XX IMEKO World Congress
Metrology for Green Growth
September 9–14, 2012, Busan, Republic of Korea

A COMPARATIVE STUDY OF FORTY ALGORITHMS FOR SPECTROPHOTOMETRIC ANALYSIS OF EDIBLE OIL MIXTURES

Roman Z. Morawski and Andrzej Miekina

Warsaw University of Technology, Faculty of Electronics and Information Technology,
Institute of Radioelectronics, Nowowiejska 15/19, Warsaw 00-665, Poland

Abstract: This paper is devoted to the comparison of forty least-squares-type algorithms for spectrophotometric analysis of edible oil mixtures, *viz.* olive oil corrupted with corn oil and nut oil. The paper is providing a specification of the compared algorithms, a description of the methodology of their comparison, selected results of comparison, and conclusions drawn from them.

Keywords: food measurements, NIR spectrophotometry, estimation of concentrations

1. FORMULATION OF THE RESEARCH PROBLEM

It is assumed that an oil mixture to be analysed is composed of three known components, and that the vectors of exact data $\dot{\mathbf{s}}_j\big|_{M\times 1}$ ($j=1,2,3<M$), representative of the absorbance spectra of all those components, are available. 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:

$$\dot{\mathbf{s}} = c_1 \cdot \dot{\mathbf{s}}_1 + c_2 \cdot \dot{\mathbf{s}}_2 + c_3 \cdot \dot{\mathbf{s}}_3 \qquad (1)$$

where $\mathbf{c} = [c_1\ c_2\ c_3]^T$ is the vector of concentrations of all components, subject to the following constraints: $c_1, c_2, c_3 \in [0,1]$ and $c_1 + c_2 + c_3 = 1$. It is further 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 $c_1 + c_2 + c_3 = 1$, the spectral data, representative of the mixture, may be modelled by means of the following set of linear algebraic equations:

$$\tilde{\mathbf{s}} - \dot{\mathbf{s}}_3 = \dot{c}_1 \cdot (\dot{\mathbf{s}}_1 - \dot{\mathbf{s}}_3) + \dot{c}_2 \cdot (\dot{\mathbf{s}}_2 - \dot{\mathbf{s}}_3) + \Delta\tilde{\mathbf{s}} \qquad (2)$$

The problem under study consists in estimation of the concentrations c_1 and c_2 . Numerous algorithms have been developed and applied for solving this problem, also when the component spectra are very similar (like those of olive oil and corn oil). In this paper, forty algorithms – when applied to mixtures of nut oil, corn oil and olive oil – are compared using a consistent set of evaluation criteria.

2. COMPARED ALGORITHMS

Each of the compared algorithms is specified by the following features:

- the estimation method (five options);
- the calibration approach (two options);
- the absence or presence of data preprocessing (two options);
- the absence or presence of data standardisation (two options).

The following estimation methods, described in Appendix, have been included in the comparison:

- the ordinary least-squares estimator (OLS),
- the ridge least-squares estimator (RiLS),
- the robust least-squares estimator (RoLS),
- the total least-squares estimator (TLS),
- the partial least-squares estimator (PLS).

According to the so-called forward-model-based approach (FMA):

- The calibration consists in determination of the error-free data $\dot{\mathbf{s}}_1$, $\dot{\mathbf{s}}_2$ and $\dot{\mathbf{s}}_3$, representative of the spectra of all components.
- The estimation of the concentrations consists in solving the set of linear algebraic equations, defined by Eq.(2), using one of the above-listed estimators.

According to the so-called inverse-model-based approach (IMA):

- The calibration consists in solving a set of linear algebraic equations:

$$\mathbf{P} \cdot \left(\tilde{\mathbf{S}}^{cal} - \dot{\mathbf{S}}_3\right) \cong \dot{\mathbf{C}}^{cal} \qquad (3)$$

  with respect to the matrix of parameters $\mathbf{P}$, using one of the above-listed estimators.
- The estimation of the concentrations consists in using the formula:

$$\hat{\mathbf{c}} = \mathbf{P} \cdot \left(\tilde{\mathbf{s}} - \dot{\mathbf{s}}_3\right) \qquad (4)$$

The columns of the matrix $\tilde{\mathbf{S}}^{cal}$ in Eq.(3) contain spectral data representative of the reference mixtures used for calibration, and the columns of the matrix $\dot{\mathbf{C}}^{cal}$ – the corresponding vectors of concentrations; all columns of the

matrix $\dot{\mathbf{S}}_3$ are equal to $\dot{\mathbf{s}}_3$. Since all component spectra are very similar; consequently, the problem is ill-conditioned.

For preprocessing of spectral data, the method described by the authors in [1] has been selected. It consists in multiplying vectors of those data by a weighing matrix $\mathbf{W}|_{K\times M}$ whose rows contain the discrete Walsh functions multiplied by $1/M$ and ordered according to the number of sign changes. Consequently, the estimates of $\dot{c}_1$ and $\dot{c}_2$ are obtained by solving the following system of linear algebraic equations:

$$\mathbf{W}\cdot\left(\tilde{\dot{\mathbf{s}}}-\dot{\mathbf{s}}_3\right)\cong\dot{c}_1\cdot\mathbf{W}\cdot\left(\dot{\mathbf{s}}_1-\dot{\mathbf{s}}_3\right)+\dot{c}_2\cdot\mathbf{W}\cdot\left(\dot{\mathbf{s}}_2-\dot{\mathbf{s}}_3\right) \qquad (5)$$

in case of the forward-model-based approach. In case of the inverse-model-based approach, the system of equations:

$$\mathbf{P}\cdot\mathbf{W}\cdot\left(\tilde{\dot{\mathbf{S}}}^{cal}-\dot{\mathbf{S}}_3\right)\cong\dot{\mathbf{C}}^{cal} \qquad (6)$$

is solved with respect to $\mathbf{P}$, and the formula

$$\hat{\mathbf{c}}=\mathbf{P}\cdot\mathbf{W}\cdot\left(\tilde{\dot{\mathbf{s}}}-\dot{\mathbf{s}}_3\right) \qquad (7)$$

is used for estimation of the concentrations $\dot{c}_1$ and $\dot{c}_2$.

The method used for standardisation of data, which consists of their mean-centring followed by unit-variance scaling, is described in Appendix.

The compared algorithms have been named using the acronyms whose meaning is explained in Table 1.

3. METHODOLOGY OF COMPARISON

The study has been based on the 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}. A set of sequences of the data, representative of oil mixtures, each containing $M=512$ points, are shown in Fig. 1. The data

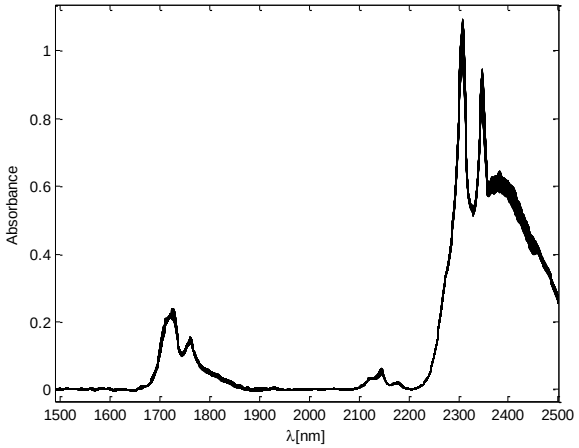

Fig. 1. The spectrophotometric data representative of olive oil, corn oil, nut oil and their mixtures (all in the grey area).

for calibration $\tilde{\mathbb{D}}^{cal}$ and for validation $\tilde{\mathbb{D}}^{val}$ have been synthesised according to the methodology described in [2], *i.e.* in a way imitating the laboratory procedure used for obtaining the real-world data. First, the reference values of the concentrations of nut oil ($\dot{c}_1$) and of corn oil ($\dot{c}_2$) have been selected, and the value of the concentration of olive oil has been calculated: $\dot{c}_3=1-\dot{c}_1-\dot{c}_2$. Next, the error-corrupted values of all concentrations have been determined by emulation of the process of sample preparation which consists in mixing up the measured-out volumes of the components, *viz.* V_1, V_2 and V_3.

Table 1. Acronyms of compared algorithms.

ESTIMATION METHOD	CALIBRATION APPROACH	PREPROCESSING OF EQUATIONS	STANDARDISATION OF DATA	ACRONYME OF ALGORITHMS
OLS	FMA	NO	NO	OLS-FMA
			YES	OLS-FMA-S
		YES	NO	OLS-FMA-P
			YES	OLS-FMA-P-S
	IMA	NO	NO	OLS-IMA
			YES	OLS-IMA-S
		YES	NO	OLS-IMA-P
			YES	OLS-IMA-P-S
RiLS	FMA	NO	NO	RiLS-FMA
			YES	RiLS-FMA-S
		YES	NO	RiLS-FMA-P
			YES	RiLS-FMA-P-S
	IMA	NO	NO	RiLS-IMA
			YES	RiLS-IMA-S
		YES	NO	RiLS-IMA-P
			YES	RiLS-IMA-P-S
RoLS	FMA	NO	NO	RoLS-FMA
			YES	RoLS-FMA-S
		YES	NO	RoLS-FMA-P
			YES	RoLS-FMA-P-S
	IMA	NO	NO	RoLS-IMA
			YES	RoLS-IMA-S
		YES	NO	RoLS-IMA-P
			YES	RoLS-IMA-P-S
TLS	FMA	NO	NO	TLS-FMA
			YES	TLS-FMA-S
		YES	NO	TLS-FMA-P
			YES	TLS-FMA-P-S
	IMA	NO	NO	TLS-IMA
			YES	TLS-IMA-S
		YES	NO	TLS-IMA-P
			YES	TLS-IMA-P-S
PLS	FMA	NO	NO	PLS-FMA
			YES	PLS-FMA-S
		YES	NO	PLS-FMA-P
			YES	PLS-FMA-P-S
	IMA	NO	NO	PLS-IMA
			YES	PLS-IMA-S
		YES	NO	PLS-IMA-P
			YES	PLS-IMA-P-S

The error-free values of concentrations are related to the error-free values of those volumes in the following way:

$$\dot{c}_j=\dot{V}_j/\dot{V} \text{ for } j=1,2,3 \qquad (8)$$

where $\dot{V}=\dot{V}_1+\dot{V}_2+\dot{V}_3$. The error-corrupted values of concentrations may be expressed in terms of their exact

values and the relative errors of measuring out the volumes, *viz.* ϑ_1, ϑ_2 and ϑ_3:

$$\tilde{c}_j = \frac{\tilde{V}_j}{\tilde{V}} = \frac{\dot{c}_j\left(1+\vartheta_j\right)}{\dot{c}_1\left(1+\vartheta_1\right)+\dot{c}_2\left(1+\vartheta_2\right)+\dot{c}_3\left(1+\vartheta_3\right)} \qquad (9)$$

where: $\tilde{V} = \tilde{V}_1 + \tilde{V}_2 + \tilde{V}_3$ and $\tilde{V}_j = \dot{V}_j\left(1+\vartheta_j\right)$ for $j = 1, 2, 3$.

The spectral data $\tilde{\mathbf{s}} \equiv [\tilde{s}_1 \ldots \tilde{s}_M]^T$, corresponding to a given triplet of concentrations – $\tilde{c}_1$, $\tilde{c}_2$ and $\tilde{c}_3$ determined after Eq.(10) – have been generated after the formula:

$$\tilde{\mathbf{s}} = \tilde{c}_1\dot{\mathbf{s}}_1 + \tilde{c}_2\dot{\mathbf{s}}_2 + \tilde{c}_3\dot{\mathbf{s}}_3 + \Delta\tilde{\mathbf{s}} \text{ for } n = 1, 2, \ldots \qquad (10)$$

For generation of the relative errors of the volumes (ϑ_k), uncorrelated pseudorandom numbers following a zero-mean normal distribution with the standard deviation $\sigma_\vartheta = 2.0 \cdot 10^{-3}$, truncated outside of the interval $[-3\sigma_\vartheta, +3\sigma_\vartheta]$, have been used; for generation of the errors introduced by the spectrophotometer – uncorrelated pseudorandom numbers following the zero-mean normal distribution with the standard deviation $\sigma_s = 1.0 \cdot 10^{-4}$, truncated outside of the interval $[-3\sigma_s, +3\sigma_s]$. Both values, σ_ϑ and σ_s, well correspond to the uncertainties of real-world mixture preparation and data acquisition.

The performance of each algorithm has been assessed with respect to the uncertainty of the final result of measurement (see [3] for details of this methodology), using a large set of validation data representative of NIR spectra of olive oil corrupted with (not more than 10 %) corn oil and nut oil (not more than 10 %).

The set of calibration data $\tilde{\mathbb{D}}^{cal}$ has been assumed to contain all the pairs ($N^{cal} = 121$) of the following values of concentrations:

$$\dot{c}_1^{cal}, \dot{c}_2^{cal} \in \left\{ k \cdot 0.01 \,\middle|\, k = 0, 1, \ldots, 10 \right\} \qquad (11)$$

and the corresponding spectral data. The set of data $\tilde{\mathbb{D}}^{val}$, used for validation of all compared algorithms, has been assumed to contain all the pairs ($N^{val} = 441$) of the following values of concentrations:

$$\dot{c}_1^{val}, \dot{c}_2^{val} \in \left\{ k \cdot 0.005 \,\middle|\, k = 0, 1, \ldots, 20 \right\} \qquad (12)$$

and the corresponding spectral data.

4. SELECTED RESULTS OF COMPARISON

The full programme of the comparative study has included:

- the data corrupted with both normally and uniformly distributed, both additive and multiplicative errors, as well as several levels of those errors;
- three indicators of estimation uncertainty: maximum value, standard deviation and bias of errors the estimates of both c_1 an c_2 are subject to.

Here, due to the limitation of space, only selected results obtained for additive errors with $\sigma_s = 1.0 \cdot 10^{-4}$ are presented. They are shown in Fig. 2 – Fig. 8. In each of Fig. 2 – Fig. 6, the standard deviations of c_1 estimation errors, under an assumption that $c_2 = 0.05$, are presented. In Fig. 7 and Fig. 8 the best results obtained for each estimator are compared.

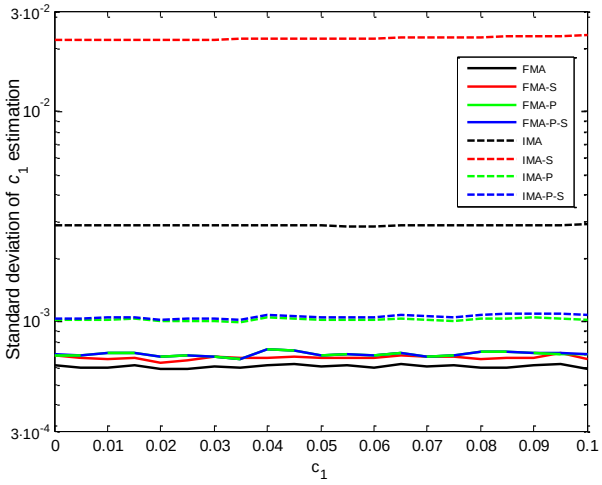

Fig. 2. The results obtained for 8 algorithms based on the OLS estimator.

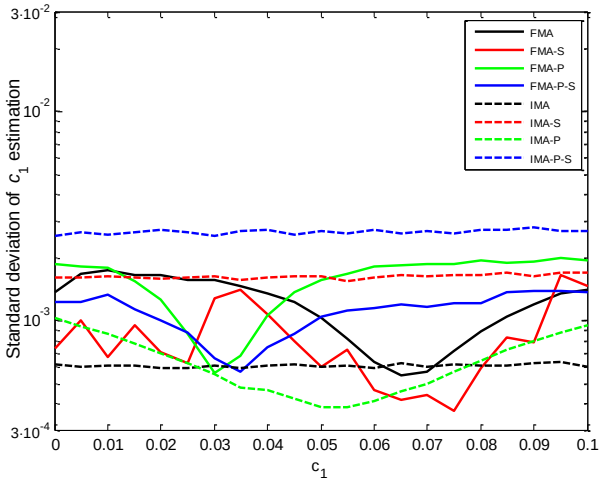

Fig. 3. The results obtained for 8 algorithms based on the RiLS estimator.

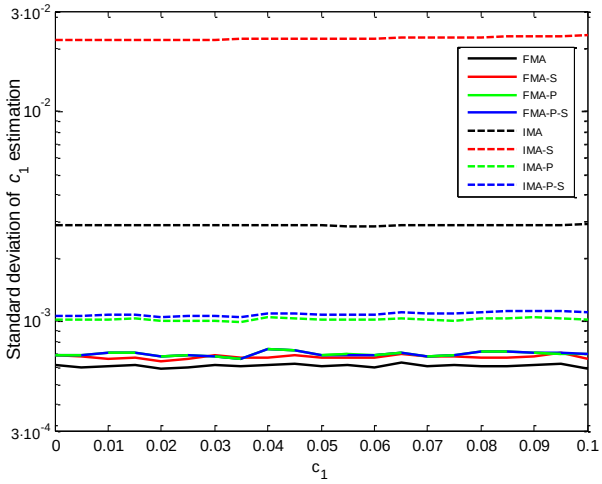

Fig. 4. The results obtained for 8 algorithms based on the RoLS estimator.

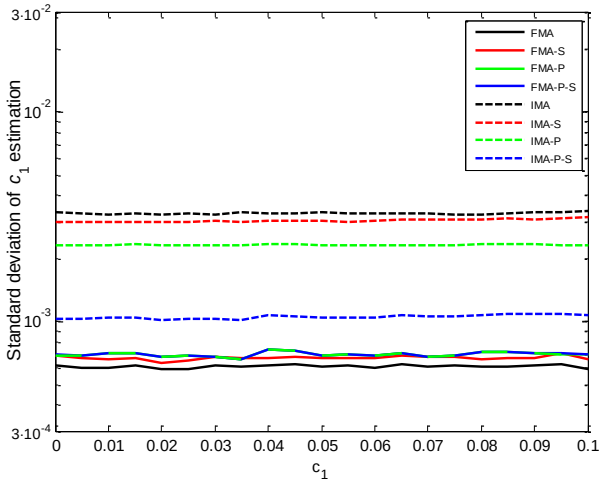

Fig. 5. The results obtained for 8 algorithms based on the TLS estimator.

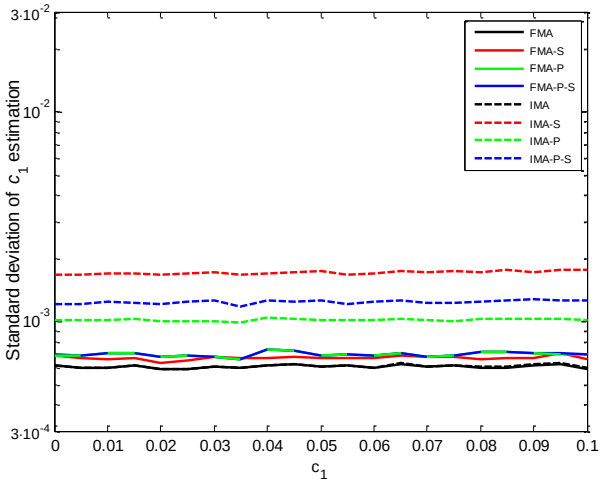

Fig. 6. The results obtained for 8 algorithms based on the PLS estimator.

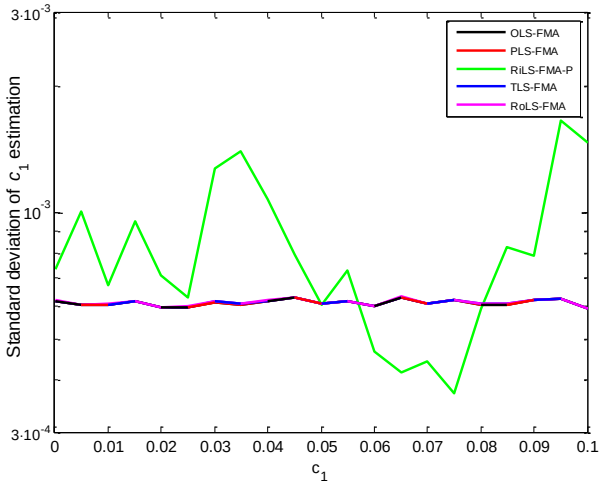

Fig. 7. The best results obtained for the FMA-based algorithms.

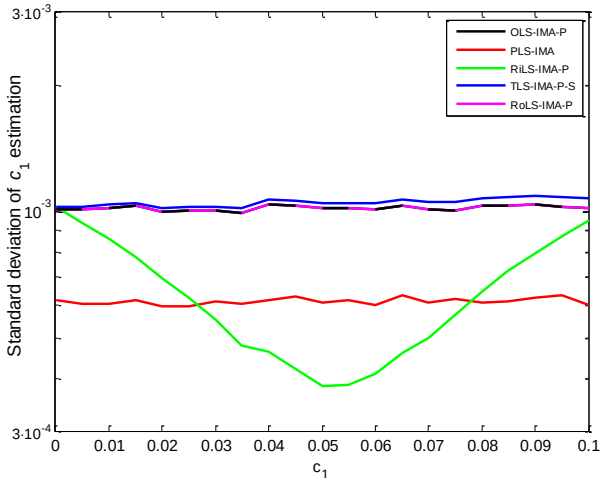

Fig. 8. The best results obtained for the IMA-based algorithms.

5. CONLUSIONS

Partial results of the comparative study, presented in this paper, do not cover all the aspects of applicability of the compared estimators for processing of NIR spectrometric data, aimed at determination of the concentrations of components of trinary oil mixtures (*e.g.* no information on bias of estimation and extended uncertainty has been provided). They are, however, representative enough to make plausible the following practical conclusions:

- The estimation bias, most important for the RiLS-based algorithms, is on the whole an order of magnitude smaller than the corresponding standard deviation.
- The FMA-based algorithms have provided better results than the IMA-based algorithms. This is not surprising since semi-synthetic data, used for comparison, are assumed to exactly follow the Lambert-Beer law. It should be, however, reminded that the FMA-based algorithms are more sensitive to any deviation from this law, and for real-world data may significantly bias the results of estimation if those deviations are non-negligible.
- For the FMA-based algorithms, the best results have been obtained without preprocessing and standardisation of data. They are very similar for all compared estimators, except the RiLS estimator. This is again not surprising since the conditioning number of the matrix of Eq.(2) is at the level of 8, and the preprocessing and standardisation of data is rather increasing than decreasing it.
- The best results for IMA-based algorithms have been obtained for the PLS estimator without preprocessing and standardisation of data. For the concentrations between 0.025 and 0.075, however, better results have been obtained for the RiLS estimator. This result seems to confirm the potential behind the RiLS estimator, already explored by the authors in [3].
- Very similar results have been obtained for the OLS, RoLS and TLS estimators with the preprocessing of data. The standardisation of data, frequently recommended in chemometrics, on the whole has not contributed to the reduction of estimation uncertainty (the only exception is the TLS-IMA algorithm).

– On the whole, the absolute standard deviation of estimation does not depend significantly on the value of the estimated concentration; so, the relative standard deviation of estimation grows hyperbolically for small values of that concentration.

Acknowledgment

This work has been supported by the Ministry of Science and Higher Education in Poland (grant No. N N505 464832).

APPENDIX: IMPLEMENTATION OF ESTIMATORS

This appendix contains the description of the least-squares-type estimators, used in comparative study, based on an assumption that they are applied to a system of linear algebraic equations of the form:

$$\dot{\mathbf{A}}\big|_{M\times N}\cdot\dot{\mathbf{x}}\big|_{N\times 1}=\dot{\mathbf{b}}\big|_{M\times 1} \qquad (13)$$

It is assumed that, instead of the exact matrix $\dot{\mathbf{A}}$ and the exact vector $\dot{\mathbf{b}}$, only their versions corrupted with random errors, *viz.*:

$$\tilde{\mathbf{A}}\equiv\dot{\mathbf{A}}+\Delta\tilde{\mathbf{A}} \text{ and } \tilde{\mathbf{b}}\equiv\dot{\mathbf{b}}+\Delta\tilde{\mathbf{b}} \qquad (14)$$

are available. The matrix $\Delta\tilde{\mathbf{A}}$ is a realisation of a random matrix $\underline{\Delta\tilde{\mathbf{A}}}$ whose columns are zero-mean random vectors; the vector $\Delta\tilde{\mathbf{b}}$ is a realisation of the zero-mean random vector $\underline{\Delta\tilde{\mathbf{b}}}$. The vector $\underline{\Delta\tilde{\mathbf{b}}}$ and the columns of the matrix $\underline{\Delta\tilde{\mathbf{A}}}$ are assumed to be pairwise statistically independent, and to have diagonal covariance matrices. Under the above assumptions, the exact solution of the system of equations, $\dot{\mathbf{x}}$, is on the whole not attainable, and various methods used for solving this system will generate its various estimates, *viz.*:

$$\hat{\mathbf{x}}\equiv\dot{\mathbf{x}}+\Delta\hat{\mathbf{x}} \qquad (15)$$

where $\Delta\hat{\mathbf{x}}$ is the error of estimation which should be evaluated to make the result of estimation $\hat{\mathbf{x}}$ meaningful.

The ordinary least-squares estimator (OLS) is derived by minimisation of the criterion:

$$J_{OLS}(\mathbf{x})\equiv\left\|\tilde{\mathbf{b}}-\tilde{\mathbf{A}}\cdot\mathbf{x}\right\|_2^2=\sum_{m=1}^{M}\left(\tilde{b}_m-\tilde{\mathbf{a}}_m^T\cdot\mathbf{x}\right)^2 \qquad (16)$$

where $\tilde{\mathbf{a}}_m^T$ are rows of the matrix $\tilde{\mathbf{A}}$. In practice, it consists in solving a system of algebraic equations (the so-called system of normal equations):

$$\hat{\mathbf{x}}_{OLS}=\arg_{\mathbf{x}}\left\{\tilde{\mathbf{A}}^T\cdot\tilde{\mathbf{A}}\cdot\mathbf{x}=\tilde{\mathbf{A}}^T\cdot\tilde{\mathbf{b}}\right\} \qquad (17)$$

In the reported study, it is implemented in MATLAB using the backslash operator ($\tilde{\mathbf{A}}\setminus\tilde{\mathbf{b}}$).

The ridge least-squares estimator (RiLS) is derived by constrained minimisation of the criterion $\|\mathbf{x}\|_2^2$:

$$\hat{\mathbf{x}}_{RiLS}=\arg_{\mathbf{x}}\inf\left\{\|\mathbf{x}\|_2^2\;\middle|\;\left\|\tilde{\mathbf{b}}-\tilde{\mathbf{A}}\cdot\mathbf{x}\right\|_2^2\le\mathrm{E}\left[\left\|\underline{\tilde{\mathbf{b}}}-\underline{\tilde{\mathbf{A}}}\cdot\dot{\mathbf{x}}\right\|_2^2\right]\right\} \qquad (18)$$

In case of additive errors:

$$\mathrm{E}\left[\left\|\underline{\tilde{\mathbf{b}}}-\underline{\tilde{\mathbf{A}}}\cdot\dot{\mathbf{x}}\right\|_2^2\right]=M\cdot\left(\sigma_b^2+\|\hat{\mathbf{x}}\|_2^2\cdot\sigma_a^2\right) \qquad (19)$$

where σ_a^2 is the variance of each element of the matrix $\underline{\tilde{\mathbf{A}}}$, and σ_b^2 is the variance of each element of the vector $\underline{\tilde{\mathbf{b}}}$. In the reported study, the RiLS estimator is implemented using the TOMLAB procedure of constrained optimisation ***snopt***.

The robust least-squares estimators (RoLS) is a modification of the OLS estimator, aimed at diminishing its sensitivity to the outliers in the data. This is achieved by replacing the optimisation criterion $J_{OLS}(\mathbf{x})$ with:

$$J_{RoLS}(\mathbf{x})=\sum_{m=1}^{M}\rho\left(\tilde{b}_m-\tilde{\mathbf{a}}_m^T\cdot\mathbf{x}\right) \qquad (20)$$

suppressing greater components of the vector $\tilde{\mathbf{b}}-\tilde{\mathbf{A}}\cdot\mathbf{x}$. In the reported study, the function $\rho(\Delta)$ is assumed to have the form:

$$\rho(\Delta_n)=\begin{cases}\Delta^2 & \text{for } |\Delta|\le\Delta_{th}\\ \Delta_{th}\left(2|\Delta|-\Delta_{th}\right) & \text{otherwise}\end{cases} \qquad (21)$$

where:

$$\Delta_{th}=\tfrac{1}{\sqrt{M}}\left\|\tilde{\mathbf{b}}-\tilde{\mathbf{A}}\cdot\hat{\mathbf{x}}_{OLS}\right\|_2 \qquad (22)$$

For more details see [4 – Chapter 3].

The total least-squares estimator (TLS) is resulting from the minimisation of the criterion:

$$J_{TLS}(\mathbf{A},\mathbf{x})=Tr\left(\left[\tilde{\mathbf{A}}-\mathbf{A}\middle|\tilde{\mathbf{b}}-\mathbf{A}\cdot\mathbf{x}\right]\cdot\left[\tilde{\mathbf{A}}-\mathbf{A}\middle|\tilde{\mathbf{b}}-\mathbf{A}\cdot\mathbf{x}\right]^T\right) \qquad (23)$$

In the reported study, it is implemented using a closed-form solution of this minimisation problem [5 – § 3.12]:

$$\hat{\mathbf{x}}_{TLS}=-\frac{1}{v_{M,L}}\begin{bmatrix}v_{1,L}\\ \vdots\\ v_{M,L}\end{bmatrix} \qquad (24)$$

where $L\in\{3,\ldots,121\}$, and $v_{m,n}$ are elements of the matrix $\mathbf{V}$ resulting from SVD of the matrix $\left[\tilde{\mathbf{A}}\,\middle|\,\tilde{\mathbf{b}}\right]$:

$$\left[\tilde{\mathbf{A}}\,\middle|\,\tilde{\mathbf{b}}\right]=\mathbf{U}\cdot\mathbf{\Sigma}\cdot\mathbf{V}^T \qquad (25)$$

The partial least-squares estimator (PLS) has been designed to overcome the collinearity of the columns of $\tilde{\mathbf{A}}$, and to make possible the identification of the components with small variances, strongly correlated with a random variable $\underline{b}$ modelling the elements of the vector $\tilde{\mathbf{b}}$. In terms of statistics, the latent variables $\underline{t}_1,\ldots,\underline{t}_K$ are sought for that maximally depend on $\underline{b}$. They are assumed to be linear combinations of the components of a random vector $\underline{\mathbf{a}}$ modelling the columns of the matrix $\tilde{\mathbf{A}}$:

$$\underline{t}_k=\mathbf{w}_k^T\cdot\underline{\mathbf{a}} \text{ with } \|\mathbf{w}_k\|=1 \qquad (26)$$

maximally correlated with $\underline{b}$, and mutually uncorrelated. Thus, the weighing vector $\mathbf{w}_1$ is determined from the equation:

$$\hat{\mathbf{w}}_1 = \arg_{\mathbf{w}_1} \sup\left\{\mathbf{w}_1^T \cdot \mathbf{c}_{ab} \,\middle|\, \|\mathbf{w}_1\| = 1\right\} \qquad (27)$$

where $\mathbf{c}_{ab} = \mathrm{E}\left[\underline{\mathbf{a}} \cdot \underline{b}\right]$. The weighing vectors $\mathbf{w}_k$ for $k > 1$ are determined from the equations:

$$\hat{\mathbf{w}}_k = \arg_{\mathbf{w}_k} \sup\left\{ \mathbf{w}_k^T \cdot \mathbf{c}_{ab} \,\middle|\, \begin{array}{l} \|\mathbf{w}_k\| = 1, \\ \mathbf{w}_k^T \cdot \mathbf{C}_{aa} \cdot \hat{\mathbf{w}}_1 = 0 \\ \cdots \\ \mathbf{w}_k^T \cdot \mathbf{C}_{aa} \cdot \hat{\mathbf{w}}_{k-1} = 0 \end{array} \right\} \qquad (28)$$

where $\mathbf{C}_{aa} = \mathrm{E}\left[\underline{\mathbf{a}} \cdot \underline{\mathbf{a}}^T\right]$. The correlation vector $\mathbf{c}_{ab}$ and matrix $\mathbf{C}_{aa}$ are estimated on the basis of the data $\tilde{\mathbf{A}}$ and $\tilde{\mathbf{b}}$:

$$\hat{\mathbf{C}}_{aa} = \tilde{\mathbf{A}}^T \cdot \tilde{\mathbf{A}} \text{ and } \hat{\mathbf{c}}_{ab} = \tilde{\mathbf{A}}^T \cdot \tilde{\mathbf{b}} \qquad (29)$$

The matrix $\hat{\mathbf{W}} = \left[\hat{\mathbf{w}}_1 \ldots \hat{\mathbf{w}}_K\right]^T$ is next used for finding the OLS solution $\hat{\mathbf{q}}_{OLS}$ of the system of equations:

$$\hat{\mathbf{W}} \cdot \tilde{\mathbf{A}} \cdot \mathbf{q} \cong \tilde{\mathbf{b}} \qquad (30)$$

and the estimate $\hat{\mathbf{x}}_{PLS}$:

$$\hat{\mathbf{x}}_{PLS} = \hat{\mathbf{W}}^T \cdot \hat{\mathbf{q}}_{OLS} \qquad (31)$$

For more details see [5 – § 3.10.14] and [6]. In the reported study, the SIMPLS version of the PLS estimator – implemented as the procedure ***pls*** from *PLS_Toolbox* [7] – has been used.

The standardisation of data, applied in the comparative study, consists in column-by-column centring and scaling of the matrix $\tilde{\mathbf{A}}$. The operation of centring an M-dimensional vector $\mathbf{v}$ is defined by the following formula:

$$\ddot{\mathbf{v}} \equiv \mathbf{v} - \mathbf{1}_M \bar{v} \qquad (32)$$

where:

$$\mathbf{1}_M \equiv \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}_{M \times 1} \qquad (33)$$

and:

$$\bar{v} \equiv \frac{1}{M} \sum_{m=1}^{M} v_m = \frac{1}{M} \cdot \mathbf{1}_M^T \cdot \mathbf{v} \qquad (34)$$

Thus:

$$\ddot{\mathbf{v}} \equiv \mathbf{v} - \frac{1}{M} \mathbf{1}_M \cdot \mathbf{1}_M^T \cdot \mathbf{v} = \mathbf{C} \cdot \mathbf{v} \qquad (35)$$

where $\mathbf{C}$ is the idempotent matrix of centring:

$$\mathbf{C} \equiv \mathbf{I} - \frac{1}{M} \mathbf{1}_{M \times M} \qquad (36)$$

with:

$$\mathbf{1}_{M \times M} \equiv \begin{bmatrix} 1 & \cdots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \cdots & 1 \end{bmatrix}_{M \times M} \qquad (37)$$

The operation of centring followed by scaling of the vector $\mathbf{v}$ is defined by the following formula:

$$\ddot{\mathbf{v}} \equiv \frac{1}{s} \cdot \mathbf{C} \cdot \mathbf{v} \qquad (38)$$

where s is an estimate of the standard deviation of the elements of the vector $\mathbf{v}$, *viz.*:

$$s \equiv \sqrt{\frac{1}{M-1} (\mathbf{C} \cdot \mathbf{v})^T \cdot (\mathbf{C} \cdot \mathbf{v})} = \sqrt{\frac{1}{M-1} \mathbf{v}^T \cdot \mathbf{C}^T \cdot \mathbf{C} \cdot \mathbf{v}} \qquad (39)$$

or:

$$s = \sqrt{\frac{1}{M-1} \mathbf{v}^T \cdot \mathbf{C} \cdot \mathbf{v}} \qquad (40)$$

If the vector $\mathbf{v}$ is subject to random additive zero-mean errors $\underline{\Delta \mathbf{v}}$, then the vector $\ddot{\mathbf{v}}$ is subject to random additive zero-mean errors $\underline{\Delta \ddot{\mathbf{v}}}$ whose covariance matrix is:

$$\mathrm{Cov}\left[\underline{\Delta \ddot{\mathbf{v}}}\right] \equiv \mathrm{E}\left[\underline{\Delta \ddot{\mathbf{v}}} \cdot \underline{\Delta \ddot{\mathbf{v}}}^T\right] = \frac{1}{s^2} \cdot \mathbf{C} \cdot \mathrm{E}\left[\underline{\Delta \mathbf{v}} \cdot \underline{\Delta \mathbf{v}}^T\right] \cdot \mathbf{C}^T \qquad (41)$$

or:

$$\mathrm{Cov}\left[\underline{\Delta \ddot{\mathbf{v}}}\right] = \frac{1}{s^2} \cdot \mathbf{C} \cdot \mathrm{Cov}\left[\underline{\Delta \mathbf{v}}\right] \cdot \mathbf{C}^T \qquad (42)$$

If $\mathrm{Cov}\left[\underline{\Delta \mathbf{v}}\right] = \sigma_v^2 \cdot \mathbf{I}$, then:

$$\mathrm{Cov}\left[\underline{\Delta \ddot{\mathbf{v}}}\right] = \frac{\sigma_v^2}{s^2} \cdot \mathbf{C} \cdot \mathbf{C}^T = \frac{\sigma_v^2}{s^2} \cdot \mathbf{C} \qquad (43)$$

This formula has been used in RiLS and RoLS estimators for computation of the variance of errors in the standardised data.

REFERENCES

[1] R. Z. Morawski, A. Miękina, J. Wagner: "A method of weighing matrix for spectrophotometric analysis of oil mixtures", [in:] *Advanced Mathematical and Computational Tools in Metrology and Testing IX* (Eds. F. Pavese, M. Bär, J.-R. Filtz, A. B. Forbes, L. Pendril, K. Shirono), World Scientific Publishing Company, Singapore-Hackensack-London 2012, pp. 276–283.

[2] A. Miękina, R. Z. Morawski: "Mathematical modelling of NIR spectral data and wavelength selection for determination of olive oil mixtures", *Journal of Physics: Conference Series*, Vol. 238, No. 1, 2010, doi:10.1088/1742-6596/238/1/012017.

[3] A. Miękina, R. Z. Morawski: "A calibration method, based on piecewise ridge LS estimator, designed for determination of olive oil mixtures on the basis of NIR spectral data", *Proc. XIX IMEKO World Congress (Lisbon, Portugal, September 6–11, 2009)*, pp. 2559–2563 (CD-ROM).

[4] H. Späth, *Mathematical Algorithms for Linear Regression*, Academic Press, London 1992.

[5] C. R. Rao, H. Toutenberg, *Linear Models – Least Squares and Alternatives*, Springer-Verlag, New York 1995.

[6] M. G. Gustafsson, "A Probabilistic Derivation of the Partial Least-Squares Algorithm", *Journal of Chemical Information and Computer Sciences*, 2001, Vol. 41, pp. 288–294.

[7] *PLS_Toolbox*, Eigenvector Research, Inc., *http://www.eigenvector.com/ software/pls_toolbox.htm* [as of June 1, 2012].