METROLOGY AND MEASUREMENT SYSTEMS
VOL. XV, NUMBER 4 (2008)

AGATA LATAŁA, ROMAN Z. MORAWSKI

Warsaw University of Technology, Faculty of Electronics and Information Technology
Institute of Radioelectronics
Warsaw, Poland
e-mail: r.morawski@ire.pw.edu.pl

COMPARISON OF LS-TYPE METHODS FOR DETERMINATION OF OLIVE OIL MIXTURES ON THE BASIS OF NIR SPECTRAL DATA

The spectrophotometric analysis of oil mixtures, containing olive oil, is the subject of this paper. Its objective is to compare six least-squares-type estimators which are potentially applicable for determination of a selected component of the mixture. The comparison presented is based on the criteria related to measurement uncertainty.

Keywords: least-squares estimators, NIR spectrophotometry, chemometrics, olive oil

1. INTRODUCTION

The quality and purity of olive oil, extensively used in the food industry, is of significant commercial importance. According to the EU regulations, in force since 2002, a manufacturer of products based on or containing olive oil, must either indicate the share of olive oil in the total weight of the product or the percentage of olive oil as percentage of the total fat. That is why increased interest in the methods for olive oil analysis has been observed for the last five years. Near-infrared (NIR) spectrophotometry, when combined with sophisticated procedures for spectrophotometric data processing, seems to be the most convenient and flexible tool for this application; consult, for example, [1-5] for more details. The comparison of numerous existing methods which are potentially suitable for this application, using both metrological and numerical criteria, seems to be of particular importance under those circumstances. This paper is devoted to the comparison of six least-squares-type methods most frequently used for estimation of the concentration of a selected component of an oil mixture, on the basis of the data representative of the NIR spectrum of this mixture, *viz.*: the ordinary least-squares estimator (OLS), the generalized least-squares estimator (GLS), the ridge least-squares estimator (RiLS), the robust least-squares estimator (RoLS), the total least-squares estimator (TLS) and the partial least-squares estimator (PLS).

[1] Received: October 20, 2008. Revised: November 21, 2008.

In the authors' conference paper [6], some preliminary results of their comparison were reported. Here the results of a more advanced stage of the comparative study are presented. The main improvements of the methodology of comparison consist in a modification of the procedure for data synthesis, in the use of the data subject to both correlated and uncorrelated measurement errors, and in more advanced optimization of the estimators RiLS, RoLS, TLS and PLS.

The following general rules are consistently used for generation of the mathematical symbols throughout this paper:
- $x, y, ...$ are real-valued scalar variables;
- $\dot{x}, \dot{y}, ...$ are exact values of the variables $x, y, ...$;
- $\hat{x}, \hat{y}, ...$ are estimated values of the variables $x, y, ...$

The diacritical signs, whose meaning has been explained above in reference to scalar variables, are applied in an analogues way with respect to vectors (x, y, ...) and matrices (X, Y, ...) of real-valued variables.

2. RESEARCH PROBLEM

It is assumed that an oil mixture to be analyzed is composed of J known components, and that the exact data $\dot{\mathbf{s}}_j|_{M\times 1}$ $(j = 1, ..., J < M)$, representative of the absorbance spectra of all those components are available. According to Lambert-Beer's law, the exact absorbance data $\dot{\mathbf{s}}$, representative of the spectrum of the mixture, satisfy the equation:

$$\dot{\mathbf{s}} = \sum_{j=1}^{J} c_j \dot{\mathbf{s}}_j \tag{1}$$

where $\mathbf{c} = [c_1 ... c_J]^T$ is the vector of (normalized) concentrations of components, subject to the following constraints:

$$\sum_{j=1}^{J} c_j = 1 \quad \text{and} \quad c_j \in [0, 1] \quad \text{for} \quad j = 1, ..., J \tag{2}$$

It is assumed that the real-world absorbance data $\tilde{\mathbf{s}}$, representative of the spectrum of a mixture, are corrupted by errors $\Delta\tilde{\mathbf{s}}$ resulting both from inaccurate preparation of the mixture and imperfections of the spectrophotometer:

$$\tilde{\mathbf{s}} = \dot{\mathbf{s}} + \Delta\tilde{\mathbf{s}} \tag{3}$$

The research problem, studied in this paper, consists in estimation of the concentration of only one component of the mixture, *viz.* c_1, by means of a linear estimator of the form:

$$\hat{c}_1 = \mathbf{p}^T \tilde{\mathbf{s}} \tag{4}$$

where $\mathbf{p} = [p_1 ... p_M]^T$ is a vector of parameters to be determined on the basis of a set of calibration data:

$$\tilde{\mathrm{D}}^{cal} = \left\{ \tilde{\mathbf{s}}_n^{cal}, \dot{c}_{1,n}^{cal} \middle| n = 1, ..., N \right\} \tag{5}$$

which – for the sake of convenience of numerical manipulations – are organized in a matrix and a vector:

$$\tilde{\mathbf{S}}^{cal} \equiv \left[\begin{array}{ccc} \tilde{\mathbf{s}}_1^{cal} & \dots & \tilde{\mathbf{s}}_N^{cal} \end{array} \right]^T \quad \text{and} \quad \dot{\mathbf{c}}^{cal} \equiv \left[\begin{array}{ccc} \dot{c}_{1,1}^{cal} & \dots & \dot{c}_{1,N}^{cal} \end{array} \right]^T \tag{6}$$

The parameters to be determined are assumed to satisfy the following approximate equality:

$$\tilde{\mathbf{S}}^{cal} \cdot \mathbf{p} \cong \dot{\mathbf{c}}^{cal} \tag{7}$$

which is the basis for development of all the methods that may be used for their estimation.

The performance of the six LS-type methods of estimation of the parameters $\mathbf{p}$ – *viz.* of OLS, GLS, RiLS, RoLS, TLS and PLS estimators – is compared on the basis of some criteria characterizing the uncertainty of the final result of analysis, defined by Eq.(4), using a set of validation data:

$$\tilde{\mathrm{D}}^{val} = \left\{ \tilde{\mathbf{s}}_n^{val}, \dot{c}_{1,n}^{val} \middle| n = 1, ..., N' \right\} \tag{8}$$

which are organized in a matrix and a vector:

$$\tilde{\mathbf{S}}^{val} \equiv \left[\begin{array}{ccc} \tilde{\mathbf{s}}_1^{val} & \dots & \tilde{\mathbf{s}}_{N'}^{val} \end{array} \right]^T \quad \text{and} \quad \dot{\mathbf{c}}^{val} \equiv \left[\begin{array}{ccc} \dot{c}_{1,1}^{val} & \dots & \dot{c}_{1,N'}^{val} \end{array} \right]^T \tag{9}$$

3. RESEARCH METHODOLOGY

The comparison of estimators has been based on the semi-synthetic data generated using the real-world data representative of corn oil, nut oil and olive oil ($J = 3$). The sequences of the latter data, each containing $N = 751$ data points, are shown in Fig. 1.

The data for calibration and validation have been synthesized in a way imitating the procedure used for obtaining the real-world data. First, the reference (exact) values of the concentrations of corn oil ($\dot{c}_1$) and of nut oil ($\dot{c}_2$) have been selected, and their error-corrupted versions calculated according to the scheme:

$$\tilde{c}_1 = \dot{c}_1 + \Delta\tilde{c}_1 \quad \text{and} \quad \tilde{c}_2 = \dot{c}_2 + \Delta\tilde{c}_2 \tag{10}$$

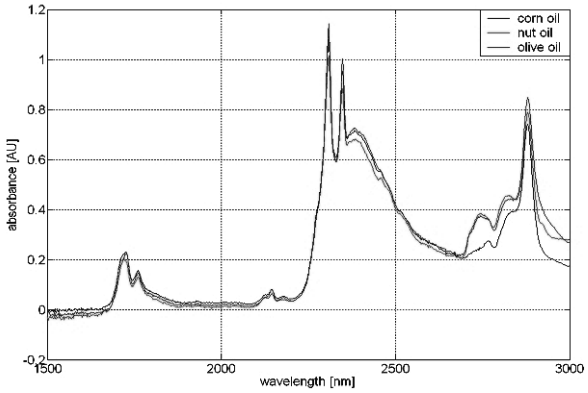

Fig. 1. The real-world data – representative of corn oil, nut oil and olive oil – acquired by means of a FTIR spectrophotometer set to a resolution of 1 cm^{-1}.

where $\Delta\tilde{c}_1$ and $\Delta\tilde{c}_2$ are variables modelling the errors of sample preparation. Next, the value of the concentration of olive oil has been calculated:

$$\tilde{c}_3 = 1 - \tilde{c}_1 - \tilde{c}_2 \tag{11}$$

Finally, the corresponding spectral data have been determined after the formula:

$$\tilde{\mathbf{s}} = \tilde{c}_1\dot{\mathbf{s}}_1 + \tilde{c}_2\dot{\mathbf{s}}_2 + \tilde{c}_3\dot{\mathbf{s}}_3 + \Delta\tilde{\mathbf{s}} \tag{12}$$

where $\dot{\mathbf{s}}_1$, $\dot{\mathbf{s}}_2$ and $\dot{\mathbf{s}}_3$ are the vectors of denoised and baseline-corrected real-world data representative of corn oil, nut oil and olive oil – respectively, and $\Delta\tilde{\mathbf{s}}$ is a vector modelling the errors of spectrum measurement. For generation of the values of $\Delta\tilde{c}_1$, $\Delta\tilde{c}_2$ and $\Delta\tilde{\mathbf{s}}$, pseudorandom numbers, following the normal distribution with the zero-mean and unit standard deviation, truncated outside of the interval [– 3, 3] have been used. Those numbers have been multiplied by the standard deviation $\sigma_c \in \left\{10^{-7}, 10^{-5}, 10^{-3}\right\}$ – to obtain $\Delta\tilde{c}_1$ and $\Delta\tilde{c}_2$ – or by the standard deviation $\sigma_s = 10^{-6}$ – to obtain elements of $\Delta\tilde{\mathbf{s}}$:

– the unprocessed sequences of pseudorandom numbers, with the standard deviation σ_c, have been used for generation of uncorrelated errors in concentration data;
– the same sequences, multiplied by the matrix:

$$\mathbf{C} = \begin{bmatrix} 0.8 & 0.6 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 0.8 & 0.6 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & 0.8 & 0.6 \\ 0 & 0 & 0 & \cdots & 0 & 0 & 0.8 \end{bmatrix}, \tag{13}$$

have been used for generation of correlated errors in concentration data;

– the sequences of random numbers, with the standard deviation σ_s, have been used for generation of uncorrelated errors in spectral data;
– the same sequences, multiplied by the matrix $\mathbf{C}$, have been used for generation of correlated errors in spectral data.

Consequently, the correlation of errors in the data has been characterized by the following covariance matrices:

$$\mathbf{\Sigma}_c = \sigma_c^2 \cdot \mathbf{C} \cdot \mathbf{C}^T \quad \text{and} \quad \mathbf{\Sigma}_s = \sigma_s^2 \cdot \mathbf{C} \cdot \mathbf{C}^T \tag{14}$$

where:

$$\mathbf{C} \cdot \mathbf{C}^T = \begin{bmatrix} 1.00 & 0.48 & 0 & \cdots & 0 & 0 & 0 \\ 0.48 & 1.00 & 0.48 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0.48 & 1.00 & 0.48 \\ 0 & 0 & 0 & \cdots & 0 & 0.48 & 0.64 \end{bmatrix} \tag{15}$$

The wavelength values to be used by the operator of concentration estimation, defined by Eq.(4), have been selected on the basis of the matrix $\tilde{\mathbf{S}}^{cal}$, containing in the consecutive rows complete sequences of spectral data synthesized for calibration. First, the eigenvalues and eigenvectors of the matrix $\tilde{\mathbf{S}}^{cal}$ have been computed and two eigenvectors, corresponding to the largest eigenvalues have been graphically represented as the functions of wavelength (*cf.* Fig. 2); the values of wavelength corresponding to the maxima of those functions have been chosen for further consideration. Next, each pair of the selected wavelength values, λ_{n1} and λ_{n2}, has been characterized by the conditioning number of the matrix composed of the columns of $\tilde{\mathbf{S}}^{cal}$ corresponding to λ_{n1} and λ_{n2}. Finally, the wavelength values, which contributed to the largest values of those conditioning numbers, have been eliminated. This procedure resulted in $M = 7$ wavelength values indicated in Fig. 2: $\lambda_1 = 1710$ nm, $\lambda_2 = 1728$ nm, $\lambda_3 = 2142$ nm, $\lambda_4 = 2336$ nm, $\lambda_5 = 2740$ nm, $\lambda_1 = 2880$ nm and $\lambda_7 = 2916$ nm.

The data for calibration $\tilde{\mathrm{D}}^{cal}$ have been synthesized using all the pairs ($N = 36$) of the following values of concentrations:

$$\dot{c}_1^{cal} \in \{0, 0.02, 0.04, 0.06, 0.08, 0.1\} \quad \text{and} \quad \dot{c}_2^{cal} \in \{0, 0.02, 0.04, 0.06, 0.08, 0.1\} \tag{16}$$

The data for validation $\tilde{\mathrm{D}}^{val}$ have been synthesized using all the pairs ($N' = 25$) of the following values of concentrations:

$$\dot{c}_1^{val} \in \{0.01, 0.03, 0.05, 0.07, 0.09\} \quad \text{and} \quad \dot{c}_2^{val} \in \{0.01, 0.03, 0.05, 0.07, 0.09\} \tag{17}$$

Thus, the validation has been carried out over an area of the $c_1 - c_2$ plane, slightly smaller that the area covered by the calibration data. In this way, the impact of bor-

der effects, masking the performance differentiation of compared methods, has been mitigated.

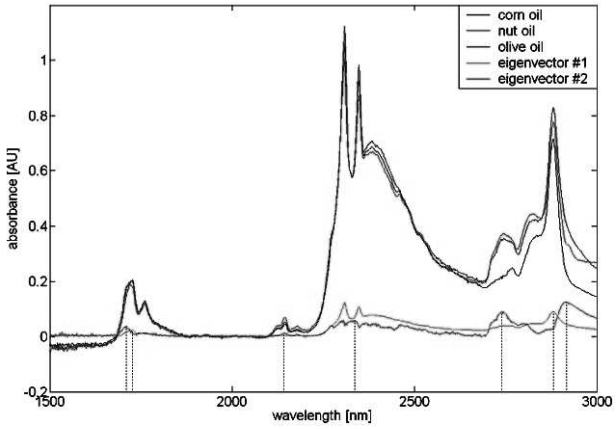

Fig. 2. The baseline-corrected spectrophotometric data, two eigenvectors of the data matrix, corresponding to its largest eigenvalues, and the wavelength values selected for experimentation (indicated with vertical dashed black lines).

For each pair of σ_s and σ_c, both for correlated and uncorrelated errors in the data:
– $R = 30$ versions of the set $\tilde{\mathrm{D}}^{cal}$, corresponding to R realisations of the errors, and
– R versions of the set $\tilde{\mathrm{D}}^{val}$, corresponding to R realisations of the errors,
have been generated. For each version $\tilde{\mathrm{D}}^{cal}(r)$, $r = 1, ..., R$, of the set $\tilde{\mathrm{D}}^{cal}$, an estimate $\hat{\mathbf{p}}(r)$ of the vector of parameters $\mathbf{p}$ has been calculated and used for validation. The validation has included the following steps:

– determination of the estimates $\hat{c}^{val}_{1,n}(r, r')$ of the concentration $\dot{c}^{val}_{1,n}$, corresponding to $\hat{\mathbf{p}}(r)$ and to each version $\tilde{\mathrm{D}}^{val}(r')$ of $\tilde{\mathrm{D}}^{val}$, $r' = 1, ..., R$;
– estimation of the bias according to the formula:

$$\hat{b}^{val}_{n}(r) = \bar{c}^{val}_{1,n}(r) - \dot{c}^{val}_{1,n} \tag{18}$$

where:

$$\bar{c}^{val}_{1,n}(r) = \frac{1}{R}\sum_{r'=1}^{R} \hat{c}^{val}_{1,n}(r, r') \tag{19}$$

– estimation of the standard deviation according to the formula:

$$\hat{\sigma}^{val}_{n}(r) = \sqrt{\frac{1}{R-1}\sum_{r'=1}^{R}\left[\hat{c}^{val}_{1,n}(r, r') - \bar{c}^{val}_{1,n}(r)\right]^2} \tag{20}$$

– computation of the expanded uncertainty according to the formula:

$$\hat{u}_n^{val}(r) = \left|\hat{b}_n^{val}(r)\right| + 3 \cdot \hat{\sigma}_n^{val}(r) \tag{21}$$

The results of validation, obtained for each $\hat{\mathbf{p}}(r)$, have been aggregated using the worst-case methodology:

$$\hat{b}^{val} = \sup\left\{\left|\hat{b}_n^{val}(r)\right| \middle| n = 1, ..., N; r = 1, ..., R\right\} \tag{22}$$

$$\hat{\sigma}^{val} = \sup\left\{\hat{\sigma}_n^{val}(r) \middle| n = 1, ..., N; r = 1, ..., R\right\} \tag{23}$$

$$\hat{u}^{val} = \sup\left\{\hat{u}_n^{val}(r) \middle| n = 1, ..., N; r = 1, ..., R\right\} \tag{24}$$

4. COMPARED METHODS OF ESTIMATION

The general definitions of the compared estimators, together with the relevant references, can be found in the review paper [7]. Here, only some specific features of their implementations, used for comparison, are characterised.

The OLS estimator has been implemented using the MATLAB operator \ according to the following formula:

$$\hat{\mathbf{p}}_{OLS} = \tilde{\mathbf{S}}^{cal} \backslash \tilde{\mathbf{c}}^{cal} \tag{25}$$

The GLS estimator has been implemented using the MATLAB operator \ according to the following formula:

$$\hat{\mathbf{p}}_{GLS} = \left[\left(\tilde{\mathbf{S}}^{cal}\right)^T \cdot \boldsymbol{\Sigma}_c^{-1} \cdot \tilde{\mathbf{S}}^{cal}\right] \backslash \left[\boldsymbol{\Sigma}_c^{-1} \cdot \left(\tilde{\mathbf{S}}^{cal}\right)^T\right] \tag{26}$$

The values of the regularization parameters in the PLS, RiLS, RoLS and TLS implementations have been selected as to minimize worst-case expanded uncertainty of concentration estimates: for each pair of σ_s and σ_c, both for correlated and uncorrelated errors in the data, the value of the regularisation parameter corresponding to the smallest value of $\hat{u}^{val}$ has been found. As a consequence, the limit estimation potential of the studied estimators has been compared rather than the performance of their particular versions corresponding to various methods applied for optimisation of regularisation parameters.

The RiLS estimator has been implemented using the MATLAB operator according to the formula:

$$\hat{\mathbf{p}}_{RiLS} = \left[\left(\tilde{\mathbf{S}}^{cal}\right)^T \cdot \tilde{\mathbf{S}}^{cal} + \alpha \cdot \mathbf{I}\right] \backslash \left(\tilde{\mathbf{S}}^{cal}\right)^T \cdot \tilde{\mathbf{c}}^{cal} \tag{27}$$

with the values of the regularization parameter α shown in Table 1.

Table 1. The values of α used in the implementation of the RiLS estimator for $\sigma_s = 10^{-6}$.

uncorrelated errors			correlated errors		
$\sigma_c = 10^{-7}$	$\sigma_c = 10^{-5}$	$\sigma_c = 10^{-3}$	$\sigma_c = 10^{-7}$	$\sigma_c = 10^{-5}$	$\sigma_c = 10^{-3}$
$1 \cdot 10^{-9}$	$1 \cdot 10^{-9}$	$3.16 \cdot 10^{-10}$	$1 \cdot 10^{-9}$	$3.16 \cdot 10^{-10}$	$1 \cdot 10^{-9}$

In the implementation of the RoLS estimator:

$$\hat{\mathbf{p}}_{RoLS} = \arg_{\mathbf{p}} \inf \left\{ \sum_{n=1}^{N^{cal}} \rho\left(\tilde{c}_{1,n}^{cal} - \mathbf{p}^T \cdot \tilde{\mathbf{s}}_n^{cal}; \Delta_{th}\right) \right\} \tag{28}$$

the Huber function of the form:

$$\rho(\Delta; \alpha) = \begin{cases} \Delta^2 & \text{for } |\Delta| \le \Delta_{th} \\ 0.01 \cdot (2 \cdot |\Delta| - 0.01) & \text{otherwise} \end{cases} \tag{29}$$

has been used with the values of the regularization parameter Δ_{th} shown in Table 2.

Table 2. The values of Δ_{th} used in the implementation of the RoLS estimator for $\sigma_s = 10^{-6}$.

uncorrelated errors			correlated errors		
$\sigma_c = 10^{-7}$	$\sigma_c = 10^{-5}$	$\sigma_c = 10^{-3}$	$\sigma_c = 10^{-7}$	$\sigma_c = 10^{-5}$	$\sigma_c = 10^{-3}$
10	10	10^{-10}	10	10	10^{-10}

The implementation of the PLS estimator has been based on the MATLAB procedures *pls* and *plspred* from the *Chemometrics Toolbox ver. 2.20* (1993). The selected values of the regularisation parameter L (number of latent variables) are shown in Table 3.

Table 3. The values of L used in the implementation of the PLS estimator for $\sigma_s = 10^{-6}$.

uncorrelated errors			correlated errors		
$\sigma_c = 10^{-7}$	$\sigma_c = 10^{-5}$	$\sigma_c = 10^{-3}$	$\sigma_c = 10^{-7}$	$\sigma_c = 10^{-5}$	$\sigma_c = 10^{-3}$
3	3	3	3	3	3

The values of the regularisation parameter K (the index of the selected eigenvector of the combined matrix $\left[\tilde{\mathbf{S}}^{cal}\tilde{\mathbf{c}}^{cal}\right]$), used in the implementation of the TLS estimator, are shown in Table 4.

Table 4. The values of K used in the implementation of the TLS estimator for $\sigma_s = 10^{-6}$.

uncorrelated errors			correlated errors		
$\sigma_c = 10^{-7}$	$\sigma_c = 10^{-5}$	$\sigma_c = 10^{-3}$	$\sigma_c = 10^{-7}$	$\sigma_c = 10^{-5}$	$\sigma_c = 10^{-3}$
6	4	4	4	5	4

5. RESULTS OF STUDY

The results of study are summarized in Tables 5-8. In Table 5, the limit error of the concentration estimates, obtained by means of the compared estimators for exact calibration and validation data, is shown. Its values characterize the accuracy of computation, i.e. the asymptotic level of estimation accuracy that may be approached if $\sigma_c \to 0$ and $\sigma_s \to 0$. In Tables 6-8, the performance of the studied estimators for non-zero σ_s and σ_c is compared, using the indicators $\hat{b}^{val}$, $\hat{\sigma}^{val}$ and $\hat{u}^{val}$, defined by Eqs.(22-24).

Table 5. The limit error of the concentration estimates, obtained by means of the compared estimators for exact calibration and validation data (σ_c and σ_s).

OLS	GLS	RiLS	RoLS	TLS	PLS
$8.67 \cdot 10^{-15}$	$8.67 \cdot 10^{-15}$	$8.67 \cdot 10^{-15}$	$8.67 \cdot 10^{-15}$	$4.23 \cdot 10^{-12}$	$9.74 \cdot 10^{-15}$

Table 6. The values of the performance indicator $\hat{b}^{val}$, obtained by means of the compared estimators for error-corrupted calibration and validation data ($\sigma_s = 10^{-6}$).

	uncorrelated errors			correlated errors		
	$\sigma_c = 10^{-7}$	$\sigma_c = 10^{-5}$	$\sigma_c = 10^{-3}$	$\sigma_c = 10^{-7}$	$\sigma_c = 10^{-5}$	$\sigma_c = 10^{-3}$
OLS	$3.41 \cdot 10^{-5}$	$3.99 \cdot 10^{-5}$	$1.13 \cdot 10^{-3}$	$3.02 \cdot 10^{-5}$	$3.73 \cdot 10^{-5}$	$1.45 \cdot 10^{-3}$
GLS	$3.41 \cdot 10^{-5}$	$3.99 \cdot 10^{-5}$	$1.13 \cdot 10^{-3}$	$5.26 \cdot 10^{-5}$	$5.87 \cdot 10^{-5}$	$1.52 \cdot 10^{-3}$
RiLS	$3.78 \cdot 10^{-5}$	$4.32 \cdot 10^{-5}$	$1.14 \cdot 10^{-3}$	$3.53 \cdot 10^{-5}$	$4.08 \cdot 10^{-5}$	$1.41 \cdot 10^{-3}$
RoLS	$1.91 \cdot 10^{-4}$	$1.44 \cdot 10^{-4}$	$3.46 \cdot 10^{-3}$	$1.73 \cdot 10^{-4}$	$2.16 \cdot 10^{-4}$	$3.39 \cdot 10^{-3}$
TLS	$1.00 \cdot 10^{-3}$	$2.98 \cdot 10^{-4}$	$1.06 \cdot 10^{-3}$	$6.02 \cdot 10^{-3}$	$8.93 \cdot 10^{-4}$	$1.35 \cdot 10^{-3}$
PLS	$3.15 \cdot 10^{-5}$	$3.68 \cdot 10^{-5}$	$1.14 \cdot 10^{-3}$	$3.15 \cdot 10^{-5}$	$3.68 \cdot 10^{-5}$	$1.14 \cdot 10^{-3}$

Table 7. The values of the performance indicator $\hat{\sigma}^{val}$, obtained by means of the compared estimators for error-corrupted calibration and validation data ($\sigma_s = 10^{-6}$).

	uncorrelated errors			correlated errors		
	$\sigma_c = 10^{-7}$	$\sigma_c = 10^{-5}$	$\sigma_c = 10^{-3}$	$\sigma_c = 10^{-7}$	$\sigma_c = 10^{-5}$	$\sigma_c = 10^{-3}$
OLS	$5.19 \cdot 10^{-5}$	$5.27 \cdot 10^{-5}$	$1.48 \cdot 10^{-3}$	$5.19 \cdot 10^{-5}$	$5.27 \cdot 10^{-5}$	$1.48 \cdot 10^{-3}$
GLS	$5.19 \cdot 10^{-5}$	$5.27 \cdot 10^{-5}$	$1.48 \cdot 10^{-3}$	$6.46 \cdot 10^{-5}$	$6.72 \cdot 10^{-5}$	$1.93 \cdot 10^{-3}$
RiLS	$3.78 \cdot 10^{-5}$	$4.32 \cdot 10^{-5}$	$1.14 \cdot 10^{-3}$	$3.53 \cdot 10^{-5}$	$4.08 \cdot 10^{-5}$	$1.41 \cdot 10^{-3}$
RoLS	$5.19 \cdot 10^{-5}$	$5.27 \cdot 10^{-5}$	$1.49 \cdot 10^{-3}$	$5.31 \cdot 10^{-5}$	$5.27 \cdot 10^{-5}$	$1.36 \cdot 10^{-3}$
TLS	$1.39 \cdot 10^{-3}$	$7.55 \cdot 10^{-4}$	$1.22 \cdot 10^{-3}$	$1.31 \cdot 10^{-3}$	$1.35 \cdot 10^{-3}$	$1.21 \cdot 10^{-3}$
PLS	$4.69 \cdot 10^{-5}$	$4.80 \cdot 10^{-5}$	$1.21 \cdot 10^{-3}$	$4.69 \cdot 10^{-5}$	$4.80 \cdot 10^{-5}$	$1.21 \cdot 10^{-3}$

Table 8. The values of the performance indicator $\hat{u}^{val}$, obtained by means of the compared estimators for error-corrupted calibration and validation data ($\sigma_s = 10^{-6}$).

	uncorrelated errors			correlated errors		
	$\sigma_c = 10^{-7}$	$\sigma_c = 10^{-5}$	$\sigma_c = 10^{-3}$	$\sigma_c = 10^{-7}$	$\sigma_c = 10^{-5}$	$\sigma_c = 10^{-3}$
OLS	$1.70 \cdot 10^{-4}$	$1.75 \cdot 10^{-4}$	$4.74 \cdot 10^{-3}$	$1.73 \cdot 10^{-4}$	$1.78 \cdot 10^{-4}$	$4.81 \cdot 10^{-3}$
GLS	$1.70 \cdot 10^{-4}$	$1.75 \cdot 10^{-4}$	$4.74 \cdot 10^{-3}$	$2.07 \cdot 10^{-4}$	$2.25 \cdot 10^{-4}$	$6.53 \cdot 10^{-3}$
RiLS	$1.56 \cdot 10^{-4}$	$1.65 \cdot 10^{-4}$	$4.22 \cdot 10^{-3}$	$1.61 \cdot 10^{-4}$	$1.61 \cdot 10^{-4}$	$4.20 \cdot 10^{-3}$
RoLS	$2.95 \cdot 10^{-4}$	$2.68 \cdot 10^{-4}$	$6.85 \cdot 10^{-3}$	$2.95 \cdot 10^{-4}$	$3.17 \cdot 10^{-4}$	$6.59 \cdot 10^{-3}$
TLS	$4.32 \cdot 10^{-3}$	$2.38 \cdot 10^{-3}$	$4.35 \cdot 10^{-3}$	$4.12 \cdot 10^{-3}$	$4.41 \cdot 10^{-3}$	$4.62 \cdot 10^{-3}$
PLS	$1.61 \cdot 10^{-4}$	$1.61 \cdot 10^{-4}$	$4.20 \cdot 10^{-3}$	$1.61 \cdot 10^{-4}$	$1.61 \cdot 10^{-4}$	$4.20 \cdot 10^{-3}$

6. DISCUSSION AND CONCLUSIONS

The comparison of the six LS-type estimators of concentration on the basis of spectrophotometric data, presented in this paper, has been intended for establishing a kind of benchmark for evaluation of more advanced – variational and nonlinear – estimators of concentration. This comparison is still of preliminary nature due to some methodological limitations:

– The selection of mixtures for data generation has been not optimized, and the selection of wavelength values for this purpose has been based on the simplest numerical criteria of ill-conditioning.
– The results of comparison have been obtained by means of semi-synthetic data and not yet confirmed by means of real-world data.

Despite the above-mentioned methodological limitations, the accomplished study enables the authors to conclude that:

- The problem under study (trinary oil mixture analysis) may be solved satisfactorily using five out of six compared estimators (OLS, GLS, RiLS, RoLS and PLS), provided the random noise in the spectral data is reduced to the level $\sigma_s = 10^{-6}$, which requires averaging of thousands of data records even in case of low-noise spectrophotometers.
- For $\sigma_s = 10^{-6}$, almost the same results are obtained for $\sigma_c = 10^{-7}$ and $\sigma_c = 10^{-5}$; so, there is no need to prepare calibration samples with extreme accuracy ($\sigma_c = 10^{-7}$) since the effect of this endeavour will be neutralised by the impact of errors in spectral data.
- The smallest (and almost identical) values of the performance indicator $\hat{u}^{val}$ have been obtained for the RiLS and PLS estimators; so, those two estimators should be used as a reference in the future study.
- For correlated data, larger values of the performance indicator $\hat{u}^{val}$ have been obtained for the GLS estimator than for the OLS estimator. This apparent anomaly may be explained by the way of data synthesis: the calibration set contains error-free concentration data and spectral data corrupted by both instrumental errors and errors due to inaccuracy of preparation of calibration samples.

The study, reported in this paper, has been carried out according to a new methodology whose main distinctive features are the following:

- The generation of semi-synthetic data has been performed according to a scheme imitating the procedure used for obtaining real-world data.
- The performance of the methods of calibration has been assessed on the basis of the criteria related the uncertainty of the final result of analysis, *i.e.* of the estimates of concentration.
- The worst-case performance indicators have been found over the space of data used for calibration and the space of data used for validation.

Thus, the results of this comparative study should be considered as a significant step towards experimentation with more advanced nonlinear estimators of concentration, both from algorithmic and methodological point of view.

ACKNOWLEDGEMENTS

The study presented in this paper has been supported by the Ministry of Science and Higher Education in Poland (grant No. N505 011 31/1428). 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 in the paper for numerical experimentation.

REFERENCES

1. Tay A., Singh R. K., Krishnan S. S., Gore J. P.: "Authentication of Olive Oil Adulterated with Vegetable Oils Using Fourier Transform Infrared Spectroscopy". *Lebensm.-Wiss. u.-Technol.*, vol. 35, 2002, pp. 99–103.
2. Christy A., Kasemusumran S., Du Y., Ozaki Y.: "The Detection and Quantification of Adultaration in Olive oil by Near-Infrared Spectroscopy and Chemometrics". *Anal. Sci.*, vol. 20, June 2004, pp. 935–940.
3. Armenta S., Garrigues S., de la Guardia M.: "Determination of Edible Oil Parameters by Near Infrared Spectrometry". *Anal. Chim. Acta*, vol. 596, 2007, pp. 330-337.
4. Özdemir D., Öztürk B.: "Near Infrared Spectroscopic Determination of Olive Oil Adulteration with Sunflower and Corn Oil". *J. Food & Drug Anal.*, vol. 15, no. 1, 2007, pp. 40–47.
5. Sinelli N., Casiraghi E., Tura D., Downey G.: "Characterization and Classification of Italian Virgin Olive Oils by Near and Medium Infrared Spectroscopy". *Proc. 13th Int. Conf. Near Infrared Spectroscopy (Umeå-Vasa, Sweden & Finland, June 15–21, 2007)*, paper 4–6.
6. Latała A., Morawski R. Z.: "A comparative study of nine methods of estimation when applied for determination of olive oil mixtures on the basis of NIR spectrophotometric data". *Proc. IMEKO TC19-TC23-TC24 Int. Conf. on Metrology of Environmental, Food and Nutritional Measurements (Budapest, Hungary, September 10–12, 2008)*, CD-ROM.
7. Morawski R. Z.: "Spectrophotometric Applications of Digital Signal Processing". *Measurement Science and Technology*, vol. 17, August 2006, pp. R117–R144.