www.czasopisma.pan.pl PAN POLSKA AKADEMIA NAUK www.journals.pan.pl

Metrol. Meas. Syst., Vol. XXI (2014), No. 2, pp. 191–204.

PAN POLSKA AKADEMIA NAUK

METROLOGY AND MEASUREMENT SYSTEMS

Index 330930, ISSN 0860-8229

www.metrology.pg.gda.pl

M&MS

COMPARISON OF FIVE SVD-BASED ALGORITHMS FOR CALIBRATION OF SPECTROPHOTOMETRIC ANALYZERS

Jakub Wagner, Roman Z. Morawski, Andrzej Miękina

Warsaw University of Technology, Faculty of Electronics and Information Technology, Institute of Radioelectronics, Nowowiejska 15/19, 00-665 Warsaw, Poland (✉ r.morawski@ire.pw.edu.pl, +48 22 234 7721)

Abstract

Spectrophotometry is an analytical technique of increasing importance for the food industry, applied *i.a.* in the quantitative assessment of the composition of mixtures. Since the absorbance data acquired by means of a spectrophotometer are highly correlated, the problem of calibration of a spectrophotometric analyzer is, as a rule, numerically ill-conditioned, and advanced data-processing methods must be frequently applied to attain an acceptable level of measurement uncertainty. This paper contains a description of four algorithms for calibration of spectrophotometric analyzers, based on the singular value decomposition (SVD) of matrices, as well as the results of their comparison – in terms of measurement uncertainty and computational complexity – with a reference algorithm based on the estimator of ordinary least squares. The comparison is carried out using an extensive collection of semi-synthetic data representative of trinary mixtures of edible oils. The results of that comparison show the superiority of an algorithm of calibration based on the truncated SVD combined with a signal-to-noise ratio used as a criterion for the selection of regularisation parameters – with respect to other SVD-based algorithms of calibration.

Keywords: spectrophotometry, chemometrics, singular value decomposition, regularisation, food analysis, edible oils analysis.

1. Introduction

Spectrophotometry is an analytical technique based on reflection or transmission of light – or more precisely: ultraviolet (UV), visible (Vis) or infrared (IR) radiation – by analyzed substances [1]. Due to a significant technological progress in the domain of spectrophotometric instrumentation, achieved during the last few decades – in particular, the development of mini- and micro-spectrophotometers – its importance is increasing in numerous branches of industry, especially in food industry – see, for example, [2–28]. Further progress in this area is critically conditioned by the development of software dedicated to spectrophotometry, in particular – to calibration of spectrophotometric analyzers. In the 2013 paper [29], the authors compared forty algorithms for calibration, here four more are included in the comparison following the same methodology.

1.1. Basic concepts of spectrophotometric analysis

The intensity spectrum of light is usually understood as a function modeling the dependence of light intensity on wavelength. It is an adequate characteristic of a source of light, but insufficient for characterizing absorption of light by analyzed substances; the concept of transmittance spectrum is more appropriate in this case. The latter is usually understood as a function representative of the ratio of two intensity spectra, *viz.* the intensity spectrum of light entering into a sample of the analyzed substance and the intensity spectrum

Article history: received on Oct. 14, 2013; accepted on Jan. 29, 2014; available online on May. 15, 2014; DOI: 10.2478/mms-2014-0017.

of light leaving it. The decimal logarithm of the first of them divided by the second one is called absorbance spectrum; as a rule, it has the form of a sequence of Gaussoid-like peaks, whose magnitudes and locations on the wavelength axis carry information on the chemical contents of the sample. Due to technical imperfections of spectrophotometric devices and instruments – in particular, their limited optical and digital resolution – the measurement data provided by them represent spectra in an approximate way. As a rule, those data are abundant and highly correlated. Therefore, the increased interest in spectrophotometric analysis in food industry is an important driving force for the development of advanced methods for data processing. Such methods can be oriented on the qualitative analysis of food (*e.g.* classification of samples according to their geographical or temporal origin, based on pattern recognition methods), as well as on the quantitative analysis of food (*e.g.* estimation of concentrations of the compounds in a sample, based on regression-type methods). A spectrophotometric device or instrument, designed for quantitative analysis of a pre-defined class of substances, will be called hereinafter a spectrophotometric analyzer, or briefly – the analyzer.

Any spectrophotometric analyzer is composed of two principal parts: a block of data acquisition and a block of data processing. The first of them usually consists of a broadband light source, a sample holder, a dispersive element, an array of photodetectors, and an analog-to-digital converter; the second – a digital signal processor or a personal computer. The estimation of the measurand on the basis of raw spectrophotometric data, performed by the latter block, is based on some assumptions concerning the mathematical model of the first block – the model relating the data to the measurand (forward model) or *vice versa* (inverse model). That model is identified during the calibration of a spectrophotometric analyzer, using a set of reference data, *viz.* the data representative of some samples for which the values of the measurand are known. If the absorbance spectra of samples used for calibration are similar, the problem is numerically ill-conditioned, and quite sophisticated methods of estimation are necessary for attaining an acceptable level of measurement uncertainty (*cf.*, for example, [30]).

The *forward-model-based* approach of calibration consists in identification of the operator M modelling the dependence of the data $\tilde{\mathbf{s}}$, representative of the absorbance spectrum, on the measurand $\mathbf{c} \equiv [c_1 \ldots c_J]^{\mathrm{T}}$:

$$\tilde{\mathbf{s}} \cong \mathsf{M}\left(\mathbf{c}; \mathbf{p}_{\mathsf{M}}\right), \tag{1}$$

where $\mathbf{p}_{\mathsf{M}}$ is the vector of parameters to be estimated during calibration. The *inverse-model-based* approach consists in identification of the operator R modelling the dependence of the measurand $\mathbf{c}$ on the data $\tilde{\mathbf{s}}$ representative of the absorbance spectrum:

$$\mathbf{c} \cong \mathsf{R}\left(\tilde{\mathbf{s}}; \mathbf{p}_{\mathsf{R}}\right), \tag{2}$$

where $\mathbf{p}_{\mathsf{R}}$ is the vector of parameters to be estimated during calibration. Both approaches are illustrated in Fig. 1, where $\hat{\mathbf{c}}$ is an estimate of $\dot{\mathbf{c}}$, being an exact value of $\mathbf{c}$; $\hat{\mathbf{p}}_{\mathsf{M}}$ is an estimate of the value of $\mathbf{p}_{\mathsf{M}}$, resulting from the forward-model-based calibration; and $\hat{\mathbf{p}}_{\mathsf{R}}$ is an estimate of the value of $\mathbf{p}_{\mathsf{R}}$, resulting from the inverse-model-based calibration.

Metrol. Meas. Syst., Vol. XXI (2014), No. 2, pp. 191–204.

1.2. Research assumptions and objectives

The study reported in this paper has been oriented on the comparison of several methods of calibration following the inverse-model-based approach. It has been assumed that the relationship $\mathbf{c} \leftrightarrow \tilde{\mathbf{s}}$ may be adequately modeled using a linear operator M or R. This assumption is justified in many practically important applications, in particular – in many cases where the analysis is aimed at estimation of the concentrations of the components of mixtures. The forward model has then the form:

$$\tilde{\mathbf{s}} \cong \sum_{j=1}^{J} c_j \dot{\mathbf{s}}_j \,. \tag{3}$$

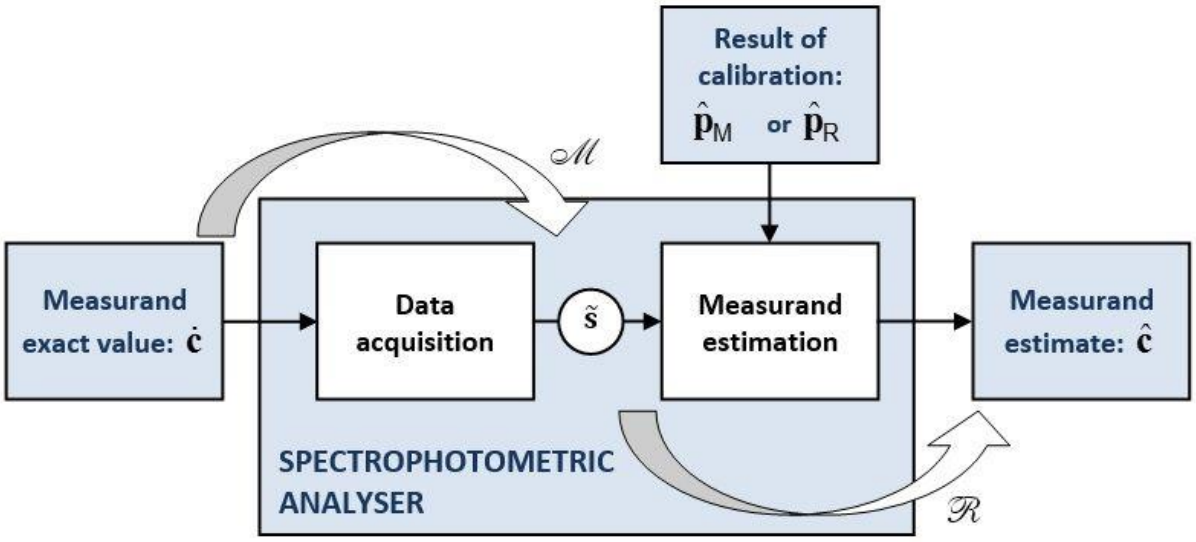

Fig. 1. Two approaches of calibration.

where $\dot{\mathbf{s}}_j$ are N-dimensional vectors of data representative of the absorbance spectra of the components – the denoised data acquired by means of the analyzer to be calibrated. The inverse model can be in this case assumed to be linear as well:

$$c_j \cong \mathbf{p}_j^{\mathrm{T}} \tilde{\mathbf{s}} \quad \text{for} \quad j = 1, ..., J \,. \tag{4}$$

For the sake of simplicity, a single equation of the above form will be considered hereinafter, *viz.* the one corresponding to a fixed value of j; therefore, this index will be omitted in the next section where the compared algorithms are described. The calibration may be now redefined as a procedure aimed at estimation of the vector of parameters $\mathbf{p}_R \equiv \mathbf{p}$ on the basis of a set of calibration (reference) data:

$$\left\{ \left\langle \dot{c}_m^{cal}, \tilde{\mathbf{s}}_m^{cal} \right\rangle \,\middle|\, m = 1, ..., M \right\}, \tag{5}$$

where each pair $\left\langle \dot{c}_m^{cal}, \tilde{\mathbf{s}}_m^{cal} \right\rangle$ contains a value of concentration and the corresponding spectral data acquired by means of the analyzer to be calibrated for M calibration samples, *i.e.* reference mixtures. For the sake of simplicity, the reference values of the concentration

$\dot{c}_m^{cal}$ will be organized in the M-dimensional vector $\dot{\mathbf{c}}$, and the reference spectral data in the $M \times N$-dimensional matrix $\tilde{\mathbf{S}}$ whose each column corresponds to a single wavelength value, and each row – to a single mixture. The objective of calibration is, therefore, the solution of the system of linear algebraic approximate equations:

$$\dot{\mathbf{c}} \cong \tilde{\mathbf{S}}\,\mathbf{p}, \tag{6}$$

with respect to $\mathbf{p}$. Since the number of samples used for calibration is usually significantly smaller than the number of data in a single absorbance spectrum, $M < N$, the above system of equations is, as a rule, underdetermined, and has infinite, apparently equivalent, solutions. However, the estimates of the concentration, determined for an unknown sample using different vectors $\mathbf{p}$ satisfying Eq.(6), may differ significantly in terms of accuracy. Thus, an important element of the procedure of calibration is the choice of a vector $\mathbf{p}$ resulting in the best accuracy of measurand estimation. It should be noted that the values of the measurand in the calibration data are subject to errors, introduced in the procedure of preparation of the samples. Therefore, the vector $\dot{\mathbf{c}}$ contains, in fact, assumed values of the measurand, slightly different from the real ones.

2. Compared algorithms of calibration

The compared algorithms of calibration are structurally similar, and differ only in the methods used for estimation of the parameters $\mathbf{p}$; therefore, the names of those methods are used here as identifiers of the compared algorithms of calibration.

2.1. Ordinary Least Squares

The algorithm of calibration labelled with the acronym OLS refers to the estimator of ordinary least squares which consists in solving the following problem of unconstrained optimisation:

$$\hat{\mathbf{p}}_{OLS} \equiv \arg_{\mathbf{p}} \inf \left\{ \left\| \tilde{\mathbf{S}}\,\mathbf{p} - \dot{\mathbf{c}} \right\|_2 \right\}. \tag{7}$$

For $M < N$, when it has more than one solution, the minimum-norm solution is selected:

$$\hat{\mathbf{p}}_{OLS} \equiv \arg_{\mathbf{p}} \inf \left\{ \left\| \mathbf{p} \right\|_2 \,\middle|\, \left\| \tilde{\mathbf{S}}\,\mathbf{p} - \dot{\mathbf{c}} \right\|_2 = 0 \right\}. \tag{8}$$

This solution may be calculated using the Moore–Penrose pseudoinverse of the matrix $\tilde{\mathbf{S}}$ [31]:

$$\hat{\mathbf{p}}_{OLS} = \tilde{\mathbf{S}}^{+}\dot{\mathbf{c}} = \tilde{\mathbf{V}}\,\tilde{\mathbf{\Sigma}}^{+}\tilde{\mathbf{U}}^{\mathrm{T}}\,\dot{\mathbf{c}}, \tag{9}$$

where the matrices $\tilde{\mathbf{V}}$, $\tilde{\mathbf{\Sigma}}$ and $\tilde{\mathbf{U}}$ are components of the singular value decomposition of the matrix $\tilde{\mathbf{S}}$, *i.e.*:

$$\tilde{\mathbf{S}}\big|_{M\times N} = \tilde{\mathbf{U}}\big|_{M\times M}\,\tilde{\mathbf{\Sigma}}\big|_{M\times N}\,\tilde{\mathbf{V}}^{\mathrm{T}}\big|_{N\times N}, \tag{10}$$

Metrol. Meas. Syst., Vol. XXI (2014), No. 2, pp. 191–204.

where:

$$\tilde{\mathbf{U}}^{\mathrm{T}}\tilde{\mathbf{U}} = \mathbf{I}, \ \tilde{\mathbf{V}}^{\mathrm{T}}\tilde{\mathbf{V}} = \mathbf{I}, \tag{11}$$

$$\tilde{\mathbf{\Sigma}} = \left[\mathrm{diag}\{\tilde{\sigma}_1, \ldots, \tilde{\sigma}_M\} \ \mathbf{0}|_{M\times(N-M)}\right] \ \text{with} \ \ \tilde{\sigma}_1 \geq \ldots \geq \tilde{\sigma}_M > 0, \tag{12}$$

and:

$$\tilde{\mathbf{\Sigma}}^{+} = \begin{bmatrix} \mathrm{diag}\left\{\dfrac{1}{\tilde{\sigma}_1}, \ldots, \dfrac{1}{\tilde{\sigma}_M}\right\} \\ \mathbf{0}|_{(N-M)\times M} \end{bmatrix}. \tag{13}$$

It follows from the above equation that the errors present in the data matrix $\tilde{\mathbf{S}}$, propagated to the singular values $\tilde{\sigma}_m$ ($m = 1, \ldots, M$), may be significantly amplified during the calculation of the pseudoinverse $\tilde{\mathbf{\Sigma}}^{+}$, especially for the smallest singular values. This negative phenomenon may be reduced by various methods of regularization (reduction of random errors at the cost of slight increase of systematic errors) described in the following subsections.

2.2. Truncated Singular Value Decomposition

The algorithm of calibration labelled with the acronym TSVD-SN refers to the estimator of the truncated singular-value decomposition which consists in zeroing a selected number of the smallest singular values: $\tilde{\sigma}_{M'+1} = \ldots = \tilde{\sigma}_M = 0$ ($M' < M$). The TSVD-SN solution to Eq.(6) takes on the form:

$$\hat{\mathbf{p}}_{TSVD}(M') \equiv \tilde{\mathbf{V}}\,\tilde{\mathbf{\Sigma}}^{+}_{TSVD}(M')\,\tilde{\mathbf{U}}^{\mathrm{T}}\,\dot{\mathbf{c}}, \tag{14}$$

where M' is a regularization parameter whose value has to be determined in a way as to reach a good trade-off between the reduction of the error amplification and the loss of information corresponding to the smallest singular values. In the reported study, this value has been selected using a criterion related to the signal-to-noise ratio, proposed by the authors in [32]. For each singular value $\tilde{\sigma}_m$ of the matrix $\tilde{\mathbf{S}}$, a ratio:

$$\tilde{v}_m \equiv \frac{\tilde{\sigma}_m^2}{M} \ \ \text{for} \ \ m = 1, \ldots, M, \tag{15}$$

has been calculated to characterize the "amount of variance" in the data to be lost if the singular value $\tilde{\sigma}_m$ is set to 0. The value of M' has been selected using the reverse accumulated variances:

$$\tilde{a}_m \equiv \sum_{\mu=m}^{M} \tilde{v}_\mu \ \ \text{for} \ \ m = 1, \ldots, M, \tag{16}$$

as the smallest integer m satisfying the following inequality:

$$\tilde{a}_m < \frac{\tilde{a}_1}{SNR}, \tag{17}$$

where SNR is the signal-to-noise ratio defined by the formula:

$$SNR = \frac{\sum_{n=1}^{N}\sum_{m=1}^{M}\left(\tilde{s}_{m,n}\right)^2}{\sum_{n=1}^{N}\sum_{m=1}^{M}\mathrm{Var}\left[\tilde{s}_{m,n}\right]}, \tag{18}$$

with $\mathrm{Var}\left[\tilde{s}_{m,n}\right]$ being the variances of $\tilde{s}_{m,n}$, derived from the *a priori* knowledge about the random errors corrupting the spectral data used for calibration.

2.3. Ridge Least Squares with Discrepancy Principle

The algorithms of calibration labelled with the acronym RiLS-DP and RiLS-GCV refer to the estimator of ridge least squares, defined by the following formula:

$$\hat{\mathbf{p}}_{RiLS}(\alpha) \equiv \arg_{\mathbf{p}} \inf \left\{ \left\| \tilde{\mathbf{S}}\,\mathbf{p} - \dot{\mathbf{c}} \right\|_2^2 + \alpha^2 \left\| \mathbf{p} \right\|_2^2 \right\}, \tag{19}$$

where α is a real-valued parameter of regularization. It can be expressed using the SVD in the following way [33]:

$$\hat{\mathbf{p}}_{RiLS}(\alpha) = \left(\tilde{\mathbf{V}}\, \tilde{\mathbf{\Sigma}}^{+}_{RiLS}(\alpha)\, \tilde{\mathbf{U}}^{\mathrm{T}} \right) \dot{\mathbf{c}}, \tag{20}$$

where $\tilde{\mathbf{\Sigma}}^{+}_{RiLS}(\alpha)$ is obtained from $\tilde{\mathbf{\Sigma}}^{+}_{OLS}$ by replacing each $1/\tilde{\sigma}_m$ with $\tilde{\sigma}_m / \left(\tilde{\sigma}_m^2 + \alpha^2 \right)$. If *a priori* knowledge about the errors corrupting the calibration data is available, then the value of the regularization parameter α may be selected using the discrepancy principle [34], [35] usually formulated as the following problem of constrained optimisation:

$$\alpha_{DP} \equiv \arg_{\alpha} \inf \left\{ \left\| \hat{\mathbf{p}}_{RiLS}(\alpha) \right\|_2 \,\middle|\, \left\| \dot{\mathbf{c}} - \tilde{\mathbf{S}}\, \hat{\mathbf{p}}_{RiLS}(\alpha) \right\|_2^2 \cong \Delta_c^2 \right\}, \tag{21}$$

where Δ_c is an empirical measure of total disturbances due to both errors in the data and non-adequacy of the mathematical model underlying the method of calibration. There are various methodologies for fixing the value of Δ_c. In the SVD-based algorithms of calibration considered here, a statistical approach – developed by the authors in [36, Chapter 8] – has been used:

$$\Delta_c^2 = \mathrm{E}\left[\left\| \tilde{\mathbf{c}} - \tilde{\mathbf{S}}\, \hat{\mathbf{p}}_{RiLS} \right\|_2^2 \right], \tag{22}$$

Metrol. Meas. Syst., Vol. XXI (2014), No. 2, pp. 191–204.

where E[◦] denotes the operator of expected value. The parameter Δ_c^2 has been evaluated under an assumption that some estimates of the first two moments of the random errors corrupting the data $\tilde{\mathbf{c}}$ and $\tilde{\mathbf{S}}$ are available.

2.4. Ridge Least Squares with Generalised Cross Validation

The algorithm of calibration labelled with the acronym RiLS-GCV refers to the estimator of ridge least squares combined with the generalized cross-validation strategy used for selection of the value of the parameter α. The latter consists in solving the following problem of unconstrained optimisation [37]:

$$\alpha_{GCV} \equiv \arg_{\alpha} \inf \left\{ \frac{\left\| \dot{\mathbf{c}} - \hat{\mathbf{c}}(\alpha) \right\|_2^2}{\left(M - \mathrm{tr}(\mathbf{H}(\alpha)) \right)^2} \right\}, \tag{23}$$

where:

$$\hat{\mathbf{c}}(\alpha) \equiv \tilde{\mathbf{S}}\, \hat{\mathbf{p}}_{RiLS}(\alpha), \tag{24}$$

and $\mathbf{H}(\alpha)$ is the so-called hat matrix defined by the condition:

$$\hat{\mathbf{c}}(\alpha) = \mathbf{H}(\alpha)\dot{\mathbf{c}}. \tag{25}$$

Since

$$\hat{\mathbf{c}}(\alpha) = \left(\tilde{\mathbf{U}}\, \tilde{\mathbf{\Sigma}}\, \tilde{\mathbf{V}}^{\mathrm{T}}\right)\left(\tilde{\mathbf{V}}\, \tilde{\mathbf{\Sigma}}^{+}_{RiLS}(\alpha)\, \tilde{\mathbf{U}}^{\mathrm{T}}\right)\dot{\mathbf{c}} = \tilde{\mathbf{U}}\, \tilde{\mathbf{\Sigma}}\, \tilde{\mathbf{\Sigma}}^{+}_{RiLS}(\alpha)\, \tilde{\mathbf{U}}^{\mathrm{T}}\dot{\mathbf{c}}, \tag{26}$$

the matrix $\mathbf{H}(\alpha)$ may be given the form:

$$\mathbf{H}(\alpha) = \tilde{\mathbf{U}}\, \tilde{\mathbf{\Sigma}}\, \tilde{\mathbf{\Sigma}}^{+}_{RiLS}(\alpha)\tilde{\mathbf{U}}^{\mathrm{T}}. \tag{27}$$

If the optimization problem defined by Eq.(23) has multiple solutions, the largest value of α is selected to avoid overfitting.

2.5. Multi-parameter Ridge Least Squares

The algorithm of calibration labelled with the acronym MRiLS-GCV refers to the estimator of ridge least squares with multiple parameters of regularization, defined by the following formula [38]:

$$\hat{\mathbf{p}}_{MRiLS}(\boldsymbol{\alpha}) = \tilde{\mathbf{V}}\, \tilde{\mathbf{\Sigma}}^{+}_{MRiLS}(\boldsymbol{\alpha})\tilde{\mathbf{U}}^{\mathrm{T}}\, \dot{\mathbf{c}}, \tag{28}$$

where $\boldsymbol{\alpha} \equiv [\alpha_1 \ldots \alpha_M]^{\mathrm{T}}$, and $\tilde{\mathbf{\Sigma}}^{+}_{MRiLS}(\boldsymbol{\alpha})$ is obtained from $\tilde{\mathbf{\Sigma}}^{+}_{OLS}$ by replacing $1/\tilde{\sigma}_m$ with $1/\left(\tilde{\sigma}_m + \alpha_m\right)$ for $m = 1, \ldots, M$; $\alpha_m > 0$ for $m = 1, \ldots, M$. The values of the parameters α_m are chosen using the generalized cross-validation strategy, *i.e.* by solving the following problem of unconstrained optimization:

$$\alpha_{GCV} \equiv \arg_{\boldsymbol{\alpha}} \inf \left\{ \frac{\left\| \dot{\mathbf{c}} - \hat{\mathbf{c}}(\boldsymbol{\alpha}) \right\|_2^2}{\left(M - \mathrm{tr}(\mathbf{H}(\boldsymbol{\alpha})) \right)^2} \right\}, \tag{29}$$

by means of an evolutionary procedure of global optimization.

3. Methodology of comparison

The algorithms of calibration, described in Section 2, have been compared using the same methodology as described in two papers already published by the authors: [29] and [32]. It is based on the use of semi synthetic data representative of the NIR absorbance spectra of trinary mixtures of edible oils (*cf.* Subsection 3.1) and three statistical criteria characterizing measurement uncertainty (*cf.* Subsection 3.2).

3.1. Generation of semi-synthetic data

Semi-synthetic data used for experimentation have been generated using the denoised absorbance data representative of the absorbance spectra of nut oil, corn oil and olive oil (Fig. 2), and pseudorandom numbers for simulation of measurement errors. The data were acquired in the near-infrared range of radiation, *i.e.* for the wavelength values from 1500 nm to 2500 nm, at the 2-nm intervals; so, the number of data representative of a single spectrum has been $N = 501$. Since the concentrations of the components satisfy the equation:

$$c_1 + c_2 + c_3 = 1, \tag{30}$$

only two variables are independent, *e.g.*: c_1 – the concentration of nut oil, and c_2 – the concentration of corn oil. The procedure of data synthesis has comprised four steps: generation of pairs of exact values of concentrations $\langle \dot{c}_1, \dot{c}_2 \rangle$ for each hypothetical mixture, introduction of random errors in the concentration data, generation of absorbance data based on the error-corrupted concentration data, and introduction of the errors modeling imperfections of the spectrophotometer.

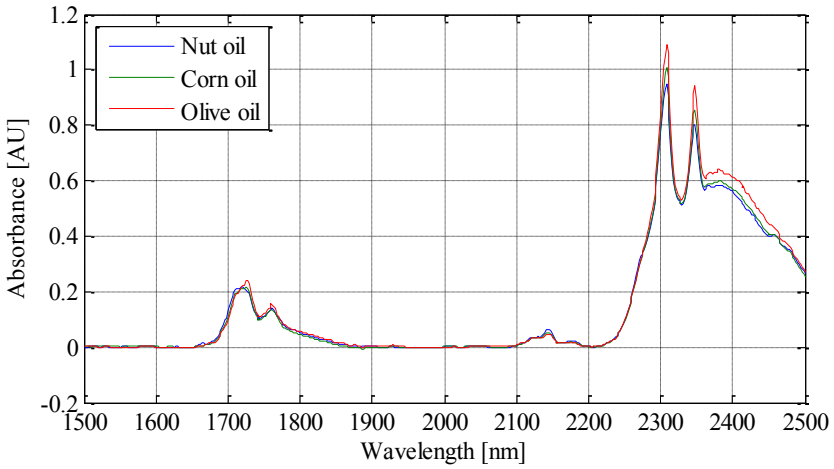

Fig. 2. Data representative of the absorbance spectra of three edible oils.

Metrol. Meas. Syst., Vol. XXI (2014), No. 2, pp. 191–204.

The exact concentration of olive oil has been calculated for each mixture according to the formula:

$$\dot{c}_3 = 1 - \dot{c}_1 - \dot{c}_2 . \tag{31}$$

Each vector of spectral data has been synthesized according to the methodology imitating the laboratory procedure applied for obtaining the real-world data; in particular, uncertainty resulting from imperfect measuring out of the volumes of oils has been taken into account. Those volumes are related to the concentrations in the following way:

$$\dot{c}_j = \frac{\dot{V}_j}{\dot{V}} \quad \text{for} \quad j = 1, 2, 3, \tag{32}$$

where $\dot{V} = \dot{V}_1 + \dot{V}_2 + \dot{V}_3$. The error-corrupted values of the volumes have been obtained using the equation:

$$\tilde{V}_j = \dot{V}_j \left(1 + \delta\tilde{V}_j\right) \quad \text{for} \quad j = 1, 2, 3, \tag{33}$$

where $\delta\tilde{V}_j$ are pseudorandom numbers modeling relative errors of the volumes and following the zero-mean normal distribution with the standard deviation $\sigma_V = 2.0 \cdot 10^{-3}$, truncated outside of the interval $\left[-3\sigma_V, +3\sigma_V\right]$. The corresponding error-corrupted values of the concentration data have been generated according to the formula:

$$\tilde{c}_j \equiv \frac{\tilde{V}_j}{\tilde{V}_1 + \tilde{V}_2 + \tilde{V}_3} = \frac{\dot{c}_j \left(1 + \delta\tilde{V}_j\right)}{\dot{c}_1 \left(1 + \delta\tilde{V}_1\right) + \dot{c}_2 \left(1 + \delta\tilde{V}_2\right) + \dot{c}_3 \left(1 + \delta\tilde{V}_3\right)}, \tag{34}$$

and the absorbance data – according to Eq.(3). The errors caused by imperfections of the analyzer have been taken into account using the formula:

$$\tilde{\mathbf{s}} = \left[\tilde{s}_1 \ldots \tilde{s}_N\right]^{\mathrm{T}} = \left[\tilde{s}'_1 + \Delta\tilde{s}_1 \ldots \tilde{s}'_N + \Delta\tilde{s}_N\right]^{\mathrm{T}}, \tag{35}$$

where $\Delta\tilde{s}_n$ are independent pseudorandom numbers modeling absolute errors and following the zero-mean normal distribution with the standard deviation σ_s, truncated outside of the $\left[-3\sigma_s, +3\sigma_s\right]$ interval.

The truncation of the range of pseudorandom numbers, used for simulation of errors, has been done by replacing with -3σ the numbers smaller than -3σ and with $+3\sigma$ the numbers greater than $+3\sigma$ (where σ denotes the assumed value of the standard deviation).

3.2. Criteria of comparison

The performance of the compared algorithms of calibration has been assessed using a calibration data set, generated according to the methodology described in Subsection 3.1 for:

$$\dot{c}_1^{cal}, \dot{c}_2^{cal} \in \left\{ k \cdot 0.01 \,\middle|\, k = 0, 1, \ldots, 10 \right\}, \tag{36}$$

and a validation data set, generated according to the same methodology for:

$$\dot{c}_1^{val}, \dot{c}_2^{val} \in \left\{ k \cdot 0.005 \,\middle|\, k = 0, 1, \ldots, 20 \right\}, \tag{37}$$

which means that $M^{cal} = 121$ mixtures have been used for calibration and $M^{val} = 441$ for validation. The concentration values in the validation data set have not been corrupted with errors; they have been assumed to be known exactly since such errors would influence the results obtained by compared algorithms in the same way, thus – not change the result of comparison.

Each set of calibration data and each set of validation data have been generated in $R = 100$ versions corresponding to different realisations of pseudorandom numbers modelling the errors. The calibration has been performed with the same algorithm for each of the R calibration data sets, and each calibration result has been tested using each of the R validation data sets, resulting in $R^2 = 10^4$ independent numerical experiments. The vector of absolute errors of estimation, performed during validation, has been determined according to the formula:

$$\Delta\hat{\mathbf{c}}_j\left(r^{cal}, r^{val}\right) = \hat{\mathbf{c}}_j^{val}\left(r^{cal}, r^{val}\right) - \dot{\mathbf{c}}_j^{val} = \tilde{\mathbf{S}}^{val}\left(r^{val}\right)\hat{\mathbf{p}}_j\left(r^{cal}\right) - \dot{\mathbf{c}}_j^{val},$$
$$\text{for} \quad j = 1, 2;\ r^{cal}, r^{val} = 1, \ldots, R. \tag{38}$$

For each element of this vector, corresponding to one mixture used for validation, the following indicators of uncertainty have been calculated:

– the mean of the estimation errors:

$$\hat{m}\left[\Delta\hat{c}_{j,m}\right] \equiv \frac{1}{R^2} \sum_{r^{cal}=1}^{R} \sum_{r^{val}=1}^{R} \Delta\hat{c}_{j,m}\left(r^{cal}, r^{val}\right), \tag{39}$$

– the standard deviation of the estimation errors:

$$\hat{s}\left[\Delta\hat{c}_{j,m}\right] \equiv \sqrt{\frac{1}{R^2 - 1} \sum_{r^{cal}=1}^{R} \sum_{r^{val}=1}^{R} \left[\Delta\hat{c}_{j,m}\left(r^{cal}, r^{val}\right) - \hat{m}\left[\Delta\hat{c}_{j,m}\right]\right]^2}, \tag{40}$$

– the worst-case estimation error:

$$\hat{e}\left[\Delta\hat{c}_{j,m}\right] \equiv \sup\left\{ \left| \Delta\hat{c}_{j,m}\left(r^{cal}, r^{val}\right) \right| \, r^{cal}, r^{val} = 1, \ldots, R \right\}, \tag{41}$$

4. Selected results of comparison

The full program of the comparative study has included:

– several levels of errors corrupting the spectrophotometric data;
– three indicators of uncertainty ($\hat{m}\left[\Delta\hat{c}_{j,m}\right]$, $\hat{s}\left[\Delta\hat{c}_{j,m}\right]$ and $\hat{e}\left[\Delta\hat{c}_{j,m}\right]$) evaluated for both $\dot{c}_1$ and $\dot{c}_2$;
– several versions of the strategies for selection of regularization parameters.

Metrol. Meas. Syst., Vol. XXI (2014), No. 2, pp. 191–204.

The results of study, accomplished according to this programme, have been systematically presented in a Master's Thesis [39]. Here, due to the space limitation, only selected results are shown, *viz.* the values of an indicator R_{OLS}^{SVD} being the ratio of the worst-case error of estimation of c_1 obtained for the compared SVD-based algorithm and the reference OLS-based algorithm, computed for the lowest ($\sigma_s = 10^{-6}$) and the highest ($\sigma_s = 10^{-3}$) level of errors, are presented in Fig. 3. In Fig. 4 the dependence of this indicator on σ_s is depicted. The average time of execution on a reference computer has turned out to be *ca.* 2 times greater for the algorithm TSVD-SN than for the algorithm OLS; the corresponding values of this indicator of computational complexity for other algorithms – as follows: *ca.* 7 for RiLS-GCV, *ca.* 150 for RiLS-DP and *ca.* 1300 for MRiLS-GCV.

5. Conclusions

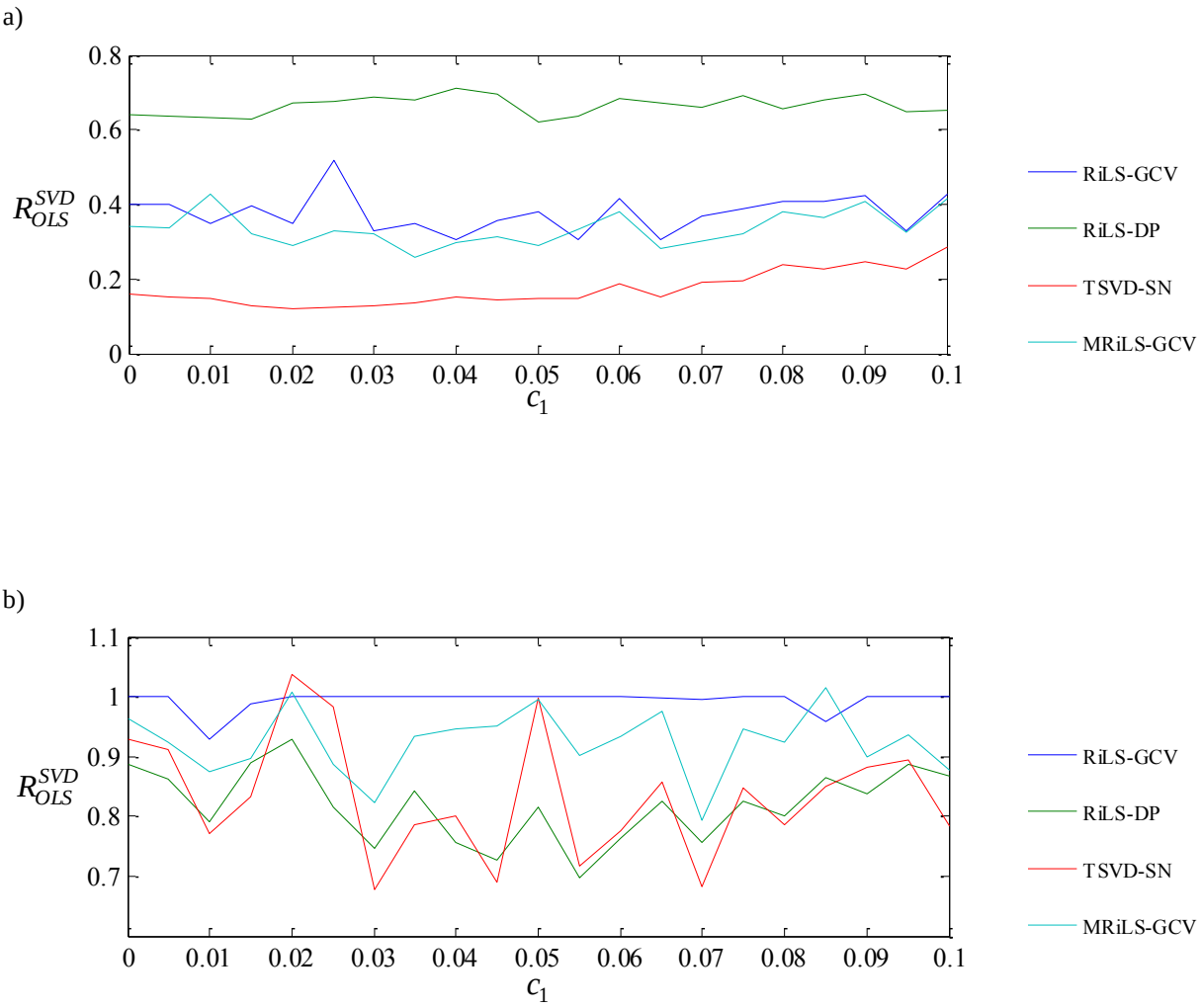

Fig. 3. Dependence of the indicator R_{OLS}^{SVD} on c_1 for the extreme values of the variance of instrumental errors σ_s: (a) $\sigma_s = 10^{-6}$, (b) $\sigma_s = 10^{-3}$.

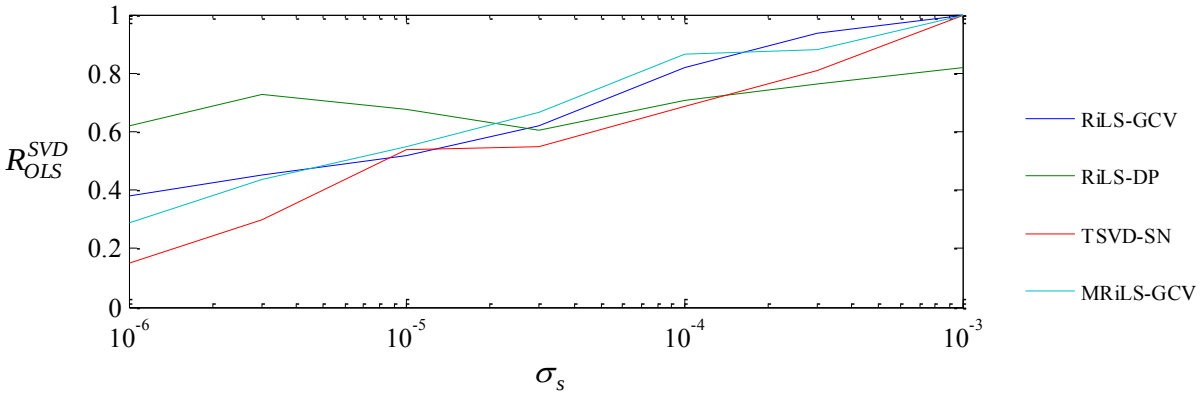

Fig. 4. Dependence of the indicator R_{OLS}^{SVD} on the variance of instrumental errors for $c_1 = c_2 = 0.05$.

Five algorithms for calibration of spectrophotometric analyzers, based on the singular value decomposition of matrices, have been compared using an example of the analysis of trinary mixtures of edible oils. The obtained results support the following conclusions:

– The algorithm TSVD-SN seems to be the most promising and worth further investigation. It makes possible a significant reduction of estimation uncertainty, if compared to the algorithm OLS, at the cost of a very moderate increase in computational complexity. The superiority of the algorithm TSVD-SN is particularly pronounced if the instrumental errors corrupting the absorbance data are significantly less important than the errors corrupting the concentration data.
– All tested strategies for selection of regularization parameters make possible the automatic adjustment of the algorithms to the level of errors in measurement data (without symptoms of under- or over-regularization), and therefore enable the use of those algorithms without the need of frequent visualisation and inspection of intermediate results.
– The SN-based strategy for selection of regularization parameters seems to be the most reliable since it is providing repeatable and accurate results for all the tested levels of errors in the data. The GCV-based strategy is providing better results than the DP-based strategy for lower levels of errors corrupting the absorbance data.

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] NIST Physical Measurement Laboratory: Spectrophotometry, *http://www.nist.gov/pml/div685/grp03/spectrophotometry.cfm* [2014.01.31].

[2] Balabin, R. M., Safieva, R. Z., Lomakina, E. I. (2010). Gasoline classification using near infrared (NIR) spectroscopy data: comparison of multivariate techniques, *Anal. Chim. Acta*, 671(1-2), 27–35.

[3] Brunt, K., Drost, W. C. (2010). Design, Construction, and Testing of an Automated NIR In-line Analysis System for Potatoes, *Potato Res.*, 53(1), 25–39 (Part I) and 41–60 (Part II).

[4] Cynkar, W., Dambergs, R., Smith, P., Cozzolino, D. (2010). Classification of Tempranillo wines according to geographic origin: Combination of mass spectrometry based electronic nose and chemometrics, *Anal. Chim. Acta*, (660), 227–231.

[5] Dupuy, N., Galtier, O., Le Dréau, Y., Pinatel, C., Kister, J., Artaud, J. (2010). Chemometric analysis of combined NIR and MIR spectra to characterize French olives, *Eur J. Lipid. Sci. Technol.*, 112(4), 463–475.

[6] Ferrer-Gallego, R., Hernández-Hierro, J. M., Rivas-Gonzalo, J. C., Escribano-Bailón, M. T. (2011). Determination of phenolic compounds of grape skins during ripening by NIR spectroscopy, *LWT – Food Sci. Technol.*, 44(4), 847–853.

[7] González-Martín, I., Hernández-Hierro, J. M., Salvador-Esteban, J., González-Pérez, C., Revillab, I., Vivar-Quintanab, A. (2011), Discrimination of seasonality in cheeses by near-infrared technology, *J. Sci. Food Agric.*, (91), 1064–1069.

[8] Guerrero, E. D., Mejías, R. C., Marín, R. N., Lovillo, M. P., Barroso, C. G. (2010). A new FT-IR method combined with multivariate analysis for the classification of vinegars from different raw materials and production processes, *J. Sci. Food Agric.*, (90), 712–718.

[9] Guy, F., Prache, S., Thomas, A., Bauchart, D., Andueza, D. (2011). Prediction of lamb meat fatty acid composition using near-infrared reflectance spectroscopy (NIRS), *Food Chem*, 127(3), 1280–1286.

[10] Hao Lin, Jiewen Zhao, Li Sun, Quansheng Chen, Fang Zhou (2011). Freshness measurement of eggs using near infrared (NIR) spectroscopy and multivariate data analysis, *Innov. Food. Sci. Emerg. Technol.*, 12(2), 182–186.

[11] Horikawa, Y., Imai, T., Takada, R., Watanabe, T., Takabe, K., Kobayashi, Y., Sugiyama, J. (2011). Near-infrared chemometric approach to exhaustive analysis of rice straw pretreated for bioethanol conversion, *Appl. Biochem. Biotechnol.*, 164(2), 194–203.

[12] Jin Hwan Lee, Myoung-Gun Choung (2011), Nondestructive determination of herbicide-resistant genetically modified soybean seeds using near-infrared reflectance spectroscopy, *Food Chem.*, 126(1), 368–373.

[13] Kapper, C., Klont, R. E., Verdonk, J. M. A. J., Urlings, H. A. P. (2012). Prediction of pork quality with near infrared spectroscopy (NIRS) – 1. Feasibility and robustness of NIRS measurements at laboratory scale, *Meat Sci.*, 91), 294–299 (Part I) and 300–305 (Part II).

[14] Lanzhen Chen, Xiaofeng Xue, Zhihua Ye, Jinghui Zhou, Fang Chen, Jing Zhao (2011). Determination of Chinese honey adulterated with high fructose corn syrup by near infrared spectroscopy, *Food Chem.*, 128(4), 1110–1114.

[15] Lerma-García, M. J., Ramis-Ramos, G., Herrero-Martínez, J. M., Simó-Alfonso, E. F. (2010). Authentication of extra virgin olive oils by Fourier-transform infrared spectroscopy, *Food Chem.*, 118(1), 78–83.

[16] Louw, E. D., Theron, K. I. (2010). Robust prediction models for quality parameters in Japanese plums using NIR spectroscopy, *Postharvest Biol. Technol.*, 58(3), 176–184.

[17] Magwaza, L. S., Opara, U. L., Nieuwoudt, H., Cronje, P. J. R., Saeys, W., Nicolaï, B. (2012). NIR Spectroscopy Applications for Internal and External Quality Analysis of Citrus Fruit – A Review, *Food Bioprocess Technol.*, (5), 425–444.

[18] Moghimi, A., Aghkhani, M. H., Sazgarnia, A., Sarmad, M. (2010). Vis/NIR spectroscopy and chemometrics for the prediction of soluble solids content and acidity (pH) of kiwifruit, *Biosystems Eng.*, 106(3), 295–302.

[19] Mutlu, A. C., Boyaci, I. H., Genis, H. E., Ozturk, R., Basaran-Akgul, N., Sanal, T., Evlice, A. K. (2011). Prediction of wheat quality parameters using near-infrared spectroscopy and artificial neural networks, *Eur. Food Res. Technol.*, 233(2), 267–274.

[20] Ntsame Affane, A. L., Fox, G. P., Sigge, G. O., Manley, M., Britz, T. J. (2011). Simultaneous prediction of acidity parameters (pH and titratable acidity) in Kefir using near infrared reflectance spectroscopy, *Int. Dairy J.*, 21(11), 896–900.

[21] Queji, M. D., Wosiacki, G., Cordeiro, G. A., Peralta-Zamora, P. G., Nagata, N. (2010). Determination of simple sugars, malic acid and total phenolic compounds in apple pomace by infrared spectroscopy and PLSR, *Int. J. Food Sci. Technol.*, 45), 602–609.

[22] Rohman, A., Man, Y. B. C. (2010). Fourier transform infrared (FTIR) spectroscopy for analysis of extra virgin olive oil adulterated with palm oil, *Food Res. Int.*, 43(3), 886–892.

[23] Sinelli, N., Pagani, M. A., Lucisano, M., D'Egidio, M. G., Mariotti, M. (2011). Prediction of semolina technological quality by FT-NIR spectroscopy, *J. Cereal Sci.*, 54(2), 218–223.

[24] Sundaram, J., Kandala, C. V. K., Butts, C. L. (2010). Classification of in-shell peanut kernels nondestructively using VIS/NIR reflectance spectroscopy, *Sens. Instrum. Food Qual. Saf.*, 4(2), 82–94.

[25] Torrecilla, J. S., Rojo, E., Domínguez, J. C., Rodríguez, F. (2010). A Novel Method To Quantify the Adulteration of Extra Virgin Olive Oil with Low-Grade Olive Oils by UV−Vis, *J. Agric. Food. Chem.*, 58(3), 1679–1684.

[26] Ulissi, V., Antonucci, F., Benincasa, P., Farneselli, M., Tosti, G., Guiducci, M., Tei, F., Costa, C., Pallottino, F., Pari, L., Menesatti, P. (2011). Nitrogen Concentration Estimation in Tomato Leaves by VIS-NIR Non-Destructive Spectroscopy, *Sensors*, 11(6), 6411–6424.

[27] Yanez, L., Saavedra, J., Martınez, C., Cordova, A., Ganga, M. A. (2012). Chemometric Analysis for the Detection of Biogenic Amines in Chilean *Cabernet Sauvignon* Wines: A Comparative Study between Organic and Nonorganic Production, *J. Food Sci.*, 77(8), T143–T150.

[28] Yi-Tao Liao, Yu-Xia Fan, Fang Cheng (2010). On-line prediction of fresh pork quality using visible/near-infrared reflectance spectroscopy, *Meat Sci.*, 86(4), 901–907.

[29] Morawski, R. Z., Miękina, A. (2012). A comparative study of forty algorithms for spectrophotometric analysis of edible oil mixtures, *Proc. XXth IMEKO World Congress (Busan, South Korea, September 9–14, 2012)*, 1–6.

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

[31] LAPACK Users' Guide, Linear Least Squares (LLS) Problems, *http://www.netlib.org/lapack/lug/node27.html* [2014.01.31].

[32] Morawski, R. Z., Miękina, A. (2013). PCA-based algorithm for calibration of spectrophotometric analysers of food, *J. Phys.: Conf. Ser.*, 459, *http://iopscience.iop.org/1742-6596/459/1* [2014.01.31].

[33] Kalivas, J. (1999). Interrelationships of multivariate regression methods using eigenvector basis sets, *J. Chemom.*, 13, 111–132.

[34] Schneider, T. (2013). Choice of regularization parameter, *www.gps.caltech.edu/~tapio/acm118/Handouts/regparameter.pdf* [2013.10.09].

[35] Lukas, M. A. (1998). Comparisons of parameter choice methods for regularization with discrete noisy data, *Inverse Prob.*, 14, 161–184.

[36] Morawski, R. Z., Miękina, A. (2013). *Methods and Algorithms for Processing Spectrophotometric Data, Volume 1 – Fundamentals and Applications in Food Analysis*, Research Report, Institute of Radioelectronics, Faculty of Electronics and Information Technology, Warsaw University of Technology.

[37] Yuedong Wang (2013). Cross-Validation and Generalized Cross-Validation, *http://sfb649.wiwi.hu-berlin.de/fedc_homepage/xplore/ebooks/html/csa/node123.html* [2013.10.09].

[38] Hines, J. W., Gribok, A. V., Urmanov, A. M., Buckner, M. A. (2002). Selection of Multiple Regularization Parameters in Local Ridge Regression Using Evolutionary Algorithms and Prediction Risk Optimization, *http://citeseer.ist.psu.edu/hines02selection.html* [2013.06.04].

[39] Wagner, J. (2013). *Algorithms for calibration of spectrophotometric analyser of edible oils mixtures, based on the singular value decomposition of matrices*, M.Sc. Thesis, Faculty of Electronics and Information Technology, Warsaw University of Technology.