can therefore be used to construct estimators that trade off variance for bias, a line very popular in multivariate calibration. Equation (75) is expressed in vector notation as

$$\vec{\sigma}_{\pi_n} = \sum_{q=1}^{F}\sum_{i=1}^{S}\sum_{k=1}^{W}\sum_{p=1}^{F}\tilde{W}_{qn}\tilde{U}_{iq}(\vec{\sigma}_{M_{ik}} - \tilde{\pi}_n\vec{\sigma}_{Q_{ik}})\tilde{V}_{kp}\tilde{\theta}_{Q,p}^{-1}\tilde{Z}_{pn} \tag{76}$$

The expression for the covariance matrix of the eigenvalues is rather complicated in the heteroscedastic case (uncorrelated noise). The expression for the homoscedastic case reveals more about the properties of the error estimates:

$$\mathrm{COV}(\pi_n, \pi_m) = |\vec{\sigma}_{\pi_n} \cdot \vec{\sigma}_{\pi_n}| = [\sigma_{\mathrm{M}}^2(1 - \tilde{\pi}_n - \tilde{\pi}_m) + \sigma_{\mathrm{Q}}^2\tilde{\pi}_n\tilde{\pi}_m]\sum_{q=1}^{F}\tilde{W}_{qn}\tilde{W}_{qm}\sum_{p=1}^{F}\left(\frac{\tilde{Z}_{pn}\tilde{Z}_{pm}}{\tilde{\lambda}_{Q,p}}\right) \tag{77}$$

The terms with $-\tilde{\pi}_n$ and $-\tilde{\pi}_m$ result from the fact that the noise in corresponding elements of **M** and **Q** is correlated, since **M** and **N** are artificially added in **Q**, so $\vec{\sigma}_{Q_{ik}} = \vec{\sigma}_{M_{ik}} + \vec{\sigma}_{N_{ik}}$.

Further simplification is possible by assuming that $\sigma_M = \sigma_N$. This assumption is reasonable if the data matrices are collected under identical experimental circumstances, since homoscedastic noise implies concentration independence. It follows that $\sigma_Q^2 = 2\sigma_M^2$ and consequently the standard errors are given by

$$\sigma_{\pi_n} = \sigma_M \left((1 - 2\tilde{\pi}_n + 2\tilde{\pi}_n^2) \sum_{q=1}^{F} \tilde{W}_{qn}^2 \sum_{p=1}^{F} \frac{(\tilde{Z}_{pn})^2}{\tilde{\lambda}_{Q,p}} \right)^{1/2} = \sigma_M \left((1 - 2\tilde{\pi}_n + 2\tilde{\pi}_n^2) \sum_{p=1}^{F} \frac{(\tilde{Z}_{pn})^2}{\tilde{\lambda}_{Q,p}} \right)^{1/2} \quad (78)$$

Here a final simplification resulted from normalizing the *left* eigenvector matrix $\tilde{\mathbf{W}}$. Consequently, the right eigenvectors needed for the evaluation of (78) must be calculated as $\tilde{\mathbf{Z}} = \tilde{\mathbf{W}}^{-T}$ ('inverse transpose'). This is not necessary for the reconstruction of $\mathbf{X}$ in (46a), since the columns of $\mathbf{X}$ are found by normalizing the matrix $\mathbf{U}_Q\mathbf{Z}$ anyway.

Using (78) derived for the 'simulated' standard additions, it is easy to derive the standard errors in the eigenvalues if the matrix $\mathbf{Q}$ is obtained by performing 'real' standard additions in the unknown sample. Now the cross-terms as well as the factor of two should vanish (assuming that $\sigma_Q = \sigma_M$) and the standard errors are consequently given by

$$\sigma_{\pi_n} = \sigma_M \left((1 + \tilde{\pi}_n^2) \sum_{p=1}^{F} \frac{(\tilde{Z}_{pn})^2}{\tilde{\lambda}_{Q,p}} \right)^{1/2} \quad (79)$$

Since the decomposition of $\mathbf{Q}$ and the resulting eigenvalue problem do not depend on the way $\mathbf{Q}$ has been constructed, a considerable reduction of variance may be achieved for the eigenvalues.

REFERENCES

1. Z. Z. Hugus Jr. and A. A. El-Awady, *J. Phys. Chem.* **75**, 2954 (1971).
2. J. T. Bulmer and H. F. Shurvell, *J. Phys. Chem.* **77**, 2085 (1973).
3. H. Bubert and H. Jenett, *Z. Anal. Chem.* **335**, 643 (1989).
4. J. L. Horn, *Psychometrika*, **30**, 179 (1965).
5. J. Mandel, *Technometrics*, **13**, 1 (1971).
6. J. F. Overland and R. W. Preisendorfer, *Mon. Weather Rev.* **110**, 1 (1982).
7. E. R. Malinowski, *J. Chemometrics*, **1**, 33 (1987).
8. S. J. Press, *Applied Multivariate Analysis*, Holt, Rinehart and Winston, New York (1971).
9. M. A. Girshick, *Ann. Math. Stat.* **10**, 203 (1939).
10. D. N. Lawley, *Biometrika*, **43**, 128 (1956).
11. E. R. Malinowski, *Anal. Chim. Acta*, **122**, 327 (1980).
12. B. A. Roscoe and P. K. Hopke, *Anal. Chim. Acta*, **132**, 89 (1980).
13. B. A. Roscoe and P. K. Hopke, *Anal. Chim. Acta*, **135**, 379 (1982).
14. E. Sánchez and B. R. Kowalski, *J. Chemometrics*, **2**, 247 (1988).
15. E. Sánchez and B. R. Kowalski, *J. Chemometrics*, **2**, 265 (1988)
16. J. E. Jackson and F. T. Hearne, *Technometrics*, **15**, 601 (1973).
17. W. J. Krzanowski, *J. Stat. Comput. Simul.* **18**, 299 (1983).
18. S. Wold, *Technometrics*, **20**, 397 (1978).
19. H. T. Eastment and W. J. Krzanowski, *Technometrics*, **24**, 73 (1982).
20. R. M. Wallace and S. M. Katz, *J. Phys. Chem.*, **68**, 3890 (1964).
21. B. Efron, *The Jackknife, the Bootstrap and Other Resampling Plans, Monograph 38*, SIAM, Philadelphia, PA (1982).
22. S. Brandt, *Statistical and Computational Methods in Data Analysis*, North-Holland, Amsterdam (1970).
23. M. G. Moran and B. R. Kowalski, *Anal. Chem.* **56**, 562 (1984).
24. G. Bauer, W. Wegscheider and H. M. Ortner, *Spectrochim. Acta B*, **46**, 1185 (1991).
25. S. D. Frans and J. H. Harris, *Anal. Chem.* **57**, 2680 (1985).
26. T. F. Brown and S. D. Brown, *Anal. Chem.* **53**, 1410 (1981).

27. R. F. Hirsch, G. L. Wu and P. C. Tway, *Chemometrics Intell. Lab. Syst.* **1**, 256 (1987).
28. E. R. Malinowski, *Anal. Chem.* **49**, 606 (1977).
29. E. A. Sylvestre, W. H. Lawton and M. S. Maggio, *Technometrics*, **16**, 353 (1974).
30. J. Öhman, P. Geladi and S. Wold, *J. Chemometrics*, **4**, 135 (1990).
31. W. O. McReynolds, *J. Chromatogr. Sci.* **8**, 685 (1970).
32. S. Wold and K. Andersson, *J. Chromatogr.* **80**, 43 (1973).
33. I. T. Joliffe, *Principal Component Analysis*, Springer, New York (1986).
34. E. R. Malinowski, *Anal. Chem.* **49**, 612 (1977).
35. J. H. Wilkinson, *The Algebraic Eigenvalue Problem*, Clarendon, Oxford (1965).
36. D. L. Duewer, B. R. Kowalski and J. L. Fasching, *Anal. Chem.* **48**, 2002 (1976).
37. A. Lorber and B. R. Kowalski, *J. Chemometrics*, **2**, 93 (1988).
38. A. Lorber and B. R. Kowalski, *Appl. Spectrosc.* **42**, 1572 (1988).
39. B. R. Kowalski and M. B. Seasholtz, *J. Chemometrics*, **5**, 129 (1991).
40. T. V. Karstang, J. Toft and O. M. Kvalheim, *J. Chemometrics*, **6**, 177 (1992).
41. B. E. Wilson, W. Lindberg and B. R. Kowalski, *J. Am. Chem. Soc.* **111**, 3797 (1989).
42. C.-N. Ho, G. D. Christian and E. R. Davidson, *Anal. Chem.* **50**, 1108 (1978).
43. A. Lorber, *Anal. Chim. Acta*, **164**, 293 (1984).
44. E. Sánchez and B. R. Kowalski, *Anal. Chem.* **58**, 496 (1986).
45. C.-N. Ho, G. D. Christian and E. R. Davidson, *Anal. Chem.* **52**, 1071 (1980).
46. E. R. Malinowski, *Factor Analysis in Chemistry*, Wiley, New York (1991).
47. F. P. Zscheile, H. C. Murray, G. A. Baker and R. G. Peddicord, *Anal. Chem.* **34**, 1776 (1962).
48. X. M. Tu, D. S. Burdick, D. W. Millican and L. B. McGown, *Anal. Chem.* **61**, 2219 (1989).
49. C. Shen, T. J. Vickers and C. K. Mann, *J. Chemometrics*, **5**, 417 (1991).
50. B. E. Wilson, E. Sanchez and B. R. Kowalski, *J. Chemometrics*, **3**, 493 (1989).

PART II PSEUDORANK ESTIMATION

Part II consists of two papers that are devoted to the subject of pseudorank estimation. The pseudorank of a matrix is defined as the mathematical rank in the absence of noise. Pseudorank estimation is a problem of major concern in multivariate data analysis. Sometimes it is even the only problem encountered. This is the case if, for example, the number of components under a chromatographic peak has to be determined (peak purity). Consequently it should not come as a surprise that many methods have been proposed in the past to tackle it. Section 2 discusses methods that are based on (functions of) the eigenvalues of principal component analysis of random matrices. The basic assumption behind these methods is the similarity of the secondary eigenvalues of the test matrix and the eigenvalues of a random matrix. (The random matrix plays the role of a reference matrix.) These methods are commonly denoted as parallel analysis. Several aspects have to be considered when using these methods because a number of approximations is implicitly made. In this section the importance of a number of aspects is studied by Monte Carlo simulations and a modification is proposed for methods that are based on the principle of parallel analysis. It is important to note that the methods discussed here can be used in the general situation that an estimate of the size of the measurement error is not available. Section 3 discusses methods that explicitly need this knowledge. They are therefore restricted in their use to favorable cases. A *t*-test is developed that determines whether a singular value of the data matrix is significantly larger than a reference value obtained from a random matrix. A comparison is made with Malinowski's *F*-test and it is found that the *t*-test leads to sharper significance levels. Since Malinowski's *F*-test does not need prior knowledge about the noise the conclusion is justified that the proposed *t*-test is capable of exploiting the extra knowledge. It is shown that a plot of the singular values constitutes a promising graphical pseudorank estimation method in the general case where the prior knowledge about the error is not available. Finally, it is important to note that the methods discussed in Section 2 can also provide a significance level if the appropriate simulations are carried out. Thus the real advantage of the *t*-test introduced in Section 3 is the fact that it supplies a significance level in the formal context of statistical hypothesis testing. This approach should be more acceptable for the practical worker.

References 3, 5 and 13 in Section 2 correspond to Sections 5, 3 and 1, respectively. References 15 and 16 in Section 3 correspond to Sections 2 and 1, respectively.

ASPECTS OF PSEUDORANK ESTIMATION METHODS BASED ON THE EIGENVALUES OF PRINCIPAL COMPONENT ANALYSIS OF RANDOM MATRICES

N.M. FABER, L.M.C. BUYDENS and G. KATEMAN

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

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.

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 is 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 can not 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 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 is 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.

Is 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 ourself to the effect of the measurement error. For second-order data like the popular spectro-chromatograms 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 adapted throughout this paper. Bold upper-case letters will denote matrices, e.g. **M**. Bold lower-case letters will denote column vectors, e.g. **v**. Matrix and vector transposition are indicated by a superior 'T', e.g. $\mathbf{M}^T$ and $\mathbf{v}^T$. 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. 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.

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.

(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 **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:

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

Since $\mathbf{M}^{\mathrm{T}}\mathbf{M}$ is symmetric, the columns of $\mathbf{V}$ are orthogonal eigenvectors in $\mathbb{R}^J$ and the diagonal elements of $\boldsymbol{\Lambda}$, the eigenvalues λ_a, are real numbers arranged in nonincreasing order. Furthermore, in the presence of noise, $\mathbf{M}^{\mathrm{T}}\mathbf{M}$ 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:

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

where $Tr[\bullet]$ 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 $\mathbf{v}_a$. 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:

$$\|\mathbf{M}^{\mathrm{T}}\mathbf{M}\mathbf{v}_a\|_E = \|\mathbf{z}_a\|_E = \lambda_a \tag{3}$$

where $\| \bullet \|_E$ represents the Euclidean vector norm and $\mathbf{z}_a$ is the converged iterate. Since the eigenvalue λ_a is found from the Euclidean vector norm of $\mathbf{z}_a$, some noise averaging will take place. Thus λ_a is expected to be more precise than the individual elements of $\mathbf{v}_a$, the normalized converged iterate. The effect should be particularly notable if $J >> 1$ (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 Θ, the singular values θ_a, are (by convention) the positive square roots of the eigenvalues λ_a. Using the SVD, the elementwise 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 equation (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 $\mathbf{v}_a$ associated with a small singular value θ_a. This is immediate from

$$\|\mathbf{M}\mathbf{v}_a\|_E = \|\theta_a\mathbf{u}_a\|_E = \theta_a \tag{6}$$

The vector $\mathbf{M}\,\mathbf{v}_a$ 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) Number of degrees of freedom

According to equation (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 $(I\text{-}A)$ $(J\text{-}A)$ 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:

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

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 'results and discussion' section.) Note that this number is used in cross-validation for the primary as well as the secondary PCs [23].* (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 $(J\text{-}A)$. Several derivations can be found in the literature. A simple proof is given by Paatero and Tapper [12]: the I x J pseudorank A data matrix is reproduced by the product of the I x A score matrix $\mathbf{S} = \mathbf{U}\,\mathbf{\Theta}$, and the A x J loading matrix $\mathbf{L} = \mathbf{V}^T$. This reproduction is fixed up to an A x A transformation matrix. As a result one finds for the number of free variables for the A-dimensional PC model: I x J - I x A - A x J + A x A = $(I\text{-}A)$ $(J\text{-}A)$. 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 equation (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

---

*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 crossvalidation 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 crossvalidation the PCs are examined in decreasing order of importance, while the methods discussed in this paper proceed *backwards* through the list of PCs.

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

where $E[\bullet]$ 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 equation (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 equation (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 crossvalidation 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 equation (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.)

(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 $1 \leq a \leq A$ 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 chi squared distribution with $(I\text{-}A)$ $(J\text{-}A)$ degrees of freedom if τ approaches ∞. Thus in the limiting case the number of degrees of freedom for the residuals previously given as $(I\text{-}A)$ $(J\text{-}A)$ becomes essentially correct.

Immediately the question rises 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 x J pseudorank A test data matrix can adequately be approximated by the eigenvalues of an $(I\text{-}A)$ $(J\text{-}A)$ 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 like equation (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 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.

(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. $I \rightarrow \infty$, $J \rightarrow \infty$ and $d = I / J =$ constant). It was shown that the probability density function, $f(\lambda)$, only depends on the divergence coefficient d:

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

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 Figure 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 equation (9) because this expression is much simpler to interpret than the expression obtained from the normal assumption.

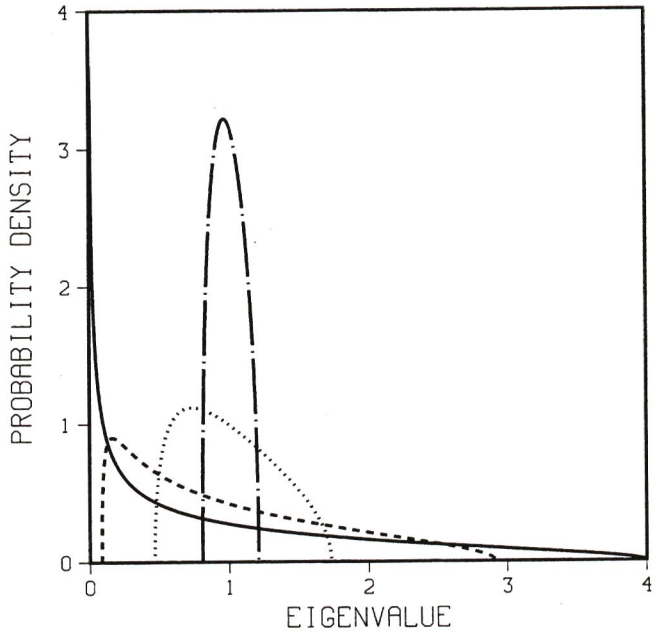


Figure 1. Distribution of eigenvalues for divergence coefficient d = 1 (——), 2 (----), 10 (.....) and 100 (.——.)

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.* Still, it is interesting to introduce this concept because the postulated properties of the logarithm of the eigenvalues and reduced 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 spectro-chromatograms 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.

(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 x 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 x 31). Kormos and Waugh [32] applied the method to simulated data with varying signal-to-noise ratio (matrix size 520 x 10 and 600 x 10) and real data (matrix size 195 x 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.

(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 $(I\text{-}a+1)$ $(J\text{-}a+1)$, where a is the number of the extracted PC. A pseudorank estimation method was developed by constructing reduced eigenvalues as

---

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

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

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 equation (8) is established as follows. Since the reduced eigenvalues should be constant for the secondary PCs, the denominator in equation (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.

(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].

(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.

(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 x J test data matrix one should generate reference matrices of size $(I-a)$ x $(J-a)$ 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 equations (7) and (8) yield two independent estimates for the standard deviation of the measurement error. It is seen that the estimate provided by equation (7) is based entirely on the test data while equation (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 equation (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.

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.

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 x 10, 20 x 20, 40 x 20 and 40 x 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 $\sqrt{3} \approx 1.732$. 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 10000 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 10000 matrices is a factor 100 smaller.

Literature data matrix

The data matrix taken from the literature consists of simulated mass spectra [36]. The size of the matrix is 20 x 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 $\sigma_M = 1/6\sqrt{3} \approx 0.289$. However, the residuals left after extracting 5 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 $(I\text{-}A)$ $(J\text{-}A) = 75$ according to equation (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.

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

Table 1. Eigenvalues of a 20x10 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	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)

standard error in the eigenvalues resulting from noise.*

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 x 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 equation (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.** (It is worth mentioning that the real error function underestimates $\hat{\sigma}_M$ by a factor $\sqrt{75/100} \approx 0.866$.)

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

*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.

**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].

Table 2. Standard errors in the first eigenvalue of a 20x10 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

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

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

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 ρ = 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 ρ = 0.5 to ρ = 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 ρ = 0.5, the level for which the approximation starts to work well, one finds σ_M/θ_1 = 0.11, which is a reasonable value since $0.11 << 1$. 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 ρ = 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 x 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 49.3 - 45.4 = 3.9. Again, this difference should be compared with the precision of the eigenvalue that is to be scrutinized for significance.

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]:

$$\sigma_\lambda = 2\lambda^{1/2}\sigma_M \tag{13}$$

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

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}
10x10	32.14	22.47	16.18	11.47	7.83	5.06	2.94	1.45	0.50	0.07
11x10	33.86	24.07	17.62	12.78	8.97	5.97	3.68	1.99	0.84	0.20
12x10	35.66	25.66	19.00	13.96	10.05	6.90	4.42	2.57	1.22	0.37
13x10	37.48	27.31	20.51	15.27	11.17	7.87	5.22	3.16	1.63	0.58
14x10	39.28	28.86	21.89	16.53	12.21	8.75	5.96	3.76	2.06	0.84
15x10	41.02	30.39	23.24	17.72	13.32	9.70	6.77	4.38	2.52	1.10
16x10	42.67	31.91	24.57	18.98	14.39	10.61	7.55	5.02	2.99	1.41
17x10	44.26	33.37	25.88	20.12	15.42	11.54	8.34	5.68	3.50	1.72
18x10	46.11	34.93	27.24	21.37	16.51	12.53	9.17	6.33	4.01	2.07
19x10	47.68	36.33	28.62	22.55	17.56	13.42	9.91	6.97	4.51	2.41
20x10	49.22	37.74	29.85	23.71	18.64	14.38	10.77	7.72	5.09	2.82
30x10	64.62	51.73	42.69	35.40	29.24	23.89	19.19	14.96	11.09	7.37
40x10	79.27	65.01	54.99	46.77	39.65	33.43	27.81	22.65	17.74	12.81
50x10	93.25	78.12	67.06	58.03	50.11	43.00	36.60	30.57	24.80	18.77
60x10	106.8	90.47	78.70	68.90	60.38	52.60	45.46	38.77	32.16	25.16
70x10	120.3	103.1	90.64	80.09	70.86	62.41	54.67	47.18	39.77	31.85
80x10	133.1	115.1	101.9	90.82	81.00	71.98	63.59	55.51	47.46	38.67
90x10	146.3	127.3	113.4	101.7	91.31	81.76	72.74	64.10	55.42	45.82
100x10	158.9	139.2	124.8	112.6	101.5	91.48	81.95	72.81	63.49	53.11

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 x 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. Figure 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 (I-$a+1$) (J-$a+1$) and λ_a/σ_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 x 10, 20 x 20, 40 x 20 and 40 x 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 x 10 and 100 x 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 x 10 and 20 x 20 (Fig.2a,b) with the plots obtained after doubling the matrix size (Fig.2c,d) shows that the

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}
10x10	6.14	4.11	3.12	2.44	1.86	1.43	1.02	0.68	0.35	0.11
11x10	6.22	4.23	3.19	2.53	2.00	1.54	1.16	0.79	0.47	0.20
12x10	6.40	4.34	3.30	2.63	2.10	1.66	1.26	0.90	0.59	0.30
13x10	6.53	4.48	3.44	2.72	2.20	1.76	1.38	1.01	0.70	0.39
14x10	6.56	4.55	3.52	2.85	2.33	1.88	1.47	1.12	0.80	0.49
15x10	6.90	4.66	3.70	2.98	2.44	1.98	1.58	1.22	0.90	0.59
16x10	6.88	4.71	3.76	3.07	2.52	2.03	1.66	1.31	0.99	0.67
17x10	6.83	4.82	3.81	3.11	2.58	2.14	1.74	1.41	1.09	0.77
18x10	7.10	4.94	3.90	3.18	2.67	2.23	1.84	1.50	1.18	0.86
19x10	7.11	5.00	3.99	3.32	2.76	2.30	1.91	1.56	1.26	0.95
20x10	7.17	5.10	4.07	3.33	2.82	2.38	1.99	1.66	1.35	1.05
30x10	8.08	5.88	4.79	4.08	3.50	3.05	2.71	2.36	2.08	1.85
40x10	9.01	6.51	5.43	4.65	4.10	3.69	3.29	2.98	2.71	2.57
50x10	9.50	7.14	5.92	5.17	4.60	4.13	3.81	3.48	3.26	3.19
60x10	10.2	7.50	6.32	5.53	5.02	4.57	4.23	3.94	3.70	3.77
70x10	10.5	8.05	6.81	6.01	5.41	4.93	4.62	4.41	4.25	4.31
80x10	11.0	8.29	7.12	6.24	5.80	5.41	4.99	4.74	4.63	4.77
90x10	11.7	8.79	7.54	6.80	6.17	5.68	5.34	5.18	5.04	5.30
100x10	11.9	9.13	7.93	7.09	6.42	6.14	5.72	5.53	5.38	5.68

agreement does not seem to improve by going to larger matrix sizes.

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 Figures 3 and 4. In Figures 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.* 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 Figure 3 if the number of primary PCs is not too large.

The reduced eigenvalues displayed in Figure 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 behavior is to be expected from the varying shapes of the distribution functions displayed in Figure 1.

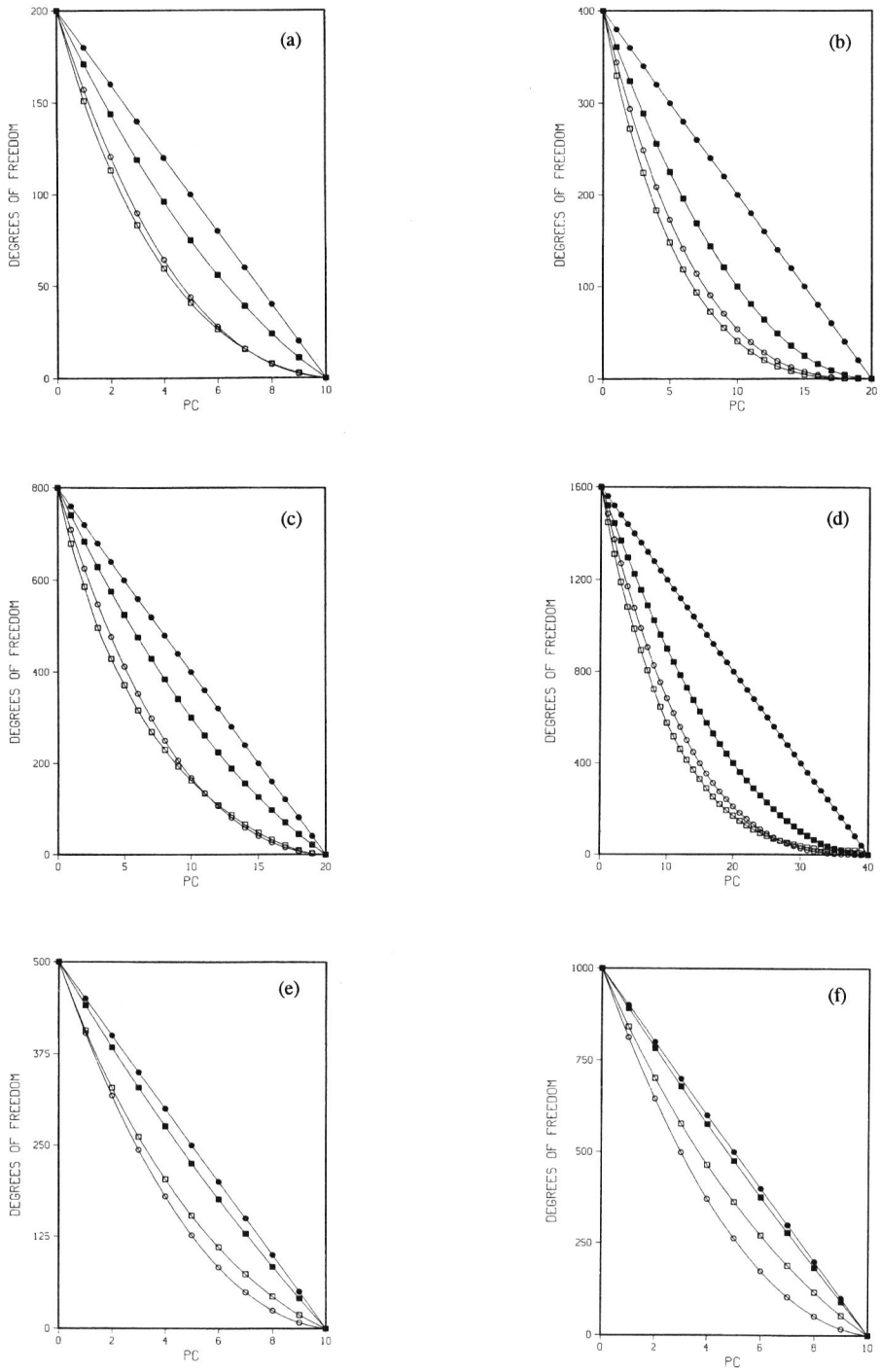


Figure 2. Number of degrees of freedom left after extracting a PCs for (a) 20x10, (b) 20x20, (c) 40x20, (d) 40x40, (e) 50x10 and (f) 100x10 matrix. The number of degrees of freedom for the ath PC is estimated by I (●), $I+J-2a+1$ (■), N $(I-a+1)$ $(J-a+1)$ (○) and λ_a/σ_M^2 (□)

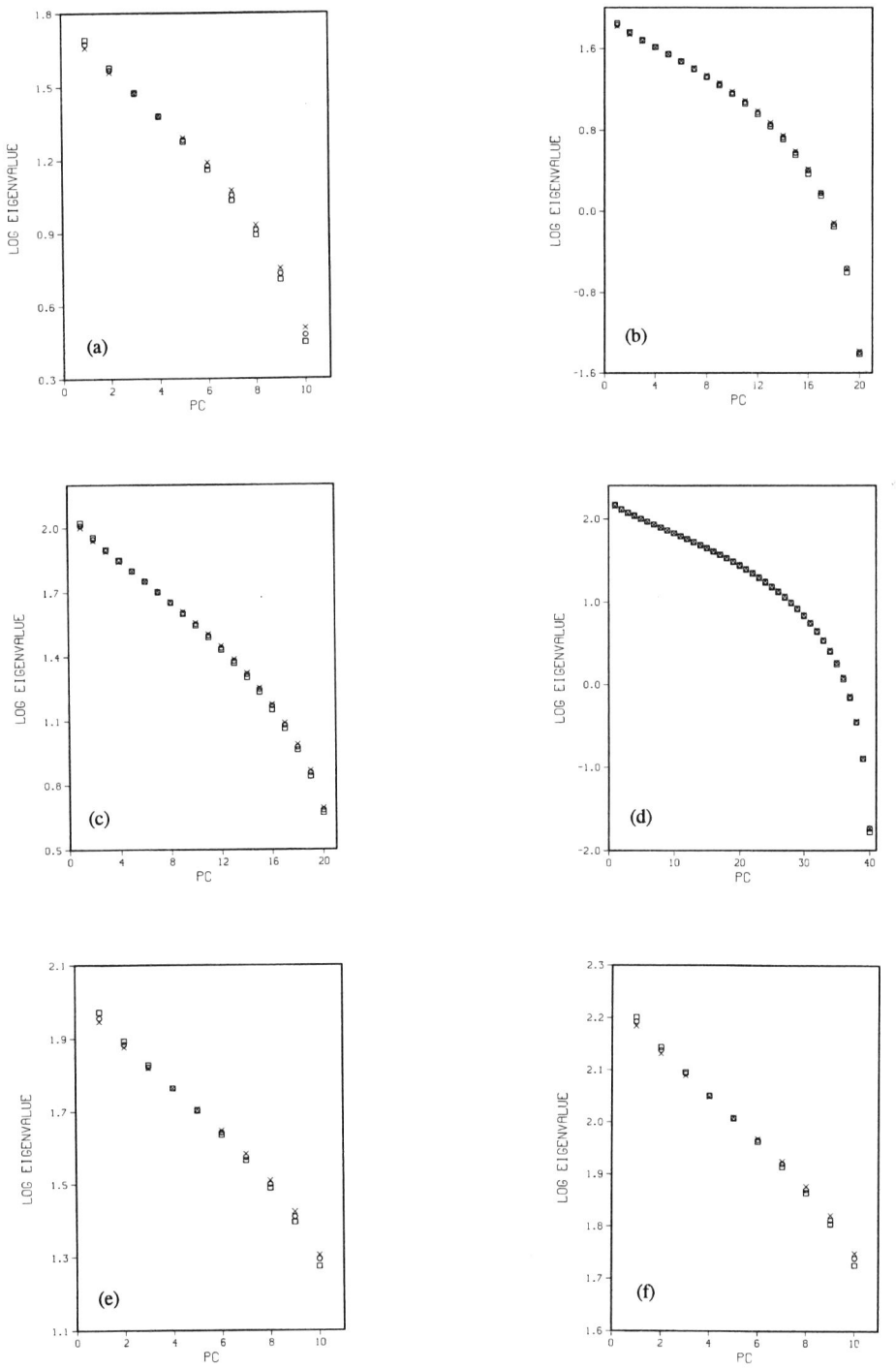


Figure 3. Logarithm of eigenvalues for a) 20x10, (b) 20x20, (c) 40x20, (d) 40x40, (e) 50x10 and (f) 100x10 random matrix with normal (□), uniform (○) and random sign (×) distribution.

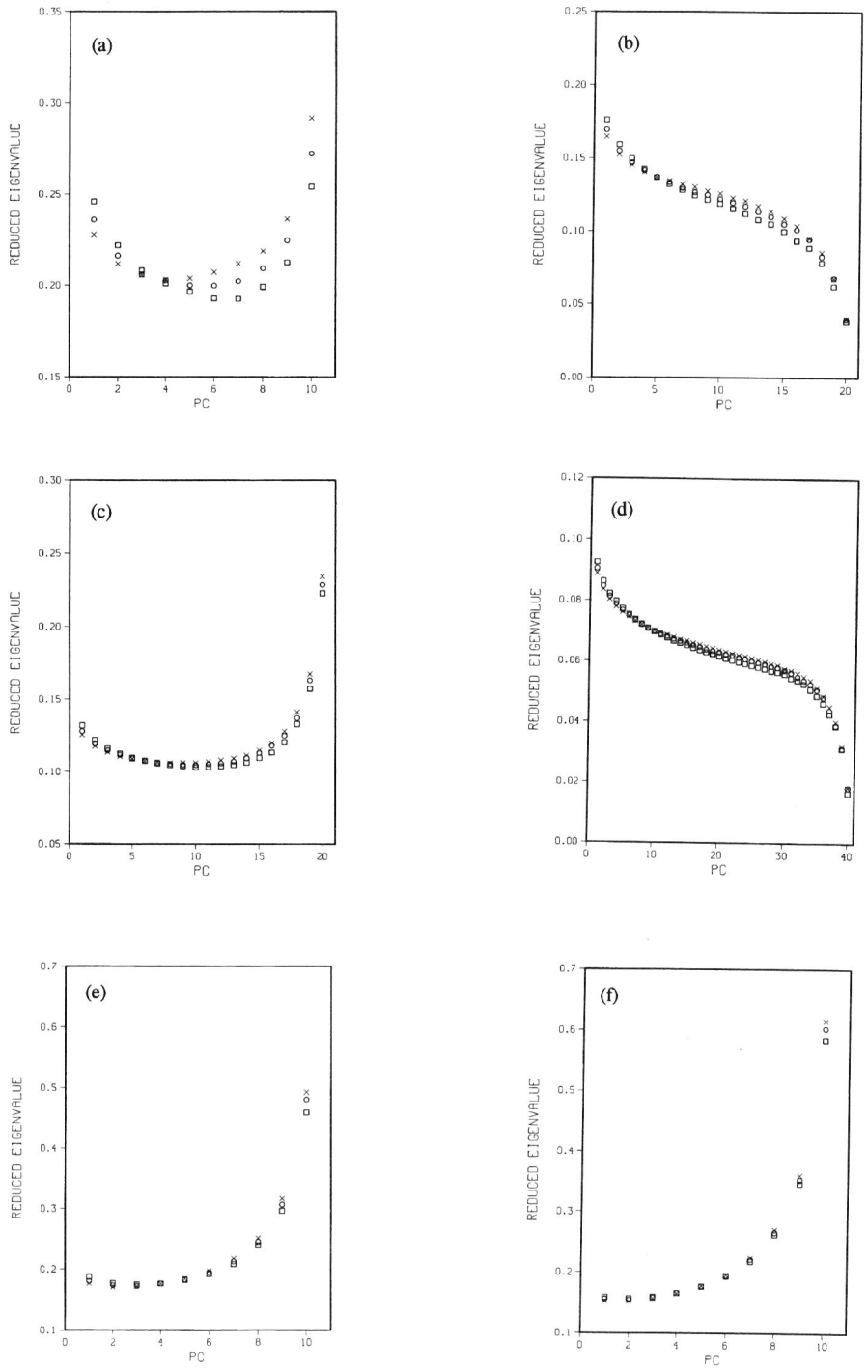


Figure 4. Reduced eigenvalues for a) 20x10, (b) 20x20, (c) 40x20, (d) 40x40, (e) 50x10 and (f) 100x10 random matrix with normal (□), uniform (○) and random sign (×) distribution.

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}
10x10	1.90	1.43	**0.99**	**0.59**	**0.25**	**-0.04**	**-0.33**	**-0.58**	**-0.90**	-1.26
11x10	1.82	1.33	**0.90**	**0.51**	**0.18**	**-0.12**	**-0.36**	**-0.60**	**-0.82**	-1.06
12x10	1.72	1.23	**0.78**	**0.39**	**0.09**	**-0.17**	**-0.41**	**-0.57**	**-0.72**	**-0.82**
13x10	1.63	1.12	**0.68**	**0.30**	**0.01**	**-0.23**	**-0.41**	**-0.55**	**-0.62**	**-0.63**
14x10	1.57	1.02	**0.57**	**0.20**	**-0.10**	**-0.31**	**-0.46**	**-0.52**	**-0.54**	**-0.41**
15x10	1.43	**0.90**	**0.44**	**0.09**	**-0.17**	**-0.35**	**-0.45**	**-0.50**	**-0.44**	**-0.24**
16x10	1.36	**0.79**	**0.32**	**-0.00**	**-0.25**	**-0.43**	**-0.48**	**-0.47**	**-0.35**	**-0.08**
17x10	1.27	**0.67**	**0.20**	**-0.12**	**-0.35**	**-0.47**	**-0.50**	**-0.42**	**-0.25**	**0.07**
18x10	1.16	**0.56**	**0.08**	**-0.23**	**-0.43**	**-0.51**	**-0.50**	**-0.41**	**-0.17**	**0.20**
19x10	1.08	**0.45**	**0.00**	**-0.31**	**-0.50**	**-0.57**	**-0.54**	**-0.39**	**-0.10**	**0.32**
20x10	**0.96**	**0.30**	**-0.16**	**-0.45**	**-0.60**	**-0.63**	**-0.55**	**-0.32**	**0.00**	**0.47**
30x10	**-0.13**	**-0.92**	-1.33	-1.47	-1.41	-1.14	**-0.68**	**-0.06**	**0.70**	1.50
40x10	-1.16	-2.11	-2.43	-2.44	-2.15	-1.58	**-0.82**	**0.15**	1.25	2.28
50x10	-2.21	-3.18	-3.49	-3.32	-2.82	-2.04	**-0.95**	**0.31**	1.72	2.95
60x10	-3.16	-4.35	-4.57	-4.27	-3.49	-2.44	-1.09	**0.48**	2.18	3.53
70x10	-4.23	-5.34	-5.48	-5.03	-4.12	-2.85	-1.20	**0.62**	2.50	4.06
80x10	-5.14	-6.46	-6.48	-5.94	-4.69	-3.14	-1.33	**0.75**	2.87	4.58
90x10	-5.96	-7.36	-7.33	-6.51	-5.24	-3.54	-1.46	**0.86**	3.20	4.99
100x10	-6.93	-8.34	-8.16	-7.27	-5.87	-3.80	-1.56	**0.97**	3.53	5.48

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 (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 x 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.

Table 6. Eigenvalues and reduced eigenvalues for literature data matrix

PC	*EV*	*REV*[a]	*REV*[b] (20x10)[c]	*REV*[b] (15x5)[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

Literature data matrix

From the results of the simulations it has become clear that the size of the literature data matrix (20 x 10) 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 5-dimensional PC model. According to equation (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 Figure 5. It is important to note that the choice of a 5-dimensional model from the log-eigenvalue diagram can not be based on a straight part in the plot since there are too many primary PCs. Instead, this conclusion should now 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] 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 x 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 equation (8). The value found is $\hat{\sigma}_M = 0.786$. This value is not close to the value estimated from the residuals using equation (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' can not 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) x (10-5). The 'reduced eigenvalues' calculated from the eigenvalues of 15 x 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 calculated from equation (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.

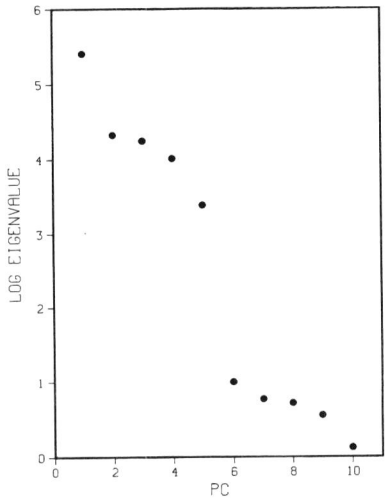


Figure 5. Logarithm of eigenvalues for literature data matrix

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 5-dimensional PC model is shown in Figure 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 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].

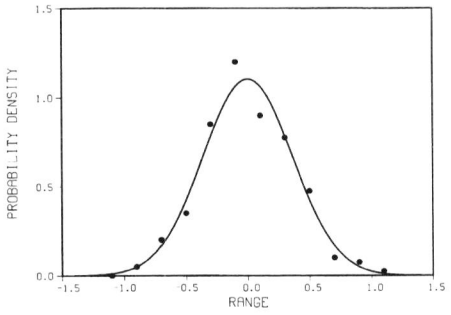


Figure 6. Distribution of residuals after extracting five PCs for literature data matrix. The dots represent normalized frequency counts based on a binsize of 0.2. The normal distribution function with the same mean and standard deviation is drawn as guide to the eye

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 x 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 ($d >> 1$) 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 equations (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*, in press.

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

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

6 J.E. Jackson, *A User's Guide to Principal Components*, John 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, 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 *Chemometrics, Theory and Application* (B.R. Kowalski, Ed), American Chemistry Society Symposium Series, No. 52 (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*, John 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.

ASPECTS OF PSEUDORANK ESTIMATION METHODS BASED ON AN ESTIMATE OF THE SIZE OF THE MEASUREMENT ERROR

N.M. FABER, L.M.C. BUYDENS and G. KATEMAN

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

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. Method A and B are modifications of methods that are well known in analytical chemistry. However, these methods can not 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.

KEY WORDS Principal component analysis Pseudorank estimation

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. $\mathbf{M}$. Bold lower-case letters will denote column vectors, e.g. $\mathbf{u}$. 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 $\mathbf{M}$. 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.

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.

(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 neighborhood 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) 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 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 [22]. Essentially the same approach can be followed for chromatographic data. Here the zero-component regions should provide the necessary information [11].

(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

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

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, ½, ¼, ⅛, ... 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 equation (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 x n leading principal submatrix of **M'** and **E'** respectively. The pseudo rank of **M'** should be at least n if $\tau_n < 1$.

(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:

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

where **U** is an r x s orthogonal matrix whose columns $\mathbf{u}_n$ are the left singular vectors, **V** is an s x c orthogonal matrix whose columns $\mathbf{v}_n$ are the right singular vectors and Θ is an s x 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 $\mathbf{M}\,\mathbf{M}^T$ and $\mathbf{M}^T\,\mathbf{M}$:

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

The singular vectors $\mathbf{u}_n$ and $\mathbf{v}_n$ are seen to be eigenvectors of the cross-product matrices. The eigenvalue decompositions of $\mathbf{M}\,\mathbf{M}^T$ and $\mathbf{M}^T\,\mathbf{M}$ 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):

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

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

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

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 *n*th PC from the test data is $(r\text{-}n)$ $(c\text{-}n)$. Since we need the number of degrees of freedom *before* extracting the *n*th PC, the number of degrees of freedom of the reference matrix is therefore found to be $(r\text{-}n+1)$ $(c\text{-}n+1)$. 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 $(r\text{-}n+1)$ x $(c\text{-}n+1)$. 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 x c random matrix, the second eigenvalue of the test data matrix is corrected by subtracting the first eigenvalue of an $(r\text{-}1)$ x $(c\text{-}1)$ random matrix, and so on. In general, the correction for the *n*th eigenvalue under scrutiny is found as the dominant eigenvalue of an $(r\text{-}n+1)$ x $(c\text{-}n+1)$ 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

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

where $\lambda_{n,\text{ref}}$ denotes the *n*th 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 $\lambda_{n,\text{ref}}$ are negligible since we do not apply a correction in Method A.

In a previous publication we showed that the standard errors predicted by Equation (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 Equation (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.

(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 Method A and B because their sampling distribution is unknown. For Method B this is caused by the fact that the numerator and denominator in Equation (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* 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 Equation (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.** 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 [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 Equation (9) by the corresponding singular values. It is easily shown that the resulting statistic is always larger. However, the Equation (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 Equation (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. $\nu = (r\text{-}n+1)\,(c\text{-}n+1)$. The ratio of Equation (10) is designed to test the null-hypothesis

$$H_0 : \theta_n = \theta_{n,\mathrm{ref}}$$

against the alternative hypothesis (a one-sided test)

$$H_1 : \theta_n > \theta_{n,\mathrm{ref}}$$

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 $t_\nu(1\text{-}\alpha)$. 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

---

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

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.

singular value of an $(r\text{-}s+1)$ x 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 $n = s\text{-}1$, and so 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 Equation (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.

(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 $\nu_2 = s\text{-}n$. The procedure consists of testing λ_n against the pool of $(s\text{-}n)$ 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].

EXPERIMENTAL SECTION

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

(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 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 x 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 unfavorable 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

Table 1. Characterization of literature test data

Data	FIAL80[a]	GUTM68[b]	HAVE85[c]	MALI82[d]	RITT76[e]	WEIN70[f]	WEIN71[g]
Size	20x5	9x5	10x8	20x10	17x7	14x9	22x6
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.

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 $C_{ik} = h_k \cdot \exp[-½ (i-\mu_k)^2 / \sigma_k^2]$ where the symbols have the meaning as indicated above.

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 $\sigma(\mathbf{M})$.

(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 equation (8) was tested. A three-component HPLC-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 36x36. Artificial Gaussian noise with standard deviation 0.5mAU 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 Figure 1. The unfavorable ratio of peakheights and high overlap of the pure component responses (in both instrumental modes) are apparent. These data 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) 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 $\%\alpha = 100\times(1\text{-DFDF}(F,\nu_1,\nu_2))$. 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 $\%\alpha = 100\times(1\text{-DTDF}(t,\nu))$.

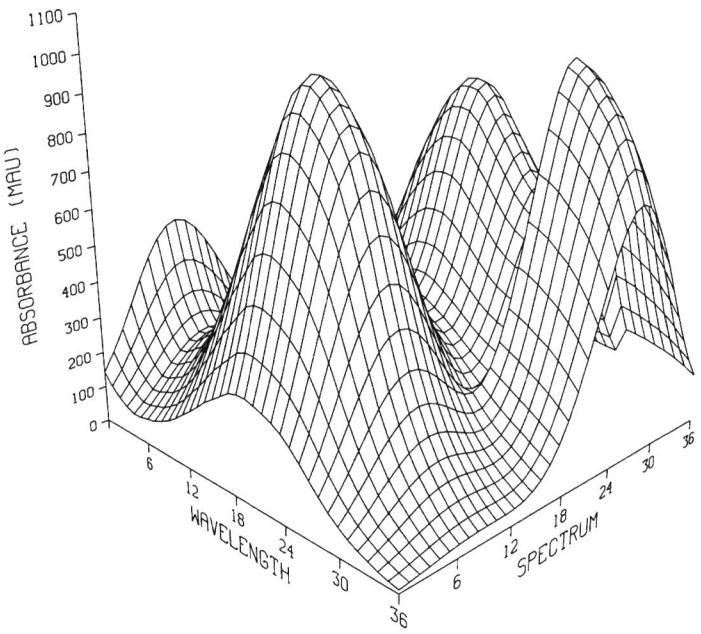


Figure 1. Simulated three-component HPLC-UV data matrix EXP1

RESULTS AND DISCUSSION

(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.

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.* 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 Equation (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 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.

---

*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 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}	9.67x10	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}	**9.55x10**	1.13x10
4	3.53	1.51×10^{-2}	3.30×10^{-3}	1.02×10^{4}	3.35×10^{-1}	1.16x10	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}	1.00x10	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 ($x10^{-2}$)	HAVE85 ($x10^{-3}$)	MALI82	RITT76 ($x10^{-2}$)	WEIN70 ($x10^{-1}$)	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				-			

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	$2.60\text{x}10^{3}$	4.20	1.21	$1.28\text{x}10^{3}$	2.05x10	$8.13\text{x}10^{4}$	$3.71\text{x}10^{3}$
2	$2.45\text{x}10^{2}$	**$7.43\text{x}10^{-3}$**	$7.32\text{x}10^{-3}$	$1.24\text{x}10^{2}$	**3.48**	4.59	**5.27x10**
3	**4.69x10**	$1.51\text{x}10^{-3}$	**$8.54\text{x}10^{-4}$**	$1.23\text{x}10^{2}$	$1.01\text{x}10^{-2}$	**1.14**	0.14
4	0.10	$1.26\text{x}10^{-3}$	$9.42\text{x}10^{-5}$	$8.60\text{x}10^{1}$	$0.60\text{x}10^{-2}$	0.18	0.15
5	0.21	$0.61\text{x}10^{-3}$	$7.80\text{x}10^{-5}$	**$2.52\text{x}10^{1}$**	$0.66\text{x}10^{-2}$	0.18	0.10
6			$2.27\text{x}10^{-5}$	0.13	$0.50\text{x}10^{-2}$	0.11	0.09
7			$3.72\text{x}10^{-5}$	0.11	$0.52\text{x}10^{-2}$	0.07	
8			$0.04\text{x}10^{-5}$	0.13		0.10	
9				0.15		0.12	
10				0.12			

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	$5.10\text{x}10^{2}$	1.374x10	9.835	$5.06\text{x}10^{2}$	4.93x10	$3.20\text{x}10^{3}$	$7.00\text{x}10^{2}$
2	$1.37\text{x}10^{2}$	**0.488**	0.679	$1.46\text{x}10^{2}$	**1.83x10**	2.18x10	**7.44x10**
3	**5.03x10**	0.178	**0.202**	$1.33\text{x}10^{2}$	0.87	**9.77**	3.35
4	1.88	0.123	0.057	$1.01\text{x}10^{2}$	0.58	3.41	2.92
5	1.85	0.055	0.043	**$4.92\text{x}10^{1}$**	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			

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).* 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. $(r\text{-}n^*)$ $(c\text{-}n^*)$ = (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 Method 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 (0.5Hz) which is 30% smaller that the standard deviation estimated from the residuals of the 3-dimensional PC model (0.65Hz). 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 Equation (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 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 Equation (8) is given in the next column. The resulting ratio calculated from Equation (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 Equation (3) and (9) respectively. It is seen that the agreement is very well in all cases. This result is remarkable since the evaluation of Equation (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 *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

---

*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.

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]	$\rho(\epsilon)$	λ	λ_{ref}[a]	$\sigma(\lambda)$	$\rho(\lambda)$
1	456	0.50	911	10200000	9.28	3200	3200
2	14.9	0.55	27.1	4770	8.34	21.8	21.5
3	7.11	0.83	8.61	95.5	7.38	9.77	**9.02**
4	-2.99	0.98	**3.04**	11.6	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	$\%\alpha$	*REV*	F	ν_1	ν_2	$\%\alpha$
1	3200	3.04	0.26	4520	126	0.0	81300	52000	1	8	0.0
2	21.8	2.88	0.27	26.8	104	0.0	4.59	10.4	1	7	1.4
3	9.77	2.70	0.27	10.0	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 $\hat{\sigma}(\mathbf{M}) = 0.50$.

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

Data	MethodA	MethodB	*t*-test			*F*-test		
	$\rho(\epsilon)$	$\rho(\lambda)$	t	ν	$\%\alpha$	F	ν_2	$\%\alpha$
FIAL80	40.3	55.4	74.9	54	0.0	337	2	0.3
GUTM68	4.37	4.18	4.24	32	0.09	5.66	3	9.8
HAVE85	7.19	7.19	7.89	48	0.0	12.5	5	1.7
MALI82	41.6	41.6	55.1	96	0.0	199	5	0.003
RITT76	60.1	63.8	122	96	0.0	466	5	0.0004
WEIN70	6.58	6.52	6.79	84	0.0	7.96	6	3.0
WEIN71	62.6	66.4	89.8	105	0.0	400	4	0.004

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

(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.

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.

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 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 favorable 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.

(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 Figure 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.* 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

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	717	4.22	3.29	2.84	210	39.5	18.2
GUTM68	7.51	4.20	3.09	2.59	45.2	13.4	7.35
HAVE85	12.4	3.19	2.41	2.05	16.2	6.57	4.04
MALI82	242	2.85	2.26	1.98	19.8	8.04	4.94
RITT76	439	2.23	1.76	1.54	15.3	6.23	3.82
WEIN70	8.21	2.80	2.19	1.91	14.2	6.18	3.90
WEIN71	492	2.48	1.99	1.75	26.1	9.48	5.58

*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.

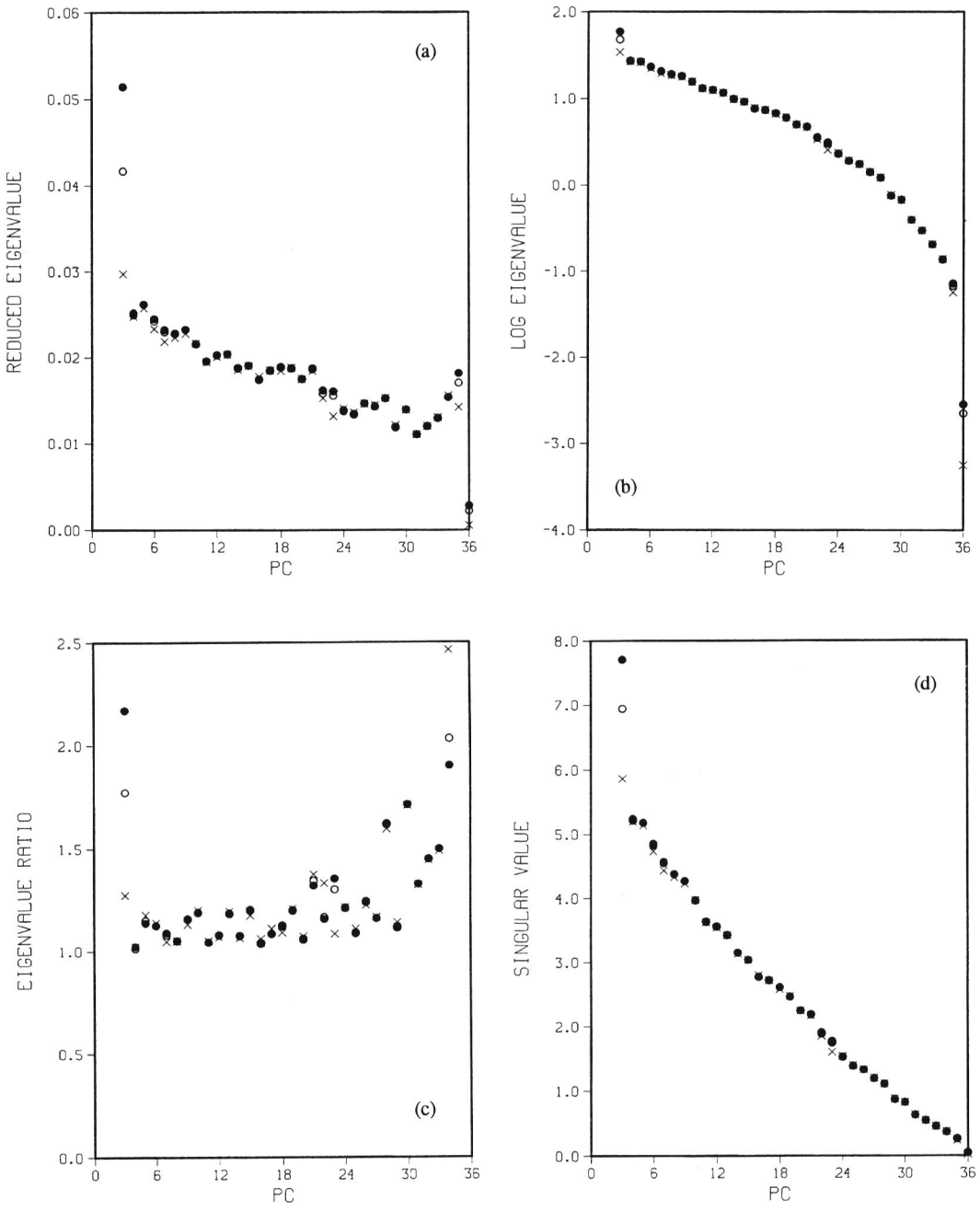


Figure 2. (a) Reduced eigenvalues, (b) logarithm of eigenvalues, (c) eigenvalue ratios and (d) singular values for EXP1 (●), EXP2 (◯) and EXP3 (×)

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, 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 applied to this matrix only a plot of the singular vectors indicates the presence of the third dimension (see Figure 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 (%α=34.5). The results for the simulated data sets support the results obtained for the literature data. The *t*-test is makes effective use of the extra knowledge. The estimated pseudoranks are summarized in Table 13.

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	%α	*REV*	F	ν_1	ν_2	%α
1	4230	5.75	0.22	5970	1296	0.0	13800	462	1	35	0.0
2	668	5.66	0.22	936	1225	0.0	364	15400	1	34	**0.0**
3	7.71	5.58	0.22	3.01	1156	**0.1**	0.051	2.44	1	33	12.8
4	5.24	5.49	0.22	-0.35	1089	63.7	0.025	1.22	1	32	27.9
5	5.18	5.40	0.22	-0.31	1024	62.2	0.026	1.30	1	31	26.4
6	4.85	5.31	0.23	-0.65	961	74.2	0.025	1.24	1	30	27.4

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

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

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 can not 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 order 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

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. equation (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 $E[\bullet]$ 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

equation (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 equation (13). We will return to this question in more detail in another publication [15]. It is tempting to evaluate equation (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 $\nu_2 = \sum_{n+1}^{s} N\,(r-j+1)\,(c-j+1)$ 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 equation (14). This problem can not 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. Chemometrics, 3 (1988) 49.
5 E.R. Malinowski, J. Chemometrics, 4 (1990) 102.
6 E.R. Malinowski, J. Chemometrics, 1 (1987) 33.
7 L. Ståhle and S. Wold, Chemometrics 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. Chemometrics, 5 (1991) 333.
10 S.P. Koinis, A.T Tsatsas and D.F. Katakis, J. Chemometrics, 5 (1991) 21.
11 Y.-Z. Liang, O.M. Kvalheim and A. Höskuldsson, J. Chemometrics, 7 (1993) 277.
12 V. Tomišić and V. Simeon, J. Chemometrics, 7 (1993) 381.
13 R.M. Wallace and S.M. Katz, J. Phys. Chem. 68 (1964) 3890.
14 Z.Z. Hugus and A.E. El-Awady, J. Phys. Chem. 75 (1971) 2954.
15 N.M. Faber, L.M.C. Buydens and G. Kateman, Chemometrics Intell. Lab. Syst. submitted.
16 N.M. Faber, L.M.C. Buydens and G. Kateman, J. Chemometrics, 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. Chemometrics, 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, John 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, 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. Chemometrics, 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. Japan, 45 (1972) 2643.
36 J. Havel, 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, Chemometrics Intell. Lab. Syst. 1 (1987) 167.
43 R.F. Hirsch, G. Lam Wu and P.C. Tway, Chemometrics 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, 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. Chemometrics, 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.

PART III GENERALIZED RANK ANNIHILATION METHOD

Part III consists of three papers. The first paper (Section 4) aims at discussing the various formulations that have been published until now for the generalized rank annihilation method (GRAM). More importantly, a shorter and more comprehensive derivation of the characteristic eigenvalue problem is presented.* Two problems that are inseparably connected to the method, i.e. the possibility of degenerate or complex solutions, are discussed using the alternative derivation and some possible solutions are proposed. The second paper (Section 5) is focussed on theoretical error estimates for the eigenvalues of GRAM (variance as well as bias). Extensive use is made of the alternative formulation presented in Section 4 and results known in the literature for univariate and multivariate calibration. Theoretical estimates may relieve the burden of experimental validation studies if the data follow the assumed model. Possible extensions are proposed in the section 'Conclusions and future research'. There is certainly reason for optimism, since some of the necessary extensions are straightforward from the current simple theory. This remark holds, for example, for the realistic noise model very recently discussed by Booksh and Kowalski[1] using Monte Carlo simulations. The third paper (Section 6) aims at showing that the numerical properties of the various implementations of GRAM are equivalent. It is important to point out that the statistical properties are far more important than the numerical properties because in many discussions GRAM is treated more like an algorithm than a method that solves a particular system of equations. This paper also tries to discuss the consequences of analyzing data with model errors. For example, if the data are corrupted by model errors the derived error estimates are completely useless. Thus there is also a good reason for pessimism although theoretical error estimates have been shown to work for real data in the case of multivariate calibration.[2]

Reference 10 in Section 4 corresponds to Section 1. References 2 and 7 in Section 5 correspond to Sections 4 and 1, respectively. References 8, 9 and 22 in Section 6 correspond to Sections 4, 5 and 1, respectively.

REFERENCES

1 K. Booksh and B.R. Kowalski, *J. Chemometrics*, **8**, 45 (1994).
2 G. Bauer, W. Wegscheider and H.M. Ortner, *Spectrochim. Acta B*, **46**, 1185 (1991).

*It is important to note that the current derivation closely resembles the algebraic start solution for PARAFAC given by Sands and Young in *Psychometrika*, **45**, 39 (1980). This was discovered after completion of the thesis.

GENERALIZED RANK ANNIHILATION METHOD. I: DERIVATION OF EIGENVALUE PROBLEMS

N. M. FABER, L. M. C. BUYDENS AND G. KATEMAN

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

SUMMARY

Rank annihilation factor analysis (RAFA) is a method for multicomponent calibration using two data matrices simultaneously, one for the unknown and one for the calibration sample. In its most general form, the generalized rank annihilation method (GRAM), an eigenvalue problem has to be solved. In this first paper different formulations of GRAM are compared and a slightly different eigenvalue problem will be derived. The eigenvectors of this specific eigenvalue problem constitute the transformation matrix that rotates the abstract factors from principal component analysis (PCA) into their physical counterparts. This reformulation of GRAM facilitates a comparison with other PCA-based methods for curve resolution and calibration. Furthermore, we will discuss two characteristics common to all formulations of GRAM, i.e. the distinct possibility of a complex and degenerate solution. It will be shown that a complex solution–contrary to degeneracy–should not arise for components present in both samples for model data.

KEY WORDS RAFA GRAM Eigenvalue problem Complex solution Degenerate solution

INTRODUCTION

Rank annihilation factor analysis (RAFA) is a method for multicomponent calibration using two data matrices simultaneously, one for the unknown and one for the calibration sample. In order to apply the technique of rank annihilation, the measured signal must be linear and additive, e.g. high-performance liquid chromatography with a diode array–UV/visible spectrophotometer as a detector (HPLC–DA–UV) or fluorescence excitation–emission spectroscopy. Data constructed in this way are called bilinear. For bilinear data the rank of a one-component data matrix is one in the absence of noise. Rank annihilation further demands that the signal for the analyte of interest be identical for both samples and finally it must be independent of the signal of the remaining substituents, i.e. the presence of the analyte of interest will raise the rank of the data matrix by one. If the data follow the assumed model, rank annihilation can be used to quantitate the analyte of interest without calibrating for the interferents.

The method was originally developed by Ho *et al.*[1] as an iterative procedure, but the latest developments in rank annihilation have their origin in Lorber's non-iterative reformulation of the calibration problem.[2] Lorber found a direct solution for the case where the calibration sample contains only one component. He derived a standard eigenvalue problem by projecting the calibration matrix on the significant principal components of the unknown data matrix.

The concentration ratio of the component was found as the only non-zero eigenvalue. Lorber's method was generalized by Sánchez and Kowalski[3] to the case of several components that are not necessarily present in both samples. This method, introduced as generalized rank annihilation factor analysis (GRAFA), is now known as the generalized rank annihilation method (GRAM). Wilson *et al.*[4] modified the procedure of Sánchez and Kowalski by projecting both matrices on a common low-dimensional subspace.

In this first paper we will compare different formulations of GRAM and present an alternative derivation that results in an eigenvalue problem for which the eigenvectors have a very simple interpretation: they form the transformation matrix that rotates the abstract factors found by a principal component analysis (PCA) to their physical counterparts. In this way a relationship can be established with other PCA-based curve resolution methods,[5,6] e.g. iterative target-testing factor analysis (ITTFA),[7] evolving factor analysis (EFA)[8] and the recently described window factor analysis (WFA).[9] Every method has its specific features, but a drawback common to all these curve resolution methods is the difficulty in constraining the (final) solution for the transformation matrix in an objective way. Finally we will discuss two characteristics of GRAM, i.e. the distinct possibility of a complex and degenerate solution. It will be shown that a complex solution should not arise for components that are present in both samples if the data follow the assumed linear additive model. However, the situation may be markedly different if the data are affected by model errors. Degeneracy constitutes a fundamental problem for ideal as well as non-ideal data. Whereas self-modeling curve resolution methods seem to work best when used for the calibration of samples that are very similar, at least part of the information may be lost if rank annihilation is applied.

We will start by introducing the relevant symbols and decompositions in the context of PCA-based curve resolution and calibration.

PCA-BASED CURVE RESOLUTION AND CALIBRATION

The goal of curve resolution is the decomposition of a data matrix into the pure contributions of the individual components. Without loss of generality we will assume throughout this paper that the data are obtained by the spectral detection of a chromatographic separation process. Then, if Beer's law is valid, the $S \times W$ data matrix $\mathbf{M}$ of S mixture spectra measured at W wavelengths can be written as

$$\mathbf{M} = \mathbf{H}\mathbf{Y}^{\mathrm{T}} \tag{1}$$

where $\mathbf{H}(S \times K)$ contains the pure elution profiles of the K components and $\mathbf{Y}(W \times K)$ contains the pure spectra. Usually the spectra in $\mathbf{Y}$ are normalized so that the concentration dependency is absorbed in $\mathbf{H}$.

In curve resolution one is primarily concerned with the reconstruction of $\mathbf{H}$ and $\mathbf{Y}$ (qualitative solution). The problem of calibration is more difficult, because the denormalized elution profiles in $\mathbf{H}$ have to be related to real concentration values (quantitative solution). This can be done directly if a theoretical relationship exists between the measured response and the concentrations. Otherwise an empirical relationship has to be built by estimating calibration factors from the response of a standard sample. Different calibration schemes are outlined in Reference 7. In calibration the following notation for $\mathbf{M}$ is often preferred:

$$\mathbf{M} = \mathbf{X}\mathbf{C}_{\mathrm{M}}\mathbf{Y}^{\mathrm{T}} \tag{2}$$

where the columns in $\mathbf{X}(S \times K)$ represent normalized elution profiles and $\mathbf{C}_{\mathrm{M}}$ is a $K \times K$ diagonal matrix proportional to the concentrations.

PCA-based curve resolution proceeds in two steps.[6] First, PCA is applied in order to define the solution space in terms of a set of orthogonal base vectors. This results in the following decomposition of **M**:

$$\mathbf{M} = \mathbf{A}\mathbf{B}^{\mathrm{T}} \tag{3}$$

where the matrices $\mathbf{A}(S \times W)$ and $\mathbf{B}^{\mathrm{T}}(W \times W)$ are usually denoted as score and loading matrices respectively. (It is assumed that $S \geqslant W$.) The decomposition in scores and loadings is equivalent to the singular value decomposition (SVD) of **M**:

$$\mathbf{M} = \mathbf{U}\Theta\mathbf{V}^{\mathrm{T}} \tag{4}$$

where **U** is the $S \times W$ matrix of left singular vectors, Θ is the $W \times W$ diagonal matrix of singular values and $\mathbf{V}^{\mathrm{T}}$ is the $W \times W$ matrix of right singular vectors. Scores and loadings are related to the singular vectors by*

$$\mathbf{A} = \mathbf{U}\Theta \tag{5a}$$

$$\mathbf{B} = \mathbf{V} \tag{5b}$$

Next **M** is reproduced using only the significant PCs and this decomposition is rewritten by means of a transformation matrix **T** as

$$\overline{\mathbf{M}} = \bar{\mathbf{A}}\bar{\mathbf{B}}^{\mathrm{T}} = \bar{\mathbf{A}}\mathbf{T}\mathbf{T}^{-1}\bar{\mathbf{B}}^{\mathrm{T}} \tag{6}$$

The 'overbar' denotes that the corresponding decomposition (PCA or SVD) is truncated. If F is the number of PCs retained for the reproduction, **T** is an $F \times F$ matrix. A successful transformation yields the physical decomposition of **M** up to a normalization constant:

$$\mathbf{H} = \bar{\mathbf{A}}\mathbf{T} \tag{7a}$$

$$\mathbf{Y}^{\mathrm{T}} = \mathbf{T}^{-1}\bar{\mathbf{B}}^{\mathrm{T}} \tag{7b}$$

Differences between PCA-based methods come down to differences in estimating the transformation matrix **T**. The problem of curve resolution and subsequent calibration is therefore translated to finding the correct number of factors F in the PCA step and determining a successful transformation matrix **T**. It is important to note that overfactoring the model will not change the PCs but will certainly affect the estimate of **T**. However, one frequently reported advantage of rank annihilation is the relative insensitivity of the solution to the number of PCs included in the model. This fact can very well be explained by the standard errors for the eigenvalues we recently derived using the method of error propagation.[10]

Until now the treatment has been restricted to the analysis of a single data matrix. In the next section it will become clear how the availability of a second data matrix, obtained under identical experimental circumstances, can help in determining **T**.

DIFFERENT FORMULATIONS OF RANK ANNIHILATION

As outlined in the Introduction, RAFA comprises a number of related methods. We will restrict ourselves to the discussion of methods that can be derived from the direct solution to the one-component problem, first published by Lorber.[2]

---

* The definition of scores and loadings mentioned before is purely conventional. A more operational definition has been given by Malinowski:[5] 'Attention is focused on either the row designees or the column designees. Where attention is focused is called the *scores*; the counterpart is called the *loadings*.' For the discussion of rank annihilation only the relation to the SVD is important.

Lorber's method

The method of rank annihilation can be applied if a calibration matrix **N** is present:

$$\mathbf{N} = \mathbf{X}\mathbf{C}_N\mathbf{Y}^T \tag{8a}$$

Equation (8a) is the preferred notation for the calibration matrix in the literature of rank annihilation. The alternative derivation presented in a later section will make use of the transcription

$$\mathbf{N} = \mathbf{H}\Pi\mathbf{Y}^T \tag{8b}$$

where $\Pi = \mathbf{C}_M^{-1}\mathbf{C}_N$. The components that are absent in the calibration sample (all but one) are indicated by a corresponding zero on the diagonal of $\mathbf{C}_N$ and Π respectively. Lorber[2] showed that combination of equations (1) and (8a) gives the generalized eigenvalue problem

$$\mathbf{N}\mathbf{Z} = \mathbf{M}\mathbf{Z}\Pi \tag{9}$$

where $\mathbf{Z} = (\mathbf{Y}^T)^+$. The matrices **N** and **M** are, however, not necessarily square and consequently the common eigenvalue-problem-solving routines[11] cannot be used. Approximating **M** by the truncated SVD of equation (4), i.e. $\bar{\mathbf{M}} = \bar{\mathbf{U}}\bar{\Theta}\bar{\mathbf{V}}^T$, and making use of the orthogonality properties of **U** and **V** leads to the standard eigenvalue problem

$$(\bar{\mathbf{U}}^T\mathbf{N}\bar{\mathbf{V}}\bar{\Theta}^{-1})\mathbf{Z}^* = \mathbf{Z}^*\Pi \tag{10}$$

where $\mathbf{Z}^* = \bar{\Theta}\bar{\mathbf{V}}^T\mathbf{Z}$. The concentration ratio of the analyte of interest is found as the only non-zero eigenvalue.

Generalization by Sánchez and Kowalski

If the calibration sample contains several components that are also present in the unknown sample, the eigenvalues found have to be identified, i.e. the qualitative solution is needed. Sánchez and Kowalski[3] recognized that the pure component responses, necessary for the identification step, could be reconstructed by

$$\mathbf{H} = \bar{\mathbf{U}}\mathbf{Z}^* \tag{11a}$$

$$\mathbf{Y}^T = (\mathbf{Z}^*)^{-1}\bar{\Theta}\bar{\mathbf{V}}^T \tag{11b}$$

Combination of (11a) and (11b) gives, as expected,

$$\bar{\mathbf{M}} = \bar{\mathbf{U}}\bar{\Theta}\bar{\mathbf{V}}^T \tag{11c}$$

In the general case where both samples contain unique components, the solution space has to be derived from a matrix that is a combination of the unknown and calibration data matrices. Sánchez and Kowalski propose to decompose the sum matrix $\mathbf{Q} = \mathbf{N} + \mathbf{M}$ and solve the standard eigenvalue problem with **M** substituted for **N**. This results in an eigenvalue matrix $\Pi = (\mathbf{C}_N + \mathbf{C}_M)^{-1}\mathbf{C}_M$ and a possible division by zero is avoided. With the necessary substitutions made, equations (10) and (11) describe the generalized rank annihilation method (GRAM).

Generalization by Wilson *et al.*

Wilson *et al.*[4] devised a different algorithm for carrying out the rank annihilation. Equation (9) is converted to the usual generalized eigenvalue problem by approximating **M** using

orthogonal bases **F** and **G** for the column and row space as follows:

$$\bar{\mathbf{M}} = \bar{\mathbf{F}}\bar{\mathbf{M}}_{FG}\bar{\mathbf{G}}^T \tag{12a}$$

$$\bar{\mathbf{N}} = \bar{\mathbf{F}}\bar{\mathbf{N}}_{FG}\bar{\mathbf{G}}^T \tag{12b}$$

If **M** spans the space of **N**, then **F** and **G** can be calculated from **M**. In the general case it is recommended to calculate **F** and **G** from the column and row augmented matrices $(\mathbf{N} \mid \mathbf{M})$ and $\binom{\mathbf{N}}{\mathbf{M}}$ respectively. Introducing $\mathbf{Z}_G = \bar{\mathbf{G}}\mathbf{Z}$ leads to

$$\bar{\mathbf{N}}_{FG}\mathbf{Z}_G = \bar{\mathbf{M}}_{FG}\mathbf{Z}_G\boldsymbol{\Pi} \tag{13}$$

The eigenvalue matrix $\boldsymbol{\Pi}$ is obtained from the QZ algorithm[12,11] in the form $\pi_k = \alpha_k/\beta_k$, where α_k and β_k are scalars, possibly zero or near zero. The divisions in $\pi_k = \alpha_k/\beta_k$ become the responsibility of the program's user.[12] The pure component responses are now reconstructed by

$$\mathbf{H} = \bar{\mathbf{F}}\bar{\mathbf{M}}_{FG}\mathbf{Z}_G \tag{14a}$$

$$\mathbf{Y}^T = (\mathbf{Z}_G)^{-1}\bar{\mathbf{G}}^T \tag{14b}$$

Combination of (14a) and (14b) gives equation (12a). The authors claim a better stability for their algorithm, since only unitary transformations are involved in solving equation (13).

Alternative derivation

Assuming that the components present in the calibration sample are a subset of those present in the unknown, a standard eigenvalue problem similar to Lorber's eigenvalue problem is derived. (The generalization is obtained by replacing $\mathbf{Q} = \mathbf{N} + \mathbf{M}$ for **M** and **M** for **N**.) The derivation has as its main advantage that it is short and leads to simple reconstruction expressions for the pure component responses.

Premultiplying **N** of equation (8b) by the pseudoinverse of **H**, postmultiplying by the pseudoinverse of $\mathbf{Y}^T$ and introducing $\mathbf{H}^+ = (\bar{\mathbf{A}}\mathbf{T})^+ = \mathbf{T}^{-1}\bar{\mathbf{A}}^+$ and $(\mathbf{Y}^T)^+ = (\mathbf{T}^{-1}\bar{\mathbf{B}}^T)^+ = \bar{\mathbf{B}}\mathbf{T}$ from equation (7) immediately leads to the standard eigenvalue problem* in scores and loadings

$$\mathbf{H}^+\mathbf{N}(\mathbf{Y}^T)^+ = \mathbf{T}^{-1}(\bar{\mathbf{A}}^+\mathbf{N}\bar{\mathbf{B}})\mathbf{T} = \boldsymbol{\Pi} \tag{15}$$

that has already been derived by Öhman *et al.*[13] by manipulations similar to those employed by Lorber.[2] (It is interesting to note that Kubista[14] arrives at the transpose problem by correlating the data matrices by means of a Procrustes rotation.)

Substitution of $\bar{\mathbf{A}}^+ = (\bar{\mathbf{U}}\bar{\boldsymbol{\Theta}})^+ = \bar{\boldsymbol{\Theta}}^{-1}\bar{\mathbf{U}}^T$ and $\bar{\mathbf{B}} = \bar{\mathbf{V}}$ and premultiplication by **T** yields

$$(\bar{\boldsymbol{\Theta}}^{-1}\bar{\mathbf{U}}^T\mathbf{N}\bar{\mathbf{V}})\mathbf{T} = \mathbf{T}\boldsymbol{\Pi} \tag{16}$$

The matrices **H** and $\mathbf{Y}^T$ are now calculated from equation (7).

There is a remarkable difference between equation (16) and equation (10) with respect to the position of the matrix $\bar{\boldsymbol{\Theta}}^{-1}$. The origin of this difference lies in the attribution of the singular values to the score matrix according to equation (5), which is merely a conventional choice

---

* In a subsequent paper we will show that the eigenvalues found by GRAM are biased estimates of the concentration ratios $\boldsymbol{\Pi}$ in equation (8). Deriving an expression for the bias obviates the introduction of a new symbol for the estimated concentration ratio, e.g. $\hat{\boldsymbol{\Pi}}$. Making a distinction between the concentration ratio actually present, $\boldsymbol{\Pi}$, and the concentration ratio estimated from the response matrices, $\hat{\boldsymbol{\Pi}}$, is not necessary for the present discussion.

from the point of view of this paper. It is, however, well known from matrix algebra that premultiplication of a matrix by a diagonal matrix followed by postmultiplication by the inverse matrix constitutes a similarity transformation and therefore leaves the eigenvalues unchanged.[15] The eigenvectors **T**, $\mathbf{Z}^*$ and $\mathbf{Z}_G$ are related according to

$$\mathbf{T} = \bar{\Theta}^{-1}\mathbf{Z}^* = (\bar{\mathbf{G}}^T\bar{\mathbf{V}})^{-1}\mathbf{Z}_G \tag{17}$$

The matrix **T** is found by multiplying the rows of $\mathbf{Z}^*$ by the inverse singular values of **M**. $\mathbf{Z}_G$ reduces to **T** if for the orthogonal bases **F** and **G** the singular vectors **U** and **V** of **M** are taken. This is immediate after premultiplication of equation (9) by $\bar{\Theta}^{-1}\bar{\mathbf{U}}^T$ and insertion of $\mathbf{Z} = \bar{\mathbf{V}}\mathbf{Z}_G = \bar{\mathbf{V}}\mathbf{Z}_V$. The manipulations are still very straightforward but the eigenvector matrix $\mathbf{Z}_V$ is identified as the desired transformation matrix **T** *a posteriori*. For the derivation of (16) the transformation matrix was the starting point.

CHARACTERISTICS OF RANK ANNIHILATION

We have seen that when applying the method of rank annihilation, an eigenvalue problem has to be solved. This leads to difficulties that are characteristic for this method. It would be nice if these difficulties already become apparent from the formulation of the model. Stated differently: rather then deriving an eigenvalue problem and discussing the properties of this eigenvalue problem, it should become clear why the formulation of the calibration problem leads to an eigenvalue problem with all its inherent properties.

The eigenvalue matrix $\mathbf{\Pi}$ enters the derivation by the transcription of equation (8a). It is seen that both (8a) and (8b) are misleading with respect to the information displayed. For the components not present in the calibration sample we have zeros on the diagonal of $\mathbf{C}_N$ or $\mathbf{\Pi}$ and the corresponding profiles in **X**, **H** and **Y** are in no way restricted to correlate with the profiles that reproduce **M**. They may even be complex, since they do not contribute to the data anyway. Furthermore, it is apparent from equation (8b) that under certain circumstances some (non-zero) diagonal elements of $\mathbf{\Pi}$ may be identical and in that case the corresponding profiles may be linear combinations. Therefore the characteristic difficulties of rank annihilation already show up from the alternative formulation of the model. How these difficulties must be encountered follows from the study of the eigenvalue problem.

Complex eigensolution

If the data follow the assumed linear additive model, i.e. equations (1) and (8) hold, the eigenvalues and eigenvectors corresponding to the calibrated components should be real. This is easily verified, since only the nth row and column of $\mathbf{H}^+$ and $(\mathbf{Y}^T)^+$ respectively are involved in the estimation of the nth eigenvalue in equation (15). This row and column are orthogonal to the remaining columns and rows in **H** and $\mathbf{Y}^T$ respectively, leading through equation (7) to a transformation vector $\mathbf{t}_n$ that is real. Therefore the eigenvector matrix **T** can be partitioned into a real and (possibly) complex part as $(\mathbf{T}_{real} \mid \mathbf{T}_{complex})$ or, using the terminology of Liang *et al.*,[16] into a white and black part as $(\mathbf{T}_{white} \mid \mathbf{T}_{black})$, since the unknown background remains unresolved.

It is in fact logical that the calibration matrix can only provide information on the components that are present in that sample ($c_{N,n} > 0$). The opposite must also be expected to hold: the components that are not present in the calibration sample (irrelevant part of the solution) cannot *misinform* about the components that are present in both samples (relevant part). For perfectly bilinear data the solution for the analytes of interest should always be real.

Two small problems remain. First, the *calculated* eigenvectors may be arbitrary to the extent of a complex multiplier of modulus one depending on the normalization used.[15] This problem is eliminated by normalizing the vectors in such a way that the largest component becomes one. It is easily verified that this solution amounts to the first similarity transformation of Li *et al.*[17] Second, for the reconstruction of **Y** the inverse of **T** must be calculated. It is easily verified that performing the calculations in complex arithmetic[18] will provide a real solution for the calibrated components. Two alternatives have been proposed in the chemometrical literature. The first alternative is to calculate $\mathbf{Y}^T$ from the transpose problem (so replacing **N** and **M** by $\mathbf{N}^T$ and $\mathbf{M}^T$) instead of using the inverse eigenvector matrix.[13] The second alternative is to convert the partially complex eigenvalue and eigenvector matrices to real matrices. This procedure amounts to the second similarity transformation of Li *et al.*[17] The three procedures should all lead to identical results for the analytes of interest.

Until now the discussion has been restricted to the analysis of model data. Model errors may cause the transformation vectors for the calibrated components to be complex as well.[13] Li and co-workers[17,19] have shown that in that case acceptable results can only be obtained if GRAM is modified by the second similarity transformation, thereby increasing the applicability of GRAM to very difficult practical situations.

Degenerate eigensolution

If some of the eigenvalues are identical, the direction of the corresponding eigenvectors is not fixed. In fact, every linear combination of vectors belonging to the eigenspace of the degenerate eigenvalues forms a valid eigenvector and the eigenvector actually calculated by the computer will even depend on the algorithm used. As a result, part of the qualitative solution of GRAM is not unique. In this case it is useful to have more than one calibration sample in order to obtain the complete qualitative solution.[20] It is also possible to remove the degeneracy by imposing extra noise on the data matrices. We have good experience with this procedure for simulated as well as real data. Although essentially correct and easily implemented, the calibration by rank annihilation now becomes a matter of trial and error that we find difficult to recommend for routine purposes. Moreover, using different calibration samples has the additional advantage that the assumed linear model can be validated.[21] Finally, it is possible to subtract the resolved components and use a self-modeling curve resolution method on the residual matrices.

A special form of degeneracy is known as a defect eigensystem:[15]

$$\lim_{\beta \to \alpha} \begin{pmatrix} \alpha & 1 \\ 0 & \beta \end{pmatrix} = \begin{pmatrix} \alpha & 1 \\ 0 & \alpha \end{pmatrix}$$

In the limit of $\beta \to \alpha$ the two eigenvalues are $\lambda_1 = \lambda_2 = \alpha$ and the corresponding eigenvectors coincide: $\mathbf{z}_1^T = \mathbf{z}_2^T = (1, 0)$. The eigenspace is one-dimensional and the matrix cannot be reduced to diagonal form, since the eigenvector matrix is not invertible. Now the direction of the eigenvectors is fixed and the pure component responses can be recovered. (The numerical consequences of eigenvalues belonging to so-called quadratic elementary divisors are discussed in Reference 12.) This situation, which could be expected to result from severe collinearity in the data, will, however not occur. If components are undistinguishable within the noise level, this will result in a reduction of the number of factors found from PCA and the reconstructed component will simply be an average of the components that were too similar. The selectivity of the experiment must be improved in that case in order to obtain the complete quantitative solution (e.g. by employing bimodal data[22]). This situation applies to all PCA-based techniques and is therefore not typical for rank annihilation.

CONCLUSIONS

We have presented a simple derivation of the standard eigenvalue problem that arises if the method of rank annihilation is used for calibration. The availability of two data matrices, obtained under identical circumstances, enables the direct evaluation of the matrix that transforms the abstract decomposition of a matrix into the desired physical one. In this way a relationship is established with other PCA-based curve resolution and calibration methods. The transformation matrix found by rank annihilation may consist of a real and (possibly) complex part if some of the components in the unknown sample are not calibrated. However, in the absence of model errors the relevant part of the transformation matrix is not affected by the irrelevant part. In the presence of model errors (e.g. irreproducibility of chromatographic data) the relevant part of the eigensolution may also become complex.[13,17,19] In that case it is mandatory to follow up the solution of the eigenvalue problem with unitary transformations as recommended in References 17 and 19. It remains an important issue as to how the accuracy and precision of the transformed solution must be estimated.

REFERENCES

1. C.-N. Ho, G. D. Christian and E. R. Davidson, *Anal. Chem.* **50**, 1108 (1978).
2. A. Lorber, *Anal. Chim. Acta*, **164**, 293 (1984).
3. E. Sánchez and B. R. Kowalski, *Anal. Chem.* **58**, 496 (1986).
4. B. E. Wilson, E. Sánchez and B. R. Kowalski, *J. Chemometrics*, **3**, 493 (1989).
5. E. R. Malinowski and D. G. Howery, *Factor Analysis in Chemistry*, Wiley, New York (1991).
6. W. Windig, *Chemometrics Intell. Lab. Syst.* **16**, 1 (1992).
7. B. G. M. Vandeginste, F. Leyten, M. Gerritsen, J. W. Noor and G. Kateman, *J. Chemometrics*, **1**, 57 (1987).
8. H. Gampp, M. Maeder, C. J. Meyer and A. D. Zuberbuhler, *Talanta*, **32**, 1133 (1985).
9. E. R. Malinowski, *J. Chemometrics*, **6**, 29 (1992).
10. N. M. Faber, L. M .C. Buydens and G. Kateman, *J. Chemometrics*, **7**, 495 (1993).
11. IMSL, International Mathematical and Statistical Libraries, Inc., 7500 Bellaire Blvd., Houston, TX 77036, U.S.A.
12. C. B. Moler and G. W. Stewart, *SIAM J. Numer. Anal.* **10**, 241 (1973).
13. J. Öhman, P. Geladi and S. Wold, *J. Chemometrics*, **4**, 135 (1990).
14. M. Kubista, *Chemometrics Intell. Lab. Syst.* **7**, 273 (1990).
15. J. H. Wilkinson, *The Algebraic Eigenvalue Problem*, Clarendon, Oxford (1965).
16. Y.-Z. Liang, O. M. Kvalheim and R. Manne, *Chemometrics Intell. Lab. Syst.* **18**, 235 (1993).
17. S. Li, J. C. Hamilton and P. J. Gemperline, *Anal. Chem.* **64**, 599 (1992).
18. J. R. Westlake, *A Handbook of Numerical Matrix Inversion and Solution of Linear Equations*, Wiley, New York (1968).
19. S. Li and P. J. Gemperline, *J. Chemometrics*, **7**, 77 (1993).
20. E. Sánchez and B. R. Kowalski, *J. Chemometrics*, **2**, 265 (1988).
21. A. Lorber, *Anal. Chem.* **57**, 2395 (1985).
22. L. S. Ramos, E. Sánchez and B. R. Kowalski, *J. Chromatogr.* **385**, 165 (1987).

GENERALIZED RANK ANNIHILATION METHOD. II: BIAS AND VARIANCE IN THE ESTIMATED EIGENVALUES

N. M. FABER, L. M. C. BUYDENS AND G. KATEMAN

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

SUMMARY

Rank annihilation factor analysis (RAFA) is a method for multicomponent calibration using two data matrices simultaneously, one for the unknown and one for the calibration sample. In its most general form, the generalized rank annihilation method (GRAM), an eigenvalue problem has to be solved. In this second paper expressions are derived for predicting the bias and variance in the eigenvalues of GRAM. These expressions are built on the analogies between a reformulation of the eigenvalue problem and the prediction equations of univariate and multivariate calibration. The error analysis will also be performed for Lorber's formulation of RAFA. It will be demonstrated that, depending on the size of the eigenvalue, large differences in performance must be expected. A bias correction technique is proposed that effectively eliminates the bias if the error in the bias estimate is not too large. The derived expressions are evaluated by Monte Carlo simulations. It is shown that the predictions are satisfactory up to the limit of detection. The results are not sensitive to an incorrect choice of the dimension of the factor space.

KEY WORDS RAFA GRAM Eigenvalues Bias Variance

INTRODUCTION

RAFA is a calibration and curve resolution technique that enables the quantification of a component in the presence of an unknown background.[1] The method can be used if the signal of the analytes of interest is bilinear and identical in the unknown and calibration sample. Furthermore, the analytes of interest should raise the rank of the data matrices by one. The model is represented as

$$\mathbf{M} = \mathbf{X}\mathbf{C}_M\mathbf{Y}^T = \mathbf{H}\mathbf{Y}^T \tag{1a}$$

$$\mathbf{N} = \mathbf{X}\mathbf{C}_N\mathbf{Y}^T = \mathbf{H}\boldsymbol{\Pi}\mathbf{Y}^T \tag{1b}$$

where $\mathbf{M}$ and $\mathbf{N}$ denote the unknown and calibration data matrices respectively. The pure component profiles $\mathbf{X}$ and $\mathbf{Y}$ are normalized and the diagonal matrices $\mathbf{C}_M$ and $\mathbf{C}_N$ contain the scaling factors. In the following we will assume that the calibration sample does not contain components that are missing in the unknown sample. In that case it is possible to absorb the matrix $\mathbf{C}_M$ into the scaled profiles. The conventional choice is to denormalize the column profiles leading to the matrices $\mathbf{H} = \mathbf{X}\mathbf{C}_M$ and $\boldsymbol{\Pi} = \mathbf{C}_M^{-1}\mathbf{C}_N$. In order to resolve (the relevant part of) the data matrices and obtain the corresponding concentration ratios $\boldsymbol{\Pi}$, the data matrices are combined in one equation. This can be accomplished in several ways. Representing $\mathbf{M}$ by

the significant PCs

$$\bar{\mathbf{M}} = \bar{\mathbf{A}}\bar{\mathbf{B}}^{\mathrm{T}} = \bar{\mathbf{U}}\bar{\mathbf{\Theta}}\bar{\mathbf{V}}^{\mathrm{T}} \tag{2}$$

leads to the following decompositions of $\mathbf{\Pi}$:[2]

$$\mathbf{\Pi} = \mathbf{H}^{+}\mathbf{N}(\mathbf{Y}^{\mathrm{T}})^{+} \tag{3a}$$

$$= \mathbf{T}^{-1}(\bar{\mathbf{A}}^{+}\mathbf{N}\bar{\mathbf{B}})\mathbf{T} \tag{3b}$$

$$= \mathbf{T}^{-1}(\bar{\mathbf{\Theta}}^{-1}\bar{\mathbf{U}}^{\mathrm{T}}\mathbf{N}\bar{\mathbf{V}})\mathbf{T} \tag{3c}$$

$$= \mathbf{Z}^{-1}(\bar{\mathbf{U}}^{\mathrm{T}}\mathbf{N}\bar{\mathbf{V}}\bar{\mathbf{\Theta}}^{-1})\mathbf{Z} \tag{3d}$$

where **A**, **B**, **U** and **V** denote the scores, loadings and left and right singular vectors respectively, $\mathbf{\Theta}$ is the diagonal matrix with singular values and the matrices **T** and **Z** contain eigenvectors. The 'overbar' indicates that the decomposition in equation (2) is truncated. Equations (3b)–(3d) are eigenvalue problems that directly provide the desired information, whereas equation (3a) shows the relation to the physical decomposition of the data matrices. Equation (3d) was first derived by Lorber[3] for the case where the calibration sample contains only one component. The generalization of his method is achieved by constructing a factor space that spans both data matrices. Sánchez and Kowalski[4] have solved this problem by decomposing the sum matrix and recently Wilson *et al.*[5] have recommended decomposing the row and column adjoined matrices.

The decomposition in real profiles has a clear interpretational advantage over the decomposition in abstract profiles. The presence of the pseudoinverse matrices makes the relation with multivariate calibration obvious. From multivariate calibration it is well known that the pseudoinverse of a stochastic calibration matrix will lead to biased concentration estimates.[6] It follows that the eigenvalues must be expected to be biased also. In order to derive expressions for the prediction of bias it will therefore be advantageous to apply the results already obtained for multivariate calibration. Furthermore, we have recently derived standard errors in the eigenvalues by performing first-order error propagation on equation (3d).[7] In the derivation we make use of approximations that are more transparent when error propagation is performed on equation (3a). We will derive expressions for bias and variance in the eigenvalues for Lorber's method, the generalization of Sánchez and Kowalski and the generalization of Wilson *et al.* If possible, the expressions will be presented in the same way as already known for multivariate calibration. Bias and variance resulting from errors in the concentrations can be predicted from results derived for univariate calibration.[8]

The remaining part of this paper will be organized as follows. First we will give a general introduction to the effect of random errors on an estimated parameter. Next we will derive the bias expressions for the eigenvalues. In the following section we will summarize the relevant expressions for the variance in the eigenvalues. Next we will introduce a procedure for the correction of bias. This procedure should perform well if the bias estimate is reasonable. Next we will show how these bias-corrected eigenvalues can be used to set up a probability range. In the following sections we will briefly discuss the bias and variance in the unknown concentrations. It will be shown from the variance expression for the unknown concentration that for the methods discussed the generalization of Sánchez and Kowalski will always give the smallest (relative) variance in the eigenvalues. Finally, the adequacy of the derived expressions is tested by performing Monte Carlo simulations. We will restrict ourselves to examining the effect of random noise in the response matrices, since the effect of noise in the concentrations is trivial. It will be shown that the expressions are reliable up to the limit of detection. Furthermore, the results do not seem to be sensitive to the choice of dimensionality of the factor space.

NOTATION AND CONVENTIONS

Boldface uppercase letters represent matrices, e.g. $\mathbf{A}$. For a given matrix $\mathbf{A}$ the matrices $\mathbf{A}^{\mathrm{T}}$, $\mathbf{A}^{-1}$ and $\mathbf{A}^{+}$ stand for its transpose, inverse and pseudoinverse respectively. The 'inverse transpose' and 'pseudoinverse transpose' matrices will be denoted by the shorthand notation $\mathbf{A}^{-\mathrm{T}} = (\mathbf{A}^{-1})^{\mathrm{T}} = (\mathbf{A}^{\mathrm{T}})^{-1}$ and $\mathbf{A}^{\ddagger} = (\mathbf{A}^{+})^{\mathrm{T}} = (\mathbf{A}^{\mathrm{T}})^{+}$. The matrix element in row i and column j of $\mathbf{A}$ will be specified by a row and column index as A_{ij}. The nth row and nth column of $\mathbf{A}$ will be denoted by $\mathbf{A}_{n\text{-row}}$ and $\mathbf{A}_{n\text{-col}}$ respectively. In order to discuss the effect of random error on variance and bias in the estimated eigenvalues, it is necessary to include the random error in the model equations. If the elements of $\mathbf{A}$ are unbiased, $\mathbf{A}$ can be decomposed as $\tilde{\mathbf{A}} + \delta\mathbf{A}$, where the 'tilde' indicates the true values and the matrix $\delta\mathbf{A}$ contains only random error. This notation will be used, for example, for the pure component profiles in equation (1). Biased quantities such as the eigenvalues will not be decomposed this way (see Appendix III).*

BIAS AND VARIANCE RESULTING FROM RANDOM ERRORS

It is important to note that we will deal with bias and variance in the estimated parameters that result from random measurement noise. There is no bias in the data and the factor models used will not be underdimensioned. It will be shown that the estimated parameters are intrinsically biased if the calculation involves a non-linear transformation of the data, which is certainly the case for rank annihilation. For the derivation of variance expressions we will make use of first-order approximations.[7] For the derivation of bias expressions additional approximations are needed. A systematic numerical evaluation must point out whether these approximations can be justified. Usually the assumption of uncorrelated, homoscedastic noise will be made. However, the assumption of homoscedasticity is only made in order to obtain tractable expressions and these expressions are in fact obtained after simplifying the expressions derived for heteroscedastic noise. The results are distribution-free, since we only use the size of the errors, i.e. we investigate how the standard error of the parent distribution is propagated by the system of equations under study. We will illustrate these points by a simple example that can be worked out by hand, because we assume a special distribution for the errors. A good discussion of the problem can also be found in Reference 8, where the influence of random error on the result of univariate standard addition is discussed using more realistic distributions.

Assume that we measure two random variables X and Y and we want to estimate the ratio $Z = X/Y$. This problem is extensively treated in the statistics literature.[9] In our specific example the true values $\tilde{X}$ and $\tilde{Y}$ are 10 and the measurement error can only take two values, -1 and $+1$. Therefore the sets of possible outcomes for X, Y and Z are

$$\Omega_X = \Omega_Y = \{9, 11\}$$

$$\Omega_Z = \left\{\frac{9}{9}, \frac{9}{11}, \frac{11}{9}, \frac{11}{11}\right\}$$

The true bias in Z is calculated as the difference between the expected value for Z, denoted by $E[Z]$, and the true value for Z, i.e. $\tilde{Z} = \tilde{X}/\tilde{Y} = 1$. The expected value for Z is the mean for the four possible outcomes,† i.e. $E[Z] = 1{\cdot}0101$. It follows that the estimated ratio is biased

---

*The pure component profiles are also biased, but calculations show that the bias is neglible compared with the variance in the regime where the bias in the eigenvalues is already considerable.

† In general we have many possible outcomes and we evaluate newly derived expressions by Monte Carlo simulations.

upwards and the bias in Z, bias(Z), takes the value of $+0{\cdot}0101$ here. The predicted value for $E[Z]$ is $1{\cdot}0100$ (see equation (41) in Appendix I) and the predicted bias is therefore $+0{\cdot}0100$. It is seen that the agreement between predicted and 'experimental' values is excellent. The true variance in Z is given by $\text{var}(Z) = E[Z^2] - (E[Z])^2 = 0{\cdot}0205$, whereas the predicted value is $0{\cdot}0200$ (see equation (42) in Appendix I). Again the agreement is excellent. However, it should be noted that in order to evaluate the predicted value, we used the true values for X, Y and the standard deviation of the parent distribution. If instead we inserted experimental values, we should expect to predict bias and variance with an error that is dictated by the size of the experimental error (about 10%). We merely used the theoretical values to indicate that the resulting predictions do not depend on distributional assumptions for the noise. The largest errors must be expected to be caused by the evaluation of these expressions using experimental values.[10]

BIAS IN THE EIGENVALUES

In the preceding section we showed how random measurement noise induces bias in the estimated parameters if the calculation involves a non-linear transformation. Measurement noise enters the rank annihilation procedure at two different places. First, there may be errors in the concentrations. These errors are introduced, for example, by the sample preparation process or the injection volume irreproducibility. This error is independent of the second kind of error, which is caused by the detection process eventually leading to the data matrices for the unknown and calibration sample. Usually the error in the concentration is (much) larger than the error in the measured response and three possibilities for the eigenvalues found by rank annihilation are relevant. Consequently we introduce three different symbols for the eigenvalues.

(1) There are no errors in the concentrations or in the measured response:

$$\tilde{\mathbf{\Pi}} = \tilde{\mathbf{C}}_M^{-1}\tilde{\mathbf{C}}_N \tag{4a}$$

and it is evident that the true concentration ratios must be found.

(2) There are only errors in the concentrations:

$$\mathbf{\Pi} = \mathbf{C}_M^{-1}\mathbf{C}_N \tag{4b}$$

and in this case the concentration ratios found are biased upwards as shown in the preceding section.

(3) There are errors in the concentrations and in the measured response and (3a) becomes

$$\hat{\mathbf{\Pi}} = \mathbf{H}^{+}\mathbf{N}\mathbf{Y}^{\ddagger} \tag{4c}$$

Now, since the pseudoinverse matrices are not free from error, an additional bias is introduced (see Appendix I). It is important to note that the equivalent expression for multivariate calibration contains the pseudoinverse of a matrix that enters the calculation, whereas in equation (4c) the pseudoinverse matrices are calculated for the reconstructed profiles. Because of the different independent contributions to the bias, the following is immediate for practical situations:

$$E[\hat{\mathbf{\Pi}}] \neq E[\mathbf{\Pi}] \neq \tilde{\mathbf{\Pi}} \tag{5}$$

In the remaining sections we will refer to the eigenvalues defined by equations (4) as *true*, *actual* and *estimated* concentration ratios respectively. The bias in the actual concentration

ratio is easily estimated from equation (41). The biasing effect of the error in the reconstructed profiles is derived by considering the error in the scores and loadings (Appendix II). We will work out these expressions for Lorber's method and the generalizations of Sánchez and Kowalski and of Wilson *et al.*

Lorber's method

First, the model represented by equation (1) is rewritten as

$$\mathbf{M} = \mathbf{H}_\mathrm{M}\mathbf{Y}_\mathrm{M}^\mathrm{T} = (\tilde{\mathbf{H}}_\mathrm{M} + \delta\mathbf{H}_\mathrm{M})(\tilde{\mathbf{Y}} + \delta\mathbf{Y}_\mathrm{M})^\mathrm{T} \tag{6a}$$

$$\mathbf{N} = \mathbf{H}_\mathrm{N}\boldsymbol{\Pi}_\mathrm{N}\mathbf{Y}_\mathrm{N}^\mathrm{T} = (\tilde{\mathbf{H}}_\mathrm{M} + \delta\mathbf{H}_\mathrm{N})\boldsymbol{\Pi}_\mathrm{N}(\tilde{\mathbf{Y}} + \delta\mathbf{Y}_\mathrm{N})^\mathrm{T} \tag{6b}$$

The subscripts of the experimental profiles, the error terms and the eigenvalue matrix refer to the matrix from which they are derived. The concentration dependence implicit in the column profiles of $\tilde{\mathbf{H}}$ is also indicated by adding a subscript. The resulting prediction equation is

$$\hat{\boldsymbol{\Pi}}_\mathrm{N} = \mathbf{H}_\mathrm{M}^{+}\mathbf{N}\mathbf{Y}_\mathrm{M}^{+\mathrm{T}} \tag{7}$$

The errors in the profiles of $\mathbf{N}$ and $\mathbf{M}$ are uncorrelated. Therefore the expected value of the eigenvalues can be evaluated using equations (46), (56) and (57) as

$$\begin{aligned} E[\hat{\boldsymbol{\Pi}}_\mathrm{N}] &= E[\mathbf{H}_\mathrm{M}^{+}\mathbf{N}\mathbf{Y}_\mathrm{M}^{+\mathrm{T}}] \\ &= E[\mathbf{H}_\mathrm{M}^{+}]\,E[\mathbf{H}_\mathrm{N}]\,E[\boldsymbol{\Pi}_\mathrm{N}]\,E[\mathbf{Y}_\mathrm{N}^\mathrm{T}]\,E[\mathbf{Y}_\mathrm{M}^{+\mathrm{T}}] \\ &= [\mathbf{I} + (\tilde{\mathbf{H}}_\mathrm{M}^\mathrm{T}\tilde{\mathbf{H}}_\mathrm{M})^{-1}E[\delta\mathbf{H}_\mathrm{M}^\mathrm{T}\,\delta\mathbf{H}_\mathrm{M}]\,]^{-1}E[\boldsymbol{\Pi}_\mathrm{N}]\,[\mathbf{I} + E[\delta\mathbf{Y}_\mathrm{M}^\mathrm{T}\,\delta\mathbf{Y}_\mathrm{M}]\,(\tilde{\mathbf{Y}}^\mathrm{T}\tilde{\mathbf{Y}})^{-1}]^{-1} \\ &= (\mathbf{I} + S\sigma_\mathrm{M}^2\tilde{\boldsymbol{\Psi}}_\mathrm{M})^{-1}E[\boldsymbol{\Pi}_\mathrm{N}]\,(\mathbf{I} + W\sigma_\mathrm{M}^2\tilde{\boldsymbol{\Psi}}_\mathrm{M})^{-1} \end{aligned} \tag{8}$$

where the symbol $\boldsymbol{\Psi} = \mathbf{T}^{-1}\boldsymbol{\Lambda}^{-1}\mathbf{T}$ is introduced. If the errors are small, the bias in the estimated eigenvalues can be evaluated by only considering the diagonal elements of the correction terms (see equation (47)). The result is

$$\mathrm{bias}(\hat{\pi}_{\mathrm{N},n}) = E[\hat{\pi}_{\mathrm{N},n}] - E[\pi_{\mathrm{N},n}] = -(S + W)\sigma_\mathrm{M}^2\tilde{\boldsymbol{\Psi}}_{\mathrm{M},nn}E[\hat{\pi}_{\mathrm{N},n}] \tag{9}$$

where $\pi_{\mathrm{N},n} = \Pi_{\mathrm{N},nn}$. The bias induced by errors in the response is proportional to the size of the eigenvalue and, since the matrix $\tilde{\boldsymbol{\Psi}}_\mathrm{M}$ is positive definite, is expected to have a negative sign. Analogously to multivariate calibration, we can define $\boldsymbol{\Psi}$ as the variance factor[11] and again several transcriptions are possible. Summarizing gives

$$\boldsymbol{\Psi}_{nn} = (\mathbf{H}^\mathrm{T}\mathbf{H})_{nn}^{-1}(\mathbf{Y}^\mathrm{T}\mathbf{Y})_{nn}^{-1} \tag{10a}$$

$$= \|\,\mathbf{H}_{n\text{-row}}^{+}\,\|^2\,\|\,\mathbf{Y}_{n\text{-col}}^{+}\,\|^2 \tag{10b}$$

$$= \sum_{p=1}^{F}\left(\frac{T_{pn}}{\theta_{\mathrm{M},p}}\right)^2 \tag{10c}$$

$$= (\mathrm{SEL})_{\mathrm{X},n}^{-2}(\mathrm{SEL})_{\mathrm{Y},n}^{-2}c_{\mathrm{M},n}^{-2} \tag{10d}$$

where $\|\bullet\|$ is the Euclidean vector norm and $\theta_{\mathrm{M},p} = \Theta_{\mathrm{M},pp}$. The close relation between multivariate and bilinear calibration is apparent from equations (10a) and (10b). The equivalent for (50c) now contains the elements of the eigenvectors. In essence it shows that in rank annihilation the two fundamental problems of linear algebra, i.e. the least squares problem and the eigenvalue problem, are combined. It is important to note that the equivalent of (50d) contains Lorber's selectivities[12] in the separate modes of the data. The situation is different from the univariate and multivariate case, where we have seen that the error in the

determined concentration depends on the sensitivity of the measurement. The generalization of the sensitivity concept from multivariate to bilinear calibration does not seem to be straightforward. It should be noted that Ho *et al.*[13] have derived an expression for the variance factor for the iterative procedure that is similar to equation (10d). The only difference from the current result is the presence of a factor of two. A comparison shows that their 'uniqueness' q relates to Lorber's selectivity as $q = (\mathrm{SEL})$. The extension to three-way data is straightforward as shown by Appelof and Davidson.[2,14] However, these theories do not include the error in the calibration matrix, thus limiting their practical usefulness considerably.

Generalization of Sánchez and Kowalski

In the generalization of Sánchez and Kowalski the matrices **N** and **M** are replaced by **M** and **Q** = **N** + **M** respectively. The analogy of equation (6) is

$$\mathbf{Q} = \mathbf{H}_\mathrm{Q}\mathbf{Y}_\mathrm{Q}^\mathrm{T} = (\tilde{\mathbf{H}}_\mathrm{Q} + \delta\mathbf{H}_\mathrm{Q})(\tilde{\mathbf{Y}} + \delta\mathbf{Y}_\mathrm{Q})^\mathrm{T} \tag{11a}$$

$$\mathbf{M} = \mathbf{H}_\mathrm{M}\boldsymbol{\Pi}_\mathrm{M}\mathbf{Y}_\mathrm{M}^\mathrm{T} = (\tilde{\mathbf{H}}_\mathrm{Q} + \delta\mathbf{H}_\mathrm{M})\boldsymbol{\Pi}_\mathrm{M}(\tilde{\mathbf{Y}} + \delta\mathbf{Y}_\mathrm{M})^\mathrm{T} \tag{11b}$$

$$\mathbf{N} = \mathbf{H}_\mathrm{N}(\mathbf{I} - \boldsymbol{\Pi}_\mathrm{M})\mathbf{Y}_\mathrm{N}^\mathrm{T} = (\tilde{\mathbf{H}}_\mathrm{Q} + \delta\mathbf{H}_\mathrm{N})(\mathbf{I} - \boldsymbol{\Pi}_\mathrm{M})(\tilde{\mathbf{Y}} + \delta\mathbf{Y}_\mathrm{N})^\mathrm{T} \tag{11c}$$

The analogy of equation (7) is

$$\hat{\boldsymbol{\Pi}}_\mathrm{M} = \mathbf{H}_\mathrm{Q}^{+}\mathbf{M}\mathbf{Y}_\mathrm{Q}^{+} \tag{12}$$

The errors in **M** and **Q** are correlated and the same holds for the errors in the reconstructed profiles. Consequently, cross-terms containing the covariance matrix $E[\delta\mathbf{H}_\mathrm{Q}^\mathrm{T}\,\delta\mathbf{H}_\mathrm{M}]$ will arise in the evaluation of the expectation of the eigenvalues:

$$E[\hat{\boldsymbol{\Pi}}_\mathrm{M}] = E[\mathbf{H}_\mathrm{Q}^{+}\mathbf{M}\mathbf{Y}_\mathrm{Q}^{+}]$$
$$E[\mathbf{H}_\mathrm{Q}^{+}\mathbf{H}_\mathrm{M}]\,E[\boldsymbol{\Pi}_\mathrm{M}]\,E[\mathbf{Y}_\mathrm{M}^\mathrm{T}\mathbf{Y}_\mathrm{Q}^{+}] \tag{13}$$

The problem of evaluating the covariance matrix is solved by noting that to first order an element of $\mathbf{H}_\mathrm{Q}$ is given by

$$H_{\mathrm{Q},ip} = H_{\mathrm{M},ip}\pi_{\mathrm{M},p} + H_{\mathrm{N},ip}(1 - \pi_{\mathrm{M},p}) \tag{14}$$

and the errors in the profiles are related as

$$\delta\mathbf{H}_\mathrm{M} = \delta\mathbf{H}_\mathrm{Q}\boldsymbol{\Pi}_\mathrm{M}^{-1} \tag{15}$$

Working out equation (13) eventually results in

$$E[\hat{\boldsymbol{\Pi}}_\mathrm{M}] = (\mathbf{I} + S\sigma_\mathrm{Q}^2\tilde{\boldsymbol{\Psi}}_\mathrm{Q})^{-1}(\mathbf{I} + S\sigma_\mathrm{M}^2\tilde{\boldsymbol{\Psi}}_\mathrm{Q}\boldsymbol{\Pi}_\mathrm{M}^{-1})E[\boldsymbol{\Pi}_\mathrm{M}](\mathbf{I} + W\sigma_\mathrm{M}^2\hat{\boldsymbol{\Pi}}_\mathrm{M}^{-1}\tilde{\boldsymbol{\Psi}}_\mathrm{Q})(\mathbf{I} + W\sigma_\mathrm{Q}^2\tilde{\boldsymbol{\Psi}}_\mathrm{Q})^{-1} \tag{16}$$

It is seen that two cross-terms are introduced that have a cancelling effect. In fact, this observation is somewhat misleading, as follows from the analogy of equation (9),

$$\mathrm{bias}(\hat{\pi}_{\mathrm{M},n}) = [-(S + W)\sigma_\mathrm{Q}^2 + (S + W - 2F - 2)\sigma_\mathrm{M}^2\pi_{\mathrm{M},n}^{-1}]\tilde{\Psi}_{\mathrm{Q},nn}E[\hat{\pi}_{\mathrm{M},n}] \tag{17}$$

If the response error is identical for **N** and **M** and $F \ll \min(S, W)$, there will be almost no bias if the concentrations are identical for both samples. The situation is also quite favorable if the concentrations are of the same order of magnitude. However, for components that have a low concentration in the unknown sample, a large bias may be found compared with the previous case. (It should be noted that in order to compare equations (9) and (17), the normalization of the profiles in **H** must be taken into account.)

Generalization of Wilson *et al.*

In the generalization of Wilson *et al.* the matrices **N** and **M** are adjoined rather than added in order to define the factor space. The column space is found by decomposing the column augmented matrix $\mathbf{Q}_C$ as

$$\mathbf{Q}_C = (\mathbf{N} \mid \mathbf{M}) = \mathbf{U}_C \boldsymbol{\Theta}_C \mathbf{V}_C^T \tag{18a}$$

and the row space is found by decomposing the row augmented matrix $\mathbf{Q}_R$ as

$$\mathbf{Q}_R = \begin{pmatrix} \mathbf{N} \\ \mathbf{M} \end{pmatrix} = \mathbf{U}_R \boldsymbol{\Theta}_R \mathbf{V}_R^T \tag{18b}$$

We have chosen to decompose the matrices using the SVD in order to parallel the preceding discussion. In fact, any orthogonal decomposition should give identical results. The matrices **N** and **M** are now denoted as in equation (6) and the resulting eigenvalue problem is (for more details see Reference 5)

$$\mathbf{N}_{UV}\mathbf{Z}_V = \mathbf{M}_{UV}\mathbf{Z}_V\hat{\boldsymbol{\Pi}}_N \tag{19}$$

where $\mathbf{N}_{UV} = \mathbf{U}_C^T\mathbf{N}\mathbf{V}_R$ and $\mathbf{M}_{UV} = \mathbf{U}_C^T\mathbf{M}\mathbf{V}_R$. Equation (19) can be converted into the analogy of equations (7) and (12) by introducing the following expression for the reconstructed profiles:

$$\mathbf{H}_C = \mathbf{U}_C\mathbf{M}_{UV}\mathbf{Z}_V \tag{20a}$$

$$\mathbf{Y}_R^T = \mathbf{Z}_V^{-1}\mathbf{V}_R^T \tag{20b}$$

Again cross-terms arise, now containing $E[\delta\mathbf{H}_C^T\,\delta\mathbf{H}_N]$, which can be worked out by taking the correlations into account. First it is recognized that errors in $\mathbf{H}_N$ are related to errors in **N** as follows:

$$\begin{aligned} \delta\mathbf{H}_N &= \delta\mathbf{N}\mathbf{Y}_N^{\dagger}\boldsymbol{\Pi}_N^{-1} \\ &\approx \delta\mathbf{N}\mathbf{Y}_R^{\dagger}\boldsymbol{\Pi}_N^{-1} \\ &= \delta\mathbf{N}\mathbf{V}_R\mathbf{Z}_V\boldsymbol{\Pi}_N^{-1} \end{aligned} \tag{21}$$

Next $\mathbf{H}_C$ is rewritten as

$$\mathbf{H}_C = \mathbf{U}_C\boldsymbol{\Theta}_C\mathbf{T}_C \tag{22}$$

and the errors in $\mathbf{H}_C$ are related to the errors in $\mathbf{Q}_C$ using the method described in Appendix II. Finally the covariance matrix is evaluated by correlating the errors in **N** and the corresponding part of $\mathbf{Q}_C$. The same procedure is followed for the row profiles. The resulting expression is

$$\begin{aligned} E[\hat{\boldsymbol{\Pi}}_N] &= [\mathbf{I} + S\sigma_N^2(\mathbf{T}_C^{-1}\boldsymbol{\Lambda}_C^{-1}\mathbf{T}_C)]^{-1}\left(\mathbf{I} + \frac{S}{2}\,\sigma_N^2(\mathbf{T}_C^{-1}\boldsymbol{\Lambda}_C^{-1}\mathbf{V}_C^T * \mathbf{V}_R\mathbf{T}_R\boldsymbol{\Pi}_N)\right)E[\boldsymbol{\Pi}_N] \\ &= \left(\mathbf{I} + \frac{W}{2}\,\sigma_N^2(\boldsymbol{\Pi}_N\mathbf{T}_C^{-1}\boldsymbol{\Theta}_C^{-1}\mathbf{U}_C^T * \mathbf{U}_R\boldsymbol{\Theta}_R^{-1}\mathbf{T}_R)\right)[\mathbf{I} + W\sigma_N^2(\mathbf{T}_R^{-1}\boldsymbol{\Lambda}_R^{-1}\mathbf{T}_R)]^{-1} \end{aligned} \tag{23}$$

where the substitution $\mathbf{T}_R = \mathbf{Z}_V$ is made for simplifying reasons and the '*' indicates that the inner products are taken over the admissible range, since the matrices involved are not conformable. It is seen that the contributions of the column and row profiles are not symmetrical for the cross-terms. There is also an important point about the practical evaluation of equation (23). The matrices $\mathbf{T}_R$ and $\mathbf{T}_C$ must be normalized in the same way to

fix the value of these cross-terms. It is tempting to speculate on this undesirable behavior: we think that this untractable expression arises because the relation between the errors in the reconstructed column and row responses is destroyed, since they result from the decomposition of two different matrices (compare equations (56) and (57)). We have not found a way of simplifying this result (note the presence of Θ_C and Θ_R in one term) and consequently we have no analogy for equations (9) and (17).

VARIANCE IN THE EIGENVALUES

In a previous paper[7] we derived variance expressions by performing error propagation on the standard eigenvalue problem of Lorber, i.e. equation (3d). This procedure leads to standard errors expressed in the abstract decomposition of **M**. In this paper we will derive standard errors expressed in the physical decomposition of **M** by performing error propagation on equation (7). The derivation (see Appendix III) is much shorter and the assumptions and approximations made are more transparent.

Lorber's method

Using the notation developed in the preceding section, equation (62) is rewritten as

$$\mathrm{var}(\hat{\pi}_{N,n}) = \tilde{\Psi}_{M,nn}(\sigma_N^2 + \pi_{N,n}^2\sigma_M^2) \tag{24}$$

This expression can be combined with equation (9) to give the mean squared error (MSE) as in equation (49).

Generalization of Sánchez and Kowalski

Taking the correlations between the elements of **M** and **M** + **N** into account gives

$$\mathrm{var}(\hat{\pi}_{M,n}) = \tilde{\Psi}_{Q,nn}(\sigma_M^2 - 2\pi_{M,n}\sigma_M^2 + \pi_{M,n}^2\sigma_Q^2) \tag{25}$$

This expression can be combined with equation (17) to give the MSE. Comparing equations (24) and (25), one finds that the variance is always reduced. This is an important result but difficult to prove for more general cases, i.e. heteroscedastic or correlated measurement noise. We will show in a later section (Variance in the unknown concentrations) that this variance reduction must be found in general.

Generalization of Wilson *et al.*

The variance expression for the generalization of Wilson *et al.* is easily derived by noting that in fact the standard eigenvalue problem of Lorber is solved if the unknown data matrix spans the factor space. There is, however, a difference in the evaluation of equation (24), since the pure quantities have to be replaced by the experimental values. Since the profiles are now reconstructed using information from both matrices, they are expected to be more precise. Therefore one should estimate the same variance but the estimate itself may be more precise.

VARIANCE IN THE BIAS-CORRECTED EIGENVALUES

Least squares estimators are characterized by their low variance. However, with errors in the regression matrix they are no longer unbiased.[6] Taking the structure of the errors in the

regression matrix into account, so-called measurement error models can be built that have a reduced MSE.[15] For rank annihilation we also have an errors-in-variables situation. The important question is therefore: is bias negligible under practical circumstances or (if this is not the case) can we construct better estimators? Before turning to complicated measurement error models (which is far beyond the scope of this paper), we will show that if the bias estimate is adequate, we can reduce the MSE by simply correcting for the bias. We will work out the principle for Lorber's method and the generalization of Sánchez and Kowalski. Assuming that $F \ll \min(S, W)$, dropping the subscripts for simplicity and replacing the true quantities by their experimental counterparts, equations (9) and (17) can be written as

$$\hat{\text{bias}}(\hat{\pi}) = \varepsilon\hat{\pi} \tag{26a}$$

$$\hat{\text{bias}}(\hat{\pi}) = \varepsilon(1 - 2\hat{\pi}) \tag{26b}$$

leading to the (almost) unbiased estimators

$$\hat{\pi}(\varepsilon) = \hat{\pi} - \varepsilon\hat{\pi} \tag{27a}$$

$$\hat{\pi}(\varepsilon) = \hat{\pi} + 2\varepsilon\hat{\pi} - \varepsilon \tag{27b}$$

which have an approximate $\hat{\text{MSE}}$ given by

$$\hat{\text{MSE}}(\hat{\pi}(\varepsilon)) = (1 + \varepsilon^2)\hat{\text{var}}(\hat{\pi}) + (\hat{\pi})^2\ \hat{\text{var}}(\varepsilon) \tag{28a}$$

$$\hat{\text{MSE}}(\hat{\pi}(\varepsilon)) = (1 + 4\varepsilon^2)\hat{\text{var}}(\hat{\pi}) + [1 + 4(\hat{\pi})^2]\hat{\text{var}}(\varepsilon) \tag{28b}$$

These values have to be compared with the $\hat{\text{MSE}}$ for the uncorrected estimator. We will illustrate the possible beneficial effect of the proposed bias correction for the generalization of Sánchez and Kowalski. Assume that $\hat{\pi} = 0{\cdot}1$, $\hat{\text{var}}(\hat{\pi}) = 10^{-4}$ and $\hat{\text{bias}}(\hat{\pi}) = 10^{-2}$. In this specific example the standard error and bias are equal (10%). Furthermore, the relative standard error in the estimated bias is assumed to be 10%. Working out (28b) shows that $\hat{\text{MSE}}(\hat{\pi}(\varepsilon)) \approx 10^{-4}$ compared with $\hat{\text{MSE}}(\hat{\pi}) = \sqrt{2} \times 10^{-4}$. The bias correction works very well in this specific example. However, it is important to note that the error in the estimated bias is an unknown quantity and the method should only be applied if it is reasonable to assume that this error is not underestimated.

CONFIDENCE LIMITS AND LIMITS OF DETECTION FOR THE ACTUAL CONCENTRATION RATIOS

The Monte Carlo simulations described later show that the sampling distribution of the *estimated* eigenvalues is approximately Gaussian. (This conclusion was arrived at by investigating percentiles and higher moments.) This means that the error estimates derived in the preceding sections can be used to construct confidence intervals for these eigenvalues in the usual way. It is, however, more interesting to have the confidence intervals for the *actual* concentration ratios. These intervals cannot be found by simply inverting the intervals for the estimated eigenvalues, since the errors depend on the corresponding eigenvalues. This problem is solved by transforming the eigenvalues in such a way that the errors become independent of the eigenvalues.[16] We will illustrate the necessary steps for the bias-corrected eigenvalues and restrict ourself to Lorber's method. We assume that the estimate for the bias is accurate. Then neglecting the last term in equation (28a) gives

$$\hat{\text{var}}(\hat{\pi}(\varepsilon)) = \hat{\text{MSE}}(\hat{\pi}(\varepsilon)) \cong (1 + \varepsilon^2)\Psi_{nn}\hat{\sigma}_{\text{M}}^2[1 + \hat{\pi}(\varepsilon)^2] \tag{29}$$

A transformation h is needed that stabilizes the variance[16]

$$\mathrm{var}(h(\hat{\pi}(\varepsilon))) = [h'(\hat{\pi}(\varepsilon))]^2 \, \mathrm{var}(\hat{\pi}(\varepsilon)) \equiv c \tag{30}$$

This transformation is found to be

$$h(\hat{\pi}(\varepsilon)) = \frac{\ln[\hat{\pi}(\varepsilon) + \sqrt{(\hat{\pi}(\varepsilon)^2 + 1)}]}{(1+\varepsilon^2)\Psi_{nn}^{1/2}\hat{\sigma}_{\mathrm{M}}} \tag{31}$$

Smooth transformations of a normally distributed random variable are also approximately normally distributed, so:[16]

$$\mathscr{L}\{h(\hat{\pi}(\varepsilon)) - h(\pi)\} = N(0,1) \tag{32}$$

The interval for the transformed variable with confidence $1-\alpha$ is given by

$$h(\hat{\pi}(\varepsilon)) - z(\alpha) \leqslant h(\pi) \leqslant h(\hat{\pi}(\varepsilon)) + z(\alpha) \tag{33}$$

and the confidence interval for the actual concentration ratio is found by back-transformation. It must be kept in mind that the terms contributing to the far right-hand side of equation (29) are not fixed. The confidence intervals derived are therefore only approximate. They may, however, be useful in practice, since no exact analytical results have been reported until now.

Since the detection of a component depends exclusively on the size of the eigenvalue, the detection limit for rank annihilation should be based on the standard error in the eigenvalue. Detection limits have been derived for multivariate calibration by Lorber[12] for the homoscedastic case and by Bauer *et al.*[17] for the heteroscedastic case. The methods proposed by these authors can be duplicated for rank annihilation, since the relevant expressions are very similar. However, if bias is important, it should be taken into account.

BIAS IN THE DETERMINED CONCENTRATIONS

For Lorber's method and the generalization of Wilson *et al.* we have, after inserting the relevant quantities in equation (41),

$$\frac{\mathrm{bias}(\hat{c}_{\mathrm{M},n})}{\tilde{c}_{\mathrm{M},n}} = \left(\frac{\sigma_{\hat{\pi}_{\mathrm{N},n}}}{\pi_{\mathrm{N},n}}\right)^2 \tag{34}$$

where $c_{\mathrm{M},n} = C_{\mathrm{M},nn}$. For the generalization of Sánchez and Kowalski we have to introduce a covariance term.[8] The result is

$$\frac{\mathrm{bias}(\hat{c}_{\mathrm{M},n})}{\tilde{c}_{\mathrm{M},n}} = 2\left(\frac{\sigma_{\hat{\pi}_{\mathrm{M},n}}}{1-\pi_{\mathrm{M},n}}\right)^2 \tag{35}$$

It follows that the bias is reduced if $c_{\mathrm{N},n} > (\sqrt{2}-1)c_{\mathrm{M},n}$. We have seen, however, that bias can be corrected to a certain extent and this result is therefore only important if the bias correction does not work.

VARIANCE IN THE UNKNOWN CONCENTRATIONS

By performing simple error propagation on the ratio of numbers and inserting the result for the eigenvalues, we find for the variance in the unknown concentration for Lorber's method

and the generalization of Wilson *et al.*

$$\mathrm{var}(\hat{c}_{\mathrm{M},n}) = \tilde{c}^2_{\mathrm{M},n}\left[\left(\frac{\sigma_{\hat{\pi}_{\mathrm{N},n}}}{\pi_{\mathrm{N},n}}\right)^2 + \left(\frac{\sigma_{\hat{c}_{\mathrm{N},n}}}{\tilde{c}_{\mathrm{N},n}}\right)^2\right]$$

$$= \tilde{c}^2_{\mathrm{M},n}\left[(\mathrm{S\tilde{E}L})^2_{\mathrm{X},n}(\mathrm{S\tilde{E}L})^2_{\mathrm{Y},n}\left[\left(\frac{\sigma_{\mathrm{N}}}{\tilde{c}_{\mathrm{N},n}}\right)^2 + \left(\frac{\sigma_{\mathrm{M}}}{\tilde{c}_{\mathrm{M},n}}\right)^2\right] + \left(\frac{\sigma_{\hat{c}_{\mathrm{N},n}}}{\tilde{c}_{\mathrm{N},n}}\right)^2\right] \quad (36)$$

Equation (36) shows explicitly the dependence of the estimated error on the expected error sources: the error in the concentration of the calibration sample and the error in the response matrices. The dependence of the error on the value of the unknown concentration reveals the extrapolating character of calibration with only one calibration sample. For the generalization of Sánchez and Kowalski one finds that the contribution of the standard error in the eigenvalues is composed of two parts:

$$\mathrm{var}(\hat{c}_{\mathrm{M},n}) = \tilde{c}^2_{\mathrm{M},n}\left[\left(\frac{1}{1-\pi_{\mathrm{M},n}}\right)^2\left(\frac{\sigma_{\hat{\pi}_{\mathrm{M},n}}}{\pi_{\mathrm{M},n}}\right)^2 + \left(\frac{\sigma_{\hat{c}_{\mathrm{N},n}}}{\tilde{c}_{\mathrm{N},n}}\right)^2\right] \quad (37)$$

The variance in the unknown concentration must, however, remain the same, since we have not added information to the data. It follows that the relative standard error in the eigenvalues is reduced by an amount

$$(1-\pi_{\mathrm{M},n})^2 \cong \left(\frac{\tilde{c}_{\mathrm{N},n}}{\tilde{c}_{\mathrm{N},n}+\tilde{c}_{\mathrm{M},n}}\right)^2 \quad (38)$$

Since no assumptions were made on the source of the standard error in the eigenvalues, this variance reduction should be found in general, i.e. also for heteroscedastic and correlated noise.

SIMULATIONS

We simulated three-component systems by multiplying Gaussian elution profiles by experimentally obtained UV spectra for myoglobin (M), α-chymotrypsin (α-C) and carbon anhydrase (C).[18] The spectra and elution profiles are shown in Figures 1 and 2 respectively. The spectra were normalized to unit length in order to make the variance of the individual components proportional to the square of the peak height. The peak positions (20, 25 and 30) and standard deviations of the peaks (10, 10 and 10) are chosen in such a way that the chromatographic resolution, defined as $R_{ij} = (t_j - t_i)/[(w_{1/2})_i + (w_{1/2})_j]$, has a value of approximately 0·1. The normalized inner products and correlation coefficients are given in Table 1. These similarity measures are to be compared with the numbers obtained in the recognition step of the rank annihilation procedure. It is seen that although the 'chromatography' is very bad, the overlap of the spectra is much larger. The numbers in Table 1 can, however, not be used in general for obtaining an impression about multicollinearities and therefore we have additionally compiled Lorber's selectivities in Table 2, since these quantities are directly related to the errors in the eigenvalues. It is seen that the selectivity of the elution profile of α-C is only twice as large as the selectivity for the spectral mode. This is in sharp contrast with the numbers in Table 1. We added Gaussian noise with a constant standard deviation of 0·05 mAU to both data matrices, i.e. **N** and **M**. This value corresponds to the average residual standard deviation found for three commercially available diode array detectors (HP1040A, PU4021 and Waters 990). This combination of overlap and signal-to-noise ratio is comparable with the most difficult level chosen in Reference

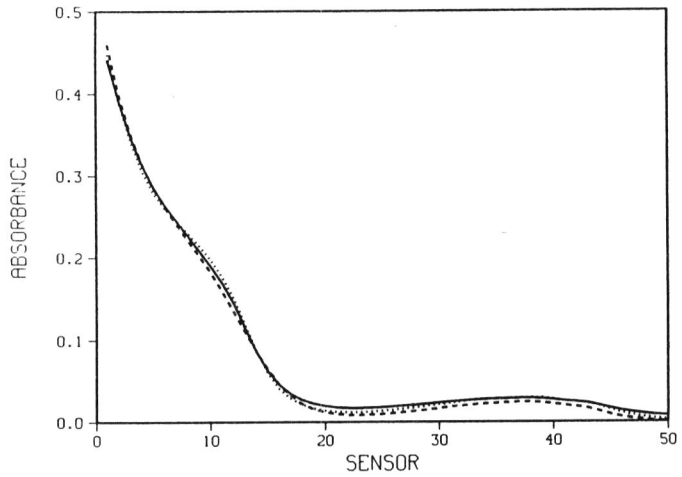


Figure 1. Normalized UV spectra for myoglobine (———), α-chymotrypsin (— — —) and carbon anhydrase (. . .)

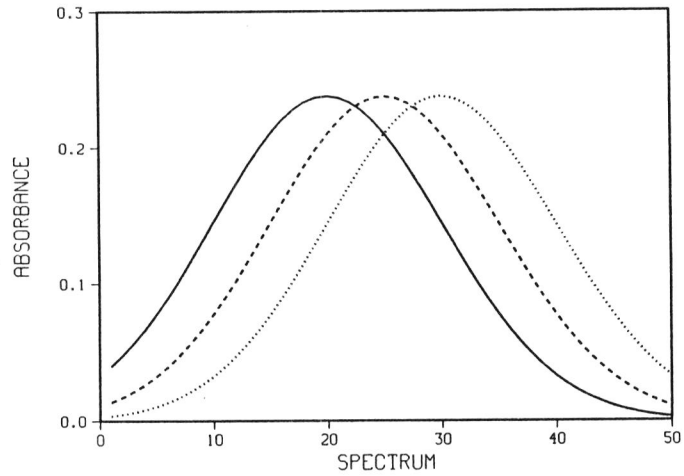


Figure 2. Normalized simulated elution profiles for myoglobine (———), α-chymotrypsin (— — —) and carbon anhydrase (. . .)

18. There the largest chromatographic overlap corresponded to a value $R_{ij} = 0{\cdot}25$ and the standard deviation for the noise was chosen as 0·12 mAU. This level proved to be very difficult for curve resolution with iterative target-testing factor analysis (ITTFA). The simulation parameters are summarized in Table 3.

Table 1. Inner product (right upper corner) and linear correlation coefficient (left lower corner) for UV spectra and elution profiles. Spectra and elution profiles are normalized

	UV spectrum			Elution profile		
Standard	M	α-C	C	M	α-C	C
M	1	0·9988	0·9996	1	0·94	0·78
α-C	0·9996	1	0·9988	0·81	1	0·94
C	0·9995	0·9989	1	0·32	0·81	1

Table 2. Selectivities of UV spectra and elution profiles

Standard	UV spectrum	Elution profile
M	0·0282	0·1577
α-C	0·0465	0·0861
C	0·0282	0·1580

Table 3. Simulation parameters

	M	α-C	C
Peak position	20	25	30
Standard deviation peak	10	10	10
Number of spectra		50	
Number of wavelengths		50	
$\sigma_N = \sigma_M$ (mAU)		0·05	

RESULTS AND DISCUSSION

This section will be divided into two parts. In the first part we will compare the bias and variance for the different models, i.e. Lorber's method, the generalization of Sánchez and Kowalski and the generalization of Wilson *et al.* It is important to realize that even for data that can be analysed by all methods, the results may be entirely different. In the second part we will study the effect of the dimensionality of the factor space on the errors in the eigenvalues.

Comparison of error estimates for different models

The developed error estimates are tested by constructing a worst case example: a large spectral and chromatographic overlap is combined with the presence of a diluted component. In Table 4 we give the peak heights encountered during these simulations. M and C are held constant while the peak of α-C is lowered from 200 to 5 mAU. Since both samples contain the same components, rank annihilation should work with the factor space of **N**, **M**, **N** + **M** and

Table 4. Peak heights in mAU for standard and unknown sample

Sample	M	α-C	C
Standard	300	300	300
Unknown	100	200–5	300

the adjoined matrices respectively. The analysis with the factor space of **M**, however, already gives bad results for the second dilution, since the third factor is spoiled by a large embedded error. Therefore we will only give the results for the other decompositions. We will restrict the discussion to the results for the diluted component, since the results obtained for the highest dilution are representative of the results obtained for the major components at all dilutions.

Table 5 summarizes the quantitative solution using the factor space of **N**, i.e. Lorber's method (reversed). The first two columns give the theoretical values for peak height and eigenvalue. The next four columns give the estimates for the eigenvalue, bias and standard error based on one simulation. The last three columns give the expected values for these quantities based on averaging the eigenvalues obtained for 10^4 simulations. Using 10^4 simulations gives a sample mean for the eigenvalues from which the bias can be accurately estimated. Since we only add noise to the data matrices, the bias is calculated with respect to the true value. (The effect of noise in the concentrations is trivial.) All error estimates are divided by the true value in order to make them comparable. It is seen that the highest dilution gives predicted results for bias and variance that are in excellent agreement with the 'experimental' results. The simple error theory we propose seems to work very well for major components. For the other cases one finds a better agreement for the standard errors than for the bias. This fact can be explained by the additional approximation needed to obtain comprehensive expressions for the bias. Since the contribution of the bias to the MSE is very small, this is of minor importance. Considering only the diagonal elements of the bias correction seems to have a marginal effect on the outcome in all cases. The detection limit is indicated by the standard error to be very near $\tilde{\pi} = 0{\cdot}033$. Since rank annihilation can give a good qualitative solution if there is only one component absent, the predicted standard error

Table 5. Summary of quantitative solution for diluted component: decomposition of **N**. The numbers printed bold indicate the limit of detection

		One simulation				10^4 simulations		
$\tilde{H}$	$\tilde{\pi}$	$\hat{\pi}$	$\hat{b}/\tilde{\pi}$[a] (%)	$\hat{b}/\tilde{\pi}$[b] (%)	$\hat{\sigma}/\tilde{\pi}$[c] (%)	$E[\tilde{\pi}]$	$E[b]/\tilde{\pi}$ (%)	$E[\sigma]/\tilde{\pi}$ (%)
200	0·667	0·641	−1·42	−1·42	1·79	0·657	−1·41	1·78
30	0·100	0·096	−1·42	−1·40	9·93	0·099	−1·32	9·92
20	0·067	0·045	−1·09	−1·06	15·7	0·066	−1·54	14·7
10	**0·033**	**0·039**	**−1·77**	**−1·71**	**29·1**	**0·033**	**−2·01**	**29·5**
5	0·017	−0·001	+0·81	+0·93	64·2	0·016	−1·49	58·3

[a] Estimated by equation (8).
[b] Estimated by equation (9).
[c] Estimated by equation (24).

is correct, since the influence of the eigenvalue itself is negligible. For the lowest dilution a negative concentration is found. This explains why the predicted bias has changed sign. We were able to locate the detection limit using the predicted standard error. An alternative is to perform target testing on the unknown data matrix. Each target-testing procedure has its own limit of detection and it would be interesting to see whether the results are consistent. It is evident that future research should be focused on reliable quantitation in the neighbourhood of the detection limit.

Table 6 summarizes the quantitative solution using the factor space of $\mathbf{N}+\mathbf{M}$, i.e. the generalization of Sánchez and Kowalski. First we notice that again the agreement is excellent for the first row. Both variance and bias are reduced substantially with respect to the previous case. For the other dilutions we find a variance reduction that is unimportant compared with the increase in bias.

Table 7 summarizes the quantitative solution using the factor space of the adjoined matrices, i.e. the generalization of Wilson *et al.* First we notice that the agreement between predicted and measured errors is generally not so good as in the previous cases. This is probably caused by the larger contribution of the eigenvalue to the error estimates. The standard error in the

Table 6. Summary of quantitative solution for diluted component: decomposition of $\mathbf{N}+\mathbf{M}$. The numbers printed bold indicate the limit of detection

		One simulation				10^4 simulations		
$\tilde{H}$	$\tilde{\pi}$	$\hat{\pi}$	$\hat{b}/\tilde{\pi}$[a] (%)	$\hat{b}/\tilde{\pi}$[b] (%)	$\hat{\sigma}/\tilde{\pi}$[c] (%)	$E[\pi]$	$E[b]/\tilde{\pi}$ (%)	$E[\sigma]/\tilde{\pi}$ (%)
200	0·400	0·396	+0·28	+0·28	1·10	0·401	+0·24	1·07
30	0·091	0·098	+8·35	+8·34	8·57	0·099	+9·06	8·53
20	0·063	0·056	+14·8	+14·8	14·5	0·072	+14·7	13·1
10	**0·032**	**0·051**	**+26·2**	**+26·2**	**26·2**	**0·043**	**+31·8**	**27·1**
5	0·016	0·003	+15·9	+15·9	63·8	0·028	+67·8	55·2

[a] Estimated by equation (16).
[b] Estimated by equation (17).
[c] Estimated by equation (25).

Table 7. Summary of quantitative solution for diluted component: decomposition of $(\mathbf{N}\,|\,\mathbf{M})$ and $\binom{\mathbf{N}}{\mathbf{M}}$. The numbers printed bold indicate the limit of detection

		One simulation			10^4 simulations		
$\tilde{H}$	$\tilde{\pi}$	$\hat{\pi}$	$\hat{b}/\tilde{\pi}$[a] (%)	$\hat{\sigma}/\tilde{\pi}$[b] (%)	$E[\pi]$	$E[b]/\tilde{\pi}$ (%)	$E[\sigma]/\tilde{\pi}$ (%)
200	1·50	1·47	−0·28	1·72	1·50	−0·02	1·81
30	10·0	9·37	+1·41	8·33	10·1	+1·22	10·5
20	15·0	15·3	+2·87	15·4	15·4	+2·67	16·7
10	**30·0**	**29·3**	**+5·39**	**28·0**	**34·6**	**+15·3**	**137**
5	60·0	408	$+3\times10^2$	3×10^3	122	$+1\times10^2$	1×10^3

[a] Estimated by equation (23).
[b] Estimated by equation (24).

Table 8. Comparison of relative $(MSE)^{1/2}$ for eigenvalue of diluted component for various decompositions. The reported values (in %) are the Monte Carlo estimates. The numbers printed bold indicate the limit of detection

H	**N**	**N + M**	**(N | M) and $\binom{N}{M}$**
200	2·27	1·10	1·81
30	10·0	12·4	10·6
20	14·8	19·7	16·9
10	**29·6**	**41·8**	**138**
5	58·3	87·4	1×10^3

first row is comparable with the standard error found for Lorber's method. The bias is, however, much smaller, which is inferred from the measured value in the last column. For the other dilutions the variance is about the same as for the other models and the bias takes values that are comparable with those found for Lorber's method. Here the detection limit is accompanied by a much more drastic increase in the variance and bias. This extreme behavior is also probably caused by the large size of the eigenvalues.

In order to compare the quantitative results for the three models, we have calculated the MSE based on the results for the extensive simulations. (The simulations must be expected to approach the true values.) The values for the MSE are given in Table 8. For the highest dilution, i.e. for major components, the generalization of Sánchez and Kowalski comes out best. For the other dilutions Lorber's method gives the lowest MSE. The results for the generalization of Sánchez and Kowalski can, however, easily be improved by applying the proposed bias correction. The bias estimates seem to be reliable enough (error less then 10%) for all cases above the detection limit. The second dilution has in fact already been treated as a simple example in the section on bias correction with slightly altered numbers. Using the bias correction for the first three dilutions will finally give the best results for the generalization of Sánchez and Kowalski.

Finally we give in Table 9 the normalized inner products of the reconstructed profiles and the corresponding standards near the detection limit of the diluted component. It is seen that the recognition is excellent even for the diluted component, which is somewhat misleading, since the eigenvalue carries such a large error. It is seen that the spectra are slightly more precise than the elution profiles, as expected. The reconstructed profiles seem to be best for the generalization of Wilson *et al.*, but the differences are very small.

Table 9. Summary of qualitative solution near limit of detection ($H = 10$): normalized inner products of reconstructed UV spectra and elution profiles and their corresponding standards

		Decomposition of **N**		Decomposition of **N + M**		Decomposition of **(N | M) and $\binom{N}{M}$**	
Component	Standard	UV spectrum	Elution profile	UV spectrum	Elution profile	UV spectrum	Elution profile
1	M	1·00000	0·99994	1·00000	0·99965	1·00000	0·99995
2	α-C	0·99999	0·99998	0·99998	0·99992	1·00000	0·99999
3	C	1·00000	0·99994	1·00000	0·99998	1·00000	0·99999

Effect of dimensionality on the error estimates

The choice of the true dimensionality is in general not a trivial problem. Therefore it is interesting to investigate how the error estimates behave when we underspan or overspan the factor space. Another reason for performing these simulations is the following. With collinear data a good alternative to multiple linear regression (MLR) is principle component regression (PCR), i.e. replace the original regressors by a selection of PCs. Underfactoring automatically introduces a bias, but since the variance is reduced at the same time very often the MSE is decreased. Thus there is a good reason for deliberately underfactoring the model. We will confine the discussion to Lorber's method.

We have simulated a system where the unknown sample contains three components and the calibration sample only one. The peak heights are given in Table 10. With only one component in the calibration sample the identity of the eigenvalue is easily established even if the profiles are distorted. In practice, distorted profiles may resemble the profiles of other components if the library is large enough, leading to a false identification. The results are given in Table 11. It is seen, that the estimated eigenvalues are severely biased if the model is underfactored. (An instructive example is also given by Malinowski and Howery.[19]) Predicted variance and bias (resulting from measurement noise) are extremely small. The predicted variance is in good agreement with the Monte Carlo result. (It should be kept in mind that the symbols in equation (10) no longer correspond to true profiles for the transformed noise factors.) We have no equivalent values for the bias, since the bias is calculated with respect to the true value. The resulting bias is therefore automatically the bias resulting from underfactoring. Increasing the dimensionality from three to five gives only small changes in the eigenvalues, variance and bias respectively. Since the maximum error in the predicted MSE (3·7%) is found to be 0·5% for

Table 10. Peak heights in mAU for standard and unknown sample

Sample	M	α-C	C
Standard	0	300	0
Unknown	100	200	300

Table 11. Summary of quantitative solution for analyte of interest for varying dimension of model: decomposition of **M**. The numbers printed bold indicate the correct dimension

	One simulation			10^4 simulations		
Dimension	$\tilde{\pi}$	$\hat{b}/\tilde{\pi}$[a] (%)	$\hat{\sigma}/\tilde{\pi}$[b] (%)	$E[\pi]$	$E[b]/\tilde{\pi}$ (%)	$E[\sigma]/\tilde{\pi}$ (%)
1	0·51	−0·0	$1·5 \times 10^{-3}$	0·51	—	$2·3 \times 10^{-3}$
2	0·92	−0·1	0·2	0·92	—	0·3
3	**1·42**	**−2·8**	**1·5**	**1·45**	**−3·4**	**1·6**
4	1·47	−3·8	1·8	1·45	−3·3	1·6
5	1·47	−3·5	1·7	1·45	−3·3	1·6

[a] Estimated by equation (8).
[b] Estimated by equation (24).

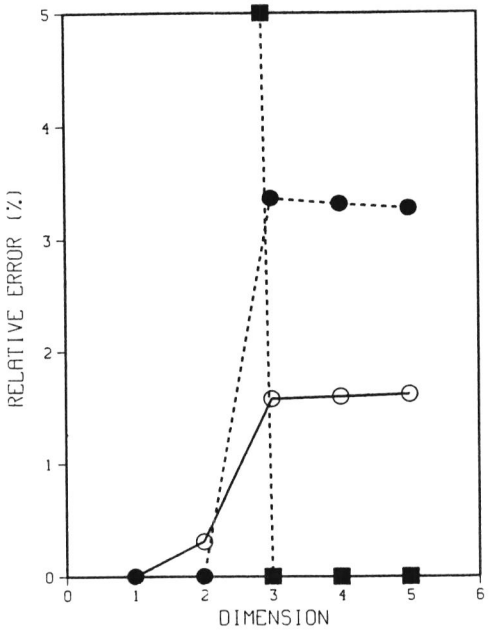


Figure 3. Contribution to mean squared error in the eigenvalues: standard error (○———○), bias from under factoring (■ ——— ■) and bias from measurement noise (● ——— ●)

the three-dimensional model, it seems that the predictions work equally well for the overdimensioned model. The different contributions to the MSE are depicted in Figure 3. For rank annihilation we always have variance and bias as a result of measurement noise. We usually find that the underfactoring bias is much larger than the reduction in the other contributions. This means that a PCR version of rank annihilation will only give an improved MSE under very extreme circumstances, since the examples worked out here were constructed to represent difficult cases.

CONCLUSIONS

Expressions have been derived for predicting the bias and variance in the eigenvalues of rank annihilation. In order to derive these expressions, extensive use has been made of analogies between a reformulation of the characteristic eigenvalue problem and the prediction equations of univariate and multivariate calibration. An important difference between the derivation of the bias and variance expressions is the fact that in order to derive the bias expressions, additional assumptions are needed. This is not the case for the variance expressions, where simple first-order error propagation is applied. The additional approximation is probably the reason why in general the bias estimate is found to be less accurate than the variance estimate. A bias correction technique has been proposed that effectively eliminates the bias if the error in the bias is not too large. The simulations discussed in this paper show that probably even an error in the predicted bias as large as 10% can be tolerated. Because the bias correction

technique introduces uncertainties, it should, however, be used with care. The necessary steps for establishing confidence limits for the actual concentration ratios have been indicated. It has been demonstrated theoretically and by performing simulations that large differences may occur in variance as well as bias for Lorber's method, the generalization of Sánchez and Kowalski and the generalization of Wilson *et al.* The generalization of Sánchez and Kowalski should always give the smallest variance. However, the bias increases quickly for this method when the eigenvalue becomes small. The best results are obtained if the concentrations are approximately equal in the unknown and calibration sample (large variance reduction and almost no bias). Lorber's method and the generalization of Wilson *et al.* behave similarly with respect to variance. Furthermore, both methods display only a moderate increase in bias when approaching the detection limit. The derived expressions also perform well if the dimension of the factor space is not correct. It is found that the bias resulting from under factoring is usually much larger than the decrease in bias and variance resulting from measurement noise. This means that a PCR version of rank annihilation will only work in very exceptional cases. A treatment of other subjects, e.g. wavelength selection, should be equally straightforward using the reformulation of the eigenvalue problem.

APPENDIX I: THE INFLUENCE OF RANDOM NOISE IN UNIVARIATE AND MULTIVARIATE CALIBRATION

Some important expressions for bias and variance are summarized for univariate and multivariate calibration. These expressions are useful for the derivation and interpretation of similar results for bilinear calibration. The usual simplifying assumptions about noise are made. These simplifications are not necessary and additional results can be found in the literature.[6,8,10]

Univariate calibration (scalar data)

With errors in both the instrumental response, denoted by r, and the sensitivity of the instrument, denoted by s, the prediction equation is

$$\tilde{r} + \delta r = (\tilde{s} + \delta s)(\tilde{c} + \delta c) \tag{39}$$

where c is the unknown concentration estimated as

$$\hat{c} = s^{-1} r \tag{40}$$

If the errors in r and s are sufficiently small, the bias[8] and variance are given by

$$\text{bias}(\hat{c}) = \tilde{s}^{-2} \tilde{c} \sigma_s^2 \tag{41}$$

$$\text{var}(\hat{c}) = \tilde{s}^{-2} (\sigma_r^2 + \tilde{c}^2 \sigma_s^2) \tag{42}$$

and combined in the MSE as

$$\text{MSE}(\hat{c}) = \tilde{s}^{-2} [\sigma_r^2 + \tilde{c}^2 \sigma_s^2 (1 + \tilde{s}^{-2} \sigma_s^2)] \tag{43}$$

In equation (43) the relative importance of variance and bias as first- and second-order properties of an estimator become apparent.

Multivariate calibration (vector data)

With errors in both the $J\times 1$ response vector, denoted by $\mathbf{r}$, and the $J\times K$ matrix of sensitivities, denoted by $\mathbf{S}$, the prediction equation is

$$\tilde{\mathbf{r}}+\delta\mathbf{r}=(\tilde{\mathbf{S}}+\delta\mathbf{S})(\tilde{\mathbf{c}}+\delta\mathbf{c}) \tag{44}$$

where $\mathbf{c}$ is the $K\times 1$ vector of unknown concentrations estimated as

$$\hat{\mathbf{c}}=\mathbf{S}^{+}\mathbf{r} \tag{45}$$

With errors in $\mathbf{S}$ the estimated concentration is biased, because the pseudoinverse of a stochastic matrix is involved in the calculation. The expectation value of the pseudoinverse matrix $\mathbf{S}^{+}$ is given (for the full-rank case) by[6]

$$\begin{aligned} E[\mathbf{S}^{+}] &= E[(\tilde{\mathbf{S}}+\delta\mathbf{S})^{+}] \\ &= [[(\tilde{\mathbf{S}}+\delta\mathbf{S})^{\mathrm{T}}(\tilde{\mathbf{S}}+\delta\mathbf{S})]^{-1}(\tilde{\mathbf{S}}+\delta\mathbf{S})^{\mathrm{T}}] \\ &= (\tilde{\mathbf{S}}^{\mathrm{T}}\tilde{\mathbf{S}}+E[\delta\mathbf{S}^{\mathrm{T}}\,\delta\mathbf{S}])^{-1}E[\tilde{\mathbf{S}}^{\mathrm{T}}+\delta\mathbf{S}^{\mathrm{T}}] \\ &= [\mathbf{I}+(\tilde{\mathbf{S}}^{\mathrm{T}}\tilde{\mathbf{S}})^{-1}E[\delta\mathbf{S}^{\mathrm{T}}\,\delta\mathbf{S}]]^{-1}\tilde{\mathbf{S}}^{+} \end{aligned} \tag{46}$$

which can be rewritten as

$$\begin{aligned} E[\mathbf{S}^{+}] &= (\mathbf{I}+J\tilde{\boldsymbol{\Phi}}\sigma_{\mathrm{S}}^{2})^{-1}\tilde{\mathbf{S}}^{+} \\ &= [\mathbf{I}-(J-K-1)\tilde{\boldsymbol{\Phi}}\sigma_{\mathrm{S}}^{2}]\tilde{\mathbf{S}}^{+} \end{aligned} \tag{47}$$

where the substitutions $(\mathbf{S}^{\mathrm{T}}\mathbf{S})^{-1}=\boldsymbol{\Phi}$ and $E[\delta\mathbf{S}^{\mathrm{T}}\,\delta\mathbf{S}]=J\sigma_{\mathrm{S}}^{2}\mathbf{I}$ have been made. The final expression is corrected for degrees of freedom.[6] The variance in the nth concentration is given by[6]

$$\mathrm{var}(\hat{c}_n)=\tilde{\boldsymbol{\Phi}}_{nn}(\sigma_{\mathrm{r}}^{2}+\|\tilde{c}\|^{2}\sigma_{\mathrm{S}}^{2}) \tag{48}$$

For small errors the off-diagonal elements of the matrix in square brackets in equation (47) contribute very little. Consequently the bias in the determined concentrations can be estimated very well by taking only the diagonal elements into account. This additional approximation leads to the following expression for the MSE:

$$\mathrm{MSE}(\hat{c}_n)=\tilde{\boldsymbol{\Phi}}_{nn}\{\sigma_{\mathrm{r}}^{2}+\|\tilde{c}\|^{2}\sigma_{\mathrm{S}}^{2}+(J-K-1)^{2}\Phi_{nn}\tilde{c}_n^{2}\sigma_{\mathrm{S}}^{4}\} \tag{49}$$

It is seen that both bias and variance are primarily influenced by the same quantity, i.e. the corresponding diagonal element of the matrix $\boldsymbol{\Phi}$. This number is usually referred to as the 'variance factor',[11] but it follows that it could also be called the 'bias factor' or more generally the 'MSE factor'.

As mentioned by other researchers, some useful transcriptions are possible for this quantity. Summarizing their results gives

$$\Phi_{nn}=(\mathbf{S}^{\mathrm{T}}\mathbf{S})_{nn}^{-1} \tag{50a}$$

$$=\|\mathbf{S}_{n\text{-row}}^{+}\|^{2} \tag{50b}$$

$$=\sum_{p=1}^{K}\left(\frac{V_{np}}{\theta_{\mathrm{S},p}}\right)^{2} \tag{50c}$$

$$=(\mathrm{SEN})_n^{-2} \tag{50d}$$

The identity of (50a) and (50b) has been proved by Bauer *et al.*,[10] thereby establishing the relation between Lorber's figures of merit[12] and well-known statistical results. Equation (50c)

is based on the SVD of **S**: $\mathbf{S} = \mathbf{U}_S\boldsymbol{\Theta}_S\mathbf{V}_S^T$. It has been used to define so-called variance inflation factors. These factors are useful as a diagnostic for identifying multicollinearities in **S**. It is a difficult problem as to whether column scaling of **S** should be applied.[20,21] Equation (50d) represents Lorber's definition of multivariate sensitivity. Inserting this symbol in the relevant expressions supplies the correspondence between univariate and multivariate calibration.

APPENDIX II: DERIVATION OF VARIANCE IN THE RECONSTRUCTED PROFILES

First the reconstructed column profiles are written as

$$\tilde{\mathbf{H}} + \delta\mathbf{H} = (\tilde{\mathbf{A}} + \delta\mathbf{A})(\tilde{\mathbf{T}} + \delta\mathbf{T}) \tag{51}$$

If the errors in the transformation matrix, $\delta\mathbf{T}$, are small, the errors in the reconstructed column profiles can be approximated by[22]

$$\delta\mathbf{H} = \delta\mathbf{A}\tilde{\mathbf{T}} \tag{52}$$

In a previous paper[7] we showed that the errors in the scores can be approximated by

$$\delta\mathbf{A} = \tilde{\mathbf{U}}\ \delta\boldsymbol{\Theta} \tag{53}$$

The resulting standard errors were in good agreement with the Monte Carlo values, but using these expressions in order to derive standard errors in the eigenvalues yielded inconsistent results. Thus using this additional approximation may result in a prediction of the bias that is less accurate than the prediction of the variance. Combining equations (52) and (53) and using $\delta\boldsymbol{\Theta} = \tilde{\mathbf{U}}^T\ \delta\mathbf{M}\tilde{\mathbf{V}}$[7] gives

$$\begin{aligned} E[\delta\mathbf{H}^T\ \delta\mathbf{H}] &= E[\tilde{\mathbf{T}}^T\ \delta\mathbf{A}^T\ \delta\mathbf{A}\tilde{\mathbf{T}}] \\ &= \tilde{\mathbf{T}}^T E[\delta\boldsymbol{\Theta}^T\tilde{\mathbf{U}}^T\tilde{\mathbf{U}}\ \delta\boldsymbol{\Theta}]\tilde{\mathbf{T}} \\ &= \tilde{\mathbf{T}}^T\tilde{\mathbf{V}}^T E[\delta\mathbf{M}^T\tilde{\mathbf{U}}\tilde{\mathbf{U}}^T\ \delta\mathbf{M}]\tilde{\mathbf{V}}\tilde{\mathbf{T}} \\ &= S\sigma_M^2(\tilde{\mathbf{T}}^T\tilde{\mathbf{T}}) \end{aligned} \tag{54}$$

The reconstructed column profiles are correlated and consequently the covariance matrix is not diagonal. This is a difference compared with the multivariate situation. Furthermore,

$$(\tilde{\mathbf{H}}^T\tilde{\mathbf{H}})^{-1} = \tilde{\mathbf{T}}^{-1}\tilde{\boldsymbol{\Lambda}}^{-1}\tilde{\mathbf{T}}^{-T} \tag{55}$$

where $\boldsymbol{\Lambda} = \boldsymbol{\Theta}^2$. Combining equations (54) and (55) leads to

$$(\tilde{\mathbf{H}}^T\tilde{\mathbf{H}})^{-1}E[\delta\mathbf{H}^T\ \delta\mathbf{H}] = S\sigma_M^2(\tilde{\mathbf{T}}^{-1}\tilde{\boldsymbol{\Lambda}}^{-1}\tilde{\mathbf{T}}) \tag{56}$$

and using equivalent expressions for **Y** shows that

$$E[\delta\mathbf{Y}^T\ \delta\mathbf{Y}](\tilde{\mathbf{Y}}^T\tilde{\mathbf{Y}})^{-1} = W\sigma_M^2(\tilde{\mathbf{T}}^{-1}\tilde{\boldsymbol{\Lambda}}^{-1}\tilde{\mathbf{T}}) \tag{57}$$

It follows that the contributions from the separate modes to the bias in the eigenvalues are intimately related. They are in fact identical for square matrices. It is important to note that this result is derived without any reference to rank annihilation. Therefore it should be valid for any method that gives precise estimates for the transformation matrix **T**. An immediate consequence of equations (56) and (57) is that the overlap and precision of the reconstructed profiles are complementary. This is important if these profiles are to be used for e.g. identification purposes, as is the case with rank annihilation. For the simulations discussed in

this paper the spectral mode is much more unselective then the chromatographic mode. If the measurement error were equally dispersed over the two modes, the reconstructed spectra would probably become useless for identification.

APPENDIX III: DERIVATION OF VARIANCE IN THE EIGENVALUES, EXPRESSED IN THE PHYSICAL DECOMPOSITION OF **M**

We will only discuss Lorber's method and drop all subscripts to simplify the notation. First equation (7) is rewritten as

$$(\tilde{\mathbf{H}} + \delta\mathbf{H})^{+}(\tilde{\mathbf{N}} + \delta\mathbf{N})(\tilde{\mathbf{Y}} + \delta\mathbf{Y})^{\ddagger} = \boldsymbol{\Pi} + \delta\boldsymbol{\Pi} \tag{58}$$

Note that the spread around $\boldsymbol{\Pi}$ instead of $\tilde{\boldsymbol{\Pi}}$ is estimated. The error in the pseudoinverse of a matrix $\mathbf{A} + \delta\mathbf{A}$ can be approximated by $(\mathbf{A} + \delta\mathbf{A})^{+} = \mathbf{A}^{+} + \delta(\mathbf{A}^{+})$.[24] This results in

$$\delta\boldsymbol{\Pi} = \tilde{\mathbf{H}}^{+}\ \delta\mathbf{N}\tilde{\mathbf{Y}}^{\ddagger} + \delta(\mathbf{H}^{+})\tilde{\mathbf{N}}\tilde{\mathbf{Y}}^{\ddagger} + \tilde{\mathbf{H}}^{+}\tilde{\mathbf{N}}\ \delta(\mathbf{Y}^{\ddagger}) \tag{59}$$

Using $\delta(\mathbf{A}^{+}) = -\mathbf{A}^{+}\delta\mathbf{A}\ \mathbf{A}^{+}$ (see Reference 10) and commutivity for diagonal matrices yields

$$\delta\boldsymbol{\Pi} = \tilde{\mathbf{H}}^{+}\ \delta\mathbf{N}\tilde{\mathbf{Y}}^{\ddagger} - \boldsymbol{\Pi}\tilde{\mathbf{H}}^{+}\ \delta\mathbf{M}\tilde{\mathbf{Y}}^{\ddagger} \tag{60}$$

leading to the following statistical error, if the measurement noise is uncorrelated:

$$\mathrm{var}(\hat{\pi}_n) = \sum_{i=1}^{r}\sum_{k=1}^{c}(\tilde{H}^{+}_{ni})^2(\tilde{Y}^{\ddagger}_{kn})^2(\sigma^2_{\mathbf{N}ik} + \pi^2_n\sigma^2_{\mathbf{M}ik}) \tag{61}$$

This result simplifies considerably if the measurement noise is homoscedastic:

$$\mathrm{var}(\hat{\pi}_n) = \| \tilde{\mathbf{H}}^{+}_{n\text{-row}} \|^2 \| \tilde{\mathbf{Y}}^{\ddagger}_{n\text{-col}} \|^2(\sigma^2_{\mathbf{N}} + \pi^2_n\sigma^2_{\mathbf{M}}) \tag{62}$$

leading through equation (10) to equation (24). The corresponding equation for the generalization of Sánchez and Kowalski is derived by working out (62) after making the necessary substitutions.

REFERENCES

1. C.-N. Ho, G. D. Christian and E. Davidson, *Anal. Chem.* **50**, 1108 (1978).
2. N. M. Faber, L. M. C. Buydens and G. Kateman, *J. Chemometrics*, **8**, 147 (1994).
3. A. Lorber, *Anal. Chim. Acta*, **164**, 293 (1983).
4. E. Sánchez and B. R. Kowalski, *Anal. Chem.* **58**, 496 (1986).
5. B. E. Wilson, E. Sánchez and B. R. Kowalski, *J. Chemometrics*, **3**, 493 (1989).
6. S. D. Hodges and P. G. Moore, *Appl. Stat.* **21**, 185 (1972).
7. N. M. Faber, L. M. C. Buydens and G. Kateman, *J. Chemometrics*, **7**, 495 (1993).
8. M. G. Moran and B. R. Kowalski, *Anal. Chem.* **56**, 562 (1984).
9. H. L. Gray and W. R. Schucany, *The Generalized Jackknife Statistic*, Marcel Dekker, New York (1972).
10. G. Bauer, W. Wegscheider and H. M. Ortner, *Spectrochim. Acta B*, **46**, 1185 (1991).
11. J. Mandel, *J. Res. NBS*, **90**, 465 (1985).
12. A. Lorber, *Anal. Chem.* **58**, 1167 (1986).
13. C.-N. Ho, G. D. Christian and E. R. Davidson, *Anal Chem.* **52**, 1071 (1980).
14. C. J. Appelof and E. R. Davidson, *Anal. Chim. Acta*, **146**, 9 (1983).
15. W. A. Fuller, *Measurement Error Models*, Wiley, New York (1987).
16. P. J. Bickel and K. A. Doksum, *Mathematical Statistics*, Holden-Day, San Fransisco, CA (1977).
17. G. Bauer, W. Wegscheider and H. M. Ortner, *J. Anal. Chem.* **340**, 135 (1991).
18. B. G. M. Vandeginste, F. Leyten, M. Gerritsen, J. W. Noor and G. Kateman, *J. Chemometrics*, **1**, 57 (1987).

19. E. R. Malinowski and D. G. Howery, *Factor Analysis in Chemistry*, Wiley, New York (1991).
20. P. K. Hopke, *Receptor Modeling in Environmental Chemistry*, Wiley, New York, (1985).
21. J. H. Kalivas, *J. Chemometrics*, **3**, 489 (1989).
22. E. R. Malinowski, *Anal. Chim. Acta*, **122**, 327 (1980).
23. B. A. Roscoe and P. K. Hopke, *Anal. Chim. Acta*, **135**, 379 (1982).
24. T. V. Karstang, J. Toft and O. M. Kvalheim, *J Chemometrics*, **6**, 177 (1992).

ERRATUM

Page 99, line 25: For GRAM the variance and bias factor are *not* identical. This is a marked difference compared to the multivariate case discussed in Appendix I. Let Φ_{nn} and Ψ_{nn} denote the variance and bias factor respectively. They are given by

$$\Phi_{nn} = (\mathbf{H}^T\mathbf{H})^{-1}_{nn}(\mathbf{Y}^T\mathbf{Y})^{-1}_{nn}$$

and

$$\Psi_{nn} = \left[(\mathbf{H}^T\mathbf{H})^{-1}(\mathbf{Y}^T\mathbf{Y})^{-1}\right]_{nn}$$

The symbol Ψ should be replaced by the symbol Φ in equations (10), (24), (25), (29) and (31). The results shown in Tables 5, 6, 7, 8 and 11 were calculated with the correct equations and therefore require no changes.

Page 100, line 6: "q = (SEL)" should be "q = $(\text{SEL})^2$"
Page 100, line 7: "$.^{2,14}$" should be "$.^{14}$"
Page 100, line 28: "$\hat{\Pi}_M^{-1}$" should be "Π_M^{-1}"
Page 105, line 12: "standard error" should be "variance"
Page 108, line 33: "$E[\tilde{\pi}]$" should be "$E[\hat{\pi}]$"
Page 109, line 21: "$E[\pi]$" should be "$E[\hat{\pi}]$"
Page 109, line 34: "$E[\pi]$" should be "$E[\hat{\pi}]$"
Page 111, line 33: "$E[\pi]$" should be "$E[\hat{\pi}]$"
Page 114, line 11: "[[" should be "E["
Page 116, line 19: "(62)" should be "(60)"

GENERALIZED RANK ANNIHILATION METHOD. III: PRACTICAL IMPLEMENTATION

N.M. FABER, L.M.C. BUYDENS and G. KATEMAN

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

SUMMARY

In this paper we discuss the practical implementation of the generalized rank annihilation method (GRAM). The practical implementation comes down to developing a computer program where two critical steps can be distinguished: the construction of the factor space and the oblique rotation of the factors. The construction of the factor space is a least-squares (LS) problem solved by singular value decomposition (SVD) whereas the rotation of the factors is brought about by solving an eigenvalue problem. In the past several formulations for GRAM have been published. The differences essentially come down to solving either a standard eigenvalue problem or a generalized eigenvalue problem. The first objective of this paper is to discuss the numerical stability of the algorithms resulting from these formulations. It is found that the generalized eigenvalue problem is only to be preferred if the construction of the factor space is not performed with maximum precision. This is demonstrated for the case where the dominant factors are calculated by the non-linear iterative partial least-squares (NIPALS) algorithm. Several performance measures are proposed to investigate the numerical accuracy of the computed solution. The previously derived bias and variance are proposed to estimate the number of physically significant digits in the computed solution. The second objective of this paper is to discuss the relevance of theoretical considerations for application of GRAM in the presence of model errors.

KEY WORDS GRAM Least-squares problem Eigenvalue problem NIPALS Performance index Condition number

INTRODUCTION

The first contributions to the development of the method of rank annihilation are characterized by *computational* problems.[1,2,3] For a long time these computational problems were thought to result from the iterative nature of the algorithm used: the solution of the problem demanded the minimization of the smallest significant eigenvalue of a residual matrix by means of trial and error. However, *model errors* caused several eigenvalues to vary simultaneously, thus leading to the difficult choice of the appropriate eigenvalue to be minimized. These computational difficulties were solved by the reformulation of the calibration problem as an eigenvalue equation by Lorber.[4] Initially restricted to the one-component case this reformulation was soon followed up by the generalization to the situation where both samples contain unique components.[5] This generalization by Sánchez and Kowalski is known as the generalized rank annihilation method (GRAM). Many of the latest contributions to GRAM (see e.g. Reference 6), however, make use of an alternative algorithm that has been claimed by Wilson *et al.* to have better numerical properties.[7]

In two previous papers we discussed the relationship between different formulations of GRAM and the effect of random measurement noise on the bias and variance in the estimated eigenvalues.[8,9] (As a by-product we also derived variance in the reconstructed profiles.) The primary result of that investigation was that the amount of bias and variance primarily depends on the specific combination of data matrices used for the construction of the factor space. It was shown that with respect to bias and variance the generalization of Sánchez and Kowalski compared favorably with the generalization of Wilson *et al.* In both papers we did not discuss the numerical properties of the different formulations. Also we refrained from making speculations as to how the derived error estimates would behave in practice. In this paper we will pay attention to both aspects, since they determine how the practical implementation should be carried out and what may be expected if the method is to be used on real data.

One of the reviewers pointed out to us that the commercially available algorithms that would be used to solve the various forms of the GRAM are stable* and will produce all the accuracy necessary for getting reliable answers. We completely agree on this and want to stress once more that the statistical properties should be brought to bear on the problem of determining which method is 'best'. However, some important additional background can be obtained by comparing different algorithms and presenting the solution in steps. The comparison will be made for the standard eigenvalue problem proposed by Sánchez and Kowalski[5] (see References 8 and 9 for notational practice):

$$(\bar{\Theta}^{-1}\bar{U}^{T}M\bar{V})T = T\Pi \tag{1}$$

and a modified form of the generalized eigenvalue problem advanced by Wilson *et al.*:[7]

$$\bar{M}_{UV}T = \bar{Q}_{UV}T\Pi \tag{2}$$

It should be noted that the method of Sánchez and Kowalski diagonalizes a slightly different matrix. However, equation (1) does not constitute a modification since the results should be identical after changing the reconstruction formulas for the pure component profiles in a straightforward manner.[8] The method of Wilson *et al.* is slightly modified in equation (2): instead of using the column and row augmented matrices to estimate the column and row space we decompose the sum matrix **Q** according to the SVD. The results presented later show that it is not necessary to decompose the augmented matrices in order to obtain a numerically stable estimate of the factor space. Thus the only difference between (1) and (2) is the inversion of the diagonal matrix of singular values. According to numerical analysis the conversion of a generalized to a standard eigenvalue problem constitutes no problem if the inverted matrix is stable (well-conditioned) with respect to inversion. In that case it is even economic to solve the standard eigenvalue problem.[10] Thus we are faced with the problem of ascertaining how instable the inverted matrix may be before we run into numerical problems when solving the standard eigenvalue problem. In a previous study we carried out calculations with data that were especially constructed in order to lead to a very instable problem.[9] These calculations should therefore be ideally suited to test the conjecture that one should not be overly concerned about the stability of a particular algorithm as long as one solves the most stable problem. This will be illustrated by means of the performance measures. A performance index is proposed in order to monitor the size of the residuals for the critical steps while the condition number is proposed to quantify the numerical stability. It is important to note that the condition number is often used to predict (an upper bound) for the propagation of data error to the estimated parameters. With the availability of standard errors for the individual components the condition number is no longer useful for this purpose. Furthermore, bias is in no way diagnosed by the condition number.

The second objective of this paper is to discuss the relevance of theoretical considerations for application of GRAM in the presence of model errors. A notable effect of model errors is the possible occurrence of a complex solution. Li *et al.* introduced similarity transformations in order to convert the complex solution into a real one.[6] This enabled the quantitation of a dilute compound in a complex matrix. For the acceptance of a method in a routine laboratory it is, however, necessary that together with the concentration also an estimate of its accuracy and precision is supplied. It is therefore imperative that the effect of the correction procedure on the final solution should be estimated. Otherwise the proposed correction will not lead to a valid method. From this example it can be concluded that a discussion of the effect of model errors is essential for a method that is quite restrictive with respect to the assumed model.

In the following section we will describe which properties are essential for the correct performance of a computer program. Next, we will introduce the performance measures that have been implemented to check the validity of the GRAM calculation. Finally, the practical consequences of the theoretical considerations and the results for model data are discussed with respect to the analysis of data where model errors have to be taken into account.

---

*It is important to distinguish between the stability of the numerical problem and the stability of the algorithm. We will discuss this matter at length in later sections.

PROPERTIES OF A COMPUTER PROGRAM

The solution of a numerical problem is preceded by the development of a computer program. A large number of aspects play a role in developing a computer program. Speed, efficient use of memory, simple structure, portability, friendly user interface, availability of a manual, precision, stability and reliability are all aspects that have to be considered.[11] However, only the last three aspects are essential if the computer program is designed for scientific research. The other aspects become increasingly important if the program is to be converted into a commercial software package. We will therefore restrict ourselves to the last three concepts, since the solution of the problem is directly affected by them.

Precision

The precision of a solution refers to the number of exact figures. The simplest way to estimate the precision is to perform the calculation with a larger number of digits. The first digit that differs between the two solutions indicates that the remaining digits of the former solution are random. Figure 1 shows the relevance of the precision concept if the input data is of experimental origin. Because of measurement noise only the first p digits of the input data are *physically* significant. The next q digits are also significant in the *numerical* sense. The last r digits constitute the representation error. (This error is caused by rounding the input decimal number to the nearest binary number.) During subsequent calculations rounding errors will affect the numerically significant digits $(p+q)$. As a result the output number has become numerically significant in $(p'+q')$ digits. If the calculation is performed in double precision floating point arithmetic 14 digits, for example, may have survived the rounding process. However, this precision may be of little value if the number of physically significant digits p' has become too low for the practical application as a result of error propagation. It may turn out that we have calculated a meaningless number in 14 'exact figures'. If the problem itself is unstable a different algorithm is not going to improve the situation. The only thing that may never happen is that the physically significant digits in the solution are affected by rounding errors. Stated differently: the solution to a possibly stable problem may never be ruined by an unstable algorithm. The precision of the solution is therefore intimately related to the following concepts.

Stability and reliability

A stable algorithm is insensitive to small changes in the input data. Reliability indicates that it is improbable that the program fails, or worse, that it gives the wrong answer without a warning. Reliability is obtained by building in controls to screen the input. There is a relation between stability and reliability: with a stable program the failure depends in a *predictable* way on small changes in the numerical problem.

A well-known example of an unstable algorithm is Gaussian elimination.[12] This algorithm solves a system of linear equations by means of sweeping rows. If the element that is involved in the elimination process (the pivot) happens to be zero, the algorithm will return as answer that the input matrix is singular, even in cases where the input matrix is clearly invertible. This instability can be removed by interchanging rows and columns in such a way that the maximum element is used for the elimination (pivoting). The following example, taken from Reference 12, may be illustrative:

$$\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \qquad \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

I: before pivoting II: after pivoting

Both matrices are non-singular with ideal condition with respect to inversion (contrary to the eigenvalue problem) but without pivoting matrix I can not be inverted by means of Gaussian elimination. Interchanging rows will immediately lead to the inverse. (It should be noted that Golub and van Loan[13] *define* Gaussian elimination to use pivoting.)

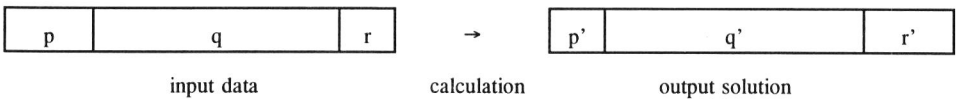


Figure 1. Schematic representation of the loss of physically significant and numerically significant digits for data with measurement error

Recently, the stability of the NIPALS algorithm was discussed using the following example:[14,15]

$$\mathbf{A} = \begin{pmatrix} 1 & -1 \\ 0.8 & 0.7 \\ 0.7 & 0.8 \end{pmatrix} \qquad \mathbf{A}^T\mathbf{A} = \begin{pmatrix} 2.13 & 0.12 \\ 0.12 & 2.13 \end{pmatrix}$$

The two eigenvalues of $\mathbf{A}^T\mathbf{A}$ are $\lambda_1 = 2.25$ and $\lambda_2 = 2.01$ and the corresponding eigenvectors are $\mathbf{x}_1^T = (1,1)$ and $\mathbf{x}_2^T = (1,-1)$ respectively. We notice that this example is characterized by a very high symmetry. It is shown in Reference 16 that if a matrix is symmetrical around both diagonals, it will have eigenvectors that can be divided in two subsets. One subset contains the eigenvectors that are symmetric with respect to the midpoint, i.e. the point that divides the components of the vector in two groups, ($\mathbf{x}_1$ in this example) and the other subset contains the eigenvectors that are anti-symmetric with respect to the midpoint ($\mathbf{x}_2$ in this example). This symmetry is preserved by multiplication with the matrix and the outcome of the iteration procedure therefore depends on the particular choice of the starting vector. Starting exactly on an eigenvector will lead to convergence in one step. It is therefore essential to implement the proposed modifications in References 14 and 15 if this particular symmetry is pertinent to the problem at hand. It follows that NIPALS is an *unstable* algorithm because small changes in the input* may lead to large changes in the output. It is also an *unreliable* algorithm because the failure is not diagnosed in all cases by convergence in one step (see Reference 15). It will, however, be clear that the examples referred to are very different from the data matrices usually encountered during a GRAM analysis.

PERFORMANCE MEASURES

It is well-known for the LS problem that small residuals in the data vector do not automatically correspond to small errors in the solution vector. The primary quantity for assessing the number of significant digits lost is given by the condition number of the inverted matrix. If the data are known to s ($= p + q + r$ in Figure 1) significant digits in base ß, if the condition number is of the order $ß^t$, and if a stable implementation of Gaussian elimination is used, then the number of significant digits in the computed solution will be s-t or 0, whichever is bigger.[17] A similar reasoning holds for the eigenvalue problem but now the invertibility of the eigenvector matrix has to be quantified (see Golub and van Loan[13] and Reference 10). It follows that the numerical accuracy of the GRAM solution can be validated by monitoring both the residuals and the condition number of the critical steps of the calculation.

Some problems remain with respect to the practical evaluation of the following performance measures. First, the choice of a particular norm must be made. It is shown in Reference 13 that this choice is not critical (see also Reference 17). In order to obtain the smallest condition number it is usually calculated by taking the largest singular value as matrix norm, i.e. the L_2 matrix norm, but other norms are simpler to evaluate, e.g. the L_1 and L_∞ matrix norms. Second, the value of the condition number (and therefore the estimated numerical precision of the solution) depends on the scaling of the matrix. General scaling strategies are unreliable and a problem-oriented approach should be followed considering measurement units and data error.[13] The difficult matter of scaling will be further discussed in the 'Results and discussion' section. (See also Reference 18 where all calculations were performed on scaled data.)

Performance index

The residuals of the SVD are monitored by calculating a performance index (*PI*) for the individual factors. These indices are calculated from the size of the residual vector $\mathbf{Q}\,\mathbf{v}_i - \Theta_i\,\mathbf{u}_i$ as follows:

$$(PI)_{SVD,I} = \frac{\|\mathbf{Q}\mathbf{v}_i - \theta_i\mathbf{u}_i\|_1}{10n\epsilon\,\|\mathbf{Q}\|_1\,\|\mathbf{v}_i\|_1} \tag{3}$$

where $\| \bullet \|_1$ denotes the L_1-norm, n is the length of the residual vector and ϵ denotes the machine precision. The performance indices for the standard eigenvalue problem (SEP) are calculated from the size

*It is seen that convergence depends on the choice of the initial starting vector. Since the particular choice made is part of the input data the algorithm is by definition instable.

of the residual vector $\mathbf{A}\,\mathbf{x}_i - \lambda_i\,\mathbf{x}_i$:

$$(PI)_{SEP,i} = \frac{\|\mathbf{A}\mathbf{x}_i - \lambda_i\mathbf{x}_i\|_1}{10n\epsilon\,\|\mathbf{A}\|_1\,\|\mathbf{x}_i\|_1} \tag{4}$$

whereas the performance indices for the generalized eigenvalue problem (GEP) are calculated from the size of the residual vector $\beta_i\,\mathbf{A}\,\mathbf{x}_i - \alpha_i\,\mathbf{B}\,\mathbf{x}_i$:

$$(PI)_{GEP,i} = \frac{\|\beta_i\mathbf{A}\mathbf{x}_i - \alpha_i\mathbf{B}\mathbf{x}_i\|_1}{10n\epsilon\{|\beta_i|\,\|\mathbf{A}\|_1 + |\alpha_i|\,\|\mathbf{B}\|_1\}\|\mathbf{x}_i\|_1} \tag{5}$$

where $|\bullet|$ denotes the modulus of a possibly complex number. The performance indices in (4) and (5) are adapted from the performance indices given in Reference 10 for the complete eigenvalue problem. (This reference was also the motivation of the use of the L_1 matrix norm.) This overall performance index is defined as the maximum of these indices. A value less then 1 is considered to be excellent, a value between 1 and 100 is considered good and a value larger than 100 is considered poor. The exact value is, however, machine dependent. Since only the part of the eigensolution that corresponds to the calibrated components is interesting (see Reference 8), we advise to calculate the performance index for each eigenvector. It should be noted that the expression for the performance index of the generalized eigenvalue problem in Reference 10 is modified by adding the factor '10 n' to the denominator. Results presented later indicate that more consistent values are obtained this way. The performance index for the SVD is new to our knowledge and is presented here as the natural analogy to the performance indices for the eigenvalue problem.

Condition number

For the inversion in equation (1) the condition number is given by

$$\mathrm{cond}(\bar{\Theta}) = \|\bar{\Theta}\|\,\|\bar{\Theta}^{-1}\| \tag{6}$$

where $\|\bullet\|$ denotes a suitable matrix norm. Since Θ is a diagonal matrix, the usual matrix norms (L_1, L_2, L_∞) lead to identical numerical values.

For the eigenvalue problem the condition number is given by

$$\mathrm{cond}(\mathbf{T}) = \|\mathbf{T}\|\,\|\mathbf{T}^{-1}\| \tag{7}$$

A small advantage of the condition number approach in combination with the eigenvalue problem is the fact that column scaling is automatically performed since contrary to the LS problem the matrix to scrutinize is a *result* of the calculation. However, another ambiguity arises because of the existence of equivalent representations of the eigenvalue problem.[8] Although leading to identical eigenvalues they yield eigenvector matrices that are related to each other by premultiplication with a diagonal matrix. Multiplication of the rows of the eigenvector matrix by a diagonal matrix may have a large influence on the numerical value of the condition number. This is not necessarily a serious problem, since only semi-quantitative statements are possible on the basis of condition numbers.

EXPERIMENTAL

Data

Details about the generation of the data are given in Reference 9.

Calculations

The crucial steps in the GRAM calculation consist of the decomposition of a matrix and the solution of an eigenvalue problem. These steps are performed in part by black box routines from the IMSL library[10] and in part by self-written or translated routines. An important advantage of not just using a black box routine is the possibility of studying the convergence properties of the algorithm. The complete SVD of $\mathbf{Q}$ is calculated with IMSL subroutine DLSVRR. The partial SVD is calculated with the NIPALS algorithm.

The NIPALS algorithm is self-written and uses as starting vector the row of the data matrix with largest variance as recommended by Wold.[19] The convergence criterion, i.e the squared length of the difference of two subsequent normalized iterates, was set to the machine precision $\epsilon = 2.22 \times 10^{-16}$. The standard eigenvalue problem of equation (1) is solved with the modified Jacobi algorithm from Reference 20. The algol procedure EIGEN was translated to FORTRAN77 for this purpose. The convergence criterion in EIGEN was adapted to the machine precision. The generalized eigenvalue problem of equation (2) is solved with the QZ-algorithm implemented in IMSL subroutine DGVCRG. Details about the algorithms used are summarized in Table 1. All calculations were performed on a HDS-EX60 mainframe computer. In Table 2 the calculation times for a complete GRAM analysis of the previously described simulated three-component system[9] are compared to elementary floating point operations. These numbers will give an indication of the calculation time on another computer. It follows from Table 2 that calculating the first three factors with the NIPALS algorithm reduces the overall calculation time by a factor of two but calculating an extra noise factor eliminates this advantage in speed because convergence is much slower for noise factors. The exact number of factors is usually not known in advance and it follows that calculation of the complete SVD will actually save time in many practical situations. A notable exception is provided by the method of cross-validation in principal component analysis (PCA).[21] The basic method of cross-validation comes down to divide the data matrix into a number of groups. Each group is deleted in turn from the data and a PCA performed on the reduced data set. The deleted values are then predicted from the PC model parameters. Since after the deletion of a group only the first PC has to be calculated, the computational gain may be considerable.

Table 1. Subroutines used for different calculations

Calculation	Subroutine	Reference
Complete SVD	DLSVRR	10
Partial SVD	NIPALS	19
SEP (equation (1))	EIGEN	20
GEP (equation (2))	DGVCRG	10

Table 2. Execution time (CPU seconds) on the HDS-EX60 mainframe computer

Calculation	Execution time
10^7 Empty DO-loops	1.7
10^7 Multiplications	0.5
10^7 Additions	0.5
10^7 Square roots	15.5
10^4 GRAM[a]	2700
10^4 GRAM[b]	1500
10^4 GRAM[c]	5600

[a]Complete SVD by DLSVRR
[b]Three dominant factors by NIPALS
[c]Four dominant factors by NIPALS

RESULTS AND DISCUSSION

We have divided this section into two parts. In the first part we give the results that were obtained calculating the full SVD of the sum data matrix **Q** with the IMSL-subroutine DLSVRR. In the second part we give the results that were obtained calculating only the first three factors with the NIPALS-algorithm.

Complete singular value decomposition of Q by IMSL-subroutine DLSVRR

The matrices $\overline{\mathbf{Q}}_{UV}$ and $\overline{\mathbf{M}}_{UV}$ of equation (2) are given in Table 3. It is seen that the matrix $\overline{\mathbf{Q}}_{UV}$ is essentially a diagonal matrix. The diagonal elements would constitute the singular values if the calculations were exact. The effect of the rounding errors is reflected by the size of the off-diagonal elements. Now the question arises whether it is allowed to convert the generalized eigenvalue problem into a standard one by simply inverting $\overline{\mathbf{Q}}_{UV}$. In order to answer this question we must investigate the effect of measurement noise in the data and the effect of rounding errors on the computed solution seperately (see Figure 1). In a previous paper we showed that the standard error in a singular value is equal to the standard deviation of the measurement noise.[22] (Deterministic and therefore less efficient bounds for the perturbation of singular values are given by Lawson and Hanson.[23]) The standard deviation of the measurement noise is 0.05 for both matrices added in **Q** (see Reference 9). This means that the standard error in each singular value is $0.05\sqrt{2} = 0.07$. It is seen that even the smallest singular value is about 50 times larger then its standard error and it follows that in spite of the large difference in scale the columns of the matrix $\overline{\mathbf{Q}}_{UV}$ (and its inverse) are orthogonal. This means that in order to assess the stability of the calculation, i.e. to predict the effect of rounding errors, the matrix is already perfectly 'scaled' because the uncertainty in every singular value is the same. A value of 1336 is obtained for the unscaled matrix. If, however, the condition number is to be used to predict that the numerical problem is stable and there is actually no propagation of data error on basis of the condition number one should scale the matrix $\overline{\mathbf{Q}}_{UV}$, giving the ideal value of one. (This lends credit to the opproach of Otto and George to scale the matrix.[18])

Table 3. Matrices $\overline{\mathbf{Q}}_{UV}$ and $\overline{\mathbf{M}}_{UV}$. Decomposition of **Q** with DLSVRR

$\overline{\mathbf{Q}}_{UV}$			$\overline{\mathbf{M}}_{UV}$		
5.2746×10^{3}	1.1×10^{-13}	1.6×10^{-12}	1.6410×10^{3}	-1.0989×10^{1}	1.5014×10^{1}
-4.2×10^{-11}	2.0197×10^{1}	-6.3×10^{-15}	-1.9356×10^{2}	7.4047	-1.0935
-7.1×10^{-13}	-2.0×10^{-13}	3.9470	2.2122×10^{1}	8.3195	0.49980

Table 4. Results of eigenanalysis. Decomposition of **Q** with DLSVRR

n	π			$(PI)_{SVD}$	$(PI)_{EVP}$
1	0.5001	6652058010	4648[a]	0.008	0.06[a]
	0.5001	6652058013	6595[b]		0.01[b]
2	0.2530	1075726999	2989[a]	0.0004	0.03[a]
	0.2530	1075727002	9987[b]		0.02[b]
3	0.0511	9064208029	71263[a]	0.001	0.007[a]
	0.0511	9064208027	60555[b]		0.007[b]

[a]Standard eigenvalue problem of equation (1)
[b]Generalized eigenvalue problem of equation (2)

The results of the eigenanalysis with GRAM are summarized in Table 4. The eigenvalues are divided in three parts according to Figure 1. The number of physically significant digits is estimated on basis of the standard errors and bias.[9] The number of numerically significant digits ('exact figures') is estimated by performing the same calculation on the data without measurement noise. (Actually this caused a floating-point underflow in subroutine DLSVRR.) The solutions are seen to be very precise for both eigenvalue problems. This is satisfactorily predicted by the proposed performance measures. The performance indices indicate that the residuals are extremely small. This is particularly true for the SVD. This information should be combined with the value of the relevant condition number. The condition number for the eigenvector matrix is 72.86. (A value of 82.75 is found for the scaled spectrum matrix that was used to construct the data.) It should be noted that the loss of 4-5 decimal digits can easily be tolerated because we used double precision arithmetic.

Partial singular value decomposition of Q by NIPALS

The matrices $\overline{\mathbf{Q}}_{\mathbf{UV}}$ and $\overline{\mathbf{M}}_{\mathbf{UV}}$ are given in Table 5. There are small differences with respect to the numbers in Table 3. It is seen that the off-diagonal elements of $\overline{\mathbf{Q}}_{\mathbf{UV}}$ are slightly larger. The elements of $\overline{\mathbf{M}}_{\mathbf{UV}}$ are identical to the precision given apart from pairwise changes of sign (that cancel in the SVD).

The results of the eigenanalysis with GRAM are summarized in Table 6. The influence of the loss of precision of the calculated SVD due to the use of NIPALS becomes clear from comparing Table 6 and Table 4. The performance index for the first factor is slightly better for the NIPALS algorithm but the indices for the other factors show that three additional digits are lost. (Their value is no longer 'excellent' but 'good'.) Surprisingly, only the solution to the standard eigenvalue problem has seriously degraded. The reason is that when solving the generalized eigenvalue problem both matrices are projected on the same, possibly slightly rotated, space. As a result the errors effectively cancel. The standard eigenvalue problem does not treat the two matrices symmetrically, since the projected matrix is replaced by its diagonal.

Table 5. Matrices $\overline{\mathbf{Q}}_{\mathbf{UV}}$ and $\overline{\mathbf{M}}_{\mathbf{UV}}$. Decomposition of **Q** with NIPALS

$\overline{\mathbf{Q}}_{\mathbf{UV}}$			$\overline{\mathbf{M}}_{\mathbf{UV}}$		
$5.2746\text{x}10^{3}$	$-2.1\text{x}10^{-10}$	$2.7\text{x}10^{-9}$	$1.6410\text{x}10^{3}$	$1.0989\text{x}10^{1}$	$1.5014\text{x}10^{1}$
$1.5\text{x}10^{-10}$	$2.0197\text{x}10^{1}$	$-6.0\text{x}10^{-9}$	$1.9356\text{x}10^{2}$	7.4047	1.0935
$7.7\text{x}10^{-10}$	$-1.2\text{x}10^{-9}$	3.9470	$2.2122\text{x}10^{1}$	-8.3195	0.49980

Table 6. Results of eigenanalysis. Decomposition of **Q** with NIPALS

n	π			$(PI)_{\text{SVD}}$	$(PI)_{\text{EVP}}$
1	0.5001	6652052	0310229[a]	0.003	0.05[a]
	0.5001	6652058025	2849[b]		0.01[b]
2	0.2530	1075732	8689523[a]	1	0.03[a]
	0.2530	1075726986	0345[b]		0.01[b]
3	0.0511	9064207	39448005[a]	4	0.006[a]
	0.0511	9064207970	10615[b]		0.007[b]

[a]Standard eigenvalue problem of equation (1)
[b]Generalized eigenvalue problem of equation (2)

Discarding the residuals thus leads to additional loss of significant digits. It follows that in combination with the NIPALS algorithm the solution of the generalized eigenvalue problem must be preferred. This conclusion presumably also holds if the partial least-squares (PLS) algorithm of Reference 6 is used to calculate the orthogonal base vectors. It is important to note that there is still some freedom of choice because the calculations were performed in double precision. If the calculations would have been performed in single precision part of the results in Table 6 would have been meaningless. The stability of a calculation depends on the algorithm as well as the precision of the arithmetic.[13]

CONSEQUENCES OF MODEL ERRORS

The preceding considerations about performance measures and results obtained for ideal bilinear data are perhaps interesting in theory but relatively useless in practice if model errors play a dominant role. Among the most frequently encountered model errors are matrix effects, interactions, non-linearities, heteroscedastic noise and retention time shifts (in chromatography). There are basically two ways to encounter model errors: a hardware and a software approach. The hardware approach comes down to improving the experiment in such a way that the assumed model for the data becomes essentially correct. This approach was followed by McCue and Malinowski who optimized a liquid chromatograph with UV detection for rank annihilation and obtained excellent results.[24] The alternative is to use more realistic models.

One may also simply ignore the fact that the assumptions about the model are violated and try to quantitate the effect in practice without doing anything about it. In order to facilitate the evaluation of the performance of multivariate methods on non-ideal data so-called realistic simulations were developed at our department.[25] These simulations come down to adding one-component data matrices with a specific weight (concentration) after background subtraction. The one-component data matrices are not approximated by the first factor as in the original work of Ho *et al.* to allow for all kind of model errors that can give rise to additional significant factors. The results can however only be used to predict the errors for simple systems since the number of factors to control by simulation rapidly increases with the number of overlapped components.

Matrix effects

Matrix effects can be corrected by the method of standard addition and this procedure is completely in the spirit of rank annihilation. Adding the analyte of interest to the unknown sample provides all the necessary data for the quantitation of this analyte. Departure from unity of the relative concentrations of the remaining components will give an indication about the quantitative aspects of the complete analytical procedure (i.e. starting from sample preparation until obtaining the raw data: e.g. injection volume irreproducibility in chromatography). Target testing[1] of the added standards will provide useful information about the number of significant factors to use for the transformation. For data obtained by high-performance liquid chromatography with ultraviolet detection (HPLC-UV) target testing proved to give reasonable estimates of the number of absorbing species while other well-known methods failed completely.[25] Furthermore, if chromatographic data is used, standard addition will simplify the synchronization of the elution profiles.[6]

Interaction effects

Interaction effects generate a cross term between two concentrations. Contrary to matrix effects, this problem can not be solved by, for example, standard additions. Neither can it be modeled within the framework of GRAM, since GRAM actually leads to an eigenvalue problem by excluding interactions. The influence on the calculated solution can, however, always be evaluated by means of realistic simulations.

Non-linearities

Non-linearities have been reported for data of entirely different sources. For fluorescence data Ho *et al.* found that several of the largest eigenvalues varied linearly with concentration. Eigenvalues corresponding to dilute components were dominated by secondary eigenvalues of the strong absorbing components. This introduced the problem of correctly interpreting the resulting eigenvalue plots. Lorber enabled the exact minima to be found by the solution of an eigenvalue problem. Since then, the eigenvector plots have no longer been reported in the literature. It is however clear that valuable information may be gained from these eigenvalue plots. We therefore suggest to construct these eigenvalue plots around the exact minima found by the solution of the eigenvalue problem since the main disadvantage connected with the iterative method, i.e. locating the minimum, has already been solved. Dose and Guiochon[26] have shown that non-linearities can not be avoided for UV-detectors. As a result more factors will be needed in order to describe the signal within the noise level.[25] However, Li *et al.*[6] have shown that the stability of the

problem is affected by the dimension of the transformation for real data. The dimension of the transformation matrix can be reduced to a minimum if non-linear factors[27] are used for the projections. This could lead to a significant improvement of the stability of the problem (i.e reduce standard errors).

Retention time shifts
Very recently Poe and Rutan[28] have evaluated the sensitivity of the GRAM solution to retention time shifts by means of realistic simulations. Ironically, they find a *negative* bias in the estimated eigenvalues while the theoretically predicted bias is *positive*. This illustrates both the usefulness of these simulations as well as the inadequacy of predicted errors in the presence of model errors.

Heteroscedastic noise
It is well-known that heteroscedastic noise can lead to additional significant eigenvalues in principal component analysis (PCA).[29] In order to obtain a good fit with the 'correct' number of factors weighted PCA must be carried out, i.e. the model must be adapted. Very recently Keller *et al.* have shown how the proper pretreatment of LC-UV data greatly improves the performance of evolving factor analysis (EFA).[30] Since EFA assumes the same model as GRAM this should be an important result for users of GRAM. Furthermore LC-UV data is representative of the most important class of data nowadays being analyzed with GRAM. It follows that in order to use GRAM on this kind of data the same pretreatment should be used. It is perhaps interesting to note that the contribution of the separate modes to the total measurement error is not necessarily reflected in the generally asymmetric apportionment of error to the reconstructed profiles since the error in the reconstructed profiles primarily depends on the overlap in the other mode.[9]

CONCLUSIONS

The practical implementation of GRAM has been discussed with respect to the choice of the algorithms and the expected influence of model errors. A number of performance measures is proposed for the assessment of the reliability of the computed solution of the GRAM. The performance indices are a measure for the size of the residuals that arise during the most important steps of the calculation, i.e. the construction of the factor space and the rotation of the factors. This information should be combined with the relevant condition number. The reliability of the computed solution is satisfactorily predicted in this way. It is shown that for the amount of overlap to be expected in practice the number of numerically significant digits should greatly exceed the number of physically significant digits (predicted by the previously derived bias and variance expressions.) Therefore, the different algorithms that apply to this problem should all perform equivalently. This has been illustrated for the modified Jacobi algorithm (applied to the standard eigenvalue problem) and for the QZ algorithm (applied to the generalized eigenvalue problem). An exception must be made if the factor space used for the projections is not calculated with maximum numerical precision. It has been found that calculating the partial SVD by the NIPALS algorithm leads to an excessive loss of numerically significant digits if the standard eigenvalue problem is evaluated but has little effect on the solution of the generalized eigenvalue problem since now both data matrices are projected on the same, less precise, factor space and the resulting errors tend to cancel out.

The effect of model errors on the outcome of the analysis has been discussed from a general point of view. Although good results are reported in the literature for the estimated concentrations and reconstructed profiles the same can not be expected for the predicted bias and variance. The reason for this unfortunate situation is that the concentration estimates are primarily determined by the adequacy of the model description of the data whereas the estimate of the confidence interval is primarily determined by the model description of the noise. The confidence levels predicted by theory may however offer a reference point indicative of what could be possible in absence of model errors. In the light of the fast development of theory we therefore agree with the prophetic statement of Sánchez and Kowalski[31] that 'The greatest potential for GRAM, and in general second-order methods, is perhaps in future second-order instruments which are yet to be built.'

ACKNOWLEDGEMENTS

Dr. Sijmen de Jong and the reviewers are acknowledged for rooting out errors and unclear passages.

REFERENCES

1 C. -N. Ho, G. D. Christian and E. R. Davidson, *Anal. Chem.* **50**, 1108 (1978).

2 C. -N. Ho, G. D. Christian and E. R. Davidson, *Anal. Chem.* **52**, 1071 (1980).
3 C. -N. Ho, G. D. Christian and E. R. Davidson, *Anal. Chem.* **53**, 92 (1981).
4 A. Lorber, *Anal. Chim. Acta* **164**, 293 (1984).
5 E. Sánchez and B. R. Kowalski, *Anal. Chem.* **58**, 496 (1986).
6 S. Li , J. C. Hamilton and P. J. Gemperline, *Anal. Chem.* **64**, 599 (1992).
7 B. E. Wilson, E. Sánchez and B. R. Kowalski, *J. Chemometrics* **3**, 493 (1989).
8 N. M. Faber, L. M. C. Buydens and G. Kateman, *J. Chemometrics*, **8**, 147 (1994).
9 N. M. Faber, L. M. C. Buydens and G. Kateman, *J. Chemometrics*, **8**, 181 (1994).
10 *IMSL MATH/LIBRARY User's Manual, Version 1.1*, IMSL, Houston, TX (1989).
11 J. R. Rice, *Matrix computation and mathematical software*, McGraw-Hill, New York, 1981.
12 J. H. Wilkinson, *The Algebraic Eigenvalue Problem*, Clarendon, Oxford, 1965.
13 G. H. Golub and C. van Loan, *Matrix Computations*, John Hopkins University Press, Baltimore, MD (1983).
14 Y. Miyashita, T. Itozawa, H. Katsumi and S-I. Sasaki, *J. Chemometrics* **4**, 97 (1990).
15 M. B. Seasholtz, R. J. Pell and K. E. Gates, *J. Chemometrics* **4**, 331 (1990).
16 M. Crampin, *Teach. Math.* **8**, 1 (1989).
17 J. H. Kalivas and P. Lang, *J. Chemometrics* **3**, 443 (1989).
18 M. Otto and T. George, *Anal. Chim. Acta* **200**, 379 (1987).
19 S. Wold, C. Albano, W. J. Dunn, K. Esbensen, S. Hellberg, E. Johansson and M. Sjöström, in *Food research and Data Analysis*, ed. by H. Martens and H. Russwurm, pp. 147-188, Applied Science, London, (1983).
20 P. J. Eberlein and J. Boothroyd, in *Handbook for Automatic Computation*, Vol. II, *Linear Algebra*, ed. by J. H. Wilkinson and C. Reinsch, pp. 327-338, Springer, Berlin (1970).
21 S. Wold, *Technometrics*, 20, 397 (1978).
22 N. M. Faber, L. M. C. Buydens and G. Kateman, *J. Chemometrics* **7**, 495 (1993).
23 C. L. Lawson and R.J. Hanson, *Solving Least Squares Problems*, Prentice-Hall, Englewood Cliffs, NJ (1974).
24 M. McCue and E. R. Malinowski, *J. Chromatogr. Sci.* **21**, 229 (1983).
25 M. J. P. Gerritsen, N. M. Faber, M. van Rijn, B. G. M. Vandeginste and G. Kateman, *Chemometrics Intell. Lab. Syst.* **12**, 257 (1992).
26 E. V. Dose and G. Guiochon, *Anal. Chem.* **61**, 2571 (1989).
27 C. Jochem and B. R. Kowalski, *Anal. Chim. Acta* **133**, 583 (1981).
28 R.B. Poe and S.C. Rutan, *Anal. Chim. Acta*, 283 (1993) 845.
29 R. N. Cochrane and F. H. Horne, *Anal. Chem.* **49**, 846 (1977).
30 H. R. Keller, D. L. Massart, Y. Z. Liang and O. M. Kvalheim, *Anal. Chim. Acta*, **263**, 29 (1992).
31 E. Sánchez and B. R. Kowalski, *J. Chemometrics*, 2, 265 (1988).

FUTURE RESEARCH

From the present work many directions for future research can be given. I will confine myself to speculate on possible future research connected with GRAM, since the research performed on this method is characterized by most of the unfinished work.

Complex eigensolution (p.92-93)
It has been shown that a complex eigensolution does not constitute a problem for model data. Several procedures can be applied for obtaining the solution for the calibrated components and this part of the overall solution should be real. The procedures mentioned here can be divided into two categories. The first category contains the 'mathematical' solutions of Öhman *et al.*[1] and of Westlake.[2] The second category contains the 'chemical' solution of Li *et al.*[3] Solutions belonging to the first category are termed mathematical because they are perfectly appropriate for model data. However, if the solution happens to be complex for one of the calibrated components, the straightforward mathematical solutions are no longer sufficient and GRAM should be modified as shown by Li *et al.* Furthermore, if the GRAM solution is to be used as a starting solution for an alternating least squares (ALS) procedure,[4] the mathematical solutions are inappropriate as well since the ALS procedure needs a real input. Forcing the complete solution to be real could be a serious advantage of the modification of Li *et al.* However, this potential advantage still needs to be investigated since, obviously, the advantage is illusory if the ALS procedure fails to converge to an acceptable solution in this case.

Degenerate eigensolution (p.93)
In the case that the eigenvalues found by GRAM are degenerate, i.e. a particular eigenvalue occurs more than once, the solution to the calibration problem is not complete. This is an inevitable problem connected with the method. Two causes for degeneracy are possible if measurement noise is present. Firstly, degeneracy can be a result of two concentration ratios that are *exactly* equal and secondly, degeneracy can be a result of two concentration ratios that are *approximately* equal. The first situation is very unlikely to occur in practice (for the analytes of interest) and will not be considered any further. The second situation, however, constitutes a fundamental drawback of GRAM. In fact, a degenerate eigensolution that is a combined effect of almost equal concentration ratios and noise is a very disturbing property, since a calibration method is expected to work best if the calibration and unknown samples are very much alike. Instead, when using GRAM on data obtained for similar samples will actually lead to an increased chance of producing an incomplete solution. Since the problem of a degenerate eigensolution is so fundamental, possible solutions should receive more attention in the future. The possible solution of adding (a small amount of) noise to the data is not so far fetched as it might seem, since noise is part of the origin of the problem. (It could be termed a 'homeopathic' solution.) Further research is needed before such a procedure can be recommended for routine use.

Bias in the eigenvalues (p.98-99)
Two different kind of errors are discussed here. The first error, the error in the concentrations, can be interpreted as a *sampling error*. For chromatographic data, for example, it may the result of an irreproducible injection volume. The second error, the error in the responses, is a *measurement error*. In this section we only focus on the effect of the measurement error. It is important to notice that in Kubista's method both data matrices are measured for the *same* mixture. Thus the sampling error is absent. Depending on the relative size of both errors, this may constitute an immense advantage over the usual applications of GRAM that need the measurement of a reference mixture. An evaluation of Kubista's method based on simulations has very recently been published.[5] These results should be compared with the theoretical predictions given here.

Figures of merit (p.99-100)
Lorber[6] has developed an error theory for multivariate calibration that includes expressions for error propagation, signal to noise, limit of detection, precision, accuracy, sensitivity and selectivity for the individual components. The important question is: can this theory be extended to higher-order data? This question has been answered in part by Wang *et al.*[7] by defining figures of merit in a way similar to Lorber's expressions. However, some of their expressions are not confirmed by the current results. The difficulty in copying Lorber's expressions primarily lies in the fact that (dimensionless) concentration ratios are determined instead of concentrations. This is of immediate consequence for the figures of merit that are concentration dependent, e.g. the sensitivity. Further research is needed in order to elucidate this important problem (see **Variance in the unknown concentrations**).

Bias in the eigenvalues for the generalization of Wilson *et al.* (p.101-102)
The final result, i.e. Equation (23), is very uncomfortable. Perhaps a completely different approach should be followed in order to derive the expectation of the estimated eigenvalues. Very recently Poe and Rutan introduced yet another eigenvalue problem.[8] This modification constitutes a combination of the two generalizations already extensively treated. The appropriate expressions should be derived for this modification as well. No *extra* difficulties are to be expected.

Bias correction technique (p.102-103)
The proposed bias correction technique seems to work well. In fact the same technique has been proposed by Goodman and Haberman[9] for the bias in the eigenvalues of PCA. Van Huffel and Vandewalle[10] briefly mention the possibility of a bias correction in MLR with errors in the independent variables.* Thomas[11] proposes an estimator that has a decreased bias at the cost of an increased variance. Future research should be directed towards deriving adequate bias expressions (see **Variance in the reconstructed profiles**) and subsequent comparison of different estimators. Furthermore, it would be useful to derive an index that would signal the breakdown of certain approximations made for GRAM. Stewart[12] has given such an index for MLR that is closely related to Lorber's signal-to-noise ratio.[6] (A major difference is that Stewart's index includes the number of observations.)

Confidence limits (p.103-104)
This part is based on the results of Monte Carlo simulations. This is not the correct procedure known from statistics. The correct procedure would be to derive the distribution of the estimator by making either assumptions about the number of observations (large sample approximation) or about the distribution of the measurement noise. Simulations are very convenient in order to investigate the consequences of having a limited number of observations with a measurement error that does not follow the Gaussian distribution. Thus a formal treatment of this matter possibly followed up by simulations is needed.

Variance in the unknown concentrations (p.104-105)
Equation (36) can be rewritten by recognizing that the total selectivity is equal to the product of the selectivities in the individual modes. This immediately yields (for the calibrated components):

$$\mathrm{var}(\hat{c}_{M,n}) = (\mathrm{SEN})_n^{-2}\left[\left(\frac{\sigma_N}{\bar{c}_{N,n}}\right)^2 + \left(\frac{\sigma_M}{\bar{c}_{M,n}}\right)^2\right] + \left(\frac{\sigma_{\hat{c}_{N,n}}}{\bar{\pi}_{N,n}}\right)^2$$

It follows that a discussion on figures of merit (see **Figures of merit**) should also include the determined concentrations.

PCR version of rank annihilation (p.111-112)
Principal component regression (PCR) is a method that can be used if MLR gives unstable estimates as a result of collinear data.[13] The variance is reduced by replacing the original regressors by a selection of PCs. At the same time, a bias is introduced and the choice between the two methods is a trade-off between the two error contributions.[14] Since the variance reduction depends on the size of the measurement error and is likely to be important if the measurement error is large, the conclusion in this section is not justified, since the measurement error is rather small in this example. (The value for the standard deviation is chosen because it is realistic for certain kind of data.) A PCR version is more likely to be feasible for data which is relatively imprecise. There is, however, an important problem connected with this modification. The number of estimable concentration ratios is equal to the number of PCs chosen to construct the factor space. Thus it is to be expected that the PCR modification will take components together that are very similar. This is not necessarily a problem if, for example, the total concentration of a number of related substituents is to be determined (e.g. o-, m- and p-substituted benzene of a certain kind). It is obvious that further research is necessary.

RR version of rank annihilation
Ridge regression (RR) is another alternative to MLR in the case of highly collinear data.[13] The procedure comes down to stabilizing the inversion step by replacing the singular values of the inverted matrix Θ by $\Theta/(\Theta^2+\kappa)$ where κ is the ridge parameter. Since collinear data is not uncommon in analytical chemistry,

---

*The resulting method (see p.232) is called corrected least squares (CLS). Thus a suitable name for GRAM with the proposed bias correction would seem to be corrected GRAM (CGRAM).

this modification also deserves attention. It does not suffer from the drawback mentioned above for the PCR version. However, the necessary expressions should be derived.

Wavelength selection (p.113)
Wavelength selection is a very important issue in multivariate data analysis.[15] It is well known that adding wavelengths reduces the variance in MLR.[16] Still, even better results are possible by selecting wavelengths and keeping the measurement time constant, i.e. replicating the experiments at the selected wavelengths.[17,18] In all these studies only the variance was considered. However, theory indicates that bias may have an overwhelming effect if the number of observations is large. Future investigations should deal with bias as well and the relevant expressions show that it should present no difficulties.

Bias in multivariate calibration (p.114)
Equation (46) is derived using the 'small sample' approximation, i.e. letting the measurement error tend to zero while keeping the number of observations constant. Davies and Hutton[19] have derived similar results using a 'large sample' approximation, i.e. letting the number of observations tend to infinity while keeping the measurement error constant. The derived bias expressions should be extended in order to include all situations concerning the number of observations and the size of the measurement error.

Variance inflation factors (p.115)
The usefulness of these variance inflation factors for detecting multicollinearities in MLR is well established. The possibilities should be investigated for GRAM as well.

Variance in the reconstructed profiles (p.115-116)
Equations (56) and (57) are the result of an approximation proposed by Malinowski.[20] Equation (51) can be worked out by taking the standard error in the eigenvectors into account (see Reference 21, p.69-70). Here the approximation is used because it leads to sufficiently accurate results. Moreover, the resulting expressions for the bias are relatively easy to interpret. It is clear that routine use of the bias expressions can only be recommended if the appropriate derivations are completed. One should either include these extra terms or prove that they are negligible, e.g. by calculating the stochastic norm[12] of the individual contributions. It is important to note that in this section focus is completely on the error analysis of the eigenvalues. This is in contrast to Kubista's method where focus is primarily on the reconstructed profiles. Working out this section should therefore be of special importance for applications of Kubista's method.

Variance in the eigenvalues (p.116)
More accurate bounds for the perturbation of a pseudoinverse are given by Lawson and Hanson.[22] Thus the recommendation made above also goes for this result. Simulations show that the predicted variances are accurate but a rigorous proof is needed. Equation (58) should be correctly worked out at the risk of obtaining expressions that are hard to interpret and only lead to a marginal improvement.

Comparison of GRAM and PLS-RBL
Comparisons between different formulations of the same method are nice but a comparison between different methods should be more rewarding. The comparison of GRAM and partial least squares-residual bilinearization (PLS-RBL) poses several problems. GRAM has been derived for bilinear signal for the analytes whereas RBL assumes that the interferents can be bilinearly modeled.[1] GRAM has been extended to handle the same kind of data. This alternative method is called nonbilinear rank annihilation (NBRA).[23] Thus it seems worthwhile to derive standard errors and bias for NBRA. Very recently Phatak *et al.*[24] have published standard errors for PLS and a comparison based on theoretically predicted errors would become feasible. The results for PLS are, however, not complete (they do not account for bias) and this topic is really very ambitious. The subject remains interesting because, for example, very recently Wang *et al.*[7] have arrived at the following conclusion: 'Since RBL is a two-stage least-squares solution while NBRA is an eigenvalue-eigenvector solution (after data preprocessing), it is expected that RBL results will be less influenced by noise in data.' I do not agree on this point. The variance expressions derived for GRAM show that the estimated concentration ratios are very efficient. Simple inspection of Equation (48) shows that it must be expected that the errors in the concentrations estimated by PLS-RBL depend on the concentrations of all the components that contribute to the data whereas the analogous results for GRAM show that the components only interact through the variance factor. It will probably depend on the data which method performs best. However, in order to back up this conjecture some theory should be developed.

Comparison of GRAM and PARAFAC

Very recently GRAM and PARAFAC have been compared by simulations[4] and it was demonstrated that for data of a certain complexity PARAFAC results are better than GRAM results in the majority of the cases. This result is useful, since it points out that the methods are different. However, in order to arrive at this result the output of the simulations had to be compared with the original input. Such a validation is not possible in practice. Moreover, in practice, one is not interested in 'the majority of the cases' but in the particular data at hand. Theoretically predicted errors would provide the necessary information, the only prerequisite being that the data obey the model. Thus variance and bias should be derived for PARAFAC. Appelof and Davidson[25] have already given an expression for the variance. However, their variance expression is likely to be incorrect, since it only contains one concentration independent term. The reasoning is similar as for the possible comparison of GRAM and PLS-RBL. However, the amount of work involved is expected to be smaller.

Problem of heteroscedastic noise

In the study of Mitchell and Burdick[4] the noise was proportional to the data. This made the simulations more realistic. The authors emphasized that PARAFAC assumes homoscedastic noise. This is also the case for GRAM in the current implementations. An interesting topic would therefore be to derive standard errors and bias in the heteroscedastic situation. This should preferably be done for data that has been appropriately preprocessed, i.e. according to Paatero and Tapper[26] for 'optimally scaled data'. Particular problems for standard error and bias are not foreseen but perhaps the distribution of the parameters is influenced. This would have consequences for the associated confidence intervals and subsequent hypothesis testing. This is certainly an interesting and important topic for future research. It is important to note that most recently (two weeks before finishing this thesis) a paper of Books and Kowalski[27] has appeared that describes the results of simulations where the noise structure of real data is mimicked as closely as possible. (In their simulations the noise is added to the different orders separately.) The present expressions should be extended using this realistic noise model. No special difficulties are to be expected. For example, the variance expression immediately results by working out their Equation (13) and inserting the resulting statistical errors in Equation (61).

REFERENCES

1 J. Öhman, P. Geladi and S. Wold, *J. Chemometrics*, **4**, 135 (1990).
2 J.R. Westlake, A Handbook of Numerical Matrix Inversion and Solution of Linear Equations, John Wiley, New York (1968).
3 S. Li, J.C. Hamilton and P.J. Gemperline, *Anal. Chem.* **64**, 599 (1992).
4 B.C. Mitchell and D.S. Burdick, *Chemometrics Intell. Lab. Syst.* **20**, 149 (1993).
5 I. Scarminio and M. Kubista, *Anal. Chem.* **65**, 409 (1993).
6 A. Lorber, *Anal. Chem.* **58**, 1167 (1986).
7 Y. Wang, O.S. Borgen and B.R. Kowalski, *J. Chemometrics*, **7**, 439 (1993).
8 R.B. Poe and S.C. Rutan, *Anal. Chim. Acta*, **283**, 845 (1993).
9 L.A. Goodman and S.J. Haberman, *JASA*, **85**, 139 (1990).
10 S. van Huffel and J. Vandewalle, *The Total Least Squares Problem*, SIAM, Philadelphia (1991).
11 E.V. Thomas, *Technometrics*, **33**, 405 (1991).
12 G.W. Stewart, *Contemp. Math.* **112**, 171 (1990).
13 E. Sanchez and B.R. Kowalski, *J. Chemometrics*, **2**, 247 (1988).
14 T. Naes and H. Martens, *J. Chemometrics*, **2**, 155 (1988).
15 P.J. Gemperline, *J. Chemometrics*, **3**, 549 (1989).
16 A. Lorber and B.R. Kowalski, *J. Chemometrics*, **2**, 67 (1988).
17 S.D. Frans and J.M. Harris, *Anal. Chem.* **57**, 2680 (1985).
18 K. Sasaki, S. Kawata and S. Minami, *Appl. Spectrosc.* **40**, 185 (1986).
19 R.B. Davies and B. Hutton, *Biometrika*, **62**, 383 (1975).
20 E.R. Malinowski, *Anal. Chim. Acta*, **122**, 327 (1980).
21 J.H. Wilkinson, The Algebraic Eigenvalue Problem, Clarendon, Oxford (1965).
22 C.L. Lawson and R.J. Hanson, *Solving Least Squares Problems*, Prentice-Hall, Englewood Cliffs, NJ (1974).
23 B.E. Wilson and B.R. Kowalski, *Anal. Chem.* **61**, 2277 (1989).
24 A. Phatak, P.M. Reilly and A. Penlidis, *Anal. Chim. Acta*, **277**, 495 (1993).
25 C.J. Appelof and E.R. Davidson, *Anal. Chem.* **53**, 2053 (1981).
26 P.Paatero and U. Tapper, *Chemometrics Intell. Lab. Syst.*, **18**, 183 (1993).
27 K. Booksh and B.R. Kowalski, *J. Chemometrics*, **8**, 45 (1994).

GENERAL CONCLUSIONS

In this thesis attention has been paid to three important topics in multivariate data analysis in chemometrics, i.e. principal component analysis (PCA), pseudorank estimation and the generalized rank annihilation method (GRAM). Especially for GRAM some new results are presented. Most results in this thesis are theoretical in nature and still have to be tested in a real situation (perhaps in an improved form). However, it has already been discussed at several instances that model errors will make such a test useless. Still, practitioners in the field should profit from the current work because it also aims at contributing to a better understanding of the working of multivariate methods.

SUMMARY

In this thesis several topics are treated that should be of interest for practitioners in the field of multivariate data analysis in chemometrics. These topics are principal component analysis (PCA) in Part I, pseudorank estimation in Part II and the generalized rank annihilation method (GRAM) in Part III. Especially for GRAM some new theoretical results are presented. The adequacy of new expressions is always tested by analyzing large samples of Monte Carlo (MC) trials. Still, there is much room for possible improvements and some of them are already proposed in the preceding section.

In Section 1 standard errors in the eigenvalues of a cross-product matrix are derived that constitute an improvement with respect to the standard errors that have been in use up until now. The derived standard errors are appropriate for the case that the elements of the test data matrix, i.e. the data matrix under consideration, are unknown constants with only measurement error. (From multivariate statistics standard errors resulting from sampling errors have been known for a long time.) It is discussed that the derived standard errors can not be applied in a straightforward fashion for determining the pseudorank of a matrix in the most demanding situation, i.e. for low signal-to-noise ratio, since the derivation only takes the first-order term into account of a series that converges slowly. (Note that the 'old' expressions were derived for the purpose of pseudorank estimation: see Section 3.) Three applications of the derived standard errors are treated. First, the relationship between previously derived standard errors in the scores and loadings of PCA is established which have previously been considered to be different. Next, standard errors in multivariate calibration are discussed and finally, standard errors in the eigenvalues of GRAM are derived. It is shown that a variance reduction is obtained if the data matrix used for the construction of the factor space is obtained by 'simulating' standard addition rather than by performing real standard addition.

In Section 2 several aspects are discussed that are essential for the applicability of pseudorank estimation methods that are based on the eigenvalues of PCA of random matrices. These aspects are the influence of the matrix size, the ratio between the number of rows and columns, the signal-to-noise ratio and the distribution of the measurement error. Asymptotic results from the statistical literature are discussed and their practical usefulness is evaluated by means of MC simulations. Deviations from 'ideal' behavior (e.g. predicted for infinitely large matrices) are compared with the intrinsic variability of the eigenvalues, i.e. their standard error. It is shown that down to a very small signal-to-noise ratio the eigenvalues associated with the residuals of the test data matrix can be approximated by the eigenvalues of a random matrix thus leading to a valid pseudorank estimation method. However, the random matrix should preferably have the same number of degrees of freedom as the test data matrix rather than the same size as is common practice now. A possible modification of these methods is therefore proposed. Furthermore, it is found that under rather general conditions the distribution of the eigenvalues of a random matrix is primarily determined by the ratio between the number of rows and columns of the matrix. The consequences for two popular pseudorank estimation methods are discussed.

In Section 3 several aspects are discussed that are essential for the applicability of pseudorank estimation methods that are based on an estimate of the size of the measurement error. These aspects are 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 noise. They are discussed with respect to three parametric methods. The first method (Method A) is based on the standard error in the diagonal elements of the row-echelon form of the test matrix, the second method (Method B) is based on the previously derived standard error in the eigenvalues of PCA and the third method (a *t*-test) is based on the standard error in the singular values. It is shown that Methods A and B are problematic because the sampling distribution of the test statistic is unknown. Comparison with Malinowski's *F*-test shows that the *t*-test makes efficient use of the extra knowledge about the noise. Finally, it is shown that a plot of the singular values yields a promising graphical pseudorank estimation method in the case that the extra knowledge is not available. This 'new' graphical method therefore provides a natural complement to the *t*-test.

In Section 4 different formulations of GRAM are compared and a slightly different eigenvalue problem is derived. This reformulation leads to simpler reconstruction formulas for the pure component profiles and thereby enables the comparison with other PCA-based methods for curve resolution and calibration. Furthermore, it facilitates the discussion of two characteristic problems of GRAM, i.e. the distinct possibility of a complex and degenerate solution. It is shown that a complex solution - unlike degeneracy - should not arise for components that are present in the unknown as well as in the known mixture if the data follows the model assumptions. Possible solutions to these problems are discussed.

In Section 5 the effect of random measurement error on the estimated eigenvalues of GRAM is treated. First, it is shown that, in general, random errors do not only lead to a spread in the estimated parameters (variance) but also to a shift (bias) if the estimation procedure comes down to a nonlinear transformation of the data. Next, some new symbols are introduced in order to cast the previously introduced reformulation

into an appropriate probability model. Bias in the eigenvalues is derived by combining established results from least squares (LS) theory and Malinowski's 'error theory for factor loadings resulting from combination target factor analysis'. Variance is derived by performing error propagation on the previously introduced reformulation. The derivations are carried out for three different procedures that are currently in use for constructing the factor space, i.e. Lorber's method, the generalization of Sánchez and Kowalski and of Wilson, Sánchez and Kowalski. It is shown that, depending on the size of the eigenvalue, large differences in performance should be expected. A bias correction technique is proposed that effectively eliminates the bias if the bias estimate is adequate. It is shown that the predictions are satisfactory up to the limit of detection. Furthermore, the quality of the results does not seem to be sensitive to the particular choice of the dimension of the factor space.

In Section 6 the practical implementation of GRAM is discussed from two points of view, i.e. the computer program that has to be developed and the data that are currently being generated in practice. From Sections 4 and 5 it has become clear that the computer program contains two critical steps, i.e. the construction of the factor space and the oblique rotation of the factors. The first step is a LS problem whereas the second step is an eigenvalue problem. The first objective of this section is to show that up until now too much attention has been paid in the literature to the numerical properties of the different formulations. It is important to note that the numerical problems are of minor importance compared with the statistical problems. The second objective is to discuss the relevance of theoretical results (especially those obtained in Section 5) if GRAM is applied to data that are corrupted by model errrors.

SAMENVATTING

In dit proefschrift worden verschillende onderwerpen behandeld die interessant kunnen zijn voor gebruikers van multivariate data-analyse technieken in de chemometrie. Deze onderwerpen zijn principale componenten analyse (PCA) in Part I, pseudorang schatting in Part II en de 'gegeneraliseerde rang annihilatie methode' (GRAM) in Part III. Met name voor GRAM wordt een aantal nieuwe theoretische resultaten gegeven. De geschiktheid van deze uitdrukkingen is altijd getest door grote steekproeven van Monte Carlo (MC) trials te analyseren. Desalniettemin is er veel ruimte voor mogelijke verbeteringen en enkele daarvan zijn reeds voorgesteld in een voorgaand gedeelte.

In Paragraaf 1 worden standard errors afgeleid voor de eigenwaarden van een kruisproduktmatrix. Deze standard errors betekenen een verbetering ten opzichte van de standard errors die tot nu toe in gebruik zijn. De afgeleide standard errors zijn van toepassing voor het geval dat de elementen van de test datamatrix, d.i. de datamatrix onder beschouwing, opgevat kunnen worden als onbekende constanten met enkel een meetfout. (Uit de multivariate statistiek zijn sinds lang standard errors bekend die het gevolg zijn van sampling errors.) Er wordt uitgelegd dat de afgeleide standard errors juist in het meest interessante geval, d.i. voor lage signaal-ruis-verhouding niet op een voor de hand liggende manier gebruikt kunnen worden voor het bepalen van de pseudorang van een matrix, omdat in de afleiding alleen de eerste orde term wordt meegenomen van een langzaam convergerende reeks. (Merk op dat de 'oude' standard errors voor dit doel waren afgeleid: zie Paragraaf 3.) Drie toepassingen van de afgeleide standard errors worden besproken. Eerst wordt het verband gelegd tussen eerder afgeleide standard errors in de scores en loadings van PCA die voorheen als verschillend beschouwd werden. Vervolgens worden standard errors voor multivariate calibratie besproken en uiteindelijk worden standard errors voor de eigenwaarden van GRAM afgeleid. Er wordt aangetoond dat een variantiereductie wordt bereikt als de datamatrix die voor de constructie van de factorruimte wordt gebruikt, is verkregen door middel van 'gesimuleerde' standaard additie in plaats van werkelijke standaard additie.

In Paragraaf 2 worden verschillende aspecten besproken die essentieel zijn voor de toepasbaarheid van pseudorang schattingsmethoden die gebaseerd zijn op de eigenwaarden van PCA van random matrices. Deze aspecten betreffen de afmeting van de matrix, de verhouding tussen het aantal rijen en kolommen, de signaal-ruis-verhouding en de verdeling van de meetfout. Asymptotische resultaten uit de statistische literatuur worden besproken en hun praktische nut is geëvalueerd met behulp van MC simulaties. Afwijkingen van 'ideaal' gedrag (bijv. voorspeld voor oneindig grote matrices) worden vergeleken met de intrinsieke variabiliteit van de eigenwaarden, d.i. de standard error. Er wordt aangetoond, dat tot een zeer lage signaal-ruis-verhouding de eigenwaarden die geassocieerd zijn met de residuen van de test datamatrix benaderd kunnen worden door de eigenwaarden van een random matrix, aldus leidend tot een geldige pseudorangschattingsmethode. Echter, de random matrix moet bij voorkeur hetzelfde aantal vrijheidsgraden als de test datamatrix bezitten in plaats van dezelfde afmeting zoals tot nu toe gebruikelijk is. Derhalve wordt een mogelijke modificatie van deze methodes voorgesteld. Verder blijkt dat onder tamelijk algemene voorwaarden de verdeling van de eigenwaarden van een random matrix voornamelijk bepaald wordt door de verhouding tussen het aantal rijen en kolommen van de matrix. De consequenties voor twee veelgebruikte pseudorang schattingsmethoden worden besproken.

In Paragraaf 3 worden verschillende aspecten besproken die essentieel zijn voor de toepasbaarheid van pseudorang schattingsmethoden die gebaseerd zijn een schatting van de grootte van de meetfout. Deze aspecten betreffen de steekproefverdeling van de test statistic, het aantal vrijheidsgraden dat voor de test gebruikt moet worden, de nauwkeurigheid van theoretische voorspellingen en de systematische fout die het gevolg is van toevallige meetfouten. De aspecten worden besproken met betrekking tot drie parametrische methoden. De eerste methode is (Methode A) is gebaseerd op de standard error in de diagonaalelementen van de bovendriehoeksvorm van de matrix, de tweede methode (Methode B) is gebaseerd op de eerder afgeleide standard error in de eigenwaarden van PCA en de derde methode (een *t*-test) is gebaseerd op de standard error in de singuliere waarden. Er wordt aangetoond dat methodes A en B problematisch zijn omdat de steekproefverdeling van de test statistic niet bekend is. Vergelijking met Malinowski's *F*-test laat zien dat de t-test efficiënt gebruik maakt van de extra kennis. Tenslotte wordt aangetoond dat een plot van de singuliere waarden een veelbelovende pseudorangschattingsmethode geeft in het algemene geval dat de extra kennis niet voorhanden is. Deze 'nieuwe' grafische methode is derhalve het natuurlijke complement van de *t*-test.

In Paragraaf 4 worden verschillende formuleringen van GRAM met elkaar vergeleken en een enigszins verschillend eigenwaardeprobleem afgeleid. Deze herformulering leidt tot eenvoudigere reconstructieformules voor de zuivere component profielen. Verder wordt de bespreking van twee karakteristieke problemen van GRAM vergemakkelijkt, d.i. de aanzienlijke kans op een complexe en ontaarde oplossing. Het wordt aangetoond dat, mits de data aan de veronderstellingen van het model voldoet, een complexe oplossing - in tegenstelling tot ontaarding - niet kan optreden voor de componenten

die zowel in het onbekende als in het bekende mengsel aanwezig zijn. Mogelijke oplossingen voor deze problemen worden voorgesteld.

In Paragraaf 5 wordt het effect van random meetfouten op de geschatte eigenwaarden van GRAM behandeld. Om te beginnen wordt uitgelegd dat random fouten niet alleen leiden tot een spreiding in de geschatte parameters (variantie) maar tevens tot een verschuiving (bias) als de schattingsprocedure neerkomt op een niet-lineaire transformatie van de data. Vervolgens worden nieuwe grootheden ingevoerd om van de eerder ingevoerde herformulering tot een geschikt waarschijnlijkheidsmodel te komen. De bias in de eigenwaarden wordt afgeleid door bekende resultaten uit de kleinste kwadraten (KK) theorie te combineren met Malinowski's 'error theory for factor loadings resulting from combination target factor analysis'. De variantie wordt afgeleid door foutenvoortplanting toe te passen op de eerder ingevoerde herformulering. De afleidingen zijn uitgevoerd voor drie verschillende procedures die momenteel gebruikt worden voor het construeren van de factorruimte, d.i. Lorber's methode, de generalisering van Sánchez en Kowalski en van Wilson, Sánchez en Kowalski. Er wordt aangetoond dat, afhankelijk van de grootte van de eigenwaarde, grote verschillen in performantie moeten worden verwacht. Een biascorrectie techniek wordt voorgesteld die de bias effectief elimineert indien de schatting van de bias geschikt is. Er wordt aangetoond dat de voorspellingen bevredigend zijn tot het niveau van de detectielimiet. Verder lijkt de kwaliteit van de resultaten niet gevoelig te zijn voor een bepaalde keuze van de dimensie van de factorruimte.

In Paragraaf 6 wordt de praktische implementatie van GRAM vanuit twee invalshoeken besproken, d.i. het computerprogramma dat ontwikkeld moet worden en de data die momenteel in de praktijk gegenereerd worden. Uit paragrafen 4 en 5 is duidelijk geworden dat het computerprogramma twee kritieke stappen bevat, te weten de constructie van de factorruimte en de oblique rotatie van de factoren. De eerste stap is een KK probleem terwijl de tweede stap een eigenwaardeprobleem is. Het eerste doel van dit gedeelte is duidelijk te maken dat in de literatuur tot nu toe te veel aandacht is geschonken aan de numerieke eigenschappen van de verschillende formuleringen. Het is belangrijk om in te zien dat de numerieke problemen ondergeschikt zijn aan de statistische problemen. Het tweede doel van dit gedeelte is het bespreken van de relevantie van theoretische resultaten (vooral die verkregen in paragraaf 5) als GRAM wordt toegepast op data die behept is met modelfouten.

REV	Reduced EigenValue
RMS	Root Mean Square
RNA	RiboNucleic Acid
RR	Ridge Regression
SEL	SELectivity
SEN	SENsitivity
SEP	Standard Eigenvalue Problem
SIM	'SIMilariteit'
SV	Singular Value
SVD	Singular Value Decomposition
SNR	Signal-to-Noise Ratio
TLC	Thin Layer Chromatography
Tr	Trace of matrix
UV	UlraViolet
UV-vis	UltraViolet-visible
var	variance
WFA	Window Factor Analysis
XE	eXtracted Error (function)
2ND	2ND derivative method

LIST OF ABBREVIATIONS

A	Adenine
AFA	Abstract Factor Analysis
ALS	Alternating Least Squares
ANOVA	ANalysis Of VAriance
AU	Absorption Unit
BTT	Bilinear Target Testing
C	Cytidine
CGRAM	Corrected Generalized Rank Annihilation Method
CLS	Corrected Least Squares
CLT	Central Limit Theorem
COC	'Correlatie Overeenkomstige Clusters'
cond	condition number of matrix
COV	COVariance
CPU	Central Processing Unit
DA	Diode Array
DF	Degree(s) of Freedom
EFA	Evolving Factor Analysis
EV	EigenValue
EVD	EigenValue Decomposition
ER	Eigenvalue Ratio
G	Guanine
GC	Gas Chromatography
GEP	Generalized Eigenvalue Problem
GRAM	Generalized Rank Annihilation Method
GSAM	Generalized Standard Addition Method
HPLC	High Performance Liquid Chromatography
ICP	Inductively Coupled Plasma
IE	Imbedded Error (function)
IMSL	International Mathematical and Statistical Library
IND	INDicator (function)
IR	InfraRed
ITTFA	Iterative Target Testing Factor Analysis
KF	Kalman Filter
KK	'Kleinste Kwadraten'
LC	Liquid Chromatography
LEV	Log EigenValue
LP	Liquid Phase
LS	Least Squares
MC	Monte Carlo
MDA	Multivariate Data Analysis
MLR	Multiple Linear Regression
MS	Mass Spectroscopy
MSE	Mean Square Error
NBRA	Non-Bilinear Rank Annihilation
NDF	Number of Degrees of Freedom
NIPALS	Non-linear Iterative Partial Least Squares
OES	Optical Emission Spectroscopy
PARAFAC	PARAllel profiles FACtor analysis
PC	Principal Component
PCA	Principal Component Analysis
PCR	Principal Component Regression
PDF	Probability Distribution Function
PI	Performance Index
PLS	Partial Least Squares
PRESS	Predicted Residual Error Sum of Squares
RAFA	Rank Annihilation Factor Analysis
RBL	Residual BiLinearization
RE	Real Error (function)

$\mathbf{Q}_R$	row augmented matrix in GRAM (generalization of Wilson, Sánchez and Kowalski)
$\mathbf{Q}_{UV}$	low rank approximation of **Q** using **U** and **V** for projection in GRAM
q	uniqueness
R	matrix of responses for calibration samples (first-order data)
R	cross-validation ratio (Wold)
R_{ij}	chromatographic resolution of components i and j
r	vector of responses for unknown sample (first-order data)
r	number of rows of (data) matrix; response for unknown sample (zero-order data)
S	sample covariance of random vector **x**; matrix of sensitivities; matrix of absorbances
S	number of spectra
s	minimum of number of rows r and columns c; sensitivity (zero-order data)
T	transformation matrix; matrix of right eigenvectors of reformulation of standard eigenvalue problem of GRAM
t	elution time; Student's t-variable
U	matrix of left singular vectors
V	matrix of right singular vectors
W	matrix of left eigenvectors of standard eigenvalue problem of GRAM
W	cross-validation ratio (Eastment and Krzanowski); number of wavelengths
$w_{½}$	peakwidth at half height
X	normalized column profiles of data matrix (second-order data)
x	random vector
Y	normalized row profiles of data matrix (second-order data)
Z	matrix of right eigenvectors of standard eigenvalue problem of GRAM; matrix of right eigenvectors of non-square generalized eigenvalue problem of GRAM
$\mathbf{Z}_G$	matrix of right eigenvectors of generalized eigenvalue problem of GRAM using **F** and **G** for the projection
$\mathbf{Z}_V$	matrix of right eigenvectors of generalized eigenvalue problem of GRAM using **U** and **V** for the projection
$\mathbf{Z}^*$	matrix of right eigenvectors of standard eigenvalue problem of GRAM

Symbols beginning with a Greek letter

α	significance level
β	base of computer arithmetic
δ	Kronecker delta
ϵ	machine precision
$\mathbf{\Lambda}$	eigenvalue matrix of PCA
λ	eigenvalue of PCA
μ	population mean of random vector **x**
μ	position of chromatographic peak
ν	number of degrees of freedom
$\mathbf{\Pi}$	eigenvalue matrix of GRAM (relative concentrations)
π	eigenvalue of GRAM
ρ	linear correlation coefficient; signal-to-noise ratio
$\mathbf{\Sigma}$	population covariance of random vector **x**
σ	standard deviation of measurement error; standard error in estimated quantity; standard deviation of chromatographic peak (width)
$\mathbf{\Theta}$	singular value matrix
θ	singular value
$\mathbf{\Phi}$	matrix of variance and covariance factors (see Erratum on page 119)
$\mathbf{\Psi}$	matrix of bias factors (see Erratum on page 119)

Symbols beginning with a Roman letter

$\mathbf{A}$	score matrix
A	pseudorank
$\mathbf{B}$	loading matrix; matrix of background responses for calibration samples (first-order data)
$\mathbf{b}$	vector of background responses for unknown sample (first-order data)
b	bias
$\mathbf{C}$	matrix of concentrations of calibration samples (first-order data); matrix of elution profiles
$\mathbf{C}_M$	matrix of concentrations of unknown sample (second-order data)
$\mathbf{C}_N$	matrix of concentrations of calibration sample (second-order data)
$\mathbf{c}$	vector of concentrations of unknown sample (first-order data)
c	number of columns of (data) matrix; concentration of unknown sample (zero-order data)
$\mathbf{D}_Q$	Q-mode cross-product matrix $\mathbf{MM}^T$
$\mathbf{D}_R$	R-mode cross-product matrix $\mathbf{M}^T\mathbf{M}$
d	divergence coefficient (ratio of number of rows and columns of a matrix)
$\mathbf{E}$	matrix of response residuals; matrix of estimated errors in data matrix
$\mathbf{F}$	loading matrix; orthogonal base for column space
F	number of PCs in truncated SVD (estimate of pseudorank); Fisher's F-variable
$\mathbf{G}$	orthogonal base for row space
$\mathbf{H}$	denormalized column profiles of data matrix (second-order data)
H, h	height of chromatographic peak
H_0	null-hypothesis
H_1	alternative hypothesis
$\mathbf{I}$	identity matrix
I	number of calibration samples; number of rows of (data) matrix
J	number of responses; number of columns of (data) matrix
$\mathbf{K}$	calibration matrix in GSAM
K	pseudorank; number of chemical species
$\mathbf{M}$	(data) matrix; matrix of responses for unknown sample (second-order data)
$\mathbf{M}_{FG}$	low rank approximation of $\mathbf{M}$ using $\mathbf{F}$ and $\mathbf{G}$ for projection in GRAM
$\mathbf{M}_{UV}$	low rank approximation of $\mathbf{M}$ using $\mathbf{U}$ and $\mathbf{V}$ for projection in GRAM
$\mathbf{m}$	sample mean of random vector $\mathbf{x}$
$\mathbf{N}$	matrix of responses for calibration sample (second-order data)
$\mathbf{N}_{FG}$	low rank approximation of $\mathbf{N}$ using $\mathbf{F}$ and $\mathbf{G}$ for projection in GRAM
$\mathbf{N}_{UV}$	low rank approximation of $\mathbf{N}$ using $\mathbf{U}$ and $\mathbf{V}$ for projection in GRAM
$\mathbf{N}_0$	matrix of initial concentrations in GSAM
$\mathbf{N}_{est}$	final estimate of pseudorank
N	sample size; normalization constant
$N(\mu,\sigma)$	normally distributed number with mean μ and standard deviation σ
n	dimension of PC model
n^*	pseudorank
$\mathbf{O}$	null matrix
p	number of components of random vector $\mathbf{x}$
$\mathbf{Q}$	matrix of volume corrected responses in GSAM; sum matrix $\mathbf{M}+\mathbf{N}$ in GRAM (generalization of Sánchez and Kowalski)
$\mathbf{Q}_C$	column augmented matrix in GRAM (generalization of Wilson, Sánchez and Kowalski)

CURRICULUM VITAE

Klaas Faber werd op 25 september 1957 geboren te Aarle-Rixtel. In mei 1975 haalde hij zijn eindexamen atheneum B aan het Carolus Borromeus College te Helmond. In dat zelfde jaar begon hij aan de Katholieke Universiteit van Nijmegen met de studie natuurkunde. Deze studie werd in januari 1976 afgebroken. Na een periode in loondienst gewerkt te hebben, begon hij in 1977 met de studie scheikunde. Deze studie werd in mei 1981 gestaakt. In september 1982 begon hij aan een opleiding MO A natuurkunde en scheikunde aan de Katholieke Leergangen te Tilburg. Deze opleiding werd in juni 1983 afgesloten. In september 1983 werd de studie scheikunde aan de Katholieke Universiteit hervat. Tijdens de doctoraalfase deed hij een (tot hoofdvak uitgebreid) bijvak Vaste Stof Chemie (Prof. P. Bennema), een hoofdvak Analytische Chemie (Prof. G. Kateman) en een bijvak bestaande uit vier doctoraaltentamens (Algemeen Gedeelte). In 1990 sloot hij deze studie af.

Hij trad per 1 september 1990 in dienst bij de Katholieke Universiteit Nijmegen in de functie van assistent in opleiding (AIO) en werd verbonden aan de afdeling Analytische Chemie. In deze hoedanigheid werkte hij onder leiding van Prof. G. Kateman en Dr. L. Buydens aan diverse aspecten van multivariate data-analyse in de Chemometrie. Het hierbij voor ogen staande doel bestond uit het ontwikkelen van theoretisch inzicht in zowel nieuwe (chemometrische) als wel oude (statistische) methoden. De resultaten van dat onderzoek worden in dit proefschrift beschreven.