e-PS, 2009, **6**, 75-80
ISSN: 1581-9280 web edition
ISSN: 1854-3928 print edition

www.Morana-rtd.com

e-PRESERVATIONScience

published by MORANA RTD d.o.o.

SCIENTIFIC PAPER

CLASSIFICATION OF SYNTHETIC ORGANIC PIGMENTS BY MULTIVARIATE DATA ANALYSIS OF FTIR SPECTRA

Anke Schäning[1,2*], Kurt Varmuza[3], Manfred Schreiner[1,4]

This paper is based on a presentation at the 8th international conference of the Infrared and Raman Users' Group (IRUG) in Vienna, Austria, 26-29 March 2008.

Guest editor:
Prof. Dr. Manfred Schreiner

1. Institute of Science and Technology in Art, Academy of Fine Arts Vienna, Schillerplatz 3, 1010 Vienna, Austria
2. Institute of Conservation-Restoration, Academy of Fine Arts, Schillerplatz 3, 1010 Vienna, Austria
3. Laboratory for Chemometrics, Institute of Chemical Engineering, Vienna University of Technology, Getreidemarkt 9/166, 1060 Vienna, Austria
4. Institute of Chemical Technologies and Analytics, Vienna University of Technology, Getreidemarkt 9/164, 1060 Vienna, Austria

corresponding author:
a.schaening@akbild.ac.at

For the identification of pure synthetic organic pigments Fourier Transform Infrared (FTIR) spectroscopy has been established as a powerful analytical technique. Usually the identification of such pigments by FTIR is processed by comparing the spectrum of an unknown sample with the spectra present in a database. Those samples are difficult to be identified if they don't match with one of the library spectra, and even class assignments can be performed by experts' analysis only.

In this paper an approach is presented to facilitate classification of organic pigments by multivariate statistical analysis of their FTIR spectra. Partial least-squares discriminant analysis (PLS-DA) was applied to the pigments FTIR spectra, using 11 binary *y*-variables for pigment class definition. The threshold of the discriminant variable $\hat{y}$ could be optimized for each class. It is shown that application of the classification models to FTIR spectra of unknown paint samples gained very similar results as obtained from experts.

1 Introduction

Synthetic organic pigments are nowadays known as the largest group of colorants used in contemporary artists' paints. Thus, investigation of those pigments in works of art has become increasingly important. For the identification of the pure pigments or colorants - already separated from binding medium - Fourier Transform Infrared (FTIR) spectroscopy has been established as a powerful analytical technique, as such spectra are characterized by very sharp and characteristic multitude of absorptions in the fingerprint region.[1, 2] Due to this fact, the identification of synthetic organic pigments by FTIR spectroscopy is processed usually by comparing the spectrum of an unknown sample with the spectra present in a database,[3] which has to be as complete as possible. Interpretation of pigment spectra in terms of chemical structural units is rarely processed, as the organic molecules normally produce lots of overlaid vibrations which can hardly be assigned to

received: 26.06.2008
accepted: 23.03.2009

key words:
Synthetic organic pigments, FTIR, classification, partial least squares, PLS-DA

distinct structures. Therefore, spectra of unknown synthetic organic pigment samples are difficult to be identified if they don't match one of the library spectra. In that case, even class assignment can be performed only by expert analysis. The identification of such unknowns might be facilitated by considering characteristic absorption bands; however, this task is rather time consuming due to the high number of pigments in some of the pigment classes.

In this study we propose a multivariate data analysis method for classification of synthetic organic pigments in order to assist such identification tasks. Multivariate classification procedures can be categorized as supervised pattern recognition methods, whereas a variety of different approaches is known.[4] In comparison to univariate discriminant data analysis multivariate methods refer to more than only one measurement or feature of a single sample. Even for the evaluation of the spectroscopic data the application of multivariate data analysis has been proven to be especially suitable, hence a lot of data were obtained already by the scan of one spectrum.[5] Relating to FTIR spectroscopy, the spectra consist often - depending on range and resolution - of more than 3000 variables respectively data points (absorption data per wave number). In this work FTIR data are used to group the pigments in classes usually defined according to chemical structure criteria.[6] Thus, especially IR absorption data of the fingerprint region are useful, as these data refer to the overlaid vibrations of the whole molecule and to the basic chemical structure, which is similar within one class.

The multivariate chemometric method proposed here is known as Partial Least-Square Discriminant Analysis (PLS-DA).[7] This classification method is based on a PLS regression where class membership is the property, which has to be predicted from the multivariate FTIR data. A similar approach for the classification of inorganic pigments by PLS-DA of dual domain data of Raman and XRF spectra was proposed by Ramos et al.[8] For the study presented here PLS-DA has been performed to compute regression models for prediction of class membership of synthetic organic pigments. The aim of the study is the application of these classification models to FTIR spectra of unknown paint samples and comparing these findings with the results obtained by expert analysis.[9]

2 Materials and Methods

2.1 Samples and Instrumentation

The synthetic organic pigment samples are part of a historical material collection at the Institute of Science and Technology in Art (ISTA) at the Academy of Fine Arts in Vienna, which dates back to the late 19th century. Nowadays the collection contains more than 1300 samples of different artists´ materials, most of them inorganic and organic colorants. About 300 samples belong to the synthetic organic pigments category, some of them with identical Colour Index Number (CI) but from various manufacturers and different periods.

The FTIR spectra of the organic pigments were collected using a Perkin Elmer instrument, type Spectrum 2000, with an attached microscope (Perkin Elmer, i-series). The pigment samples were prepared on a diamond cell and measured in transmission mode (% T, spectral range: 4000-580 cm^{-1}, data interval = 1 cm^{-1}, 100 µm aperture, 100 scans averaged). As a sufficiently large size of the data set is important in multivariate data analysis, the sample set of the ISTA collection has been extended by a set of externally collected reference FTIR spectra.[10] In summary FTIR spectra of 281 (144 from ISTA; 137 from Tate[10]) synthetic organic pigments of the reference spectra collection have been included in the calculations.

Data pre-processing was carried out by applying baseline correction, normalizing and smoothing to all spectra (SPECTRUM®-Software by Perkin Elmer[3]). Due to the fact that all external spectra were collected at a different spectral range and have a resolution other than the own spectra, the external spectra have been interpolated to a data interval of 1 cm^{-1} and the used range has been set to 4000-650 cm^{-1}. All calculations have been performed on transmission spectra with transmission T (in %) being defined by $T = 100\ I\ /\ I_0$, with I_0 for the intensity of the radiation, and I the intensity after passing the sample. No improvement was obtained by using absorbances or derivative spectra.

For the development of classification models the FTIR data had to be transformed into ASCII files, and imported into the software The Unscrambler®.[11]

2.2 Data

The pigment classes were defined according to the usual classification system of organic colorants

based on their chemical constitution.[9] Synthetic organic pigments were usually classified into azo and non-azo or polycyclic pigments and each group into various sub-classes. In this study sub-categories have only been considered if containing at least 12 samples. According to this requirement, samples of eleven pigment classes could be included in the calculations.

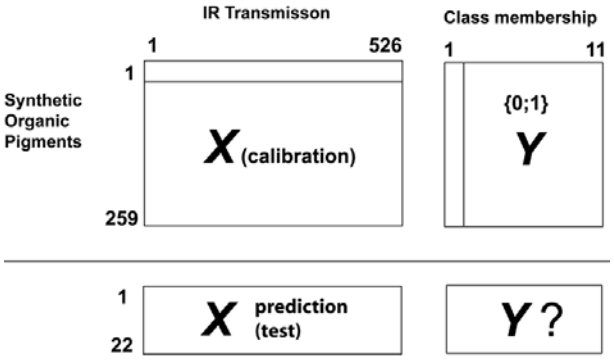

Figure 1: Matrices for PLS-DA.

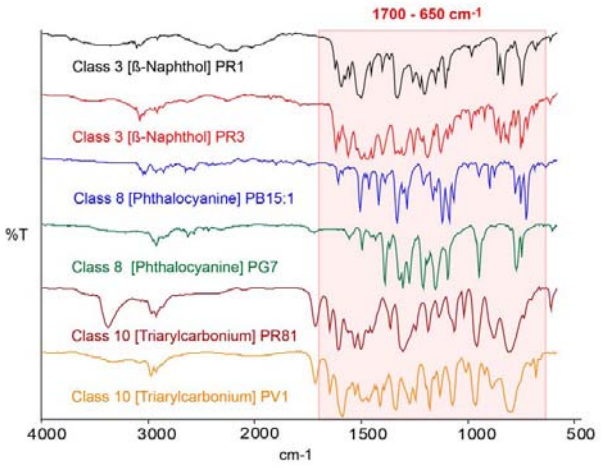

Figure 2: FTIR spectra of synthetic organic pigments; the used spectral range 1700-650 cm^{-1} is highlighted.

No.	Pigment Class	$n_{training}$	n_{test}
1	Monoazo yellow pigments	26	2
2	Disazo yellow pigments	26	2
3	ß-Naphthol pigments	17	2
4	Naphthol_AS pigments	41	2
5	ß-Naphthol Lake pigments	16	2
6	BONS Lake pigments	26	2
7	Benzimidazolone pigments	11	2
8	Phthalocyanine pigments	25	2
9	Quinacridone pigments	9	2
10	Triarylcarbonium pigments	47	2
11	Hydroxyanthrachinon pigments	15	2
	Sum	259	22

Table 1: Pigment classes and number of samples.

The data were prepared as matrices (Figure 1) where in the ***X*** block the rows represent the objects (*n* pigment samples) and the columns the features (*x* variables/transmission data). The used spectral range was 1700-650 cm^{-1}, and data interval of 2 cm^{-1} resulted in *m* = 526 variables. The spectral range has been selected due to the high information content; the region of higher wavenumbers has been omitted (Figure 2). Further variable selection procedures were not applied.

The class membership of the samples is defined by a ***Y*** block. The 11 classes have been binary encoded in 11 *y*-variables (1 and 0 denotes membership to the class and no membership, respectively). The total of 281 pigment samples were split into a training set and a test set (containing two selected samples per class). Thus the training set contained 259 samples and the test set 22 samples (Table 1).

2.3. Multivariate classification method

The multivariate classification method applied was Partial Least-Squares Discriminant Analysis (PLS-DA).[4, 7] PLS regression - as widely used in chemometrics - gives classifiers of the form

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \ldots b_{526} x_{526} \quad (1)$$

with $\hat{y}$ for the predicted *y* (class membership, discriminant variable), x_1 to x_{526} the transmission values at the selected 526 wavenumbers, b_1 to b_{526} the regression coefficients, and b_0 the intercept. PLS allows the use of ***X*** data sets with more variables than samples, highly correlating variables, and optimizes the complexity of the classifier model for a good performance for new cases by using a small set of latent variables (the PLS components) for regression instead of the original variables. In this study separate classifiers have been developed for each of the 11 classes, resulting in 11 sets of regression coefficients. The optimal number of PLS components was estimated by cross-validation (cv), using 10 randomly selected segments. All calculations were performed with the software The Unscrambler®.[11]

Assignment of a sample to a class is based on the value of the discriminant variable $\hat{y}$ by using a critical value (threshold) *t* as follows.

$\hat{y} < t \rightarrow$ class 0
(not a member of the considered class) (2)

$\hat{y} \geq t \rightarrow$ class 1
(member of the considered class) (3)

As the values 0 and 1 are used for *y* (the true class membership), a threshold of 0.5 is a first approxi-

mation for separating "class membership" and "no class membership". However, in this study the threshold of the discriminant variable was optimized separately for each of the 11 pigment classes to achieve maximum prediction performance in cross validation. Varying the threshold between 0.2 and 0.8 (in steps of 0.1) the predictive abilities P_0 and P_1 have been determined for the samples of the calibration set as follows.

$P_0 = 100\ n_{00} / n_0$ (4)

$P_1 = 100\ n_{11} / n_1$ (5)

with n_0 the number of samples not belonging to the considered class, n_1 the number of samples belonging to the considered class, n_{00} the number of correctly assigned samples not belonging to the considered class, and n_{11} the number of correctly assigned samples belonging to the considered class. P_0 and P_1 are the percent correctly assigned samples from "class 0" and "class 1". Note that P_0 and P_1 are independent from the number of samples belonging to the considered class or not. The arithmetic mean P_{mean} of P_0 and P_1 is an appropriate single measure for the total prediction performance of a classifier.[12]

$P_{mean} = (P_0 + P_1) / 2$ (6)

Results show that in general P_0 increases with increasing t, and P_1 increases with decreasing t. P_{mean} typically shows a weak maximum between t = 0.3 and 0.5 (see Results). From this maximum the critical value was determined for each pigment class. Results from cross validation give only a first estimation of the prediction performance; a more reliable measure could be obtained from an independent test set (see Results). The use of 11 classifiers in parallel may result in ambiguous class assignments; e.g. if the value of the discriminant variable is above the critical value for more than one class. Such a situation can be interpreted as warning of a possibly wrong classification.

			Training set n = 259 (cv)				Test set		
CLASS	No.	t	n_1	P_0 %	P_1 %	$P_{01\ mean}$ %	n_1	Classification result	
								correct	wrong
Monoazo yellow pigments	1	0.3	26	97.9	96.2	97.0	2	2	0
Disazo yellow pigments	2	0.5	26	99.6	100.0	99.8	2	2	0
ß-Naphthol pigments	3	0.3	17	99.6	100.0	99.8	2	2	0
Naphthol_AS pigments	4	0.4	41	99.6	100.0	99.8	2	2	0
ß-Naphthol Lake pigments	5	0.5	16	100.0	93.8	96.9	2	2	0
BONA Lake pigments	6	0.4	26	99.1	100.0	99.6	2	2	0
Benzimidazolone pigments	7	0.5	11	100.0	90.9	95.5	2	2*	0
Phthalocyanine pigments	8	0.5	25	100.0	100.0	100.0	2	2	0
Quinacridone pigments	9	0.5	9	100.0	100.0	100.0	2	2	0
Triarylcarbonium pigments	10	0.3	47	99.5	100.0	99.8	2	2	0
Hydroxyanthrachinon pigments	11	0.5	15	100.0	100.0	100.0	2	2	0
			259				22		

Table 2: Classification results for cross validation with training set and independent test set. One of the test set samples (*) of the Bezimidazolone pigment class (No. 7) was classified ambiguously. It was assigned correctly as member of class 7 but also misattributed as a member of class 4.

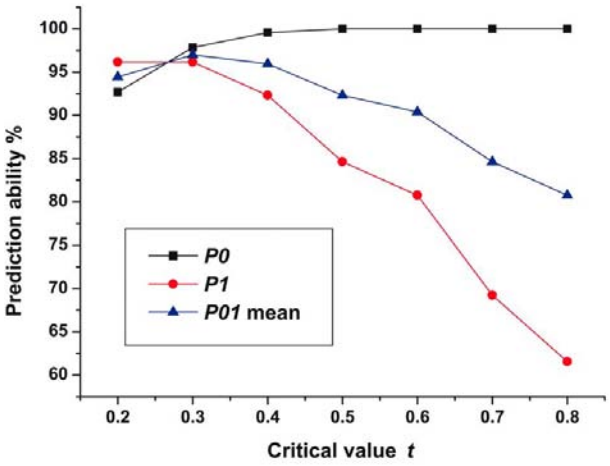

Figure 3. From the broad maximum of P_{mean} an optimum value for the threshold of 0.3 has been selected. The results from cross validation with the training set are summarized in Table 2. For eight pigment classes the membership to the class was correctly determined for all samples of the training set (P_1 = 100%). For three other pigment classes P_1 is between 90.9 and 96.2%. P_0 and P_{mean} are for all pigment classes near 100%.

3 Results

3.1 Model calibration and threshold optimization with training set

Eleven PLS models - one for each pigment class - have been computed using the training set. The classification threshold has been optimized for each pigment class as described in 2.3; results for the monoazo yellow pigment class are shown in

3.2 Model Validation

The quality of the prediction performance of the models has been evaluated by the independent test set of the 22 pigment samples sorted out of the data set previously. The PLS-DA models were applied to the test set samples by using the optimal number of PLS-components and the optimum threshold for the discriminant variable (both obtained from the training set). Results in Table 2 show that for 10 of the 11 pigment classes both samples were correctly assigned. One test set sample of class 7 (Benzimidazolone pigments) was partially misclassified as that sample was assigned to two classes: one correct and another wrong. These results indicate a presumably high classification power of the developed method; however, the rather small test set does not allow a final evaluation.

3.3 Classification of Unknown Samples

Since a good classification performance of the PLS-DA models could be seen from the results obtained for the test set samples, capability of the multivariate data analysis of the synthetic organic pigments FTIR spectra should be verified on real paint samples, taken from works of art.

For that purpose PLS-DA classification models of the 11 pigment classes were applied to the FTIR spectra of six unknown paint samples of two paintings by My Ullmann (Figures 4 and 5) dated about 1925 and the results were compared with those obtained by expert analysis[9] (considered as true results).

Spectra of the unknown samples have to be preprocessed in the same way as the FTIR spectra in the training set. As some of the samples depicted only weak intensities of relevant absorption bands, the normalizing function has been modified. The parameter 'zero point', which is adjusting the baseline, had to be adapted by setting the abscissa value manually to the point at which the baseline had to be scaled. The wavenumber selected is the minimum absorbance in the spectral range of 1700-650 cm^{-1}, which has been considered for multivariate classification.

The classification results for the six paint samples are presented in Table 3. Three of the samples (P1, P3, P4) have been classified correctly; one sample (P6) was assigned to class 1 and 3, both azo pigment classes. Samples P2 and P5 yielded no assignment to any class, as no predicted value

Unknown Paint Samples		Class (expert)	n	correct	wrong
P1	Yellow (Painting 1)	1	2	1	1
P2	Light Green (Painting 1)				
P3	Dark Red (Painting 1)	11	3	2	1
P4	Dark Red (Painting 1)				
P5	Dark Red (Painting 2)				
P6	Red (Painting 2)	3	1	1*	0
sum			6	4	2

Table 3: Prediction of unknown paint samples; P6 was classified inconclusively (*) as class 1 and class 3 member.

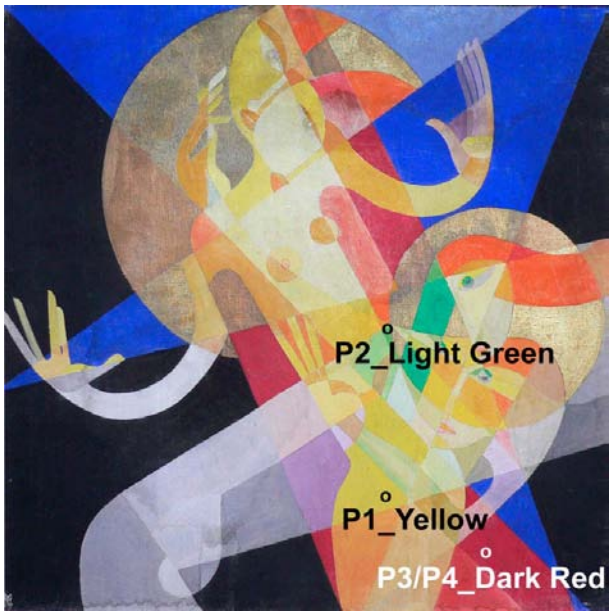

Figure 4: Painting 1 „Composition with Two Nudes", My Ullmann, 1925, 80 x 79.8 cm, gouache on canvas, Wien Museum, Vienna.

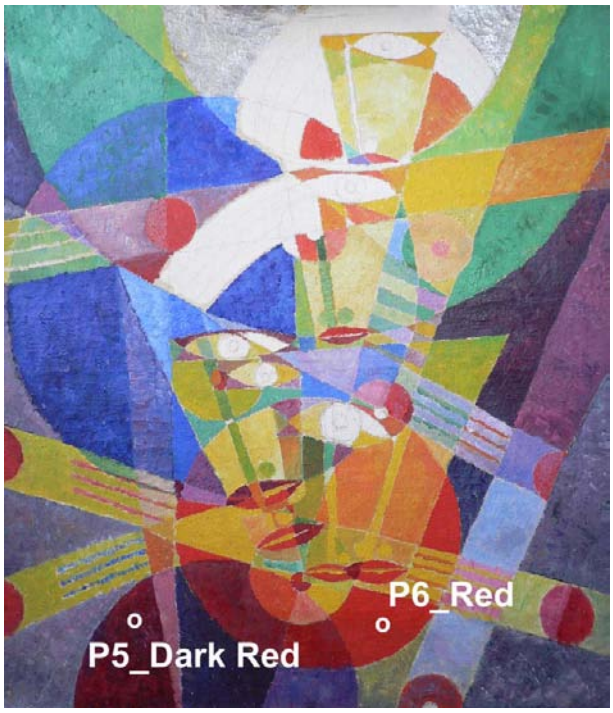

Figure 5: Painting 2 „Faces in a Circle Figuration", My Ullmann, 1923, 65 x 55 cm, oil on canvas, Wien Museum, Vienna.

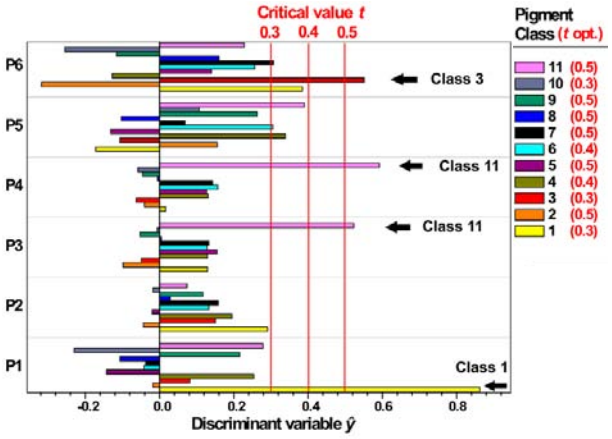

Figure 6: Classification results of six paint samples taken from the paintings by My Ullmann (see Figure 4, 5): The diagram shows for each sample the values of the discriminant variables, $\hat{y}$, obtained from 11 classification models. The sample is classified as class member, if $\hat{y}$ is equal or higher than the critical value, which is enclosed in the legend for each class.
Samples from painting 1 (Fig. 4): P1/Yellow, P2/ Pale Green, P3/Dark Red, P4/Dark Red and from painting 2 (Fig. 5): P5/Dark Red, P6/Red.

$\hat{y}$ was above the class-specific threshold. The latter result can be explained by the relative weak absorption bands of these FTIR spectra in the relevant spectral region. In these cases even the manually adapted normalizing procedure did not gain a better result. Interestingly, prediction results $\hat{y}$ for samples P2 and P5 were highest for the pigment class suggested in the expert analysis (Figure 6). Note that the value of the discriminant variable may be negative due to the applied regression approach in which the target value for class 0 has been set to value 0; because all critical values are positive, such a sample is assigned to class 0.

4 Conclusion

Multivariate data analysis of FTIR-spectra from synthetic organic pigments is a promising approach to support the classification of unknown samples which do not match library spectra. It was shown that the class of synthetic organic pigments can be determined with a high success rate by multivariate classification of their FTIR spectra, and furthermore classification of unknowns was possible by PLS-DA.

It is important to note that the spectra of unknowns must have strong absorption bands for the applicability of the developed method. In many cases this can be achieved by pre-processing of the spectral data, in particular by the normalizing function implemented in the IR software which is providing individually adjustments of the parameters.

Future work should consider determination of the main- and subclass categories of synthetic organic pigments to provide better class separation. Further improvement of the method could be achieved by including more pigment classes, an expansion of the number of pigment samples available per class, and by application of variable selection procedures to optimize class separation.

5 Acknowledgement

The authors gratefully acknowledge Dr. T. Learner (GCI) for providing FTIR reference spectra of the Tate Organic Pigment Archive, Millbank, London SW1P 4 RG.

6 References

1. M.R. Derrick, D. Stulik, *Infrared spectroscopy in conservation science,* Scientific tools for conservation series, Getty Conservation Institute, Los Angeles, 1999, 113-114.

2. T. Learner, *Analysis of Modern Paints*, Research in Conservation Series, Getty Conservation Institute, Los Angeles, 2004, 92-97.

3. Software *Spectrum Search Plus 7.5*, Perkin Elmer, 40 Winter Street, Waltham, Massachusetts 02451, USA, http://www.perkinelmer.com

4. R.G. Brereton, *Chemometrics. Data Analysis for the Laboratory and Chemical Plant*, Wiley & Sons, Chichester, 2006, 183-269.

5. T. Naes, T. Isaksson, T. Fearn, T. Davies, A user-friendly guide to multivariate calibration and classification, NIR Publications, Chichester, 2004.

6. W. Herbst, K. Hunger, *Industrielle Organische Pigmente, Herstellung, Eigenschaften, Anwendung*, VCH, Weinheim, 1995, 4-11.

7. L. Eriksson, H. Antti, E. Holmes, E. Johansson, T. Lundstedt, J. Shockcor, S. Wold, *Partial least squares (PLS)*, in: J. Gasteiger (ed.) *Handbook of Chemoinformatics*, Vol. 3, Wiley-VCH, Weinheim, 2003, 1134-1166.

8. P. M. Ramos, I. Ruisanchez, *Data fusion and dual-domain classification analysis of pigments studied in works of art*, Anal. Chim. Acta, 2006, **558**, 274-282.

9. A. Schäning, M. Schreiner, M. Mäder, U. Storch, *Synthetische organische Pigmente in Künstlerfarben des frühen 20. Jahrhunderts: Möglichkeiten und Grenzen ihrer Identifizierung am Beispiel von zwei Gemälden um 1925 von My/Marianne Ullmann,* Zeitschrift für Kunsttechnologie und Konservierung, 2007, **1/2007**, 87-110.

10. FTIR reference spectra of the Tate Organic Pigment Archive, London, UK.

11. *The Unscrambler v 9.7*, CAMO Software AS, Norway.

12. K. Varmuza, P. Filzmoser, *Introduction to Multivariate Statistical Analysis in Chemometrics*, CRC Press, Boca Raton, 2009.