Prediction chicchi di grano

Relazione di Francesco Malferrari

Il set di dati preso in analisi è proveniente da 415 campioni di chicchi di grano, rappresentanti 43 differenti varietà, o miscele di varietà, provenienti da due diverse località in Danimarca, e da altri 108 campioni di chicchi di grano, rappresentanti 11 diverse varietà provenienti da una singola località.

Lo scopo è creare un modello di regressione predittivo basato su PLS per poter trovare la direzione multidimensionale nello spazio X che spiega la massima direzione di varianza multidimensionale nello spazio Y. Per realizzare ciò è stato usato Matlab con il PLS toolbox.

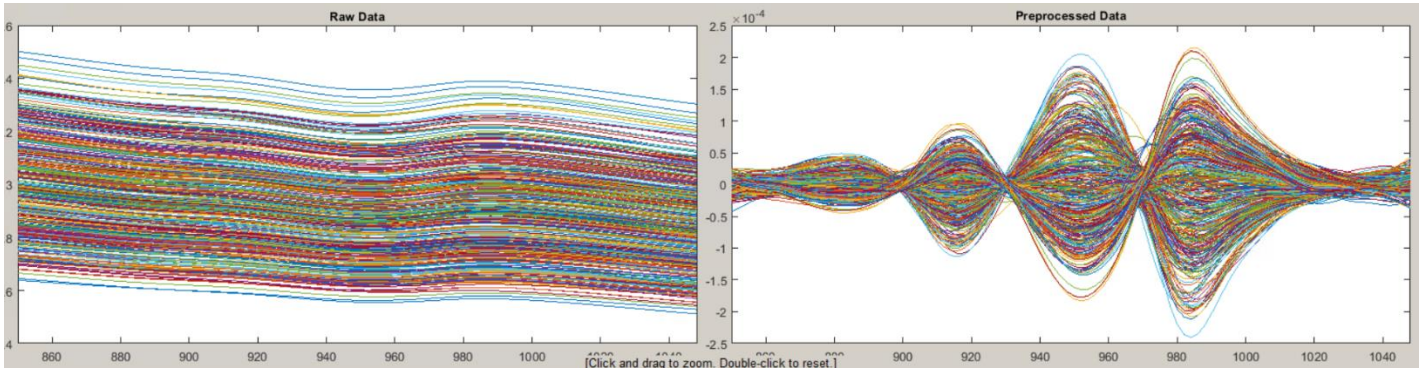
Nel caso delle X, ovvero i chicchi di grano, si possono distinguere tra i primi 415 che costituiscono il calibration set e i successivi 108 che formano il test set. Tutti i chicchi sono stati scelti casualmente da campioni sfusi. Inoltre, sono stati acquisiti nello stesso momento, ma quelli del test set sono stati conservati per circa 2 mesi aggiuntivi prima della misurazione al fine di verificare la dipendenza dal tempo nei campioni e nelle apparecchiature.

Le Y contengono il contenuto proteico.

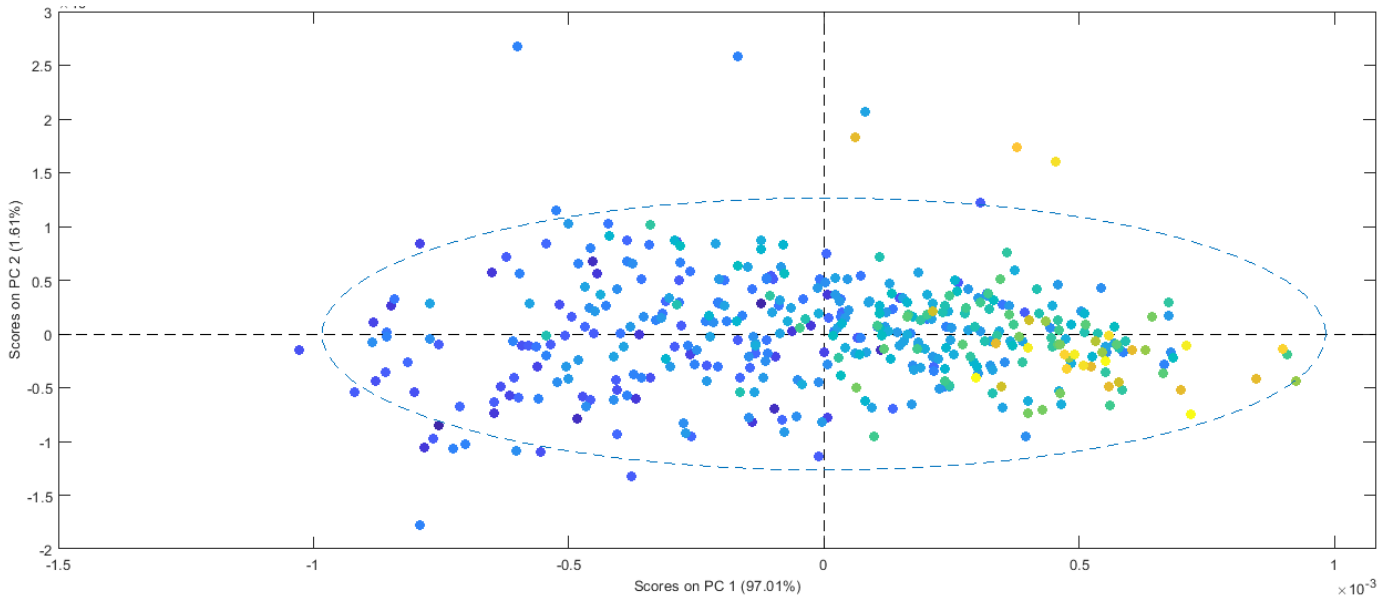
Il primo passo è fare preprocessing per le X e le Y del calibration set. Per le Y basta il mean centering, mentre per le X si sono provate diverse alternative in aggiunta al mean centering tra cui:

- Weighted baseline di ordine 3;
- MSC;
- 2^{nd} Derivative;
- Combinazioni di questi metodi.

Il risultato che nello score plot ha mostrato una divisione migliore tra i vari elementi in base alle Y è il Weighted baseline di ordine 3 combinato con il 2^{nd} Derivative e il mean centering. Qua è presente l'immagine dei plot con a sinistra i dati grezzi e a destra i dati preprocessati.

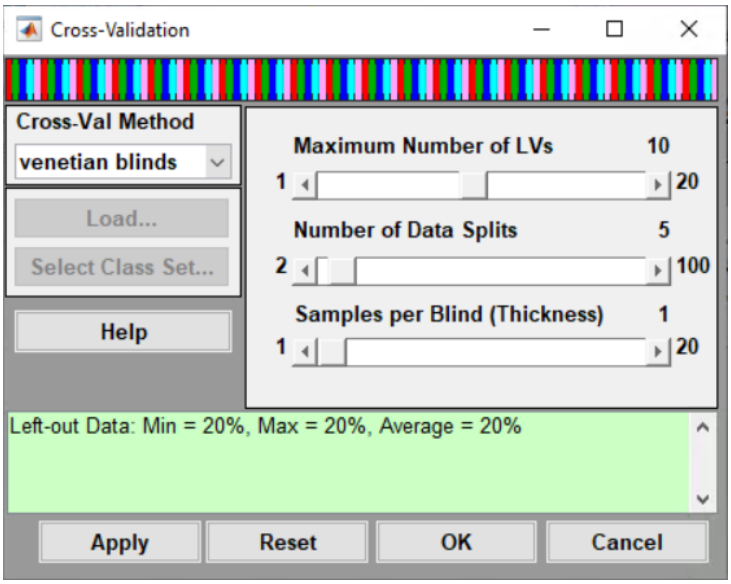


Qua sotto invece viene mostrato lo score plot accennato prima.

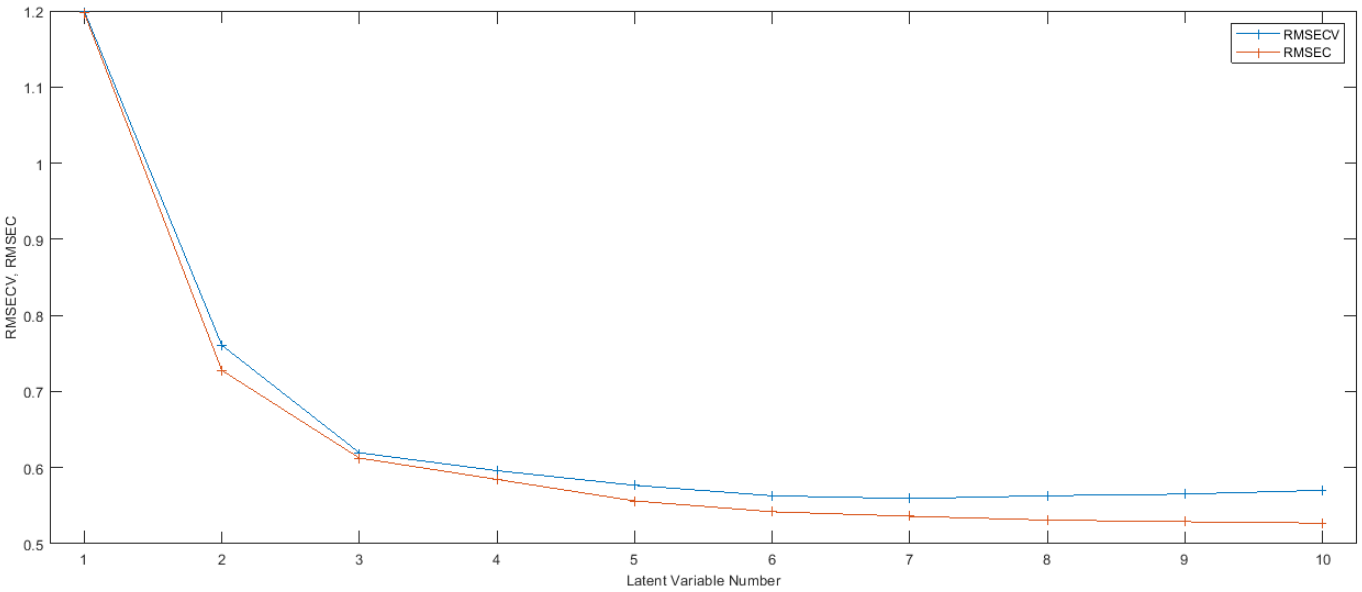


Per quanto concerne la cross-validazione è stata scelta la tecnica “Venetian blinds”, che consiste in una preliminare divisione del dataset in splits e per ciascuno dei raggruppamenti si porta fuori un campione per volta. Ciò permette il miglioramento della capacità di generalizzazione del modello. In questo caso il numero scelto di splits è 5, quindi si fa la cross-validazione togliendo 83 (415/5) campioni alla volta per un totale di 332 elementi su cui si ricalcola per ogni ciclo fino a che ogni campione non è stato lasciato fuori una volta.

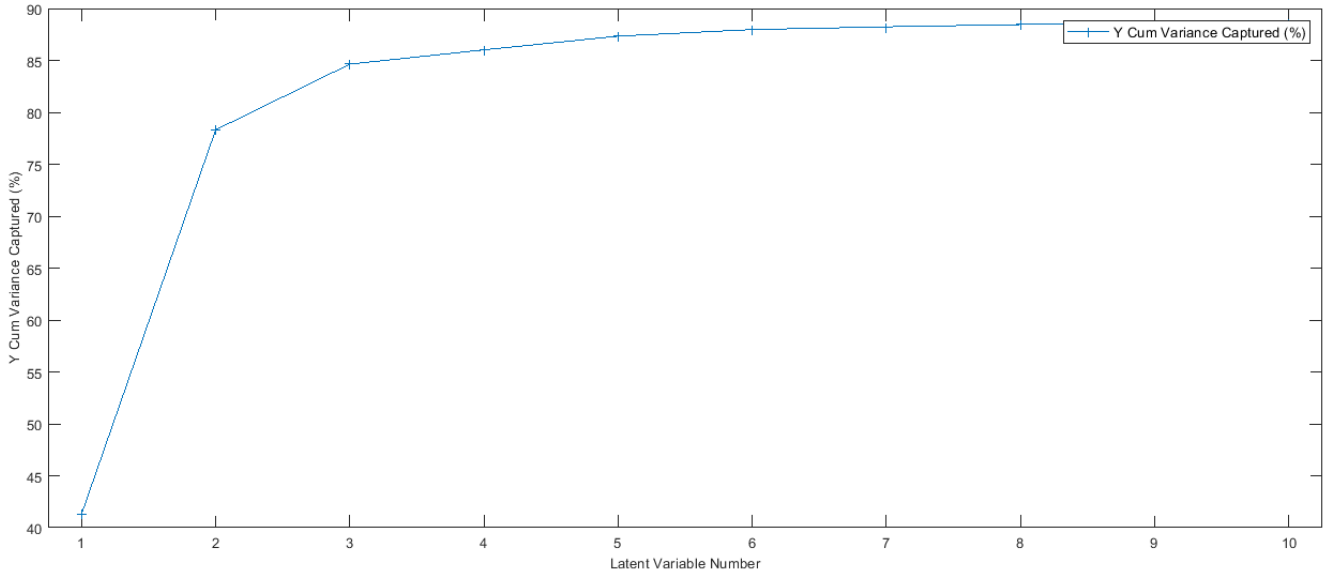
Mentre il numero di componenti principali deve essere ancora deciso, ma per stare larghi si prendono i primi 10.



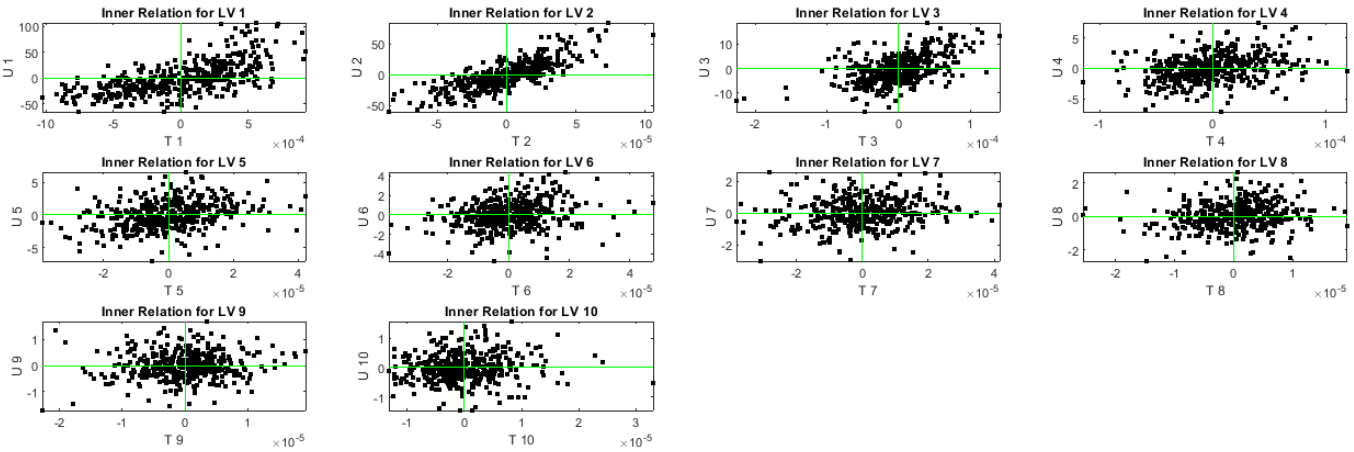
Per la scelta delle componenti principali, si è guardato per prima cosa il grafico dell'errore in fit (RMSEC) e in cross-validazione (RMSECV) e da lì si è capito che un buon numero può essere 3 o 4.



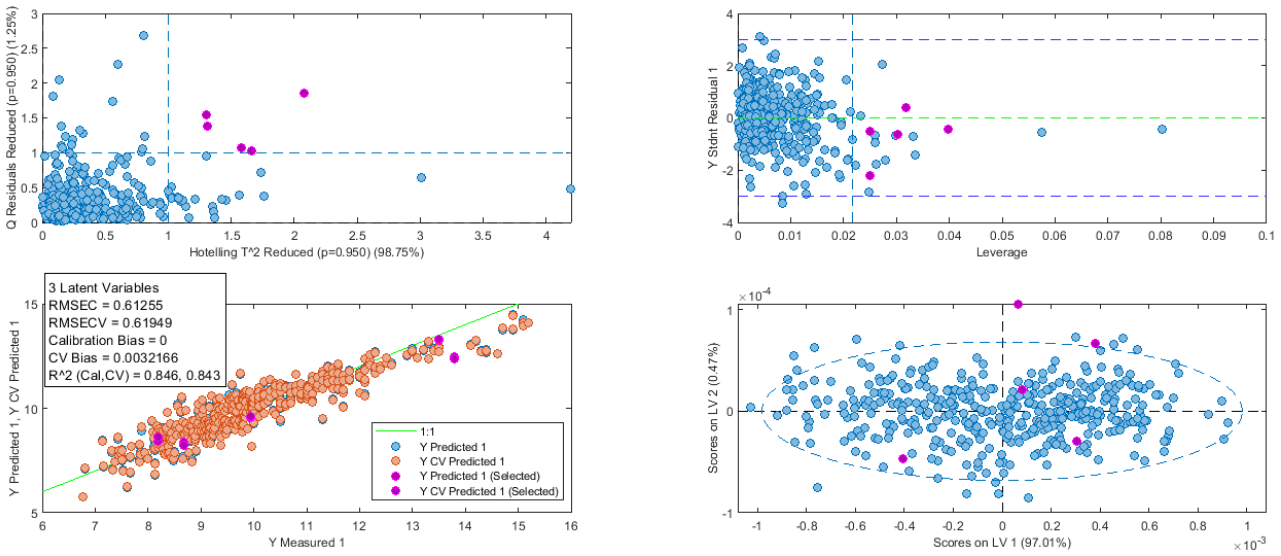
Per esserne più certi si è osservato anche il grafico della varianza cumulativa delle Y e risulta più chiaro che il numero giusto di componenti principali sia 3, dato che la quarta non è troppo differente.



Inoltre, sono stati fatti grafici delle relazioni interne per ciascuna componente principale e da questi si può notare che le prime tre seguono un trend approssimativamente lineare, ma dalla quarta in poi comincia ad essere tutto molto sparso, il che conferma la scelta delle tre componenti.



Una volta confermato di volere tre componenti principali, si comincia a guardare i grafici per vedere come sono distribuiti i campioni.

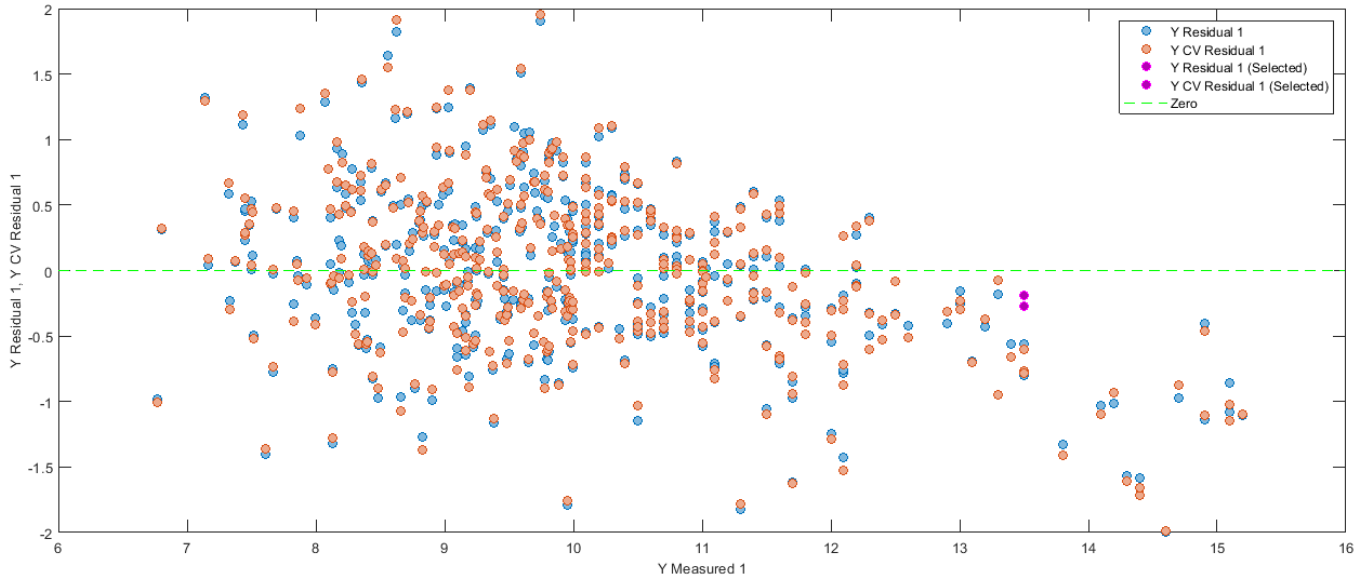


Dal grafico T^2 contro Q (in alto a sinistra) si possono individuare dei campioni (che sono quelli evidenziati in viola) che possono essere dei potenziali outliers, ma si è scelto di non rimuoverli perché dagli altri grafici si possono vedere comportarsi come gli altri. In particolare, dal grafico dei residui standardizzati delle Y (in alto a destra) si possono vedere sempre dentro al range (come quasi tutti gli altri campioni), quindi sono estremi in X ma non comportano una distorsione del modello che uso per prevedere la Y.

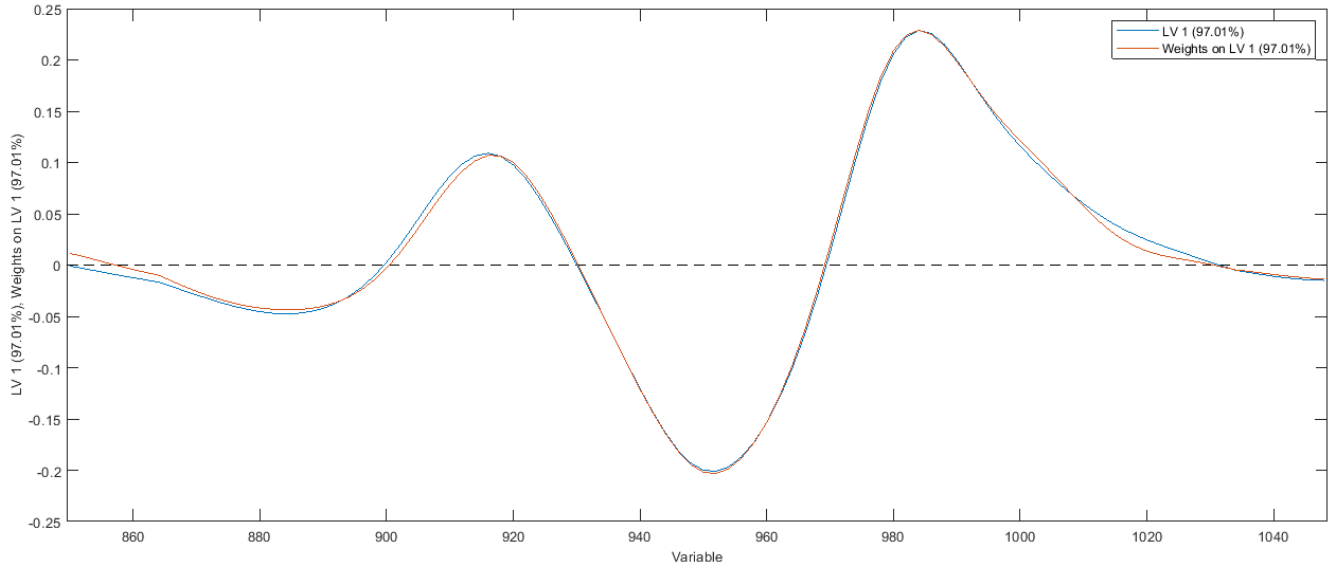
Nel grafico delle Y misurate contro le predette e le predette in cross-validazione (in basso a sinistra), si può notare come:

- I bias hanno valori molto ridotti, in fit è 0 mentre in cross-validazione è molto piccolo, il che è normale visto che si sta introducendo dell'errore sistematico;
- L'errore in fit e in cross-validazione hanno valori molto simili, quindi l'errore è ottimale;
- I campioni in fit e in cross-validazione sono disposti in maniera molto simile intorno alla retta.

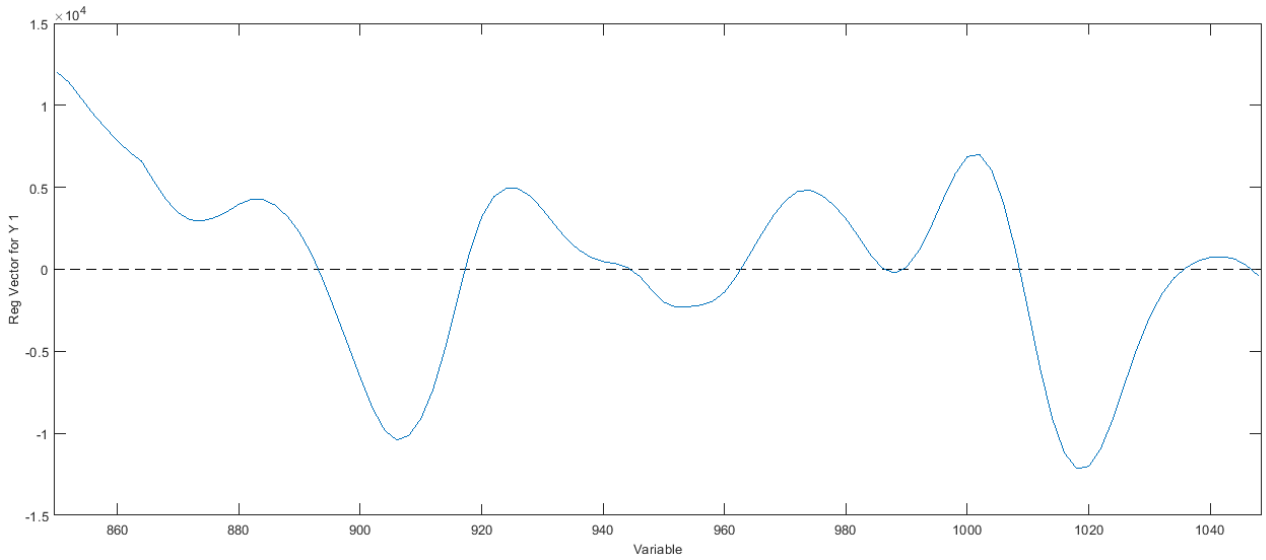
Anche il grafico dei residui mostra una disposizione molto simile tra i campioni in fit e quelli in cross-validazione.



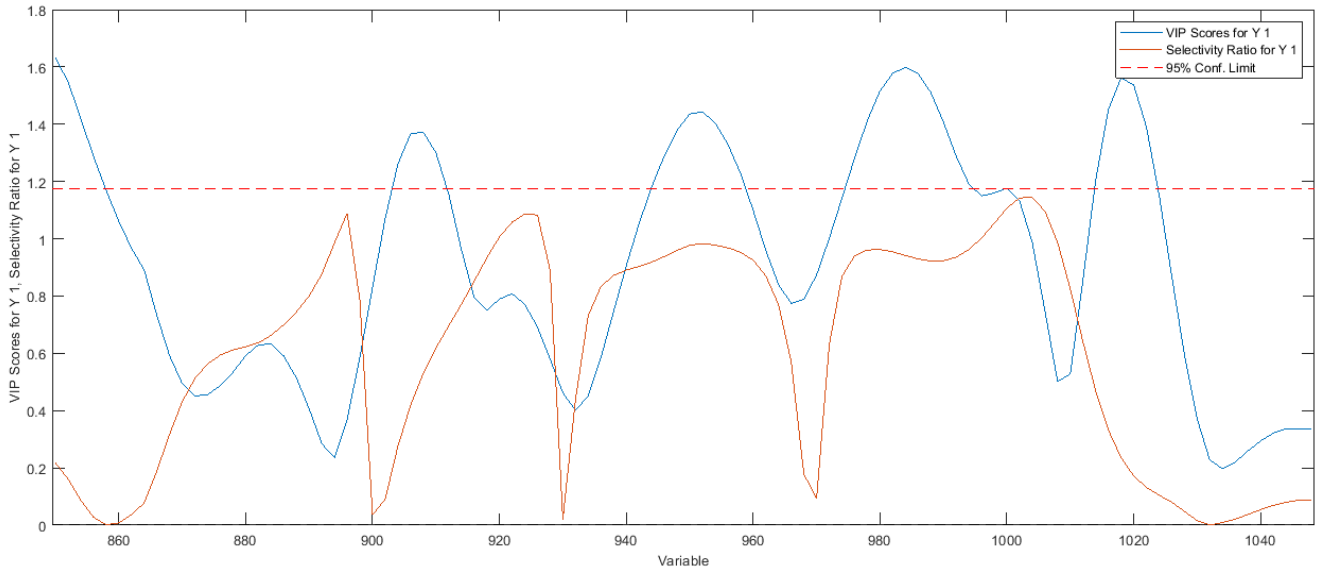
Un altro controllo che si può fare è verificare che non ci siano regioni rumorose poco significative, per fare ciò si può vedere il grafico dei loadings delle X con i pesi, che non mostra nulla che meriti una rimozione.



Anche i coefficienti di regressione sono poco rumorosi.



Nemmeno nel grafico dei VIP (Variable Influence in Projection) con la selectivity ratio si notano regioni rumorose che conviene rimuovere.



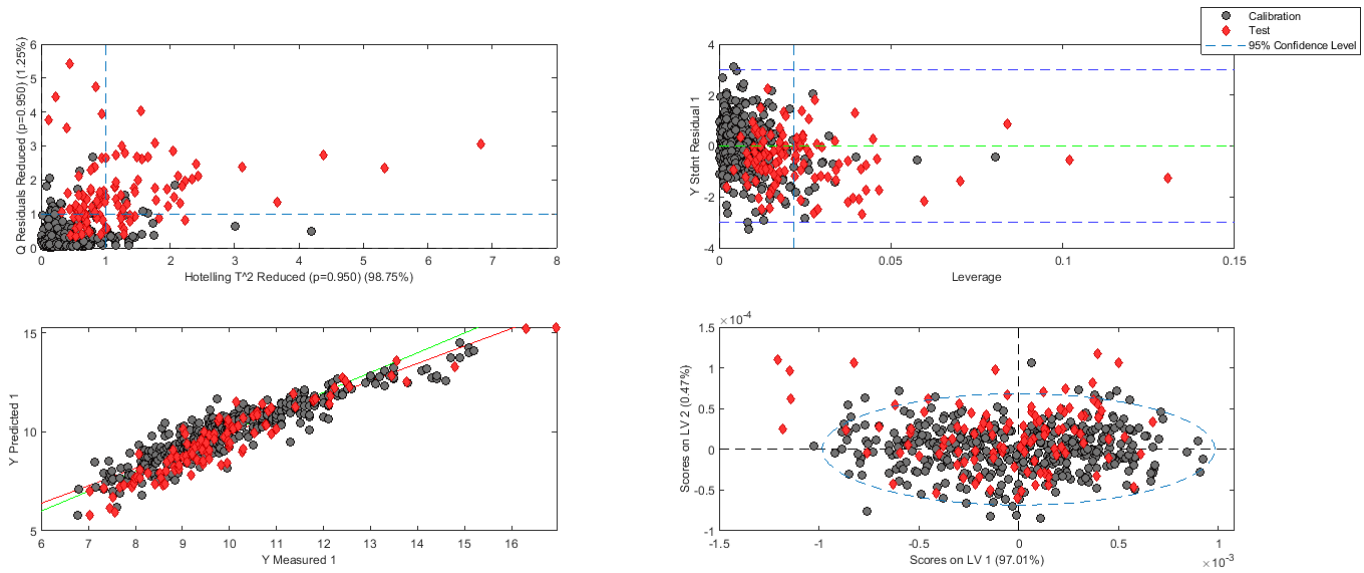
Successivamente, con l'introduzione delle X e delle Y del test set, si può applicare il modello in predizione. Il risultato è che l'errore in predizione (RMSEP) è di circa 0.1 peggiore dell'errore in fit (RMSEC) e in cross-validazione (RMSECV). Quindi questo si può definire un buon modello in quanto in predizione commette un errore molto simile a quelli commessi sul set di calibrazione e sulla cross-validazione.

```
Linear regression model using
Partial Least Squares calculated with the SIMPLS algorithm
Developed 22-Dec-2023 00:19:09.785
Author: Utente@DSCG-W10-006
X-block:   108 by 100 (Utente@DSCG-W10-006@20231222T001844.12998527 m:20231222001844.129)
Included: [ 1-108 ]  [ 1-100 ]
Included (in axis units): [ n/a ]  [ 850-1048 ]
Preprocessing: Baseline (Automatic Weighted Least Squares, order=3), 2nd Derivative (order: 2, window: 15 pt, tails: polyinterp), Mean Center
Y-block:   108 by 1 (Utente@DSCG-W10-006@20231222T001848.55303507 m:20231222001848.553)
Included: [ 1-108 ]  [ 1 ]
Preprocessing: Mean Center
Num. LVs: 3
Cross validation: venetian blinds w/ 5 splits and blind thickness = 1
Label: Y Column 1
mean : 9.8399
std  : 1.7489
n    :   108
min  : 7.03188
max  : 16.9509
RMSEC: 0.612551
RMSECV: 0.619489
RMSEP: 0.711282
Bias:     0
CV Bias: 0.00321661
Pred Bias:-0.332533
R^2 Cal: 0.846416
R^2 CV: 0.842928
R^2 Pred: 0.88164

    Percent Variance Captured by Regression Model

          -----X-Block-----    -----Y-Block-----
   Comp     This      Total      This      Total
   ----    -------   -------    -------   -------
     1      97.01     97.01      41.32     41.32
     2       0.47     97.48      37.00     78.32
     3       1.27     98.75       6.32     84.64
```

L'ultima osservazione che si vuole fare è basata sui grafici di score ricalcolati con l'aggiunta dei campioni test. Questi presentano una situazione simile a quella descritta precedentemente, però è osservabile che nel grafico T^2 contro Q (in alto a sinistra) una buona parte dei campioni di test eccedono di questi valori, ma rimangono regolari nel grafico dei residui standardizzati delle Y (in alto a destra) il che li rende campioni accettabili, anche se un po' diversi da quelli di calibrazione.



In conclusione, si è costruito un modello predittivo con PLS con un errore sul test set molto simile a quelli sul calibration set e sulla cross-validazione e per ottenere ciò ha avuto queste caratteristiche:

- Preprocessing sulle X con Weighted baseline di ordine 3, il 2nd Derivative e il mean centering;
- Preprocessing sulle Y con mean centering;
- Cross-validazione con “Venetian blinds” e 5 come numero dei data splits;
- 3 componenti principali;
- Non presenta regioni rumorose.