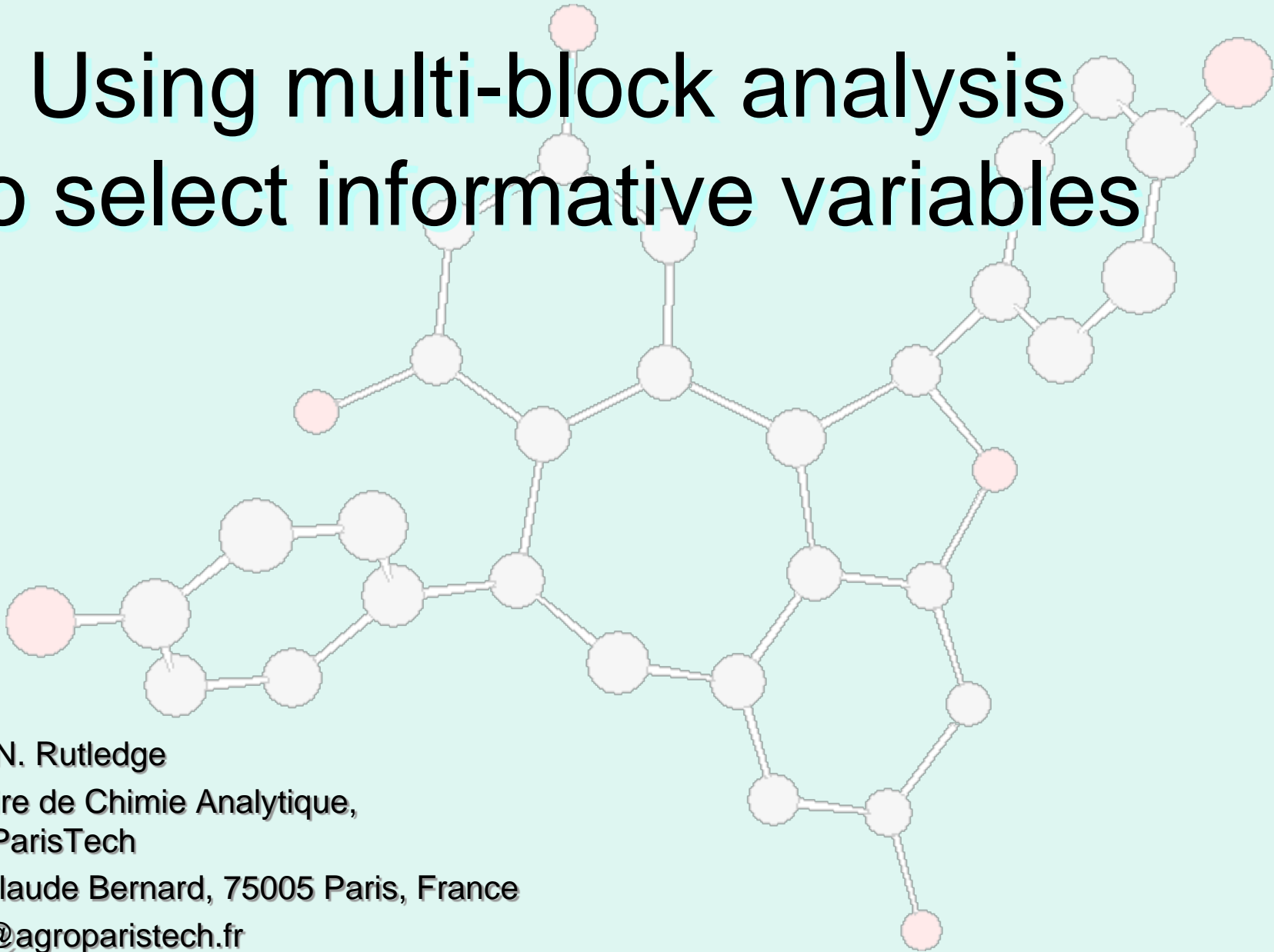


Using multi-block analysis to select informative variables



Douglas N. Rutledge

Laboratoire de Chimie Analytique,
AgroParisTech

16, rue Claude Bernard, 75005 Paris, France

rutledge@agroparistech.fr



Variable Selection

The quality of multivariate predictive models is increased by eliminating uninformative variables.

For discriminant models, pp-ANOVA is often used :

- test each variable separately
- varies more between groups than within groups ?

For regression analysis, many methods :

- Uninformative Variable Elimination-PLS [1]
- Genetic Algorithm-PLS [2]
- iPLS [3] ...

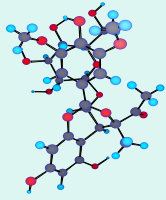
[1] V. Centner, D. L. Massart, O. E. deNoord, S. deJong, B. M. Vandeginste, C. Sterna, Elimination of uninformative variables for multivariate calibration.

Analytical Chemistry 1996, 68, 3851-3858.

[2] A.S. Bangalore, R.E. Shaffer, G.W. Small, M.A. Arnold, Genetic algorithm-based method for selecting wavelengths and model size for use with partial least-squares regression: Application to near-infrared spectroscopy. *Analytical Chemistry* 1996, 68, 4200-4212.

[3] L. Norgaard, A. Saudland, J. Wagner, J.P. Nielsen, L. Munck, S.B. Engelsen, Interval partial least-squares regression (iPLS): A comparative chemometric study with an example from near-infrared spectroscopy.

Applied Spectroscopy 2000, 54, 413-419.



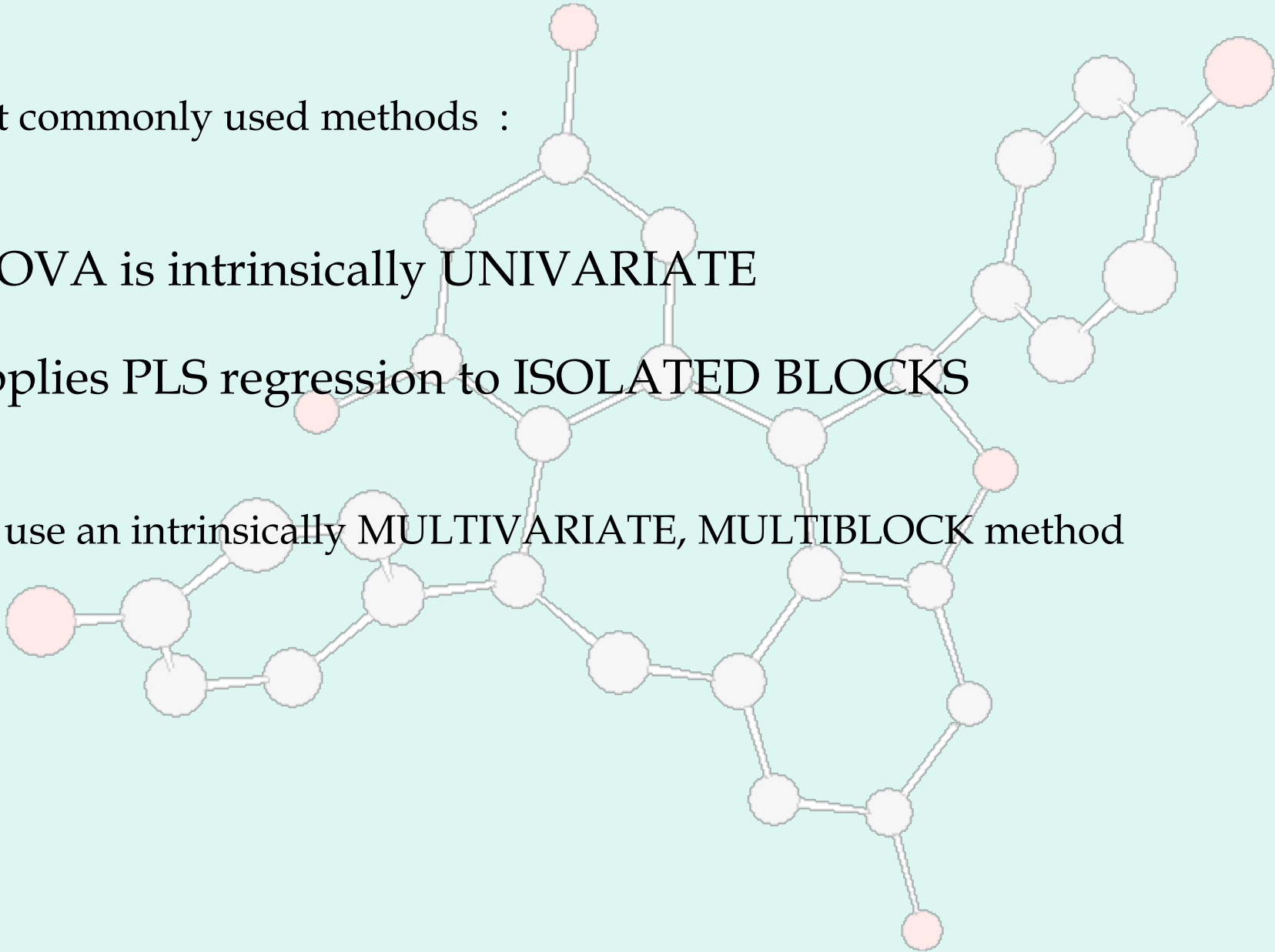
pp-ANOVA and iPLS

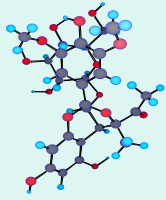
The most commonly used methods :

pp-ANOVA is intrinsically UNIVARIATE

iPLS applies PLS regression to ISOLATED BLOCKS

Better to use an intrinsically MULTIVARIATE, MULTIBLOCK method





Multi-block analysis

"Common Components and Specific Weights Analysis" - CCSWA [4]

Simultaneously study several matrices

- with different variables describing the same samples

Describe m data tables observed for the same n samples :

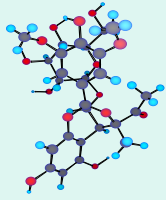
- a set of m data matrices (\mathbf{X}) each with n rows,
- but not necessarily the same number columns

Determine a common space for all m data table,

- each matrix has a specific contribution ("salience") to the definition of each dimension of this common space

[4] E. Qannari, I. Wakeling, P. Courcoux, H.J.H MacFie,

Defining the underlying sensory dimensions. *Food Quality and Preference* 2000, 11, 151-154.



Multi-block analysis

Start with p matrices \mathbf{X}_i of size $n \times k_i$ ($i = 1$ to p)

Each \mathbf{X}_i column-centered and scaled by dividing by matrix norm :

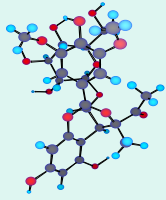
$$\mathbf{X}_{si}$$

For each \mathbf{X}_{si} , an $n \times n$ scalar product matrix \mathbf{W}_i can be computed as :

$$\mathbf{W}_i = \mathbf{X}_{si} \cdot \mathbf{X}_{si}^T$$

\mathbf{W}_i reflect the dispersion of the *samples* in the space of that table

The common dimensions of all the tables are computed iteratively
At each iteration, a weighted sum of the p \mathbf{W}_i matrices is computed,
resulting in a global \mathbf{W}_G matrix



Multi-block analysis

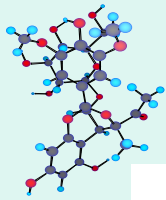
For each successive Common Dimension, calculate a scores vector q (coordinates of the n samples along the common dimension)

$$W_i = \sum_{j=1}^{j=n} \lambda_j^{(i)} q_j q_j^T$$

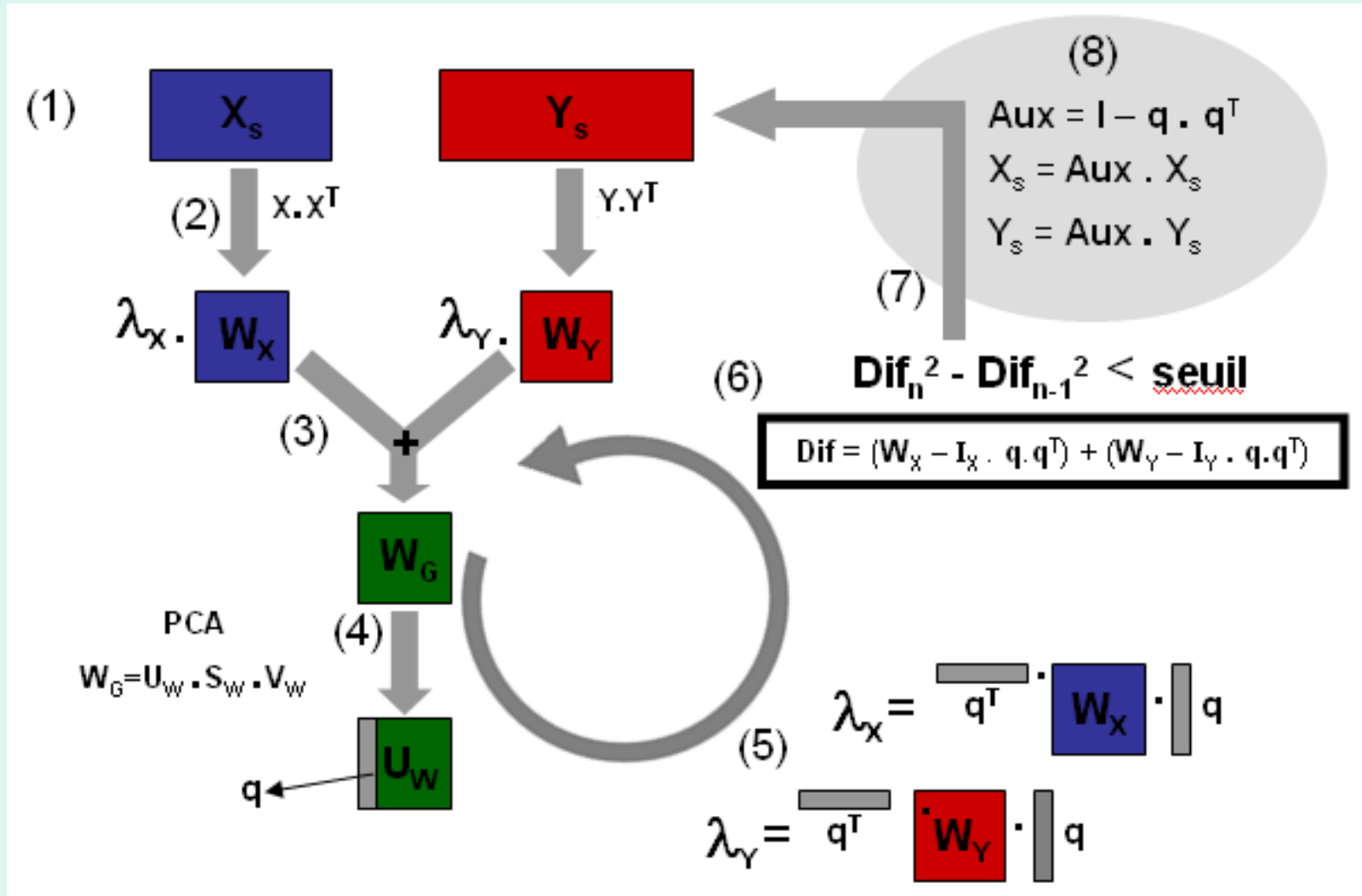
$\lambda_j^{(i)}$ is the *specific weight* ("salience") associated with the i^{th} table for the j^{th} Common Dimension generated by q_j

Differences in the values of the *specific weights* for a dimension :
- information present in some tables but not others

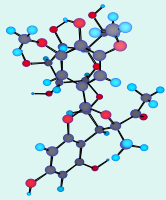
Subsequent components calculated after deflating the data tables



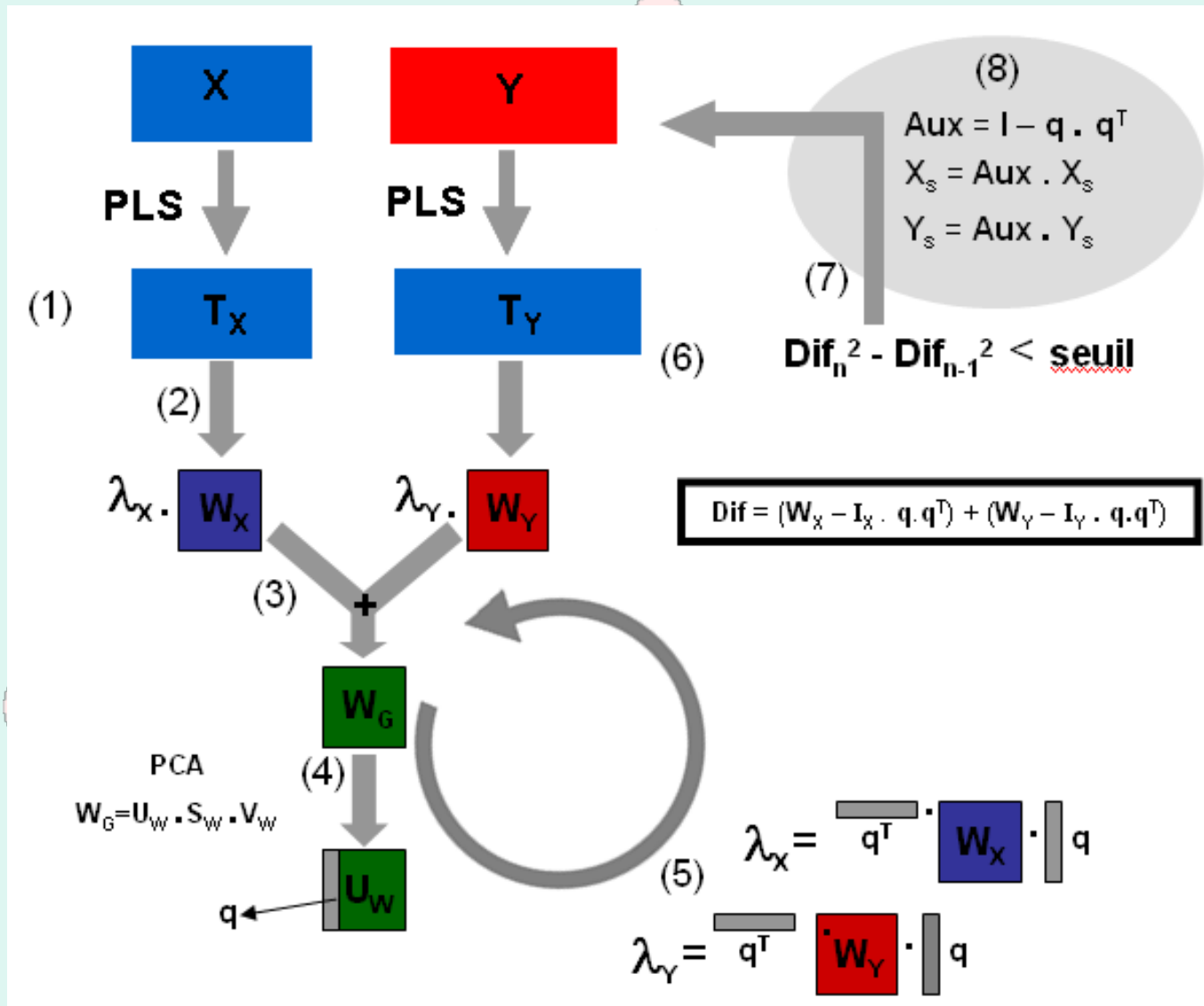
Classical ComDim

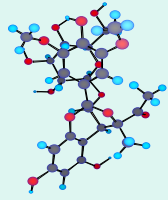


"ComDim" the implementation of CCSWA used here is part of the SAISIR toolbox
 SAISIR (2008): Statistics Applied to the Interpretation of Spectra in the InfraRed
 Dominique Bertrand (bertrand@nantes.inra.fr)



PLS-ComDim





1) Starch-Lignin mixtures

Samples

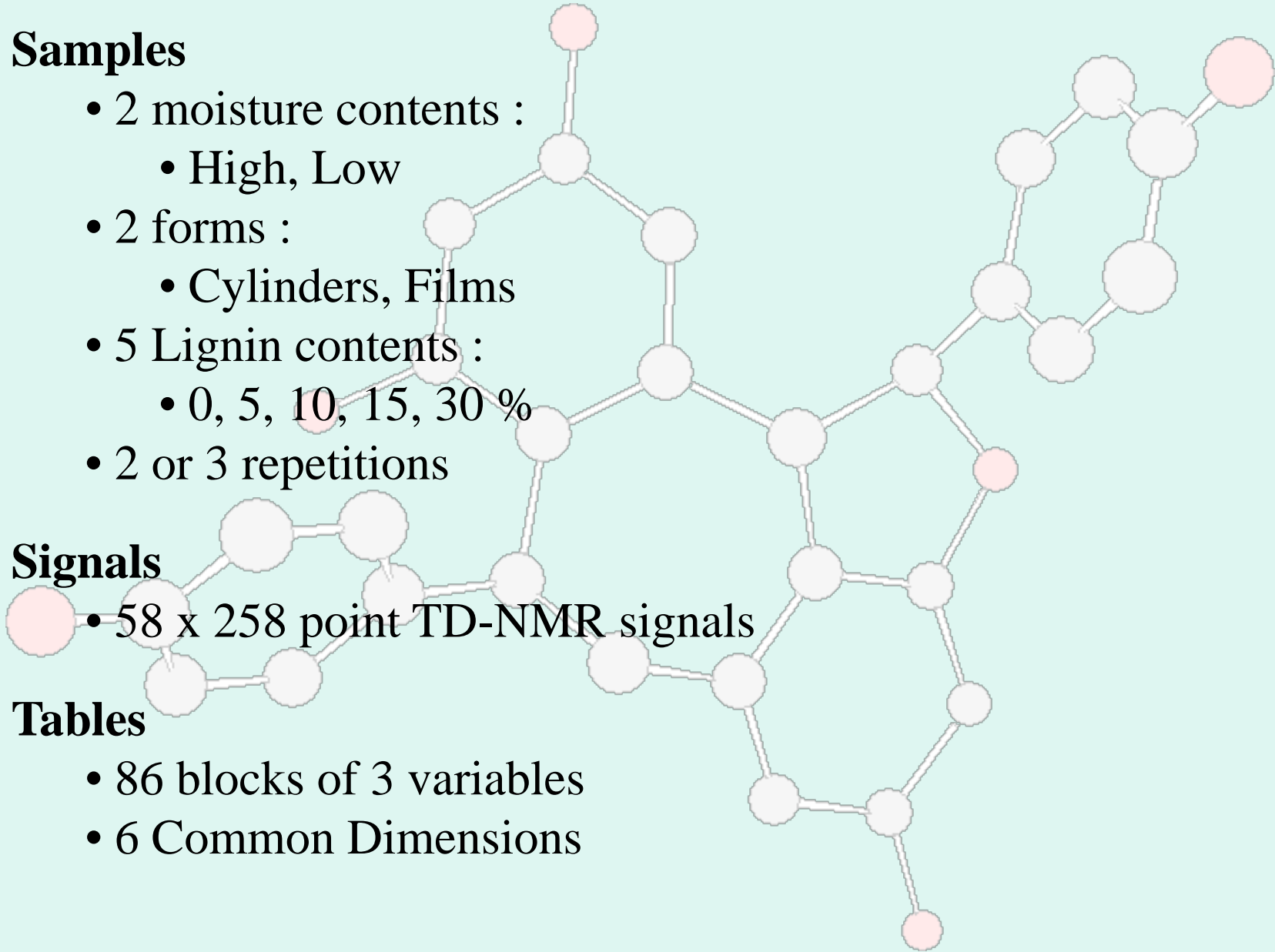
- 2 moisture contents :
 - High, Low
- 2 forms :
 - Cylinders, Films
- 5 Lignin contents :
 - 0, 5, 10, 15, 30 %
- 2 or 3 repetitions

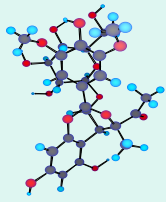
Signals

- 58 x 258 point TD-NMR signals

Tables

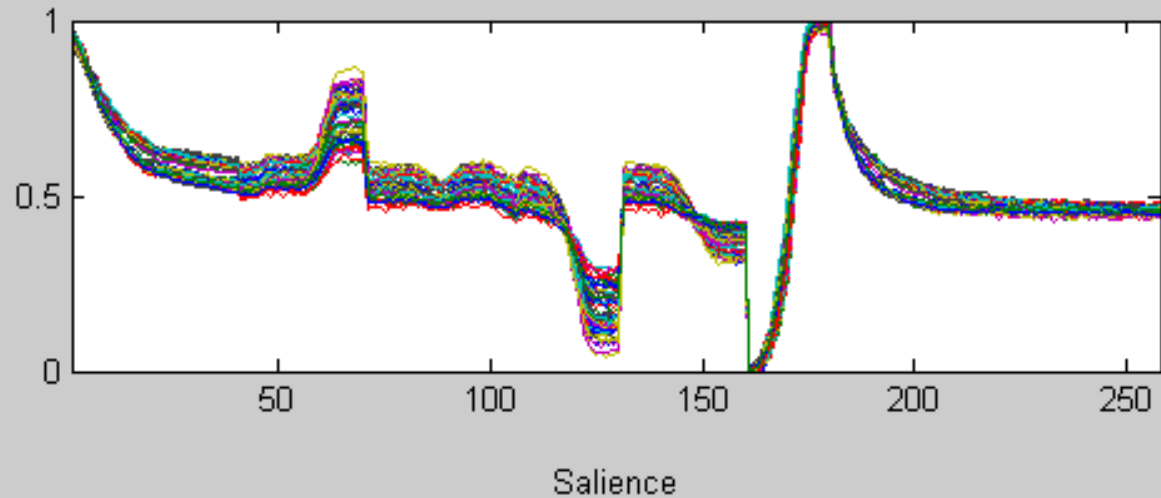
- 86 blocks of 3 variables
- 6 Common Dimensions



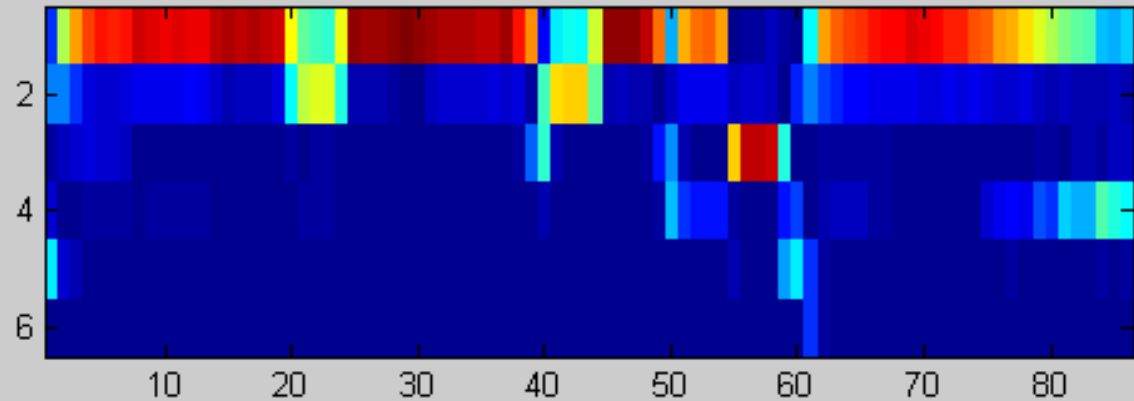


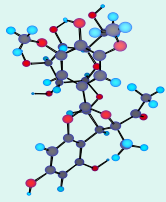
TD-NMR signals

Spectra

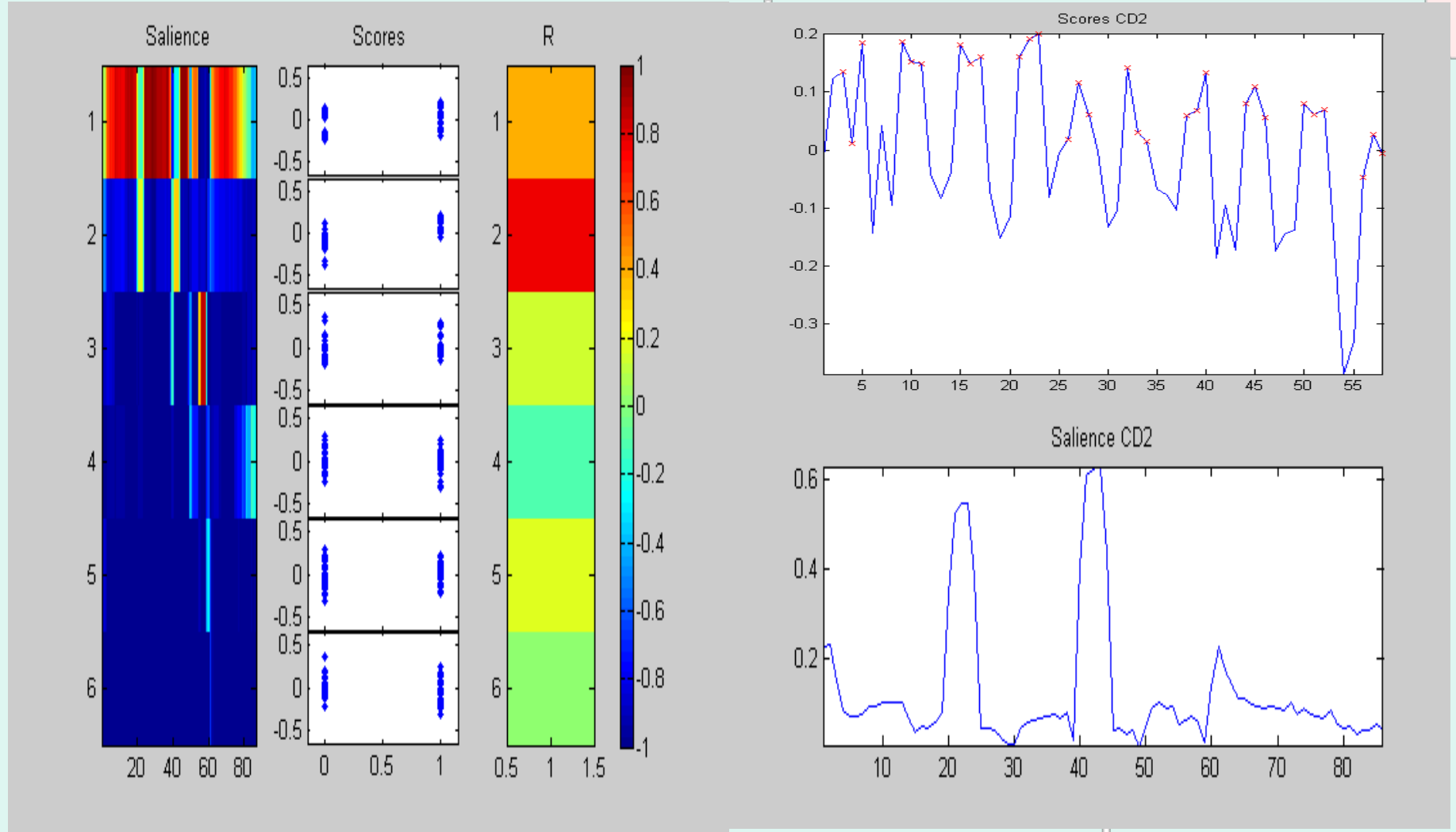


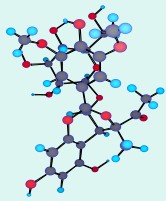
ComDim
Saliences



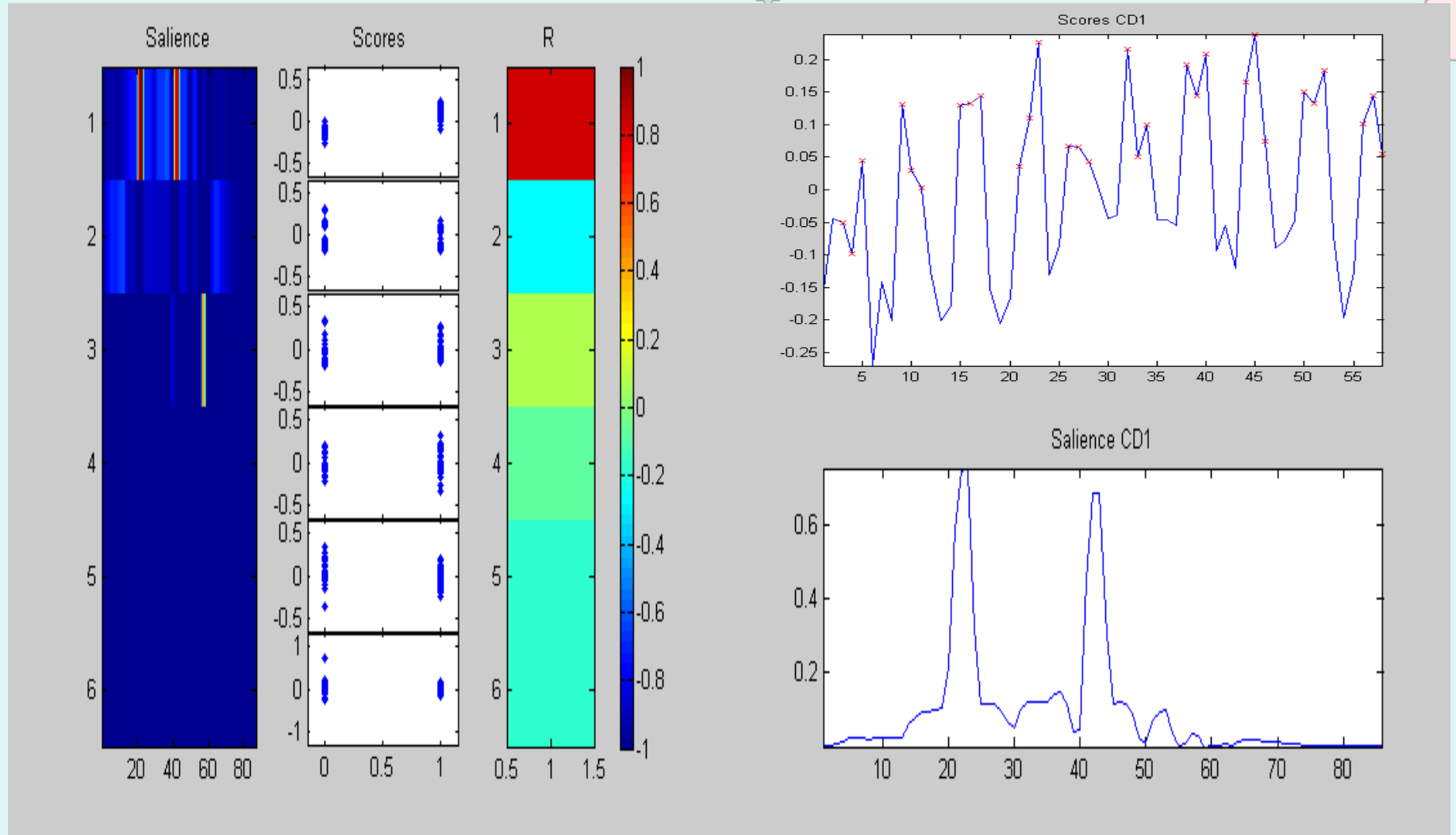


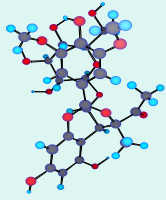
Correlation between ComDim Scores and "Form"





Correlation between PLS-ComDim Scores and "Form"

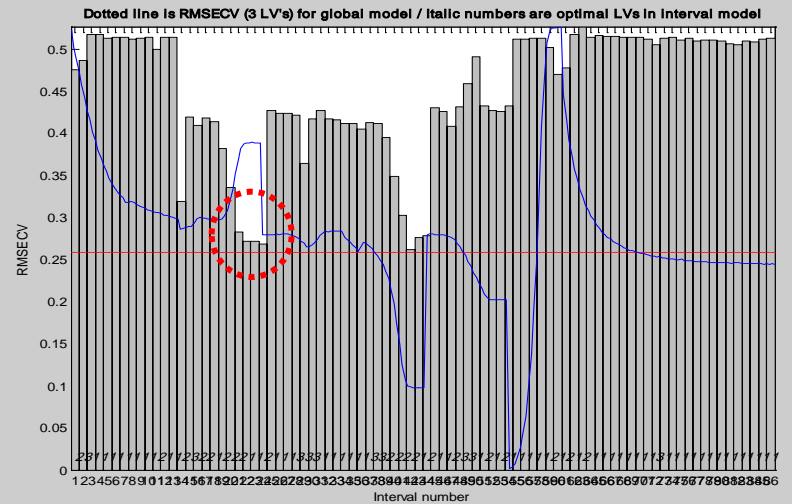
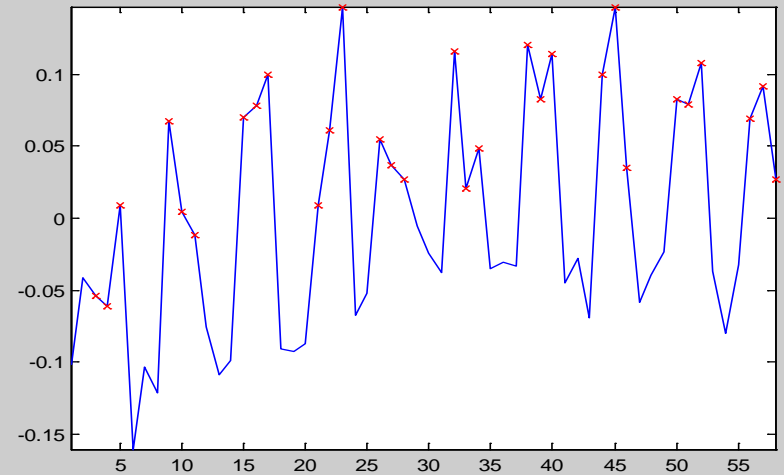
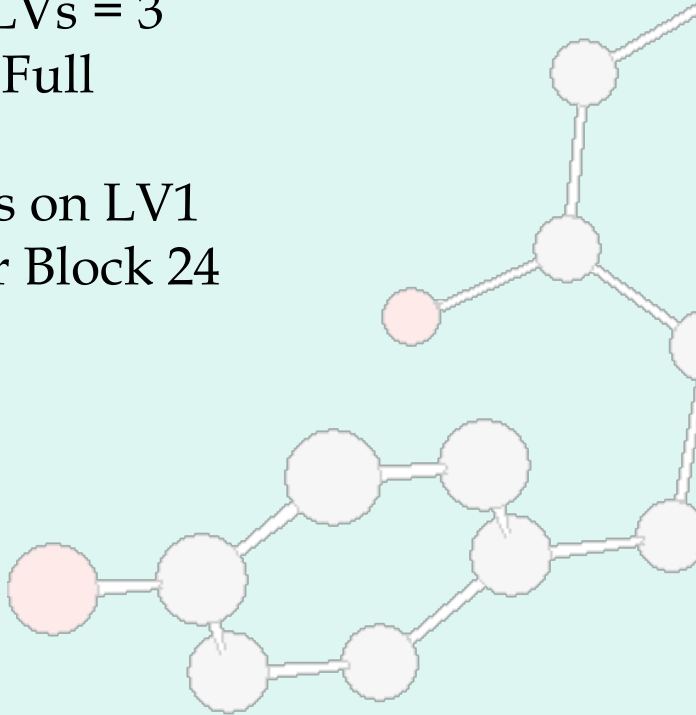


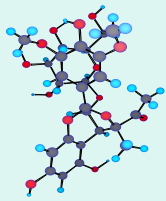


i-PLS between NIR and "Form"

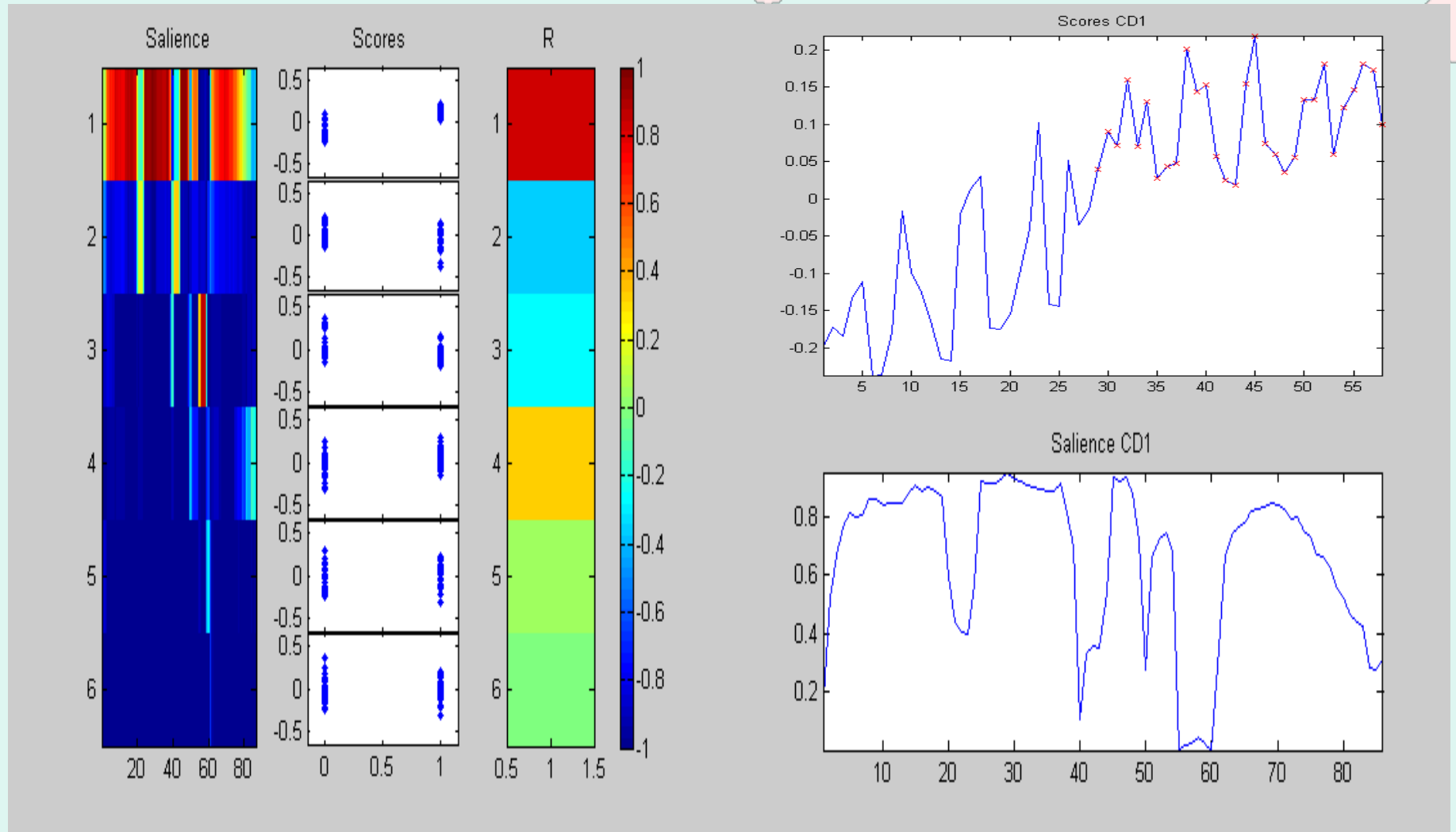
- Blocks = 86
- Mean centred
- Max LVs = 3
- CV = Full

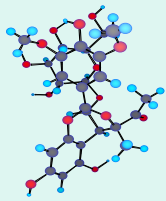
- Scores on LV1
for Block 24



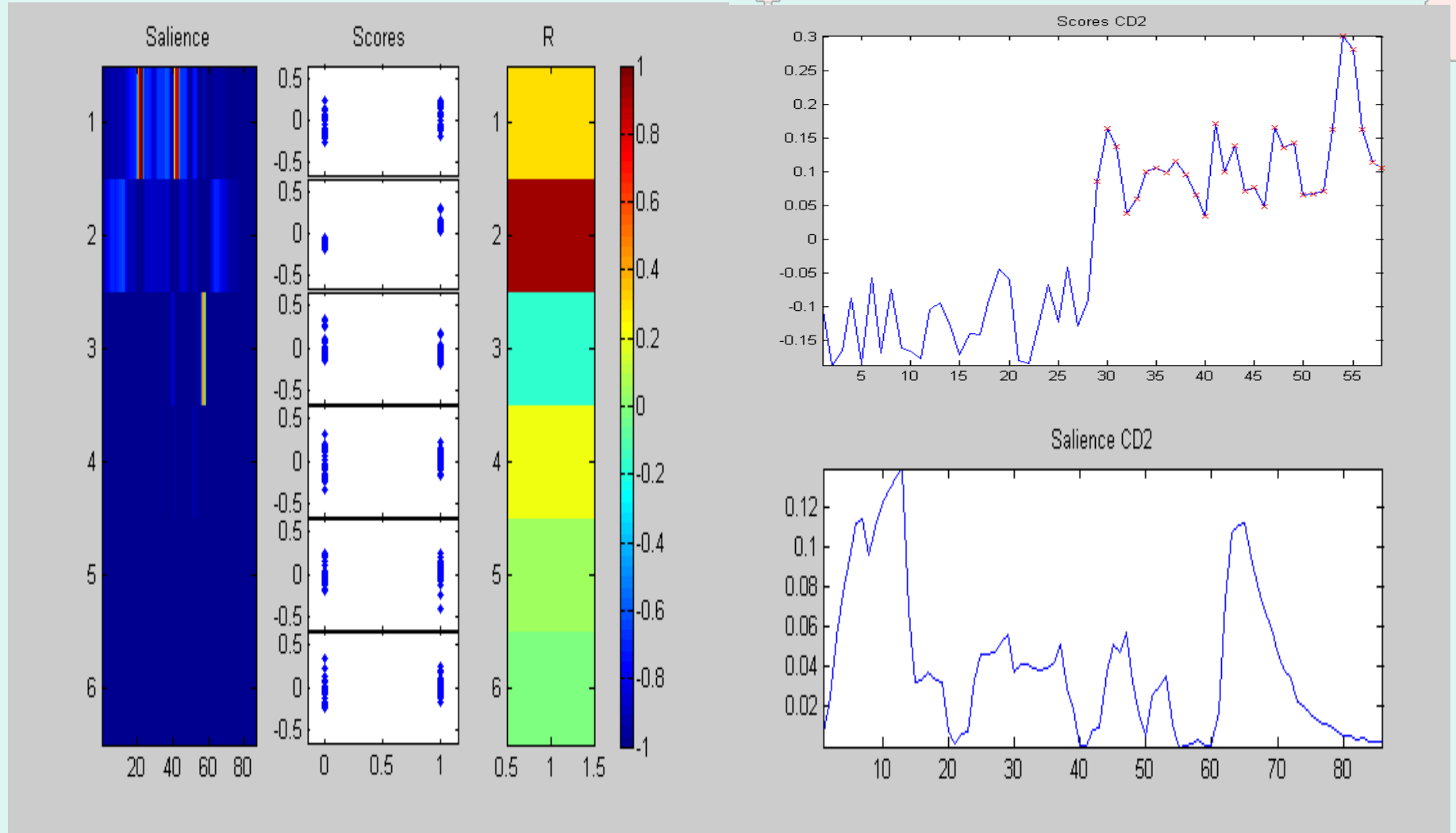


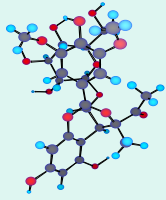
Correlation between ComDim Scores and "H2O"





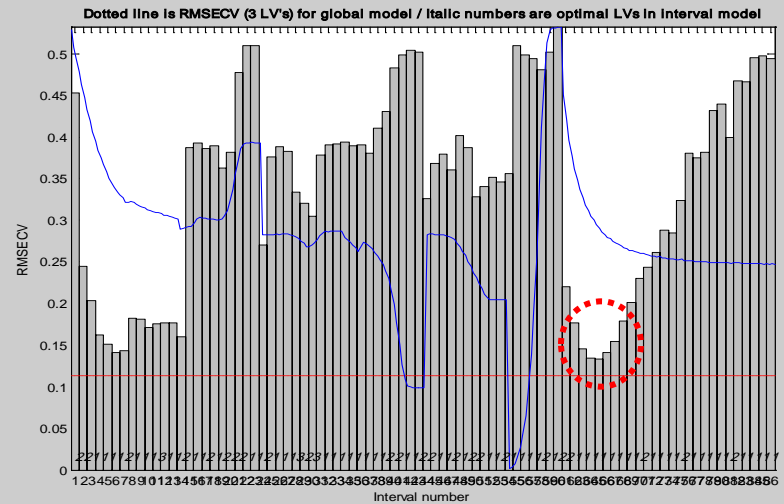
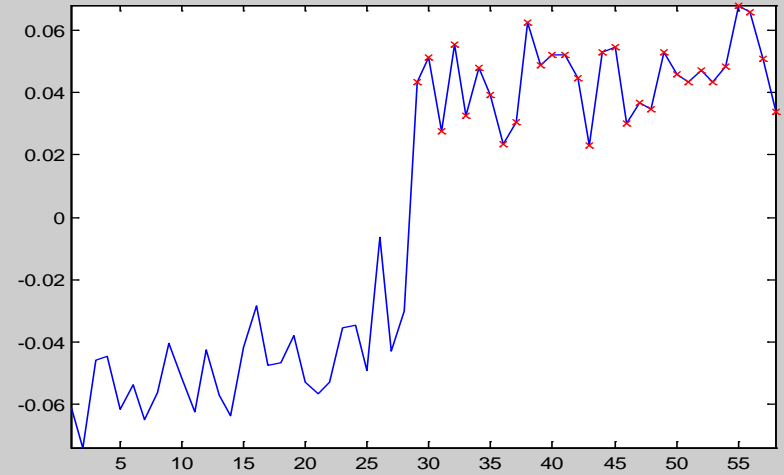
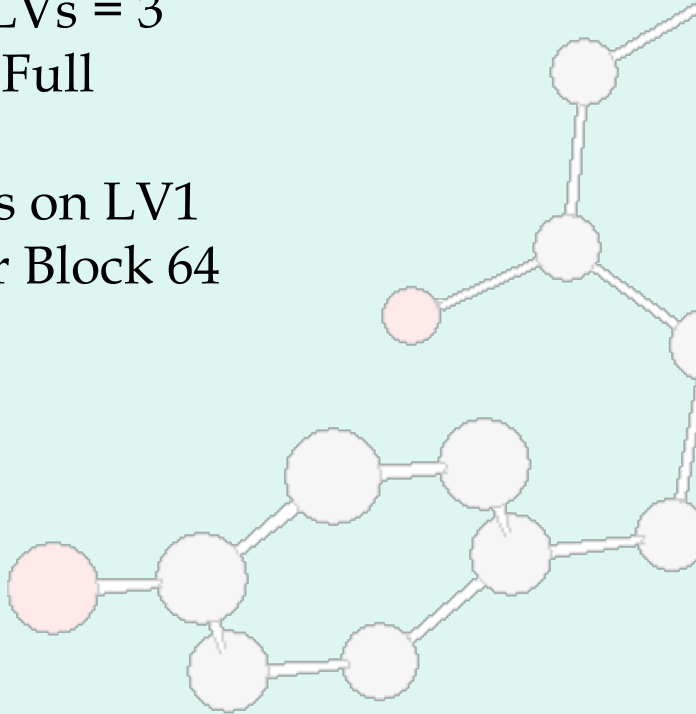
Correlation between PLS-ComDim Scores and "H2O"

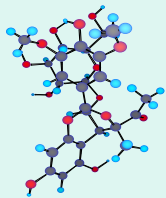




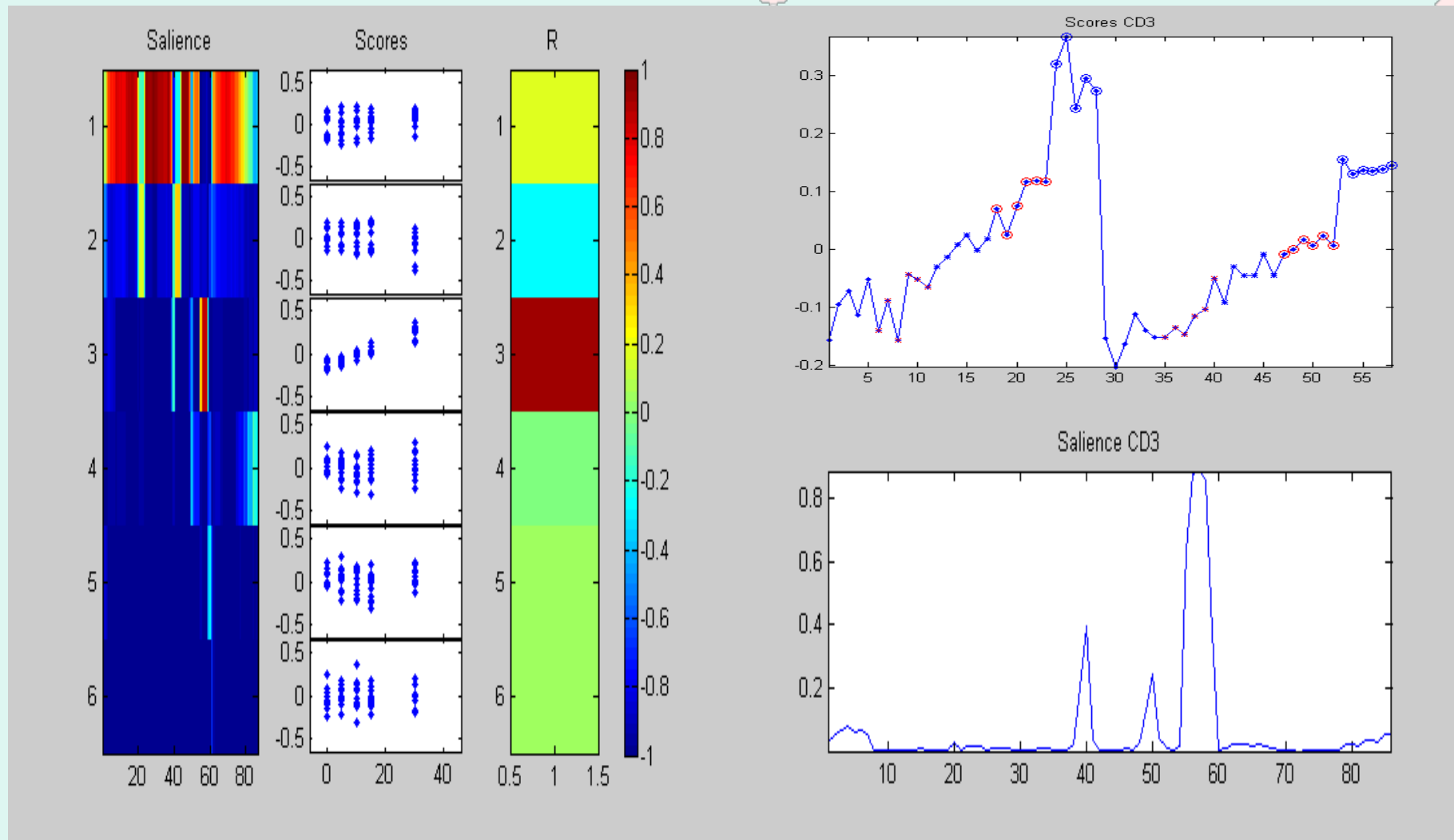
i-PLS between NIR and "H2O"

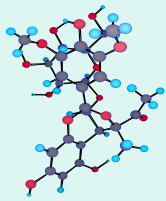
- Blocks = 86
 - Mean centred
 - Max LVs = 3
 - CV = Full
- Scores on LV1
for Block 64



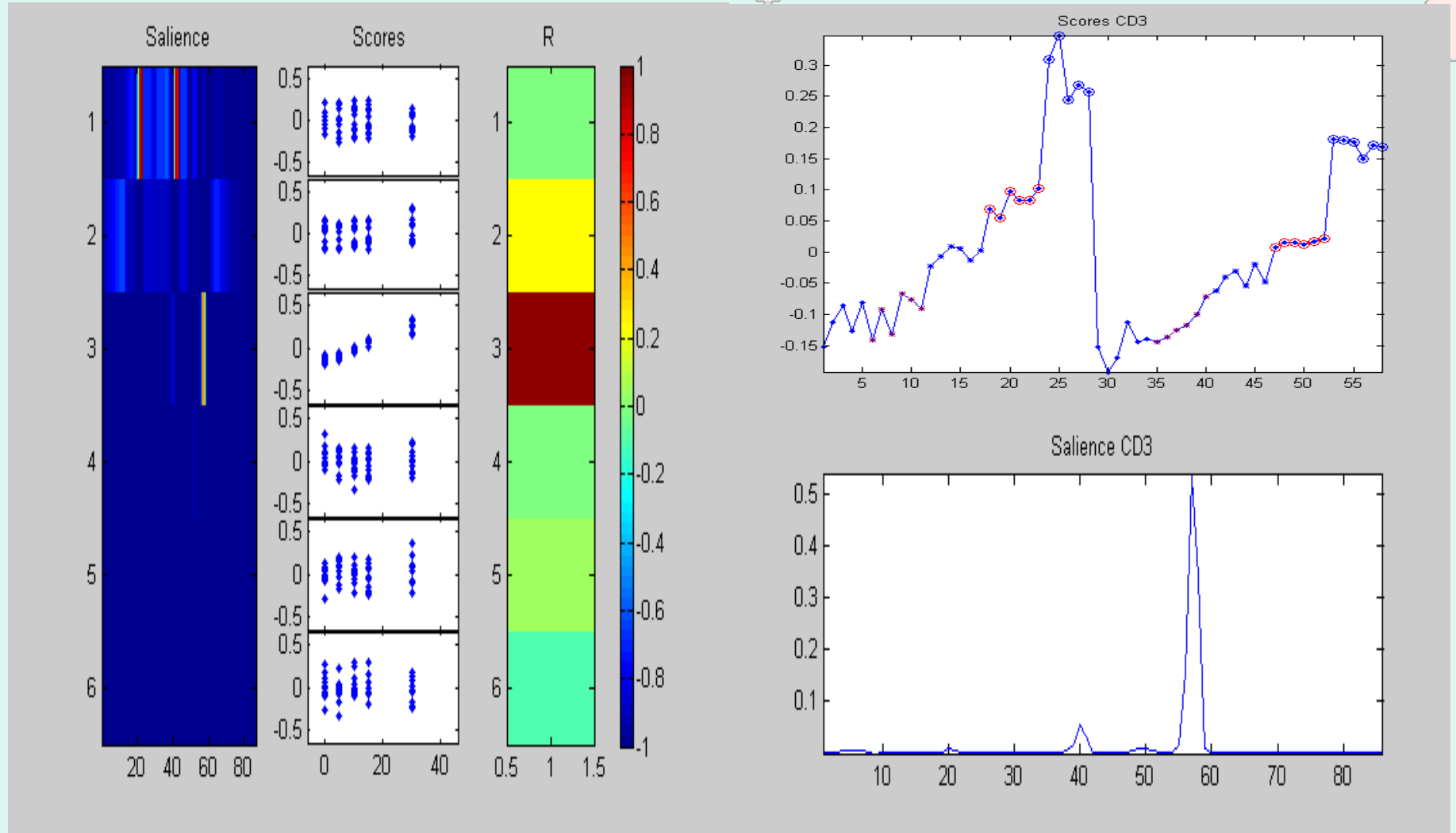


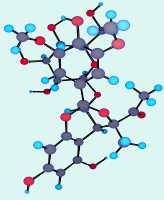
Correlation between ComDim Scores and "Lignin"





Correlation between PLS-ComDim Scores and "Lignin"

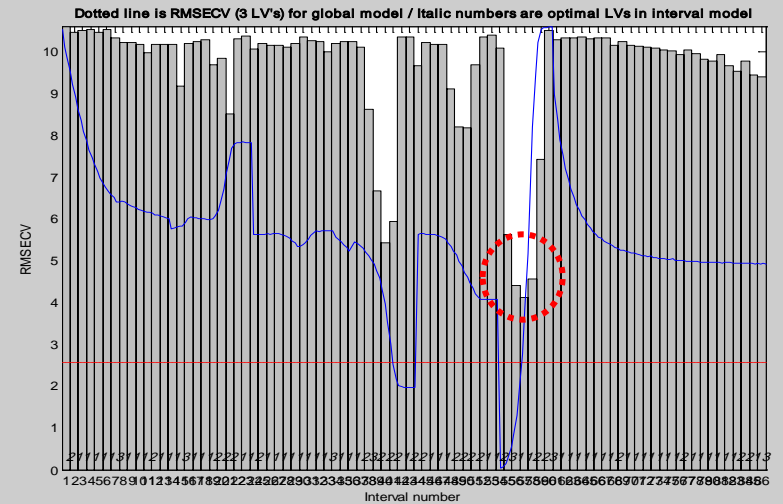
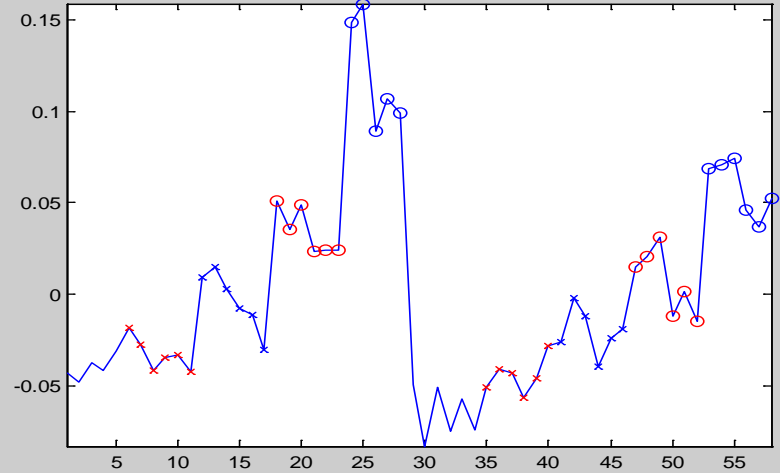
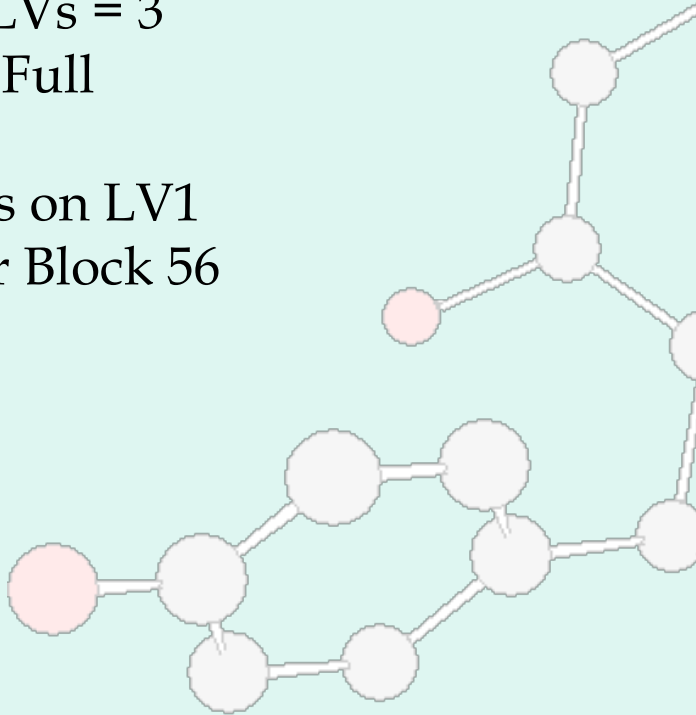


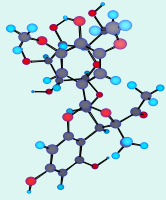


i-PLS between NIR and "Lignin"

- Blocks = 86
- Mean centred
- Max LVs = 3
- CV = Full

- Scores on LV1
for Block 56





2) NIR on apples

Samples

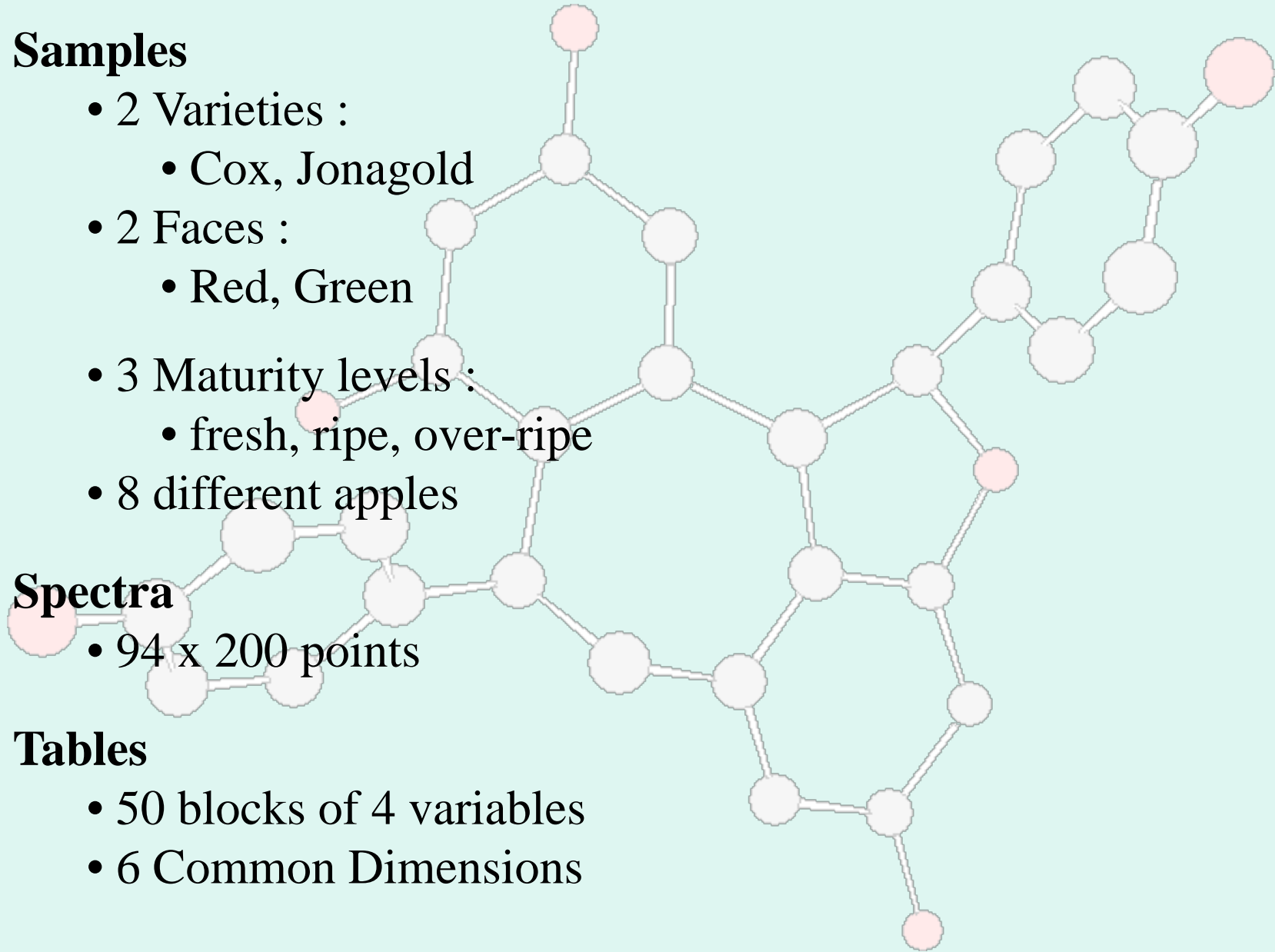
- 2 Varieties :
 - Cox, Jonagold
- 2 Faces :
 - Red, Green
- 3 Maturity levels :
 - fresh, ripe, over-ripe
- 8 different apples

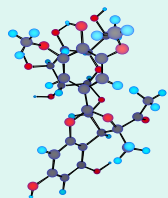
Spectra

- 94 x 200 points

Tables

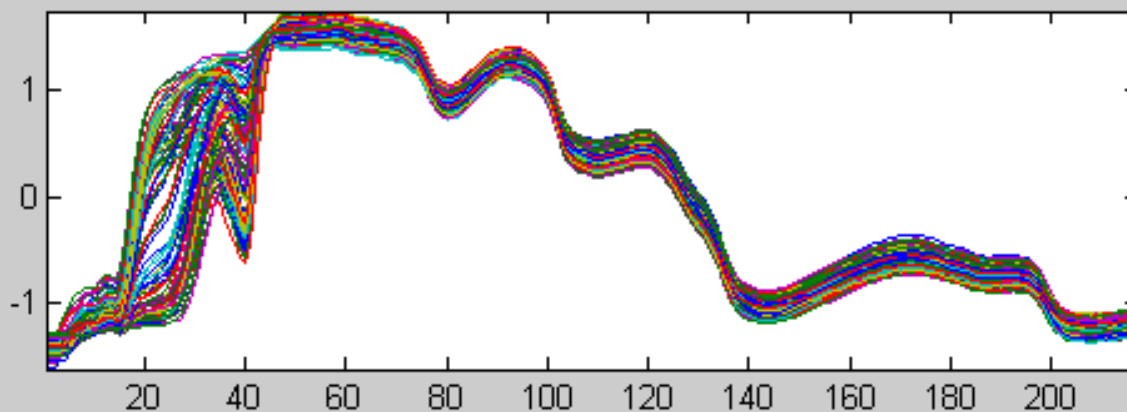
- 50 blocks of 4 variables
- 6 Common Dimensions



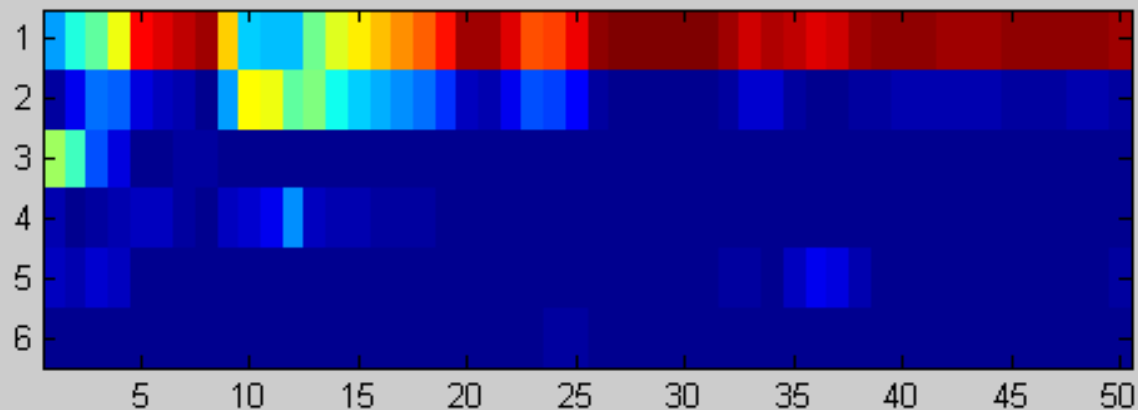


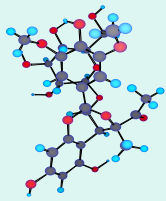
NIR Spectra

Spectra

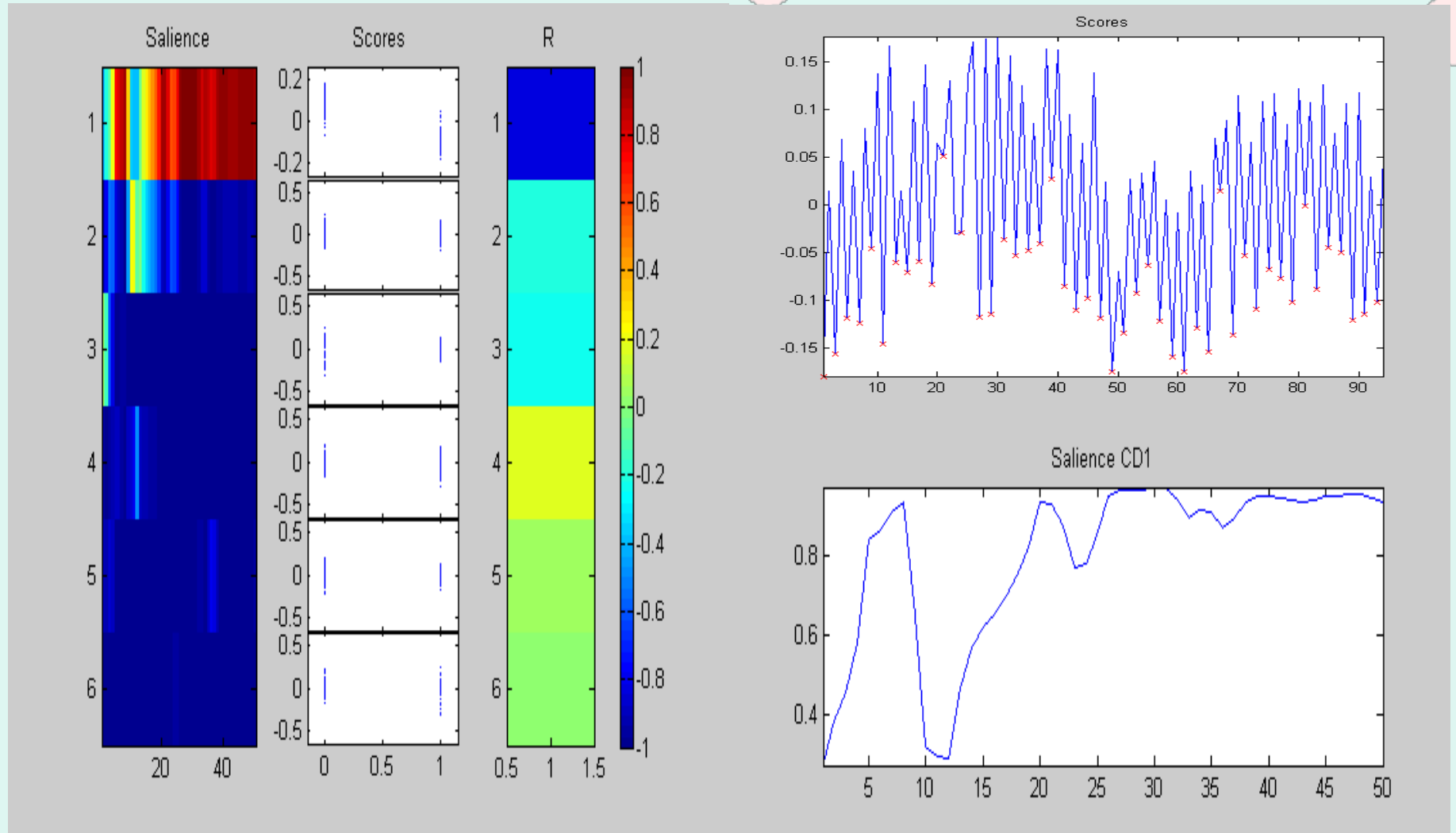


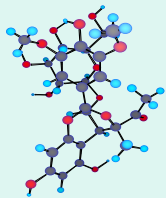
ComDim
Saliences



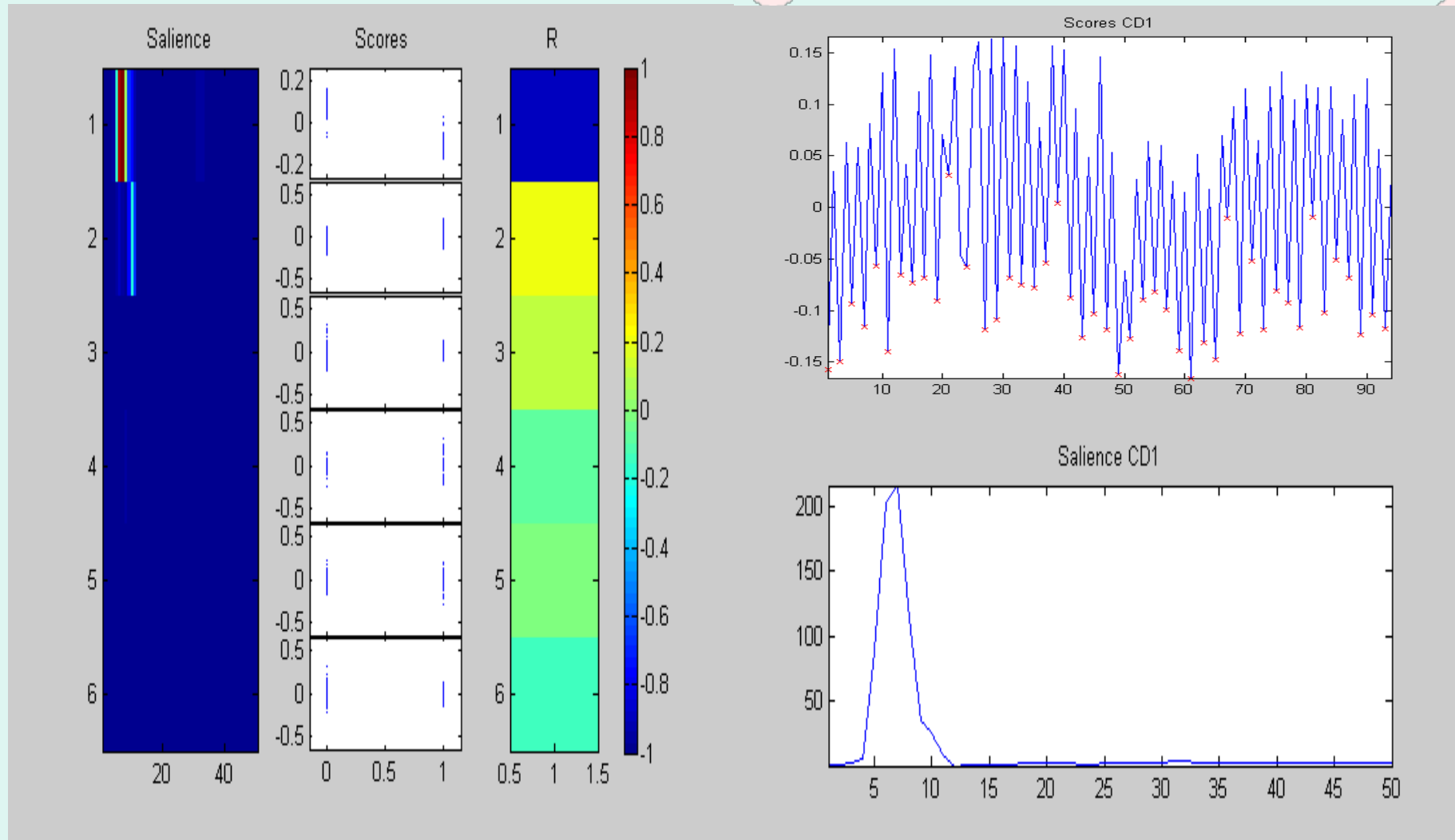


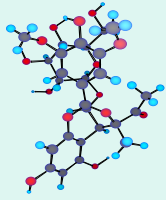
Correlation between ComDim Scores and "Face"





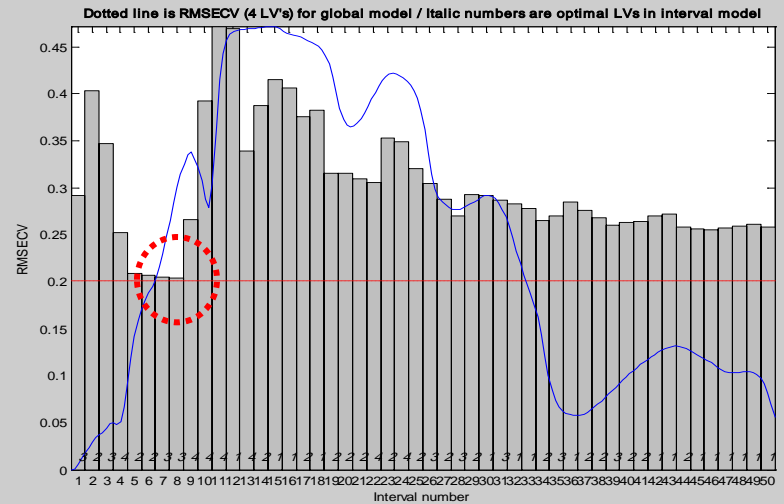
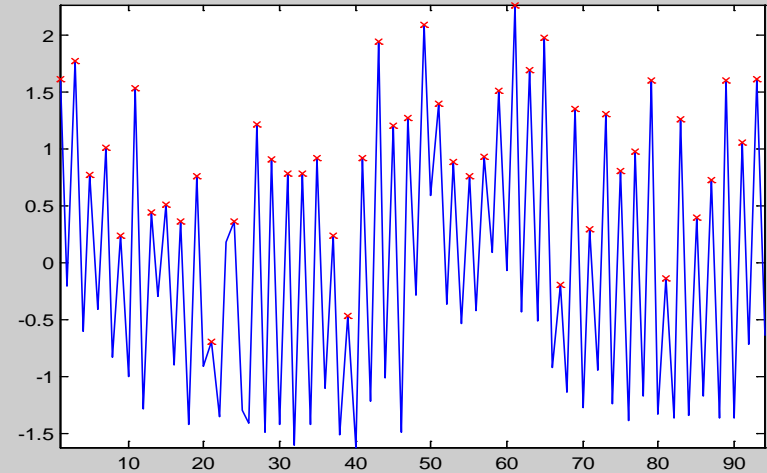
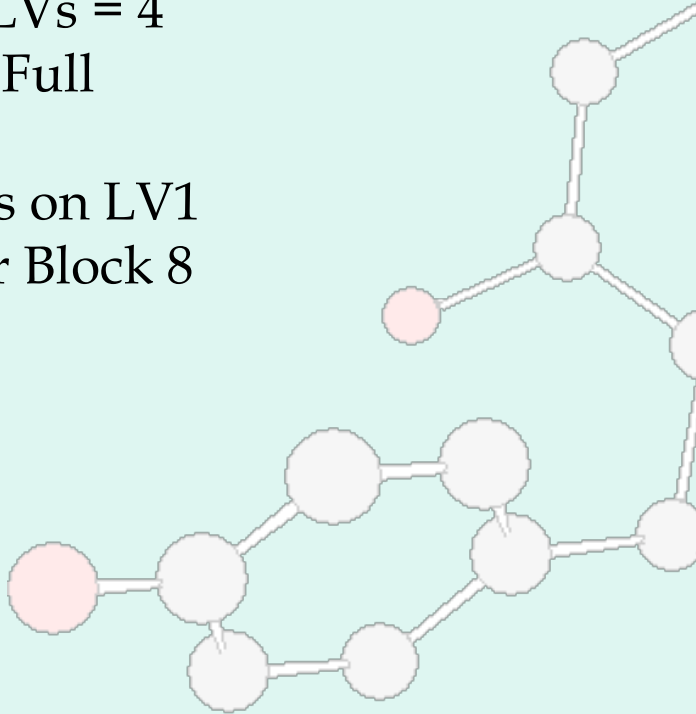
Correlation between PLS-ComDim Scores and "Face"

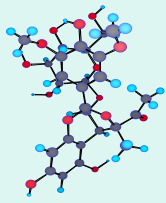




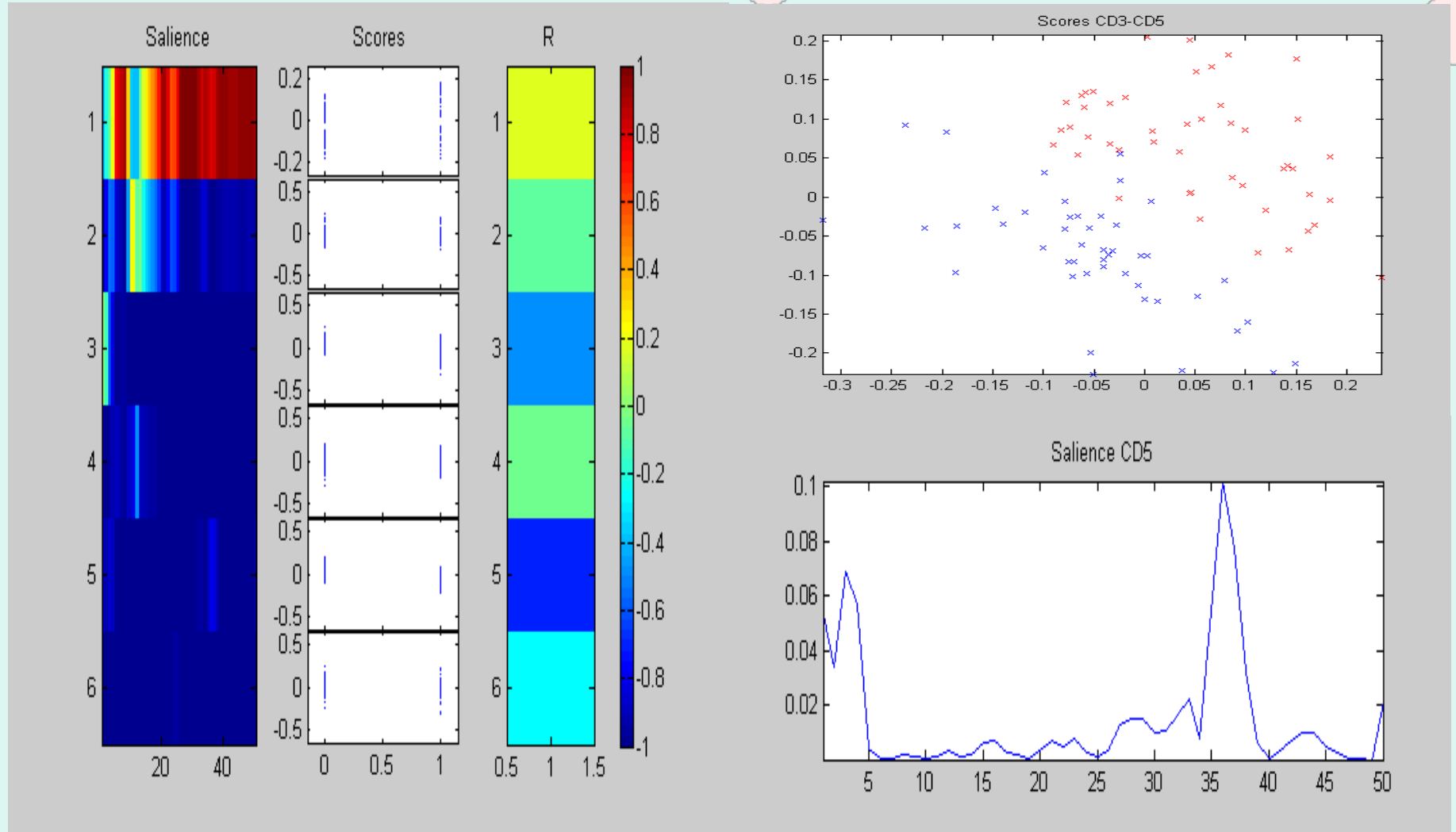
i-PLS between NIR and "Face"

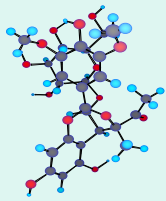
- Blocks = 50
 - Mean centred
 - Max LVs = 4
 - CV = Full
- Scores on LV1
for Block 8



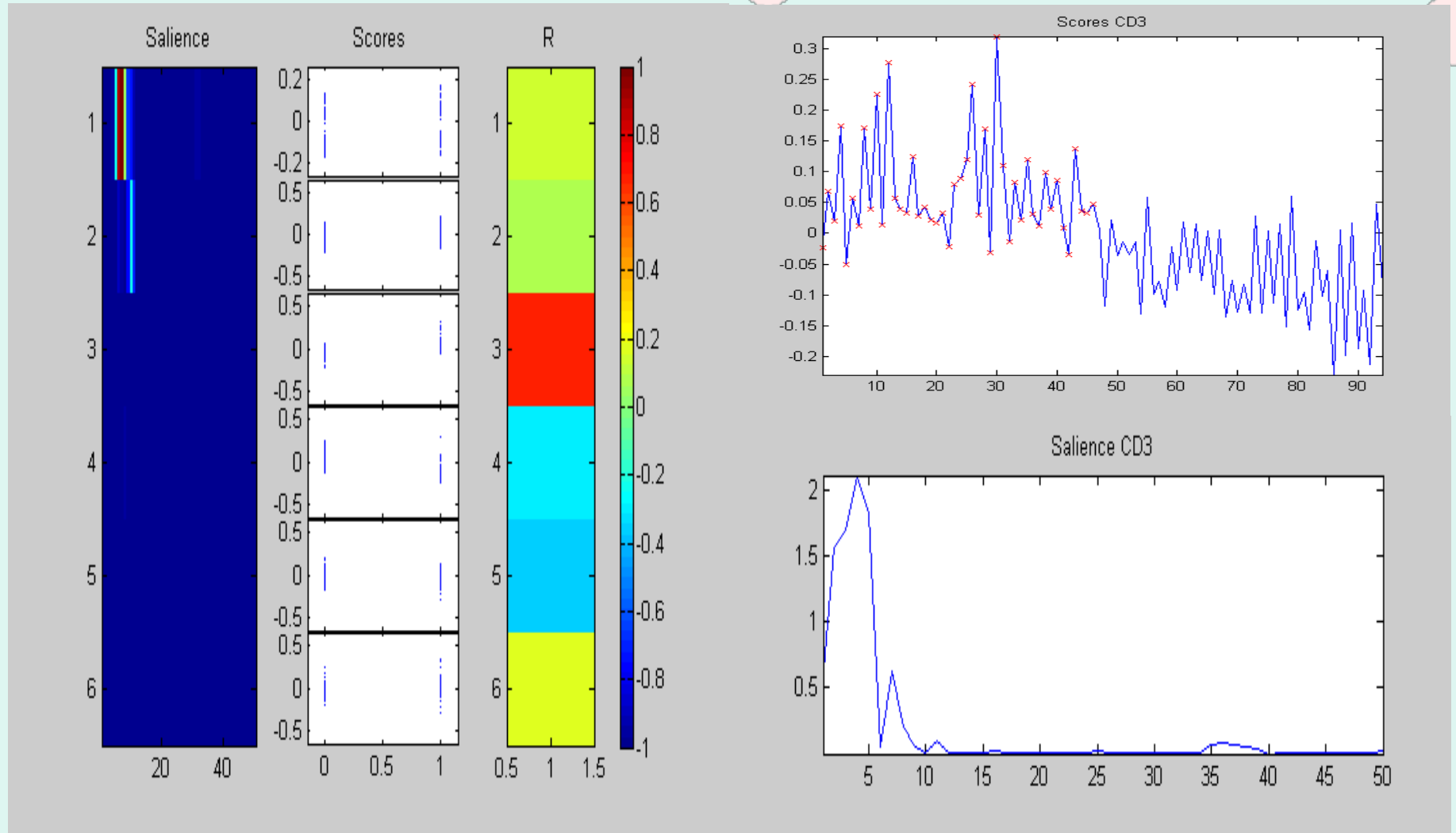


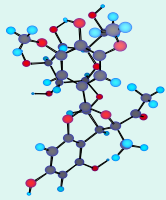
Correlation between ComDim Scores and "Variety"





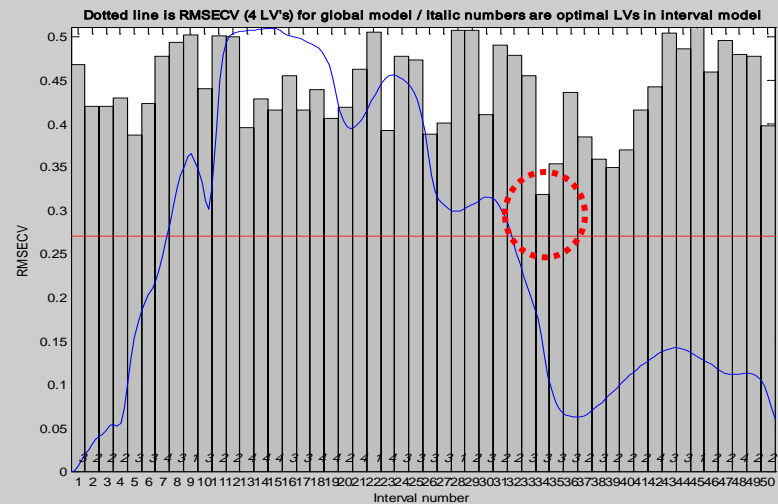
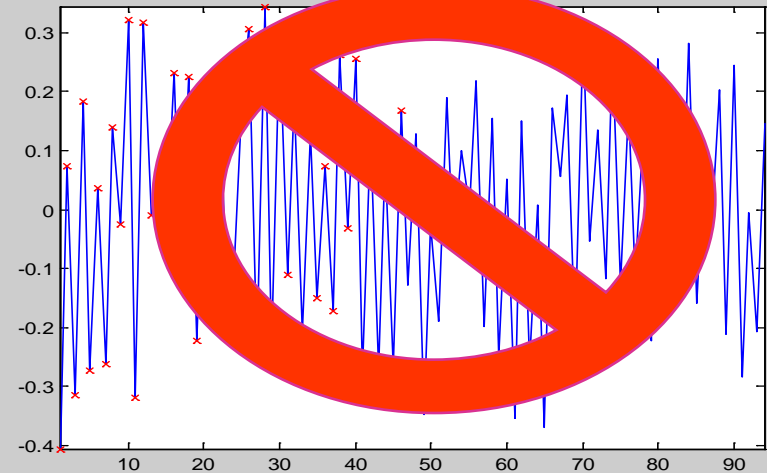
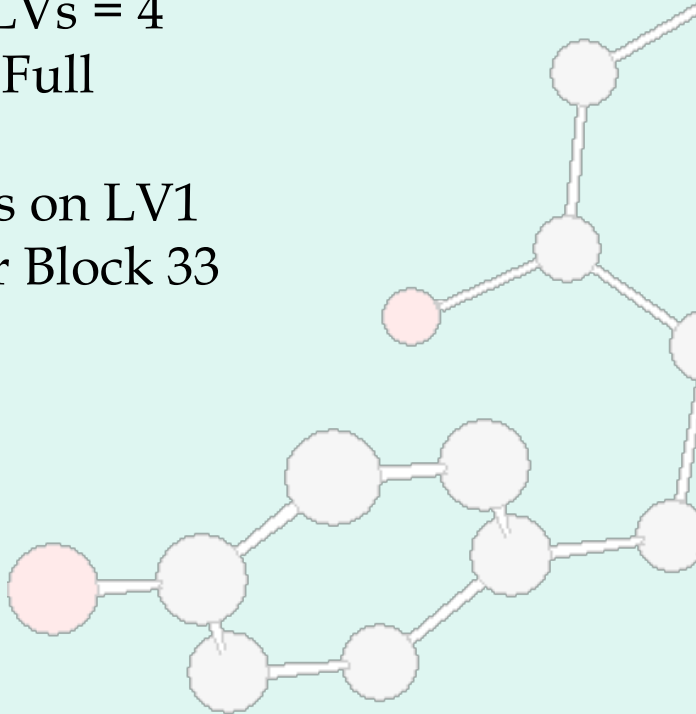
Correlation between PLS-ComDim Scores and "Variety"

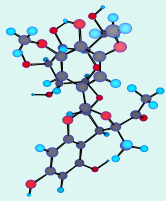




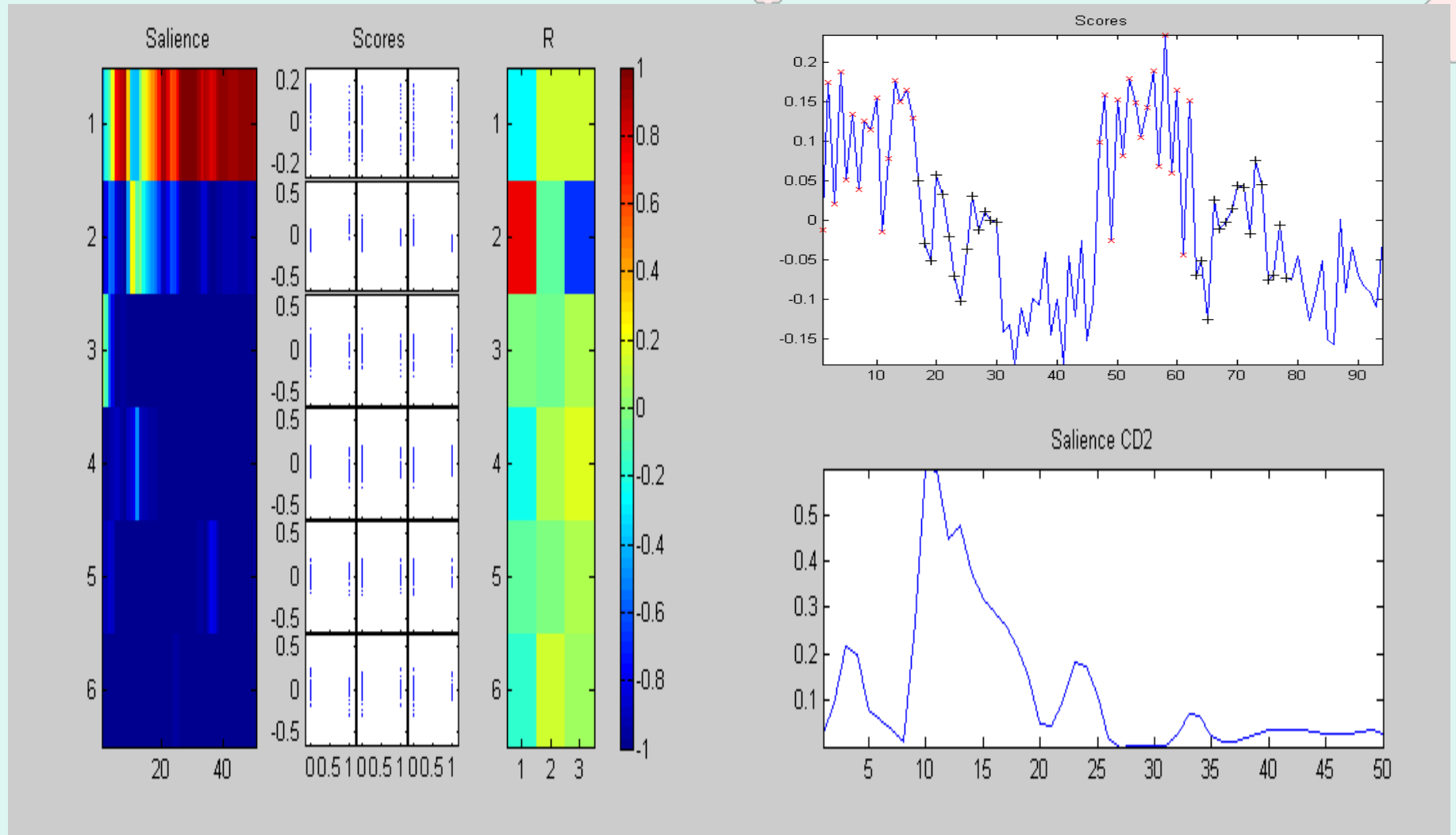
i-PLS between NIR and "Variety"

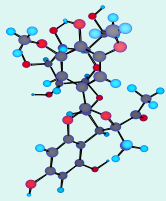
- Blocks = 50
 - Mean centred
 - Max LVs = 4
 - CV = Full
- Scores on LV1
for Block 33



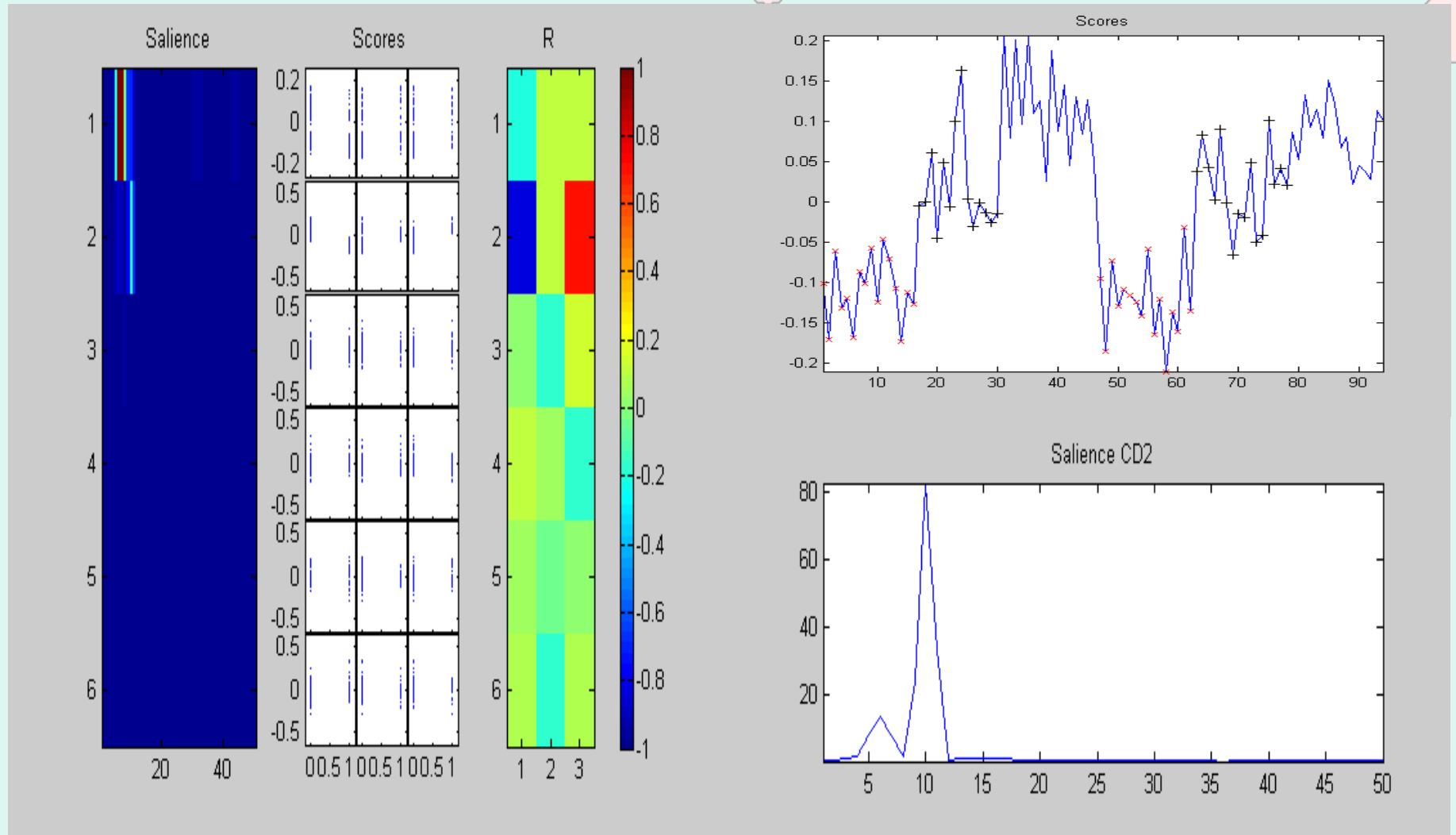


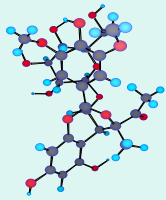
Correlation between ComDim Scores and "Maturity"





Correlation between PLS-ComDim Scores and "Maturity"

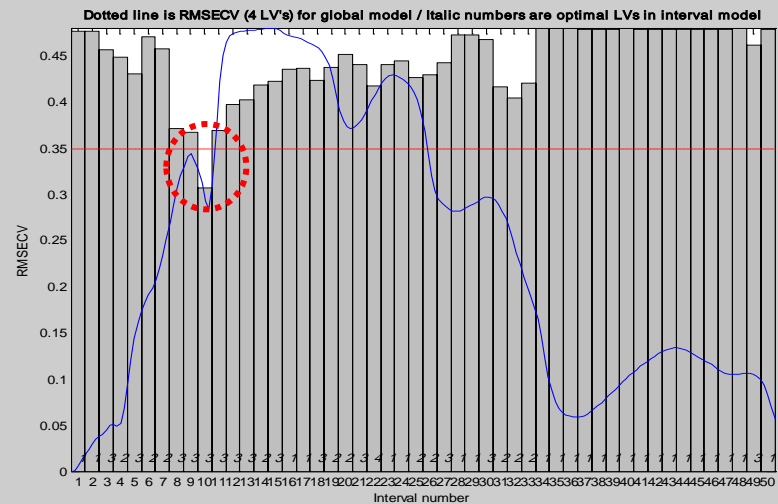
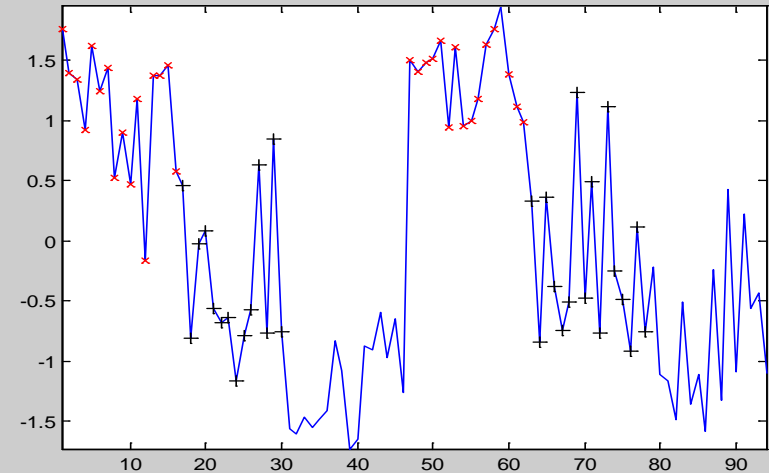
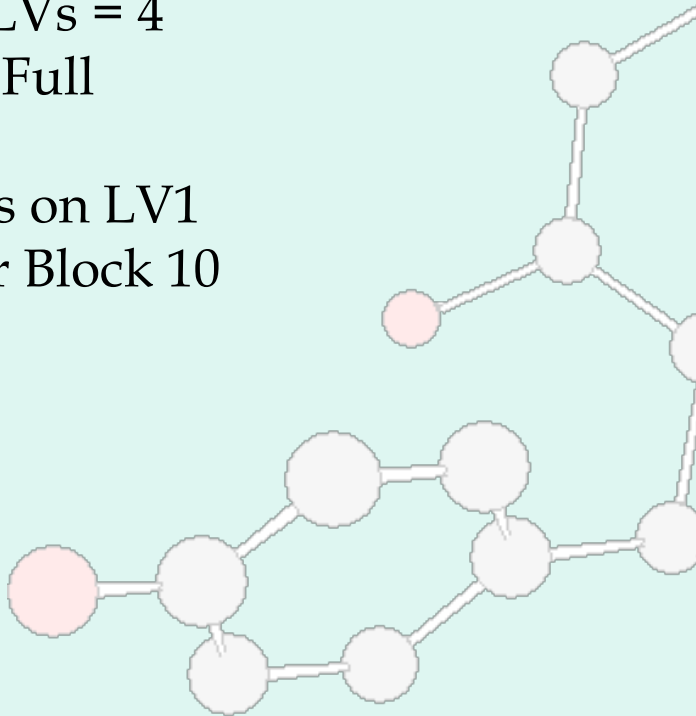


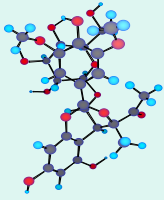


i-PLS between NIR and "Maturity"

- Blocks = 50
- Mean centred
- Max LVs = 4
- CV = Full

- Scores on LV1
for Block 10





3) 2D-Fluorescence on wines

Samples

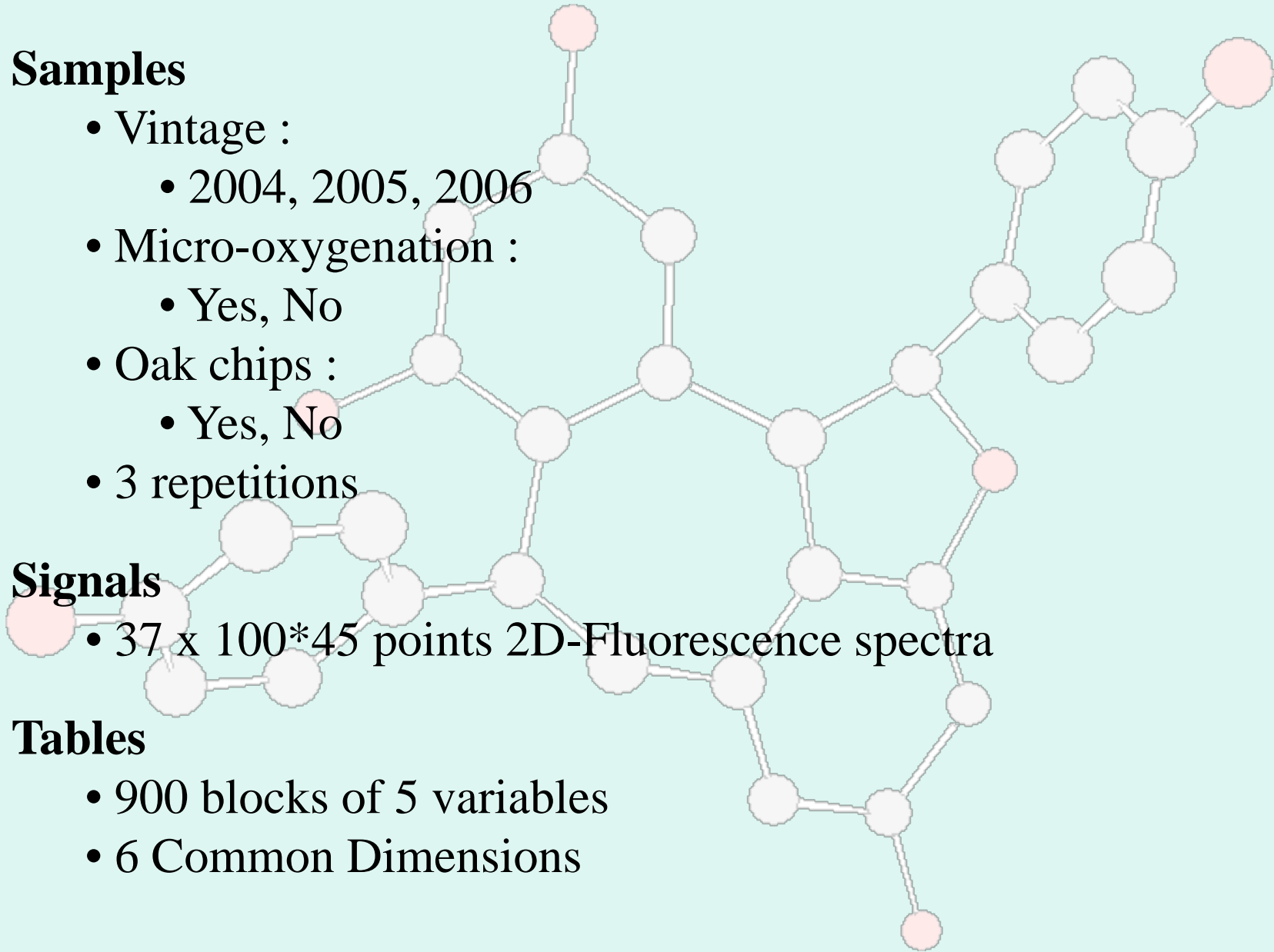
- Vintage :
 - 2004, 2005, 2006
- Micro-oxygenation :
 - Yes, No
- Oak chips :
 - Yes, No
- 3 repetitions

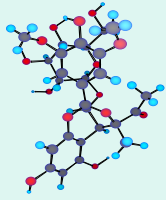
Signals

- 37 x 100*45 points 2D-Fluorescence spectra

Tables

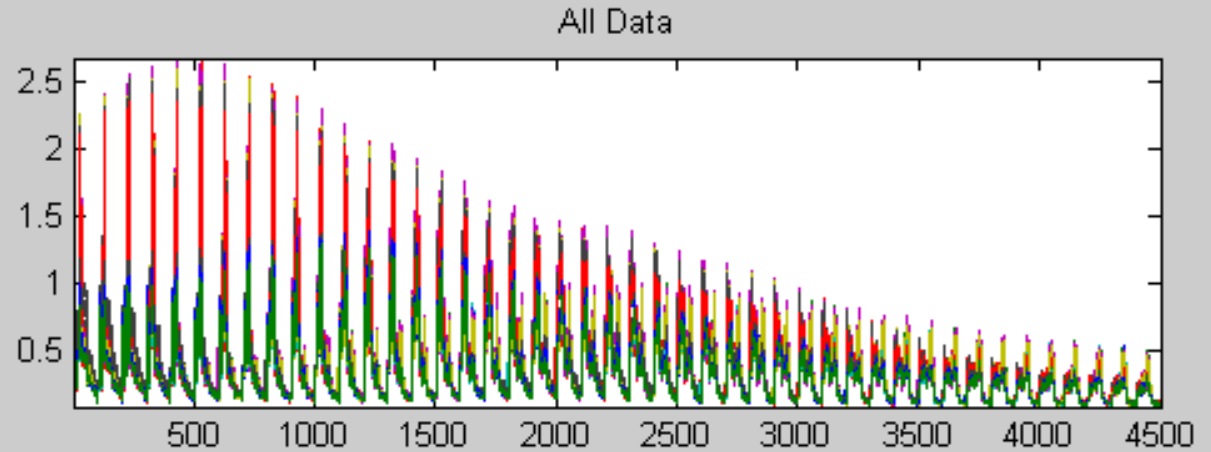
- 900 blocks of 5 variables
- 6 Common Dimensions



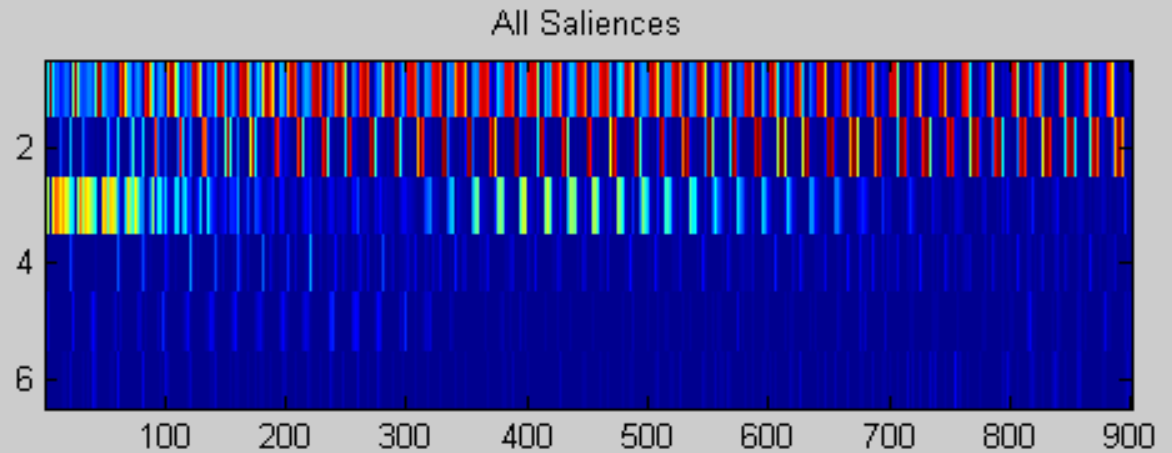


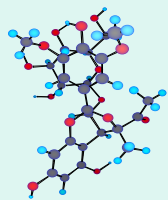
Unfolded 3D-Fluorescence Spectra

Spectra

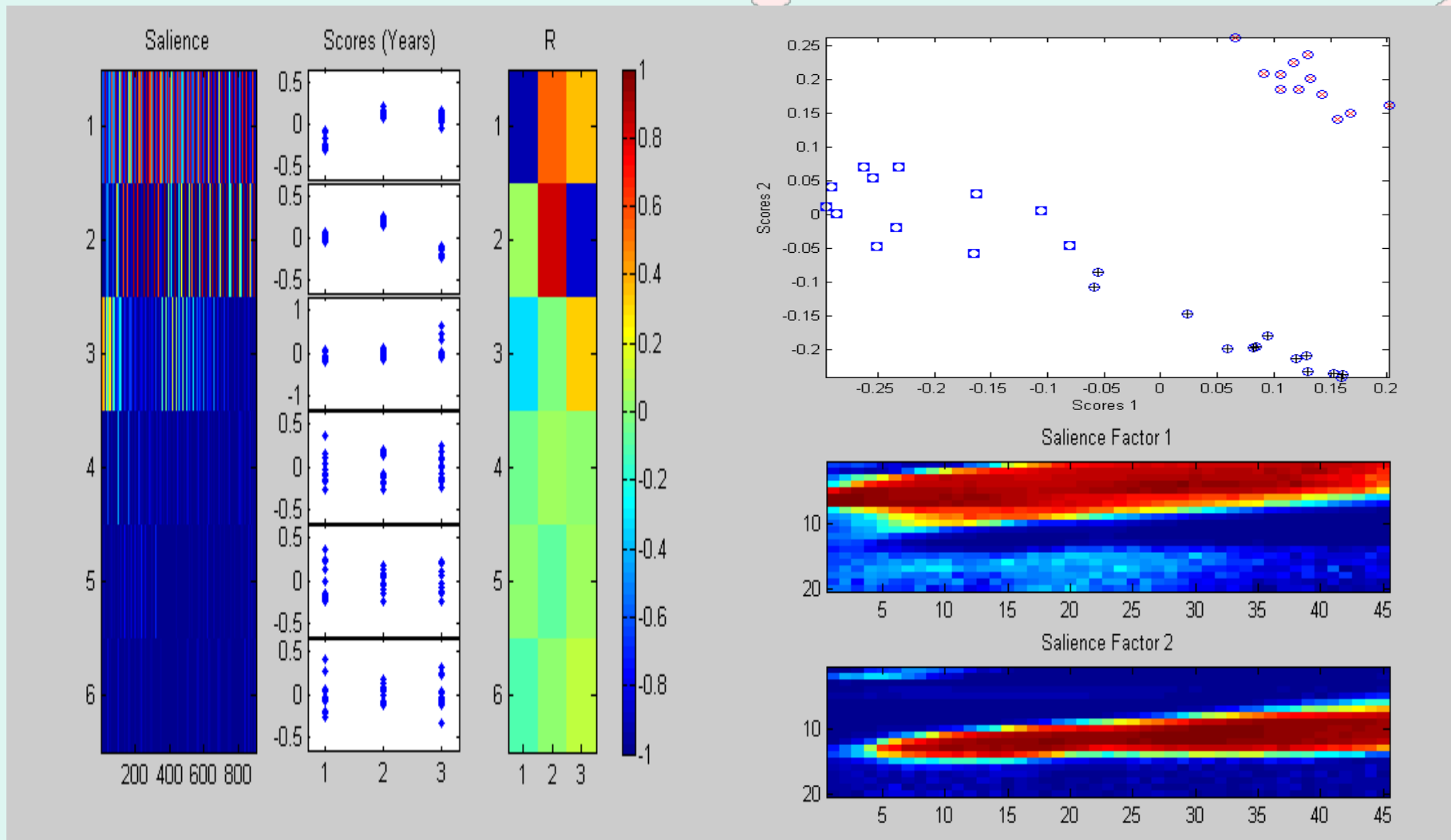


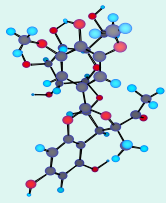
ComDim
Saliences



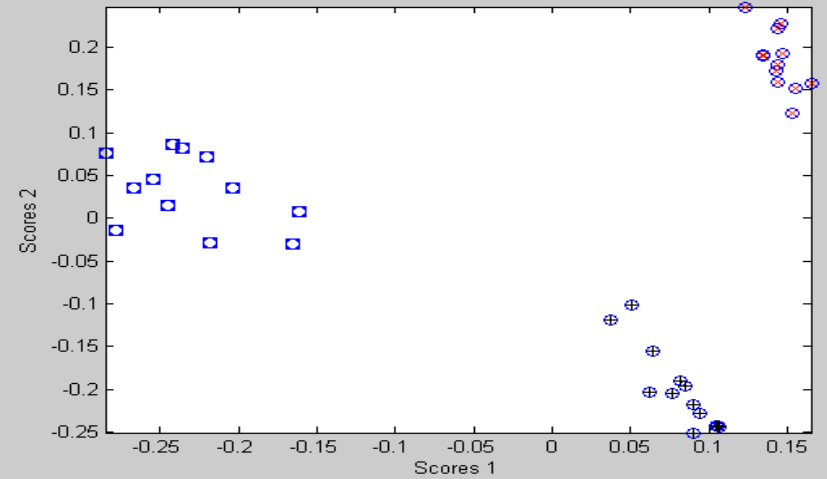
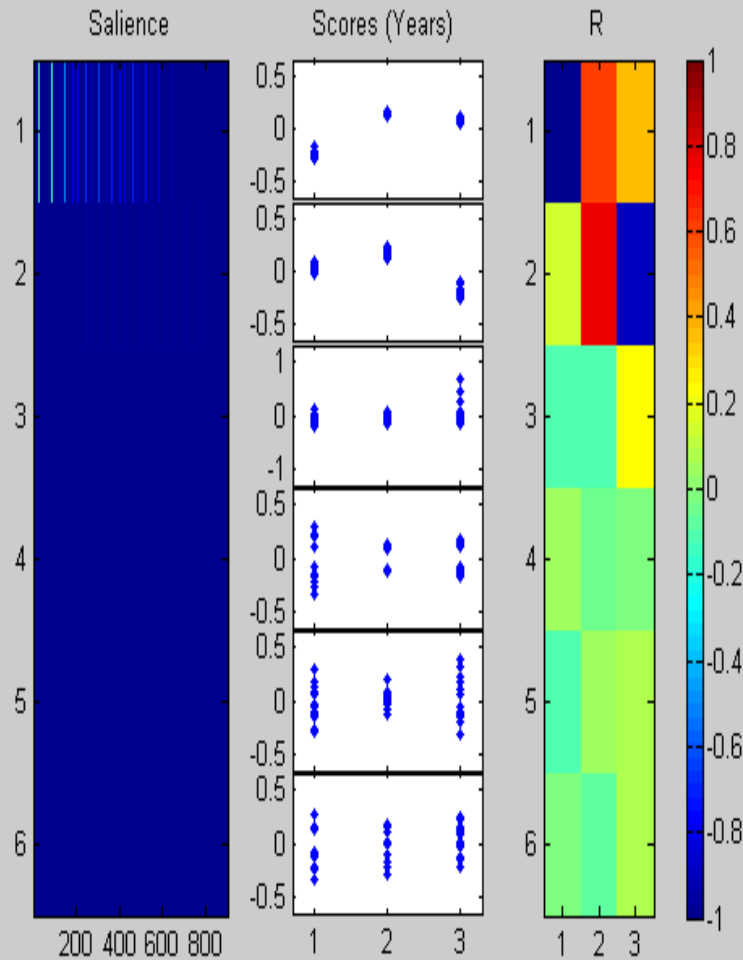


Correlation between ComDim Scores and "Year"

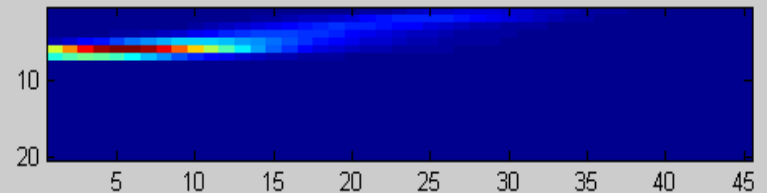




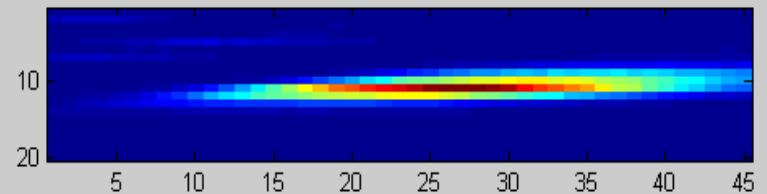
Correlation between PLS-ComDim Scores and "Year"

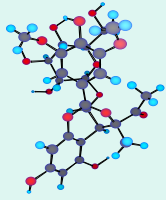


Salience Factor 1



Salience Factor 2



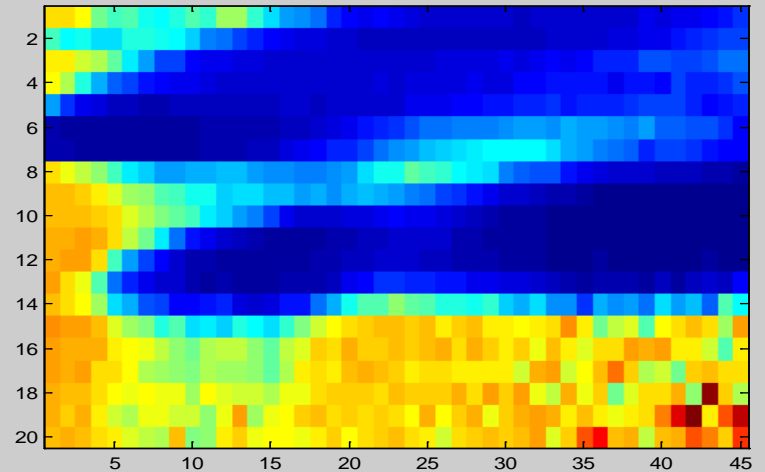
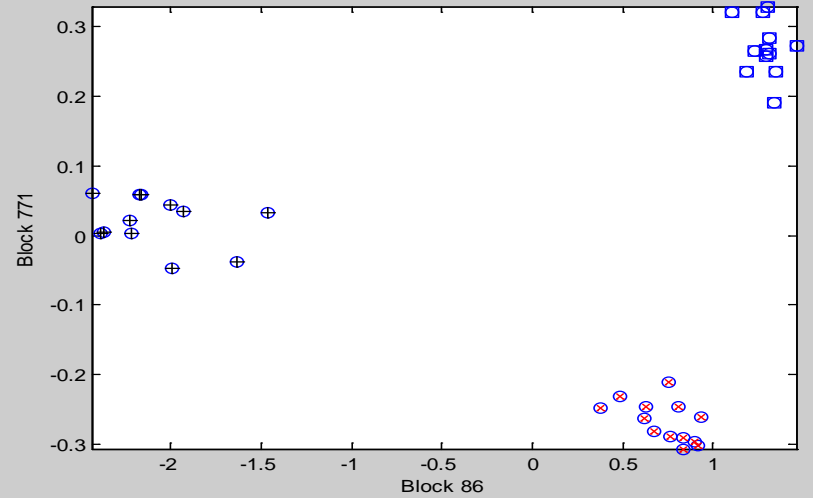
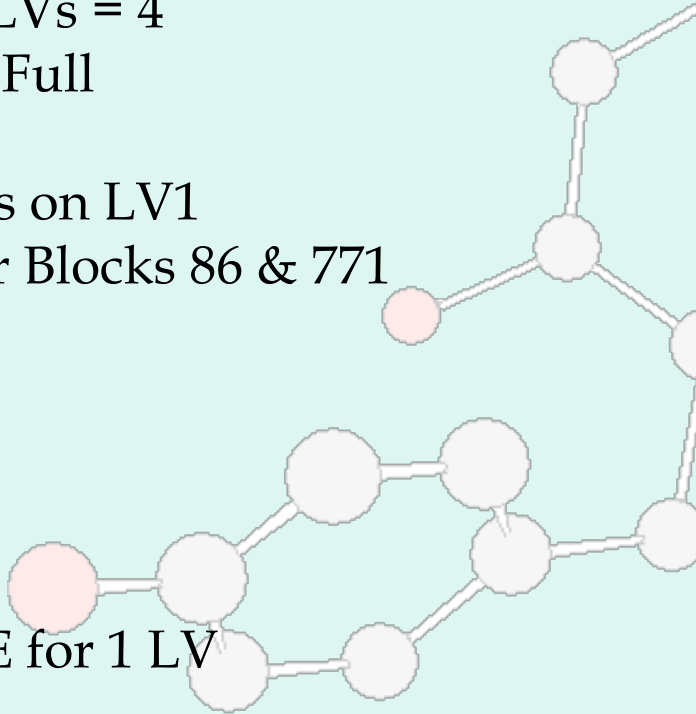


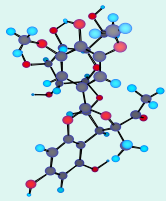
i-PLS between NIR and "Year"

- Blocks = 900
- Mean centred
- Max LVs = 4
- CV = Full

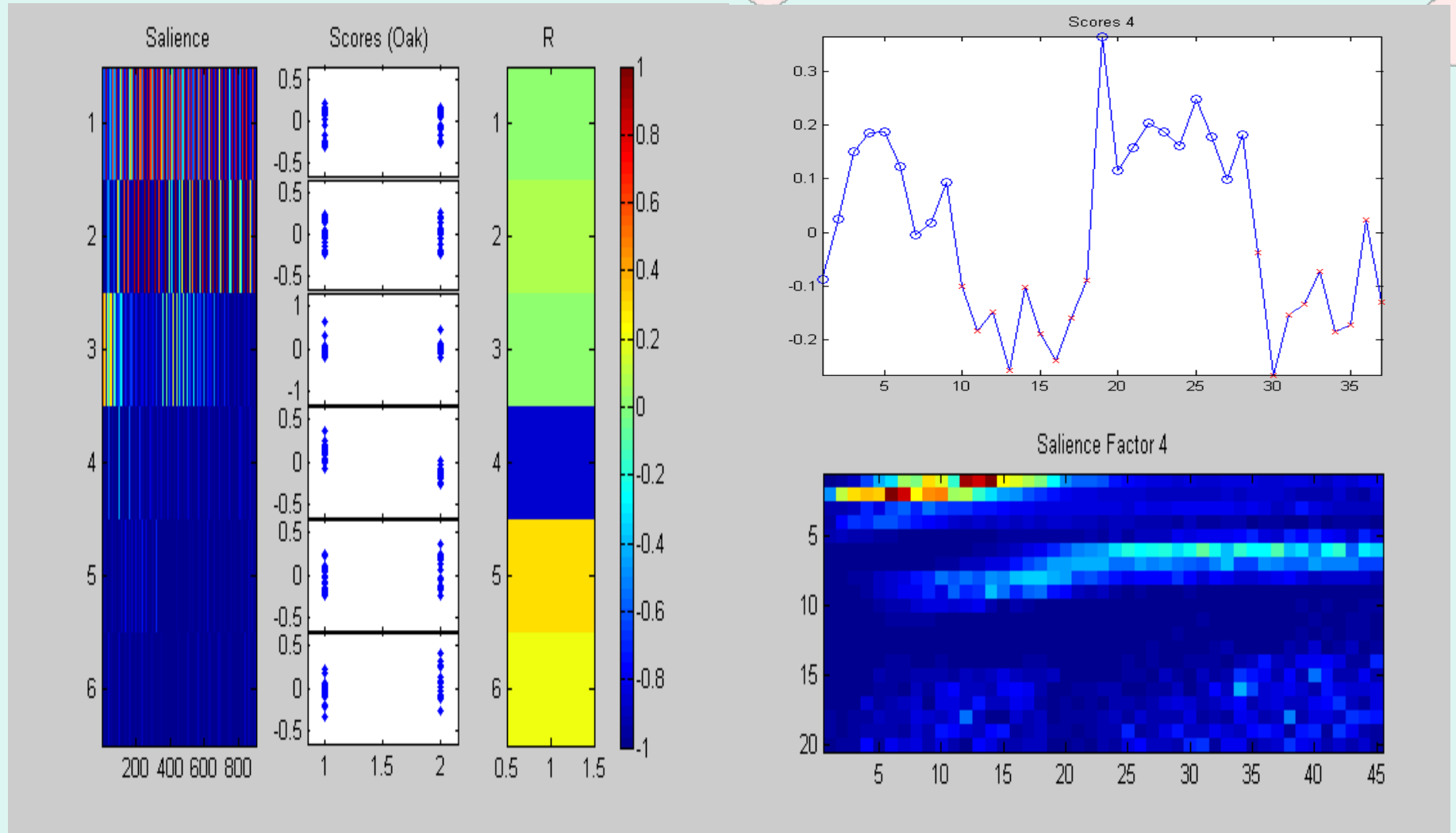
- Scores on LV1
for Blocks 86 & 771

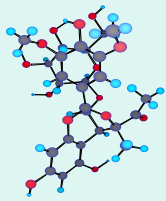
- RMSE for 1 LV



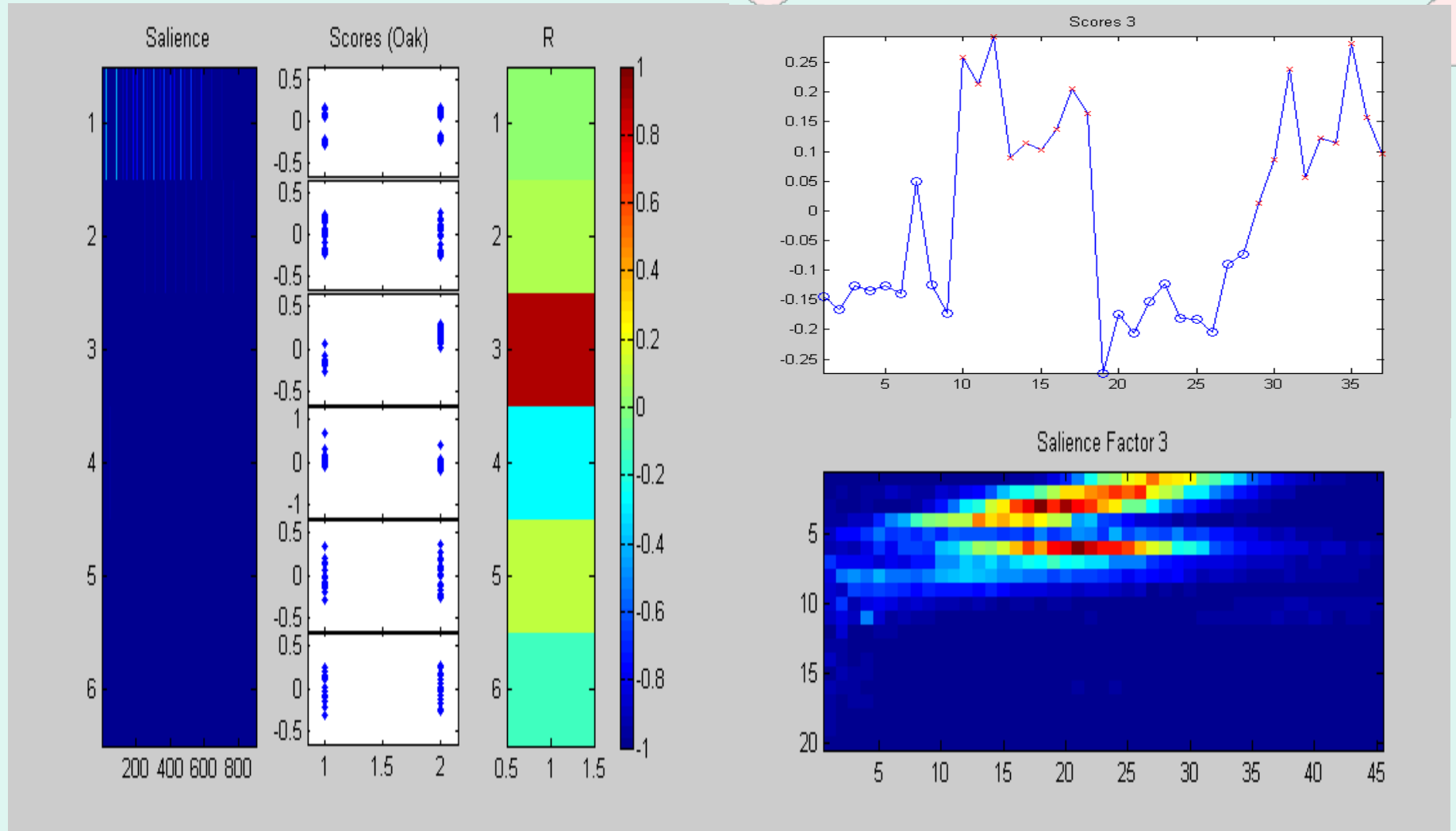


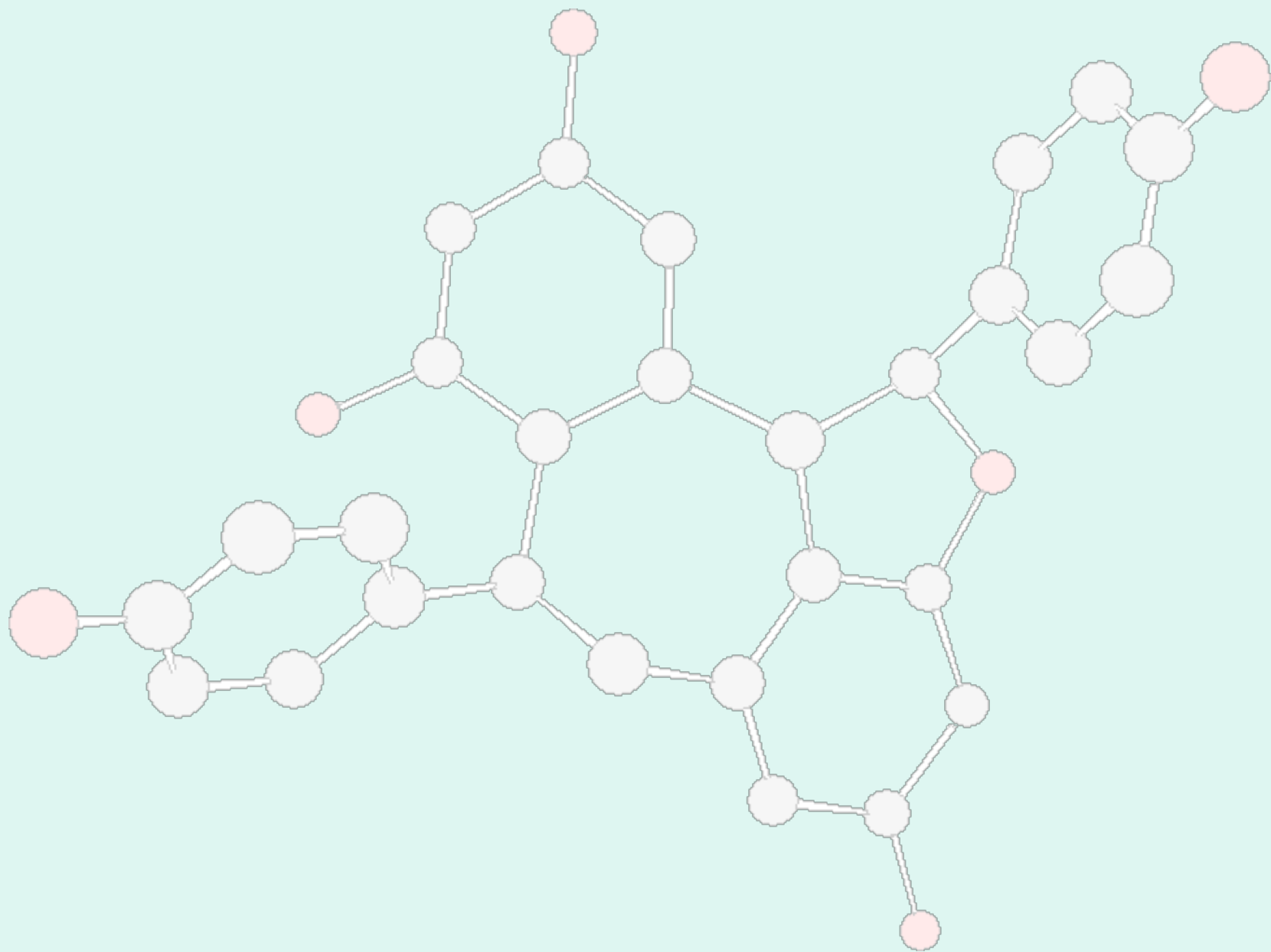
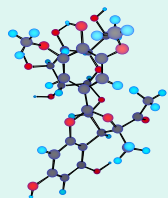
Correlation between ComDim Scores and "Oak"

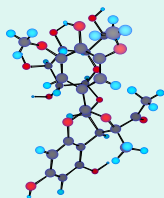




Correlation between PLS-ComDim Scores and "Oak"







First African-European Conference on Chemometrics

Mining School of Rabat

Morocco, 20th to 24th of September 2010



www.afrodata.org