# Multivariate Projection of Spectral Data Considering Correlations to the Corresponding Chemical Structures

*W. Werther[1] and K. Varmuza[2]*

[1] University of Vienna, Inst. of Analytical
Chemistry, Lab. for Mass Spectrometry
Währingerstrasse 38, A-1090 Vienna
Wolfgang.Werther@univie.ac.at

[2] Vienna University of Technology, Inst. of
Chem. Engineering, Lab. for Chemometrics
Getreidemarkt 9, A-1060 Vienna
kvarmuza@email.tuwien.ac.at

## Introduction

Different spectroscopic methods play an important role for the identification and structural interpretation of unknown chemical compounds. In organic analytical chemistry mass spectrometry (MS), infrared spectroscopy (IR) and nuclear magnetic resonance spectroscopy (NMR) supply most of the information for this task.

In a mathematical sense a spectrum is a multivariate signal - that means a series of single measurement values - for one compound. In multivariate data analysis such a series of values can be imagined as a point in a multidimensional data space spanned by these measurement variables.

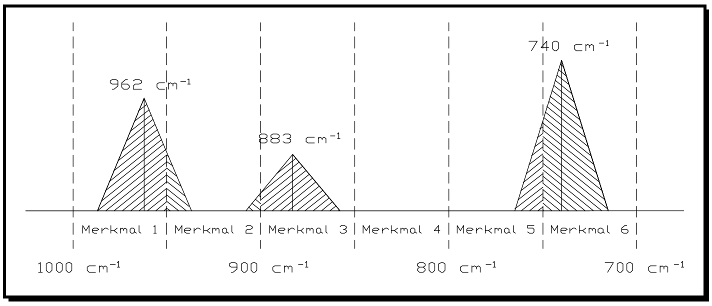
Exploratory data analysis of multivariate spectral data is often done by linear mapping methods enabling the human visual exploration and the interpretation of clustering and class separation. Mostly principal component analysis (PCA) is used for mapping purposes. As a major disadvantage of the variance-maximizing properties of PCA  the mapping structure can be dominated by parts of the spectral data that show no or only weak correlations to the chemical structure.

A "partial-least-squares" (PLS) based mapping can be utilized to overcome this drawback. Spectral features are applied as x-variables and binary structural descriptors as y-variables. The result of this method are two sets of pairwise correlated latent variables which can be used to span corresponding projection planes from both data spaces.

# Spectral X-Variables

## *IR spectroscopy*

The infrared spectra used in this work are peak tables. Calculation of **"fuzzy" wavenumber interval sums" (FWIS)** [1] has been applied in this work to consider the "fuzziness" of the IR peak position of related compounds.



## *$^{13}$C-NMR spectroscopy*

**"Fuzzy chemical shift interval sums" (FCSIS)** are calculated in analogy to the concept in IR spectroscopy. Five different sets of FCSIS can be calculated considering either all peaks, or only singulets, doublets, triplets or quartets.

## *Mass spectrometry*

Fragmentation pathways are very complex and sensitive to little structural changes. It is often difficult to formulate a direct and always existing relationship between a structural element and a spectroscopic signal as possible in $^{13}$C-NMR and IR spectroscopy. Therefore more complex calculations have been applied to get **"spectral features"** in mass spectrometry [2]:

**modulo-14 spectra**
**autocorrelation spectra**
**logarithmic intensity ratios**

# Structural Y Variables

In this work **binary structural descriptors** are applied to transform chemical structures into numerical variables. To derive binary descriptor variables a set of substructures or structural categories is defined. If a substructure is present in the compound, the binary descriptor value is set to 1, if the substructure is absent, the value is set to 0.

Atom-centered **"HOSE code"** substructures[3] are used as binary structural descriptors: HOSE codes can cover the whole chemistry and such substructures are terms which are easy to interprete for the chemist.

# PLS mappings

PLS is a special case of a common principle to calculate vectors spanning a mapping plane in data space. This common principle is the decomposition of the data matrix into latent variables:

$$\mathbf{X} = \mathbf{T}\ \mathbf{P}^{\mathbf{T}} + \mathbf{E} \qquad (1)$$

A decomposition equation similar to (1) can be formulated for the block of the y variables:

$$\mathbf{Y} = \mathbf{U}\ \mathbf{Q}^{\mathbf{T}} + \mathbf{F} \qquad (2)$$

Equations (1) and (2) are called "outer relationships" and are linked together with the "inner relationship":

$$\mathbf{U} = \mathbf{T}\ \mathbf{B} + \mathbf{H} \qquad (3)$$

It is interesting and informative to compare the role of the error residual matrices **E**, **F** and **H** within three different latent variable methods: principal component analysis (PCA), canonical correlation analysis (CCA) and partial-least-squares (PLS). In PCA only outer relationships influence the result minimizing independently **E** and **F**. In CCA only the inner relationship is important minimizing the regression error **H**. PLS takes account of both outer and inner relationships minimizing **E**, **F** and **H** simultaneously.

For a PLS mapping as a two-dimensional data model the inner relationship (3) can be formulated into two vector equations:

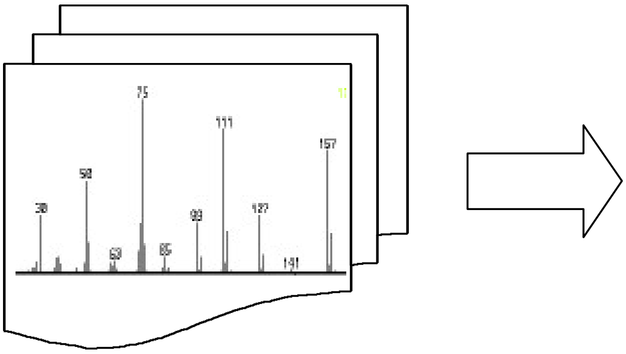
$$\mathbf{u}_1 = b_1\ \mathbf{t}_1 + \mathbf{h}_1 \qquad (4)$$
$$\mathbf{u}_2 = b_2\ \mathbf{t}_2 + \mathbf{h}_2 \qquad (5)$$

In a simplified view x data space mappings of CCA and PLS can be seen as an estimate of the y-mapping with an certain error. Ideally objects would have the same position in both mappings after scaling the x scores with the regression coefficients. The differences between the positions are the two-dimensional residual vectors $\mathbf{h}_1$ and $\mathbf{h}_2$.

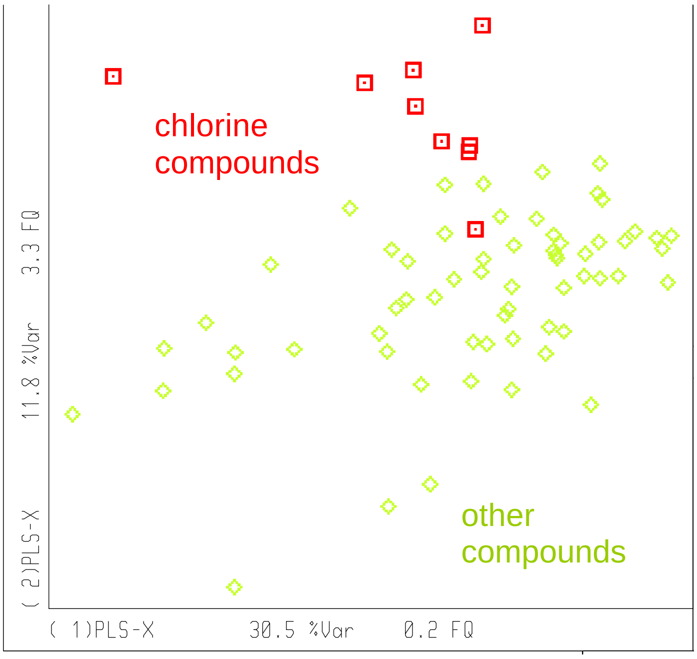
There exists no comparable geometric connection between the PCA mappings of x and y data space, because the regression residual matrix **H** plays no role in the calculation of the principal components.

Considering the PLS y-mapping of the structural descriptors, the position of the objects is directed by correlations to spectral variables and can reveal structural clusters relevant in spectroscopic terms.

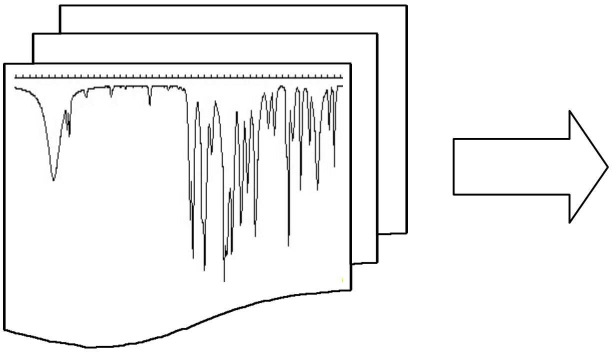
mass spectral variables



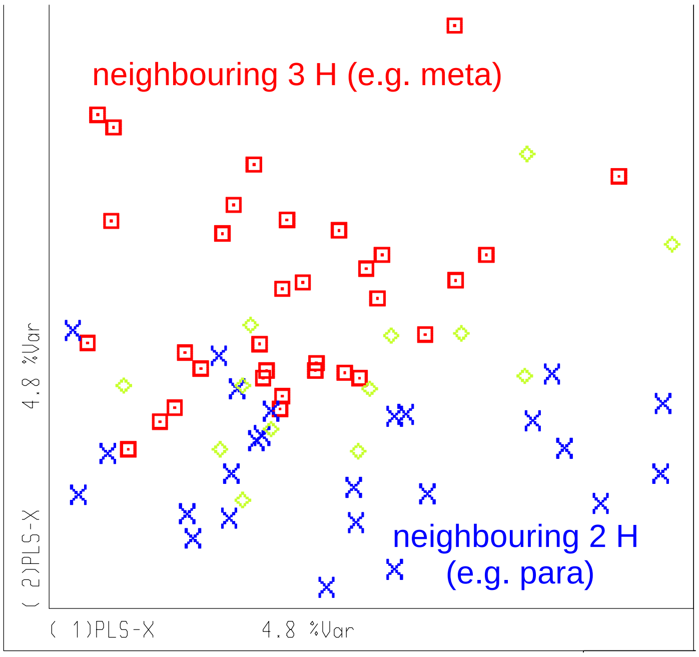
autocorrelation features
($\Delta m/z$ 1 - 50)



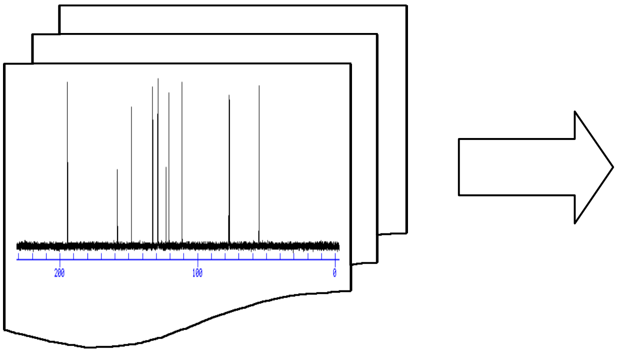
infrared spectral variables



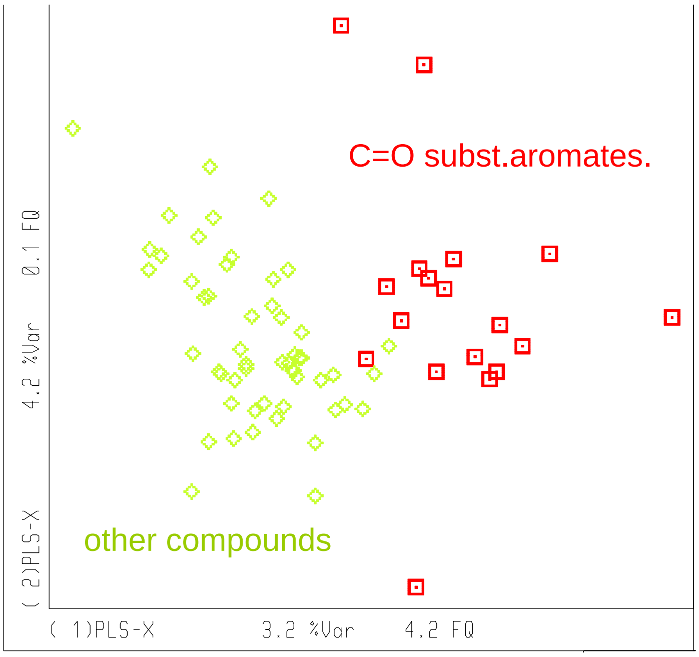
fuzzy wavenumber interval sums (500-1000cm$^{-1}$)



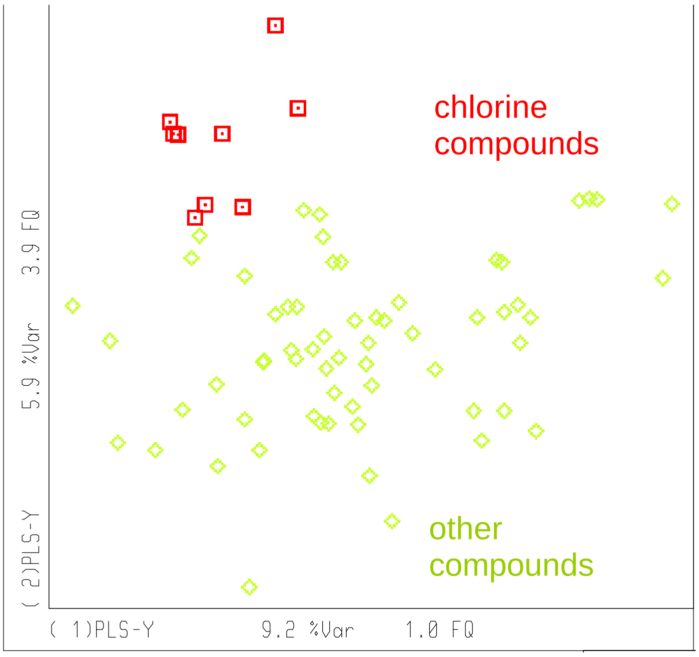
$^{13}$C-NMR spectral variables



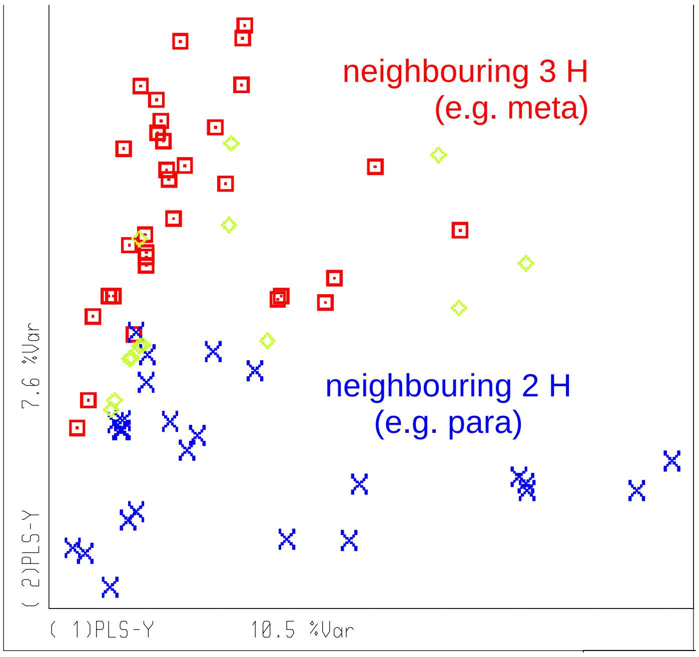
220 fuzzy chemical shift interval sums (0-220ppm)

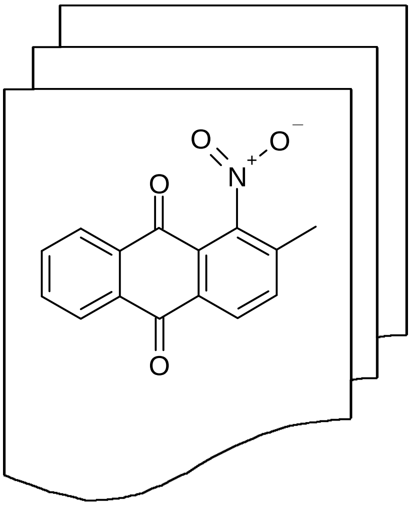


The structural map is an estimate of the spectral map (and vice versa).



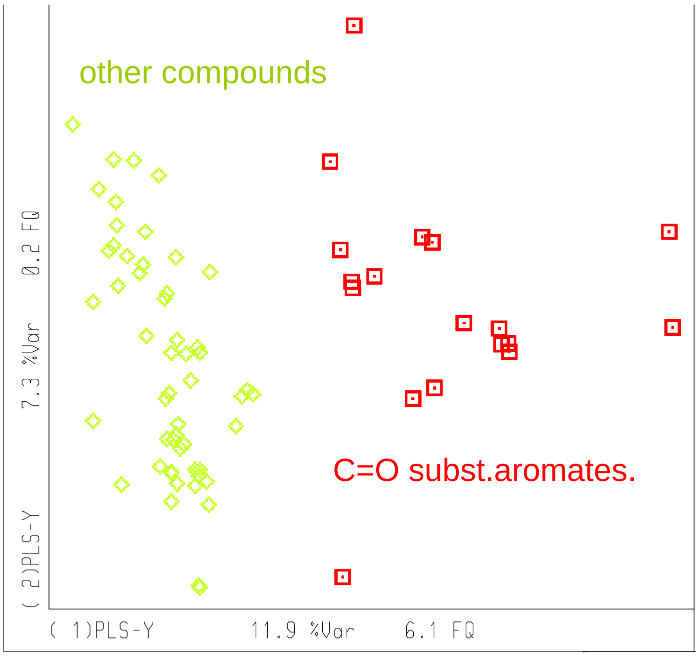
71 nitro compounds





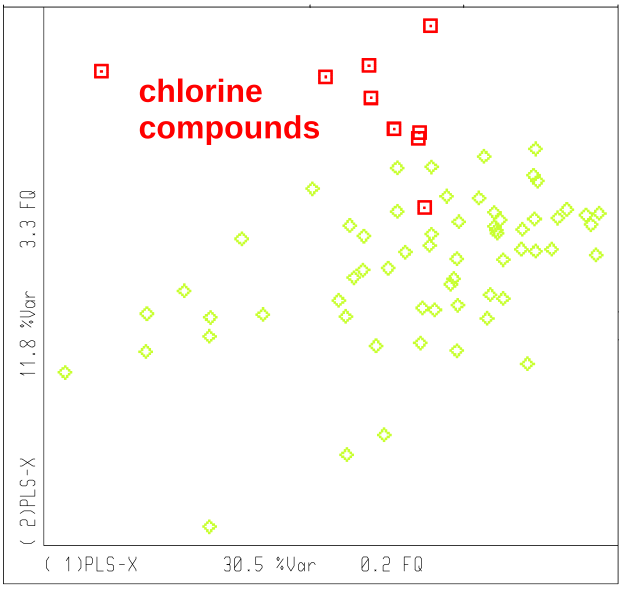
binary structural descriptors

179 HOSE codes

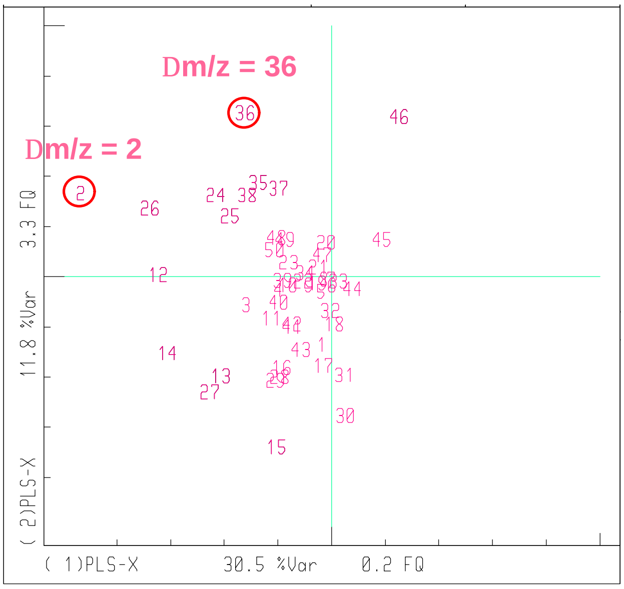


Different spectroscopic methods (and particularly different spectral features) lead to different structural maps = projection planes in structural data space.

# Loading interpretation



mass spectral map



mass spectral loadings

# Conclusions

Finding "structural similarity clusters" (contiguous regions in a mapping dominated by objects of one chemical structure class) is a major step for the detection of relevant spectra-structure-relationships. Interpretation of the loadings of the latent variables of both data spaces can then reveal spectral reasons for clustering and for separation of structure classes. For a complete utilization of a PLS mapping all four available plots (spectral and structural map as well as the corresponding loading-loading-plots) should be examined in common.

## *Acknowledgements:*

The work has been supported by the Austrian "Bundesministerium für Wissenschaft und Forschung". The authors wish to thank "Chemical Concepts" (Weinheim, Germany) and the "Max-Planck-Institut für Kohlenforschung" (Mülheim a. d. Ruhr, Germany) for software support. Discussions about PLS with Heidrun and Peter Filzmoser (Technical University of Vienna) have been an important help for the success of this work.

## *References*

[1] W. Werther, K. Varmuza, Fresenius J. Anal. Chem., **344** (1992), 223.

[2] W. Werther et al., J. Chemom., **16** (2002), 99.

[3] W. Bremser, Anal. Chim. Acta **103** (1978), 355.