Presentation at Chimiometrie 2009

Bringing the Chemical Knowledge into Empirical Models
ou
Introduire de l'information théorique dans les modèles empiriques.

Marion Cuny & Frank Westad

CAMO Software AS
mc@camo.no

CAMO

Outline

1. Limits of empirical modeling and necessity of bringing chemical knowledge in the model
2. Methodology
   a) NIR and MIR data
   b) Cross-model validation
   c) Background information
   d) LPLS
3. Results
4. Conclusions

Outline

1. Limits of empirical modeling and necessity of bringing chemical knowledge in the model
2. Methodology
   a) NIR and MIR data
   b) Cross-model validation
   c) Background information
   d) LPLS
3. Results
4. Conclusions

Limits of empirical modeling

For many years, chemometrics has focused on new, better methods for classification and regression, and using variable selection and pretreatments for the optimal classification or the lowest RMSEP.

- examples of variable selection methods: iPLS, moving window iPLS, genetic algorithms, best combination search, uncertainty estimates by resampling (bootstrap, jack-knifing)…

- examples of pretreatments: MSC, SNV, Detrending, OSC, EPO…

32 marzipan samples

- 32 marzipan samples made from 9 recipes were obtained from an industrial batch production.
- Differences in recipes:
  - processing times
  - amounts of almonds,
  - amounts of apricot kernels,
  - amounts of water,
  - amounts of sucrose,
  - amounts of invert sugar,
  - amounts of glucose syrup
  - presence and amount of additives in the marzipan masses.

  Details about the experimental setup can be found in [1].

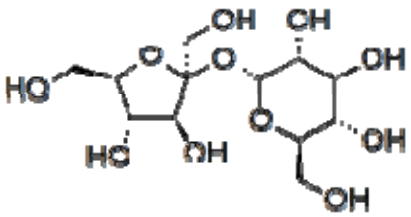
→ Goal: Online measurement of the sugar content ranging from 35-70% and water content varying from 6-18% with NIR.

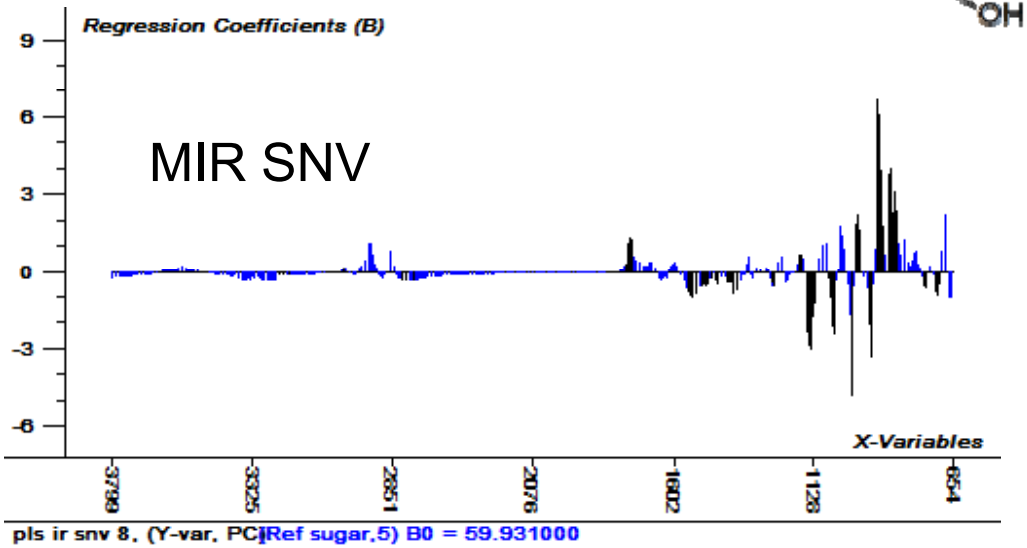
[1] J Christensen, L Nørgaard, H Heimdal, JG Pedersen, SB Engelsen. Rapid Spectroscopic Analysis of Marzipan – Comparative Instrumentation. *Journal of Near Infrared Spectroscopy*, vol 12 (1), 2004.

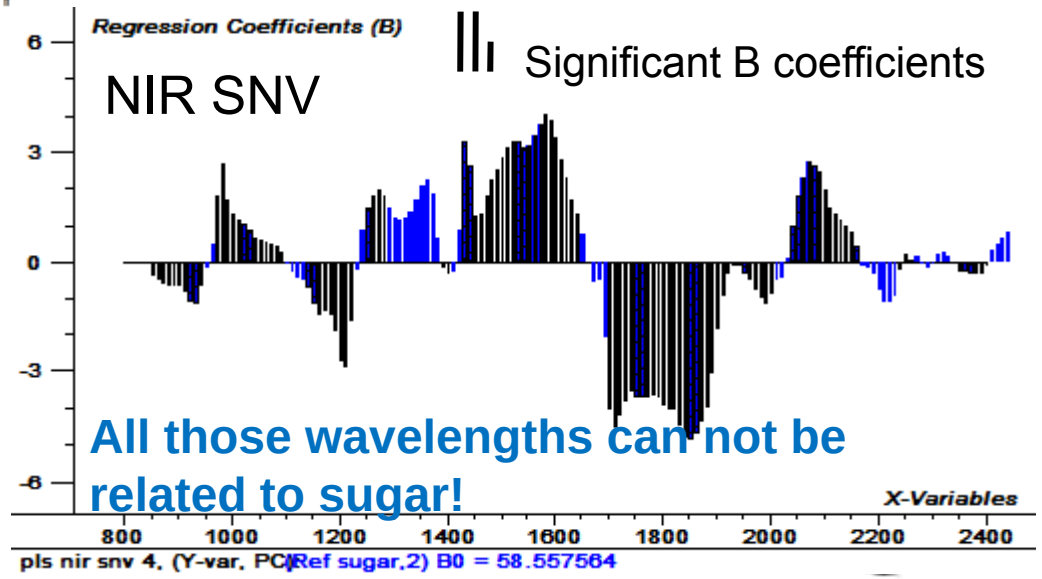
Regression results for sugar content

Type of data	Pretreatment	# LVs	RMSECV (Full)	R2 (Validation Full CV)
NIR	No pretreatment	5	1.57	0.986
	SNV	4	1.43	0.988
MIR	No pretreatment	8	2.01	0.976
	SNV	8	1.77	0.982

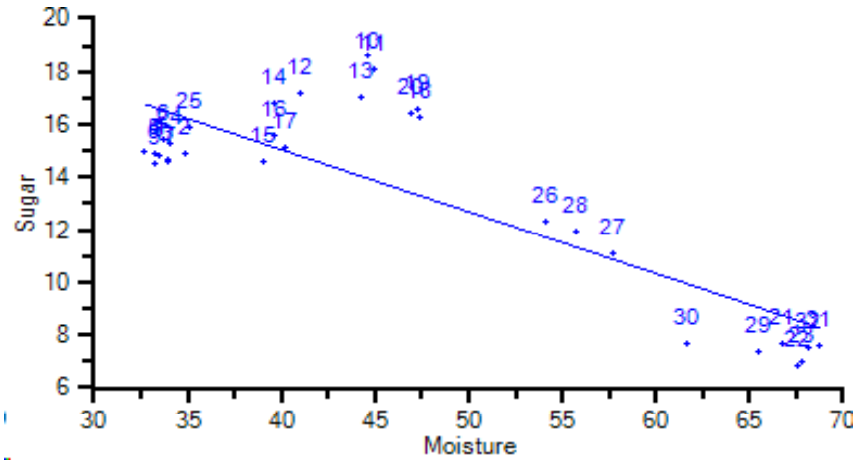
Sugar molecule:

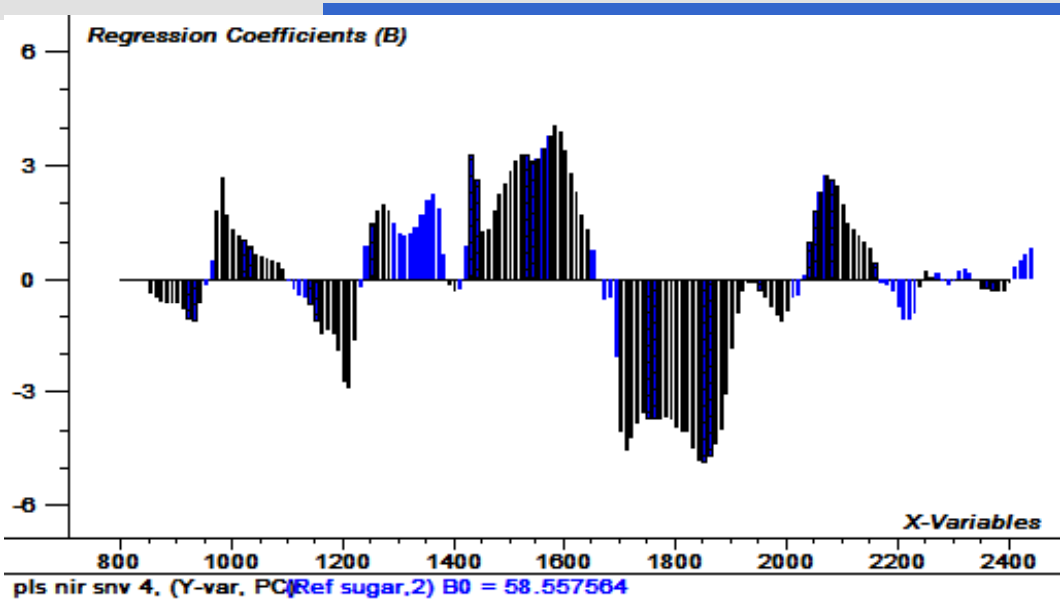






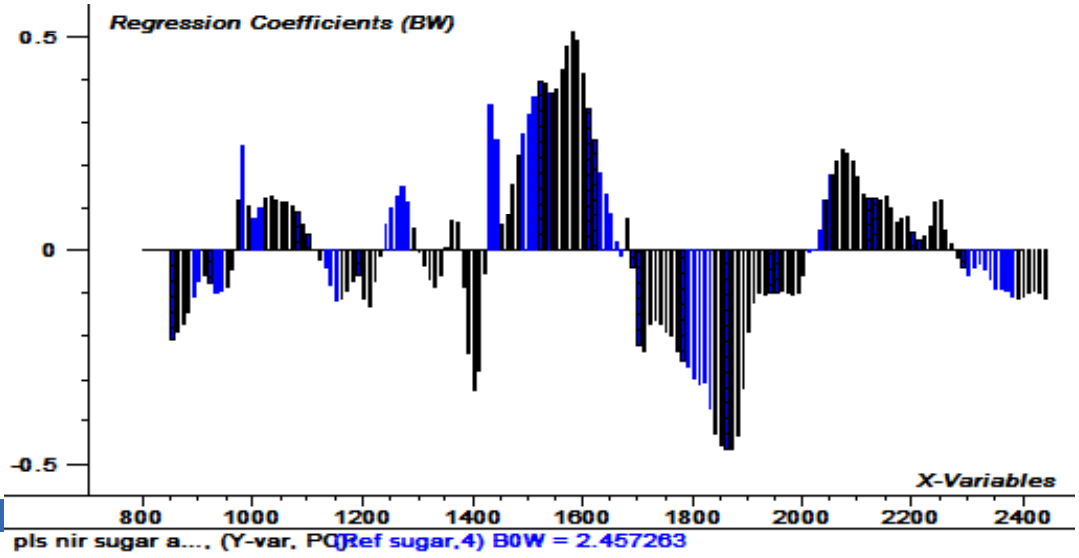
How do you deal with very correlated variables?





PLS2 algorithm.

More selective but still too many wavelengths!



Goals

- To find a good model of sugar content that can be explained by chemistry
  - What is the chemistry in the system that is being observed?

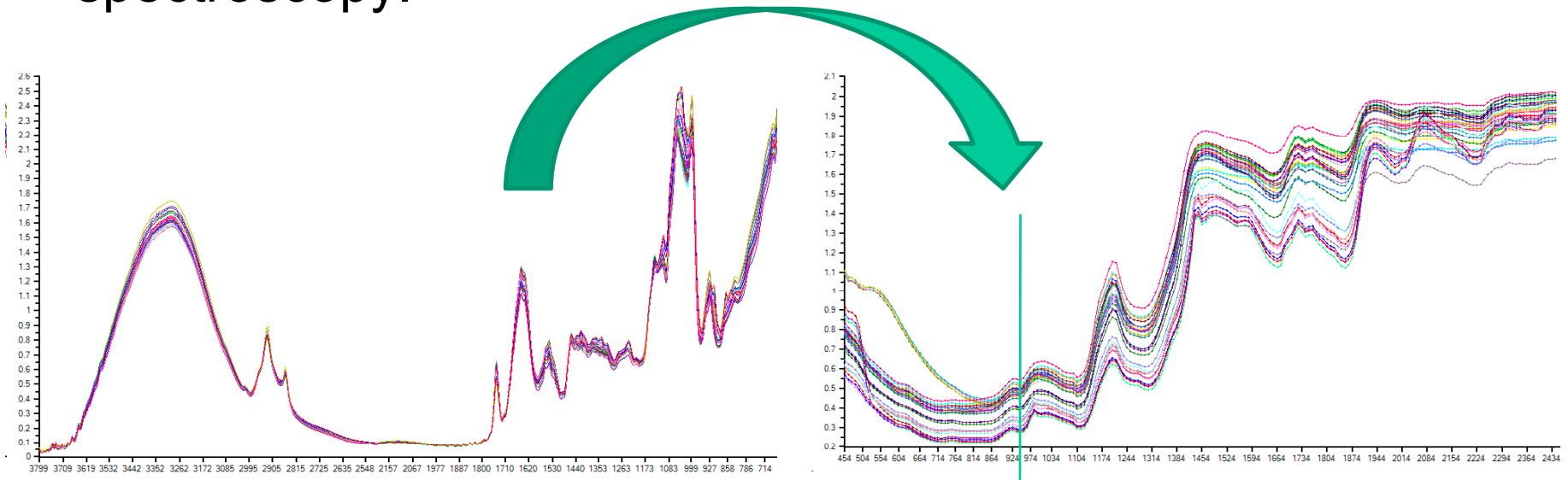
- To do so: include more of the chemistry in the models:
  - Combined the available instrumental data, to get a better understanding: NIR explained by MIR.
  - At the same time, insert the chemical background knowledge to confirm the interpretation from the models.

Outline

1. Limits of empirical modeling and necessity of bringing chemical knowledge in the model
2. Methodology
   a) NIR and MIR data
   b) Cross-model validation
   c) Background information
   d) LPLS
3. Results
4. Conclusions

MIR and NIR spectroscopy of marzipan samples

- FT-IR (Perkin-Elmer System 2000 1700-600 cm^{-1}) and VIS/NIR (FOSS NIR System 6500 instrument 400-2500 nm) spectroscopy.



MIR spectra (left) and VIS/NIR spectra (right) of marzipan samples

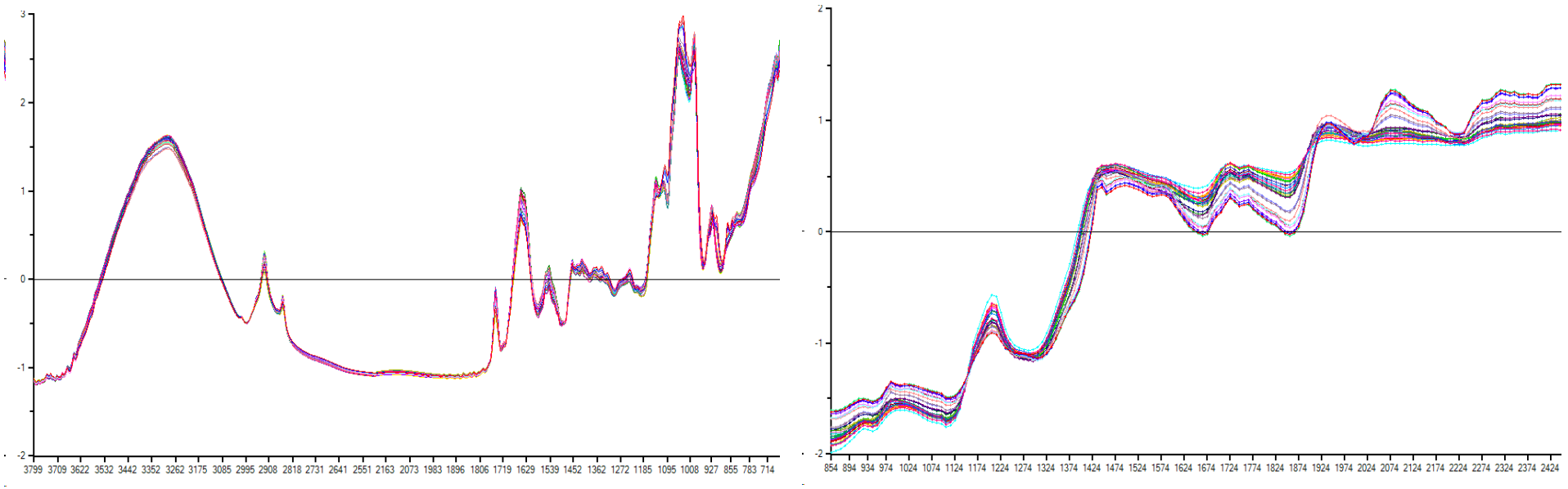
MIR spectra are considered as containing the fundamental chemical information

NIR spectra contain overtones of the fundamental chemical information

Preprocessing

- Standard normal variates (SNV), which is equivalent to centering and normalizing each spectrum to unit variance.

$$x_{ik}^{SNV} = (x_{ik} - m_i) / s_i$$

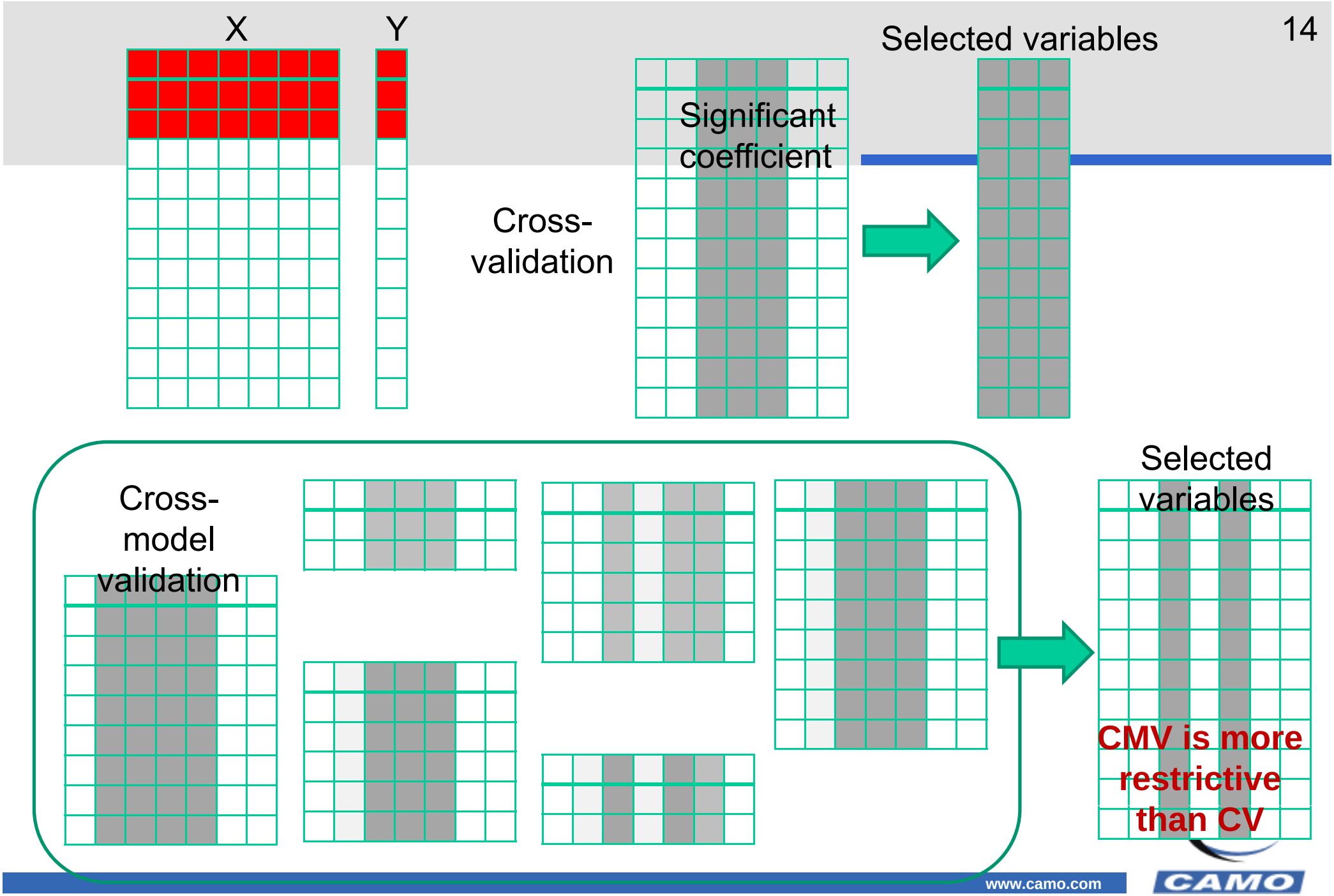


Selection of variables

- PLS regression between X=MIR and Y=NIR
- Study of the significance of the coefficient tested by jack-knifing.
- Validation method=cross-model validation (CMV)

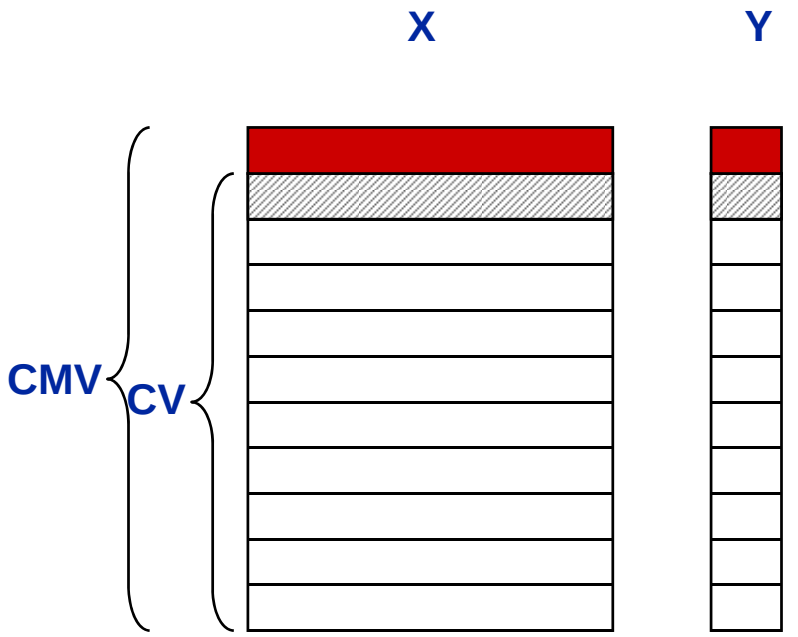
Outline

1. Limits of empirical modeling and necessity of bringing chemical knowledge in the model
2. Methodology
   a) NIR and MIR data
   b) Cross-model validation
   c) Background information
   d) LPLS
3. Results
4. Conclusions


X
Y
Cross-validation
Significant coefficient
Selected variables
Cross-model validation
Selected variables
CMV is more restrictive than CV

Cross-model validation (CMV) with variable selection based on jack-knifing

0. Cross-validation on all objects
1. Take out e.g. 10% of the objects
2. Cross-validate the remaining
3. Find significant variables
4. Predict the objects that were kept out
5. Estimate RMSE (or explained variance)
6. Repeat 1 - 5 until all objects have been taken out
7. Show frequency and significance for all variables
9. Collect and predict an independent test-set!



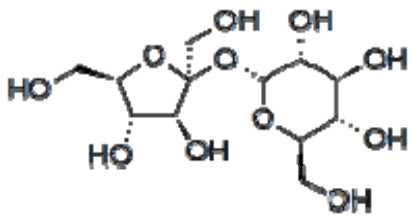
Outline

1. Limits of empirical modeling and necessity of bringing chemical knowledge in the model
2. Methodology
   a) NIR and MIR data
   b) Cross-model validation
   c) Background information
   d) LPLS
3. Results
4. Conclusions

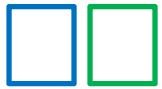
Band assignment table

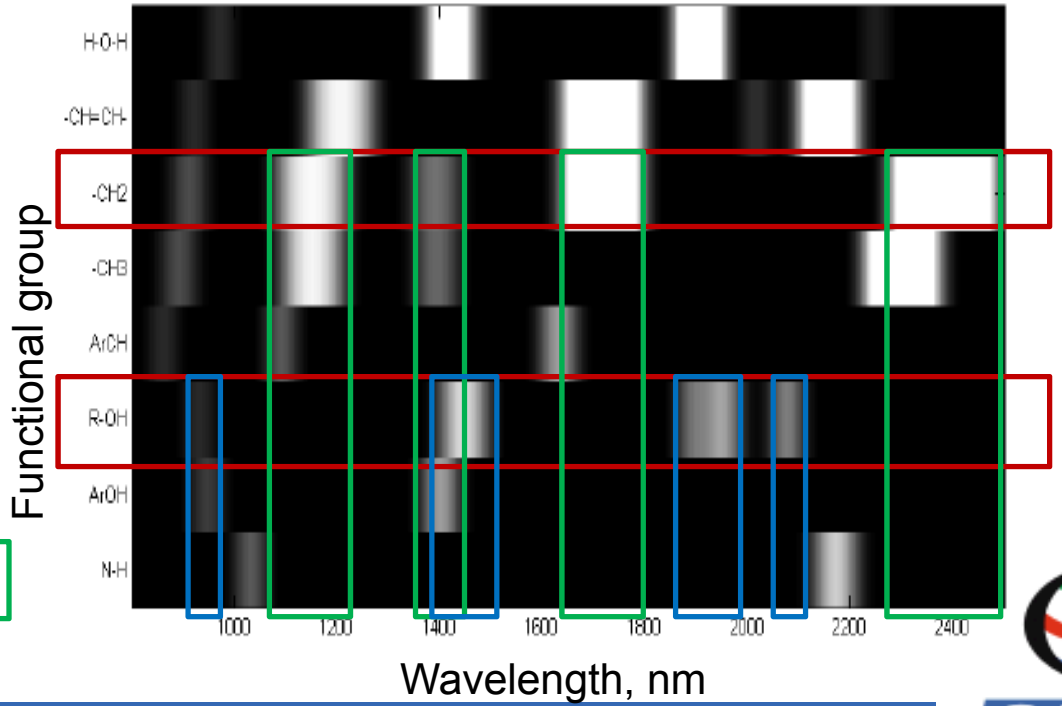
- 8 group frequencies were assigned to their respective vibrational regions from an NIR-chart, and an initial binary matrix was built using 1 to indicate peaks and 0 elsewhere.
- In addition, peaks were weighted either as weak (0.5), normal (1), strong (2) or very strong (4). The band assignment matrix was then convolved with a Gaussian filter of size N=30 (corresponding to 60 nm when the spectral resolution is equal to 2 nm) for each group.

Sugar molecule:



Expected wavelengths



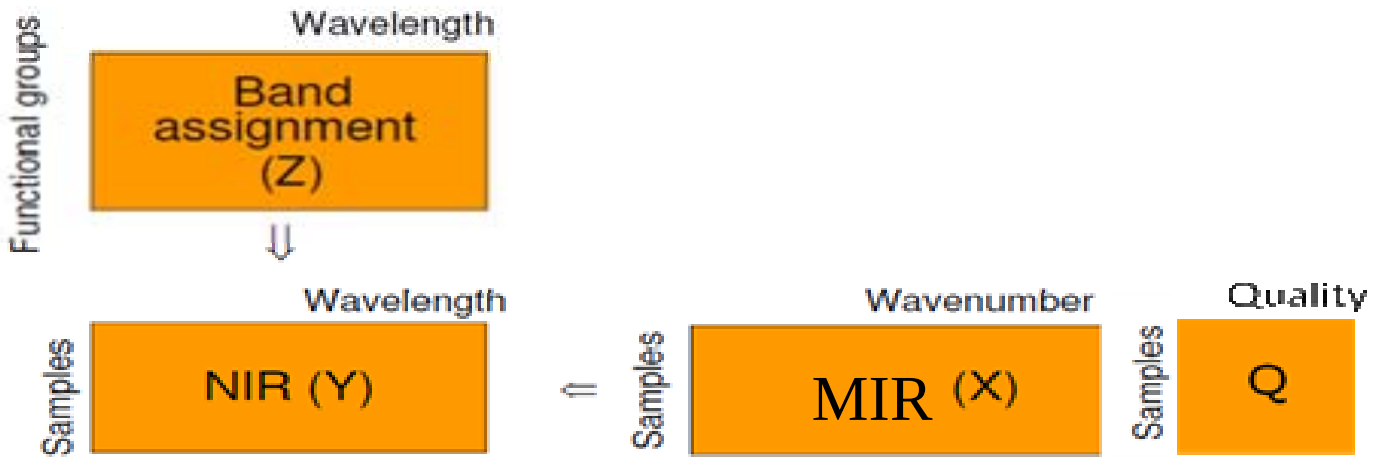


Outline

1. Limits of empirical modeling and necessity of bringing chemical knowledge in the model
2. Methodology
   a) NIR and MIR data
   b) Cross-model validation
   c) Background information
   d) LPLS
3. Results
4. Conclusions

Structure of the L-PLS data

- One interesting aspect of L-PLSR is that the regression model is based on the inherent link between the actual spectra and theoretical band assignment, giving direct “chemical” interpretation.
- With broad bands like in NIR, the assignments are rather crude, and one should always interpret the results in the light of the chemical background knowledge. By applying this procedure in e.g. MIR spectroscopy, a more detailed interpretation compared to NIR would then be possible.



L-shape PLS regression

- Regular two-block PLS:

  X-weights (used to calculate scores and loadings) can be obtained from the A first eigenvectors of:

$$\mathbf{X}_{a-1}{}^{T}\mathbf{Y}_{a-1}\mathbf{Y}_{a-1}{}^{T}\mathbf{X}_{a-1}\mathbf{w}_{a} = \mathbf{w}_{a}\lambda_{a}$$

- Three-block / L-PLS:

  Weights for X and Z (used to find X and Z scores and loadings) may be obtained from the A first eigenvectors of:

$$\mathbf{X}_{a-1}{}^{T}\mathbf{Y}_{a-1}\mathbf{Z}_{a-1}\mathbf{Z}_{a-1}{}^{T}\mathbf{Y}_{a-1}{}^{T}\mathbf{X}_{a-1}\mathbf{w}_{Xa} = \lambda_{Xa}\mathbf{w}_{Xa}$$

$$\mathbf{Z}_{a-1}{}^{T}\mathbf{Y}_{a-1}{}^{T}\mathbf{X}_{a-1}\mathbf{X}_{a-1}{}^{T}\mathbf{Y}_{a-1}\mathbf{Z}_{a-1}\mathbf{w}_{Za} = \lambda_{Za}\mathbf{w}_{Za}$$

H Martens., E Anderssen, A Flatberg, L.H. Gidskehaug, M Høy, F Westad, A Thybo, M Martens. Regressing a matrix on descriptors of both its rows and of its columns, by low-rank L-PLS Regression. *Computational Statistics and Data Analysis*, **48**, 103-125, 2005.

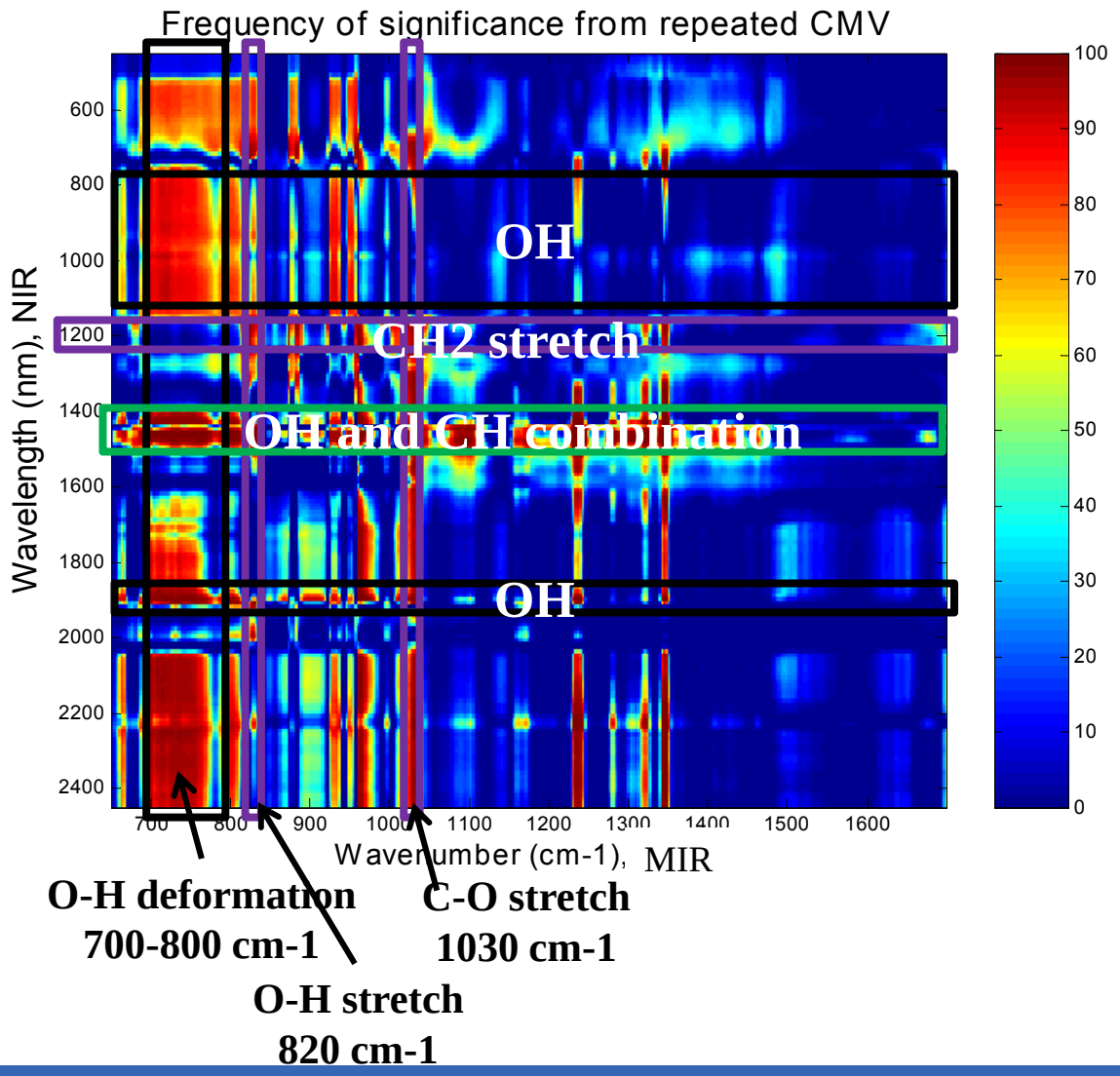
Outline

1. Limits of empirical modeling and necessity of bringing chemical knowledge in the model
2. Methodology
   a) NIR and MIR data
   b) Cross-model validation
   c) Background information
   d) LPLS
3. Results
4. Conclusions

Selection of variable by CMV

- The 32 marzipan samples were subject to repeated CMV (100 runs. Uncertainty estimates from jack-knifing).
- 3 PLSR components was the basis for significance tests at 5% level.
- The main results are shown as a map of frequency of significance for the regression coefficient matrix **B**. Sugar and water constitute the chemical compounds of interest.

Frequency of significance from repeated CMV



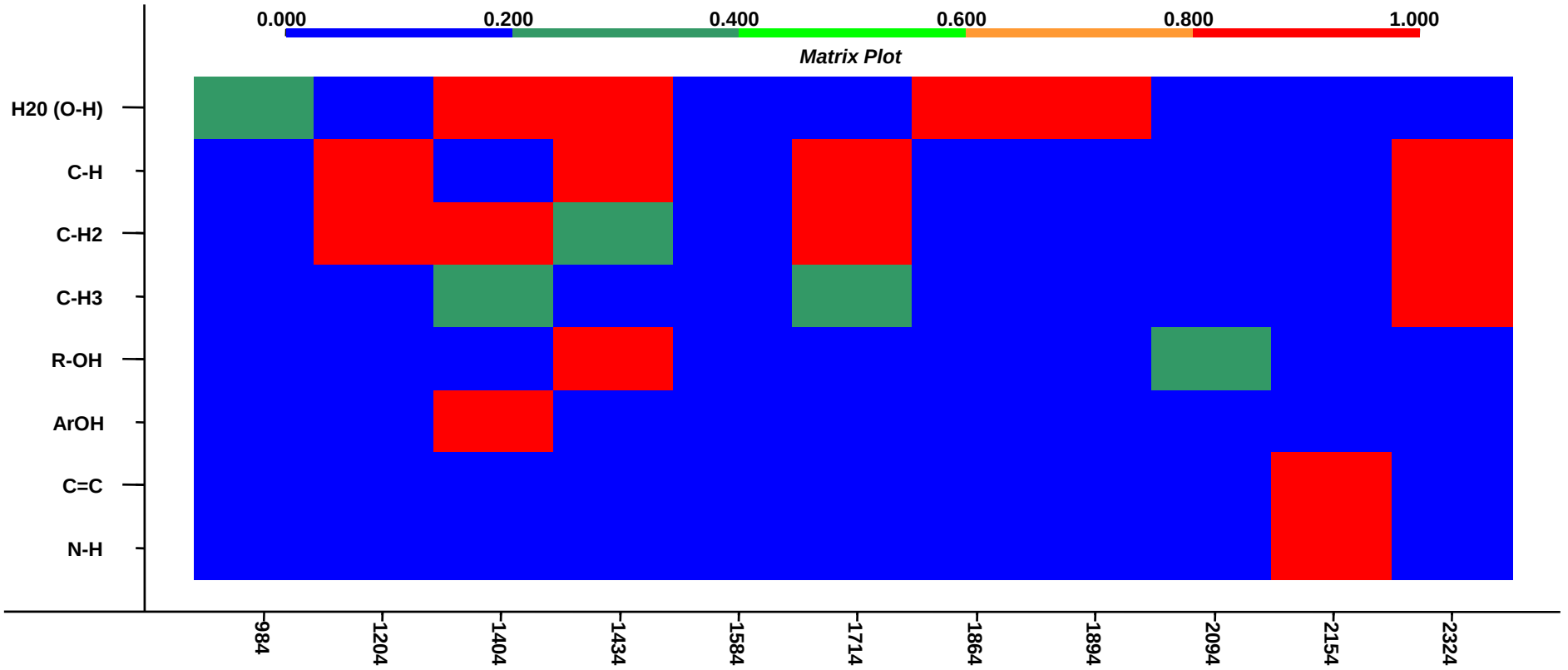
C-H deformation at 820 cm^{-1} corresponds to the NIR region around 2200 nm.

The region 1400–1500 nm is related to both O-H stretch vibrations and C-H combinations.

The main peaks in the bands that were significant were selected as input to the L-PLS regression.

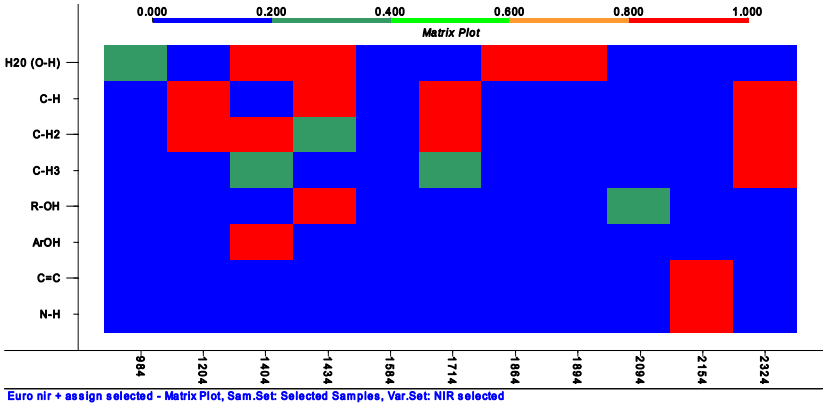
Background: selected peaks

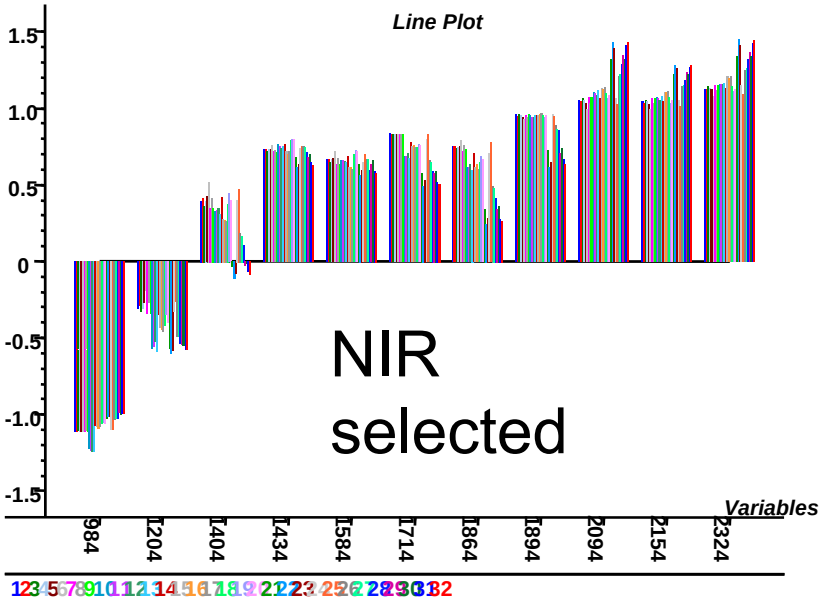
Selection of the wavelengths representing the top of the peak.

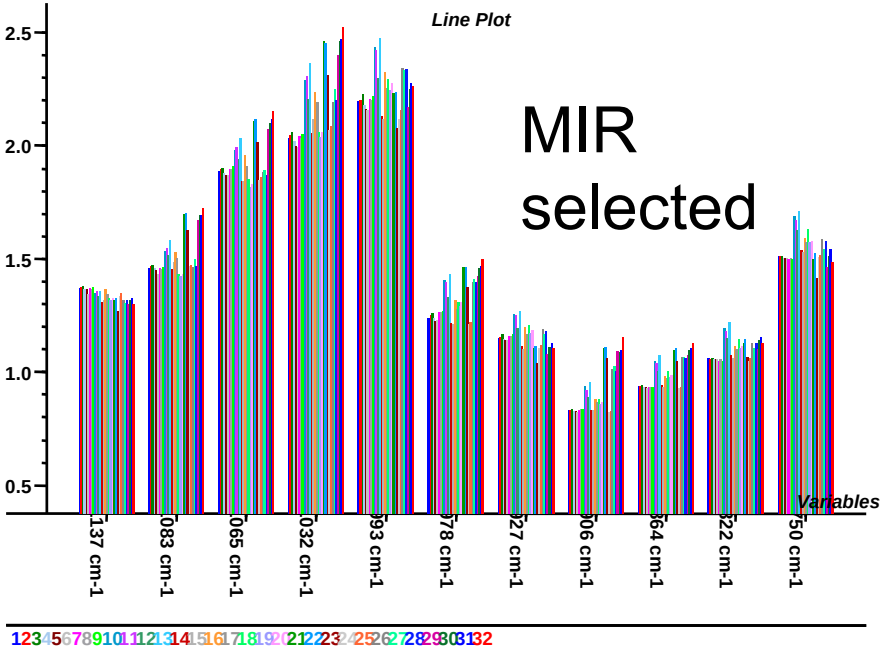


Euro nir + assign selected - Matrix Plot, Sam.Set: Selected Samples, Var.Set: NIR selected

Data structure



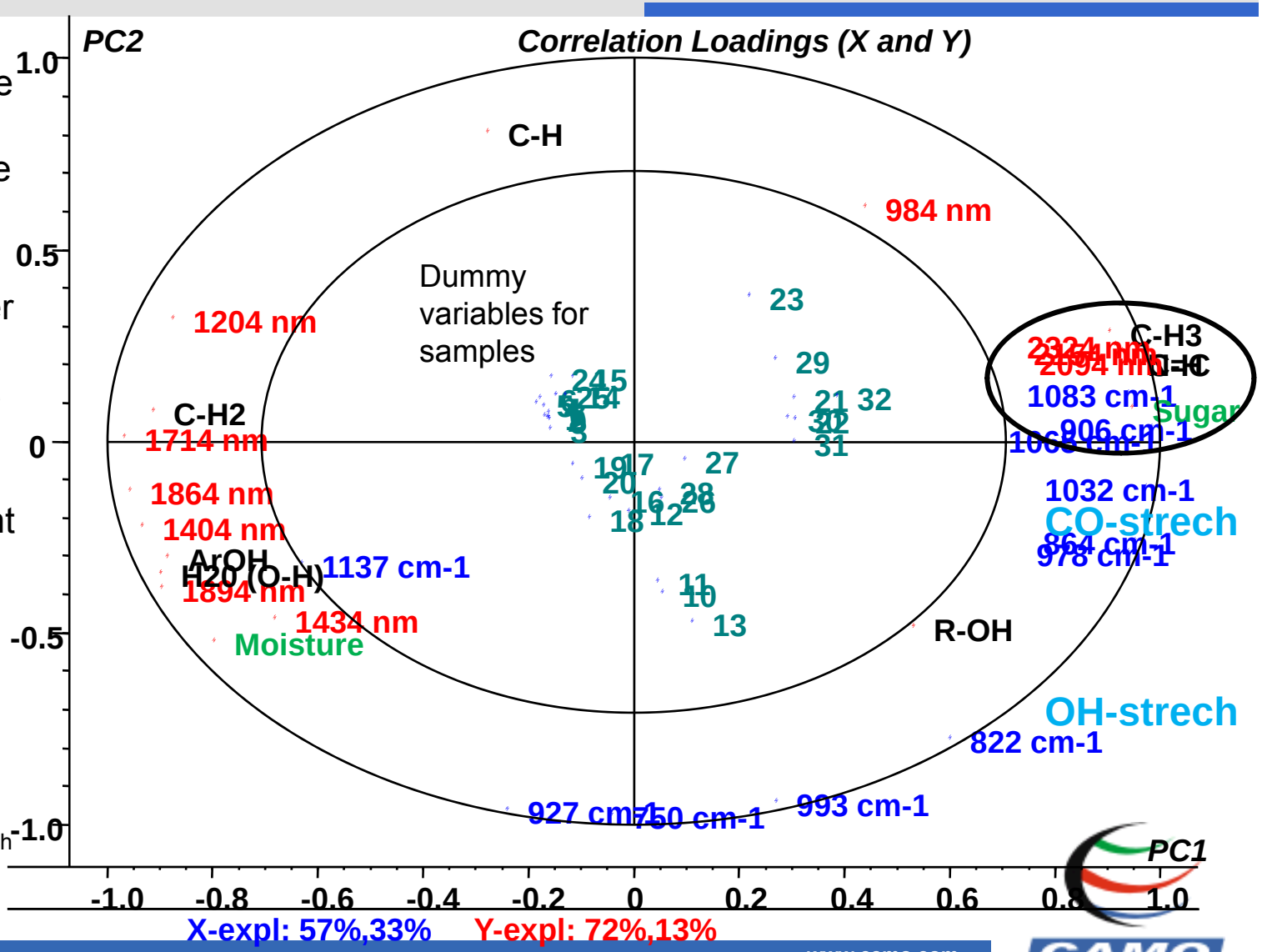




Moisture	Sugar
33	34
14.7000	33.90
14.9000	33.20
14.9000	33.20
15.4000	33.70
14.6000	34.00
14.8000	33.50
14.5000	33.20
18.6000	44.60
18.1000	45.00
17.2000	41.00
17.0000	44.30
16.8000	39.60
14.6000	39.10
15.6000	39.60
15.1000	40.20
16.3000	47.40
16.6000	47.30
16.4000	46.90
7.7000	66.70
6.8000	67.60
7.0000	67.80
15.3000	34.10
15.9000	35.10
12.3000	54.10
11.1000	57.70
11.9000	55.70
7.4000	65.50
7.7000	61.60
7.6000	68.70
7.5000	68.20

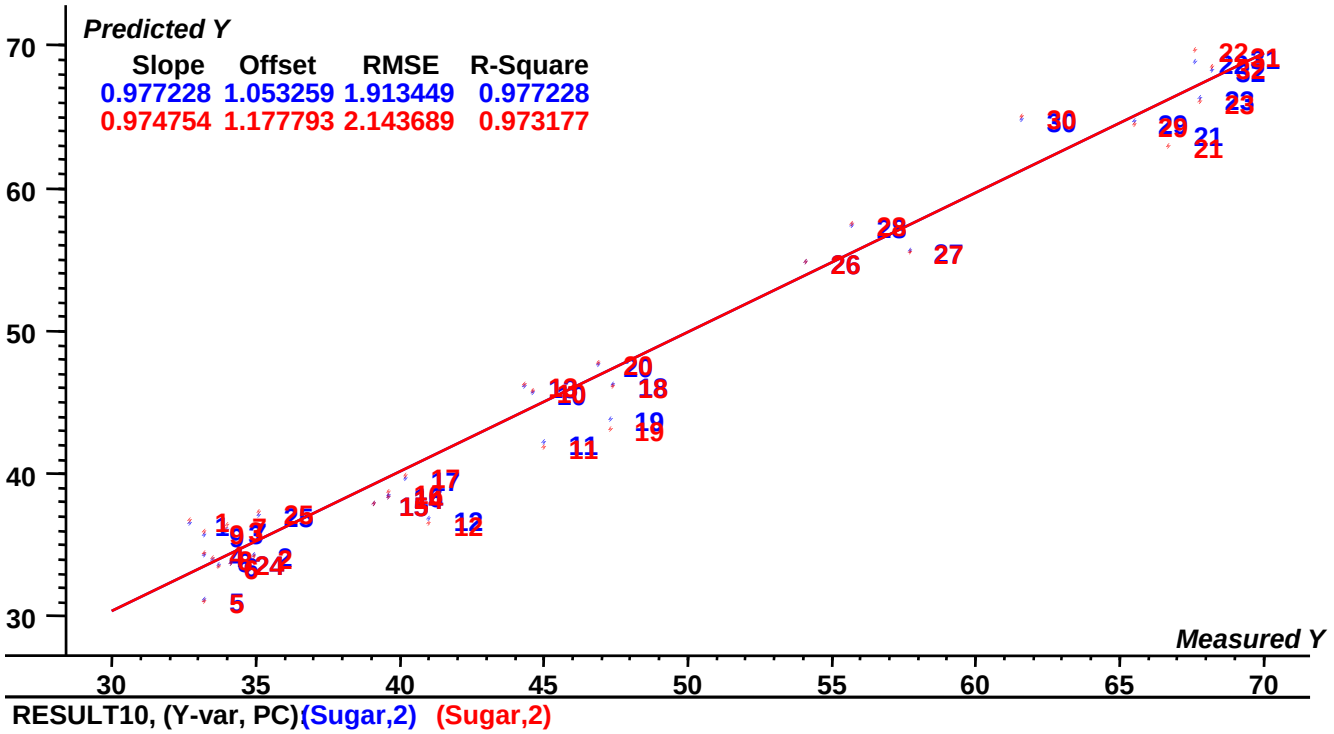
Correlation loading plot

- Direct interpretation of the chemistry and the actual spectral regions that were found to be significant.
- Note that sugar and water content are inversely correlated in the samples themselves, so there is both a concentration dependence as % content as well as a chemical dependence in terms of OH bands.
- Legend:
  - NIR wavelengths,
  - MIR wave numbers,
  - Chemical background: wavelength table,
  - Composition



Regression on the selected OH peaks

The –OH peaks to be used to measure the sugar contents are: 2094, 2154, 2324 nm.



Results may not be better (in this case they are not) but you are able to say why you are using those wavelengths!

Outline

1. Limits of empirical modeling and necessity of bringing chemical knowledge in the model
2. Methodology
   a) NIR and MIR data
   b) Cross-model validation
   c) Background information
   d) LPLS
3. Results
4. Conclusions

Conclusions

- Cross model validation with jack-knife estimates is an efficient way of removing variables that are not of interest, and the relation between two instrumental methods can be presented as a color image.
- L-PLS regression gives direct interpretation of the underlying chemistry which is useful for confirming existing knowledge but also for finding unknown phenomena which may lead to innovation and further research.
- The correlation loading plot is a very condensed way of visualizing all of interest for the three data tables.

Perspectives

- L-PLS is adapted to other types of data:
  - NMR

Z: information on shift assignment

Y: NMR measurement

X: information on the samples or other measurements

Perspectives

- L-PLS is adapted to other types of data:
  - metabolomic

Z: information on phenotype

Y: metabolomic measurement

X: information on the samples

Perspectives

- L-PLS is adapted to other types of data:
  - sensory

Z: information on consumer

Y: consumer preference

X: information on the samples

Get value out of your data

The Unscrambler®

Thank you for your attention

Marion Cuny

marion@camo.no

Other webinars: http://www.camo.com/training/webinars-seminars.html

Recorded webinars: http://www.camo.com/training/archives.html

CAMO