Selective Standard Normal Variate

G. Rabatel[1], F. Marini[2], B. Walczak[3], **J-M. Roger**[1]

[1] ITAP, Irstea Montpellier Centre, BP 5095 34196 Montpellier cedex 5, France. gilles.rabatel@irstea.fr
[2] Department of Chemistry, Univeristy of Rome "La Sapienza", P.le Aldo Moro 5, I-00185 Rome, Italy
[3] Silesian University, 9 Szkolna Street, 40-006 Katowice, Poland

Outline

- Introduction / theory
- Example on simulated data
- Example on real data
- Conclusion

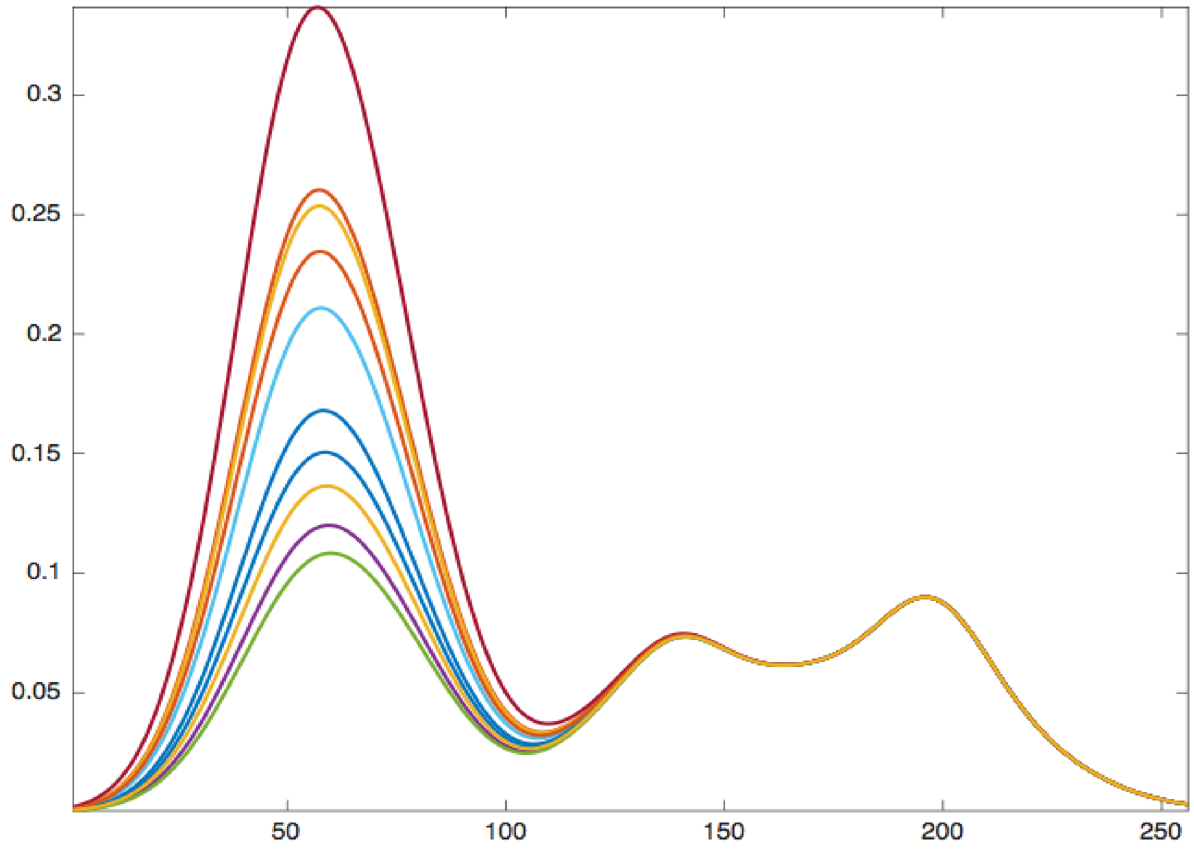
Introduction

- Beata Walczak told you at Genève in 2015
- Tom Fearn told you again at Namur in 2016

The normalisation must be used with caution

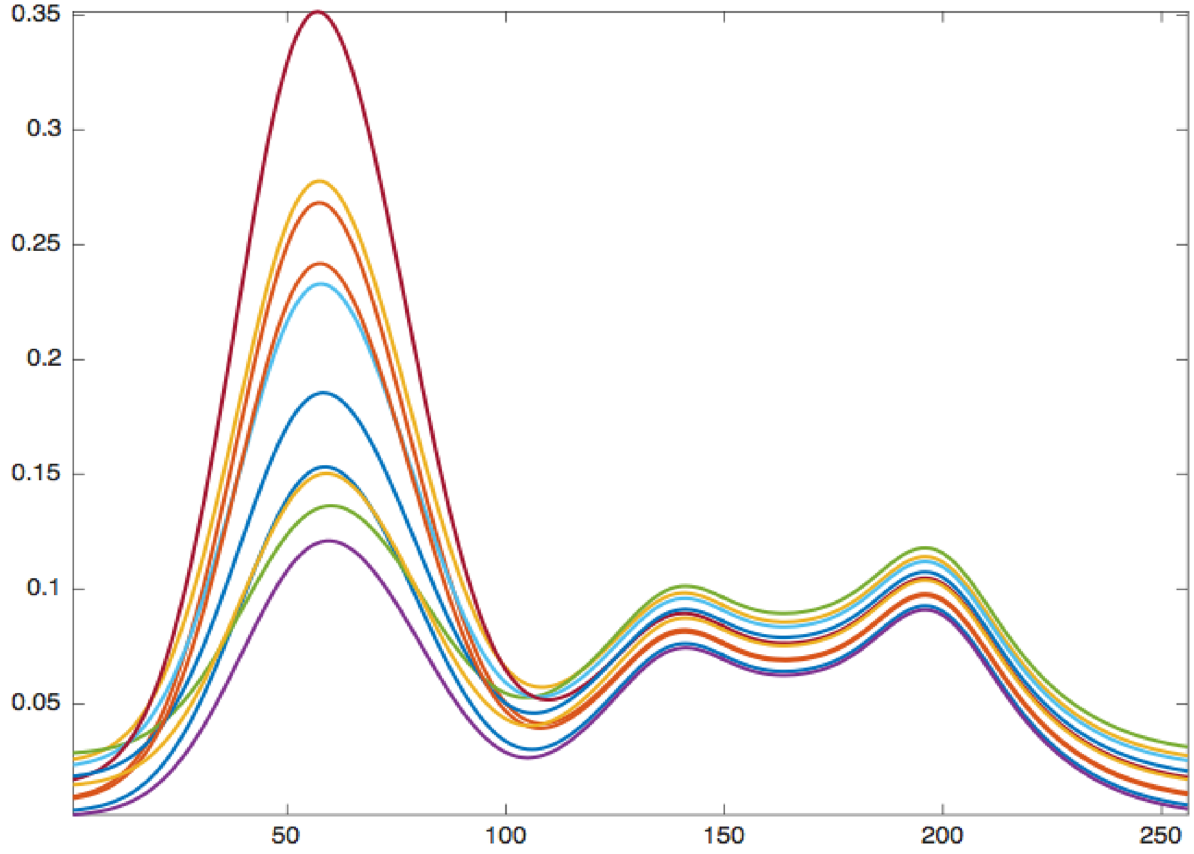
Because it has side effects !!

Introduction: a simulated example



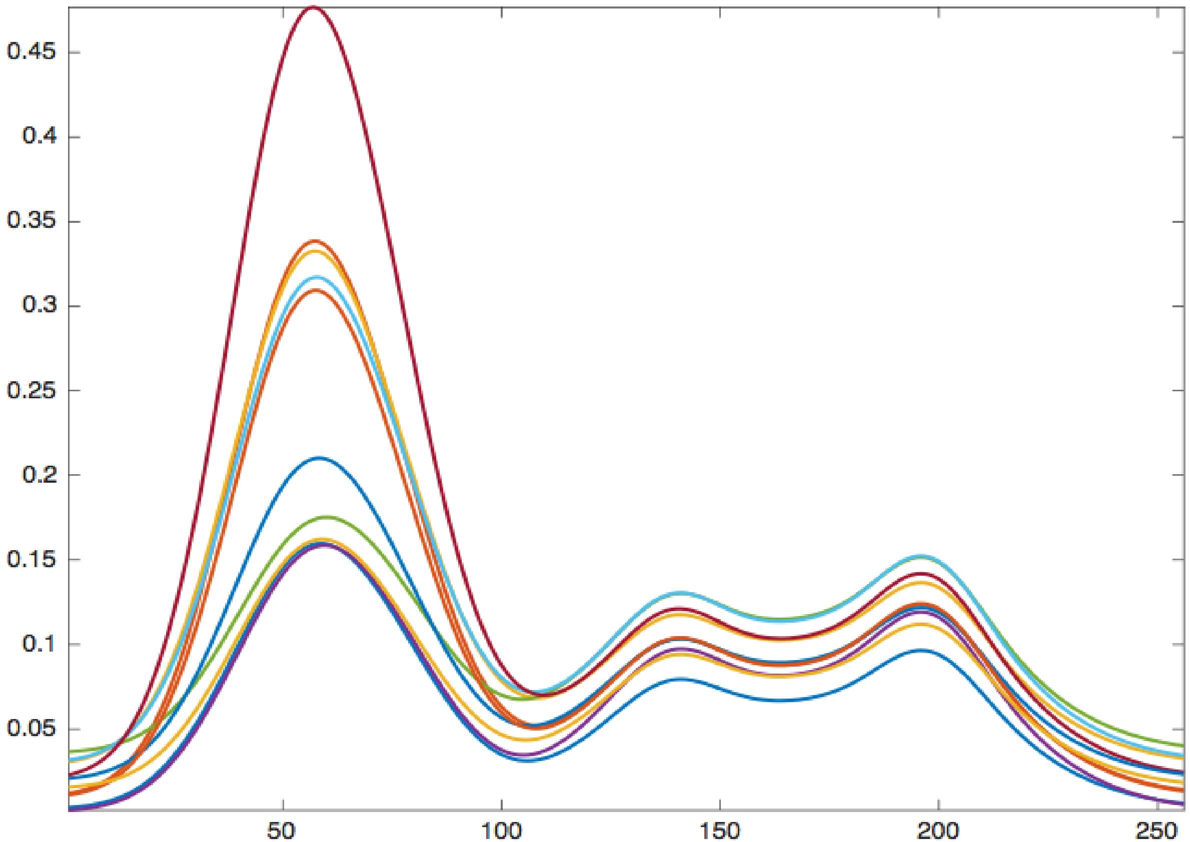
Spectra without any additive or multiplicative effect
One peak related to Y, two not

Introduction:
a simulated example



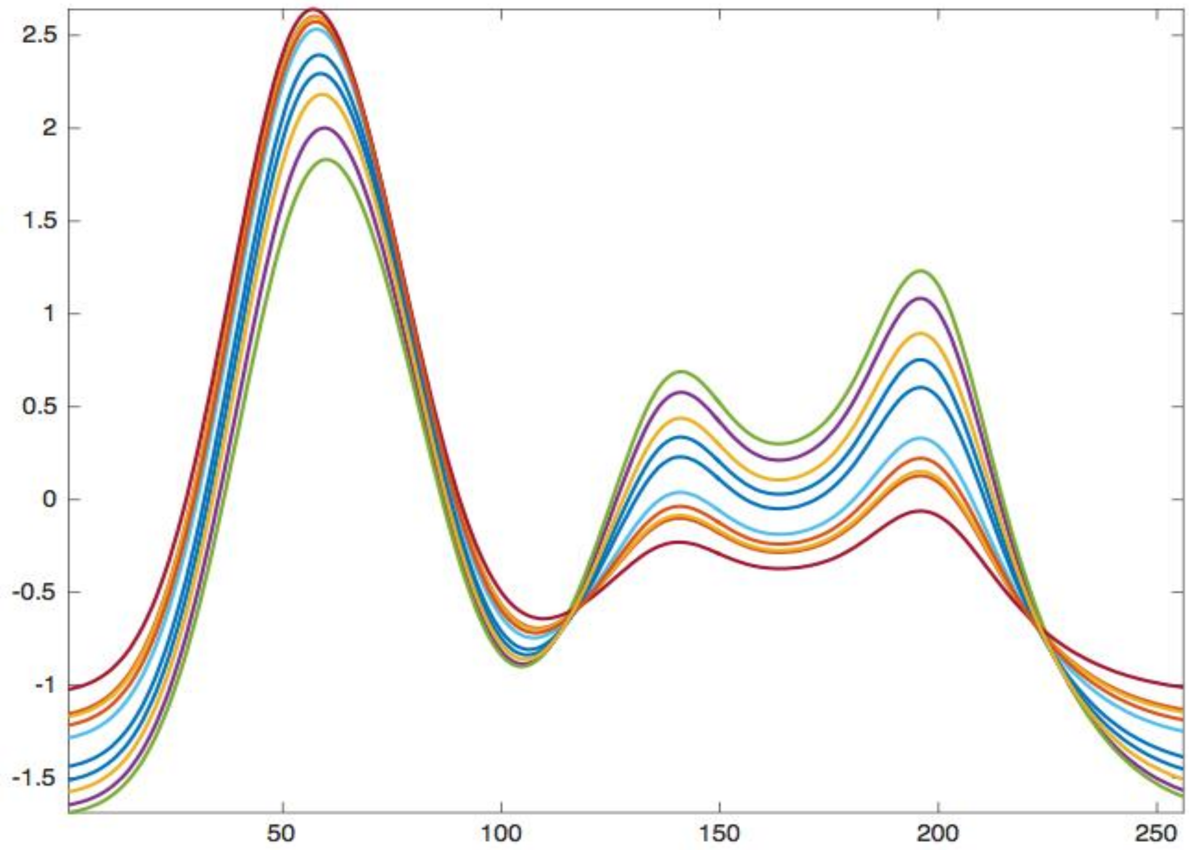
Let add baselines

Introduction:
a simulated example



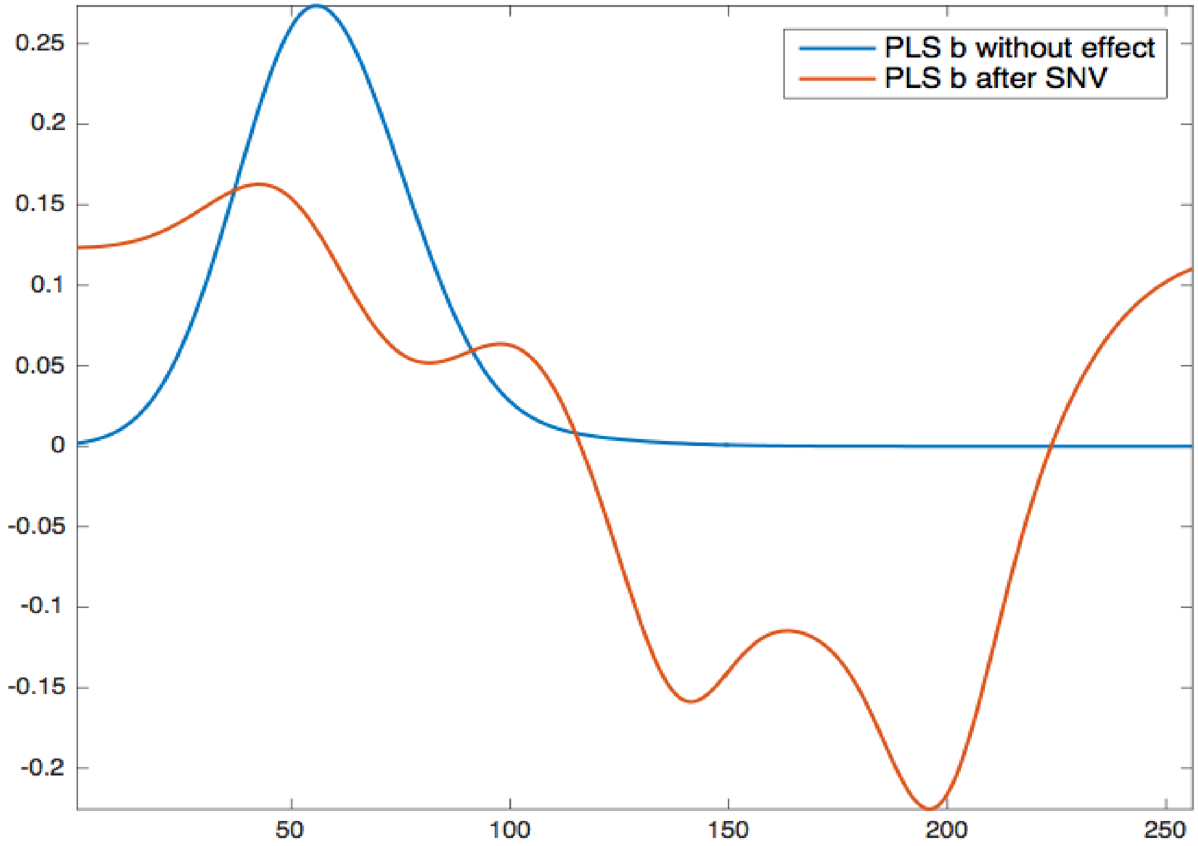
And a multiplicative effect

Introduction:
a simulated example



And let apply SNV

Introduction: a simulated example



Model performances are good (on calibration set)
But the model itself is erroneous

Theory

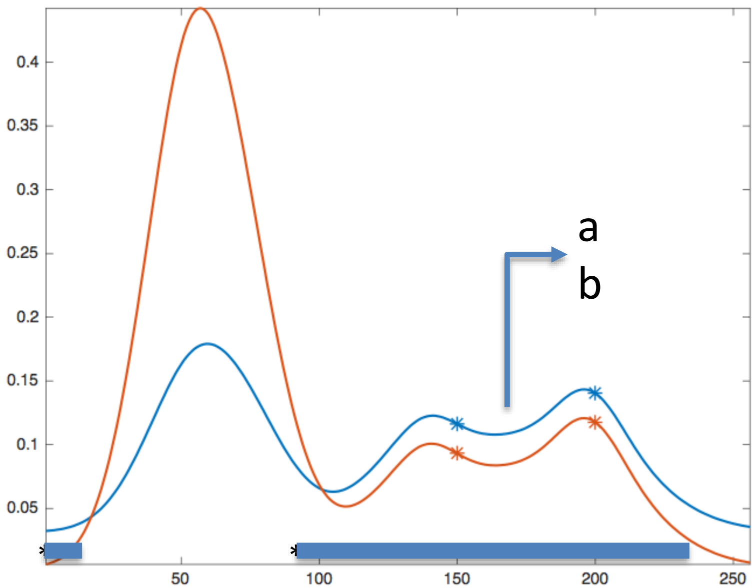
- What happens?
- SNV estimates :
  - the multiplicative effect as the standard deviation of the spectrum
  - the additive effect as the mean of the spectrum
- But these statistics depend also on Y
- SNV tends to dilute the information along the whole spectrum

Theory

- A solution :
  - To calculate standard deviation and mean on wavelengths little related to **Y**
  - To normalize the spectrum with these values
- Or, more generally:
  - To calculate diagonal matrix **W** of weights between 0 (no selection) and 1 (complete selection)
  - To calculate the normalisation on **Wx** and apply it to **x**

An algorithm using RANSAC

(Fischler and Bolles, 1981)



Let take a couple of spectra i, j

Let take a couple of wavelengths k,l

Calculate coefficients a, b so that $(x_{ik},x_{il}) = a(x_{jk},x_{jl}) + b$

Retrieve the set of wavelengths that respect the same relationship, given a tolerance

After some iterations, retain the largest set

One gets a partition of the wavelengths in two subsets:

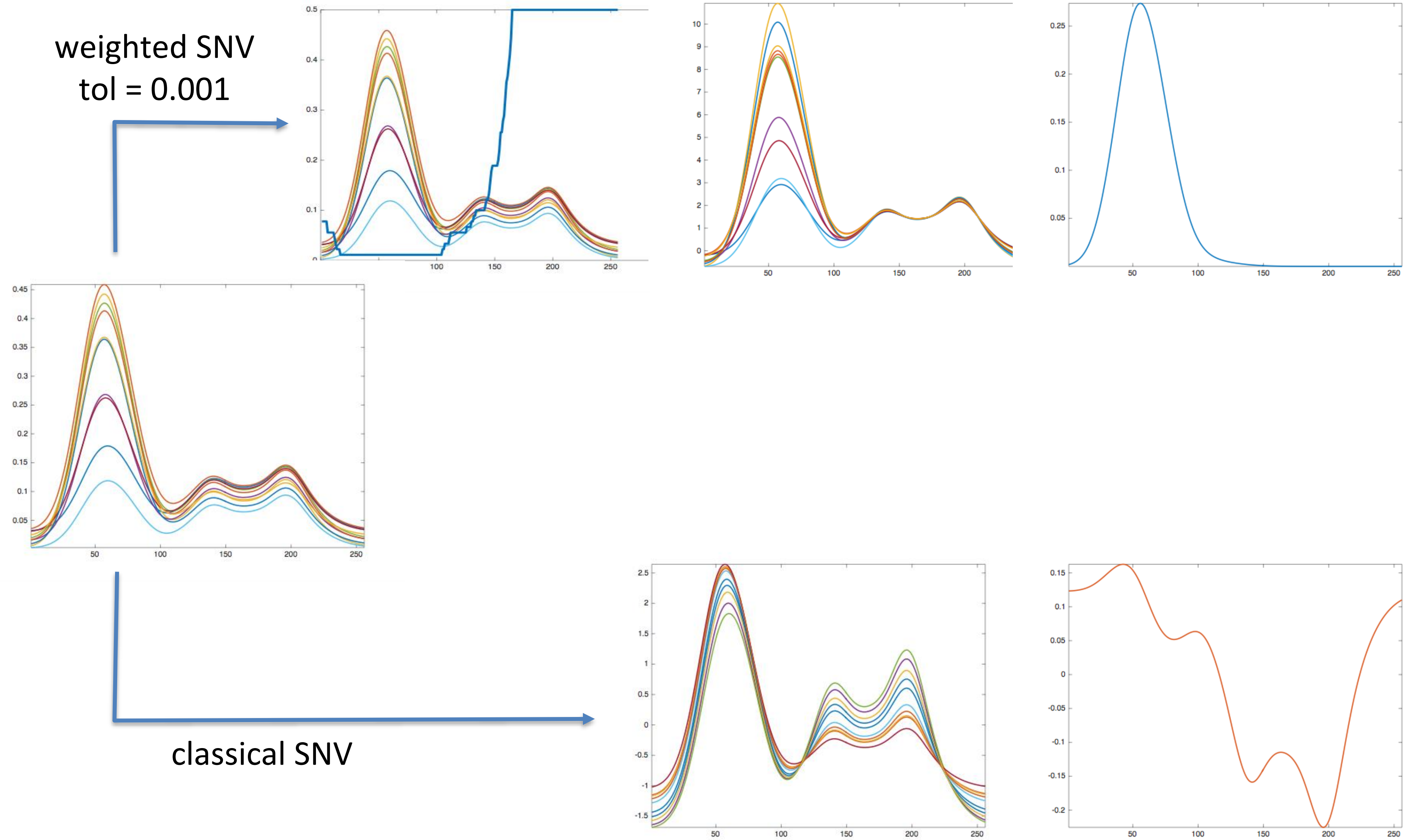
the INLIERs, which all share the same coefficients

the OUTLIERs

We propose to calculate w_i = p(wavelength i is an INLIER)

One estimate this probability by drawing couples of spectra in **X**

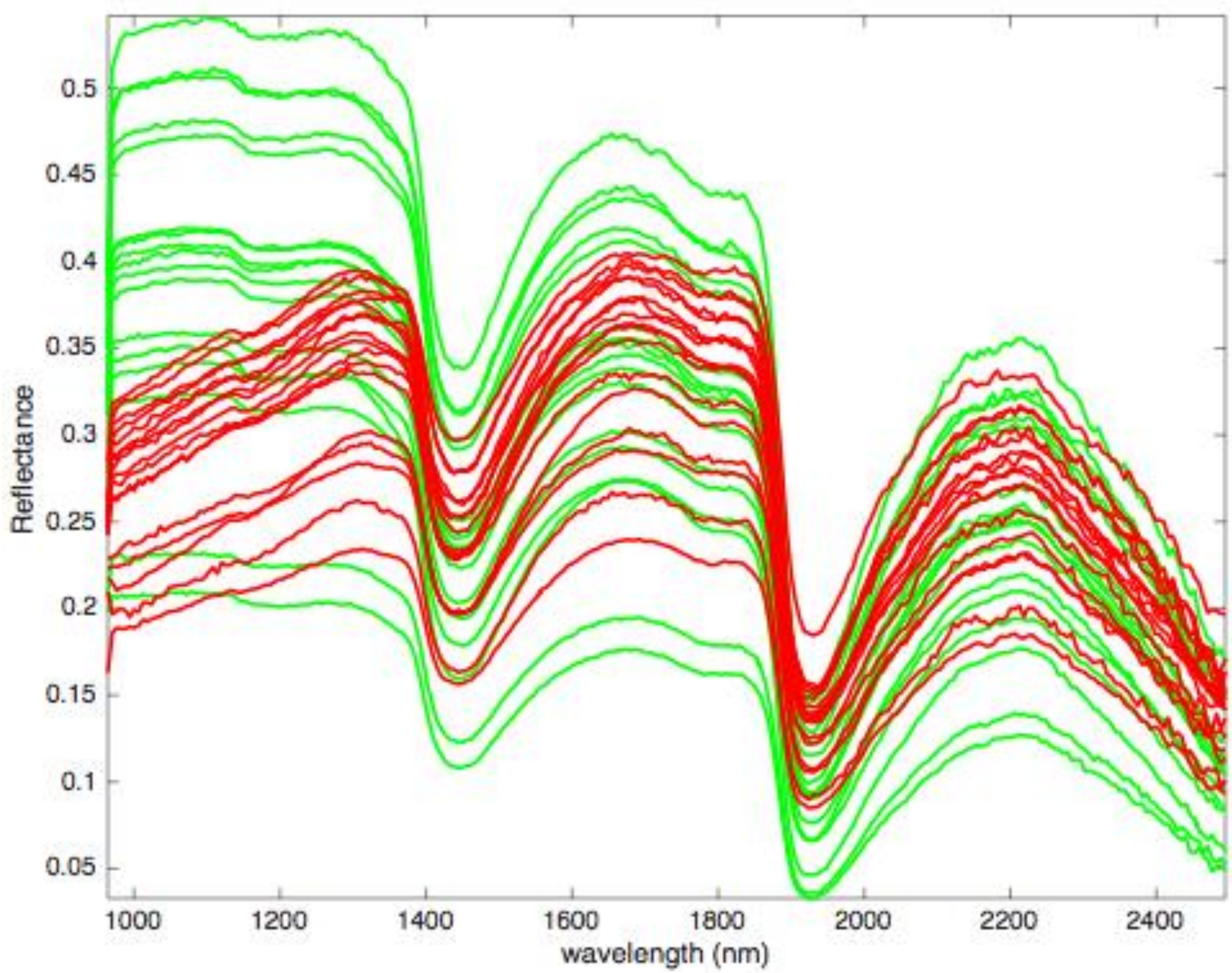
Results on simulated data



Real example

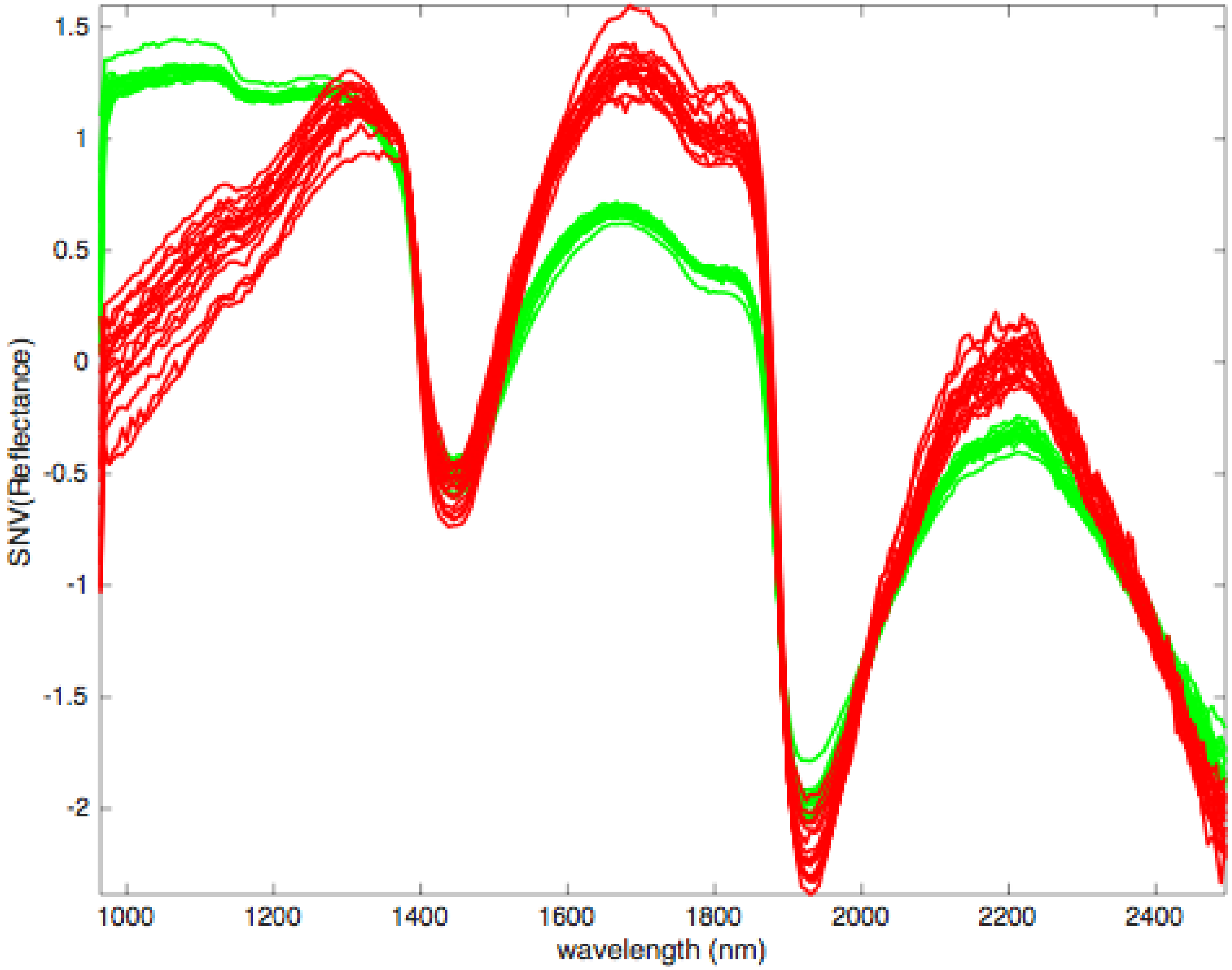
- Data : apple tree leaf spectra
- Images acquired with an NEO SWIR hyperspectral camera; 1000 - 2500 nm
- Each spectrum is the mean of pixels from an area
- Two classes :
  - healthy
  - scab disease spot

Real example



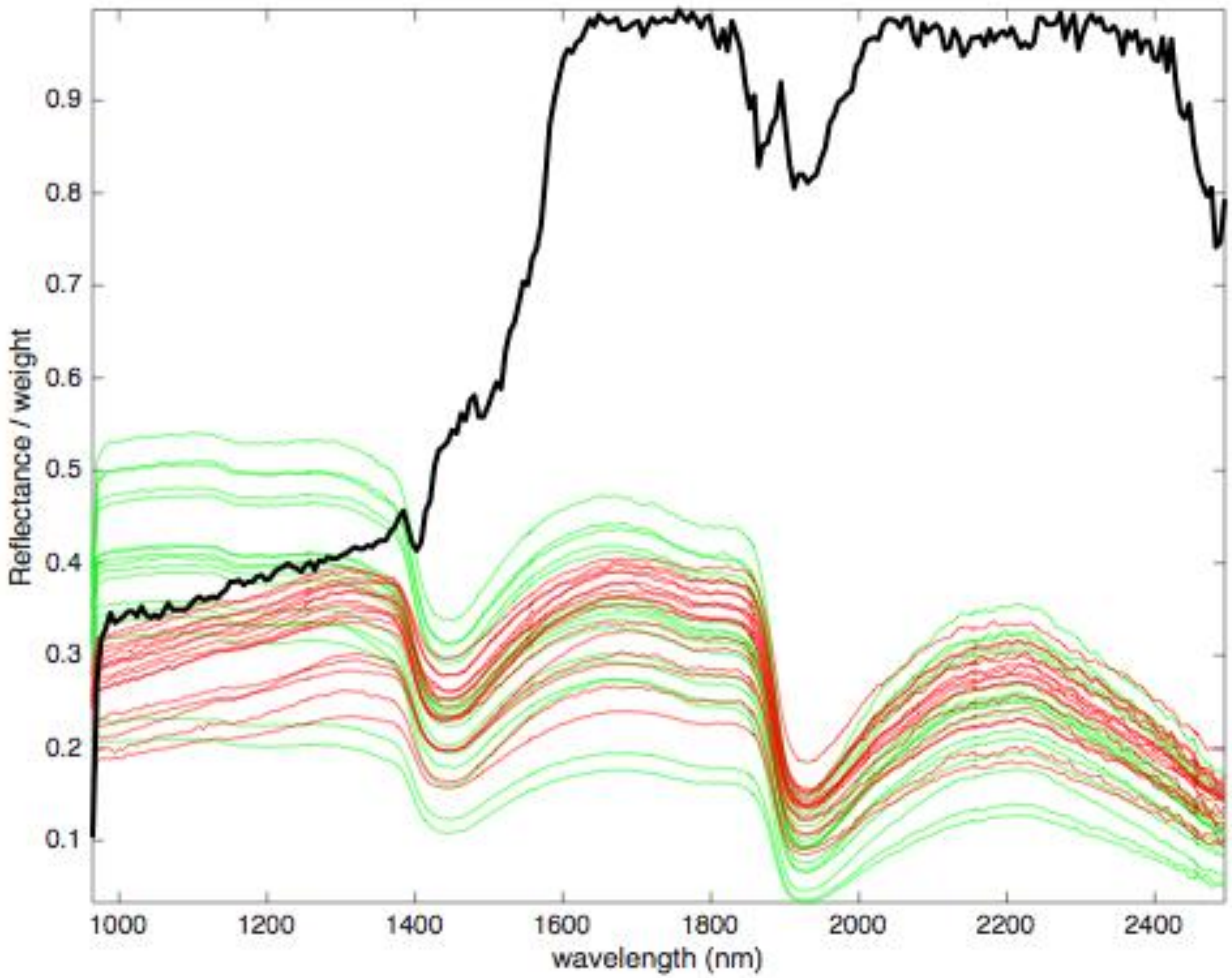
healthy (green) and scab (red) spectra

Real example



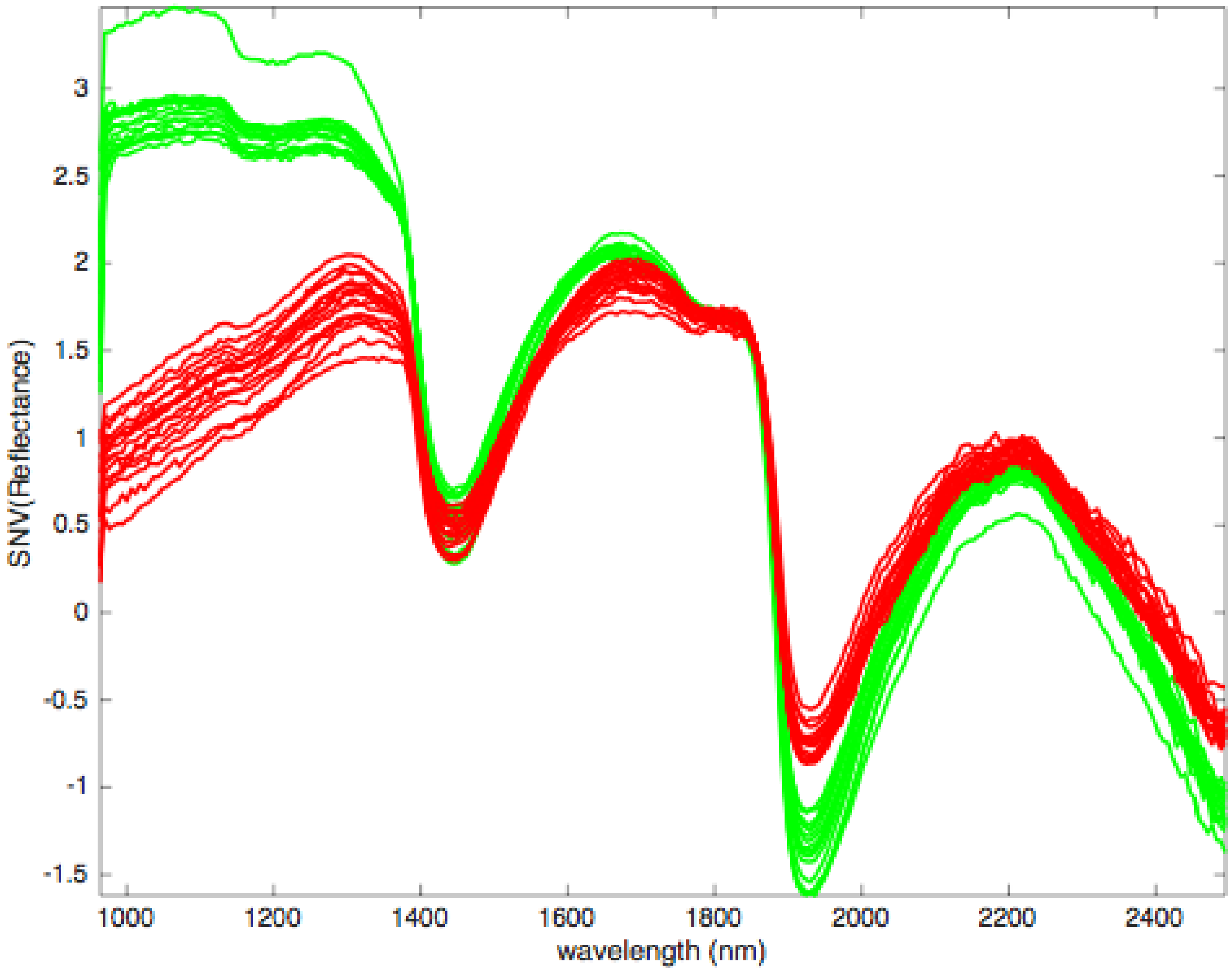
spectra processed by classical SNV

Real example



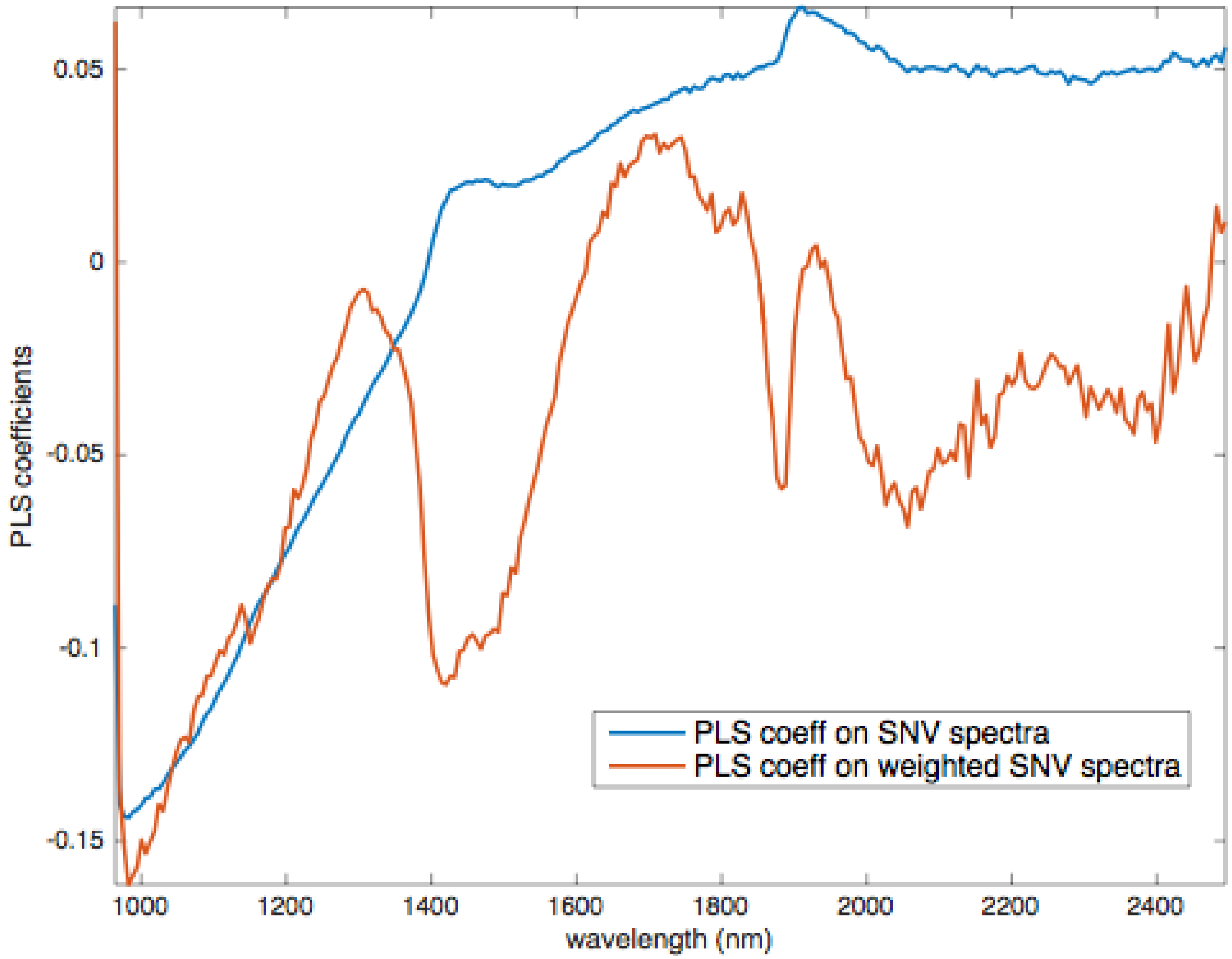
weights yielded by the algorithm ; tol = 0.01

Real example



spectra processed by weighted SNV

Real example



PLS models on the two sets, (2 latent variables)

Conclusions

- The normalisation (e.g. SNV) induces undesirable alterations
- This does not change the model performances, but can severely affect the loadings
- A solution consists of weighting the variables regarding the normalisation
- An algorithm is proposed
  - Results are satisfactory
  - Do not need reference spectrum, as MSC, PQN, …
  - The weights found can be easily applied to new spectra
  - It must be compared to other methods, as those using robust regressions (p.ex. RSNV, Guo et al, 1999)
  - It should be adapted to other type of effects
  - It must be optimized and automatized

Thanks for your attention