Acta Agrophysica, 2005, 6(1), 7-18

ORTHOGONAL SIGNAL CORRECTION TO PLS MODELLING IN APPLICATION TO SPECTRAL DATA*

Grażyna Balcerowska[1], Ryszard Siuda[1], Henryk Czarnik-Matusewicz[2]

[1]Institute of Mathematics and Physics, University of Technology and Agriculture
ul. Kaliskiego 7, 85-796, Bydgoszcz
e-mail: gbalcer@mail.atr.bydgoszcz.pl
[2]Research Institute of Clinical Pharmacology, Faculty of Pharmacy, Medical Academy
ul. Bujwida 44, 50-345 Wrocław

Abstract. A typical task in chemometrics is to estimate the linear relationship between two sets of variables, i.e. the set of spectra, X, and the concentrations of some sample constituents, Y. Among the classical regression methods, partial least squares (PLS) is one of the most commonly used tools. One of the complications which could negatively affect the interpretation of the PLS model is related to the systematic variation present in X that is unrelated with the variation in Y. This situation typically occurs when X variables represent the absorbance or reflectance measured at hundreds of wavelengths, and the measurements are possibly influenced by sources of different types of variation having nothing in common with the information of interest. Orthogonal signal correction (OSC) is a recently proposed pre-processing method that seems to be promising in this context. This approach determines and removes from spectral data X the part of information which is Y-orthogonal (i.e. not correlated with Y). The purpose of the present paper is to illustrate how the technique works in application to near infrared (NIR) spectra of rapeseed meal. The results of PLS modelling for OSC pre-processed data have been compared with those of non-pre-processed as well as with those after multiplicative scatter correction (MSC). The main noticeable advantage of the OSC approach was the simplification of the calculated PLS models. It was also found that the combination of MSC with OSC may lead to improved performance of the model.

Keywords: chemometrics, NIRS, rapeseed meal

*The paper was presented and published in the frame of activity of the Centre of Excellence AGROPHYSICS – Contract No.: QLAM-2001-00428 sponsored by EU within the 5FP.

INTRODUCTION

Near Infrared Spectroscopy (NIRS), along with multivariate calibration, are being increasingly used to infer properties of the analytes in samples. The aim is to build a calibration equation to predict analytes contents from spectra for future samples, using some of known regression methods. Pre-processing the data before the calibration is often the first step employed in order to reduce the effects which are not related to the parameters of interest. For NIR spectra of granular samples, scattering of radiation and differences in spectroscopic path length, caused by particle size distribution, often constitute the major part of the variation. Thus, signal correction of NIR spectra is a quite wide topic of investigation that includes different approaches developed to do this. Commonly used are multiplicative signal correction (MSC) [3,5,11], standard normal variate (SNV) [3,11], first and further derivative filtering, Fourier transformation, the Savitzky-Golay smoothing filtering [12,13].

It should be noted that it has been difficult to develop a signal correction method to improve the calibration model in general. Therefore, from the viewpoint of a given modelling task, it is desirable to try different variants of data correction and a few regression methods for a best variant of the model to be selected.. The pre-treatment methods mentioned above may be applied to data for which there are no reference measurements. When reference values exist, they can be used to help the choice of the pre-treatment way, so that only a minimum of relevant information included in the spectra can be removed. The orthogonal signal correction (OSC), proposed by Wold *et al.* [18], is a relatively new technique which separates strong structured (i.e. systematic) variation in X-matrix that is not correlated to the response Y-vector or matrix. To date, several algorithms for OSC as a filtering procedure to the data have been discussed [2,4,14-17]. In this paper Trygg and Wold's proposal [15] will be used for pre-processing the spectra matrix.

The OSC method is usually used together with a regression method, such as partial lest squares (PLS) or principal component regression (PCR), to build the calibration model. In this report we will compare the predictive abilities of regular PLS regression models for original data and when the data have been pre-treated with MSC and OSC methods.

MATERIAL AND METHODS

In this study the set of data consisted of NIR spectra coming from 69 samples of rapeseed meal. Reference methods determined the concentration of five constituents: dry mass, protein, oil, ash and fibre. (For more details see accompanying paper by Jankowski *et al*. [7]).

PLS has been found to be the most popular regression method for multivariate data. The theoretical basis for PLS has been widely described in literature and can be found for example in Refs. [6,9,11]. PLS tries to find a relationship between the latent structure in spectra set, X, and the latent structure in responses, Y. It is carried out by finding specific directions in data space, the so-called latent variables, and determining some new vectors called loadings and scores. Two groups of loadings and scores are under investigation; one group for the X matrix – they are commonly called loadings P and scores T, and another group for Y matrix, denoted Q and U, respectively. Loadings give information about the relationship between the original variables directions and the latent variables directions in data space. Scores are the projections of the samples (meant as points in variable data space) on the latent variables directions. Each score vector has a corresponding loading vector. The objective of PLS is to maximize covariance between the first PLS score vector of the data X and the score vector of the responses, Y. To do this one estimates the PLS weights (W) for X, and then calculates the scores blocks for X and Y. The same is then performed for subsequent scores vectors.

In practical use of modelling by latent variables methods, such as PLS, first of all the number of significant latent variables (components) has to be determined for each calibration model. The cross-validation approach provides a very reliable way for this [19]. Validation means a model testing on a data set that has not been used in the development of the model. In cross-validation, the same parts of the data are used in two different roles - once in model making, once in model testing. A number of alternations is performed accordingly to some permutations schemes and then the root mean square error of cross-validation (RMSECV) for all models with different dimensions is calculated. It is commonly accepted that the number of PLS components giving a minimum RMSECV is the proper number for the model that gives optimal prediction. Additionally, regression diagnostic is often based on other statistical parameters used in such analysis; first of all on the coefficient of multiple determination (R^2). (For more complete review the reader is referred to Ref. [1]). Besides, the proper number of PLS components can be inferred based on the analysis of cumulative variance explained in X and Y block matrices. From the viewpoint of modelling efficiency, these variances in the first few components should be as large as possible to provide the satisfactory model. In some situations, the PLS model captures a very large amount of X-variance in the first component and only a low variance on the predicted Y-value. When more components are calculated, the model improves slowly, and finally it is too complicated. In these cases applying OSC could be helpful. The main goal of OSC is to capture Y-orthogonal variation in X within a limited number of orthogonal scores (T_{ort}) and loadings (P_{ort}). Filtered data are obtained after iterative removal of the first 2-3 (usually) orthogonal components, as follows

$$X_{OSC} = X - \sum T_{ortho} \cdot P_{ortho}^{T} \quad (1)$$

The suggested number of orthogonal components is generic. There exists a viable risk of overfitting the estimated model if too many OSC components are removed. This is the first problem encountered with OSC application, being a simple consequence of inaccurate predictor values and thereby removing some relevant, systematic variation in X and leaving only weakly correlated information for the final calibrations. Trygg and Wold [15] introduce two alternative plots as good indicative measures of the correct number of orthogonal components to extract. Such plots will be described further on in this paper.

Different variants of OSC differ in the way they estimate the orthogonal scores and therefore do not give a unique solution. It is common practice to apply principal component analysis (PCA) in order to calculate the orthogonal scores [8]. PCA is a basic tool in analysis of multivariate data matrix X (see e.g. Refs [9-11]). The aim of this method is to decompose X into a limited number of scores (T) and loadings (P) vectors, plus a residual matrix (E)

$$X = \sum T \cdot P^{T} + E \quad (2)$$

In PCA, the loadings and scores have the same meaning as in PLS but the so-called principal components are used instead of latent variables.

The original approach presented by Wold et al. [18] uses the first score vector of X matrix calculated by PCA as a starting score vector orthogonal to Y, T_{ort}. This vector is then orthogonalized to Y in iterative way until convergence is reached. In each iteration, a PLS model is calculated to estimate weights, W_{ort} , and to make product $X \cdot W_{ort}$ as close to T_{ort} as possible. When a suitable Y-orthogonal T_{ort} is found, a loading vector, P_{ort}, is calculated. The OSC corrected matrix is then found according to expression (1). For additional orthogonal components the correction is performed by repeating the steps as below.

The approach proposed by Trygg and Wold [15] first seeks for a matrix W_{ort}, whose columns in the initial stage are loading weights of the regular PLS model. Next, the suitable vector of weights, W_{ort}, is calculated and the score vector orthogonal to Y, T_{ort}, corresponding to these weights, accounts for as much variability in X as possible. The resulting vectors are then used in the same way as described above. This variant of OSC is focused solely on model simplification and improvement of interpretation, which is done by removing only this part of irrelevant variation that creates problems for the PLS model. Moreover, as a rule, the total number of the final PLS components is reduced by the number of OSC components.

RESULTS

Three modified spectra sets have been prepared for the purpose of modelling in addition to the original data set of rapeseed meal spectra. The first of the modified sets consists of MSC pre-processed spectra. The second one consists of OSC corrected spectra and the last one was obtained using OSC followed by MSC (MSC+OSC). In all cases mean centring has been used prior to data modelling.

To test the performance of the OSC pre-processing, the PLS modelling has been made for all prepared data sets. Table 1 shows the percentage of cumulative variance captured by models for spectra, X, and oil as predicted variable Y. Bold figures in the Table help to follow how particular methods for pre-processing of the data influence the cumulative variance in Y space. It can be seen that for original data the model with only one PLS component accounts for a very low percentage of Y-variance. MSC pre-processing improves the model, but still less than 20% of the variance is explained. The OSC method, after removing one orthogonal component for this data set, gives calibration model with substantially better predictive ability. Note that the PLS model built on MSC+OSC corrected data captures the largest amount of Y variance in the first PLS component.

Table1. Percentage of cumulative variance captured by PLS model for X and Y data space, where Y is oil content, when different pre-processing methods are used. Bold figures are explained in the text

Latent	Original		MSC		OSC		MSC + OSC	
variable	X	Y	X	Y	X	Y	X	Y
1	92.3	**3.1**	82.3	**17.7**	77.8	**32.2**	65.8	**70.9**
2	99.0	32.2	90.9	70.9	94.5	54.8	84.8	95.4
3	99.8	54.8	96.0	95.4	97.6	90.0	93.4	97.7
4	99.9	90.9	98.3	97.7	98.7	96.2	96.7	98.1

Figure 1 depicts the calculated RMSECV values versus the number of PLS components obtained for spectra sets prepared with all the procedures, when oil content is predicted. It can be seen that for the original data the significant PLS model requires 13 components. After applying MSC, the number decreases to 12. Application of only the OSC method prior to PLS modelling when 3 orthogonal components are being removed yields 10 components. It is, however, worth noting that the resulting number of the final PLS components is reduced by 3, which is the number of significant orthogonal components. As it can be seen from Figure1, the simplest model was obtained when orthogonal components were calculated for MSC pre-processed spectra and then removed from spectral data

(MSC+OSC approach). The solution given by MSC+OSC pre-processing is advantageous over the OSC transformation in two aspects. First, the model simplification is gained, second - the lowering of RMSECV values, clearly visible in Figure 1, results in better prediction ability of such models.

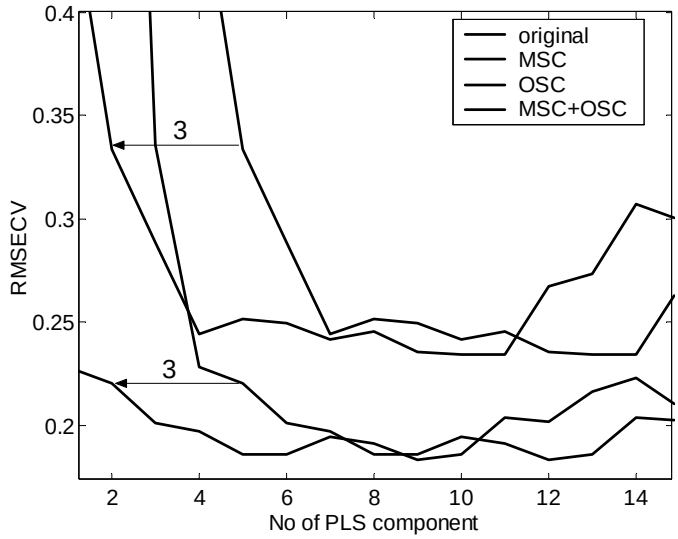

Fig. 1. RMSECV values of oil prediction using PLS model with 1-15 components after different approaches of the pre-processing have been applied

The number of three orthogonal components for modelling of oil content results from Figure 2. This plot, suggested by Trygg and Wold [15], shows the ratio $\|W_{ort}\|/\|P\|$ versus the number of orthogonal components. This ratio becomes close to zero if no orthogonal information remains in X matrix. Hence, the index of the last orthogonal component before the plot flattens is assumed to be suitable. It can be seen from Figure 2 that it happens for three orthogonal components.

Alternatively, the authors propose to follow the relation between X and Y by looking at graphs of the Y scores, U, against the X scores, T, after removing every orthogonal component. These graphs, the so-called t-u score plots, exemplify the enhancement of correlation in the first PLS component after removing from X the part of information that is not correlated with Y. The correct number of orthogonal components is assumed to be found when the pattern of the set of the points became linear. Such diagnostic plots are shown in Figure 3. It can be seen that if only one orthogonal component is removed the plot does not show much correlation, while removal of three orthogonal components makes the dependence clearly linear. The removal of four OSC components makes no noticeable improvement.

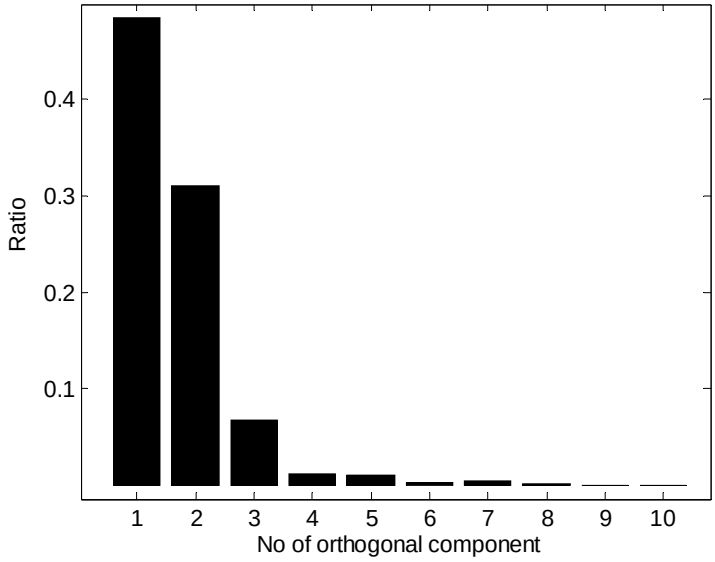

Fig. 2. Plot of the ratio $\|W_{ortho}\|/\|P\|$ versus the number of orthogonal components. Three significant components are indicated to be removed

In a similar way the analyses for the remaining constituents have been made. The numbers of significant PLS components determined from cross-validation are listed in Table 2. MSC and OSC correction methods show quite similar ability in model simplification. It should be noticed that for ash and fibre one can observe a substantial simplification of the calibration models when MSC+OSC approach is applied. The correct number of orthogonal components were obtained based on the ratio $\|W_{ortho}\|/\|P\|$ plot and t-u score plots, as illustrated earlier for oil constituent. Since Trygg and Wold's [15] OSC approach is intended to simplify the model, only the numbers of orthogonal components to be removed in each case can be inferred directly from Table 2 by comparing the values of PLS components for original data with those for OSC pre-processed data or by comparing the values of PLS components for MSC pre-processed data with those obtained when MSC+OSC pre-processing was applied.

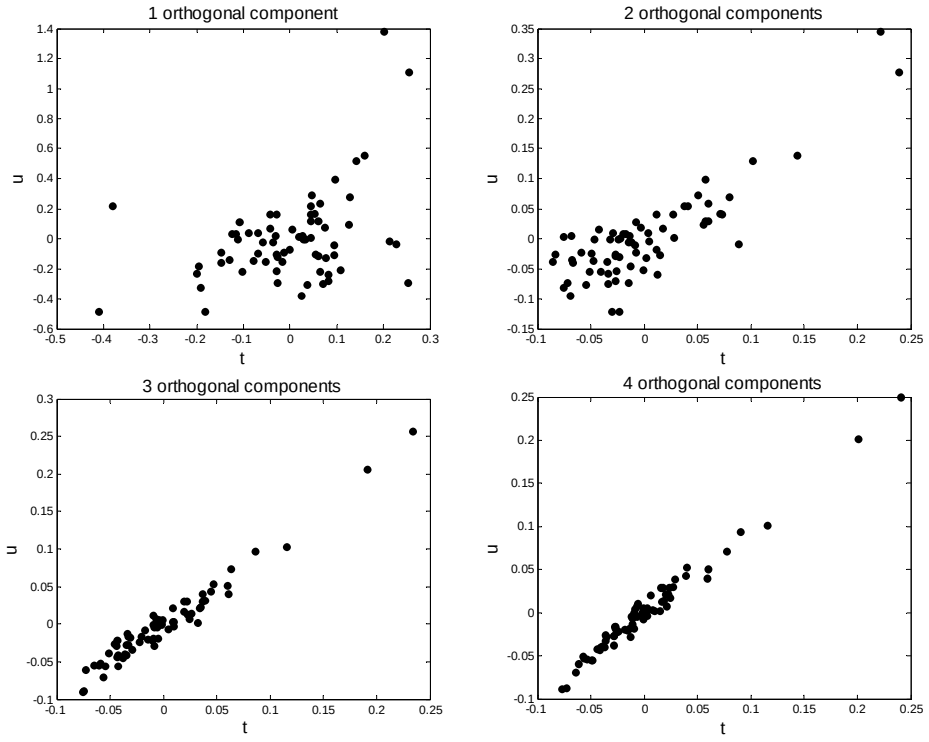

Fig. 3. t-u score plots for the first PLS component after progressive removal of 1,2,3 and 4 orthogonal components from MSC pre-processed spectral data for modelling of oil content

Table 2. Numbers of significant PLS components determined from minimum of RMSECV for models of each constituent

Component	Original	MSC	OSC	MSC+OSC
Dry mass	12	10	10	9
Protein	5	5	2	3
Oil	13	12	10	9
Ash	18	10	15	8
Fibre	9	6	6	4

Once the optimum numbers of the PLS components had been estimated, the regular PLS calibrations with the determination of the statistical parameters were made. The predictive ability of the calibration was estimated in terms of the

RMSECV (as low as possible) and the R^2 (as high as possible) obtained from cross-validation. Figure 4 shows the RMSECV and R^2 values for all analysed constituents after different approaches of pre-processing have been applied.

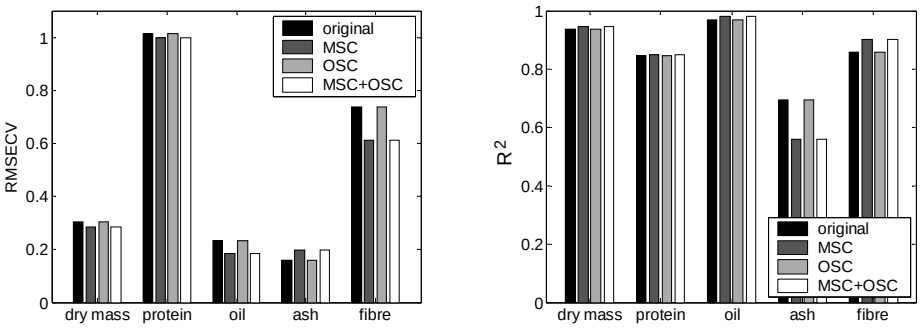

Fig. 4. The RMSECV and R^2 values for all calculated models

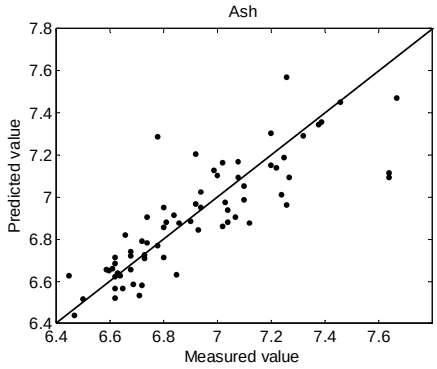

Fig. 5. The predicted versus measured (referenced) values for ash content

It can be seen that the best models for all constituents, except ash, were obtained from spectra pre-treated with MSC as well as when MSC was combined with OSC (MSC+OSC). This result is not surprising since OSC algorithm, in Trygg and Wold's variant, is focused solely on the PLS model simplification and therefore the estimated parameters are finally equivalent to those obtained by MSC. The only diffe-rence is the number of PLS com-ponents to be used. For ash, the modelling by PLS gives quite different results. The only method providing the most favourable parameters and the most simple model is OSC. In this case OSC transformation seems to work more efficiently if it is carried out on the original data. Moreover, if to follow R^2 value it can be concluded that PLS regression for this constituent has some prediction problem because only less than 70% of the variance of the response variable Y is explained by the regression relationship. This can be more directly seen from Figure 5, where rather weak correlation between predicted and measured values exists.

DISCUSSION AND CONCLUSIONS

In this paper we have illustrated on NIR rapeseed meal data set how the orthogonal signal correction of the spectra applied prior to multivariate calibration improved the effectiveness of PLS method. The use of a conventional correction method such as MSC does not require reference values as OSC does, but the use of reference values allows to focus the pre-treatment of the data by orthogonal correction on modelling the Y values. In practice this idea can encounter some problems, because neither accuracy nor precision of the reference measurements are examined. Beside, overfitting of the estimated models is likely to be achieved when too many orthogonal components are removed.

In the literature, the OSC has become an alternative, independent preprocessing method which determines and removes from spectral data X the part of information which is not correlated with Y. However, MSC reduces the additive and multiplicative effects on individual spectra which come from different sample granulations. This information is not simply related with Y, but with optical phenomena accompanying the scattering of light. Thus, OSC as well as MSC may give more or less similar results. How much the results are similar depends on the case. For the data under investigation, the calibration models displayed a viable improvement if OSC approach was combined with MSC pre-processing. However, in one case (for ash), the conventional approach that uses only OSC pre-processing appeared to be more efficient, although not satisfactory results have been obtained. This is probably due to various reasons, like low range of the referenced ash values in the investigated data.

Finally, one can conclude that the proposed OSC pre-processing offers the advantage of at least simplification of the PLS model, and in some cases combination with MSC may lead to improved performance of the model. In general, it is difficult to predict in advance the consequences for calibration task of applying the combined approach (MSC+OSC). Which approach is to be recommended – OSC or MSC+OSC - depends both upon the data and the constituent to be analysed. Further investigations with other data are required to determine the actual merit of the combined MSC and OSC pre-processing.

REFERENCES

1. ASTM Standards: Standard Practices for Infrared Multivariate Quantitative Analysis, Designation: E 1655-00, 100 Barr Harbor Drive, West Conshohocken, PA 19428-2959, United States, 2000.
2. **Andersson C. A.:** Direct orthogonalization. Chemom. Intell. Lab. Sys., 47, 51-63, 1999.

3. **Dhanoa M.S., Lister S. J., Sandeerson R. and Barnes R. J.:** The link between Multiplicative Scatter Correction (MSC) and Standard Normal Variate (SNV) transformation of NIR spectra. J. Near Infrared Spectrosc. 2, 43-47, 1994.
4. **Fearn T.:** On orthogonal signal correction. Chemom. Intell. Lab. Syst. 50, 47, 2000.
5. **Geladi P., MacDougall D., Martens H.:** Linearization and scatter-correction for near infrared reflectance spectra of meant. Appl. Spectrosc. 39, 491-500, 1985.
6. **Geladi P. and Kowalski B.R.,:** Partial Least-squares regression: a tutorial, Anal. Chim. Acta, 185, 1-17, 1986.
7. **Jankowski J., Czarnik-Matusewicz H., Siuda R.:** Comparison of MLR and PLS models in compositional analysis of rapeseed meal from NIR spectra, Acta Agrophysica, 6(1), 91-102, 2005.
8. **Jolliffe T.:** Principal Component Analysis, Springer Verlag, New York, 1986.
9. **Massart D.L., Vandeginste B.G.M., Buydens L.M.C., De Jong S., Lewi P.J., Smeyers-Verbeke J.:** Handbook of Chemometrics and Qualimetrics: Part A and B, Elsevier Science B.V., 1997.
10. **Morrison D.F.:** Wielowymiarowa analiza statystyczna, PWN, Warszawa, 1990.
11. **Naes T., Isaksson T., Fearn T, Davies T.:** Multivariate Calibration and Classification. NIR Publications, 6 Charlton Mill, Charlton, Chichester, West Sussex, 2002.
12. **Savitzky A., Golay M. J. E.:** Smoothing and differentiation of data by simplified least-squares procedures, Anal. Chem. 36, 1627-1639, 1964.
13. **Madden H.H. :** Comments on the Savitzky-Golay Convolution Method for Least-Squares Fit Smoothing and Differentiation of Digital Data, Anal. Chem. 50, 1383-1386, 1978.
14. **Sjöblom J., Svensson O., Josefson M., Kullberg H., Wold S.:** An evaluation of orthogonal signal correction applied to calibration transfer of near infrared spectra. Chemom. Intell. Lab. Syst. 44, 229, 1998.
15. **Trygg J. and Wold S.:** Orthogonal projections to latent structures (O-PLS). J. Chemom., 16, 119-128, 2002.
16. **Trygg J.:** O2-PLS for qualitative and quantitative analysis in multivariate calibration. J. Chemom., 16, 283-293, 2002.
17. **Westerhuis J. A., de Jong S., Smilde A. K.:** Direct orthogonal signal correction. Chemom. Intell. Lab. Sys., 56, 13-25, 2001.
18. **Wold S., Antti H., Lindgren F., Ohman J.:** Orthogonal signal correction of near-infrared spectra. Chemom. Intell. Lab. Sys., 44, 175-185, 1998.
19. **Wold S.:** Cross-Validatory estimation of the number of components in factor and principal components models. Technometrics, 20, 397-405, 1978.

ORTOGONALNA KOREKCJA SYGNAŁU W METODZIE PLS W ZASTOSOWANIU DO DANYCH SPEKTRALNYCH

Grażyna Balcerowska[1], Ryszard Siuda[1], Henryk Czarnik-Matusewicz[2]

[1]Zakład Fizyki Doświadczalnej, Instytut Matematyki i Fizyki, Akademia Techniczno-Rolnicza
ul. Kaliskiego 7, 85-796, Bydgoszcz
e-mail: gbalcer@mail.atr.bydgoszcz.pl
[2]Zakład Farmakologii Klinicznej, Wydział Farmacji, Akademia Medyczna
ul. Bujwida 44, 50-345 Wrocław

Streszczenie. Typowym zadaniem w chemometrii jest oszacowanie liniowej zależności pomiędzy dwoma zbiorami zmiennych: zbiorem widm, X, i zbiorem koncentracji pewnych składników, Y. Jedną z najpowszechniej stosowanych metod regresji jest metoda częściowych najmniejszych kwadratów (partial least squares – PLS). Systematyczne zmiany obecne w X, nie skorelowane ze zmianami w Y, mogą wpływać negatywnie na interpretację modelu PLS. Taka sytuacja może wystąpić w przypadku, gdy zmienne X reprezentują wartości absorbancji lub reflektancji mierzone dla bardzo wielu (setek) długości fal, a pomiary są np. obarczone zaburzeniami pochodzącymi z różnych źródeł, nie mających związku z interesującą nas informacją. W takim przypadku, zaproponowana ostatnio metoda ortogonalnej korekcji sygnału (OSC) może okazać się pomocna. Metoda ta polega na określeniu, a następnie usunięciu z macierzy, widm X tej części informacji, która jest ortogonalna do Y (tj. nie jest skorelowana z Y). Celem pracy jest zilustrowanie możliwości metody OSC, w zastosowaniu do widm śruty rzepakowej zarejestrowanych metodą NIR, poprzez porównanie wyników otrzymanych przy zastosowaniu metody PLS dla danych oryginalnych oraz danych po korekcji metodą MSC (multiplicative scatter correction) oraz OSC. Otrzymane wyniki pozwalają stwierdzić, że metoda OSC upraszcza model kalibracyjny, a gdy jest stosowana do widm po wcześniejszej korekcji MSC obserwuje się w pewnych przypadkach również poprawę statystycznych parametrów charakteryzujących model.

Słowa kluczowe: chemometria, NIRS (spektroskopia bliskiej podczerwieni), śruta rzepakowa