Some aspects of SVM Regression: an example for spectroscopic quantitative predictions

Alice Croguennoc, Jordane Lallemand,
Sylvie Roussel

Ondalys - 4 rue Georges Besse, 34830 Clapiers, France - acroguennoc@ondalys.fr

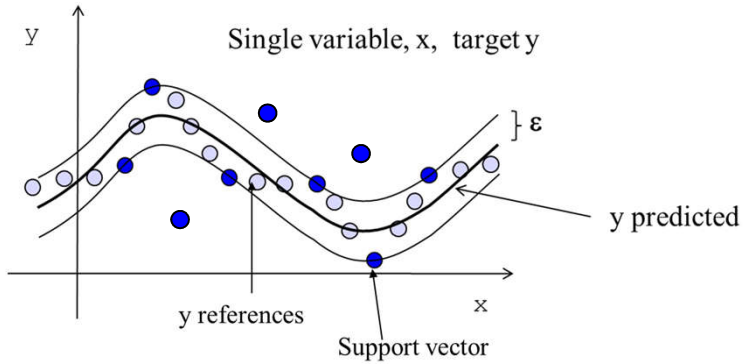
Summary

- What are SVM?
  - SVM Regression
  - Criteria to optimize
  - The risk of overfitting
  - Two algorithms
- Fit non-linear spectroscopic data with SVM
  - Data set
  - Non-linearity issues
  - Spectroscopic pretreatments
  - X-Data compression
  - Training Set size
  - Comparison of algorithms
- Conclusions

What are SVM?

- Principle of Support Vector Machines
  - Supervised methods based on margins
  - Only a few samples are used for the calculation of the final model

    = samples defining the margins = support vectors



Source : Eigenvector Research Inc.

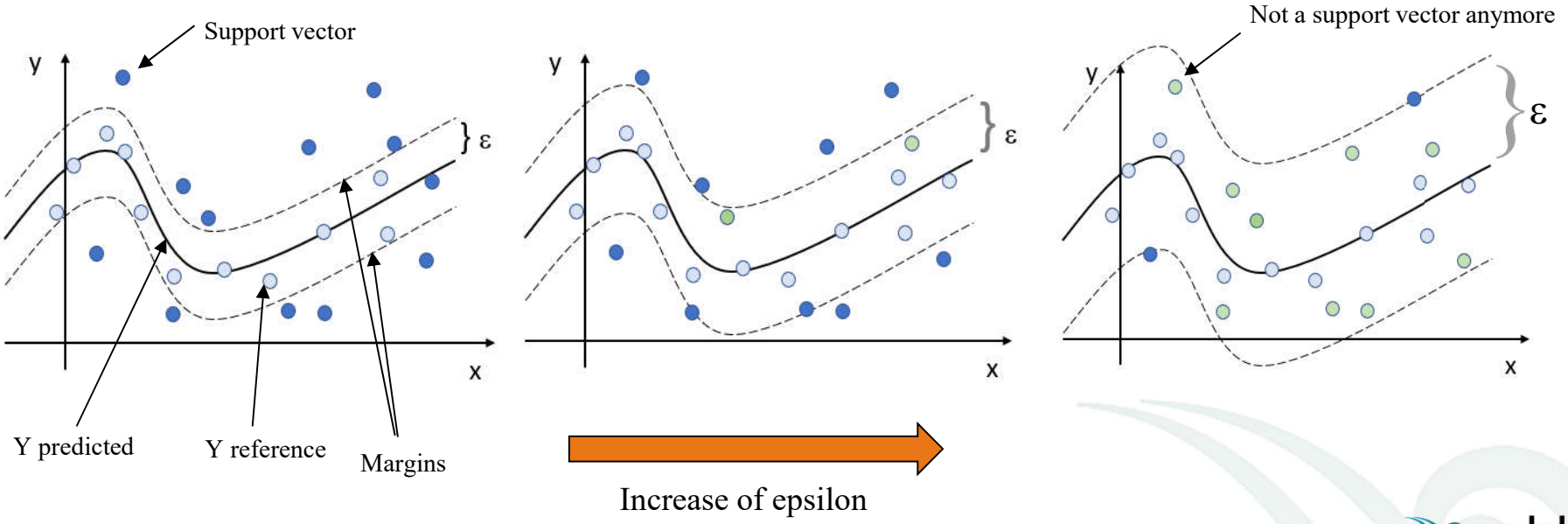
- How to cope with non-linearities
  - Non-linearities can be modeled thanks to data transformation into a kernel = similarity between samples
    - Gaussian kernel :

$$K(x_i, x_j) = e^{\frac{-\|x_i - x_j\|^2}{\sigma^2}}$$

- Parameters to adjust: C, ε, σ^2

Criteria to optimize

- Epsilon ε :
  - Has a direct effect on the size of the margin: a small epsilon leads to a tight margin (tighter fitting to the calibration set)
    - Directly correlated to the number of support vectors selected
    - Low impact on the error (on a 'reasonable' range)



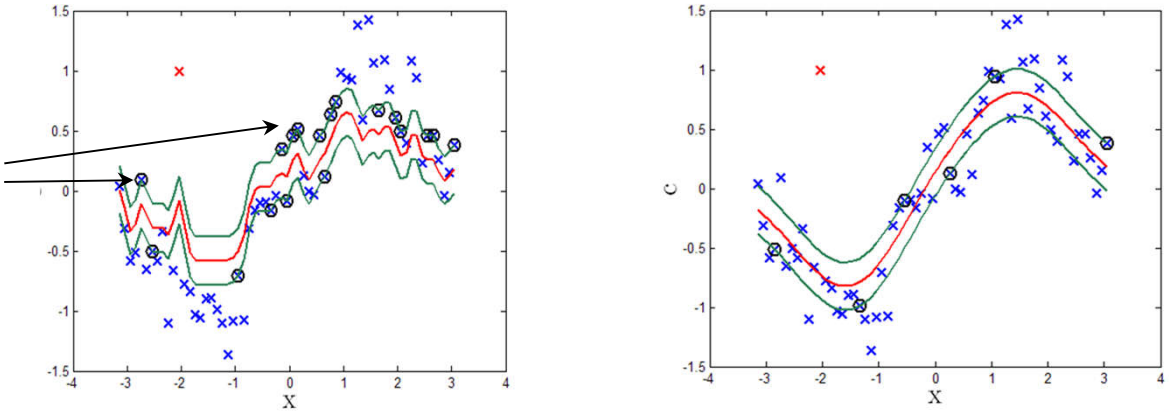
Criteria to optimize

- Sigma σ^2:
  - Determines the degree of non-linearity of the model
  - ➢ Smaller sigma allows the SVM to represent stronger non-linearity
  - ➢ Larger sigma tends towards linear kernel behavior
    - Criterion that seems the more prone to cause overfitting (or underfitting)
    - Linked to the level of X values

Small sigma

High sigma

Support vectors



Criteria to optimize

- Cost C : *penalty (or regularization) parameter*
  - High cost: errors very impactful, dangerous if outliers
  - Low cost: might lead to under-fitting

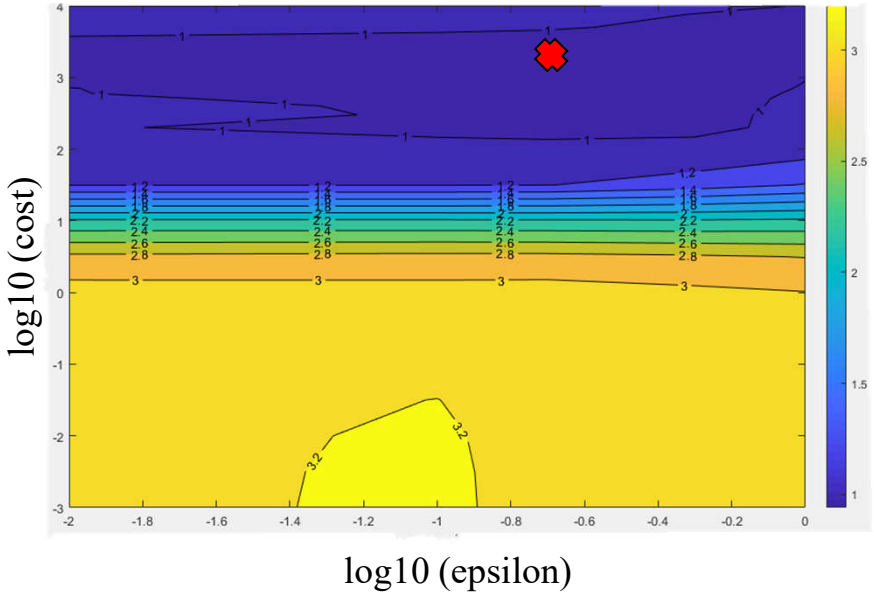
$$min\left(C\sum_{i=1}^{n}\xi_i + \frac{b^T b}{2}\right) \qquad \text{With :} \qquad \xi_i = \begin{cases} 0 & if\ |y_i - \hat{y}_i| < \varepsilon \\ |y_i - \hat{y}_i| & otherwise \end{cases}$$

b – coefficients of the regression

Optimization map

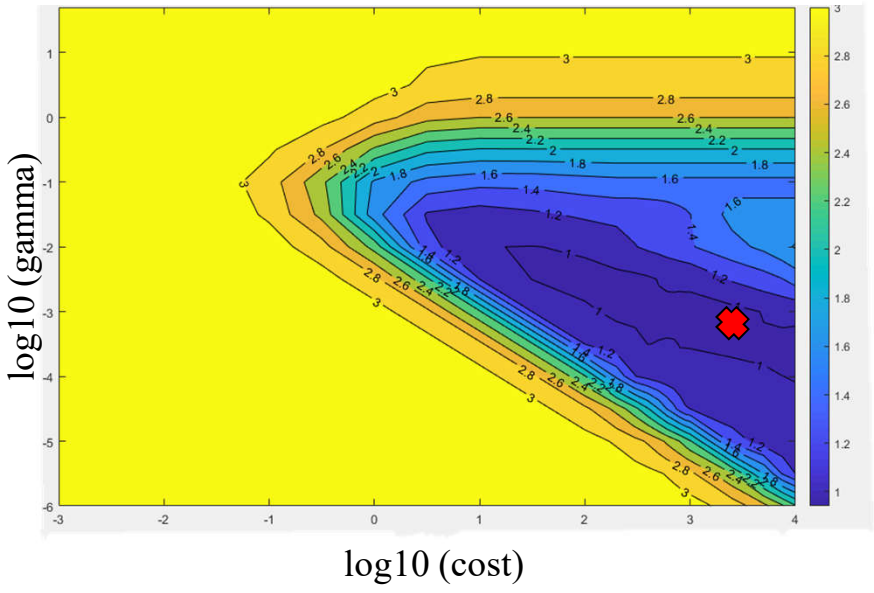
Cross validation – mean squared error

Gamma fixed



Cross validation – mean squared error

Epsilon fixed

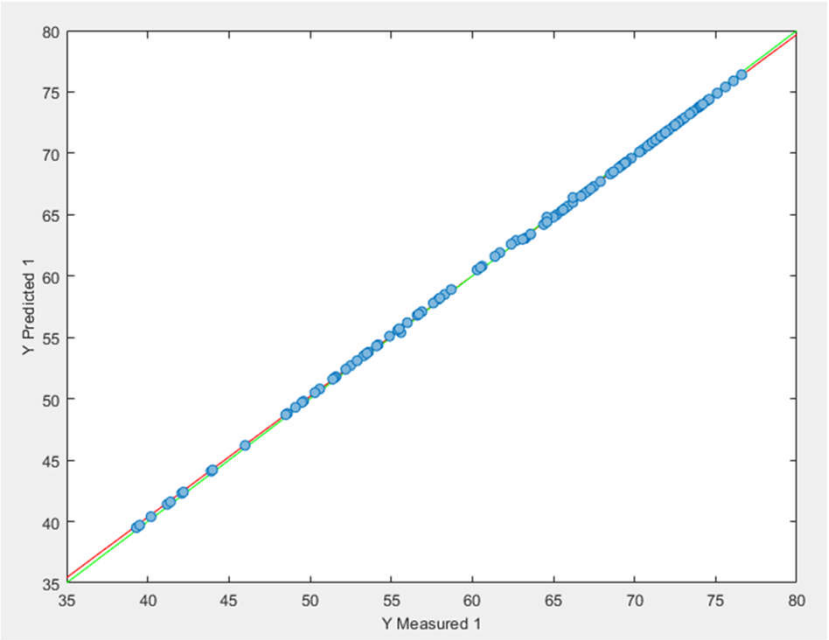


(gamma used instead of sigma : $\gamma=1/\sigma^2$)

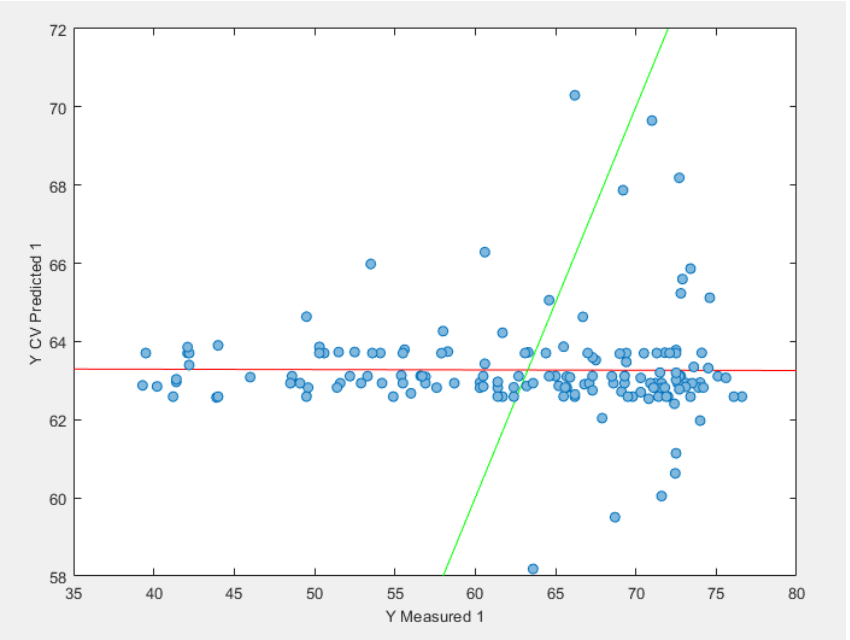
SVM Regression: the risk of overfitting

- Warning: unproperly tuned, SVM can overfit very quickly !

Example: Y randomized: still possible to achieve good results on the calibration set



Calibration set



Cross validation results

Software : PLS Toolbox (Eigenvector Research Inc.)

SVM Regression: 2 algorithms

- SVM-R (or SVR): minimization of the sum of the errors higher than epsilon

$$min\left(C\sum_{i=1}^{n}\xi_i+\frac{b^Tb}{2}\right) \qquad \text{With :} \qquad \xi_i=\begin{cases}0 & if\ |y_i-\hat{y}_i|<\varepsilon\\ |y_i-\hat{y}_i| & otherwise\end{cases}$$

- LS-SVM: minimization of the sum of squared errors
  - All samples are support vectors

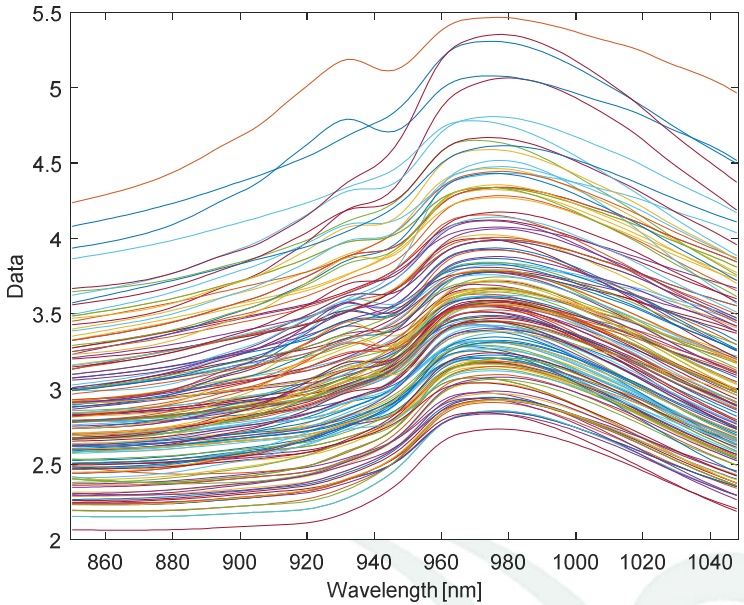
$$min\left(C\sum_{i=1}^{n}(y_i-\hat{y}_i)^2+\frac{b^Tb}{2}\right)$$

Summary

- What are SVM?
  - SVM Regression
  - Criteria to optimize
  - The risk of overfitting
  - Two algorithms
- **Fit non-linear spectroscopic data with SVM**
  - Data set
  - Non-linearity issues
  - Spectroscopic pretreatments
  - X-Data compression
  - Training Set size
  - Comparison of algorithms
- Conclusions

Data set

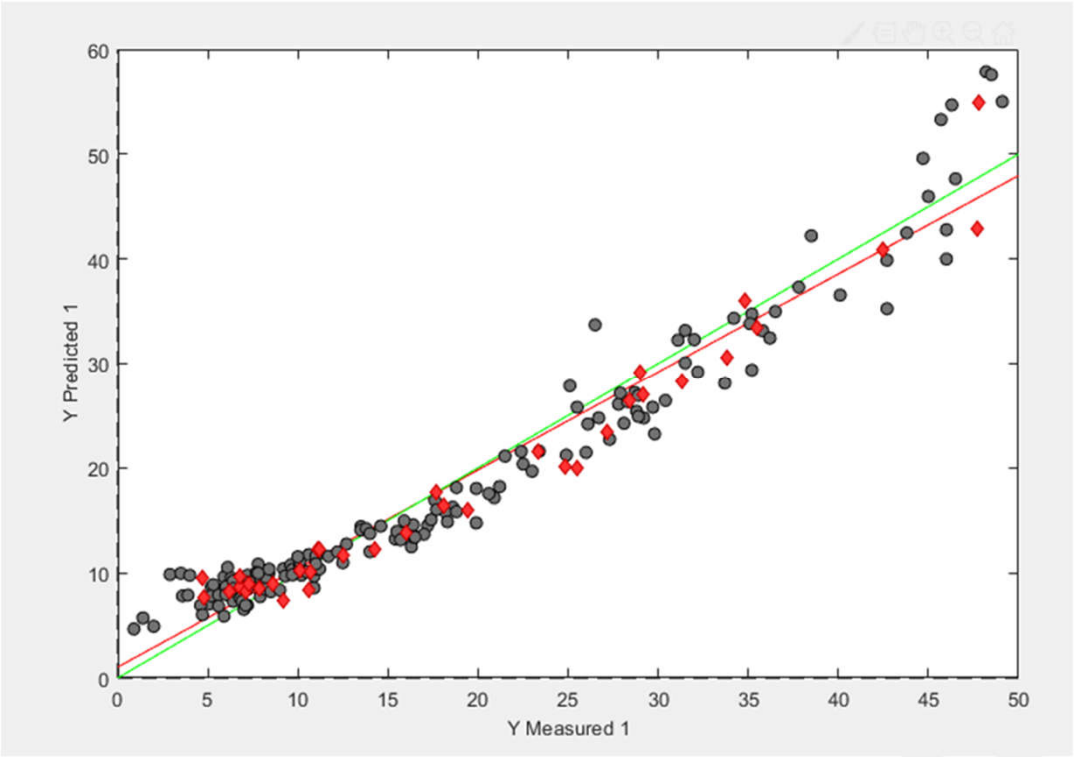
- Near Infrared spectroscopic data on raw meat. Three quantitative chemical results available: moisture content, fat content, protein content.
  - Instrument: FOSS Tecator Infratec Food and Feed Analyzer
    - Range: 850-1050 nm
  - Dataset
    - Training set : 158 samples
    - Test set : 35 samples
    - Cross-validation 2 blocks



Source : http://lib.stat.cmu.edu/datasets/tecator

Non-linearity issues

- Strong non-linearities : example of PLS linear modeling

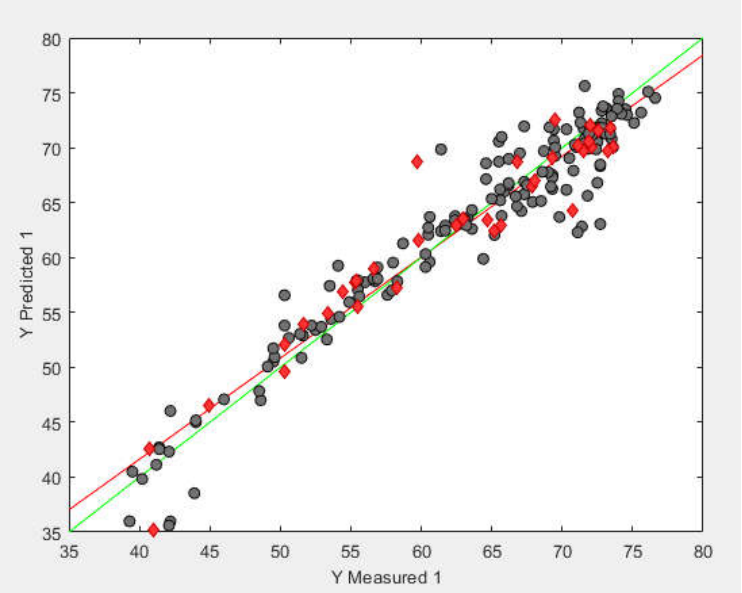


Fat prediction using PLS model

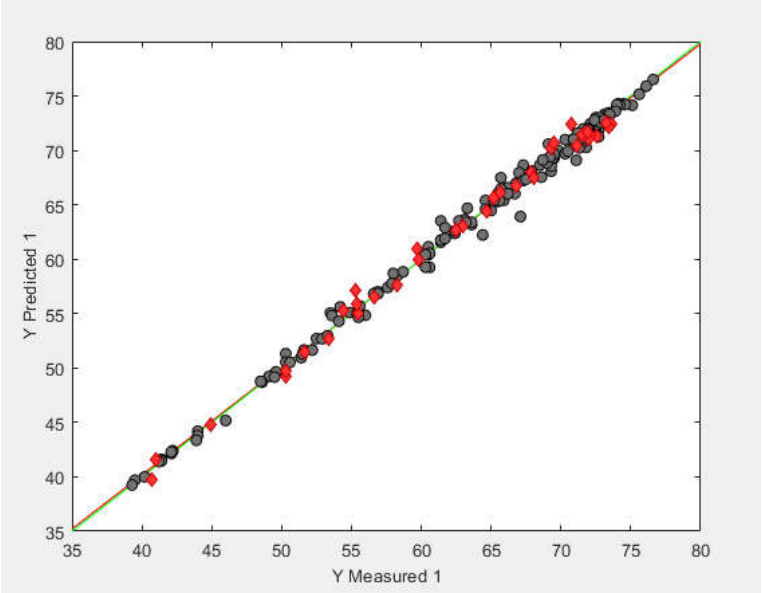
Software : PLS Toolbox (Eigenvector Research Inc.)

Spectroscopic pretreatments

- 4 modalities tested: raw, D1, D2, SNV
- Strong impact of the pretreatment on the results with SVM-R



Y = moisture
Raw X



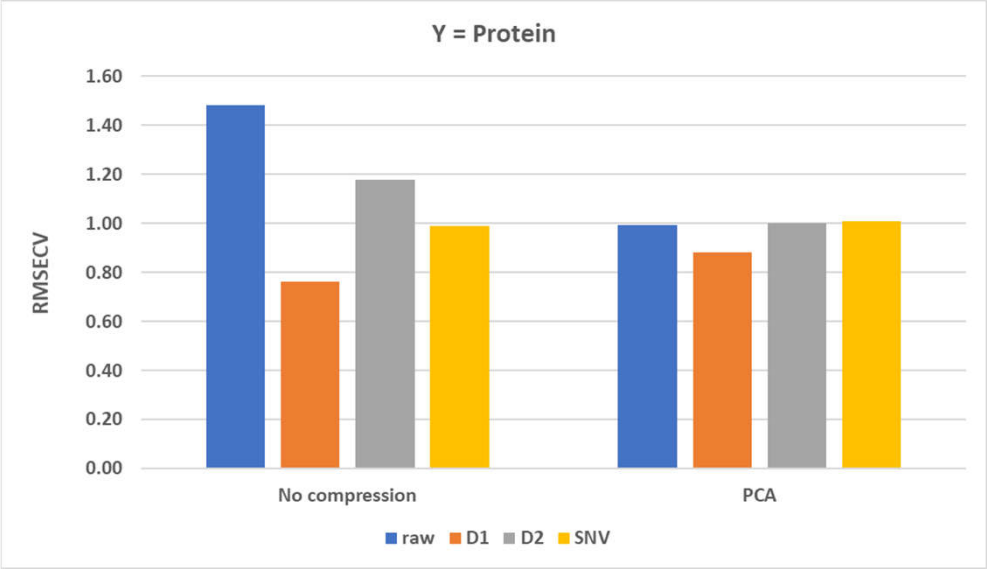
Y = moisture
SNV on X

➢ SVM are based on sample similarities, thus any perturbation should be corrected (e.g. scattering effect)

Software : PLS Toolbox (Eigenvector Research Inc.)

X-data compression

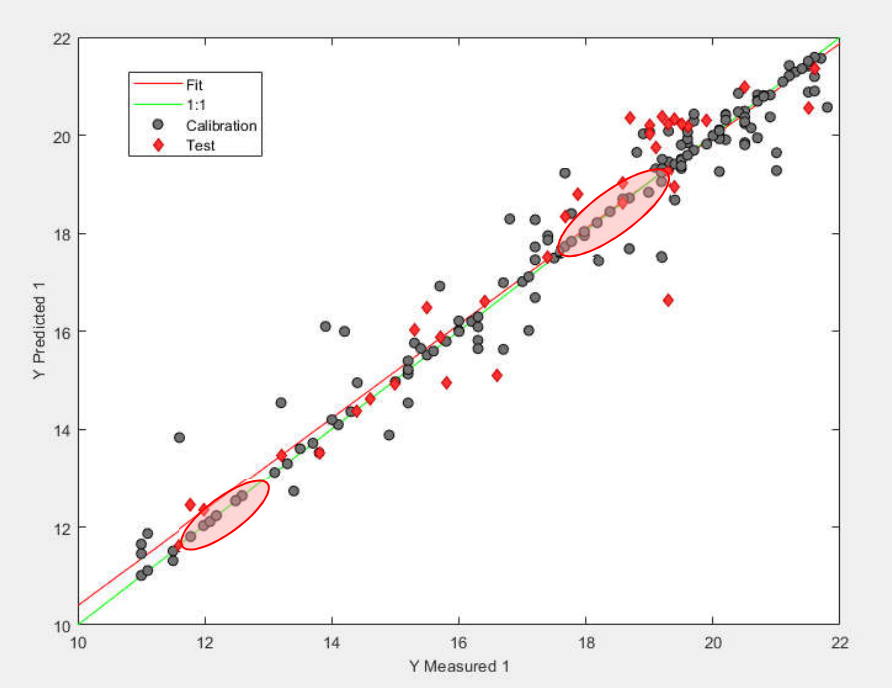
- Compression can be useful with spectroscopic data
  - Scores from PCA or PLS model used instead of spectra
  - Possibility to correct the scores with Mahalanobis distance (equivalent to scaling)
  - Careful not to select too many components to avoid overfitting



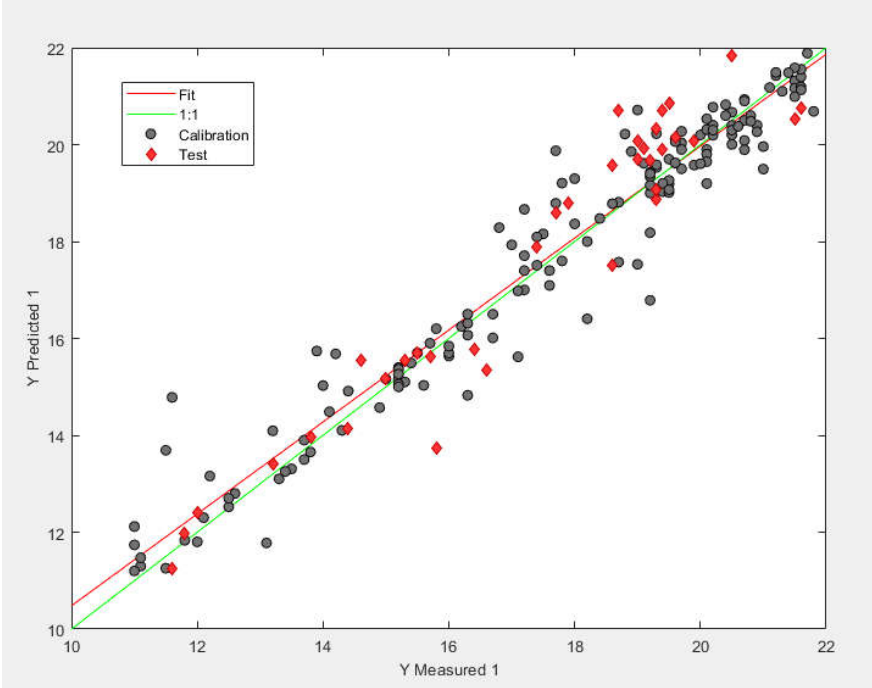
➢ In this particular example, results are globally better without compression. Compression should then not be automatically done, but investigated for each case

Compression

Too many components can lead to overfit, especially with a correction from the Mahalanobis distance (normalization)



Y = Protein – PCA compression – 10 principal components
Scores corrected from the Mahalanobis distance

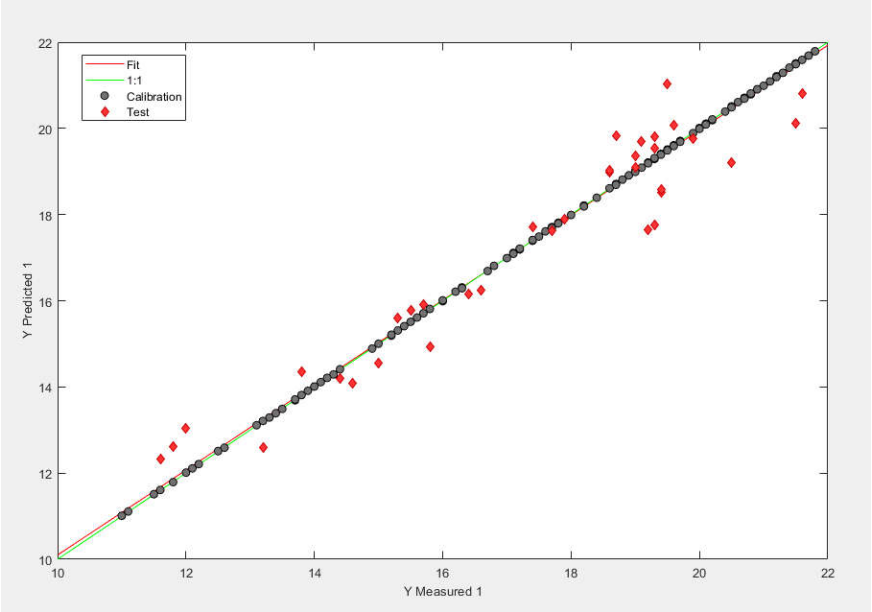


Y = Protein – PCA compression – 10 principal components
Non corrected scores

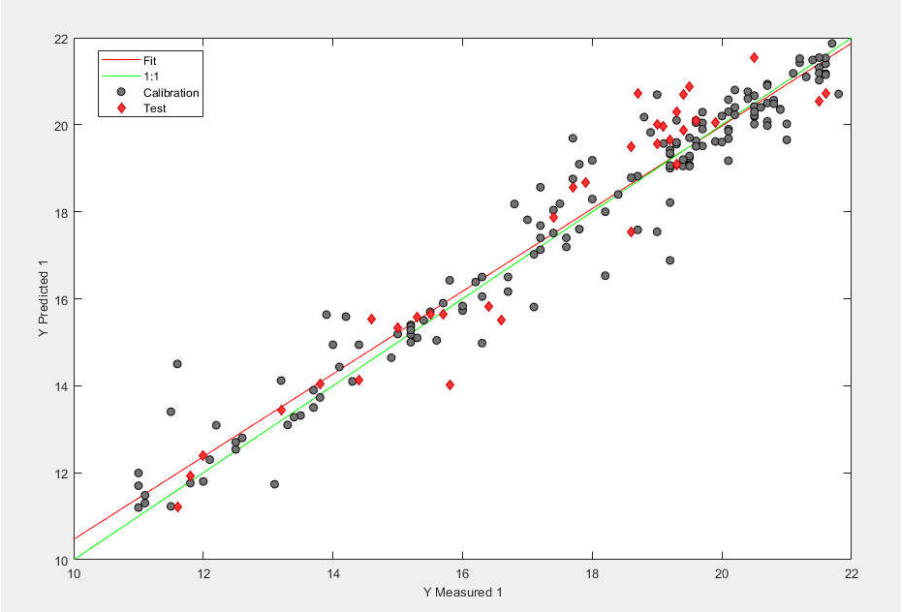
Software : PLS Toolbox (Eigenvector Research Inc.)

Compression

Example of an extreme case: PCA compression with 50 principal components



Y = Protein – PCA compression – 50 principal components
Scores corrected from the Mahalanobis distance



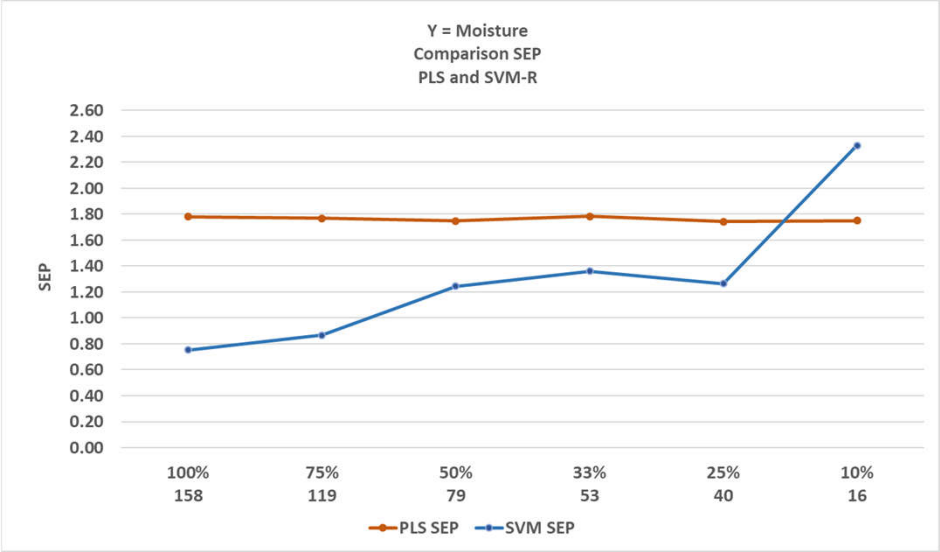
Y = Protein – PCA compression – 50 principal components
Non corrected scores

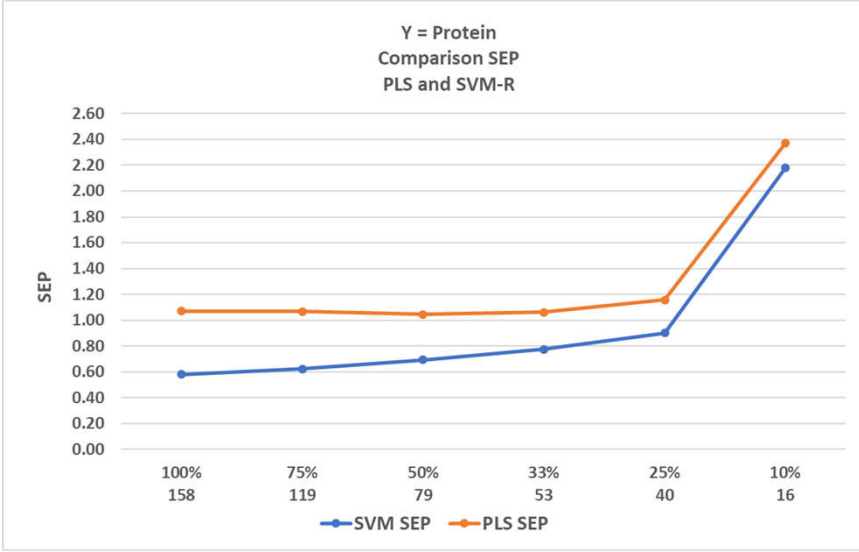
- Too many components combined with Mahalanobis correction might lead us to model noise

Software : PLS Toolbox (Eigenvector Research Inc.)

Training set size

- How does the SVM perform with fewer training samples?
- Crop of the calibration set – down to 10% of the initial set. Test set identical.





Examples on moisture – strong non-linearities – and on protein – weak non-linearity

➢ In this case, results acceptable with SVM up to 25% of the original calibration set (N = 40)

Comparison of algorithms

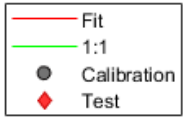
Parameter	Algorithm	Pretreatment	N LVs	N SVs	Bias test	R^2 test	RPD test	SEP	SEP (%)
fat	PLS	D1	3		-0.49	0.953	4.6	2.78	19.4%
fat	SVM-R	D1	-	40	-0.14	0.994	12.9	0.98	6.9%
fat	LS-SVM	D1	-	158	1.06	0.996	9.3	1.37	10%
moisture	PLS	SNV	3	-	0.14	0.966	5.6	1.78	2.7%
moisture	SVM-R	SNV	-	114	-0.06	0.993	12.3	0.81	1.2%
moisture	LS-SVM	SNV	-	158	-1.09	0.992	7.2	1.38	2.1%
protein	PLS	D1	3	-	0.24	0.853	2.8	1.07	5.7%
protein	SVM-R	D1	-	67	0.08	0.960	5.2	0.58	3.1%
protein	LS-SVM	D1	-	158	-0.30	0.965	4.9	0.62	3.3%

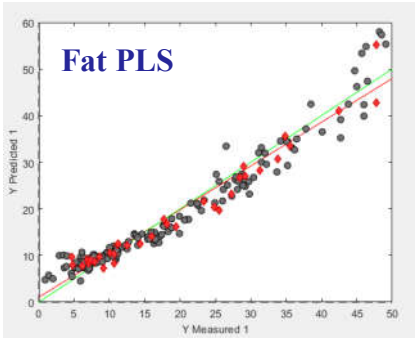
- SVM algorithms significantly better than PLS
- Even for protein content, which showed weak non-linearity

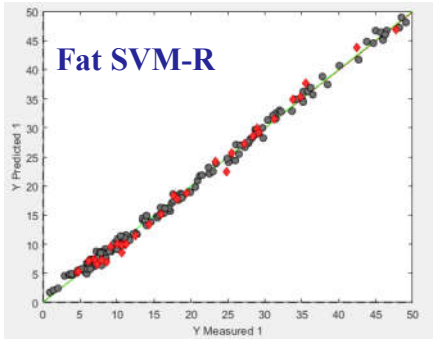
Software:

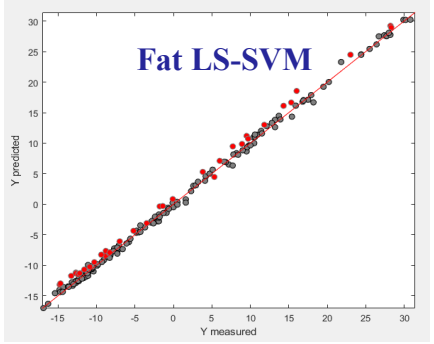
- SVM-R: PLS Toolbox (Eigenvector Research Inc.)
- LS-SVM: LS-SVMlab Toolbox, ESAT – K.U.Leuven

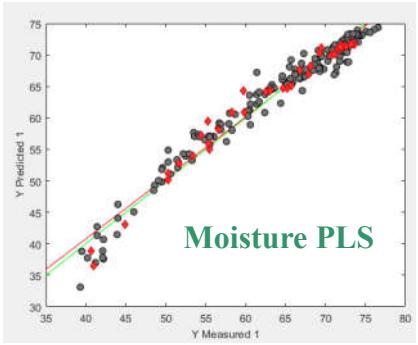
Comparison of algorithms

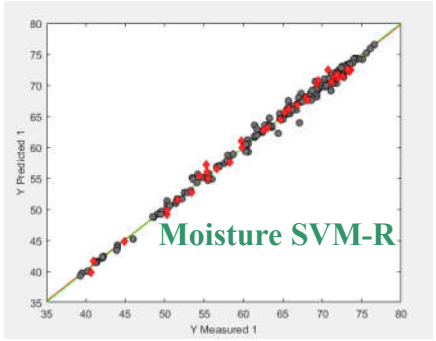


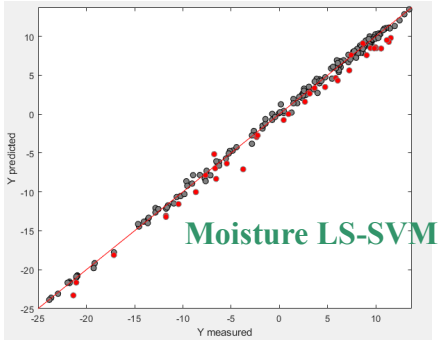


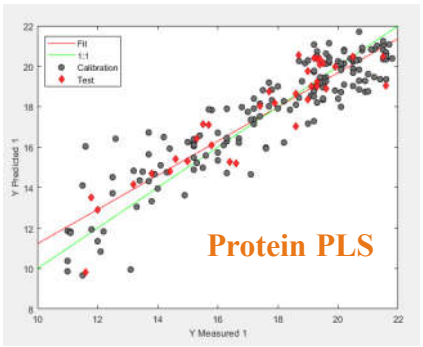


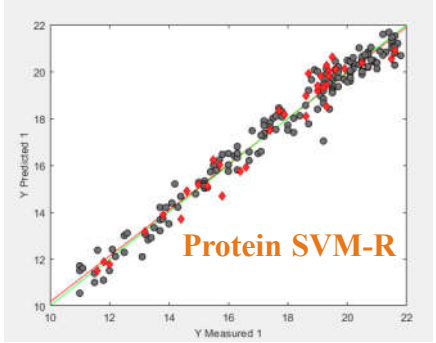


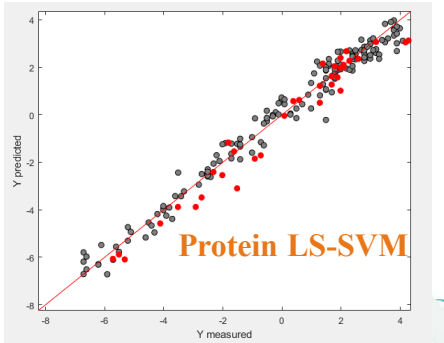












Conclusion

- SVM are very useful for non linear data
  - Much better results than PLS
  - Works also well for linear data!
- Efficient even with a low number of samples in the calibration set
  - In this particular study, satisfactory results obtained with 40 samples in the calibration set
  - Interesting alternative to ANN
- Careful however not to overfit!

www.ondalys.fr

Donnez du sens à vos données

Making sense of your data

Prestations et formation en
Analyse de données /
Chimiométrie /
Data Analytics

- Analyse exploratoire/ Data-mining
- Calibration spectroscopique
- Multivariate Process Control (MSPC)
- Multi-blocs / Fusion de données
- Plans d'Expériences

Merci pour votre attention!

www.ondalys.fr

ondalys
4, Rue Georges Besse
34830 CLAPIERS
Tél. 04 67 67 97 87
Fax 04 67 67 97 88
contact@ondalys.fr
www.ondalys.fr