ELSEVIER

Analytica Chimica Acta 304 (1995) 273–283

ANALYTICA CHIMICA ACTA

Random error bias in principal component analysis. Part II. Application of theoretical predictions to multivariate problems

N.M. Faber [a,*,1], M.J. Meinders [b], P. Geladi [a], M. Sjöström [a], L.M.C. Buydens [c], G. Kateman [c]

[a] *Research Group for Chemometrics, Umeå University, S-901 87 Umeå, Sweden*
[b] *TNO Centre for Leather and Shoe Research, P.O. Box 135, 5140 AC Waalwijk, Netherlands*
[c] *Department of Analytical Chemistry, University of Nijmegen, Toernooiveld 1, 6525 ED Nijmegen, Netherlands*

Received 1 June 1994; accepted 10 November 1994

Abstract

In the first part of this paper expressions were derived for the prediction of random error bias in the eigenvalues of principal component analysis (PCA) and the singular values of singular value decomposition (SVD). The main issues of Part I were to investigate the question whether adequate prediction of this bias is possible and to discuss how the validation and evaluation of these predictions could proceed for a specific application in practice. The main issue of this second part is to investigate how random error bias should be taken into account. This question will be treated for a number of seemingly disparate multivariate problems. For example, the construction of confidence intervals for the bias-corrected quantities will be discussed with respect to the estimation of the number of significant principal components. The consequences of random error bias for calibration and prediction with ordinary least squares (OLS), principal component regression (PCR), partial least squares (PLS) and the generalized rank annihilation method (GRAM) will also be outlined. Finally, the derived bias expressions will be compared in detail with the corresponding results for OLS and GRAM.

Keywords: Principal component analysis; Singular value decomposition

1. Introduction

In the first part of this paper [1] expressions were derived for the prediction of random error bias in the eigenvalues of principal component analysis (PCA) and the singular values of singular value decomposition (SVD). It was found that depending on the signal-to-noise ratio for a specific principal component (PC) the random error bias in the eigenvalues is adequately predicted by (see Part I for notational practice), instead of

$$b_{\lambda_a} = (I + J - A)\sigma_{\mathrm{M}}^2$$

$$\text{for } a = 1,\dots,A \tag{1}$$

Paatero [2] has pointed out that an intuitive derivation of Eq. (1) is possible as follows: "If we apply Eq. (1) to form the total bias in A eigenvalues,

[*] Corresponding author.

[1] Present address: Center for Process Analytical Chemistry, Department of Chemistry, BG-10, University of Washington, Seattle, WA 98195, USA.

SSDI 0003-2670(94)00586-9

it is $(A \times I + A \times J - A^2)\sigma_M^2$. The expression in parentheses happens to be the number of essential parameters in A factors. If $\mathbf{M} = \mathbf{GF} + \mathbf{E}$, where $\mathbf{G}$ is $I \times A$ and $\mathbf{F}$ is $A \times J$, then there are $A \times I$ parameters in $\mathbf{G}$, and $A \times J$ parameters in $\mathbf{F}$. But it is well known that if $\mathbf{G}$ and $\mathbf{F}$ are given, then A^2 parameters in either $\mathbf{G}$ or $\mathbf{F}$ may be fixed at arbitrary values, by inserting a suitable $A \times A$ rotation $\mathbf{T}$: $\mathbf{GF} = \mathbf{GTT}^{-1}\mathbf{F}$. Thus A^2 parameters are redundant, which leaves the expression in parentheses. My intuitive picture is as follows: the norm of the noise in the matrix $\mathbf{M}$ is $I \times J \times \sigma_M^2$. In factorization, this noise is distributed partly to the residuals, and partly to the factors (= to the eigenvalues). The noise in the residuals is decreased by the number of parameters fitted (degrees of freedom). The expected value of the norm of the residuals is therefore $(I \times J - A \times I - A \times J + A \times A)\sigma_M^2$. That amount of noise that is missing from the residuals will appear as a bias in the eigenvalues.''

Paatero further notes [2]: ''This argument does not prove anything, but it makes Eq. (1) quite plausible and easy to remember.'' It is clear that his derivation does not include the complications that have been discussed in detail in Part I. Since these complications should guide the validation of the bias expressions, it is not entirely sufficient for the purpose of the current paper. However, the intuitive picture should be very appealing for most readers just because it avoids complications and reduces the derivation given in Part I to the essential part. It follows that the advantage of the ''intuitive picture'' that it is ''quite plausible and easy to remember'' should not be underestimated and that is the reason why it is included here.

The random error bias in the singular values is predicted by

$$b_{\theta_a} = 1/2(I + J - A - 1)\sigma_M^2/\theta_a$$

$$\text{for } a = 1,\dots,A \qquad (2)$$

The main objective of this second part is to investigate how random error bias should be taken into account. This will be illustrated by discussing the relevance of the derived expressions with respect to a number of important problems in multivariate data analysis. These problems comprise the construction of confidence intervals for the true quantities, the determination of the number of significant PCs and the calibration and prediction with ordinary least squares (OLS), principal component regression (PCR), partial least squares (PLS) and the generalized rank annihilation method (GRAM).

It will be shown that the construction of confidence intervals for the true quantities is essentially a different problem than determining the number of significant PCs. It should be mentioned that the last problem is more relevant in analytical chemistry and therefore will receive more attention here. It will be explained that for OLS, PCR and PLS the prediction is not necessarily influenced by the random error bias in the model parameters that are used for the prediction. The situation is, however, entirely different for GRAM, since for this method the model building (calibration) and prediction step coincide. Thus prediction of and subsequent correction for random error bias is mandatory for GRAM as already follows from the work of Booksh and Kowalski [3]. This difference between on one side OLS, PCR and PLS and on the other side GRAM may lead to a better understanding of the working of these methods. It is emphasized that no new (numerical) results will be presented here. Testing the adequacy of the bias expressions was one of the subjects of Part I.

Finally, before moving on to the applications of the derived bias expressions, a peculiar consequence of random error bias in the eigenvalues is pointed out. The random error bias is always positive. It therefore automatically gives a lower bound for the primary eigenvalues or, to be more specific, for the smallest primary eigenvalue. Other lower bounds were found during the present investigation and it may be worth while to compare them with respect to their efficiency. Since deriving lower bounds for the primary eigenvalues is not considered to be a problem in multivariate data analysis, they are not presented together with the other applications of the derived expressions in the theoretical section but in the Appendix.

2. Theory

The application of the derived bias expressions comes down to answering two questions. First, one has to investigate whether bias is harmful for the particular application at hand. Second, since the bias is a constant background, one has to investigate the consequences of simply removing the bias. Before discussing the possible harm of random error bias for a number of multivariate problems the correction for bias is treated.

2.1. Bias correction

The consequences of a bias correction are as follows. Without the bias correction the estimated quantity, say $\hat{Z}$, may be expressed as (see Part I)

$$\hat{Z} = Z + \epsilon_Z + b_Z$$

and the mean square error is given by

$$mse_Z = \sigma_Z^2 + b_Z^2$$

The bias-corrected estimated quantity, $\hat{Z}_c$, may be expressed as

$$\hat{Z}_c = \hat{Z} - \hat{b}_Z = Z + \epsilon_Z + b_b$$

where $\hat{b}_Z$ denotes the estimated bias in $\hat{Z}$ and b_b is a residual bias. (For the eigenvalues of PCA the bias is systematically underestimated and the residual bias is positive.) The mean square error for the bias-corrected quantity is consequently given by

$$mse_{Z_c} = \sigma_{Z_c}^2 + b_b^2$$

The bias correction is successful if the residual bias is much smaller than the standard error in the estimated quantity. In that case an (almost) unbiased estimate of the true value is obtained and only the standard error will effectively contribute to the total error in the estimate.

For example, from the results discussed in Part I for the simulated three-component system it can be inferred that the bias correction will give excellent results for the first two eigenvalues but is not successful for the smallest one, since for this eigenvalue the signal-to-noise ratio is too unfavourable. Whether a bias correction will be successful is conveniently investigated by Monte Carlo simulations. Bias correction has already been discussed for GRAM in another paper [9].

2.2. Construction of confidence intervals for bias-corrected quantities

An important application of the derived bias expressions is already noted by Goodman and Haberman [4]. If confidence intervals are to be constructed for the true values, then correction for (non-negligible) bias is mandatory. It should be noted that in the simple illustrative example in Part I bias was a result of a skewed distribution of the calculated quantity. This leads to a troublesome situation if confidence intervals have to be derived. Fortunately, this is not (necessarily) the case here, since by the central limit theorem, linear combinations of independent numbers tend to approach normality regardless of their initial distribution. As a general guide, a number of at least 50 can be considered to be large enough in practice [5]. This principle can be applied to the sum of imbedded errors in Eq. (9) of Part I. Thus if the number of matrix elements is larger than 50 there may be a bias in the eigenvalues and singular values but one can still set up confidence intervals in the usual way after correcting for bias.

2.3. Pseudorank estimation

The construction of confidence intervals is related to the problem of pseudorank estimation but not identical. Since pseudorank estimation is one of the central problems of multivariate data analysis in analytical chemistry, this matter will be given ample consideration here.

In another paper [6] it was shown that a singular value can be tested for significance by comparing it to the first singular value obtained from a 'reference' matrix in a t-test. The reference matrix is a *random matrix* and its size, which also gives the number of degrees of freedom ν to be used in the test, is derived as follows. If we test the ath singular value of an $I \times J$ data matrix (working backwards through the list of singular values), then under the null-hypothesis, i.e. it only represents noise, this singular value should be equal to the first singular value of an $(I - a + 1) \times (J - a + 1)$ random matrix. (This becomes clear if one inserts, for example, $a = 1$.) If the singular value under test, $\hat{\theta}_a$, is equal to the reference singular value, $\hat{\theta}_{a,\mathrm{ref}}$, then it only represents noise.

The procedure depends on a reliable estimate of $\sigma_\theta = \sigma_M$ (see Eq. (4) of Part I). Taking also the standard error in $\hat{\theta}_{a,\mathrm{ref}}$ into account leads to the following t-value

$$t_{\theta_a} = \frac{\hat{\theta}_a - \hat{\theta}_{a,\mathrm{ref}}}{\sqrt{2}\,\hat{\sigma}_M}$$

$$\text{for } a = J,...,1 \qquad (3)$$

which should be compared to the tabulated $t_\nu(1-\alpha)$ in order to test at the α level of significance. Bias is not corrected for in this procedure, since it is 'implicitly taken into account' by subtracting the reference value. (It is easily seen that this principle should hold for the testing of any function of the eigenvalues.) The procedure is illustrated in Fig. 1 for the 20 × 10 matrix with constant elements M and normally distributed noise with standard deviation $\sigma_M = 1.0$ (see Part I).

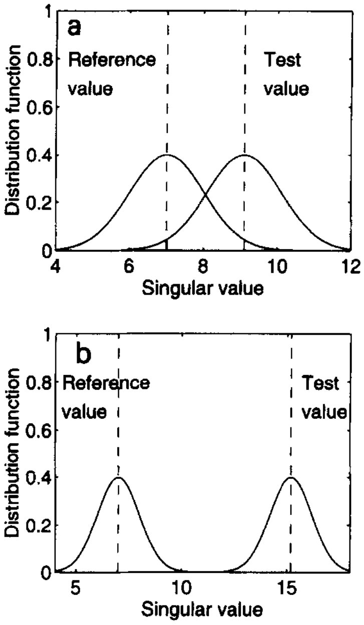

Fig. 1. Distribution functions for the singular values of the test data matrix and the reference matrix. The test data matrix (20 × 10) has constant elements with size (a) $M = 0.5$ and (b) $M = 1.0$ and normally distributed noise with variance 1 added. The reference matrix is a 20 × 10 matrix with normally distributed elements with variance 1. It is seen that the distribution functions of the singular values overlap to a certain extent. The confidence level α is supplied by the t-test. $\alpha = 15\%$ for the data set with $M = 0.5$ (not significant) and $\alpha < 0.1\%$ for the data set with $M = 1$ (highly significant).

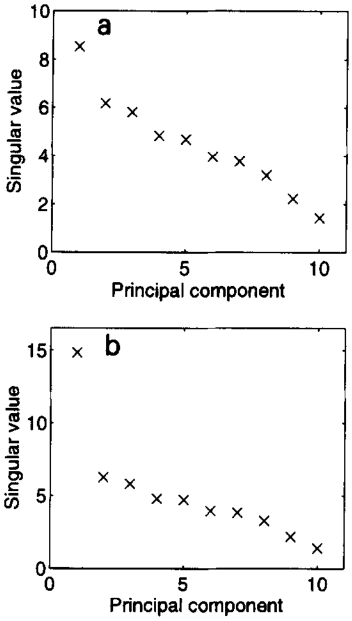

Fig. 2. Singular values of the test data matrix with constant elements of size (a) $M = 0.5$ and (b) $M = 1.0$.

The obtained significance levels were compared to those obtained by Malinowski's F-test [7]. The significance levels produced by the t-test were consistently sharper. Thus if a good estimate of σ_M is available, then the t-test is to be preferred over the F-test. However, it is well known that Malinowski's F-test works without prior knowledge of σ_M and is therefore more generally applicable.

If an estimate for σ_M is not available a graphical alternative for the t-test is obtained by plotting the singular values [6]. The singular values for the secondary PCs should lie approximately on a straight line [6]. In Fig. 2 the singular values are shown for the 20 × 10 matrix mentioned above. For $M = 0.5$ the first singular value slightly deviates from the straight line part and its significance remains doubtful. For $M = 1.0$ the significance of the first singular value is established without difficulty.

In summary, a bias correction should be applied if confidence intervals are to be constructed for the eigenvalues or singular values. However, it should not be applied if the eigenvalues of PCA or singular values of SVD are to be used for pseudorank estimation.

2.4. Random error bias in ordinary least squares (vector data)

The first method to be discussed as a PC model is ordinary least squares (OLS). The standard regression model is defined by [2]

$$\tilde{\mathbf{y}} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}_{\mathrm{y}} \tag{4}$$

where $\tilde{\mathbf{y}}$ is an $I \times 1$ vector of observations on the dependent variable, $\mathbf{X}$ is an $I \times K$ matrix of observations on the independent variables, $\boldsymbol{\beta}$ is a $K \times 1$ vector of model parameters and $\boldsymbol{\epsilon}_{\mathrm{y}}$ is an $I \times 1$ vector of errors. Then the ordinary least squares (OLS) estimate of $\boldsymbol{\beta}$ is given by

$$\hat{\boldsymbol{\beta}}_{\mathrm{OLS}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\tilde{\mathbf{y}} \tag{5}$$

These estimates for the parameters are commonly used for the prediction of properties of a new sample. Under the usual assumption that $\mathbf{X}$ is known without error, i.e. it is 'fixed', the OLS estimate for the parameter vector $\boldsymbol{\beta}$ is unbiased, since it is a linear function of $\tilde{\mathbf{y}}$ (the dependence on $\mathbf{X}$ need not be considered).

This assumption is, however, not valid in many practical situations in analytical chemistry where both dependent as independent variables are measured with a certain (non-negligible) precision. In, for example, multivariate calibration the components of $\tilde{\mathbf{y}}$ often represent the spectral responses for an unknown sample and the elements of $\tilde{\mathbf{X}}$ are the spectral responses for the pure chemical components (classical model). The immediate consequence of $\tilde{\mathbf{X}}$ not being fixed is that the estimated parameter vector is biased, since it is a non-linear function of $\tilde{\mathbf{X}}$. This bias can be related to the previously derived bias in PCA by decomposing $\tilde{\mathbf{X}}$ according to the SVD, i.e. $\tilde{\mathbf{X}} = \hat{\mathbf{U}}\hat{\boldsymbol{\Theta}}\hat{\mathbf{V}}^{\mathrm{T}}$, and rewriting Eq. (5) as

$$\hat{\boldsymbol{\beta}}_{\mathrm{OLS}} = (\hat{\mathbf{V}}\hat{\Lambda}^{-1}\hat{\mathbf{V}}^{\mathrm{T}})\tilde{\mathbf{X}}^{\mathrm{T}}\tilde{\mathbf{y}} \tag{6}$$

Hodges and Moore [8] have shown that for uncorrelated homoscedastic errors in $\tilde{\mathbf{X}}$ the expected value for the estimated parameter vector is given by

$$E\left[\hat{\boldsymbol{\beta}}_{\mathrm{OLS}}\right] = \left(\mathbf{I} + I\boldsymbol{\Phi}\sigma_{\mathrm{X}}^{2}\right)^{-1}\boldsymbol{\beta} \tag{7}$$

where $\mathbf{I}$ symbolizes the identity matrix, σ_{X}^{2} denotes the error variance in $\tilde{\mathbf{X}}$ and the factor $\boldsymbol{\Phi}$ is given by

$$\boldsymbol{\Phi} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1} = \mathbf{V}\boldsymbol{\Lambda}^{-1}\mathbf{V}^{\mathrm{T}} \tag{8}$$

If the errors in $\tilde{\mathbf{X}}$ are small Eq. (7) can be worked out by taking only the diagonal elements of $\boldsymbol{\Phi}$ into account [9]. This approximation leads to the following expression for the bias in the individual estimated parameters $\hat{\beta}_k$

$$b_{\beta_k} = E\left[\hat{\beta}_k\right] - \beta_k = -\sum_{p=1}^{K} \frac{V_{kp}^{2}}{\lambda_p}\beta_k(I - K - 1)\sigma_{\mathrm{X}}^{2}$$

$$\text{for } k = 1,\ldots,K \tag{9}$$

Eq. (9) will be discussed in detail in a later section. For the moment we only point out that the first factor on the far right-hand side of Eq. (9), i.e. $\Phi_{kk} = \Sigma V_{kp}^{2}/\lambda_p$, is an error propagation factor which is also present in the expression for the standard error in $\hat{\beta}_k$ [9]. This factor tends to be dominated by the small eigenvalues of $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ in the denominator although small elements for the eigenvectors in the numerator may cancel out the effect of small eigenvalues. Consequently, the effect of small eigenvalues is not completely certain [10]. However, if only large eigenvalues are present, then the error propagation factor should always be small, since the eigenvectors are normalized (they originate from the SVD of $\mathbf{X}$). The practical evaluation of Eqs. (7)–(9) proceeds along the line discussed in Part I.

Thus it is easily demonstrated that for many applications the estimated model parameters are biased but is this bias automatically harmful if the model is applied for the prediction of properties of a new sample? With respect to the consequences of applying biased model parameters for prediction (forecast) Hodges and Moore [8] state: "The foregoing account gives some idea of the possible effects of data errors on the estimation of regression coefficients by least squares. Regression equations are commonly used for making forecasts so it is relevant to examine how such forecasts are affected. The first point to make is well known, namely that if the

[2] The notation introduced in Part I is used here. In most presentations of OLS Eq. (4) is written, for example, as $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \epsilon_{\mathrm{y}}$. It will become clear later that it is convenient to make a distinction between quantities that are known without error and quantities that are measured with a non-negligible error.

independent variables are drawn from stationary distributions then, so long as the values used in making the forecast are also generated by the same distributions and subject to the same sorts of error, no forecasting bias is involved. Unfortunately this is often not the case, and it is never true when an independent variable follows a trend over time.''

The fact that biased model parameters will not automatically deteriorate the prediction for a new sample may come as a surprise. However, a biased model is not necessarily an incorrect model and, consequently, a bias correction is not always useful. Consider, for example, the case where a sample that has been used to build the (biased) model is remeasured as a 'future' sample. Then, obviously, the biased model parameters should be used in order to predict its properties (e.g. concentrations), since it is the correct model for this sample. It is also clear from this example that a bias correction could actually ruin the prediction. However, it is important to note that the additional assumption (stationary distribution) may be severe in many applications of regression models. Regression models are often used for the analysis of food samples, e.g. milk or meat. Then the stationary assumption is not likely to be fulfilled over a long period and, consequently, prediction will suffer from random error bias. This will be further discussed in the next section.

In summary, a bias correction should be applied if focus is on the model parameters themselves. For example, if confidence intervals are to be constructed for the parameters or if the parameters are to be used for interpretation. However, it should not be applied if the model is to be used for prediction as is most often the case.

2.5. *Random error bias in principal component regression (vector data)*

Principal component regression (PCR) is a method that has been introduced to provide an alternative to OLS in the case that severe collinearity among the independent variables (corresponding to very small eigenvalues of $\mathbf{X}^{\mathrm{T}}\mathbf{X}$) leads to excessive standard errors in the estimated parameters. We will confine the discussion to prediction. In PCR the original matrix of independent variables $\tilde{\mathbf{X}}$ is replaced by a selection of PCs in order to stabilize the inversion step in Eq. (5), i.e.

$$\hat{\boldsymbol{\beta}}_{\mathrm{PCR}} = \left(\hat{\mathbf{V}}_A \hat{\Lambda}_A^{-1} \hat{\mathbf{V}}_A^{\mathrm{T}}\right)\tilde{\mathbf{X}}^{\mathrm{T}}\tilde{\mathbf{y}} \tag{10}$$

where the subscripted A indicates that a subset of A PCs is selected from the total of K PCs. This leads to a more economic model where the standard errors are reduced at the expense of introducing an *underfactoring bias*. Depending on the data there is a favourable trade-off and, hopefully, a better predictive model is obtained this way. This trade-off principle is nicely discussed by Mandel [11].

For OLS the standard error is affected by the same error propagation factor as given in Eq. (9) for the random error bias [9]. In PCR the summation index p in Eq. (9) runs over a subset of the K PCs. Consequently, the reduction of the standard error by deleting highly contributing eigenvalues automatically translates into a reduction of the *random error bias* and vice versa. This is an additional advantage of using PCR that may be of great practical importance in the light of the troublesome stationary assumption. It should be mentioned that different strategies for the selection of PCs have been proposed in the literature. The preceding gave an account of the 'top-down' procedure, i.e. the eigenvalues are selected in order of size. Sutter et al. [10] advocate to select PCs that correlate best with the dependent variable $\tilde{\mathbf{y}}$. This will not necessarily lead to the deletion of the PCs associated to the smallest eigenvalues. In this way a model is built with optimal prediction properties rather than a small variance of the parameters.

It is important to note that the predictive power of a model is usually tested by an internal validation procedure, i.e. the available data is split up in a calibration and a test set. The calibration set is used to built a model that predicts the properties of the test set. Such a procedure will not enable the verification of the validity of the stationary assumption for future samples. With respect to the random error bias it does not matter which strategy is used as long as one is able to verify the validity of the stationary assumption. Since the bias contribution is expected to be relatively large for the smallest eigenvalues, it seems reasonable to remove their influence from the model. This reasoning would favour a top-down selection of the PCs. It is emphasized that removing bias by deleting PCs is different from removing bias by subtracting the theoretical value, which is never recommended if the model is to be used for prediction.

In summary, with respect to the correction for bias the same conclusions hold as for OLS. With respect to the deletion of PCs it is important to know whether the stationary assumption is likely to hold, i.e. the random error bias is constant. If it may be assumed to hold, then a 'best subset' procedure as recommended by Sutter et al. [10] should be followed. Otherwise, one should consider the top-down procedure as safer, i.e. delete PCs that give an excessive contribution to the overall random error bias in the parameters.

2.6. Random error bias in partial least squares regression (vector data)

The same principles discussed until now for PCR should also hold for the intimately related PLS although there are some theoretical complications, since the error analysis for this method is still in its infancy. This can be explained as follows. In PLS the inversion step in Eq. (5) is performed in a subspace that also contains predictive information about the dependent variable $\tilde{\mathbf{y}}$. The PLS factors are linear combinations of the PCs of $\tilde{\mathbf{X}}$ that successively have a maximum correlation with $\tilde{\mathbf{y}}$. This automatically leads to good predictive models by the top-down selection procedure.

Since the PLS factors contain information about the dependent variable $\tilde{\mathbf{y}}$, the estimated parameters are no longer linear functions of $\tilde{\mathbf{y}}$ and only approximate standard errors have been derived [12]. (Furthermore, these standard errors are derived for the case that $\mathbf{X}$ is fixed.) For the underfactoring bias no theoretical results have been reported yet and, finally, the random error bias is still undiscussed in the literature. However, in the light of the great difficulties that have to be solved in order to obtain (only approximate) standard errors, it is expected to be a formidable task to derive the necessary bias expressions for PLS.

It is worth mentioning that in two large simulation studies of PLS [13,14] both $\tilde{\mathbf{y}}$ and $\tilde{\mathbf{X}}$ were considered to be stochastic, whereas Frank and Friedman [15] performed simulations where $\mathbf{X}$ was considered to be fixed. This is an illustration of the fact that simulation studies (in general evaluation studies) are always carried out with certain applications in mind. It is evident that it completely depends on the kind of data that is generated whether the random error bias will be present. Moreover, even when it shows up, it may still be negligible compared to the standard error or the underfactoring bias. However, the important point is that one can only neglect this source of error after one has established that it does not influence the final result. For a number of methods the necessary information is conveniently provided by the theoretical bias predictions, as discussed in this paper. For PLS there is still a large gap in the relevant error theory to be filled.

In summary, with respect to the correction for bias the same conclusions hold as for OLS and PCR but there is no theory available to bring this insight to use. Since in PLS the model is constructed with reference to the dependent variable $\tilde{\mathbf{y}}$, the usual (automatic) top-down procedure is even more in favour than for PCR.

2.7. Random error bias in the generalized rank annihilation method (matrix data)

The last method to be discussed as a PC model is the generalized rank annihilation method (GRAM). GRAM is a method for curve resolution and calibration using two data matrices simultaneously, one for the unknown and one for the calibration sample. In order to apply this technique, 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. Without loss of generality it will be assumed that the data are obtained by the spectral detection of a chromatographic separation process.

Let the $I \times J$ unknown data matrix $\tilde{\mathbf{M}}$ be given as

$$\tilde{\mathbf{M}} = \mathbf{H}\mathbf{Y}^{\mathrm{T}} + \mathbf{E}_{\mathrm{M}} \qquad (11)$$

where $\mathbf{H}(I \times K)$ contains the errorless elution profiles of the K (detectable) components, $\mathbf{Y}(J \times K)$ contains the corresponding errorless spectra and $\mathbf{E}_{\mathrm{M}}$ is the $I \times J$ matrix of measurement errors. The spectra in $\mathbf{Y}$ are normalized so that the concentration dependence is absorbed in **H**. It will be assumed for convenience that the unknown sample contains all components present in the calibration sample and discuss the general case later. Then the $I \times J$ calibration data matrix $\tilde{\mathbf{N}}$ can be written as [9]

$$\tilde{\mathbf{N}} = \mathbf{H}\tilde{\boldsymbol{\Pi}}\mathbf{Y}^{\mathrm{T}} + \mathbf{E}_{\mathrm{N}} \tag{12}$$

where $\tilde{\boldsymbol{\Pi}}$ is a $K \times K$ diagonal matrix that contains the ratios of the concentrations in the samples, i.e. $\tilde{\pi}_k = \tilde{c}_{\mathrm{N},k}/\tilde{c}_{\mathrm{M},k}$ and $\mathbf{E}_{\mathrm{N}}$ is the $I \times J$ matrix of measurement errors. [3]

Sánchez and Kowalski [16] have shown that Eqs. (11) and (12) can be combined by decomposing the unknown data matrix according to the SVD, i.e. $\tilde{\mathbf{M}} = \hat{\mathbf{U}}\hat{\boldsymbol{\Theta}}\hat{\mathbf{V}}^{\mathrm{T}}$. Retaining only the first A principal components of $\tilde{\mathbf{M}}$ leads to the following standard eigenvalue problem

$$\hat{\boldsymbol{\Pi}} = \hat{\mathbf{T}}^{-1}\left(\hat{\mathbf{U}}_A^{\mathrm{T}}\tilde{\mathbf{N}}\hat{\mathbf{V}}_A\hat{\boldsymbol{\Theta}}_A^{-1}\right)\hat{\mathbf{T}} \tag{13}$$

where the eigenvalue matrix $\hat{\boldsymbol{\Pi}}$ contains the estimated concentration ratios and the eigenvector matrix $\hat{\mathbf{T}}$ can be used to reconstruct **H** and **Y** from the singular vectors of $\tilde{\mathbf{M}}$ as $\hat{\mathbf{H}} = \hat{\mathbf{U}}_A\hat{\mathbf{T}}$ and $\hat{\mathbf{Y}}^{\mathrm{T}} = \hat{\mathbf{T}}^{-1}\hat{\boldsymbol{\Theta}}_A\hat{\mathbf{V}}_A^{\mathrm{T}}$. The subscripted A indicates that the SVD is truncated to the A leading PCs (A is actually an estimate of the number of components K). [4]

Booksh and Kowalski [3] have demonstrated that the estimated concentration ratios, $\hat{\Pi}$, are biased. It is important to note that this bias does not result from a skewed distribution of the eigenvalues. Even for a skewed distribution of the measurement noise they found that GRAM yields (approximately) normally distributed eigenvalues. In another paper [9] the following expression is derived for the expected value of the estimated concentration ratios assuming that the measurement noise is uncorrelated and homoscedastic:

$$E[\hat{\boldsymbol{\Pi}}] = \left(\mathbf{I} + I\boldsymbol{\Psi}\sigma_{\mathrm{M}}^2\right)^{-1}E[\tilde{\boldsymbol{\Pi}}]\left(\mathbf{I} + J\boldsymbol{\Psi}\sigma_{\mathrm{M}}^2\right)^{-1} \tag{14}$$

where σ_{M}^2 denotes the error variance in $\tilde{\mathbf{M}}$ and the factor $\boldsymbol{\Psi}$ is given by

$$\boldsymbol{\Psi} = (\mathbf{H}^{\mathrm{T}}\mathbf{H})^{-1}(\mathbf{Y}^{\mathrm{T}}\mathbf{Y})^{-1} = \mathbf{T}^{-1}\boldsymbol{\Lambda}^{-1}\mathbf{T} \tag{15}$$

Under the assumption of small errors in $\tilde{\mathbf{M}}$ (cf. Eq. (9)) the bias in the individual estimated parameters can be approximated by [5]

$$b_{\pi_a} = E[\hat{\pi}_a] - E[\tilde{\pi}_a]$$
$$= -\sum_{p=1}^{A}\frac{T_{ap}^{-1}T_{pa}}{\lambda_p}E[\tilde{\pi}_a](I + J - 2A - 2)\sigma_{\mathrm{M}}^2$$
$$\text{for } a = 1,\ldots,A \tag{16}$$

where the first factor on the far right-hand side is Ψ_{aa}. It was found that the bias estimate is still accurate if A under or overestimates K. [6]

It is emphasized that predicting the (actual) concentration ratios $\tilde{\boldsymbol{\Pi}}$ in GRAM is equivalent to estimating the (true) model parameters β in OLS, PCR and PLS [8]. Since one is directly interested in the concentration ratios (there is no separate prediction step), bias should always be corrected for.

It is worth mentioning that also in GRAM one has made a restrictive assumption with respect to the model represented by Eqs. (11) and (12). Analogous to the stationary assumption discussed before, now the assumption has been made that the pure component profiles in **H** and $\mathbf{Y}^{\mathrm{T}}$ be identical for both

[3] An important difference between Eq. (12) and Eq. (4) is the presence of quantities that already carry an error. The concentrations actually present in the samples will deviate from the true values as a result of errors made, for example, during sample acquisition, sample preparation, sample injection in chromatography, etc.

[4] In GRAM the A PCs are automatically selected top-down. This makes Eq. (13) essentially different from Eq. (10).

[5] The bias resulting from measurement errors is defined with respect to the actual concentration ratio, since this would be the value found if there were no measurement errors. It is emphasized that Eq. (16) is slightly different from the expression given in [9], i.e. Eq. (9). There the right-hand side of the bias expression contains the expected value of the *estimated* concentration ratio instead of the expected value of the *actual* concentration ratio. This seems to be more appropriate if the bias expression is to be evaluated, since only a realization of the estimated value is available. However, for the purpose of this paper (discussion and comparison of bias expressions) the current formulation seems to be suitable.

[6] This is only true if a correct estimate for σ_{M} is inserted in (16). Two examples were found in the literature where error estimates are calculated from the residuals of an incorrectly dimensioned PC model [17,18]. This situation is easily recognized, since contrary to the underfactoring bias, the standard error and random error bias should always increase with model complexity.

samples. In practice, however, it should be possible to meet this requirement by measuring both samples within a short time span. This can certainly be interpreted as being an advantage of having only one calibration sample, although from a general statistical point of view a one-point calibration is very unsatisfactory.

In summary, contrary to OLS, PCR and PLS, one always has to correct for bias in the estimated eigenvalues of GRAM. An overview of the consequences of random error bias for the PC models discussed in this paper is given in Table 1.

2.8. Similarities and dissimilarities between bias expressions (1), (2), (9) and (16)

It has been shown that OLS and GRAM are directly related to PCA. One therefore expects that a useful comparison of the corresponding bias expressions is possible. It should, however, be kept in mind that Eqs. (9) and (16) were obtained by the additional assumption of small errors. The comparison will be made with respect to the elements that are present in Eqs. (9) and (16) but may be missing in Eqs. (1) and (2), i.e. (i) the overall sign of the bias, (ii) the error propagation factor, (iii) the parameters that are estimated, (iv) the number of observations made, (v) the dimension of the PC model and (vi) the size of the measurement noise.

(i) Overall sign of the bias

The overall minus sign in Eqs. (9) and (16) is easily explained, since the matrix of eigenvalues or singular values is inverted. It should be mentioned that a minus sign is not always to be expected for the bias in the eigenvalues of GRAM. In the general case where both the unknown and calibration sample have unique components, the procedure is modified by substituting the sum matrix $\tilde{\mathbf{M}} + \tilde{\mathbf{N}}$ for $\tilde{\mathbf{M}}$ and $\tilde{\mathbf{M}}$ for $\tilde{\mathbf{N}}$. Now one has the situation that the errors in $\tilde{\mathbf{M}} + \tilde{\mathbf{N}}$ and $\tilde{\mathbf{M}}$ are correlated, thereby leading to a bias expression with terms of opposite sign [9]. The overall sign of the bias will then actually depend on the specific data at hand.

(ii) Error propagation factor

There is a notable difference between Eqs. (1), (2), (9) and (16). Eqs. (9) and (16) are characterized by the presence of a factor that quantifies the amount of error propagation. This factor, which is missing in Eqs. (1) and (2), directly depends on the amount of overlap encountered in the matrices **X**, **H** and **Y**. Thus, conversely, one might conclude that there is 'no error propagation' in PCA. Error propagation in PCA is, however, directly indicated by a large difference in size of the singular values. PCs with small singular values have an unfavourable signal-to-noise ratio.

(iii) Estimated parameters

The dependence on the estimated parameter is linear for Eqs. (2), (9) and (16). This is a marked difference with respect to Eq. (1) that states that the predicted bias is independent of the size of the eigenvalue. (It has been seen in Part I that the real bias is not constant.)

Table 1
Summary of the consequences of random error bias for the multivariate problems discussed in this paper [a]

Quantity	Application	Bias correction	Equation
Eigenvalue PCA [b]	Confidence interval	Yes	(1)
	Pseudorank estimation	No	
Singular value SVD [b]	Confidence interval	Yes	(2)
	Pseudorank estimation	No	
Parameter OLS [c]	Confidence interval	Yes	(7,9)
	Interpretation	Yes	(7,9)
	Prediction	No	
Eigenvalue GRAM	Prediction	Yes	(14,16)

[a] A direct consequence of random error bias in PCA is the fact that the expressions derived for the prediction of bias in related problems have to be corrected for this bias.

[b] This holds in general for any function of the eigenvalues.

[c] The consequences are identical for PCR and PLS. The analogy for Eqs. (7) and (9) has, however, not yet been derived for PLS.

It is important to note that no other eigenvalues contribute to the bias in a specific eigenvalue of GRAM. It is well known that under certain circumstances some of the eigenvalues of GRAM may be complex [19]. Surely, it would be undesirable that a complex eigenvalue could 'spoil' a real eigenvalue by interacting through the bias expression. The same reasoning holds for the variance in the eigenvalues [20,9].

(iv) Number of observations

The number of observations is present in all bias expressions: both I and J in Eqs. (1), (2) and (16) and I in Eq. (9).

(v) Dimension of PC model

The explicit dependence on the dimension of the PC model of $\tilde{\mathbf{M}}$ in Eqs. (1), (2) and (16), i.e. A, is equivalent to the dependence on the number of independent variables in Eq. (9), i.e. K. This is to be expected, since for OLS the PC model is K-dimensional.

(vi) Size of the measurement noise

In all bias expressions one finds an identical dependence on the variance of the measurement noise, i.e. σ_{M}^2.

3. Conclusions

The main issue of this second part was to investigate how random error bias should be taken into account for a number of multivariate problems. It was found that one must be very careful in automatically applying a straightforward bias correction (see Table 1). From the present theoretical comparison the following conclusions are drawn.

The difference between the construction of confidence intervals for the bias-corrected quantities and the problem of pseudorank estimation has been explained. In order to obtain confidence intervals for the true values the estimated values must be corrected for bias [4]. The pseudorank can be estimated by comparing the singular values of the test data matrix with the first singular value of an appropriately sized random matrix in a t-test. No bias correction takes place in this procedure.

The consequences of random error bias for calibration and prediction with OLS, PCR and PLS have been discussed. If the focus is on the parameters obtained from the calibration phase, then one should correct for bias. If, however, the biased model is to be used for prediction, then a bias correction is not allowed.

It has been detailed that the prediction is not affected by the random error bias as long as the distribution of the independent variables is stationary. This assumption may have consequences for the selection of PCs in PCR. For PCR it is concluded that if the distribution for the independent variables can not be considered to be stationary, then one should consider to delete PCs for which the bias makes an important contribution. Otherwise, the 'best-subset' procedure recommended by Sutter et al. [10] should be in favour. For PLS the same principle holds but the necessary theory has not yet been developed. This means that it should be important to extend the theory that is already available for PCR to PLS.

The situation is more transparent for GRAM, since for this method the model building (calibration) and prediction stage are the same. Here it is always necessary to correct for bias.

Finally, the previously derived expressions, i.e. Eqs. (1) and (2), have been compared with the corresponding results for OLS (with errors in the independent variables), i.e. Eq. (9), and GRAM, i.e. Eq. (16). Some striking similarities (as well as differences) have been pointed out for these multivariate methods.

Acknowledgements

NMF acknowledges the financial assistance of a COMETT Ba grant during his stay at Umeå

Appendix 1

Three lower bounds for the smallest primary eigenvalue $\hat{\lambda}_A$

The first lower bound follows from the reformulation of the real error function, i.e. Eq. (6) in Part I. The smallest primary eigenvalue should be larger than the average secondary eigenvalue. Since the summation in Eq. (6) runs over *J*-*A* secondary eigenvalues, the first lower bound immediately follows as [7]

$$(I-A)\sigma_{\mathrm{M}}^2 < \hat{\lambda}_A \tag{I}$$

The second lower bound is given by the bias in the eigenvalue itself:

$$(I+J-A)\sigma_{\mathrm{M}}^2 < \hat{\lambda}_A \tag{II}$$

It is seen that the bias is always larger than the average secondary eigenvalue. For example, if the data matrix is square and highly overdetermined, i.e. $I=J \gg A$, then the bias is approximately twice as large as the average secondary eigenvalue. The third lower bound is obtained by observing that the discriminant in Eq. (16) of Part I should be positive:

$$2(I+J-A-1)\sigma_{\mathrm{M}}^2 < \hat{\lambda}_A \tag{III}$$

It follows that the third lowerbound is approximately twice as large as the second lowerbound. It is by far the most efficient of the three. An immediate consequence of this expression is that one can not predict a bias that is larger than approximately 50% of the smallest primary eigenvalue. This should be a useful result that is easy to remember. (Note, however, that the predicted bias always underestimates the real bias (see Part I).)

A numerical example may show how efficient this lower bound may be in practice. For the simulated three-component system in Part I the Monte Carlo average for the third eigenvalue was 39. The third lowerbound is $2 \times (36+36-3-1) \times (0.5)^2 = 34$. The lowerbound is found to be rather good in this case. It is, however, emphasized that this data matrix was simulated in order to have one eigenvalue with an exceptionally large bias. Evidently, this lower bound is not efficient for the largest two eigenvalues (3803 and 143, respectively).

[7] The discussion is simplified by assuming that σ_{M} is adequately estimated from the residuals of PCA.

References

[1] N.M. Faber, M.J. Meinders, P. Geladi, M. Sjöström, L.M.C. Buydens and G. Kateman, Anal. Chim. Acta, 304 (1995) in press.

[2] P. Paatero, personal communication.

[3] K. Booksh and B.R. Kowalski, J. Chemom., 8 (1994) 45.

[4] L.A. Goodman and S.J. Haberman, JASA, 85 (1990) 139.

[5] J.R. Green and D. Margerison, Statistical Treatment of Experimental Data, Elsevier, Amsterdam, 1978.

[6] N.M. Faber, L.M.C. Buydens and G. Kateman, Anal. Chim. Acta, 296 (1994) 1.

[7] E.R. Malinowski, J. Chemom., 3 (1988) 49.

[8] S.D. Hodges and P.G. Moore, Appl. Stat., 21 (1972) 185.

[9] N.M. Faber, L.M.C. Buydens and G. Kateman, J. Chemom., 8 (1994) 181.

[10] J.M. Sutter, J.H. Kalivas and P.M. Lang, J. Chemom., 6 (1992) 217.

[11] J. Mandel, Am. Stat., 36 (1982) 15.

[12] A. Phatak, P.M. Reilly and A. Penlidis, Anal. Chim. Acta, 277 (1993) 495.

[13] I.E. Frank, Tech. Rep. No. 105, Department of Statistics, Stanford University, 1989.

[14] E.V. Thomas and D.M. Haaland, Anal. Chem.. 62 (1990) 1091.

[15] I.E. Frank and J.H. Friedman, Technometrics, 35 (1993) 109.

[16] E. Sánchez and B.R. Kowalski, Anal. Chem., 58 (1986) 486.

[17] C.J. Appellof and E.R. Davidson, Anal. Chem., 53 (1981) 2053.

[18] E.R. Malinowski, Factor Analysis in Chemistry, Wiley, New York, 1991.

[19] S. Li, J.C. Hamilton and P.J. Gemperline, Anal. Chem., 64 (1992) 599.

[20] N.M. Faber, L.M.C. Buydens and G. Kateman, J. Chemom., 7 (1993) 495.