Engineering Proceedings

MDPI

Proceedings

Multivariate Spectra Analysis: PLSR vs. PCA + MLR †

Sander Vervoort [1,2] * and Marcus Wolff [1]

[1] Hamburg University of Applied Sciences, Heinrich Blasius Institute for Physical Technologies, Berliner Tor 21, 20099 Hamburg, Germany; email1@gmail.com
[2] University of the West of Scotland, School of Computing, Engineering & Physical Sciences, PA1 2BE Paisley, Scotland, UK
* Correspondence: sander.vervoort@haw-hamburg.de; Tel.: +49-40-42875-8655
† Presented at the 7th Electronic Conference on Sensors and Applications, 15–30 November 2020; Available online: https://ecsa-7.sciforum.net/.

Published: 15 November 2020

Abstract: For mixtures of compounds with very similar spectral features, common for larger organic molecules, multivariate analysis (MVA) methods can be applied to determine the concentration of the individual component. We analyzed photoacoustic spectra of mixtures of different volatile organic compounds with and without different feature selection and feature projection methods. These include: Multiple Linear Regression (MLR), Principal Component Analysis (PCA), Partial Least Squares Regression (PLSR) and Random Forest Algorithm (RFA). Even though PLSR provided the best prediction accuracy, the other techniques also exhibited some advantages.

Keywords: multivariate linear regression; partial least squares regression; feature selection; feature projection; random forest algorithm; principal component analysis; chemometrics; photoacoustic spectroscopy; VOC

1. Introduction

Spectroscopic probing of energetic transitions of molecules or atoms enables the analysis of mixtures and the selective determination of concentrations. Successfully applied laser spectroscopic methods include absorption spectroscopy, atomic emission spectroscopy, fluorescence spectroscopy and photoacoustic spectroscopy (PAS) [1].

If the spectral features of the single substances are broad and overlap strongly, the spectra evaluation requires a multivariate analysis. The general suitability of Partial Least Squares Regression (PLSR) to determine the absolute concentrations of different components of a mixture has been demonstrated [2,3]. However, the according study also revealed certain limitations of this evaluation method. Therefore, we further investigated methods of multivariate statistics and compared their prediction accuracy.

2. Materials Methods

2.1. Experiment

The investigation was performed on mixtures of five Volatile Organic Compounds (VOCs): 2-Butanone (C_4H_8O), 1-Propanol (C_3H_8O), Ethylbenzene (C_8H_{10}), Styrene (C_8H_8) and Hexanal ($C_6H_{12}O$).

A spectrum of each VOC was recorded with a photoacoustic analyzer based on an optical parametric oscillator (OPO). The system delivers highly resolved spectra in the mid-IR wavelength region between 3.2 μm and 3.5 μm [4,5].

The measured spectra of the single VOCs were weighed and additively combined in several variations in order to get a larger dataset over a wider range of concentrations. To consider the measurement uncertainty, noise is added to each of these synthetic mixtures.

2.2. Multivariate Analysis

Multivariate analysis (MVA) is used to identify the relationship between the photoacoustic spectra and the concentrations of different VOCs. The so-called response matrix $\mathbf{Y} \in \mathbb{R}^{(m,n_y)}$ contains the dependent variables (y_{il}), i.e. the concentration of VOC (l) in mixture/spectrum (i). We investigated ($n_y = 5$) components and ($m = 100$) mixtures. The predictor matrix ($\mathbf{X} \in \mathbb{R}^{(m,n_x)}$) contains the independent variables (x_{ij}), which correspond to the photoacoustic signal of mixture/spectrum (i) at wavelength (j). One measurement contains ($n_x = 200$) values, which are equally distributed over the wavelength range (3.3 µm to 3.5 µm) in 1 nm steps. For the analysis, the synthetic spectra are split into a training set of 70 spectra and a validation set of 30 spectra.

The investigated methods, including Multiple Linear Regression (MLR), Partial Least Squares Regression (PLSR) and Principal Component Analysis are linear methods. Since the absorption of the VOCs at low concentrations is relatively weak, a linear relationship between the photoacoustic signal and the concentration can be assumed. According to the simplest model, the MLR is defined as follows:

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E}_{\mathrm{pre}} \ , \tag{1}$$

$$\hat{\mathbf{Y}} = \mathbf{XB} \ , \tag{2}$$

with the model's linearity coefficients ($\mathbf{B}$), the prediction error ($\mathbf{E}_{\mathrm{pre}}$) and the predicted values (here concentrations vector) ($\hat{\mathbf{Y}}$).

2.3. Dimensionality Reduction by Feature Projection

A way to increase the accuracy of the regression can be a dimensional reduction. The PLSR performs this dimensionality reduction as feature projection prior to the actual regression. Feature projection is a technique to generate new, fewer variables, while preserving most of the information of the original dataset.

While the Principal Component Analysis (PCA) only decomposes the matrix of independent variables ($\mathbf{X}$) (Equation (3)), the PLSR also decomposes the matrix of dependent variables ($\mathbf{Y}$) into corresponding linear combinations ($\mathbf{TP}^{\top}, \mathbf{UQ}^{\top}$) (Equation (4)) [6]:

$$\mathbf{X} = \mathbf{TP}^{\top} + \mathbf{E} \ , \tag{3}$$

$$\mathbf{Y} = \mathbf{UQ}^{\top} + \mathbf{F} \ . \tag{4}$$

The noise in the data set ($\mathbf{X}$) and ($\mathbf{Y}$) is indicated by the corresponding error vectors ($\mathbf{E}$) and ($\mathbf{F}$).

The regression model described by Equation (2) also applies to PLSR. The linearity coefficients ($\mathbf{B}$) are determined by the model's weights ($\mathbf{W}$) and loadings ($\mathbf{P}$ and $\mathbf{Q}$) [6]:

$$\mathbf{B} = \mathbf{W}(\mathbf{P}^{\top}\mathbf{W})\mathbf{Q}^{\top} \ . \tag{5}$$

The Equations (3)–(5) can be solved by the Nonlinear Iterative Partial Least Squares (NIPALS) algorithm [6].

2.4. Dimensionality Reduction by Feature Selection

In addition to the feature projection we investigated the feature (subset) selection, a method of dimensionality reduction in which only the most relevant features (independent variables) from the original data set are retained [7,8]. We used the Random Forest Algorithm (RFA) implementation of the Scikit-learn library based on Python, version 0.19.1 [9].

3. Results and Discussion

For the evaluation of the individual methods, two values are considered: The Mean Absolute Error (MAE) and the standard deviation (**s**) of ($\mathbf{E}_{\mathrm{pre}} = \hat{\mathbf{Y}} - \mathbf{Y}$), both averaged over all five VOCs. The bias for the different prediction methods is 6 ppb and below.

Table 1 lists the results of the different multivariate analysis methods.

Table 1. Results of different multivariate analysis methods.

	MAE/ppm	s/ppm
MLR	6.8	9.2
RFA + MLR	6.8	9.2
PCA + MLR	5.9	7.9
PLSR	5.8	7.8

MLR is in general well suited for determining concentrations but gives less accurate results compared to the other methods. Even in combination with the RFA as feature selection method the accuracy remains the same. However, the method has a significant advantage. Applying a feature selection reduces the measuring time considerably since not the entire spectrum has to be recorded but only the ca. 70% with the most significant values. This enables sensors with approximately 30% shorter response time which is quite relevant considering that it can take several hours to record a complete spectrum.

Applying feature projection like PCA and PLSR shows a significant increase in prediction accuracy. In this case the PLSR provides the highest prediction accuracy of all methods. An advantage of PCA+MLR is that the dimensional reduction is performed independently of the regression and even data sets of completely unknown composition can be used.

Based on the first results presented here, the MVA models will be investigated in the future by cross-validation and additional test data in the form of real gas mixtures. In addition, the feature selection will be investigated in greater depth. It can also be combined with the feature projection methods which have been introduced here.

Author Contributions: Conceptualization, M.W.; methodology, S.V.; experiment, S.V.; data curation, S.V.; writing—original draft preparation, M.W. and S.V.; writing—review and editing, M.W. and S.V.; supervision, M.W.; All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Demtröder, W. *Laserspektroskopie: Grundlagen und Techniken*, 4th ed.; Springer-Verlag: Berlin/Heidelberg, Germany, 2000; ISBN 978-3-662-08266-9.
2. Loh, A.; Wolff, M. Multivariate Analysis of Photoacoustic Spectra for the Detection of Short-Chained Hydrocarbon Isotopologues. *Molecules* **2020**, *25*, 2266, doi:10.3390/molecules25092266.
3. Saalberg, Y.; Wolff, M. Multivariate Analysis as a Tool to Identify Concentrations from Strongly Overlapping Gas Spectra. *Sensors* **2018**, *18*, 1562, doi:10.3390/s18051562.
4. Bruhns, H.; Marianovich, A.; Wolff, M. Photoacoustic Spectroscopy Using a MEMS Microphone with Inter-IC Sound Digital Output. *Int. J. Thermophys.* **2014**, *35*, 2292–2301, doi:10.1007/s10765-014-1690-5.
5. Bruhns, H.; Saalberg, Y.; Wolff, M. Photoacoustic Hydrocarbon Spectroscopy Using a Mach-Zehnder Modulated cw OPO. *Sens. Transducers* **2015**, *188*, 40.
6. Kessler, W. *Multivariate Datenanalyse für die Pharma-, Bio- und Prozessanalytik: ein Lehrbuch*, 1st ed.; WILEY-VCH: Weinheim, Germany, 2008; ISBN 978-3-527-31262-7.
7. Guyon, I.; Gunn, S.; Nikravesh, M.; Zadeh, L.A. *Feature Extraction: Foundations and Applications*; Springer: Berlin/Heidelberg, Germany, 2008; ISBN 978-3-540-35488-8.
8. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32, doi:10.1023/A:1010933404324.

9. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.