

PCA-based algorithm for calibration of spectrophotometric analysers of food

Roman Z. Morawski¹ and Andrzej Miękina

Warsaw University of Technology, Faculty of Electronics and Information
Technology, Institute of Radioelectronics, Warsaw, Poland

E-mail: r.morawski@ire.pw.edu.pl

Abstract. Spectrophotometric analysers of food, being instruments for determination of the composition of food products and ingredients, are today of growing importance for food industry, as well as for food distributors and consumers. Their metrological performance significantly depends of the numerical performance of available means for spectrophotometric data processing; in particular – the means for calibration of analysers. In this paper, a new algorithm for this purpose is proposed, *viz.* the algorithm using principal components analysis (PCA). It is almost as efficient as PLS-based algorithms of calibration, but much simpler.

1. Introduction

The term *spectrophotometric analyser of food* is applied to various spectrophotometric sensors, devices, instruments, probes and testers dedicated to measurements of physical and (bio)chemical parameters characterising food products (including beverages). The author's review paper published by the end of 2012 [1] contains a comprehensive overview of basic concepts related to food analysers and of the methodology of measurement data processing in such analysers, as well as a justification of their practical importance – all these illustrated with 250 references (which are not repeated here because of the space limitation). It contains, in particular, the definition, typology and review of the methods for calibration of food analysers. In this paper, a new algorithm of calibration is proposed, the algorithm based on the method of principal components analysis (PCA) which up to now has been used in chemometrics for identification of data structures rather than for parameter estimation. This method is applied here for obtaining an approximate solution of the calibration equation of the form:

$$\dot{\mathbf{X}}|_{N \times M} \cdot \dot{\mathbf{p}}|_{M \times 1} = \dot{\mathbf{y}}|_{N \times 1} \quad (1)$$

with respect to the vector of parameters $\dot{\mathbf{p}}$, on the basis of an error-corrupted data matrix $\tilde{\mathbf{X}}$ representative of $\dot{\mathbf{X}}$ and an-error corrupted data vector $\tilde{\mathbf{y}}$ representative of $\dot{\mathbf{y}}$.

2. Principal components analysis

If the rows of the matrix $\dot{\mathbf{X}}|_{N \times M}$ are realisations of a zero-mean random vector $\underline{\mathbf{x}}$ whose covariance matrix is $\Sigma_{\mathbf{x}}$, then the covariance matrix of the vector $\mathbf{z}|_{M \times 1} \equiv \mathbf{L}|_{M \times M} \cdot \mathbf{x}|_{M \times 1}$ is:

$$\Sigma_{\mathbf{z}} \equiv \text{Cov}[\mathbf{z}] = \mathbf{L} \cdot \mathbf{E}[\underline{\mathbf{x}} \cdot \underline{\mathbf{x}}^T] \cdot \mathbf{L}^T = \mathbf{L} \cdot \Sigma_{\mathbf{x}} \cdot \mathbf{L}^T \quad (2)$$

¹ To whom any correspondence should be addressed.



Let's construct a row orthonormal matrix $\mathbf{L} \equiv [\mathbf{l}_1 \dots \mathbf{l}_N]^T$ whose rows $\mathbf{l}_n^T$ maximise the variances of the random variables $z_n \equiv \mathbf{l}_n^T \cdot \mathbf{x}$. The application of the method of Lagrange multipliers to this optimisation problem leads to the conclusion that the vectors $\mathbf{l}_n$ should satisfy the equations: $\Sigma_{\mathbf{x}} \cdot \mathbf{l}_n = \lambda_n \cdot \mathbf{l}_n$, which means that they are eigenvector of the covariance matrix $\Sigma_{\mathbf{x}}$ and λ_n are the corresponding eigenvalues of this matrix. If the matrix $\mathbf{L}$ is composed of the (normalised) eigenvectors of the matrix $\Sigma_{\mathbf{x}}$, ordered according to the decreasing magnitude of the corresponding eigenvalues λ_n , then the vector $\mathbf{z} \equiv [z_1 \dots z_N]^T$ is ordered according to the decreasing variance of its components. Since $\mathbf{z}^T \equiv \mathbf{x}^T \cdot \mathbf{L}^T$, the corresponding transformation of the realisations of the random vectors has the form: $\dot{\mathbf{Z}} \equiv \dot{\mathbf{X}} \cdot \mathbf{L}^T$, where $\dot{\mathbf{X}} \equiv [\dot{\mathbf{x}}_1 \dots \dot{\mathbf{x}}_N]^T$ and $\dot{\mathbf{Z}} \equiv [\dot{\mathbf{z}}_1 \dots \dot{\mathbf{z}}_N]^T$. The above transformation may be conveniently implemented using the singular-value decomposition (SVD) of the matrix $\dot{\mathbf{X}}$: $\dot{\mathbf{X}} = \dot{\mathbf{U}} \cdot \dot{\Sigma} \cdot \dot{\mathbf{V}}^T$, where the matrices $\dot{\mathbf{U}}|_{N \times N}$ and $\dot{\mathbf{V}}|_{M \times M}$ are orthonormal, and the matrix $\dot{\Sigma}$ has the form:

$$\dot{\Sigma}|_{N \times M} \equiv \begin{cases} \begin{bmatrix} \text{diag}\{\dot{\sigma}_1, \dots, \dot{\sigma}_M\} & \mathbf{0} \end{bmatrix} & \text{if } N > M \\ \begin{bmatrix} \text{diag}\{\dot{\sigma}_1, \dots, \dot{\sigma}_M\} \\ \mathbf{0} \end{bmatrix} & \text{if } N < M \end{cases} \quad (3)$$

Under an assumption that the singular values $\dot{\sigma}_n$ are ordered ($\dot{\sigma}_1 \geq \dot{\sigma}_2 \geq \dot{\sigma}_3 \geq \dots$), the variances of the columns of the matrix $\dot{\mathbf{Z}} \equiv \dot{\mathbf{X}} \cdot \dot{\mathbf{V}} = \dot{\mathbf{U}} \cdot \dot{\Sigma}$ are also ordered according to the decreasing magnitude. The mean values of the columns of $\dot{\mathbf{Z}}$ are all equal to zero because:

$$\sum_{n=1}^N \dot{z}_{n,m} = \sum_{n=1}^N \left(\sum_{\mu=1}^M \dot{x}_{n,\mu} \dot{v}_{\mu,m} \right) = \sum_{\mu=1}^M \left(\sum_{n=1}^N \dot{x}_{n,\mu} \right) \dot{v}_{\mu,m} = 0 \quad \text{for } m=1, \dots, M \quad (4)$$

where $\dot{z}_{n,m}$, $\dot{x}_{n,\mu}$ and $\dot{v}_{\mu,m}$ are the elements of the corresponding matrices. The m th column of the matrix $\dot{\mathbf{Z}}$ has the form $\dot{\sigma}_m \dot{\mathbf{u}}_m$, where $\dot{\mathbf{u}}_m$ is the m th column of the matrix $\dot{\mathbf{U}}$; so, its variance is $\dot{\sigma}_m^2/N$ due to the orthonormality of $\dot{\mathbf{U}}$ implying $\|\dot{\mathbf{u}}_m\|_2^2 = 1$. If the matrix $\dot{\mathbf{X}}$ is not centred, then the columns of the matrix $\dot{\mathbf{Z}}$ should be sorted to order them according to the decreasing variance.

3. Criterion of discrimination

A criterion for rejection of the least informative components is a key element of the PCA-based algorithm for solving Eq.(1). In order to derive such a criterion, let's perform the singular value decomposition of $\tilde{\mathbf{X}}$: $\tilde{\mathbf{X}} = \tilde{\mathbf{U}} \cdot \tilde{\Sigma} \cdot \tilde{\mathbf{V}}^T$, compute the matrix $\tilde{\mathbf{Z}} = \tilde{\mathbf{U}} \cdot \tilde{\Sigma}$, and the variances of its columns $\tilde{v}_{Z,m}$ for $m=1, \dots, M$. Now, the corresponding reverse accumulated variances defined as:

$$\tilde{a}_{Z,m} = \sum_{\mu=m}^M \tilde{v}_{Z,\mu} \quad \text{for } m=1, \dots, M \quad (5)$$

may be used for elimination of non-informative components which have appeared due to the insufficient number of measurements ($N < M$) or due to the presence of random errors in the x -data. If $N < M$, then the variances $v_{Z,m}$ with indices larger than M are zero; so, the signal-to-noise ratio (SNR_X) seems to be a natural criterion for the rejection of the least informative components due to the errors in the x -data. Under an assumption that those errors are realisations of zero-mean independent random variables with the variances $\sigma_{X,n,m}^2$, SNR_X may be defined and calculated as follows:

$$SNR_X \equiv \|\dot{\mathbf{X}}\|_F^2 / E[\|\Delta \tilde{\mathbf{X}}\|_F^2] \cong \sum_{n=1}^N \sum_{m=1}^M \tilde{x}_{n,m}^2 / \sum_{n=1}^N \sum_{m=1}^M \sigma_{X,n,m}^2 \quad (6)$$

where $\|\cdot\|_F$ is the Frobenius norm, $\tilde{x}_{n,m}$ are elements of the matrix $\tilde{\mathbf{X}}$. One may reasonably assume that the share of rejected components in the accumulated variance of all components should be the same as the share of the power of noise in the power of signal, i.e. that a threshold for cutting-off the least informative components should be $a_{th} = \tilde{a}_{Z,1}/SNR_x$, and the indices of rejected components should start from $m_{th} = \inf \arg_m \{ \tilde{a}_{Z,m} < a_{th} \}$.

4. Proposed algorithm of calibration

The method of principal components may be used not only for identification of data structures, but also for estimation of model parameters or for model response prediction. The corresponding estimation method is most frequently called principal component(s) regression. Here, for the sake of nomenclature consistency, the name PCLS will be used for implementation of that method. Its derivation is based on an assumption that the statistical model of the data has the form:

$$\tilde{\mathbf{y}} = \tilde{\mathbf{X}} \cdot \mathbf{p} + \Delta \tilde{\mathbf{y}} \quad \text{and} \quad \tilde{\mathbf{X}} = \dot{\mathbf{X}} + \Delta \tilde{\mathbf{X}} \quad (7)$$

where $\tilde{\mathbf{y}}$ and $\tilde{\mathbf{X}}$ are column-centred. For the given realisations $\tilde{\mathbf{y}}$ and $\tilde{\mathbf{X}}$ of $\tilde{\mathbf{y}}$ and $\tilde{\mathbf{X}}$, respectively, the following operations are performed to obtain an estimate $\hat{\mathbf{p}}$ of $\mathbf{p}$. First, the matrix $\tilde{\mathbf{Z}} = \tilde{\mathbf{U}} \cdot \tilde{\mathbf{\Sigma}}$ is determined using matrices resulting from singular value decomposition of $\tilde{\mathbf{X}}$. Next, the columns of $\tilde{\mathbf{Z}}$, whose indices are larger than m_{th} , are removed. The matrix $\tilde{\mathbf{Z}}_0$ obtained in this way is then used for solving the following OLS normal equation:

$$\tilde{\mathbf{Z}}_0^T \cdot \tilde{\mathbf{Z}}_0 \cdot \hat{\mathbf{q}} = \tilde{\mathbf{Z}}_0^T \cdot \tilde{\mathbf{y}} \quad (8)$$

with respect to $\hat{\mathbf{q}}$ which is next used for computing $\hat{\mathbf{p}} = \tilde{\mathbf{V}} \cdot \hat{\mathbf{q}}$. The latter equality is implied by the following identity:

$$\dot{\mathbf{X}} \cdot \mathbf{p} = (\dot{\mathbf{U}} \cdot \dot{\mathbf{\Sigma}}) \cdot (\dot{\mathbf{V}}^T \cdot \mathbf{p}) = \dot{\mathbf{Z}} \cdot \mathbf{q} \quad (9)$$

where $\mathbf{q} \equiv \dot{\mathbf{V}}^T \cdot \mathbf{p}$ or $\mathbf{p} = \dot{\mathbf{V}} \cdot \mathbf{q}$ (because $\dot{\mathbf{V}}$ is orthonormal).

5. Example of spectrophotometric application

5.1. Problem formulation

This section is entirely devoted to an example of application of the proposed algorithm for determination of edible oil mixtures on the basis of spectrophotometric data. It is assumed here that an oil mixture to be analysed is composed of three known components – nut oil, corn oil and olive oil – and that the M -dimensional vectors of exact data – $\dot{\mathbf{s}}_1$, $\dot{\mathbf{s}}_2$ and $\dot{\mathbf{s}}_3$ – representative of their absorbance spectra, are available. The study is based on demi-synthetic data generated using the denoised and baseline-corrected real-world data representative of nut oil ($\dot{\mathbf{s}}_1$), corn oil ($\dot{\mathbf{s}}_2$) and olive oil ($\dot{\mathbf{s}}_3$), acquired by means of the FTIR spectrophotometer *Perkin Elmer System 2000*, set to the resolution 1 cm^{-1} .

According to the Lambert-Beer's law, the vector of exact absorbance data $\dot{\mathbf{s}}$, representative of the spectrum of the mixture, satisfies the equation:

$$\dot{\mathbf{s}} = c_1 \cdot \dot{\mathbf{s}}_1 + c_2 \cdot \dot{\mathbf{s}}_2 + c_3 \cdot \dot{\mathbf{s}}_3 \quad (10)$$

where $\mathbf{c} = [c_1 \ c_2 \ c_3]^T$ is the vector of concentrations of all components satisfying the following constraints: $c_1, c_2, c_3 \in [0, 1]$ and $c_1 + c_2 + c_3 = 1$. It is moreover assumed that the real-world absorbance data $\tilde{\mathbf{s}}$, representative of the spectrum of a mixture, are corrupted with random errors $\Delta \tilde{\mathbf{s}}$ resulting both from inaccurate preparation of the mixture and from imperfections of the spectrophotometer. Since $c_3 = 1 - c_1 - c_2$, the spectral data, representative of the mixture, may be modelled by means of the following set of M linear algebraic equations:

$$\tilde{\mathbf{s}} - \dot{\mathbf{s}}_3 = c_1 \cdot (\dot{\mathbf{s}}_1 - \dot{\mathbf{s}}_3) + c_2 \cdot (\dot{\mathbf{s}}_2 - \dot{\mathbf{s}}_3) + \Delta \tilde{\mathbf{s}} \quad (11)$$

The problem under study consists in estimation of the concentrations c_1 and c_2 . It is numerically ill-conditioned because the spectra of nut oil, corn oil and olive oil are very close to each other.

5.2. Generation of semi-synthetic data

Each vector of spectral data for calibration is synthesised according to the methodology imitating the laboratory procedure used for obtaining the real-world data. First, the reference (exact) values of the concentrations of nut oil ($\dot{c}_1$) and of corn oil ($\dot{c}_2$) are selected, and the value of the concentration of olive oil is calculated: $\dot{c}_3 = 1 - \dot{c}_1 - \dot{c}_2$. Next, the error-corrupted values of all concentrations are determined by emulation of the process of sample preparation which consists in mixing the measured-out volumes of the components: $\tilde{V}_1$, $\tilde{V}_2$ and $\tilde{V}_3$:

$$\tilde{c}_j = \tilde{V}_j / \tilde{V} \quad \text{for } j=1, 2, 3 \quad \text{with } \tilde{V} = \tilde{V}_1 + \tilde{V}_2 + \tilde{V}_3 \quad \text{and } \tilde{V}_j = \dot{V}_j (1 + \mathcal{G}_j) \quad (12)$$

where $\dot{V}_j$ are error-free values of the volumes, and $\mathcal{G}_j$ are realisations of zero-mean random variables $\underline{\mathcal{G}}_j$ with the variances modelling relative errors of measuring out the volumes. The resulting concentrations $\tilde{c}_j = \dot{c}_j (1 + \zeta_j)$ are subject to relative errors:

$$\zeta_j = \frac{1 + \mathcal{G}_j}{\dot{c}_1(1 + \mathcal{G}_1) + \dot{c}_2(1 + \mathcal{G}_2) + \dot{c}_3(1 + \mathcal{G}_3)} - 1 = \frac{1 + \mathcal{G}_j}{1 + \dot{c}_1\mathcal{G}_1 + \dot{c}_2\mathcal{G}_2 + \dot{c}_3\mathcal{G}_3} - 1 \quad \text{for } j=1, 2, 3 \quad (13)$$

which may be viewed as realisations of random variables $\underline{\zeta}_j$ related in the same way to $\underline{\mathcal{G}}_1$, $\underline{\mathcal{G}}_2$ and $\underline{\mathcal{G}}_3$ as ζ_j are related in to $\mathcal{G}_1$, $\mathcal{G}_2$ and $\mathcal{G}_3$ in the above equation. The moments of those random variables depend on concentrations and on the variance $\sigma_{\mathcal{G}}^2$ of $\underline{\mathcal{G}}_j$ in the following way:

$$\mathbb{E}[\underline{\zeta}_1] \equiv (\dot{c}_1^2 + \dot{c}_2^2 + \dot{c}_3^2 - \dot{c}_1) \sigma_{\mathcal{G}}^2 \quad \text{and} \quad \mathbb{E}[\underline{\zeta}_2] \equiv (\dot{c}_1^2 + \dot{c}_2^2 + \dot{c}_3^2 - \dot{c}_2) \sigma_{\mathcal{G}}^2 \quad (14)$$

$$\text{Var}[\underline{\zeta}_1] \equiv ((1 - \dot{c}_1)^2 + \dot{c}_2^2 + \dot{c}_3^2) \sigma_{\mathcal{G}}^2 \quad \text{and} \quad \text{Var}[\underline{\zeta}_2] \equiv (\dot{c}_1^2 + (1 - \dot{c}_2)^2 + \dot{c}_3^2) \sigma_{\mathcal{G}}^2 \quad (15)$$

$$\text{Cov}[\underline{\zeta}_1, \underline{\zeta}_2] \equiv (\dot{c}_1^2 + \dot{c}_2^2 + \dot{c}_3^2 - \dot{c}_1 - \dot{c}_2) \sigma_{\mathcal{G}}^2 \quad (16)$$

The spectral data $\tilde{\mathbf{s}} \equiv [\tilde{s}_1 \ \tilde{s}_2 \ \dots]^T$, corresponding to a given triplet of concentrations – $\tilde{c}_1$, $\tilde{c}_2$ and $\tilde{c}_3 = 1 - \tilde{c}_1 - \tilde{c}_2$ determined after Eq.(12) – are generated using the formulae:

$$\tilde{\mathbf{s}} = \tilde{c}_1 \dot{\mathbf{s}}_1 + \tilde{c}_2 \dot{\mathbf{s}}_2 + \tilde{c}_3 \dot{\mathbf{s}}_3 + \Delta \mathbf{z} \quad (17)$$

where $\Delta \mathbf{z} \equiv [\Delta z_1 \ \dots \ \Delta z_M]^T$ is a realisation of a random vector $\underline{\Delta \mathbf{z}}$ modelling additive errors of spectrum measurement. By comparing Eq.(17) with Eq.(11) and randomising the result of comparison, one may get:

$$\underline{\Delta \tilde{\mathbf{s}}} = \mathbf{A} \cdot \underline{\zeta} + \underline{\Delta \mathbf{z}} \quad \text{with} \quad \mathbf{A} \equiv [\dot{c}_1 \cdot (\dot{\mathbf{s}}_1 - \dot{\mathbf{s}}_3) \ \dot{c}_2 \cdot (\dot{\mathbf{s}}_2 - \dot{\mathbf{s}}_3)] \quad \text{and} \quad \underline{\zeta} \equiv [\underline{\zeta}_1 \ \underline{\zeta}_2]^T \quad (18)$$

Under an assumption that $\underline{\Delta \mathbf{z}}$ is a zero-mean random vector whose components are statistically independent of $\underline{\zeta}_1$ and $\underline{\zeta}_2$:

$$\text{Cov}[\underline{\Delta \tilde{\mathbf{s}}}] = \mathbf{A} \cdot \Sigma_{\zeta} \cdot \mathbf{A}^T + \Sigma_{\mathbf{z}} \quad \text{with} \quad \Sigma_{\zeta} \equiv \begin{bmatrix} \text{Var}[\underline{\zeta}_1] & \text{Cov}[\underline{\zeta}_1, \underline{\zeta}_2] \\ \text{Cov}[\underline{\zeta}_1, \underline{\zeta}_2] & \text{Var}[\underline{\zeta}_2] \end{bmatrix} \quad (19)$$

where $\Sigma_{\mathbf{z}}$ is the covariance matrix of $\underline{\Delta \mathbf{z}}$, and Σ_{ζ} is the covariance matrix of $\underline{\zeta}$. Using this equation, one may determine the average power $\mathbb{E}[\underline{\Delta \tilde{\mathbf{s}}}^T \cdot \underline{\Delta \tilde{\mathbf{s}}}]$ of the noise $\underline{\Delta \tilde{\mathbf{s}}}$:

$$\mathbb{E}[\underline{\Delta \tilde{\mathbf{s}}}^T \cdot \underline{\Delta \tilde{\mathbf{s}}}] \equiv \sum_{m=1}^M \mathbf{a}_m^T \cdot \Sigma_{\zeta} \cdot \mathbf{a}_m + \sum_{m=1}^M \sigma_{z,m}^2 \quad (20)$$

where $\mathbf{a}_m^T$ is the m th row of the matrix $\mathbf{A}$. For generation of the relative errors of the volumes ($\mathcal{V}_k$), uncorrelated pseudorandom numbers, following the 3-std-truncated normal distribution $\mathcal{N}_3(0, \sigma_g^2)$, have been used; for generation of the errors of spectrum measurements (Δz_n) – uncorrelated pseudorandom numbers following the distribution $\mathcal{N}_3(0, \sigma_z^2)$. The set of data for calibration $\tilde{\mathbf{D}}^{cal}$ has been assumed to contain $N=121$ pairs of the following values of concentrations:

$$\dot{c}_1^{cal}, \dot{c}_2^{cal} \in \{k \cdot 0.01 \mid k=0, 1, \dots, 10\} \quad (21)$$

and the corresponding sequences of spectral data, each composed of $M=512$ points. The set of data for validation $\tilde{\mathbf{D}}^{val}$ has been assumed to contain $N'=441$ pairs of the following values of concentrations:

$$\dot{c}_1^{val}, \dot{c}_2^{val} \in \{k \cdot 0.005 \mid k=0, 1, \dots, 9\} \quad (22)$$

and the corresponding sequences of spectral data, each composed of $M=512$ points.

According to the inverse-model-based approach, the calibration consists in solving a set of linear algebraic equations defined by Eq.(1) with:

$$\tilde{\mathbf{X}} \equiv \tilde{\mathbf{S}}^{cal} - \dot{\mathbf{S}}_3 \quad \text{and} \quad \tilde{\mathbf{y}} \equiv \dot{\mathbf{c}}_j^{cal} \quad \text{for } j=1, 2 \quad (23)$$

where $\dot{\mathbf{c}}_1^{cal}$ and $\dot{\mathbf{c}}_2^{cal}$ are the vectors containing the reference concentrations defined by Eq.(21), $\tilde{\mathbf{S}}^{cal}$ is the matrix whose rows contain corresponding spectral data, and $\dot{\mathbf{S}}_3$ is the matrix whose all rows are equal to $\dot{\mathbf{s}}_3^T$. If $\tilde{\mathbf{y}} = \dot{\mathbf{c}}_j^{cal}$, then the procedure of calibration yields an estimate $\hat{\mathbf{p}}_j$ of the vectors of parameters $\mathbf{p}$, which may be used for computing an approximate value of $\dot{c}_j$ on the basis of any vector of acquired spectral data $\tilde{\mathbf{s}}$, using the formula: $\hat{c}_j = \hat{\mathbf{p}}_j \cdot (\tilde{\mathbf{s}} - \dot{\mathbf{s}}_3)$ for $j=1, 2$.

5.3. Methodology and selected results of comparison

The data for calibration $\tilde{\mathbf{D}}^{cal}$ have been synthesised according to the methodology described in Subsection 5.2 using the values $\sigma_g = 2.2 \cdot 10^{-3}$ and $\sigma_z = 3.0 \cdot 10^{-6}$. The data for validation $\tilde{\mathbf{D}}^{val}$ have been generated using the values $\sigma_g = 0$ and $\sigma_z = 3.0 \cdot 10^{-6}$. The latter assumption is reflecting the laboratory situation where the concentrations for validation may be measured with a significantly smaller uncertainty than that expected to be reached by the algorithm under study. It has been used to facilitate the interpretation of the results of numerical experimentation since in this way they are freed from that component of error of estimation which is propagated from validation data with the amplification coefficient equal to 1. The performance of the PCLS algorithm – with respect to the uncertainty of the final result of measurement – has been assessed using $R=100$ versions of the set of data $\tilde{\mathbf{D}}^{cal}$ and $R'=1000$ versions of the set of data $\tilde{\mathbf{D}}^{val}$, each corresponding to a different realisation of errors. For each version $\tilde{\mathbf{D}}^{cal}(r)$, $r=1, \dots, R$, of the set $\tilde{\mathbf{D}}^{cal}$, the full calibration procedure has been performed, and its results have been validated by means of R' versions $\tilde{\mathbf{D}}^{val}(r')$ of $\tilde{\mathbf{D}}^{val}$, $r'=1, \dots, R'$. For each pair of the values of concentrations $\dot{c}_1^{val}$ and $\dot{c}_2^{val}$, defined by Eq.(22), the estimates of two uncertainty indicators have been calculated, viz. of the worst-case error:

$$\hat{e}[\Delta \hat{c}_j^{val}] \equiv \sup \left\{ \left| \Delta \hat{c}_j^{val}(r, r') \right| \mid \begin{matrix} r=1, \dots, R \\ r'=1, \dots, R' \end{matrix} \right\} \quad (24)$$

and of the error standard deviation:

$$\hat{s}[\Delta \hat{c}_j^{val}] \equiv \sqrt{\frac{1}{RR'-1} \sum_{r=1}^R \sum_{r'=1}^{R'} \left(\Delta \hat{c}_j^{val}(r, r') - \hat{m}[\Delta \hat{c}_j^{val}] \right)^2} \quad \text{with} \quad \hat{m}[\Delta \hat{c}_j^{val}] \equiv \frac{1}{RR'} \sum_{r=1}^R \sum_{r'=1}^{R'} \Delta \hat{c}_j^{val}(r, r') \quad (25)$$

where $\Delta\hat{c}_j^{val}(r, r') \equiv \hat{c}_j^{val}(r, r') - \hat{c}_j^{val}$. Those indicators have been used for comparing the results of calibration obtained by means of the proposed algorithm with the corresponding results obtained by means of the OLS and PLS estimators. The first of them has been implemented according to the formula:

$$\hat{\mathbf{p}} = \arg_p \inf \left\{ \tilde{\mathbf{X}}^T \cdot \tilde{\mathbf{X}} \cdot \mathbf{p} = \tilde{\mathbf{X}}^T \cdot \tilde{\mathbf{y}} \right\} \quad (26)$$

For implementation of the PLS estimator, the procedure *pls* from *PLS_Toolbox* [2] – with manual optimisation of the number of latent variables – has been used. The indicators of estimation uncertainty $\hat{s}[\Delta\hat{c}_1^{val}]$ and $\hat{e}[\Delta\hat{c}_j^{val}]$, obtained for the PCLS algorithm and two reference algorithms, are compared in Figure 1.

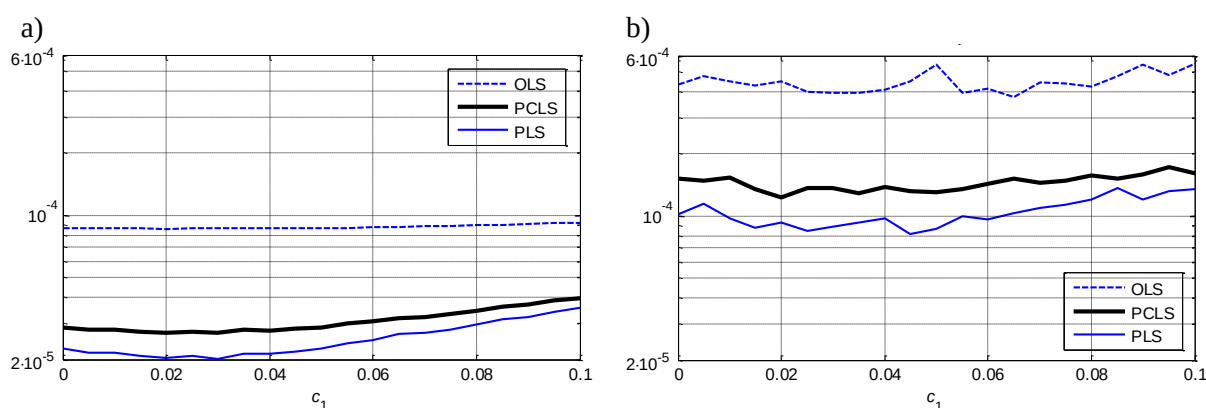


Figure 1. The uncertainty indicators: a) the error standard deviation $\hat{s}[\Delta\hat{c}_1^{val}]$,
b) the worst-case error $\hat{e}[\Delta\hat{c}_j^{val}]$.

6. Conclusion

A new PCA-based algorithm for calibration of spectrophotometric analysers of food has been proposed – the algorithm including fully automatic selection of the most informative components. It has been compared with two standard tools currently used for this purpose, viz. with the OLS estimator (being the simplest, but least accurate tool) and with the PLS estimator (being the most complex and most accurate one). The comparison has been based on a test problem consisting in determination of the composition of ternary oil mixtures. The results of comparison have demonstrated that the proposed algorithm is significantly more accurate than the OLS estimator, and only slightly less accurate than the PLS estimator. Being much simpler than the latter and fully automatic, the proposed algorithm seems to be an attractive alternative for the PLS estimator which is currently a favourite calibration tool in many areas of chemometrics.

Acknowledgements

This work has been supported by the National Science Centre in Poland (grant No. N N505 464832). The authors express their sincere gratitude to Dr. Grażyna Zofia Żukowska from the Faculty of Chemistry, Warsaw University of Technology, for the acquisition of data used for numerical experimentation reported in this paper.

References

- [1] Morawski R Z 2012 *Metrol. Meas. Syst.* **19** 623–652
- [2] *PLS_Toolbox*, Eigenvector Research, Inc., http://www.eigenvector.com/software/pls_toolbox.htm [as of February 10, 2013]