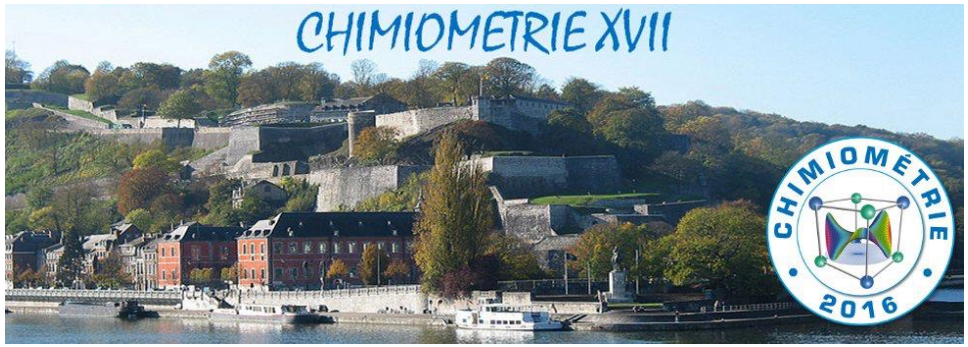
Des indices simples pour mesurer la qualité d'une base d'étalonnage

Simple indexes to assess the quality of a calibration database

JM Roger[1], A Gobrecht[1], DN Rutledge[2], JC Boulet[3]

[1] Irstea UMR ITAP, [2] Inra UMR1145 GENIAL,

[3] Inra UMR SPO,



Outline

- Introduction / theory
- First example
- Second example
- Third example
- Conclusion

Introduction

- Quality of calibration strongly depends on the quality of the database (X,Y)

- The quality assessment is based on:
  - Either expert judgement, PCA, histograms, etc.
  - Or on systematic tests, such as cross validation

- Numerous applications :
  - Comparison of devices
  - Choice of preprocessing
  - …

Introduction: an example

- How to choose the best preprocessing ?
- In (Engel et al, 2013) : « Breaking with trends in pre-processing? » three types of methods:
  - Trial + error
  - Visual / expert inspection
  - Quality parameters

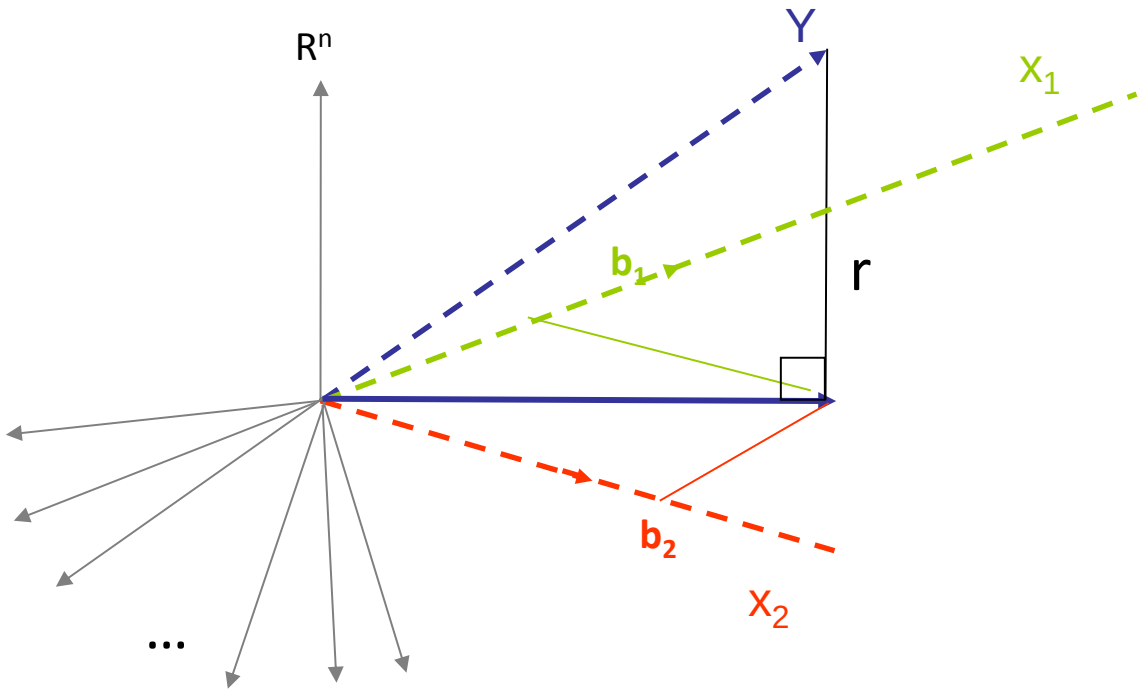
« *Quality parameters offer a robust, objective and sometimes quicker alternative to visual inspection. However, pre-processing strategy selection via quality parameters does not yet seem to be common practice.* »

Introduction

- How to choose the best preprocessing ?
- In (Engel et al, 2013) : « Breaking with trends in pre-processing? » three types of methods:
  - Trial + error
  - Visual / expert inspection
  - Quality parameters

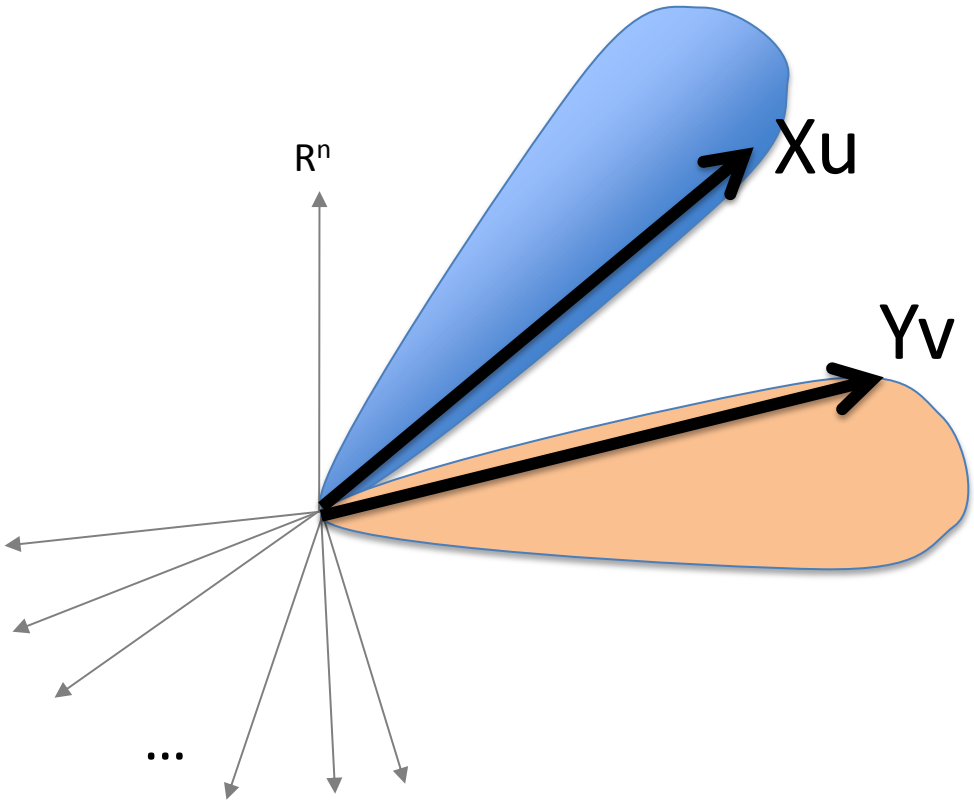
Theory

- Let X (nxp) and Y (nxq) be two matrices
- In R^n, an OLS regressing Y by X is analogous to projecting Y on X



Theory

- In R^n, a PLS consists of finding directions of X and Y as close as possible

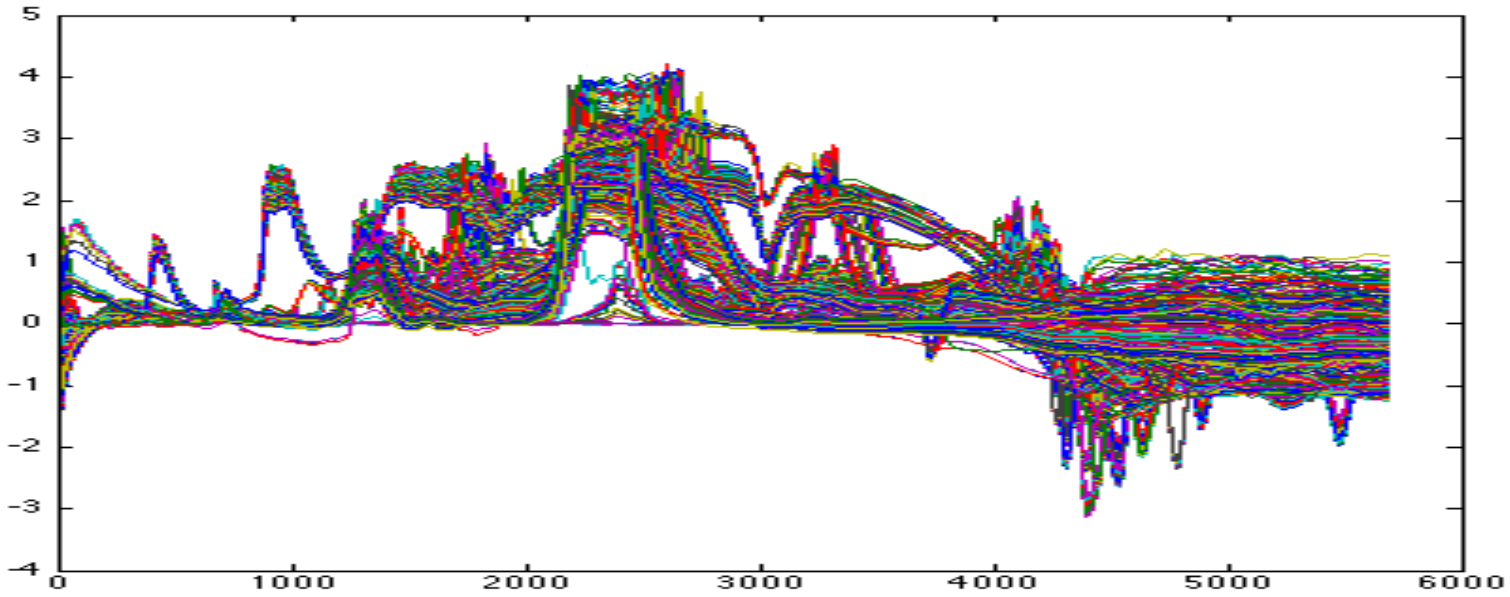


Theory

- The idea :
  - To measure the closeness of X and Y in R^n
  - By the norm of the projection of X on Y (or Y on X)
    - C1(X,Y) = trace(X'YY'X)
    - C2(X,Y) = trace(X'Y(Y'Y)$^{-1}$Y'X)
  - When used to compare preprocessing, should be independent from the X scale :
    - CN1(X,Y) = trace(X'YY'X) / trace(X'X)
    - CN2(X,Y) = trace(X'Y(Y'Y)$^{-1}$Y'X) / trace(X'X)

First example

- Challenge Chimiométrie 2015, Geneva
- The problem :
  - To calibrate a model on a very heterogeneous set



First example

- One (awarded) solution :
  - Identify the best calibration subsets : { S_1 , S_2 , … }
  - Learn one calibration by subset : { b_1 , b_2 , … }
  - Learn a discriminant model : X <-> {S_i}

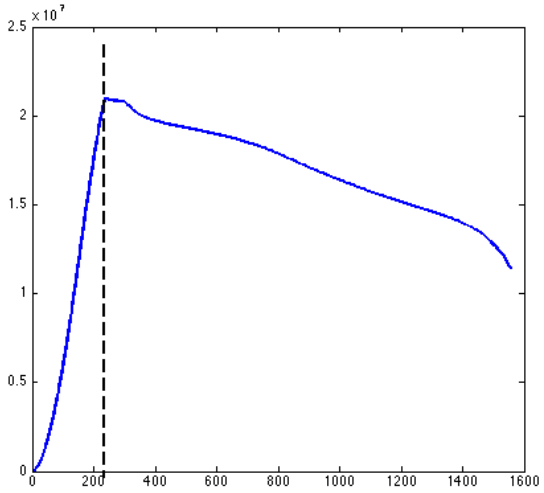
  - For a spectrum x of the test set :
    - Determine the subset S_i which x belongs to
    - Apply the corresponding model, b_i

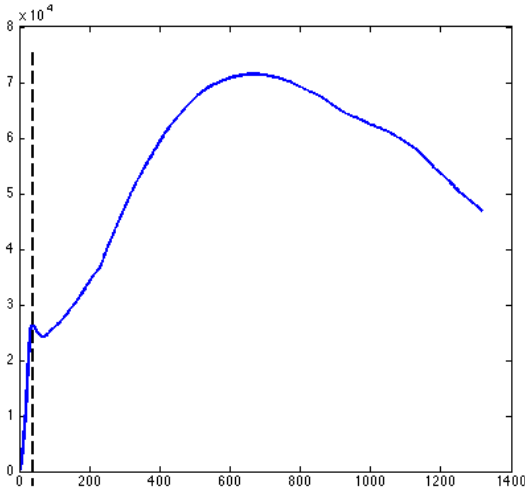
First example

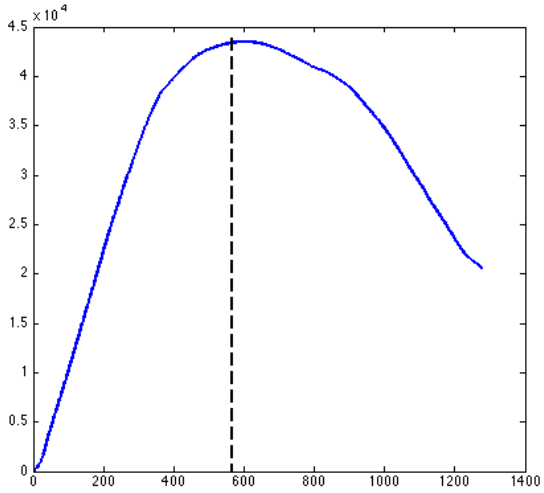
- How to find the best calibration subsets?

- Stepwise algorithm:
  1. Select one individual
  2. Select the individual which increases the most C1(X,Y) = trace(X'YY'X)
  3. As soon C1(X,Y) stops increasing, goto 1

First example







-> 6 groups:

238 34 759 395 110 19

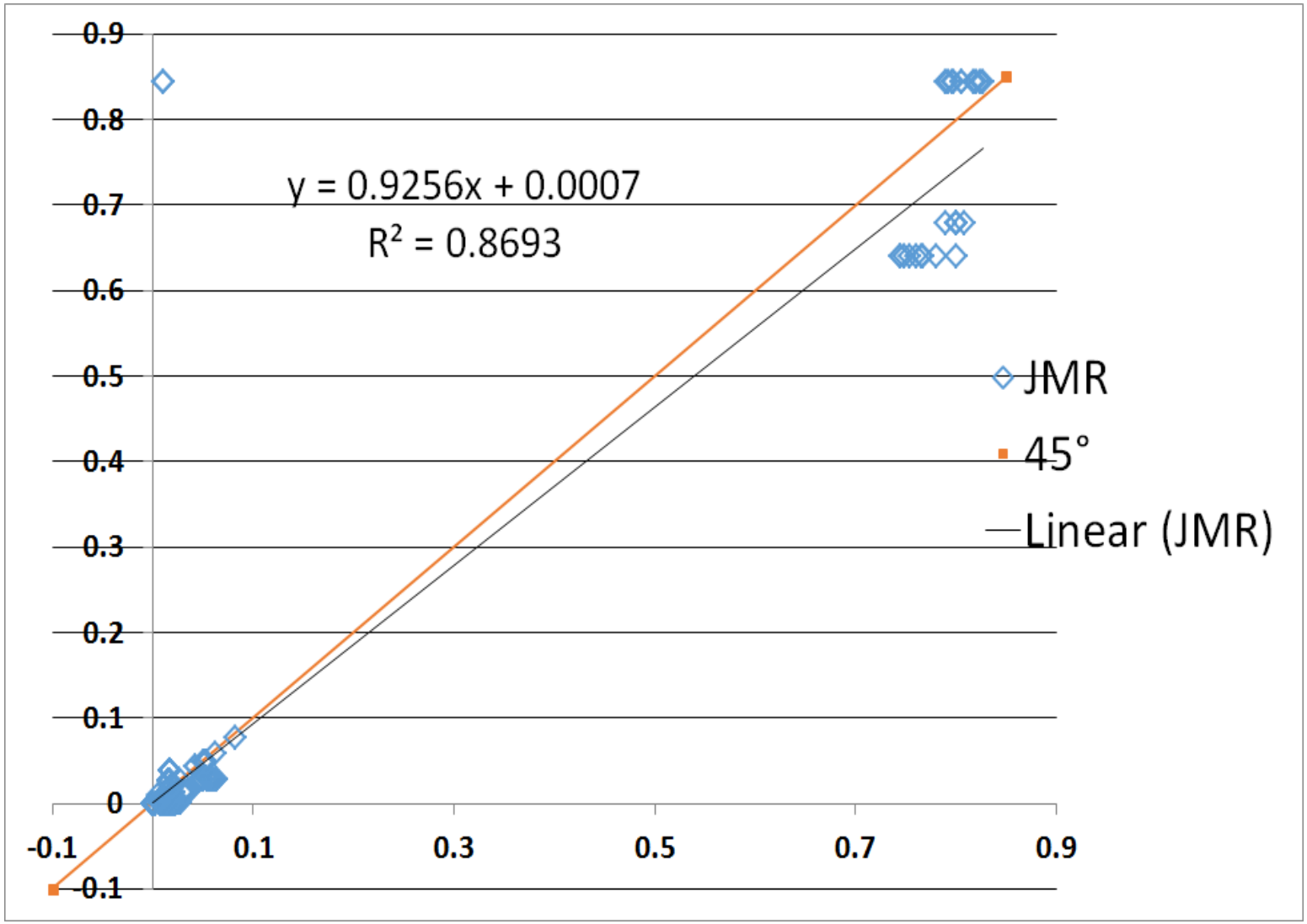
First example

Local calibrations:

Bloc	effectif	SECV	R^2	#LV
1	238	0.0238	0.997	10
2	34	0.0015	1.00	7
3	759	0.0062	0.941	10
4	395	0.0056	0.954	7
5	110	0.0074	0.972	7
6*	19	--	--	--

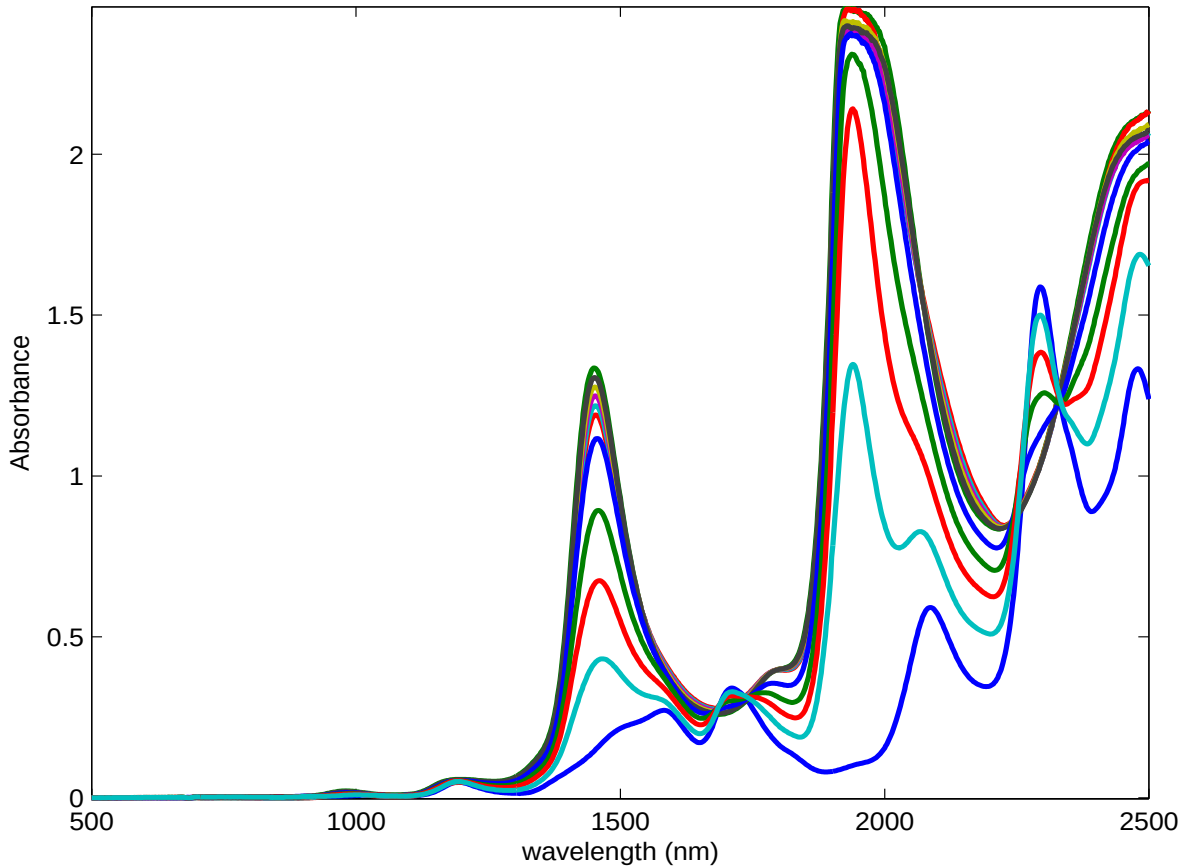
* : outliers

First example : results

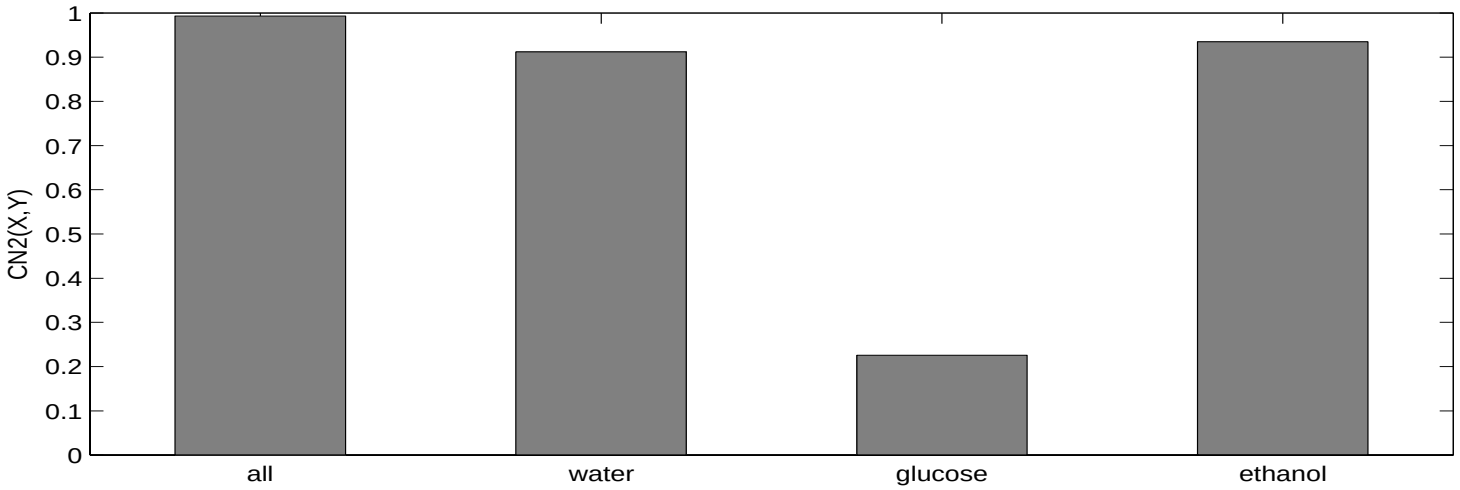


Second example

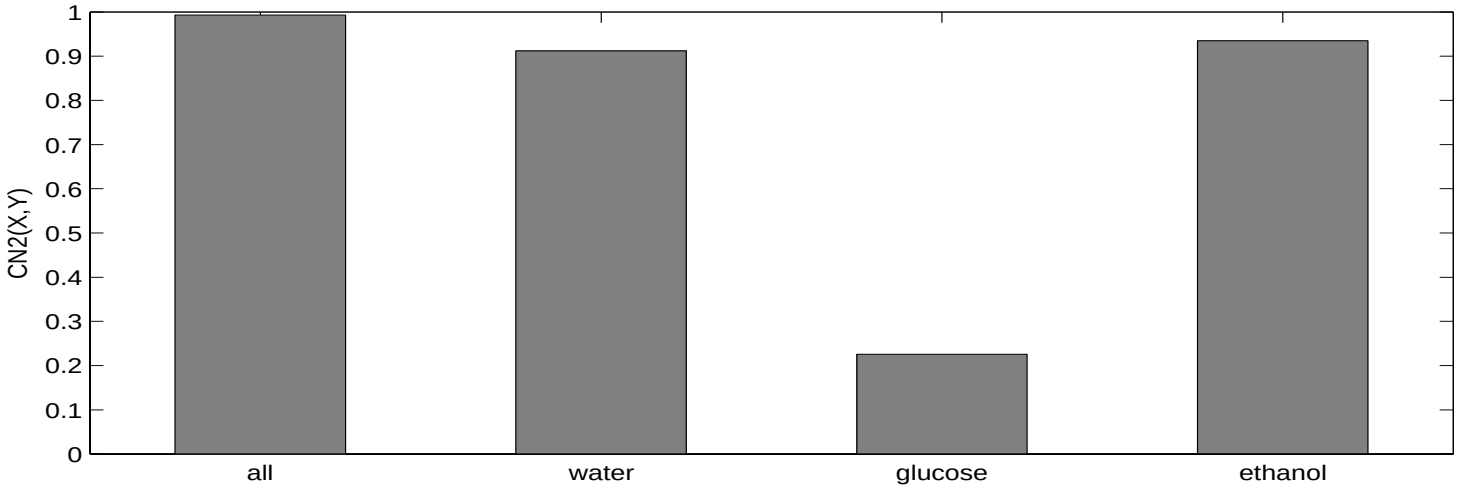
- Mix of Ethanol, glucose and water
- NIR spectra (Jasco V560, 500-2500 nm)

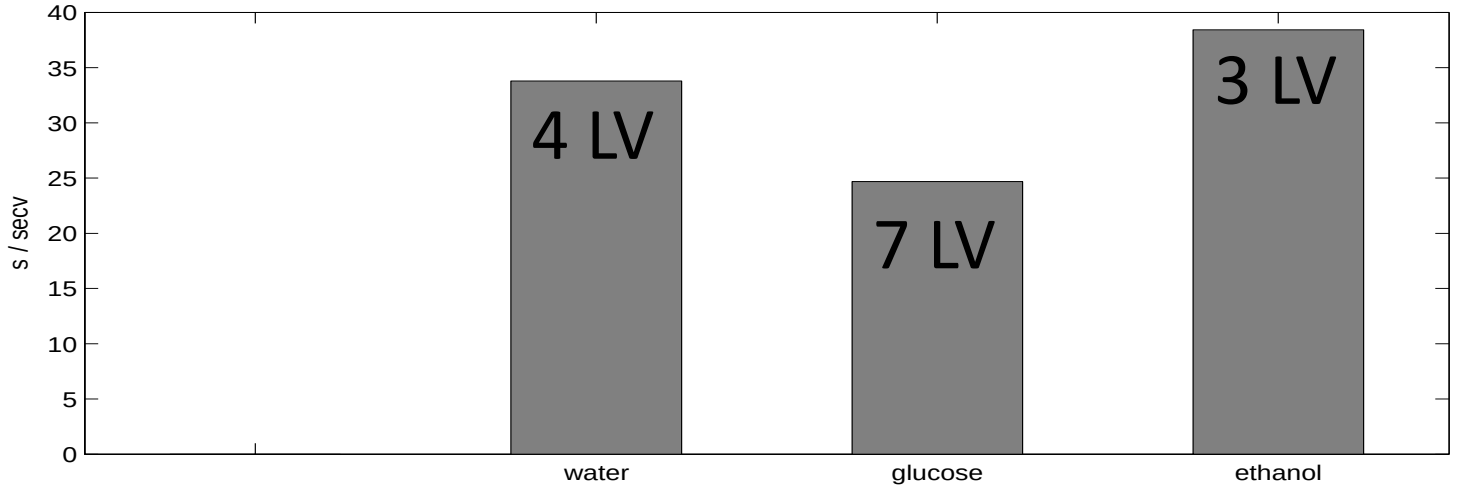


Second example



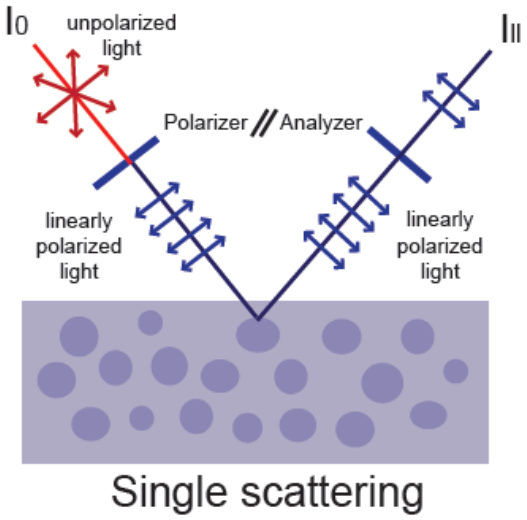
Second example



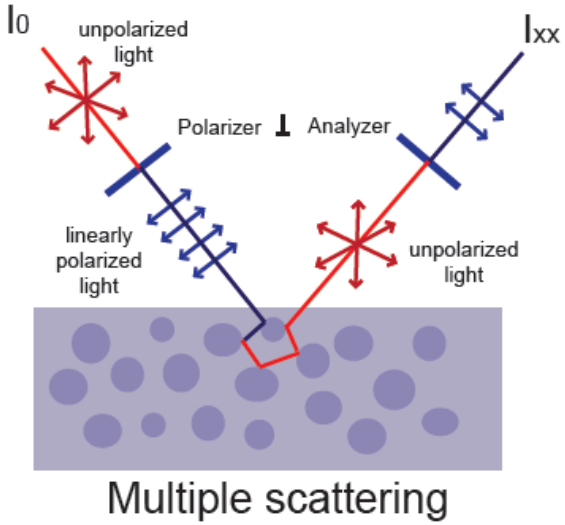


Third example

- PoLiS spectroscopy (Gobrecht *et al*, 2015) :



Single scattering



Multiple scattering

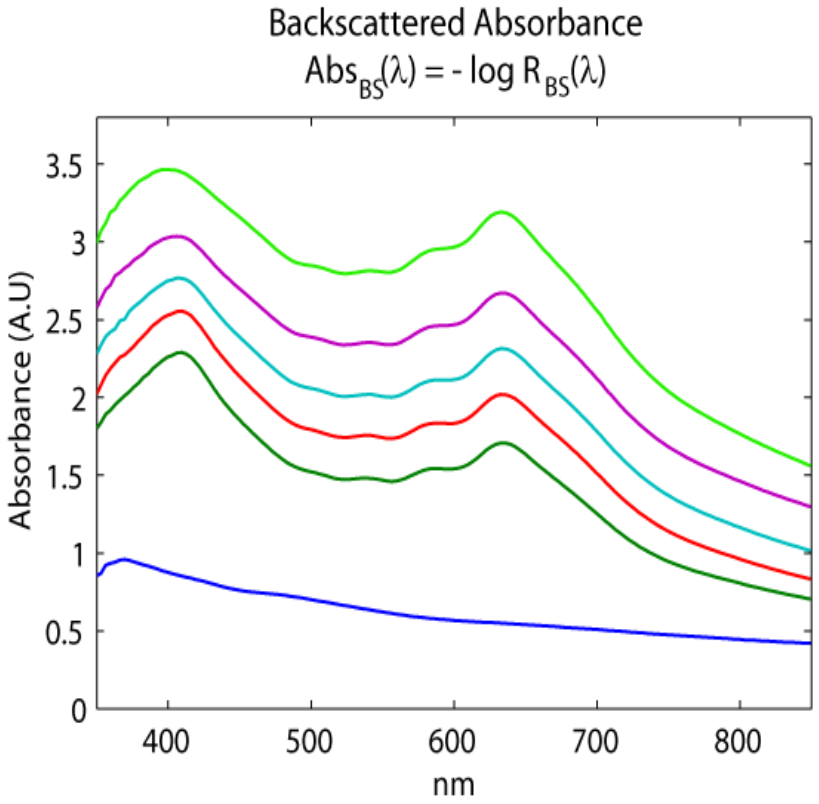
$$R_{BS}(\lambda) = R_{\parallel}(\lambda) + R_{\perp}(\lambda)$$

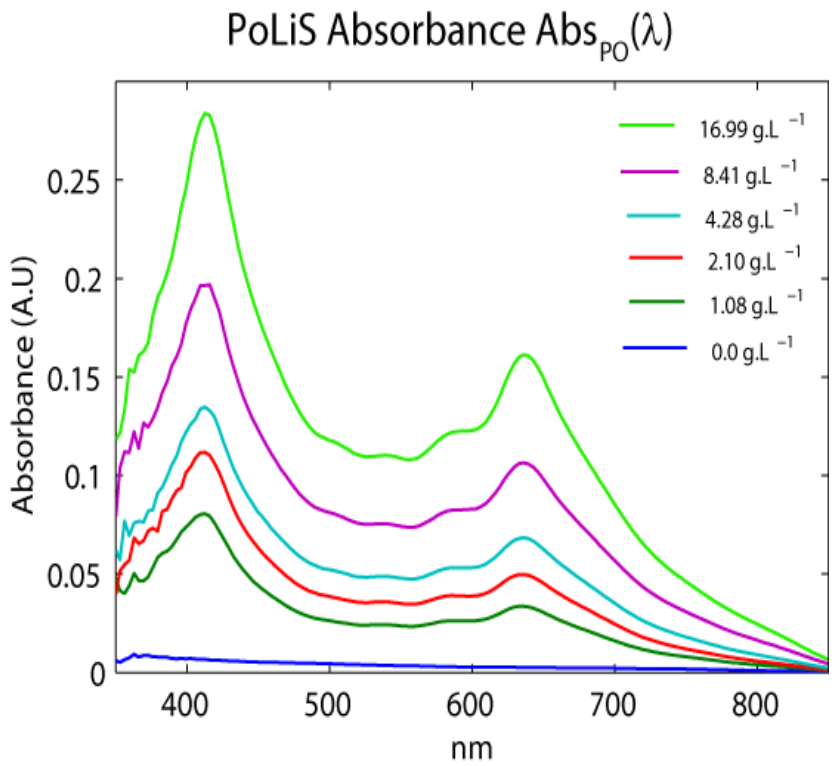
$$R_{SS}(\lambda) = R_{\parallel}(\lambda) - R_{\perp}(\lambda)$$

$$Abs_{PO}(\lambda) = -\log\left(R_{SS}(\lambda) + \sqrt{\left(1 - R_{SS}(\lambda)\right)^2 - \frac{R_{SS}(\lambda)}{R_{BS}(\lambda)}\left(1 - R_{BS}(\lambda)\right)^2} \right)$$

Gobrecht *et al*, 2015 : Analytica Chimica Acta. Vol. 853, 486-494

Third example





Raw absorbance spectra
CN2(X,Y) = 0.70

POLIS spectra
CN2(X,Y) = 0.88

$$CN2(X,Y) = tr(X'Y(Y'Y)^{-1}Y'X) / tr(X'X)$$

Conclusions

- This presentation introduces simple indexes
  - Measuring the closeness of X and Y
  - By the norm of the projection of X on Y

- Some potential to:
  - Select (sub)sets of calibration
  - Measure the sensitivity of spectral measurement
  - Compare preprocessing (not shown)

Perspectives

- Local regression ?
- Outlier detection ?
- Case of discrimination ?
  - CN2(X,Y) <-> Wilks Lambda