The following full text is a publisher’s version.

For additional information about this publication click this link.
http://hdl.handle.net/2066/112325

Please be advised that this information was generated on 2017-07-16 and may be subject to change.
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 $M (r \times c)$ is given by

$$M = U \Theta V^T$$

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

---

*The cross-product matrix $M^T M$ 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:

\[
\begin{align*}
D_R &= M^T M = V \Lambda V^T \\
D_Q &= MM^T = U \Lambda U^T
\end{align*}
\]

(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 \( \Lambda \). 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.
EIGENVALUES OF A CROSS-PRODUCT MATRIX

EIGENVALUES OF A CROSS-PRODUCT MATRIX 497

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 \(M\) in row \(i\) and column \(k\) this gives \(M_{ik} = \tilde{M}_{ik} + \delta M_{ik} = \tilde{M}_{ik} + dM_{ik}\). 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 \(\text{COV}(M_{ik}, M_{jl}) = E[\delta M_{ik} d M_{jl}] = |\tilde{\sigma}_{M_{ik}} \cdot \tilde{\sigma}_{M_{jl}}|\). 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 \(x\) with population mean \(\mu = E[x]\) and population covariance \(\Sigma = E[(x - \mu)(x - \mu)^T]\) is observed by randomly drawing a sample \(\{x_1, \ldots, x_N\}\). The sample mean \(m = \Sigma_i x_i\) and sample covariance \(S = [1/(N-1)] \Sigma_i (x_i - m)(x_i - m)^T\) are efficient estimators of the population parameters. The eigenvectors of \(\Sigma\) and \(S\) 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 \(x\) 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 \(\mu\) and covariance \(\Sigma\), independent samples can be generated. Given a sample with known mean and covariance, say \(m_k\) and \(S_k\), different realizations for the data matrices are possible. This process is called \textit{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 \textit{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 \(M_{k1}\), can be partitioned into an errorless part and the measurement error as \(M_{k1} = M_{k1} + \delta M_{k1}\). 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 \(\mu\) and \(\Sigma\) are assessed by generating samples from the population (i.e. unconditional sampling). Conditional to the sample estimates \(m_k\) and \(S_k\) for the population values, data matrices \(M_{k1}, M_{k2}, \ldots\) can be generated in order to assess the variability within a sample (i.e. conditional sampling).](image-url)
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.\textsuperscript{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.\textsuperscript{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\textsuperscript{20} and multivariate calibration.\textsuperscript{23,44} 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.\textsuperscript{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\textsuperscript{23} have succeeded in deriving the distribution for the estimated calibration matrix \( \hat{K} \) but not for the matrix of estimated initial concentrations, \( \hat{N}_0 = (\hat{K}^T)^{-1}Q \), for the generalized standard addition method (GSAM), because the elements of \( \hat{N}_0 \) are a complicated function of the elements of \( \hat{K}^T \). 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](image_url)

**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 \( \mathbf{D}_R = \mathbf{M}^T \mathbf{M} \):

\[
\mathbf{D}_R = \mathbf{V} \mathbf{A} \mathbf{V}^T \\
\mathbf{A} = \mathbf{V}^T \mathbf{D}_R \mathbf{V}
\]

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

\[
\tilde{\mathbf{A}} + \delta \mathbf{A} = (\tilde{\mathbf{V}} + \delta \mathbf{V})^T (\tilde{\mathbf{D}}_R + \delta \mathbf{D}_R) (\tilde{\mathbf{V}} + \delta \mathbf{V})
\]

which can be multiplied out to yield

\[
\delta \mathbf{A} = \delta (\mathbf{V}^T) \tilde{\mathbf{D}}_R \tilde{\mathbf{V}} + \tilde{\mathbf{V}}^T \delta \mathbf{D}_R \tilde{\mathbf{V}} + \tilde{\mathbf{V}}^T \tilde{\mathbf{D}}_R \delta \mathbf{V}
\]

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

\[
d\mathbf{A} = d(\mathbf{V}^T) \tilde{\mathbf{D}}_R \tilde{\mathbf{V}} + \tilde{\mathbf{V}}^T d\mathbf{D}_R \tilde{\mathbf{V}} + \tilde{\mathbf{V}}^T \tilde{\mathbf{D}}_R d\mathbf{V}
\]

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)

\[
d\mathbf{A} = \tilde{\mathbf{V}}^T (d(\mathbf{M}^T) \tilde{\mathbf{M}} + \tilde{\mathbf{M}} + \tilde{\mathbf{M}}^T d\mathbf{M}) \tilde{\mathbf{V}}
\]

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

\[
d\mathbf{A} = 2 \tilde{\mathbf{V}}^T \tilde{\mathbf{M}}^T d\mathbf{M} \tilde{\mathbf{V}}
\]

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

$$\bar{\sigma}_{\lambda n} = 2\bar{\theta}_n \sum_{i=1}^{r} \bar{U}_{in} \sum_{k=1}^{c} \bar{V}_{kn} \bar{\sigma}_{M_{ik}}$$

(13)

Here $\bar{\theta}_n = \bar{\theta}_{nn}$. 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:

$$\text{COV}(\lambda_n, \lambda_m) = \mid \bar{\sigma}_{\lambda n} \cdot \bar{\sigma}_{\lambda m} \mid = 4\bar{\theta}_n \delta_{nm} \left( \sum_{i=1}^{r} \bar{U}_{in} \sum_{k=1}^{c} \bar{V}_{kn} \bar{\sigma}_{M_{ik}} \right) \left( \sum_{j=1}^{r} \bar{U}_{jm} \sum_{l=1}^{c} \bar{V}_{lm} \bar{\sigma}_{M_{jl}} \right)$$

(14)

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

$$\text{COV}(\lambda_n, \lambda_m) = 4\bar{\theta}_n \delta_{nm} \left( \sum_{i=1}^{r} \bar{U}_{in} \bar{U}_{im} \sum_{k=1}^{c} \bar{V}_{kn} \bar{\sigma}_{M_{ik}} \right) + \sum_{i=1}^{r} \sum_{j=1}^{r} \sum_{k=1}^{c} \sum_{l=1}^{c} \bar{U}_{in} \bar{U}_{jm} \bar{V}_{kn} \bar{V}_{lm} \rho_{ik, jl} \sigma_{M_{ia}} \sigma_{M_{jl}}$$

(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

$$\text{COV}(\lambda_n, \lambda_m) = 4\bar{\theta}_n \delta_{nm} \left( \sum_{i=1}^{r} \bar{U}_{in} \sum_{k=1}^{c} \bar{V}_{kn} \bar{\sigma}_{M_{ik}} \right)$$

(16)

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

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

(17)

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

$$\text{COV}(\lambda_n, \lambda_m) = 4\bar{\theta}_n \delta_{nm} \sum_{i=1}^{r} \bar{U}_{in} \bar{\sigma}_{M_{i}}$$

(18)

The standard errors of the eigenvalues are given as

$$\sigma_{\lambda n} = 2\bar{\theta}_n \sigma_M = 2\lambda_n^{1/2} \sigma_M$$

(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

$$\delta_{\lambda n} = 2\lambda_n^{1/2} \left( \sum_{i=1}^{c} U_{in} \sum_{k=1}^{c} V_{kn} \bar{\sigma}_{M_{ik}} \right)^{1/2}$$

(20)

$$\delta_{\lambda n} = 2\lambda_n^{1/2} \delta_{M}$$

(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:

$$\sigma_{\theta} = \sigma_{\lambda n} = \sigma_{M}$$

(22)

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

$$\delta_{\theta} = \delta_{M}$$

(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. 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 have successfully estimated the magnitude of the noise in voltammograms from the high-frequency components of the Fourier power spectrum. Hirsch et al. 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

\[ \hat{\sigma}_M = \left( \frac{\sum_{p=K+1}^{c} \lambda_p}{(r - K)(c - K)} \right)^{1/2} \]  

(24)

Here \( K \) denotes the pseudorank. Equation (24) is equivalent to the real error (RE) function of Malinowski. The number of degrees of freedom in the denominator is, however, found by different authors 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.

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. These standard errors are derived under the assumption that the variables are Gaussian-distributed in the population:

\[ \hat{\sigma}_{\lambda_n} = \sqrt{\frac{2}{N-1}} \lambda_n \]  

(25)

Here the population eigenvalue \( \lambda_n \) 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. Because of the restrictive assumption of Gaussian-distributed variables and large samples, these expressions have had little practical evaluation.

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

$$
S = \begin{bmatrix}
2.53 & 3.50 & 2.06 & 1.45 \\
3.50 & 5.05 & 2.86 & 2.02 \\
2.06 & 2.86 & 1.68 & 1.19 \\
1.45 & 2.02 & 1.19 & 0.86
\end{bmatrix}
$$

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 ($n = 10$ solutes on 226 liquid phases). This data matrix has been studied by a large number of investigators and some of the results have been reviewed by Joliffe. Of the 226 liquid phases (LPs), 13 are identified as outliers by Wold and Andersson 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 comes to the conclusion that five PCs are significant for the 213 $\times$ 10 matrix. Eastment and Krzanowski find four significant PCs using their modified cross-validation technique for the 212 $\times$ 10 matrix. Furthermore, Joliffe 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>
<thead>
<tr>
<th>$n$</th>
<th>$\lambda$</th>
<th>$\hat{\sigma}_M$ (equation (21))</th>
<th>$\sigma_\lambda$ (equation (25))</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>10</td>
<td>0.132</td>
<td>1.414</td>
</tr>
<tr>
<td>2</td>
<td>0.1</td>
<td>0.013</td>
<td>0.014</td>
</tr>
<tr>
<td>3</td>
<td>0.02</td>
<td>0.006</td>
<td>0.003</td>
</tr>
<tr>
<td>4</td>
<td>0.01</td>
<td>0.004</td>
<td>0.001</td>
</tr>
</tbody>
</table>
Table 2. Results of PCA for gas chromatography retention index data

<table>
<thead>
<tr>
<th>n</th>
<th>( \lambda )</th>
<th>( \sigma_\lambda ) (equation (21))</th>
<th>( \sigma_\lambda ) (equation (25))</th>
<th>( \sigma_\lambda ) (bootstrap)</th>
<th>( \sigma_\lambda ) (jack-knife)</th>
<th>CUM (%)</th>
<th>( \delta_M )</th>
<th>( R^a )</th>
<th>( R^b )</th>
<th>( W^c )</th>
<th>IND</th>
<th>ER</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>( 6.30 \times 10^7 )</td>
<td>( 0.18 \times 10^6 )</td>
<td>( 6.11 \times 10^6 )</td>
<td>( 6.17 \times 10^6 )</td>
<td>98.73</td>
<td>21.7</td>
<td>0.02</td>
<td>0.05(1)</td>
<td>494.98</td>
<td>0.255</td>
<td>1.35 \times 10^2</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>( 4.68 \times 10^7 )</td>
<td>( 1.52 \times 10^4 )</td>
<td>( 5.41 \times 10^4 )</td>
<td>( 5.38 \times 10^4 )</td>
<td>99.47</td>
<td>15.3</td>
<td>0.43</td>
<td>0.63(2)</td>
<td>4.95</td>
<td>0.222</td>
<td>2.91</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>( 1.61 \times 10^7 )</td>
<td>( 0.89 \times 10^4 )</td>
<td>( 1.79 \times 10^4 )</td>
<td>( 1.90 \times 10^4 )</td>
<td>99.72</td>
<td>11.8</td>
<td>0.60</td>
<td>0.79(2)</td>
<td>1.90</td>
<td>0.226</td>
<td>2.26</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>( 7.13 \times 10^4 )</td>
<td>( 5.94 \times 10^3 )</td>
<td>( 8.41 \times 10^3 )</td>
<td>( 8.73 \times 10^3 )</td>
<td>99.83</td>
<td>9.4</td>
<td>0.70</td>
<td>0.88(3)</td>
<td>0.92</td>
<td>0.258</td>
<td>1.75</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>( 4.07 \times 10^4 )</td>
<td>( 4.49 \times 10^3 )</td>
<td>( 6.35 \times 10^3 )</td>
<td>( 6.76 \times 10^3 )</td>
<td>99.89</td>
<td>7.4</td>
<td>0.83</td>
<td>0.96(2)</td>
<td>0.41</td>
<td>0.322</td>
<td>1.48</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>( 2.76 \times 10^4 )</td>
<td>( 3.69 \times 10^3 )</td>
<td>( 4.21 \times 10^3 )</td>
<td>( 4.73 \times 10^3 )</td>
<td>99.94</td>
<td>6.5</td>
<td>0.99</td>
<td>0.90(2)</td>
<td>0.54</td>
<td>0.434</td>
<td>1.65</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>( 1.67 \times 10^4 )</td>
<td>( 2.88 \times 10^3 )</td>
<td>( 1.89 \times 10^3 )</td>
<td>( 2.01 \times 10^3 )</td>
<td>99.96</td>
<td>5.6</td>
<td>—</td>
<td>1.06(3)</td>
<td>0.13</td>
<td>0.684</td>
<td>1.46</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>( 1.15 \times 10^4 )</td>
<td>( 2.38 \times 10^3 )</td>
<td>( 2.26 \times 10^3 )</td>
<td>( 2.86 \times 10^3 )</td>
<td>99.98</td>
<td>4.8</td>
<td>—</td>
<td>0.98(8)</td>
<td>0.11</td>
<td>1.360</td>
<td>1.55</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>( 7.38 \times 10^3 )</td>
<td>( 1.91 \times 10^3 )</td>
<td>( 0.72 \times 10^3 )</td>
<td>( 1.53 \times 10^3 )</td>
<td>99.99</td>
<td>4.1</td>
<td>—</td>
<td>1.11(6)</td>
<td>0.03</td>
<td>4.916</td>
<td>1.51</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>( 4.88 \times 10^3 )</td>
<td>( 1.55 \times 10^3 )</td>
<td>( 0.48 \times 10^3 )</td>
<td>( 1.04 \times 10^3 )</td>
<td>100.00</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td></td>
</tr>
</tbody>
</table>

\( a \) Reference 18 (213 \times 10). \( b \) Average of ten cross-validations (212 \times 10). \( c \) Reference 19 (212 \times 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. $R$ should be smaller than unity for a significant PC, but it is clear that the PC with $R = 0.99$ cannot have much predictive value. The next column gives the ratios obtained by randomly drawing subgroups ten times from the 212 x 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. $W$ should be smaller than unity for a significant PC, but it is clear that the PC with $W = 0.99$ cannot have much predictive value. The next column gives the ratios obtained by randomly drawing subgroups ten times from the 212 x 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. $W$ should be smaller than unity for a significant PC, but it is clear that the PC with $W = 0.99$ cannot have much predictive value. The next column gives the ratios obtained by randomly drawing subgroups ten times from the 212 x 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. $W$ should be smaller than unity for a significant PC, but it is clear that the PC with $W = 0.99$ cannot have much predictive value. The next column gives the ratios obtained by randomly drawing subgroups ten times from the 212 x 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. $W$ should be smaller than unity for a significant PC, but it is clear that the PC with $W = 0.99$ cannot have much predictive value.
gives the values $\rho_{1.2} = 0.988$, $\rho_{1.10} = 0.988$ and $\rho_{2.10} = 0.965$. 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\textsuperscript{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.\textsuperscript{27} Although straight lines are obtained for a standard deviation larger than 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 $\sigma_M$ in (24). A value of $\sigma_M = 11.8$ 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 (△), 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.](image-url)
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 $212 \times 10$ 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: 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 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) $A$ is a general $c \times c$ matrix with linear elementary divisors and (ii) $A$ and $B$ are matrices that satisfy $|A_{ij}| \leq 1$, $|B_{ij}| \leq 1$ and $\lambda_n$ is a simple eigenvalue of $A$ (i.e. $\lambda_n$ has multiplicity one, a condition usually met in PCA), then $\lambda_n(e)$ is an eigenvalue of $A + eB$ and

$$\lambda_n(e) = \lambda_n + k_n^{(1)} e + k_n^{(2)} e^2 + \cdots$$

is a convergent power series independent of the multiplicities of the other eigenvalues. Substituting $D_R = A$ and $dD_R = eB$ leads to the following expressions for the first two error terms in the expansion:

$$k_n^{(1)} e = 2 \Theta_{n-row} d \Theta_{n-col} = d \Lambda_{nn}$$

$$k_n^{(2)} e^2 = \sum_{p \neq n} \frac{d \Lambda_{pn} d \Lambda_{np}^T}{\lambda_n - \lambda_p}$$

Here the indices ‘$n$-row’ and ‘$n$-col’ denote the $n$th row and $n$th 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 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. 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 $M$ is decomposed as

\[ M = AF \]

The matrices $A(r \times c)$ and $F(c \times c)$ are called scores and loadings; the loadings represent the principal axes of $M$ and the scores represent the co-ordinates on this rotated system respectively. The relation with the SVD of $M$ is given by the identities $A = U\Theta$ and $F = V^T$. Expressions for the errors in the scores and loadings are derived by Malinowski and Roscoe and Hopke. 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.

The error in $M$ is expressed in the contributions from $A$ and $F$:

\[ dM = \tilde{A} dF + dA\tilde{F} \tag{30a} \]

Assuming that all measurement error resides in $A$ gives

\[ dM = dA\tilde{F} \tag{30b} \]

Assuming that all measurement error resides in $F$ gives

\[ dM = \tilde{A} dF \tag{30c} \]

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

\[ dA = \tilde{A} (d\Theta \tilde{\Theta}^{-1}) \tag{31a} \]

\[ dF = (d\Theta \tilde{\Theta}^{-1})\tilde{F} \tag{31b} \]

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

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

\[ \hat{\sigma}_{F_{ik}} = \lambda_n^{-1/2} \left( \sum_{i=1}^{r} (U_{in})^2 \hat{\delta}_{M_{ik}}^2 \right)^{1/2} \tag{32b} \]
and for homoscedastic noise the indices \( k \) and \( i \) are summed out to give

\[
\delta A = \delta_{A_{n \to m}} = \tilde{\delta}_M \tag{33a}
\]

\[
\delta F_{n \to m} = \lambda_n^{-1/2} \delta M \tag{33b}
\]

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 \( \lambda_n \) are monotonously decreasing, it is expected that the instability of the eigenvectors of \( M^T M \) 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 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. Quantitation always proceeds in two steps. In the first step the instrument responses are recorded from a set of calibration samples:

\[
R = SC + E \tag{34}
\]

where \( R \) is the matrix of calibration data responses, \( S \) is the matrix of sensitivities, \( C \) is the matrix of concentrations in the calibration set and \( E \) is the matrix of response residuals. \( R \) and \( E \) are dimensioned \( J \times I \), \( S \) is \( J \times K \) and \( C \) 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

\[
\hat{S} = RC^+. \tag{35}
\]

In the second step the instrument response is recorded for the unknown sample and fitted to the model

\[
r = \hat{S}c + e \tag{36a}
\]

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

\[
\hat{c} = \hat{S}^+r = CR^+r \tag{36a}
\]

The advantage of substituting for \( \hat{S}^+ \) becomes clear from the expression for the individual
components:

\[ \tilde{c}_n = C_{n-row}R^*r \]  

(36b)

Only \( C_{n-row} \), i.e. the row in \( C \) 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 \( CC^+ \) equals the \( K \times K \) identity matrix \( I_K \). 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 \( R \) and \( r \) are replaced by \( R - B \) and \( r - b \) respectively. A model for which the complete concentration matrix \( C \) 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)

\[ \tilde{c}_n + \delta c_n = (\tilde{C}_{n-row} + \delta C_{n-row}) (\tilde{R} + \delta R)^+ (\tilde{r} + \delta r) \]  

(37)

Depending on how the pseudoinverse of \( \tilde{R} + \delta R \) 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 \( (\tilde{R} + \delta R)^+ \) according to the SVD

\[ (\tilde{R} + \delta R)^+ = \tilde{V}(\tilde{\Theta} + \delta \Theta)^+ \tilde{U}^T = \tilde{V}(\tilde{\Theta} + \delta \Theta)^{-1} \tilde{U}^T \]  

(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 \( \theta_{K+1} \). 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

\[ dc = [-S^+ (dR - dB) + dC] C^+ c + S^+ (dr - db) \]  

(39)

and

\[ dc = [-C(R - B)^+ (dR - dB) + dC] C^+ C (R - B)^+ (r - b) + C (R - B)^+ (dr - db) \]  

(40a)

for the total and partial calibration approaches respectively. Since (40a) contains the matrix \( C^+ \), it can only be evaluated if the complete matrix \( C \) 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,

\[ dc = [-C(\tilde{R} - \tilde{B})^+ (dR - dB) + dC] \tilde{C}^+ \tilde{C}(\tilde{R} - \tilde{B})^+ (\tilde{r} - \tilde{b}) + \tilde{C}(\tilde{R} - \tilde{B})^+ (dr - db) \]  

(40b)

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

\[ \tilde{V}(\tilde{\Theta} + \delta \Theta)^{-1} \tilde{U}^T \approx \tilde{V}(\tilde{\Theta} + d\Theta)^{-1} \tilde{U}^T \approx \tilde{V}(\tilde{\Theta}^{-1} - d\Theta \tilde{\Theta}^{-2}) \tilde{U}^T \]  

(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.\textsuperscript{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.,\textsuperscript{40} where the procedure of Lorber and Kowalski is modified by multiplying out (37) using \((\hat{R} + \delta R)^+ = \hat{R}^+ + \delta (R^+)\). The expressions of Karstang et al.\textsuperscript{40} differ from those of Bauer et al.\textsuperscript{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.\textsuperscript{15} We will restrict ourselves to the discussion of a method developed for bilinear data. For non-bilinear data, non-bilinear rank annihilation (NBRA)\textsuperscript{41} and residual bilinearization (RBL)\textsuperscript{30} should be more appropriate. (RBL assumes that the residuals rather than the responses are bilinear.)

Ho et al.\textsuperscript{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\textsuperscript{43} to a direct method which is generalized by Sánchez and Kowalski\textsuperscript{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:

\[
M = X C_M Y^T = \bar{X} \bar{C}_M \bar{Y}^T + E_{M, \text{real}} \quad (42a)
\]

\[
N = X C_N Y^T = \bar{X} \bar{C}_N \bar{Y}^T + E_{N, \text{real}} \quad (42b)
\]

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 \(C_M\) and \(C_N\) are \(K \times K\) diagonal matrices with calibration factors. The \(S \times W\) error matrices \(E_{M, \text{real}}\) and \(E_{N, \text{real}}\) denote the difference between the experimental and the errorless data, i.e. the real error.\textsuperscript{28}

In the most general case (i.e. both samples contain unique components) the matrices \(\bar{C}_M\) and \(\bar{C}_N\) 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 \(\bar{C}_M + \bar{C}_N\) contains no zeros:

\[
Q = M + N = \bar{X} (\bar{C}_M + \bar{C}_N) \bar{Y}^T + E_{Q, \text{real}} \quad (43)
\]

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

\[
Q = U_Q \Theta_Q V_Q^T + E_{Q, \text{extr}} \quad (44)
\]

Here the error matrix \(E_{Q, \text{extr}}\) denotes the difference between the experimental and the reproduced sum matrix, i.e. the extracted error.\textsuperscript{28} The resulting prediction equation is an \(F \times F\)
standard eigenvalue problem

\[(U_Q^T M V_Q \Theta_Q^{-1}) Z = Z \Pi\]  

(45)

where \( \Pi \) is a diagonal matrix of eigenvalues, related to the calibration factors as \( \Pi = C_M (C_M + C_N)^{-1} \), and \( Z \) 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

\[X(C_M + C_N) = U_Q Z\]  

(46a)

\[Y^T = Z^{-1} \Theta_Q V_Q^T\]  

(46b)

\[C_M = C_N \Pi(I_F - \Pi)^{-1}\]  

(46c)

where \( I_F \) 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 as well as for the eigenvalues obtained by GRAM. Inspection of these error estimates shows that only the error in the decomposed data matrix \( Q \) 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 \( M \) and \( N \), 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):

\[\hat{\sigma}_{\lambda} = \frac{\hat{\sigma}_M}{2 \pi_n + 2 \pi_n^2 \sum_{p=1}^{F} \frac{(Z_{pn})^2}{\lambda_{Q,p}^2}}^{1/2}\]  

(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 \( Q \) is obtained by performing ‘real’ standard additions in the unknown sample, the standard errors are given by (see Appendix III)

\[\hat{\sigma}_{\lambda} = \frac{\hat{\sigma}_M}{1 + \pi_n^2 \sum_{p=1}^{F} \frac{(Z_{pn})^2}{\lambda_{Q,p}^2}}^{1/2}\]  

(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. 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 coefficient (left lower corner) for UV spectra and elution profiles. Spectra and elution profiles are normalized.

<table>
<thead>
<tr>
<th>Component</th>
<th>UV spectra</th>
<th>Elution profiles</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A</td>
<td>C</td>
</tr>
<tr>
<td>Adenine</td>
<td>1</td>
<td>0.70</td>
</tr>
<tr>
<td>Cytidine</td>
<td>-0.24</td>
<td>1</td>
</tr>
<tr>
<td>Guanine</td>
<td>0.79</td>
<td>0.12</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>Experiment</th>
<th>SNR</th>
<th>Adenine</th>
<th>Cytidine</th>
<th>Guanine</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2000</td>
<td>1000</td>
<td>1000</td>
<td>1000</td>
</tr>
<tr>
<td>2</td>
<td>10</td>
<td>1000</td>
<td>5</td>
<td>1000</td>
</tr>
<tr>
<td>3</td>
<td>6</td>
<td>1000</td>
<td>3</td>
<td>1000</td>
</tr>
</tbody>
</table>

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 10000 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>
<thead>
<tr>
<th>n</th>
<th>$\lambda$ (equation (10))</th>
<th>$\sigma_\lambda^a$ (MC)</th>
<th>$\sigma_\lambda^b$ (MC)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3.79 × 10³</td>
<td>6.15 × 10³</td>
<td>6.11 × 10³</td>
</tr>
<tr>
<td>2</td>
<td>1.25 × 10⁶</td>
<td>1.12 × 10³</td>
<td>1.05 × 10³</td>
</tr>
<tr>
<td>3</td>
<td>1.84 × 10²</td>
<td>4.29 × 10²</td>
<td>4.25 × 10²</td>
</tr>
<tr>
<td>4</td>
<td>2.84 × 10</td>
<td>5.32</td>
<td>5.32</td>
</tr>
<tr>
<td>5</td>
<td>2.66 × 10</td>
<td>6.29</td>
<td>6.29</td>
</tr>
<tr>
<td>6</td>
<td>2.39 × 10</td>
<td>4.88</td>
<td>4.88</td>
</tr>
</tbody>
</table>

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

* Monte Carlo estimate from 10⁴ trials. a Monte Carlo estimate from 10² 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. 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 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 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>
<thead>
<tr>
<th>n</th>
<th>λ</th>
<th>σ_λ^a (equation (10))</th>
<th>σ_λ (equation (21))</th>
<th>σ_λ^a (MC)</th>
<th>σ_λ (MC)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.79 x 10^7</td>
<td>6.67 x 10^2</td>
<td>4.23 x 10^3</td>
<td>4.19 x 10^3</td>
<td>4.35 x 10^3</td>
</tr>
<tr>
<td>2</td>
<td>4.46 x 10^3</td>
<td>4.09 x 10^2</td>
<td>6.68 x 10^2</td>
<td>6.65 x 10^2</td>
<td>6.53 x 10^2</td>
</tr>
<tr>
<td>3</td>
<td>4.82 x 10^2</td>
<td>5.16 x 10^2</td>
<td>6.94</td>
<td>5.67</td>
<td>5.49</td>
</tr>
<tr>
<td>4</td>
<td>2.72 x 10^2</td>
<td>5.20 x 10^2</td>
<td>5.21</td>
<td>2.40</td>
<td>2.03</td>
</tr>
<tr>
<td>5</td>
<td>2.68 x 10^2</td>
<td>5.19 x 10^2</td>
<td>5.18</td>
<td>1.81</td>
<td>1.73</td>
</tr>
<tr>
<td>6</td>
<td>2.33 x 10^2</td>
<td>5.19 x 10^2</td>
<td>4.83</td>
<td>1.52</td>
<td>1.40</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>n</th>
<th>λ</th>
<th>σ_λ^a (equation (10))</th>
<th>σ_λ (equation (21))</th>
<th>σ_λ^a (MC)</th>
<th>σ_λ (MC)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.79 x 10^7</td>
<td>6.67 x 10^2</td>
<td>4.23 x 10^3</td>
<td>4.19 x 10^3</td>
<td>4.35 x 10^3</td>
</tr>
<tr>
<td>2</td>
<td>4.46 x 10^3</td>
<td>4.09 x 10^2</td>
<td>6.68 x 10^2</td>
<td>6.65 x 10^2</td>
<td>6.53 x 10^2</td>
</tr>
<tr>
<td>3</td>
<td>3.44 x 10^2</td>
<td>5.19 x 10^2</td>
<td>5.86</td>
<td>3.16</td>
<td>3.07</td>
</tr>
<tr>
<td>4</td>
<td>2.70 x 10^2</td>
<td>5.20 x 10^2</td>
<td>5.20</td>
<td>2.05</td>
<td>1.80</td>
</tr>
<tr>
<td>5</td>
<td>2.64 x 10^2</td>
<td>5.19 x 10^2</td>
<td>5.14</td>
<td>1.67</td>
<td>1.48</td>
</tr>
<tr>
<td>6</td>
<td>2.24 x 10^2</td>
<td>5.19 x 10^2</td>
<td>4.73</td>
<td>1.45</td>
<td>1.37</td>
</tr>
</tbody>
</table>

^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

<table>
<thead>
<tr>
<th>n</th>
<th>( \lambda )</th>
<th>( \sigma_{\lambda} ) (equation (21))</th>
<th>( \sigma_{\lambda} ) (MC)</th>
<th>( \lambda )</th>
<th>( \sigma_{\lambda} ) (equation (21))</th>
<th>( \sigma_{\lambda} ) (MC)</th>
<th>( \lambda )</th>
<th>( \sigma_{\lambda} ) (equation (21))</th>
<th>( \sigma_{\lambda} ) (MC)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3.79 x 10^7</td>
<td>6.15 x 10^3</td>
<td>6.10 x 10^3</td>
<td>1.79 x 10^7</td>
<td>4.23 x 10^3</td>
<td>4.19 x 10^3</td>
<td>1.79 x 10^7</td>
<td>4.23 x 10^3</td>
<td>4.19 x 10^3</td>
</tr>
<tr>
<td>2</td>
<td>1.25 x 10^6</td>
<td>1.12 x 10^3</td>
<td>1.13 x 10^3</td>
<td>4.46 x 10^5</td>
<td>6.68 x 10^2</td>
<td>6.72 x 10^2</td>
<td>4.46 x 10^5</td>
<td>6.68 x 10^2</td>
<td>6.72 x 10^2</td>
</tr>
<tr>
<td>3</td>
<td>1.84 x 10^5</td>
<td>4.29 x 10^2</td>
<td>4.27 x 10^2</td>
<td>5.60 x 10^2</td>
<td>7.48</td>
<td>5.60</td>
<td>3.82 x 10^2</td>
<td>6.13</td>
<td>2.95</td>
</tr>
<tr>
<td>4</td>
<td>3.08 x 10</td>
<td>5.55</td>
<td>2.15</td>
<td>3.07 x 10</td>
<td>5.54</td>
<td>2.10</td>
<td>3.06 x 10</td>
<td>5.53</td>
<td>1.82</td>
</tr>
<tr>
<td>5</td>
<td>2.73 x 10</td>
<td>5.23</td>
<td>1.58</td>
<td>2.56 x 10</td>
<td>5.05</td>
<td>1.55</td>
<td>2.41 x 10</td>
<td>4.91</td>
<td>1.44</td>
</tr>
<tr>
<td>6</td>
<td>2.29 x 10</td>
<td>1.32</td>
<td>0.79</td>
<td>2.30 x 10</td>
<td>1.30</td>
<td>0.79</td>
<td>2.29 x 10</td>
<td>1.24</td>
<td>0.79</td>
</tr>
</tbody>
</table>

### Table 9. Standard errors in factor scores and loadings

<table>
<thead>
<tr>
<th>n</th>
<th>( \sigma_{\lambda} ) (MC) (equation (33b))</th>
<th>( \sigma_{\lambda} ) (MC) (equation (33b))</th>
<th>( \sigma_{\lambda} ) (MC) (equation (33b))</th>
<th>( \sigma_{\lambda} ) (MC) (equation (33b))</th>
<th>( \sigma_{\lambda} ) (MC) (equation (33b))</th>
<th>( \sigma_{\lambda} ) (MC) (equation (33b))</th>
<th>( \sigma_{\lambda} ) (MC) (equation (33b))</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.5010(7)</td>
<td>8.127 x 10^{-5}</td>
<td>8.027(17) x 10^{-5}</td>
<td>0.5009(7)</td>
<td>1.182 x 10^{-4}</td>
<td>1.167(3) x 10^{-4}</td>
<td>0.5009(7)</td>
</tr>
<tr>
<td>2</td>
<td>0.4986(10)</td>
<td>4.465 x 10^{-4}</td>
<td>4.378(17) x 10^{-4}</td>
<td>0.4940(10)</td>
<td>7.489 x 10^{-4}</td>
<td>7.274(28) x 10^{-4}</td>
<td>0.4936(12)</td>
</tr>
<tr>
<td>3</td>
<td>0.4869(15)</td>
<td>1.176 x 10^{-3}</td>
<td>1.120(4) x 10^{-3}</td>
<td>0.6375(22)</td>
<td>7.203 x 10^{-2}</td>
<td>9.339(33) x 10^{-2}</td>
<td>0.8606(48)</td>
</tr>
</tbody>
</table>
Table 10. Summary of quantitative solution of GRAM analysis: eigenvalues and their standard errors

<table>
<thead>
<tr>
<th>n</th>
<th>Unknown sample</th>
<th>Standard addition</th>
<th>Expected eigenvalue</th>
<th>'Simulated' standard addition (all errors × 10^-4)</th>
<th>'Real' standard addition (all errors × 10^-4)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>( \pi ) ( \sigma_\pi ) ( \sigma_\pi^a ) ( \sigma_\pi^b )</td>
<td>( \pi ) ( \sigma_\pi ) ( \sigma_\pi^a ) ( \sigma_\pi^b )</td>
</tr>
<tr>
<td>1</td>
<td>1100</td>
<td>1000</td>
<td>0.52381</td>
<td>0.52376 2.48 2.46 2.44</td>
<td>0.52372 4.01 3.93 3.81</td>
</tr>
<tr>
<td>2</td>
<td>1000</td>
<td>1000</td>
<td>0.50000</td>
<td>0.49991 1.71 1.69 1.62</td>
<td>0.49986 2.73 2.69 2.59</td>
</tr>
<tr>
<td>3</td>
<td>900</td>
<td>1000</td>
<td>0.47368</td>
<td>0.47445 3.55 3.52 3.51</td>
<td>0.47484 5.54 5.51 5.40</td>
</tr>
</tbody>
</table>

*Monte Carlo estimate from \( 10^4 \) trials. \(^b\)Monte Carlo estimate from \( 10^2 \) trials.
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 $I_c$ is a constant matrix, so

$$dI_c = 0_c.$$  \hspace{1cm} (49)
where $0_c$ denotes the $c \times c$ null matrix. The eigenvectors are orthonormal:

$$V^TV = I_c$$  (50)

Taking derivatives gives

$$d(V^TV) = d(V^T)\tilde{V} + \tilde{V}^TdV$$  (51)

Combining (49)–(51) gives

$$d(V^T)\tilde{V} + \tilde{V}^TdV = 0_c$$  (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 $\tilde{A}$ yields

$$d(V^T)\tilde{V}\tilde{A} + \tilde{V}^TdV\tilde{A} = 0_c$$  (53)

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

$$\tilde{D}_R(\tilde{V} + dV) = (\tilde{V} + dV)\tilde{A}$$  (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

$$\tilde{D}_RdV = dV\tilde{A}$$  (55)

Inserting (55) in (53) finally results in

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

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

Premultiplication of the SVD of $M$ by $U^T$ and postmultiplication by $V$ gives

$$\Theta = U^T MV$$  (57)
The error matrix \( d\Theta \) can therefore be expanded as
\[
d\Theta = d(U^T)\hat{M}\hat{V} + \hat{U}^T dM\hat{V} + \hat{U}^T\hat{M}dV
\] (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):
\[
U^TU = I_c \quad (59)
\]
\[
d(U^TU) = d(U^T)\hat{U} + \hat{U}^TdU = 0_c \quad (60)
\]
\[
d(U^T)\hat{U} + \hat{U}^TdU = 0_c \quad (61)
\]
\[
d(U^T)\hat{U}\hat{\Theta} + \hat{U}^TdU\hat{\Theta} = 0_c \quad (62)
\]
\[
\hat{M}(\hat{V} + dV) = (\hat{U} + dU)\hat{\Theta} \quad (63)
\]
\[
\hat{M}dV = dU\hat{\Theta} \quad (64)
\]
\[
d(U^T)\hat{M}\hat{V} + \hat{U}^T\hat{M}dV = 0_c \quad (65)
\]

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
\[
d\Theta = \hat{U}^TdM\hat{V} \quad (66)
\]

The eigenvalues are the squares of the singular values, so
\[
d\Delta = 2\hat{\Theta}d\Theta \quad (67)
\]

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

Using the SVD of \( \mathbf{M} \) rather than the EVD of \( \mathbf{M}^T\mathbf{M} \), 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 \( \Pi = \Lambda \), \( \mathbf{W}^T = \mathbf{V}^T \) (\( \mathbf{W} \) is the matrix of left eigenvectors), \( \mathbf{U}^DM\mathbf{V}\Theta\Theta^{-1} = \mathbf{D}_R \) and \( \mathbf{Z} = \mathbf{V} \) in equation (7) and dropping the subscript \( \mathbf{Q} \) for simplicity gives
\[
d\Pi = d(\mathbf{W}^T)\hat{U}^T\hat{M}\hat{V}\hat{\Theta}^{-1}\hat{Z} + \hat{W}^Td(U^TM\Theta^{-1})\hat{Z} + \hat{W}^T\hat{U}^T\hat{M}\hat{V}\hat{\Theta}^{-1}d\mathbf{Z}
\]
\[
= \hat{W}^Td(U^TM\Theta^{-1})\hat{Z} \quad (68)
\]

Working out the remaining error term on the right-hand side of (68) in a straightforward manner as
\[
d(U^TM\Theta^{-1}) = d(U^T)\hat{M}\hat{V}\hat{\Theta}^{-1} + \hat{U}^TdM\hat{V}\hat{\Theta}^{-1} + \hat{U}^T\hat{M}dV\hat{\Theta}^{-1} + \hat{U}^T\hat{M}\hat{V}d(\Theta^{-1}) \quad (69)
\]
and inserting the results obtained for \( \mathbf{A} = \mathbf{U}\hat{\Theta} \) and \( \mathbf{F} = \mathbf{V}^T \) (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 \( \text{COV}(\pi_n, \pi_m) \) 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 $\mathbf{M}$ on to the row and column space of $\mathbf{Q}$ must leave $\mathbf{M}$ unchanged, since $\mathbf{Q}$
spans the space of $\mathbf{M}$ in an unbiased way. This procedure is known in the literature of rank
annihilation as bilinear target testing:\textsuperscript{44}

$$
\mathbf{M} = \mathbf{U}^T \mathbf{M} \mathbf{V} \mathbf{V}^T = \mathbf{U} \mathbf{M} \mathbf{V} \mathbf{V}^T
$$

Here the substitution $\mathbf{M}_{\mathbf{UV}} = \mathbf{U}^T \mathbf{M} \mathbf{V}$ has been made. (Wilson \textit{et al.}\textsuperscript{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

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

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

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

Furthermore,

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

Combining (68) and (71)–(73) gives

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

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

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

Now it is possible to substitute $\tilde{\mathbf{W}}^T \tilde{\mathbf{U}}^T = (\tilde{\mathbf{C}}_\mathbf{M} + \tilde{\mathbf{C}}_\mathbf{N})^{-1} \tilde{\mathbf{X}}^+$ and $\tilde{\mathbf{V}} \tilde{\Theta}^{-1} \tilde{\mathbf{Z}} = (\tilde{\mathbf{Y}}^T)^+$ (see equation
(46)). This will eventually lead to expressions that contain the physical decomposition of $\mathbf{Q}$, 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 $\mathbf{Q}$. 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

$$
\tilde{\sigma}_{\pi_n} = \sum_{q=1}^{F} \sum_{i=1}^{S} \sum_{k=1}^{W} \sum_{p=1}^{E} \overline{W}_{qn} \tilde{U}_{iq} (\tilde{\sigma}_{M_{\pi_n}} - \tilde{\sigma}_{Q_{\pi_n}}) \tilde{V}_{kp} \tilde{C}_{\pi_n}^{-1} \tilde{Z}_{pn}
$$

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:

$$
\text{COV}(\pi_n, \pi_m) = | \tilde{\sigma}_{\pi_n} \cdot \tilde{\sigma}_{\pi_m} | = [ \tilde{\sigma}_M (1 - \tilde{\pi}_n - \tilde{\pi}_m) + \tilde{\sigma}_N \tilde{\pi}_n \tilde{\pi}_m ] \sum_{q=1}^{F} \overline{W}_{qn} \overline{W}_{qm} \sum_{p=1}^{E} \left( \frac{\tilde{Z}_{pn} \tilde{Z}_{pm}}{\lambda_{Q_{\pi_n}}} \right) (77)
$$

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

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

Here a final simplification resulted from normalizing the left eigenvector matrix $\tilde{W}$. Consequently, the right eigenvectors needed for the evaluation of (78) must be calculated as $\tilde{Z} = \tilde{W}^{-1}$ (‘inverse transpose’). This is not necessary for the reconstruction of $X$ in (46a), since the columns of $X$ are found by normalizing the matrix $UQZ$ anyway.

Using (78) derived for the ‘simulated’ standard additions, it is easy to derive the standard errors in the eigenvalues if the matrix $Q$ 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

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

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

REFERENCES