Analisi componenti principali e classificazione mandorle

Francesco Malferrari 193103

Strumenti

Matlab:

- PCA e PLS/DA: PLS_Toolbox
- SIMCA: codici forniti dal prof. Federico Marini (La Sapienza) e prof.ssa Marina Cocchi (Unimore)

Dataset

Mandorle:

Training set (300 campioni):

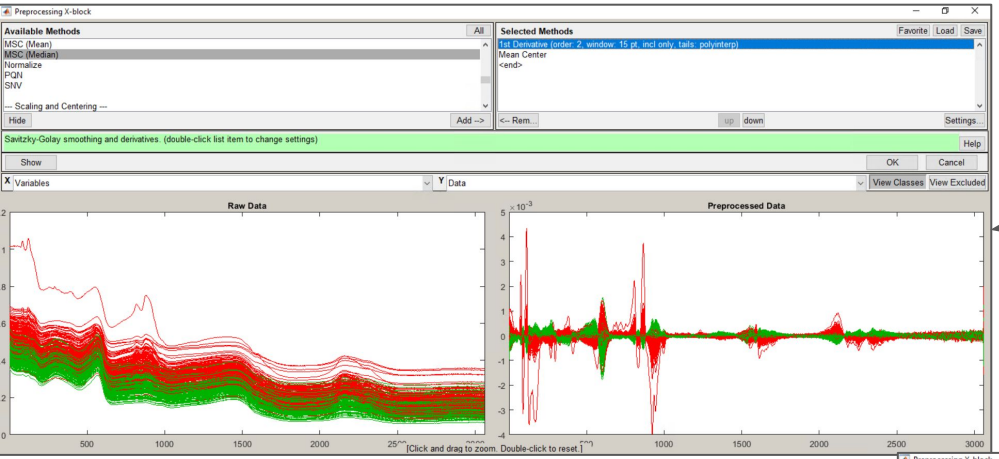
- 152 amare ‘Bitter’
- 148 dolci ‘Sweet’

Test set: 20 campioni

Variabili (3060):

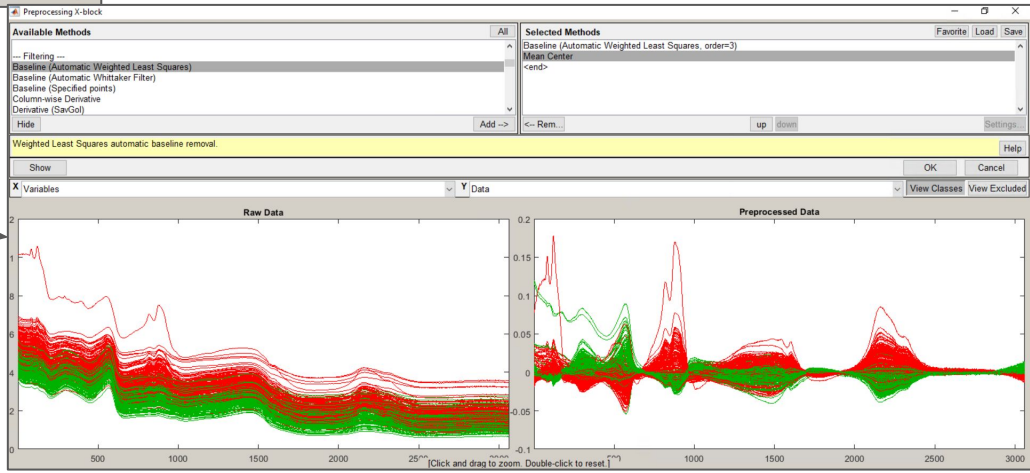
- Sono stati acquisiti spettri nel vicino infrarosso NIR in riflettanza.
- La regione spettrale di lunghezze d’onda è 1000-2500 nm.

Quale preprocessing scegliere

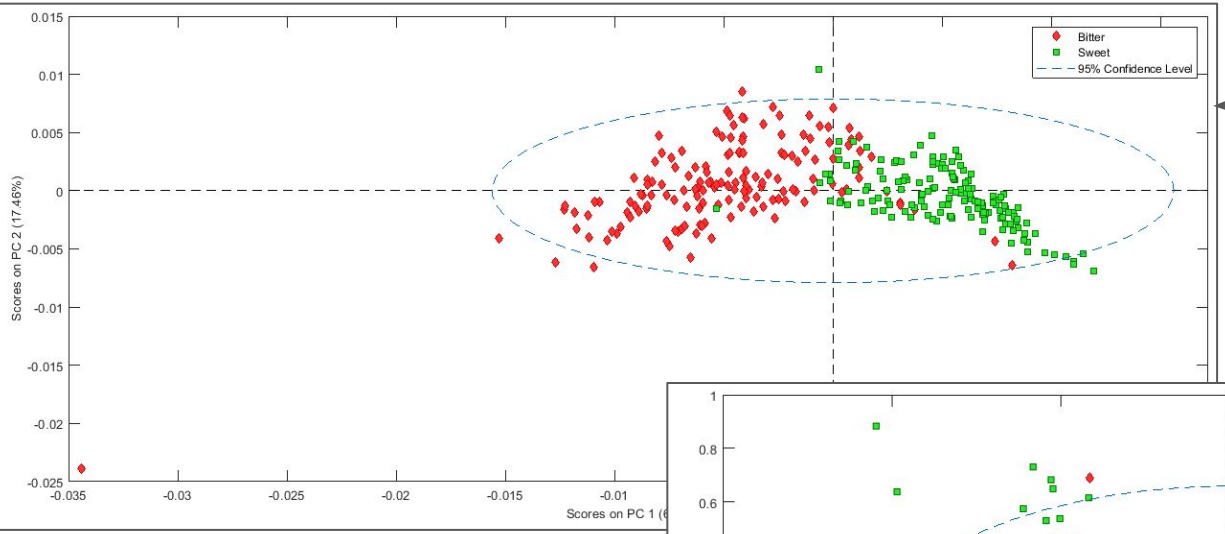


1st Derivative (order=2) + Mean Center

<u>Baseline (Weighted Least Squares, order=3) + Mean Center</u>

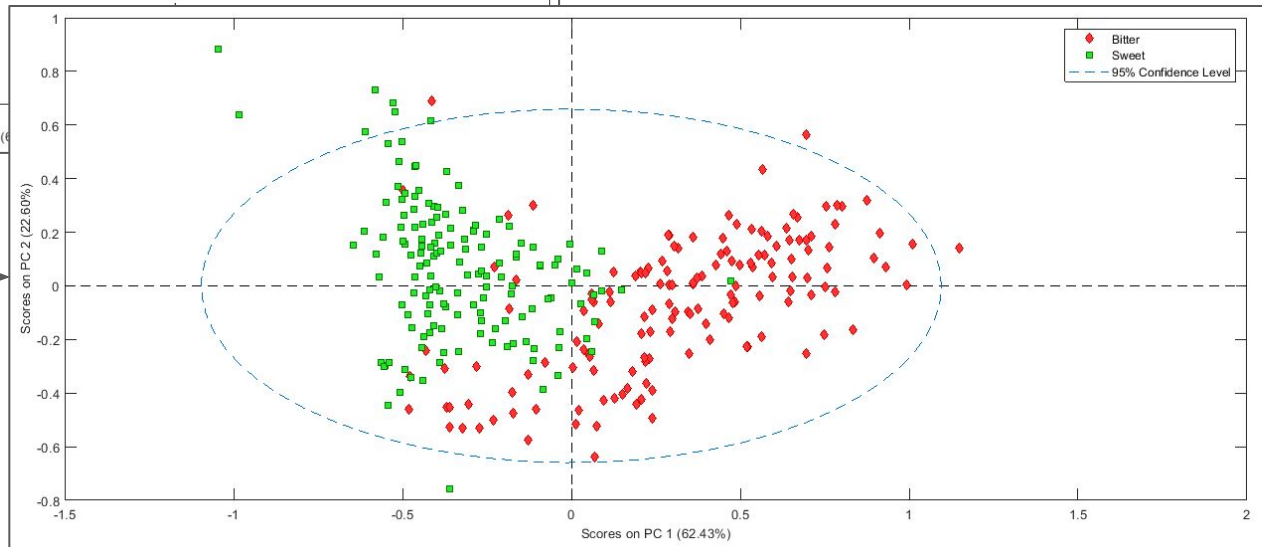


Quale preprocessing scegliere (2)



1st Derivative (order=2)
+ Mean Center

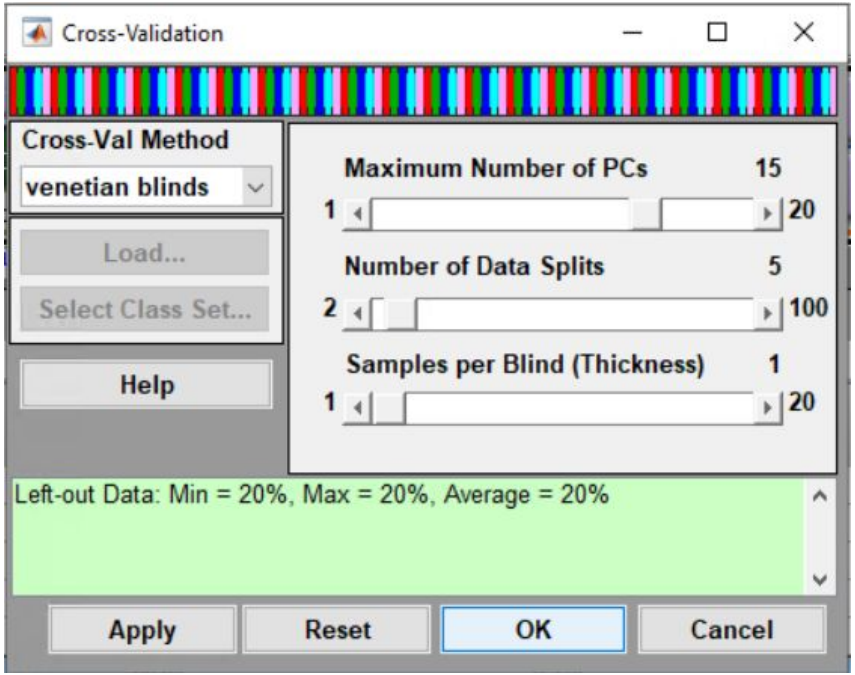
<u>Baseline (Weighted Least Squares, order=3) + Mean Center</u>



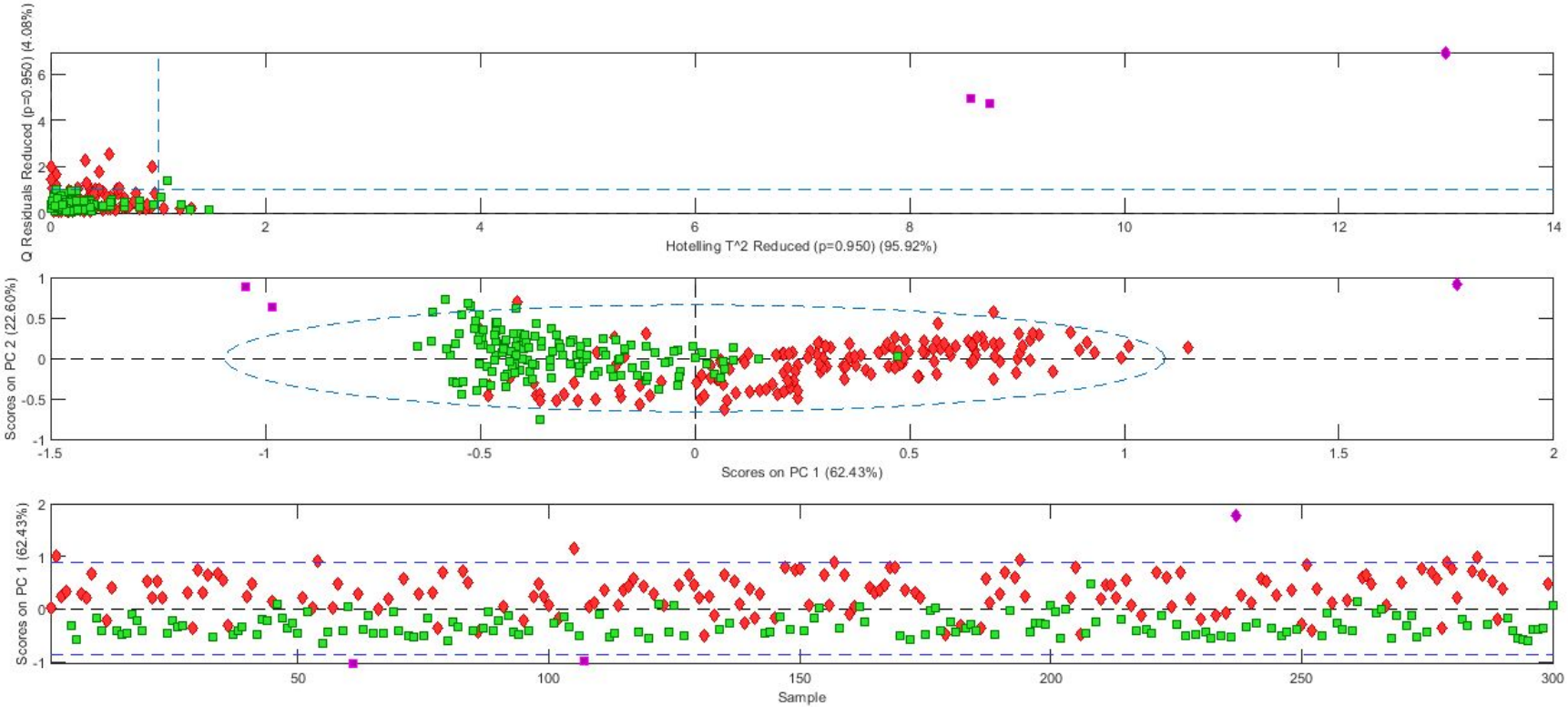
Cross-validazione

“Venetian blinds”: ⟶ Ordinamento

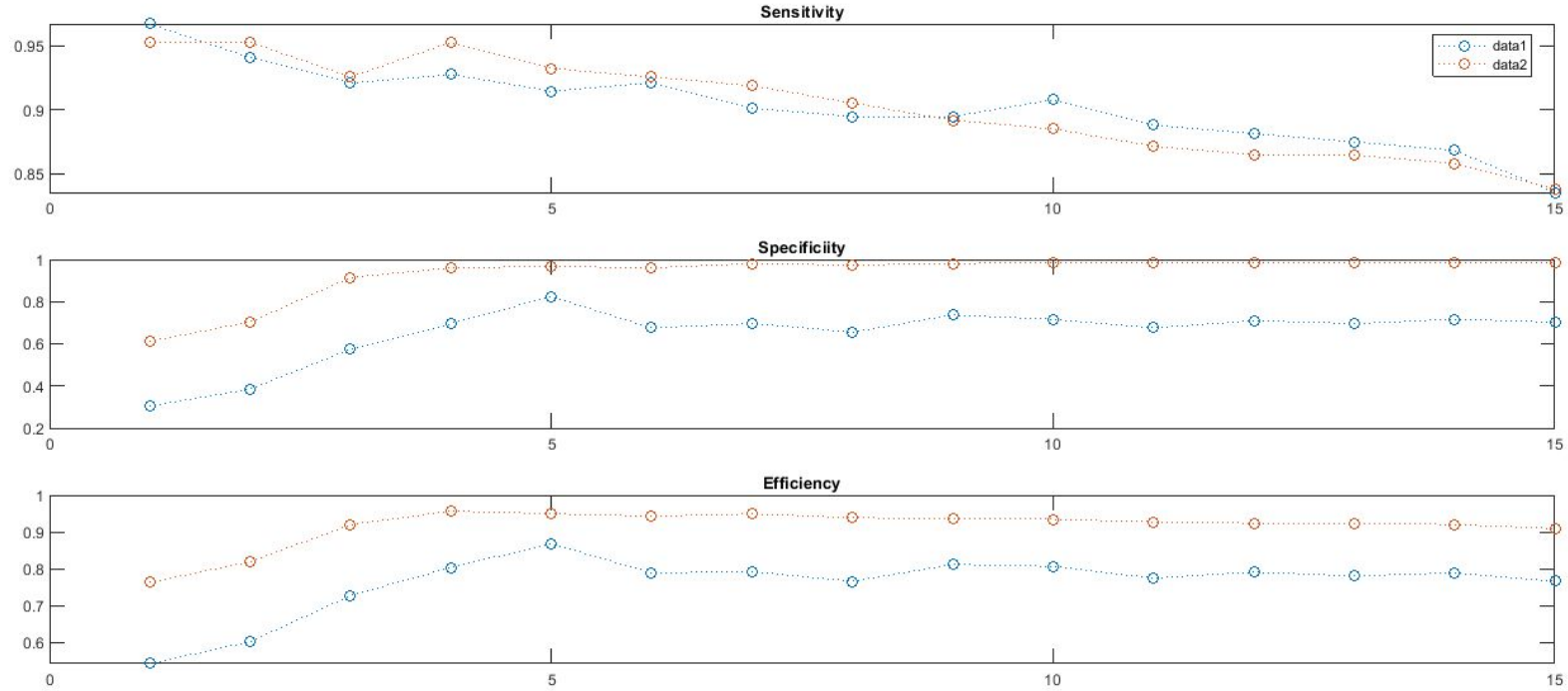
- Splits: 5 (20%)
  - 240 campioni alla volta
- PC: 15 per stare larghi (12 per PLS/DA)



Outliers


Q Residuals Reduced (p=0.950) (4.08%)
Hotelling T^2 Reduced (p=0.950) (95.92%)
Scores on PC 2 (22.60%)
Scores on PC 1 (62.43%)
Scores on PC 1 (62.43%)
Sample

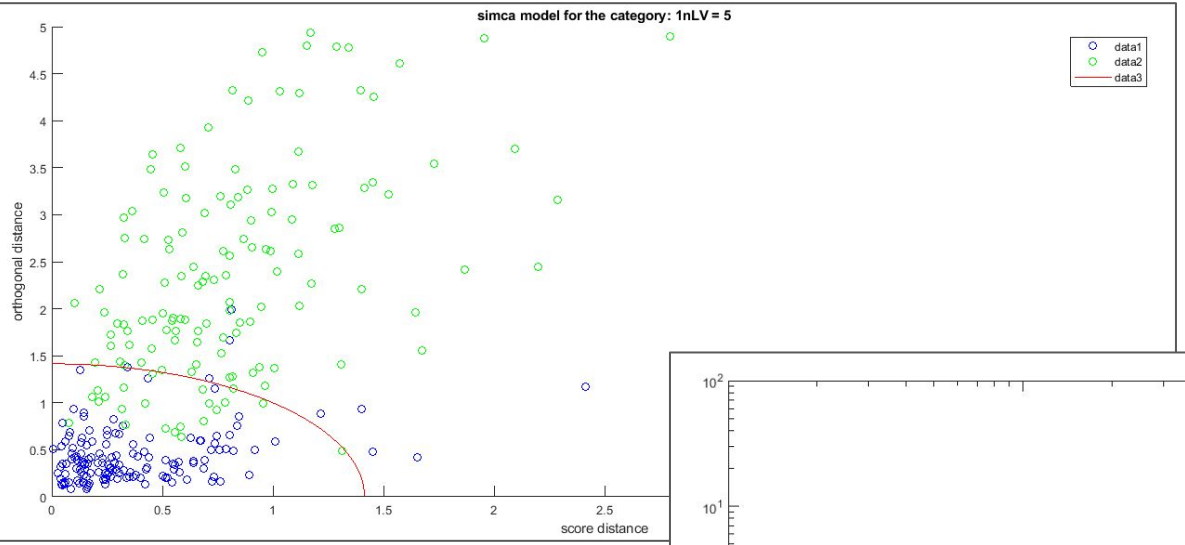
SIMCA: scelta componenti principali



Componenti:

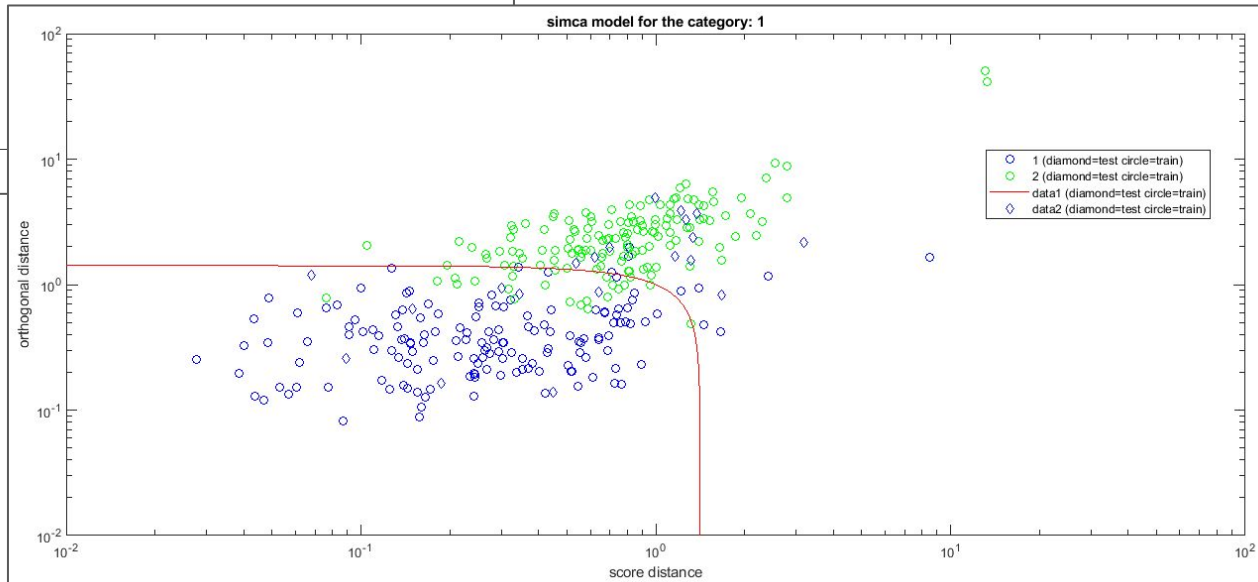
- “Bitter” (data1): 5
- “Sweet” (data2): 4

SIMCA: SD vs OD “Bitter”

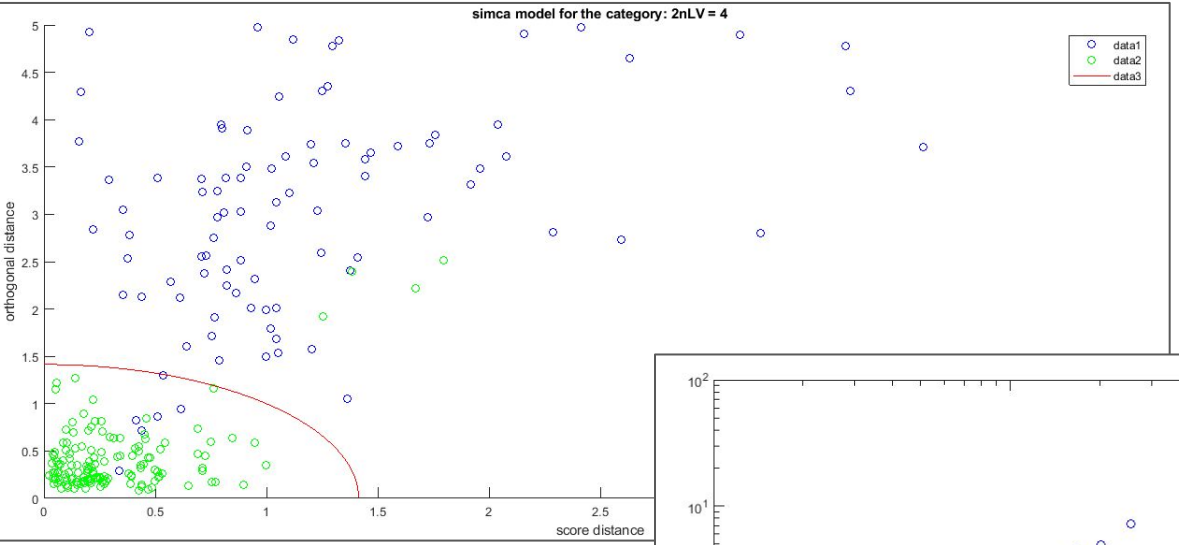


Sensitivity: 0.9342

Specificity of C1 vs C2: 0.8514

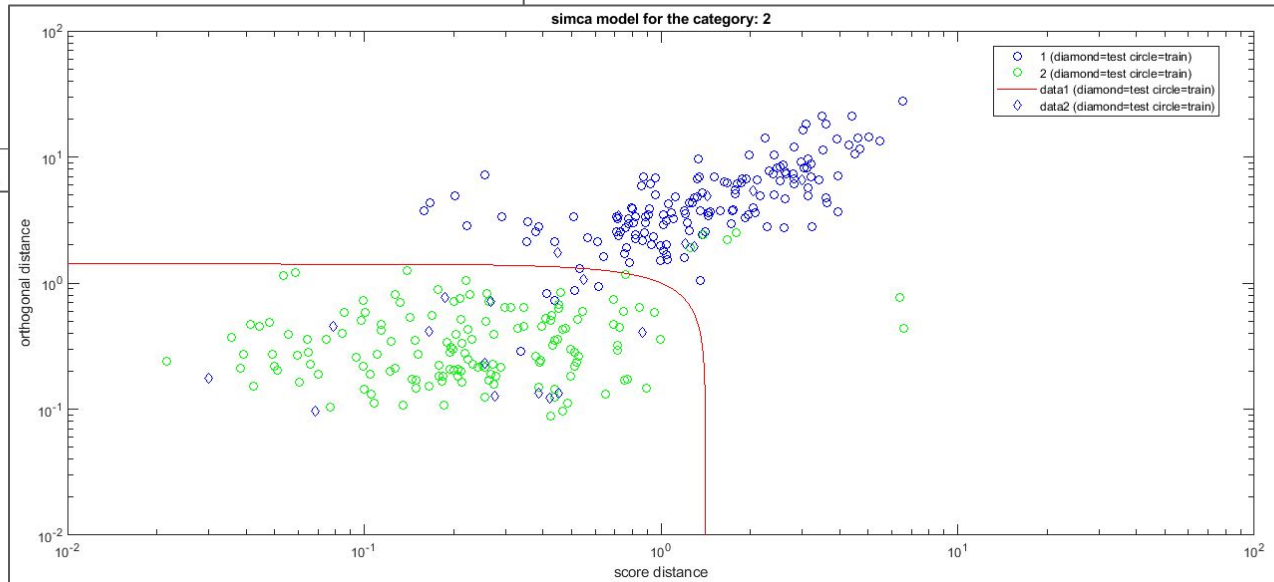


SIMCA: SD vs OD “Sweet”

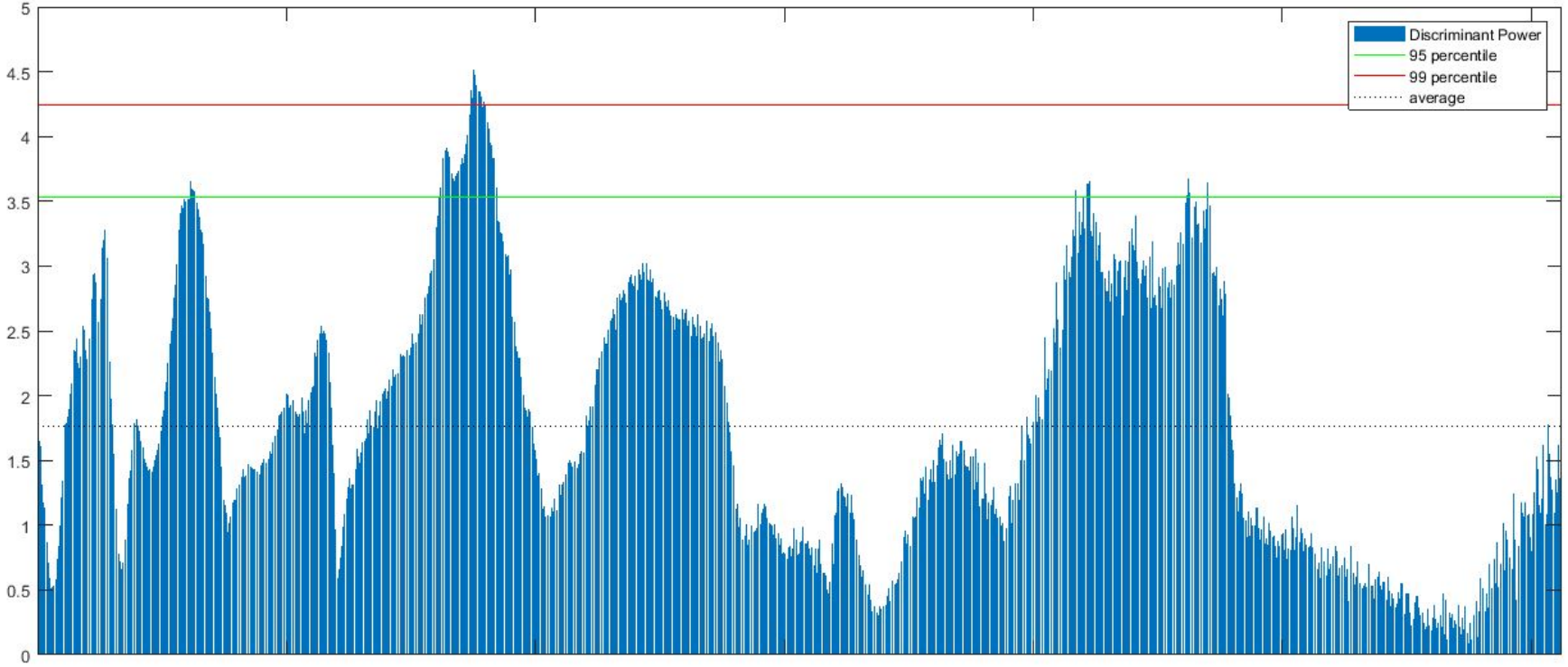


Sensitivity: 0.9595

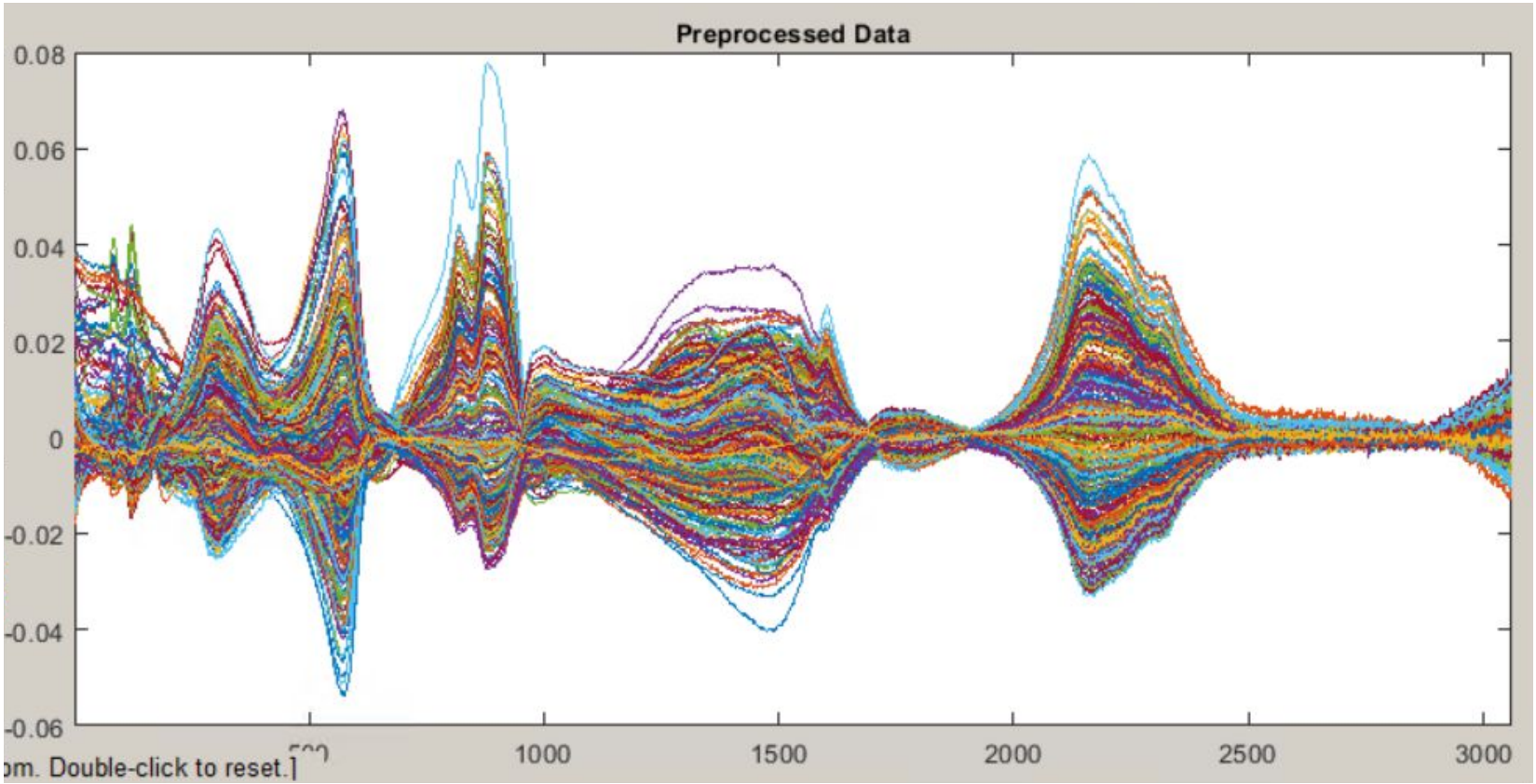
Specificity of C2 vs C1: 0.9605



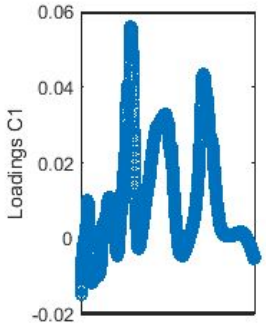
SIMCA: Discriminant power

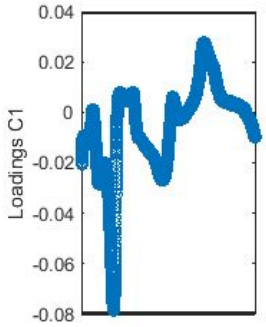


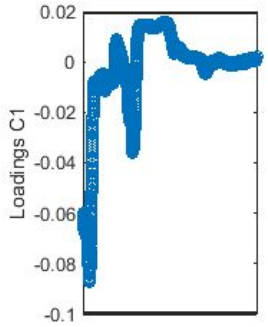
Somiglianza con lo spettro

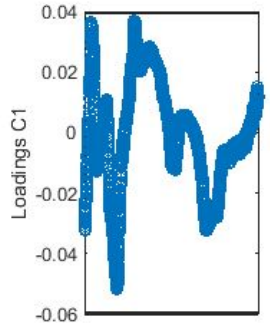

Preprocessed Data
0.08
0.06
0.04
0.02
0
-0.02
-0.04
-0.06
1000
1500
2000
2500
3000
om. Double-click to reset.]

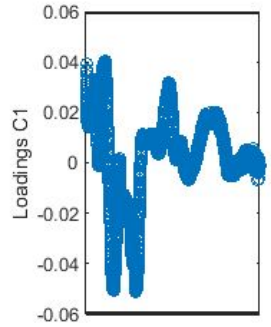
SIMCA: Loadings

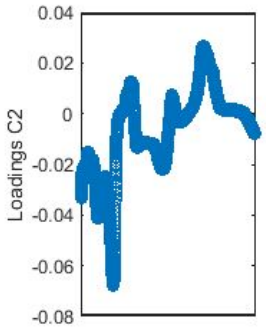


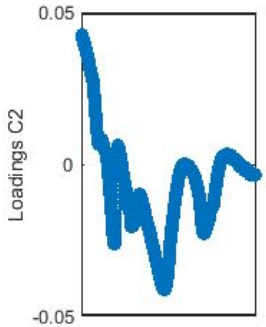


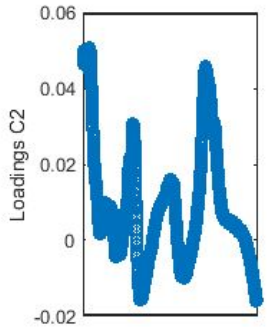


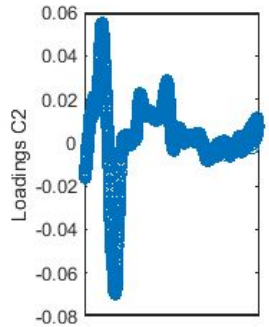




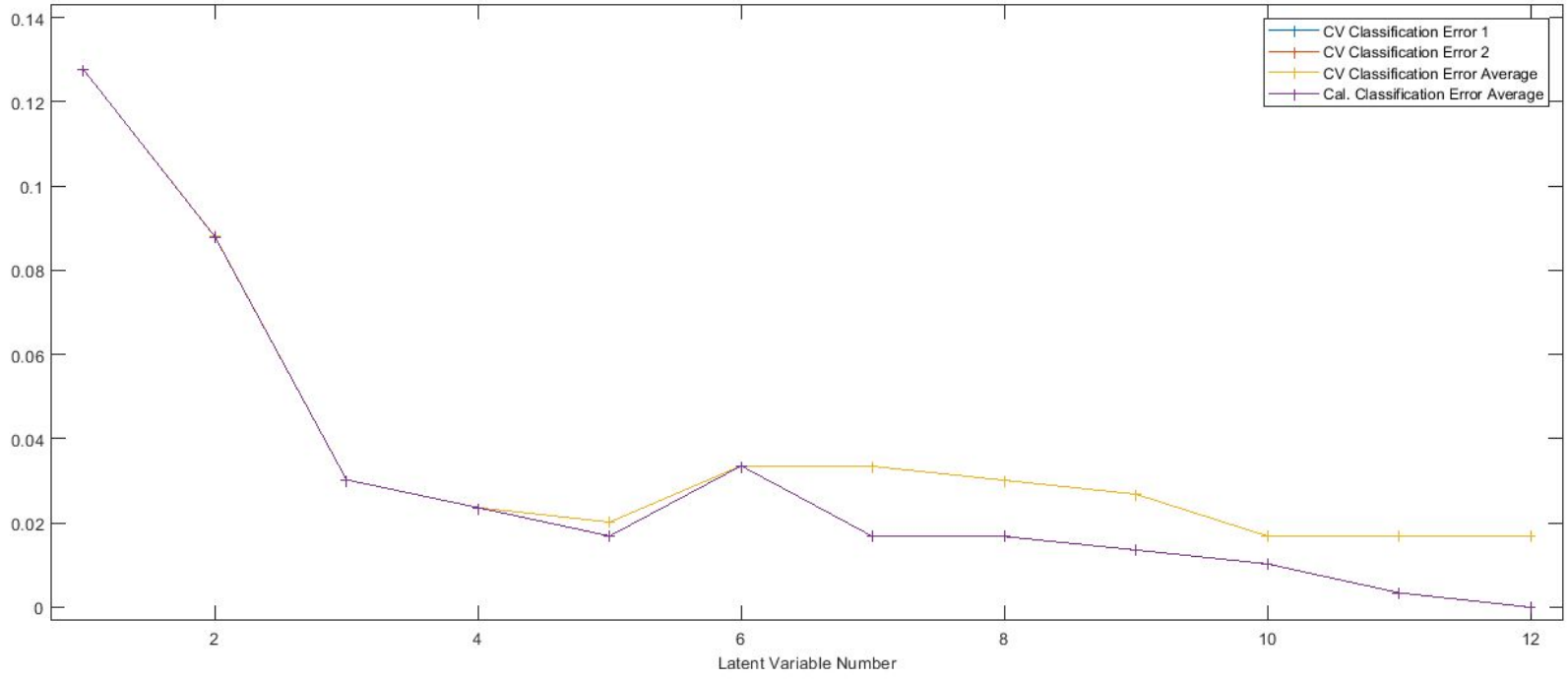






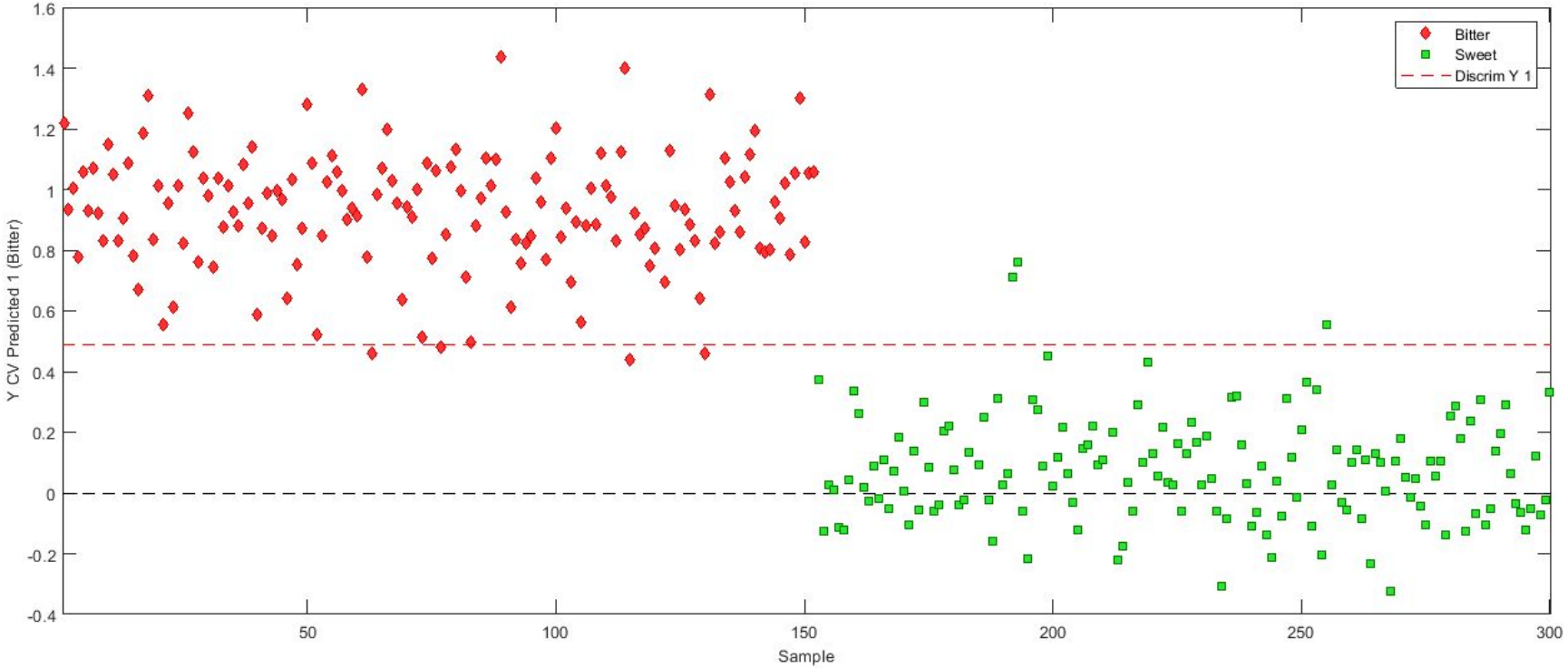


PLS/DA: scelta componenti principali

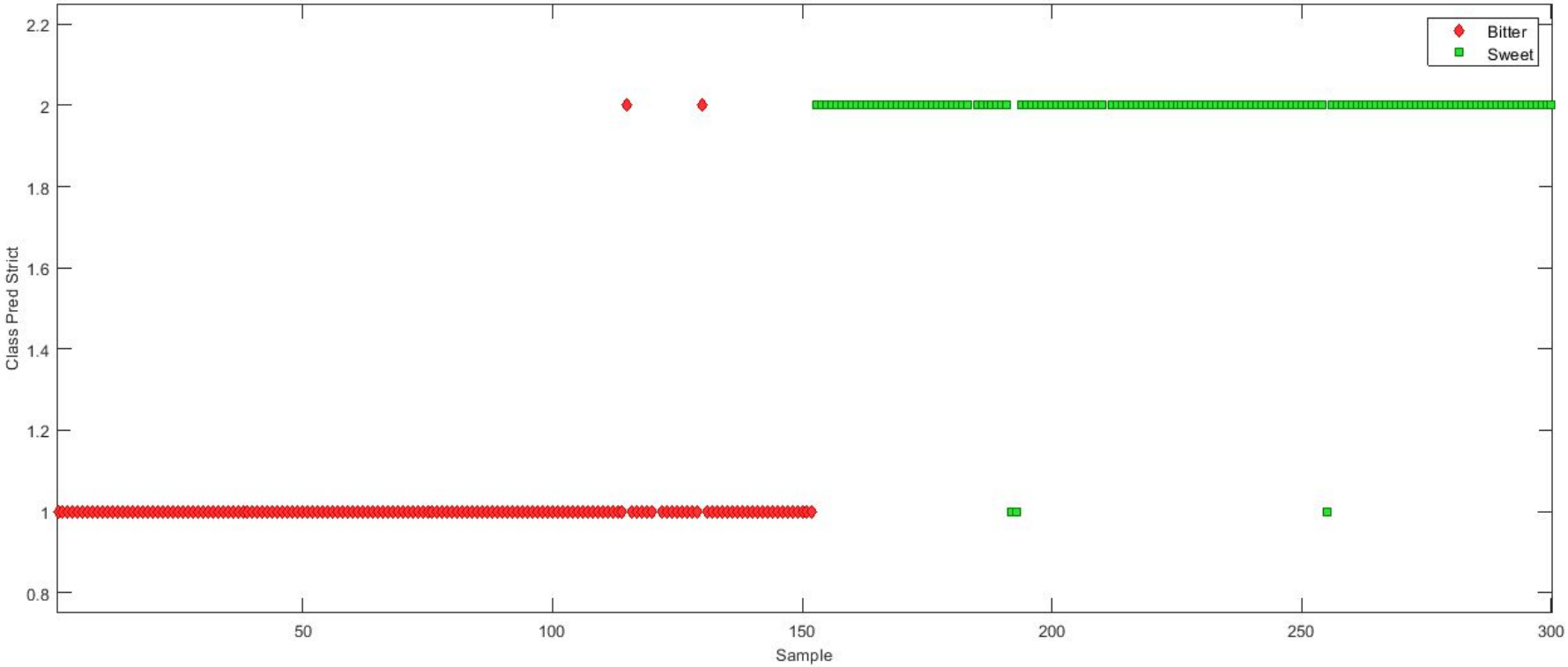


5 componenti

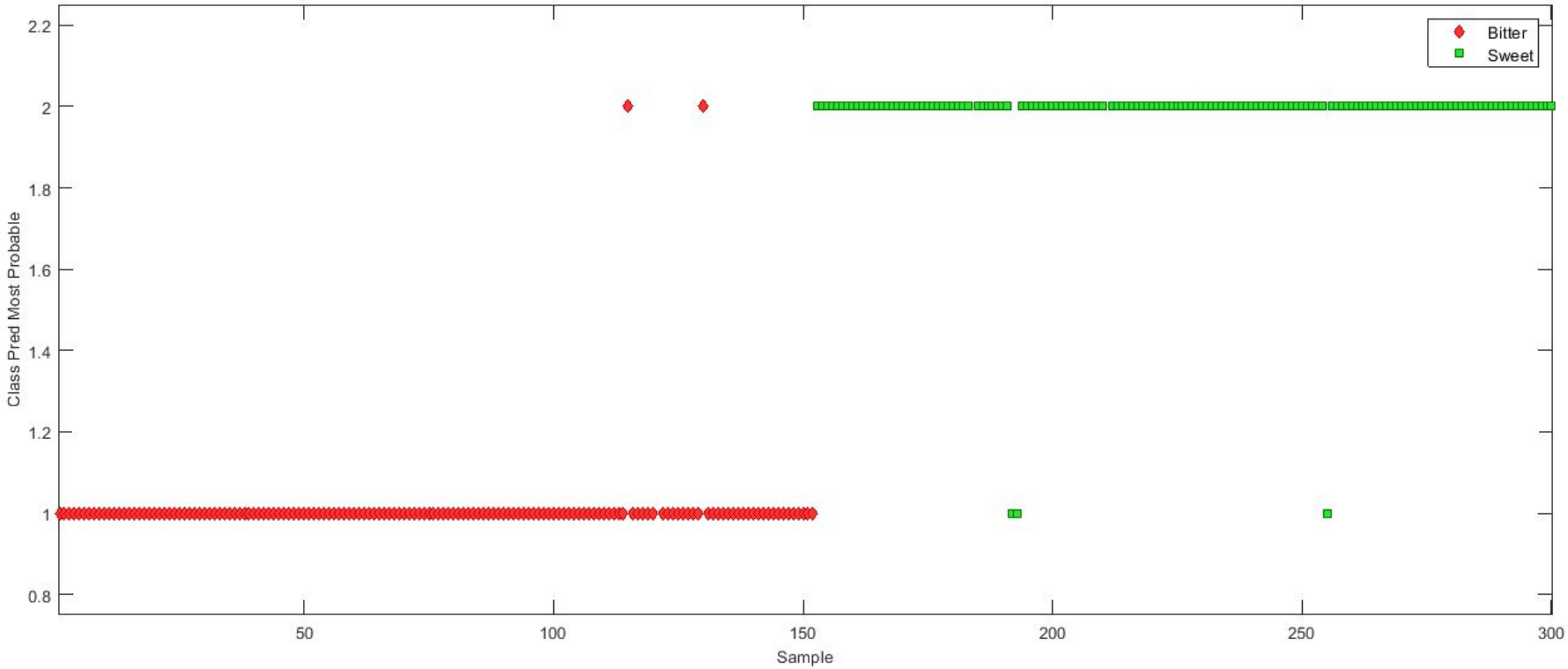
PLS/DA: Y CV predicted



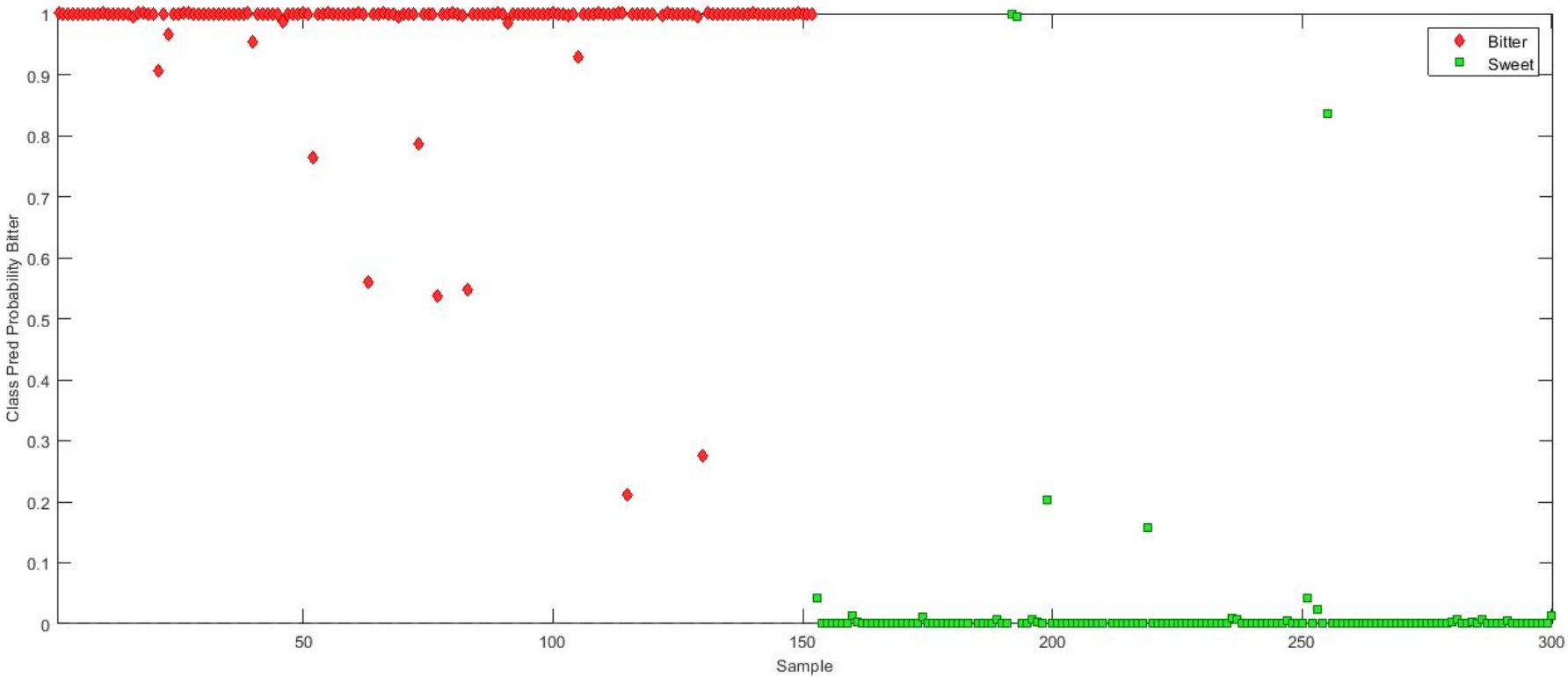
PLS/DA: Class pred. Strict



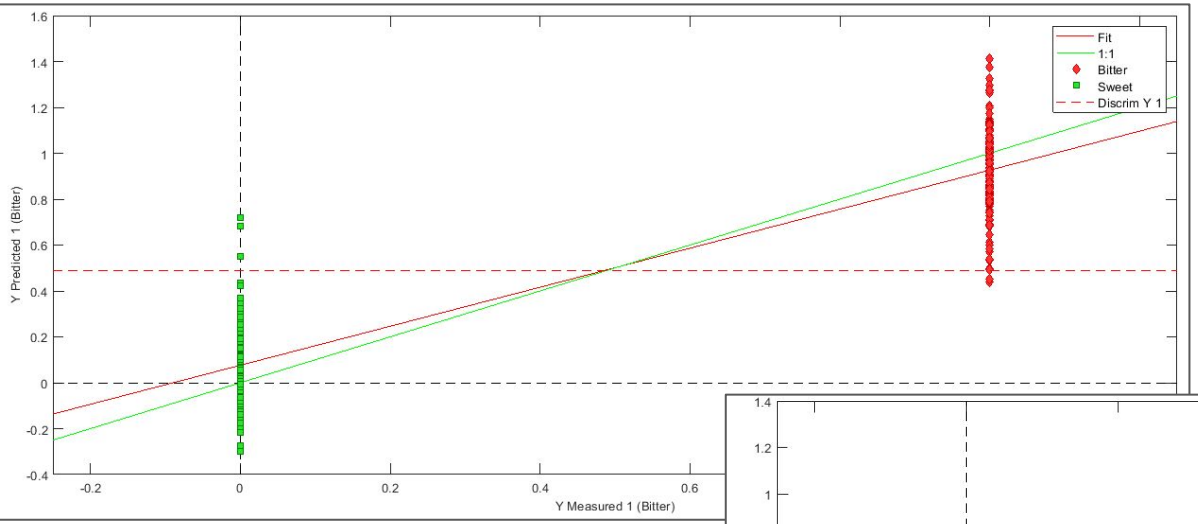
PLS/DA: Class pred. Most Probable

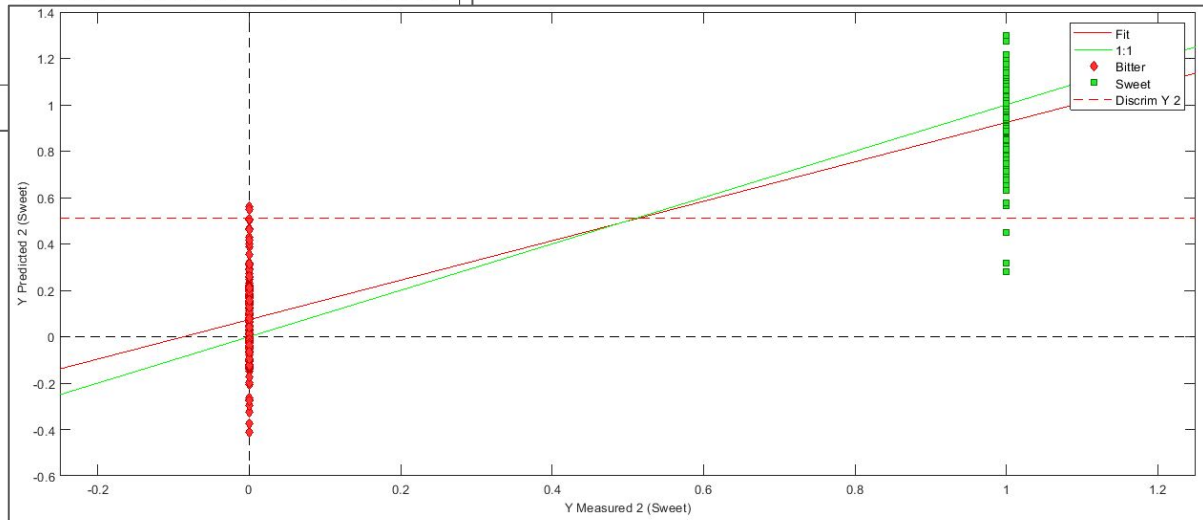


PLS/DA: Class pred. Probability

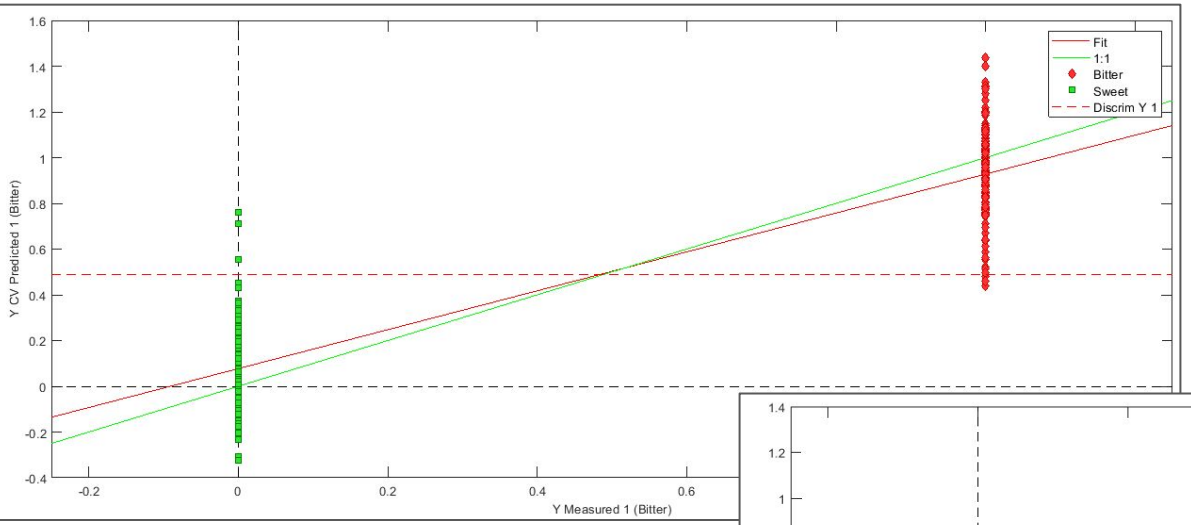


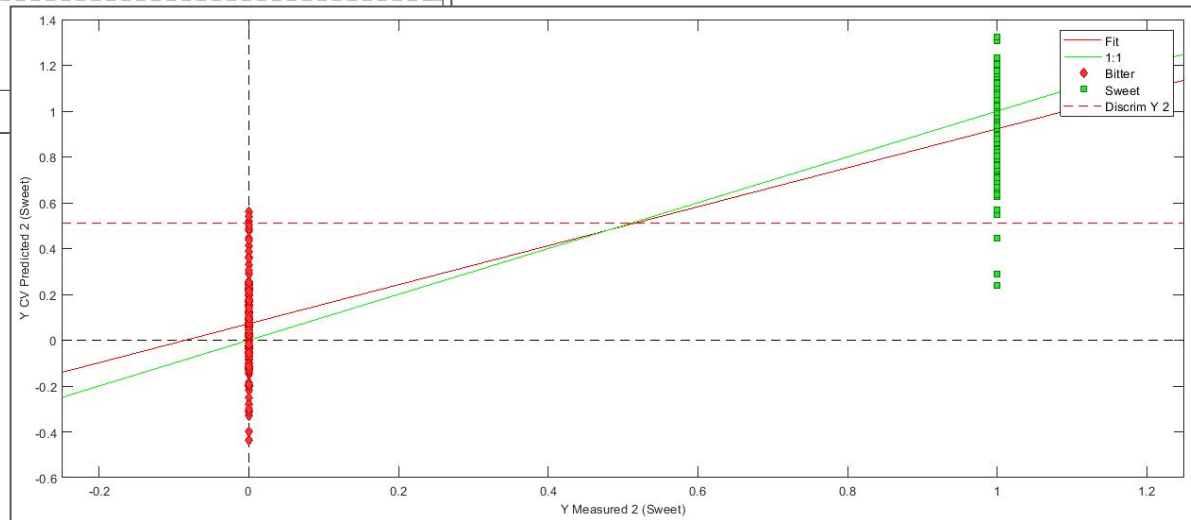
PLS/DA: Y measured/predicted



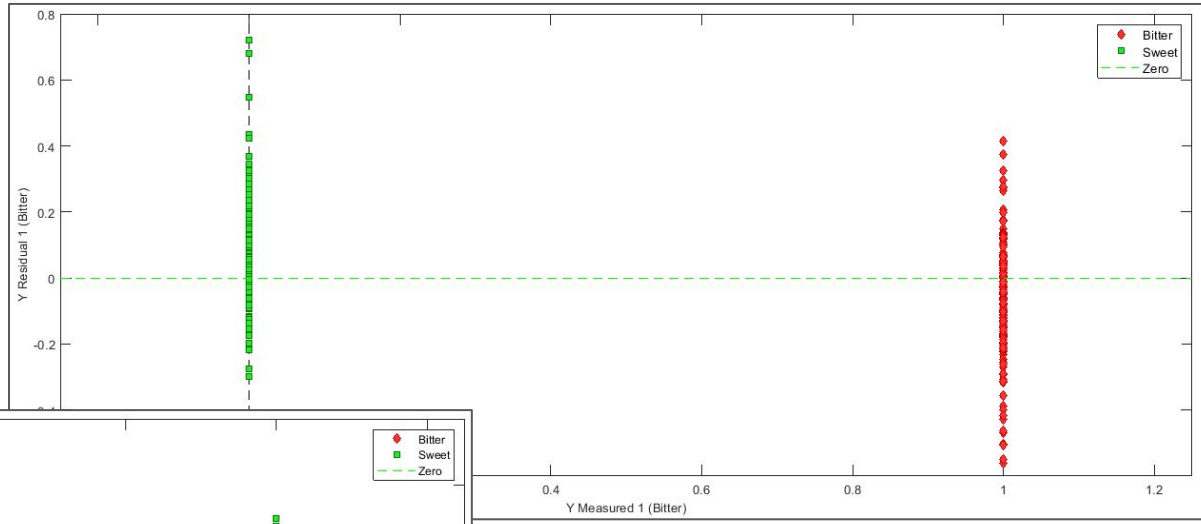


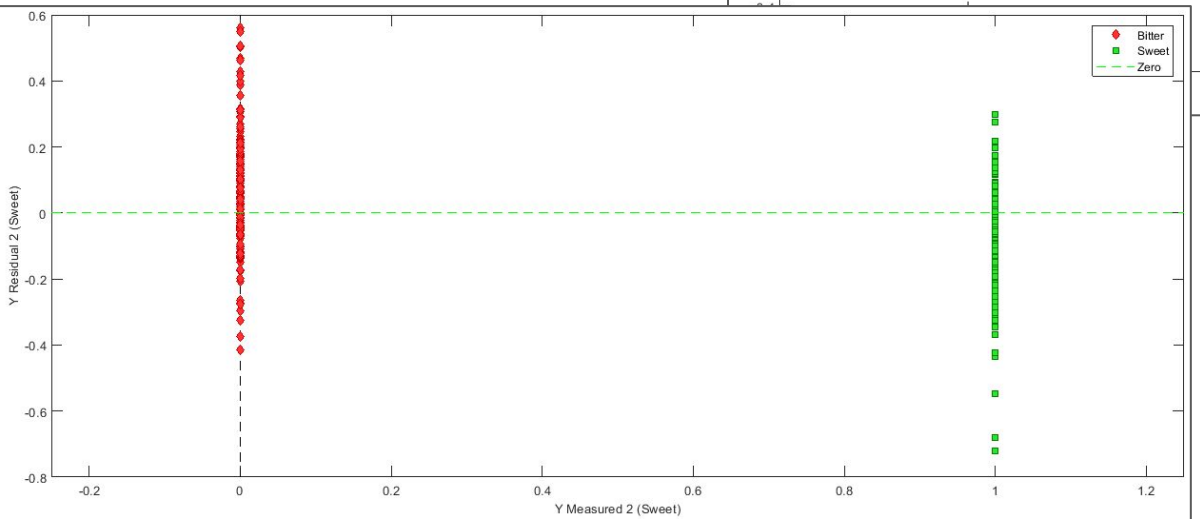
PLS/DA: Y measured/CV predicted



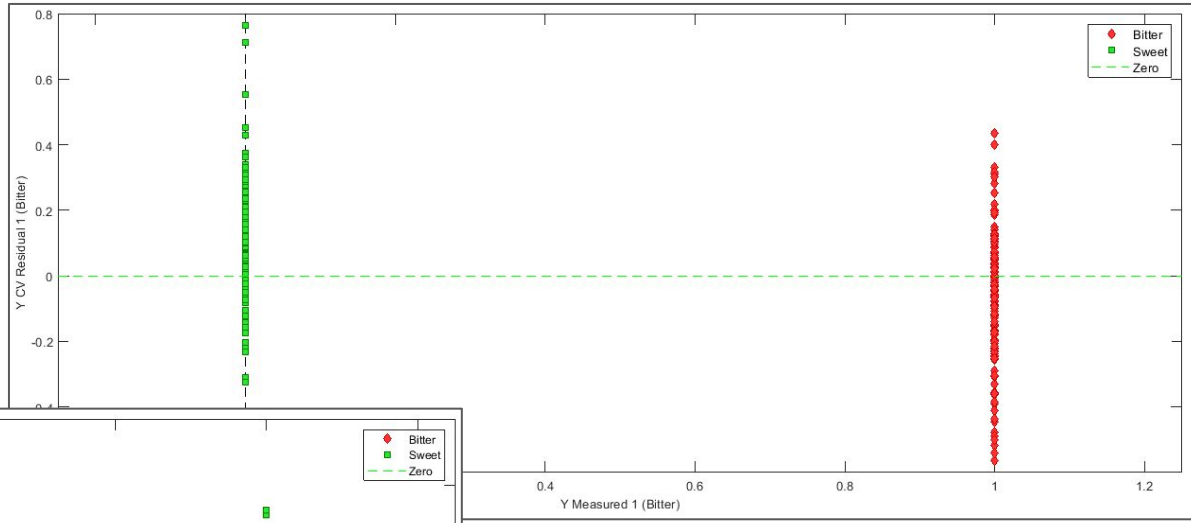


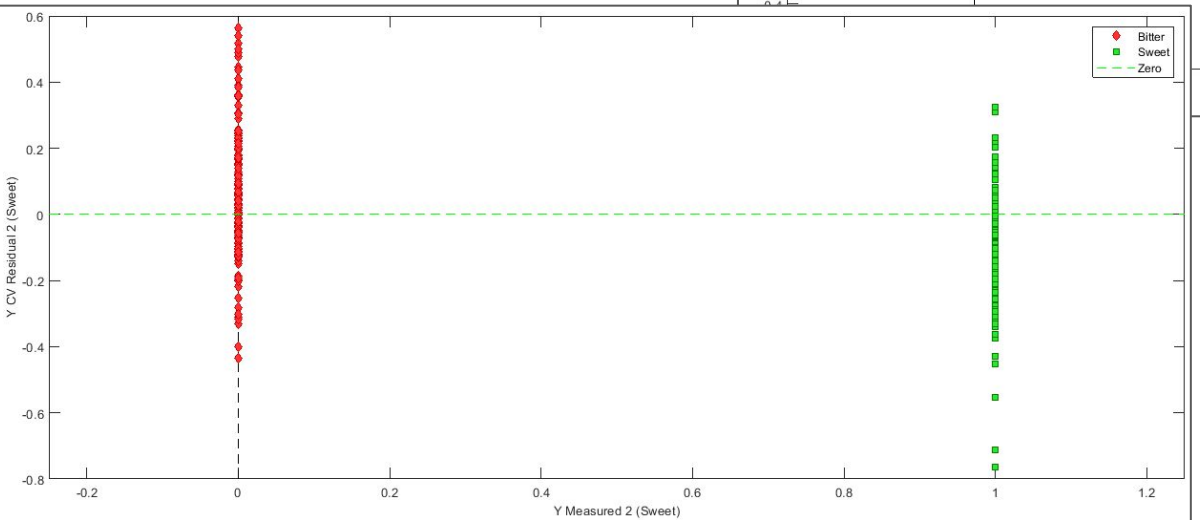
PLS/DA: Y measured/residual



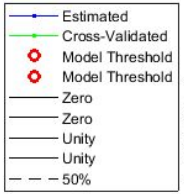


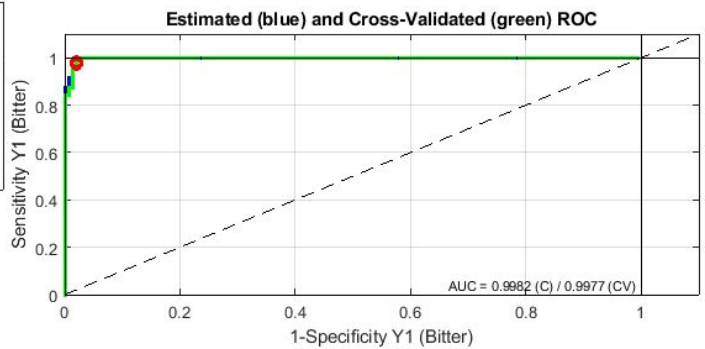
PLS/DA: Y measured/CV residual

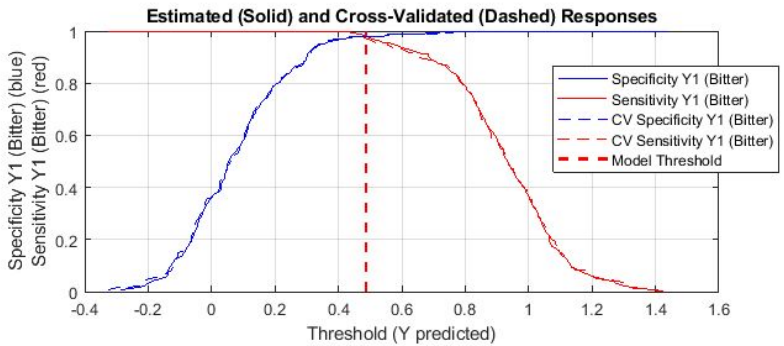


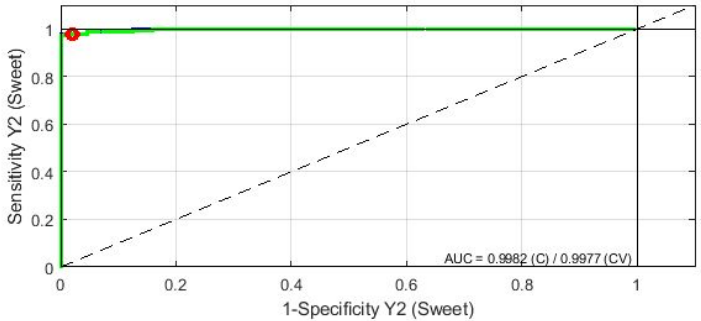


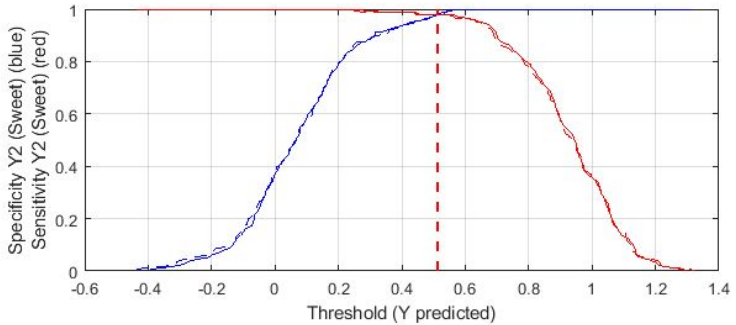
PLS/DA: Threshold



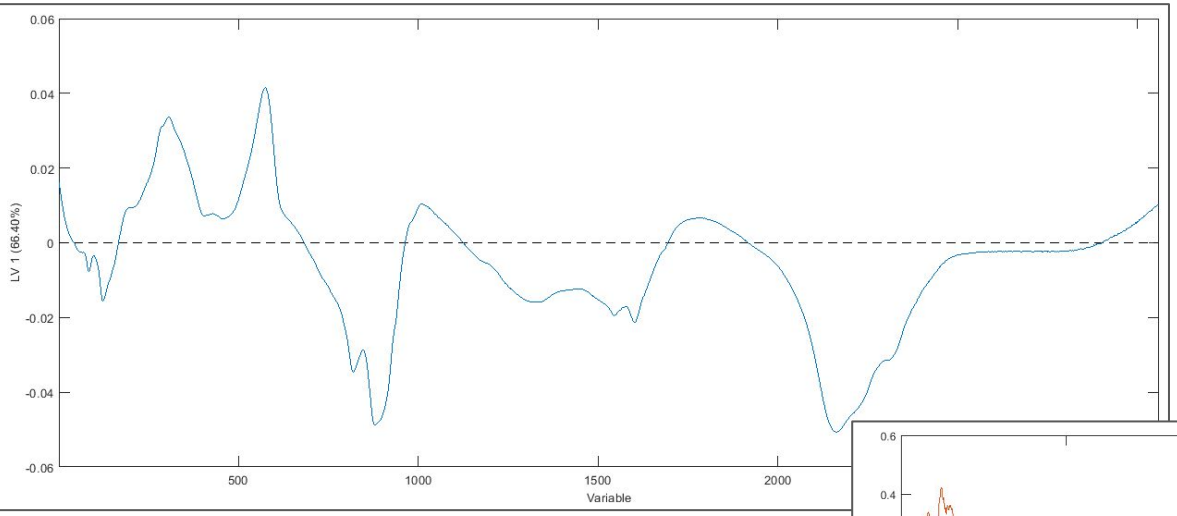




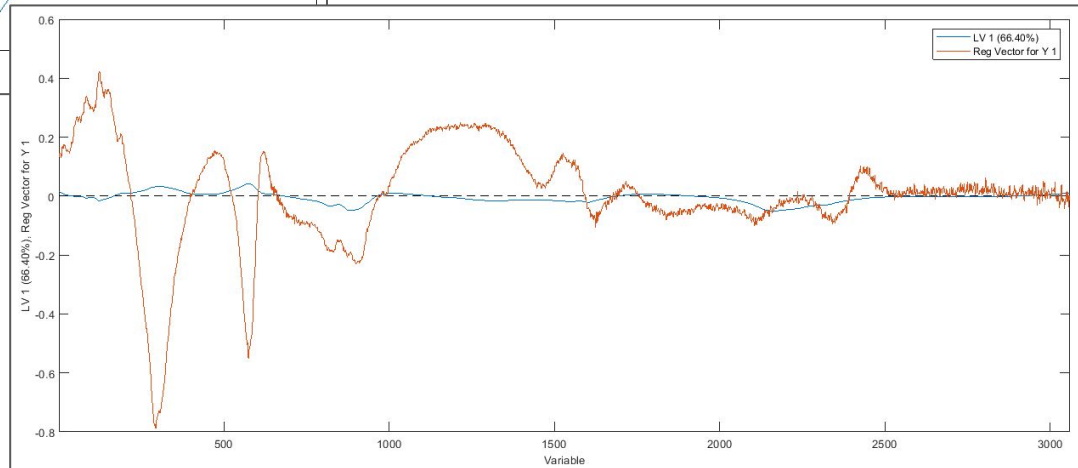




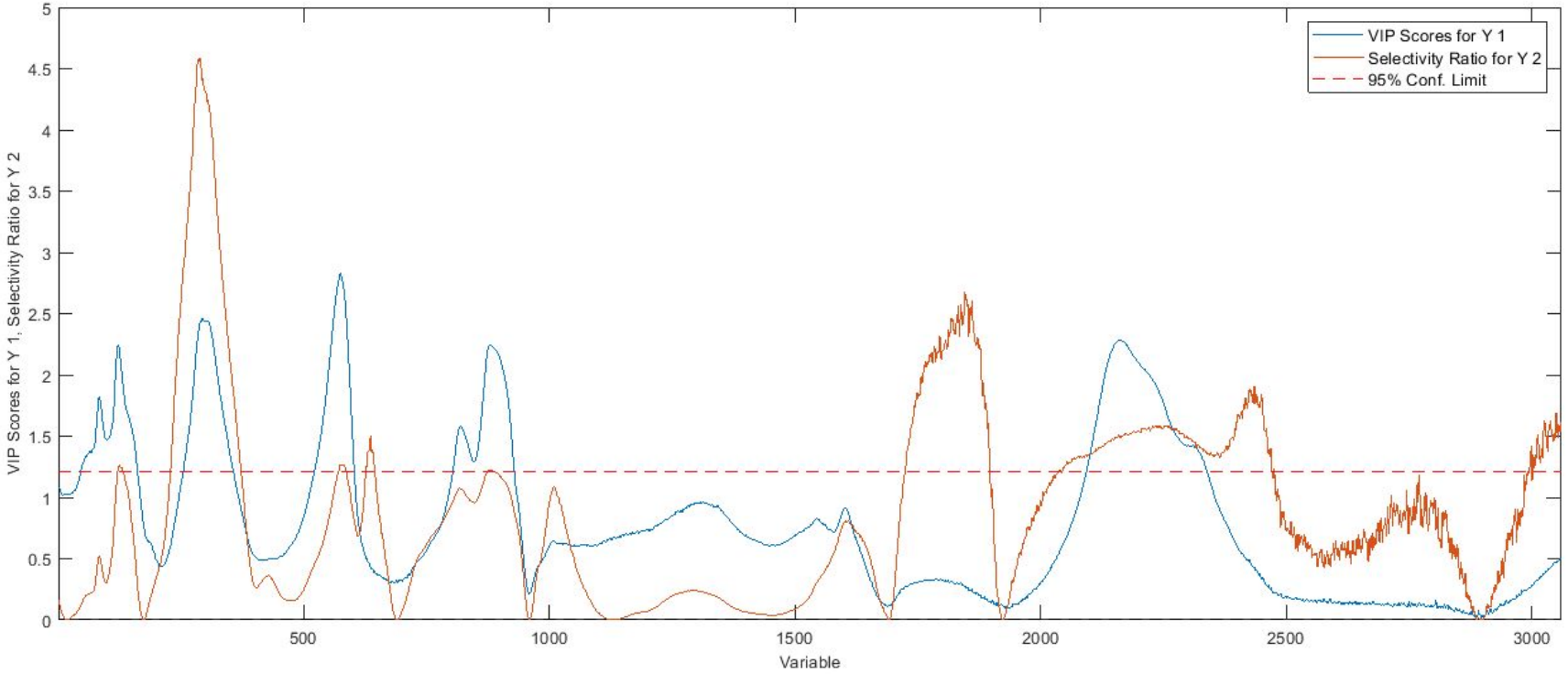
PLS/DA: Loadings + Reg Vector



Rumore da rimuovere?



PLS/DA: VIP + Selectivity Ratio



Rumore poco influente: non si è rimosso

PLS/DA: Matrice di confusione (Strict)

Fit:

Confusion Table:

	Actual Class	
	Bitter	Sweet
Predicted as Bitter	149	3
Predicted as Sweet	2	143
Predicted as Unassigned	0	0

Cross-validazione:

Strict threshold = 0.50

Confusion Table (CV):

	Actual Class	
	Bitter	Sweet
Predicted as Bitter	147	3
Predicted as Sweet	4	143
Predicted as Unassigned	0	0

PLS/DA: Matrice di confusione (Most Probable)

Fit:

Confusion Table:

	Actual Class	
	Bitter	Sweet
Predicted as Bitter	149	3
Predicted as Sweet	2	143
Predicted as Unassigned	0	0

Cross-validazione:

Confusion Table (CV):

	Actual Class	
	Bitter	Sweet
Predicted as Bitter	147	3
Predicted as Sweet	4	143
Predicted as Unassigned	0	0

PLS/DA: Predictions

```
Statistics for each y-block column:
Sensitivity (Cal):  0.987  0.979
Specificity (Cal):  0.979  0.987
Sensitivity (CV):  0.980  0.979
Specificity (CV):  0.979  0.980
Class. Err (Cal): 0.0168965 0.0168965
Class. Err (CV): 0.0202077 0.0202077
RMSEC: 0.193465 0.193465
RMSECV: 0.19982 0.19982
CV Bias: 0.00137838 -0.00137838
R^2 Cal: 0.850242 0.850242
R^2 CV: 0.840379 0.840379
```

Risultati:

- Sensibilità e specificità sia in fit che in CV
- RMSEC e RMSECV bassi
- CV bias basso
- R^2 alto

Grazie per

l’attenzione