UNIVERSITÀ
DI TORINO

Modello Bayesiano sparso a infiniti fattori latenti: applicazioni a dati di spettroscopia del vicino infrarosso

Massimo Armano

Curse of dimensionality:il paradigma Large–p,Small–n

UNIVERSITÀ
DI TORINO

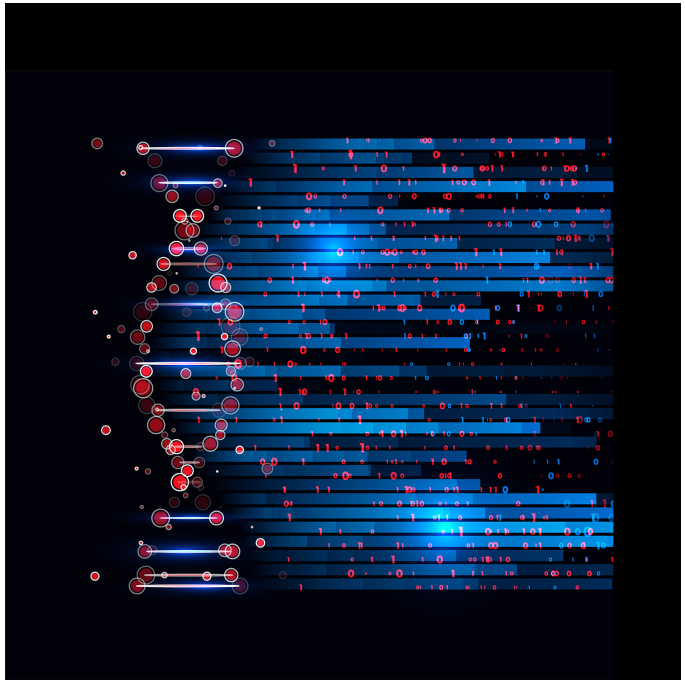
In molti contesti statistici moderni, il numero di variabili (p) supera di gran lunga il numero di osservazioni (n). Questa struttura rende difficile l'inferenza classica e richiede approcci specifici.

Curse of dimensionality:il paradigma Large–p,Small–n

In molti contesti statistici moderni, il numero di variabili (p) supera di gran lunga il numero di osservazioni (n). Questa struttura rende difficile l'inferenza classica e richiede approcci specifici.

Alcuni esempi:

- Dati genomici
- Immagini e segnali ad alta risoluzione
- Dati ambientali o sensoriali complessi
- Dati di spettroscopia



Fondamenti di statistica Bayesiana

UNIVERSITÀ DI TORINO

Nell'approccio bayesiano i parametri sono trattati come **variabili aleatorie** soggette a incertezza.

Principi essenziali:

- Si assegna una **prior** $\pi(\theta)$, che rappresenta l'informazione a priori sul parametro.
- I dati sono modellati tramite la **verosimiglianza** $p(x \mid \theta)$.
- L'inferenza si basa sulla **posterior**:

$$\pi(\theta \mid x) = \frac{p(x \mid \theta)\pi(\theta)}{p(x)}$$

La stima dei parametri deriva dalla posterior. Quando la sua forma esplicita è inaccessibile, si utilizzano metodi **Monte Carlo Markov Chain (MCMC)** per approssimarla tramite campionamento.

Modello fattoriale: struttura e riduzione dimensionale

I modelli fattoriali offrono una soluzione efficace ai contesti $p \gg n$, riducendo p variabili osservate in k fattori latenti, con $k \ll p$.

Si può definire ogni osservazione x_i p-dimensionale come:

$$x_i = \mu + \Lambda f_i + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}_p(0, \Psi)$$

dove:

- $\Lambda \in \mathbb{R}^{p \times k}$: matrice dei carichi fattoriali λ_{jh}
- $f_i \in \mathbb{R}^k$: fattori latenti, $f_i \sim \mathcal{N}(0, I_k)$
- $\Psi \in \mathbb{R}^{p \times p}$: matrice diagonale delle varianze specifiche

Di conseguenza la matrice di Varianza Covarianza:

$$\Sigma = \Lambda\Lambda^\top + \Psi$$

La struttura consente una riduzione dimensionale: invece di stimare una matrice $\Sigma \in \mathbb{R}^{p \times p}$, si stimano Λ e Ψ, con Λ di dimensione ridotta se $k \ll p$.

Lo Sparse Bayesian Infinite Factor Model

Il modello introduce prior di **shrinkage**, che riducono progressivamente l'influenza dei fattori meno rilevanti, permettendo di concentrare la varianza spiegata su meno fattori.

Prior di Shrinkage sui carichi fattoriali:

$$\lambda_{jh} \sim \mathcal{N}(0, (\phi_{jh}\tau_h)^{-1})$$

- $\phi_{jh} \sim \text{Gamma}(\frac{\nu}{2}, \frac{\nu}{2})$: *shrinkage locale* (regola l'importanza di ogni singolo carico)
- $\tau_h = \prod_{\ell=1}^{h} \delta_\ell$: *shrinkage globale* con il **Multiplicative Gamma Process (MGP)**

$$\delta_1 \sim \Gamma(a_1, 1), \quad \delta_\ell \sim \Gamma(a_2, 1), \ \ell \geq 2$$

Il MGP impone una penalizzazione crescente su λ_{jh} al crescere di h, grazie a questa struttura, il numero effettivo di fattori k non va scelto a priori ma viene **determinato automaticamente e in modo adattivo** dal modello stesso.

Stima della matrice Σ e dei coefficienti β

Dal campionamento MCMC, otteniamo stime per:

$$\Lambda \in \mathbb{R}^{p \times k}, \quad \Psi \in \mathbb{R}^{p \times p} \text{ (diagonale)}$$

Queste consentono di ricostruire la matrice di covarianza:

$$\Sigma = \Lambda\Lambda^\top + \Psi$$

Assumendo normalità congiunta tra la variabile di risposta z_i e i predittori $y_i \in \mathbb{R}^{p-1}$:

$$x_i = \begin{bmatrix} z_i \\ y_i \end{bmatrix} \sim \mathcal{N}_p \left(0, \begin{bmatrix} \Sigma_{zz} & \Sigma_{zy} \\ \Sigma_{yz} & \Sigma_{yy} \end{bmatrix} \right)$$

Dalla normalità condizionata si ottiene:

$$z_i \mid y_i \sim \mathcal{N}(y_i^\top \beta, \Sigma_{z|y}), \quad \text{con} \quad \beta = \Sigma_{yy}^{-1}\Sigma_{yz}$$

(dove: $\Sigma_{yy} \in \mathbb{R}^{(p-1)\times(p-1)}$, $\Sigma_{yz} \in \mathbb{R}^{(p-1)\times 1}$, $\beta \in \mathbb{R}^{(p-1)\times 1}$)

Valutazione del modello su dati simulati

Dati simulati da un modello fattoriale con $p = 30$ variabili e $k = 5$ fattori latenti.

$$\Lambda \in \mathbb{R}^{30\times 5}, \quad \Psi = 0.01 \cdot I_{30}$$

$$\Sigma = \Lambda\Lambda^\top + \Psi$$

Λ costruita sparsa, con valori diversi da zero solo su righe selezionate casualmente. Vettori $x_i \sim \mathcal{N}_{30}(0, \Sigma)$, con $n = 200$ osservazioni simulate.

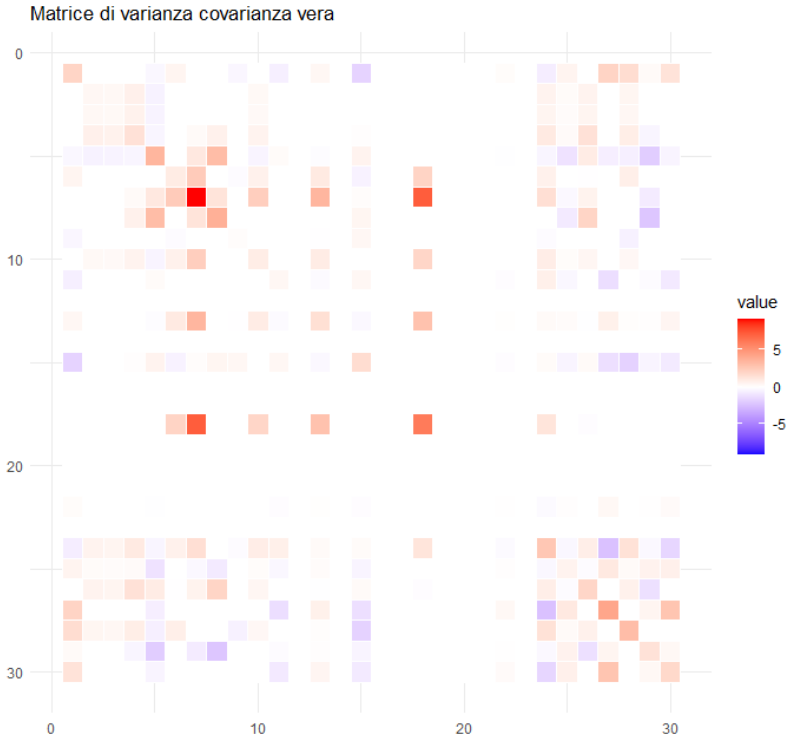
Valutazione del modello su dati simulati

Dati simulati da un modello fattoriale con $p = 30$ variabili e $k = 5$ fattori latenti.

$$\Lambda \in \mathbb{R}^{30\times 5}, \quad \Psi = 0.01 \cdot I_{30}$$

$$\Sigma = \Lambda\Lambda^\top + \Psi$$

Λ costruita sparsa, con valori diversi da zero solo su righe selezionate casualmente. Vettori $x_i \sim \mathcal{N}_{30}(0, \Sigma)$, con $n = 200$ osservazioni simulate.



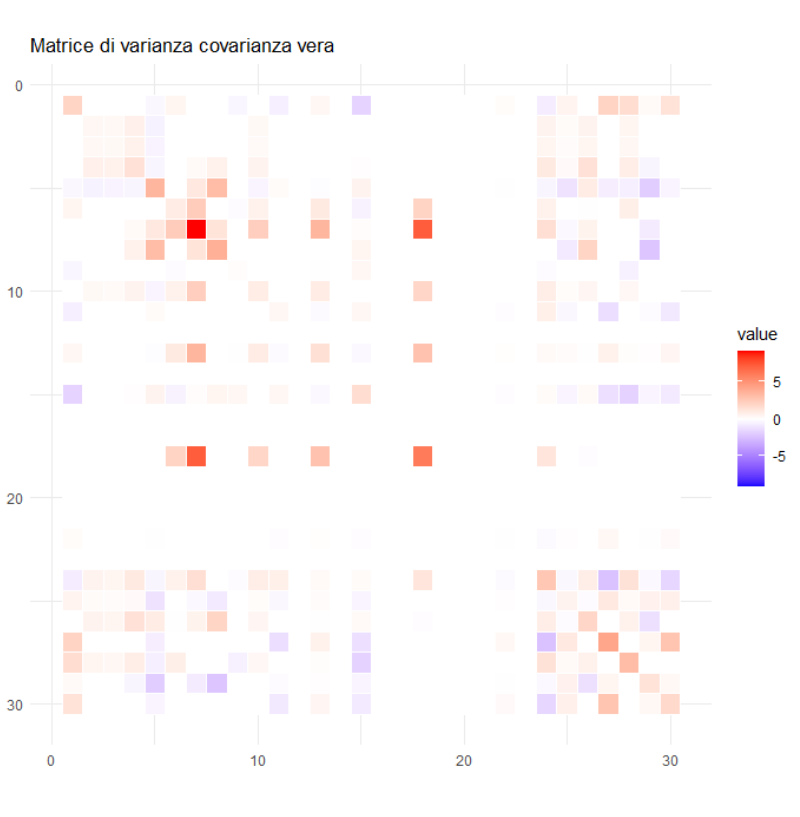
Valutazione del modello su dati simulati

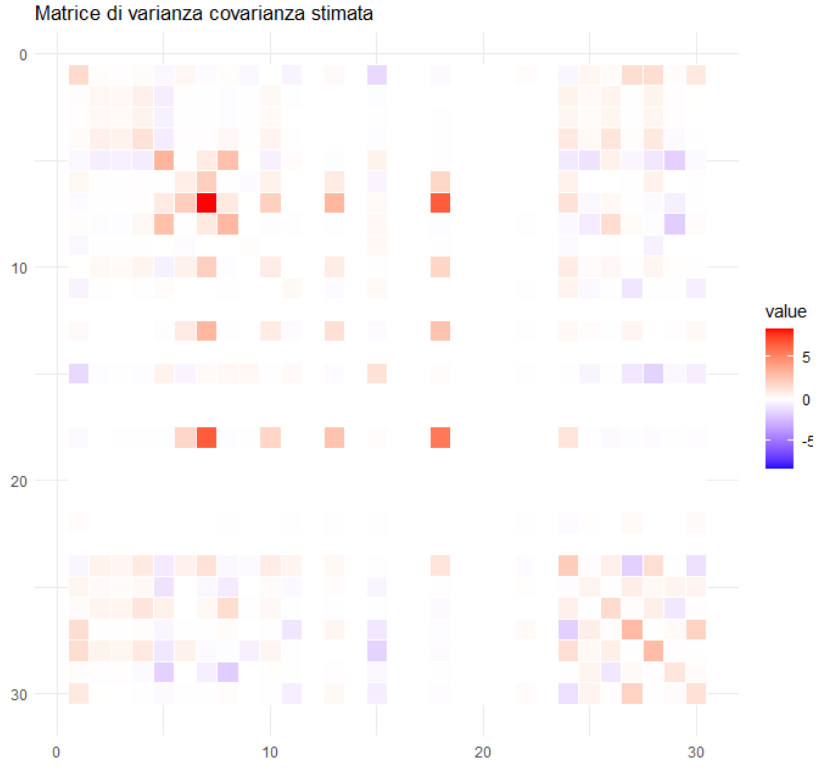
Dati simulati da un modello fattoriale con $p = 30$ variabili e $k = 5$ fattori latenti.

$$\Lambda \in \mathbb{R}^{30\times 5}, \quad \Psi = 0.01 \cdot I_{30}$$

$$\Sigma = \Lambda\Lambda^\top + \Psi$$

Λ costruita sparsa, con valori diversi da zero solo su righe selezionate casualmente. Vettori $x_i \sim \mathcal{N}_{30}(0, \Sigma)$, con $n = 200$ osservazioni simulate.



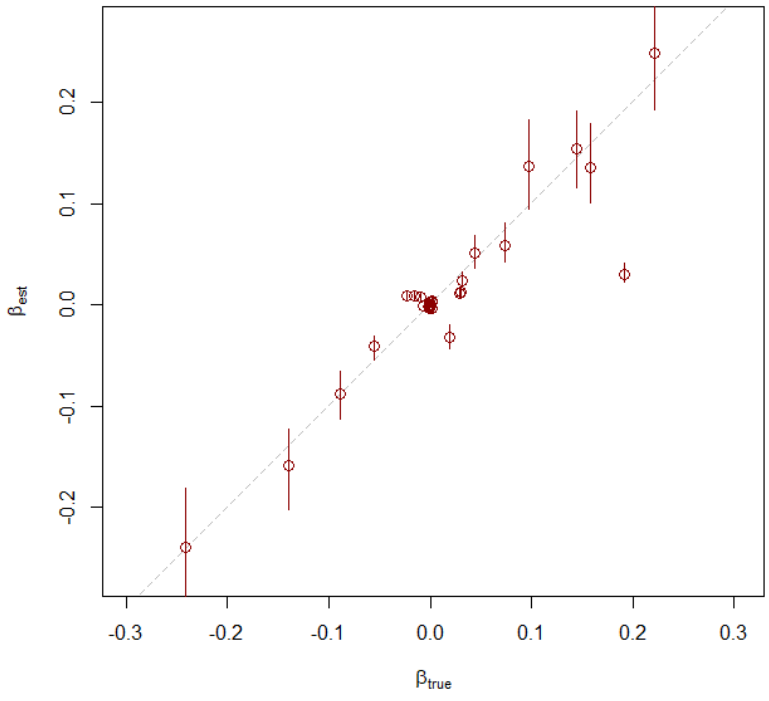


Valutazione del modello su dati simulati

Dati generati come nel caso precedente, con $p = 100$ variabili e $k = 5$ fattori latenti.

Valutazione del modello su dati simulati

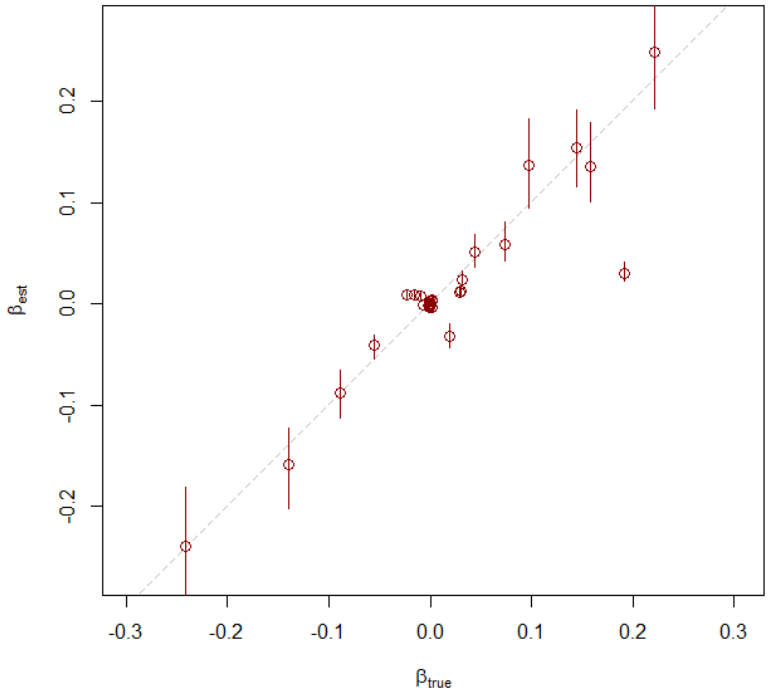
Dati generati come nel caso precedente, con $p = 100$ variabili e $k = 5$ fattori latenti.



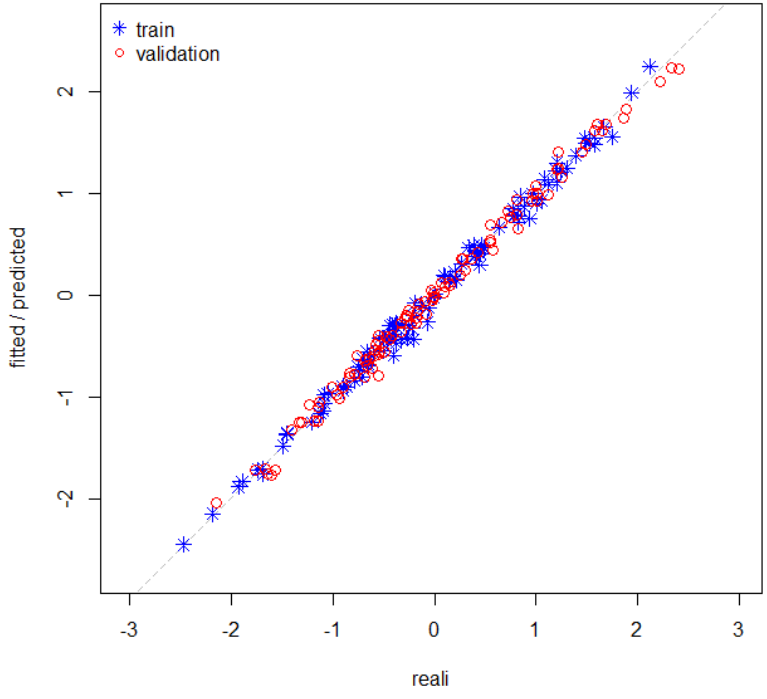
Coefficienti $\beta = \Sigma_{yy}^{-1}\Sigma_{yz}$

Valutazione del modello su dati simulati

Dati generati come nel caso precedente, con $p = 100$ variabili e $k = 5$ fattori latenti.



Coefficienti $\beta = \Sigma_{yy}^{-1}\Sigma_{yz}$



Predizioni vs. Osservati

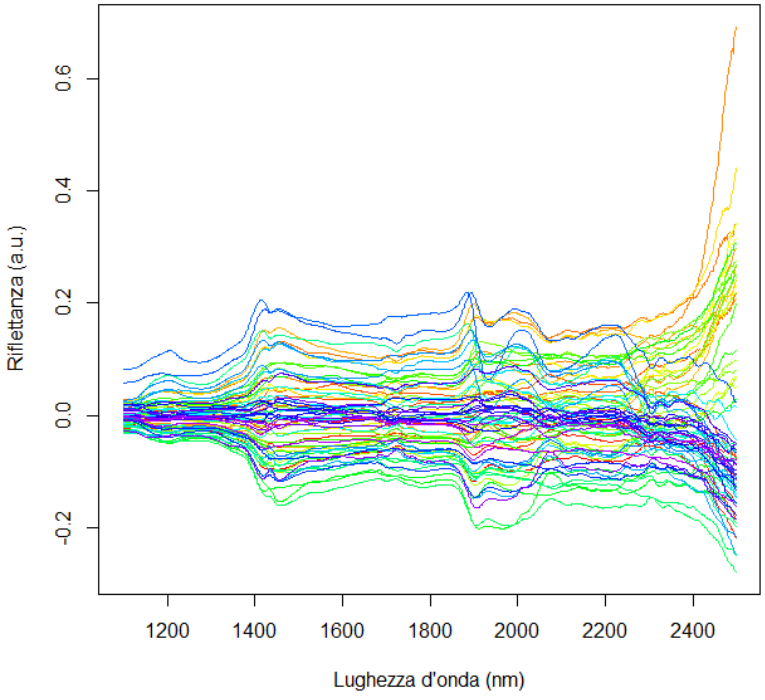
Introduzione ai dati NIRS

La spettroscopia del vicino infrarosso (NIRS) è una tecnica analitica non distruttiva, rapida e a basso costo, sempre più utilizzata in ambito scientifico e industriale.

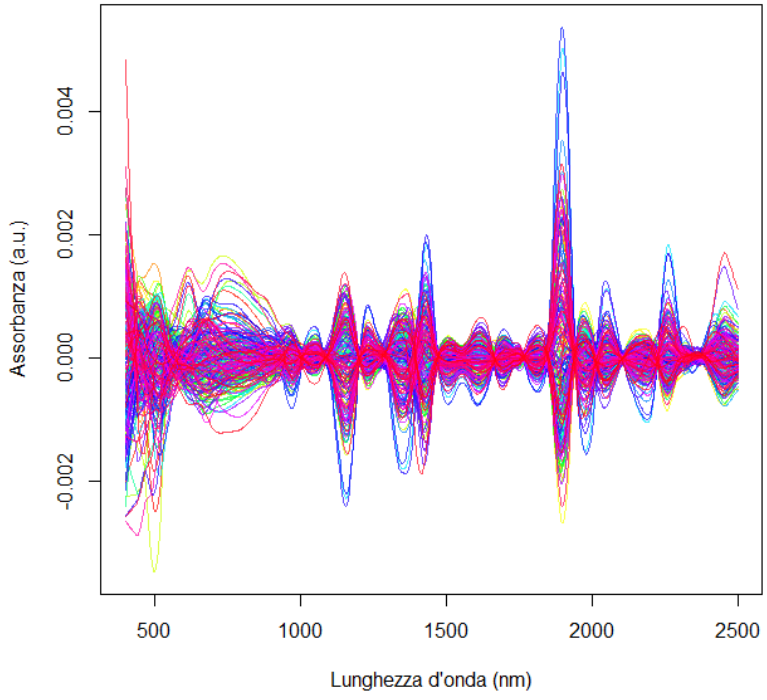
Essa consente di acquisire spettri altamente informativi, legati alla struttura e alla composizione del campione.

Sono state effettuate due applicazioni:

NIRS sull'impasto dei biscotti



NIRS sul grano



Applicazione I — Impasto di biscotti (NIRS)

Dati NIRS raccolti su impasto di biscotti (Osborne et al., 1984). Ogni campione è descritto da 256 punti spettrali (intervallo 1380–2498 nm).

Il dataset include $n = 72$ osservazioni totali:

- **Training set:** 39 campioni
- **Test set:** 32 campioni

Variabili risposta: **grassi**, **saccarosio**, **farina**, **acqua** (in percentuale).

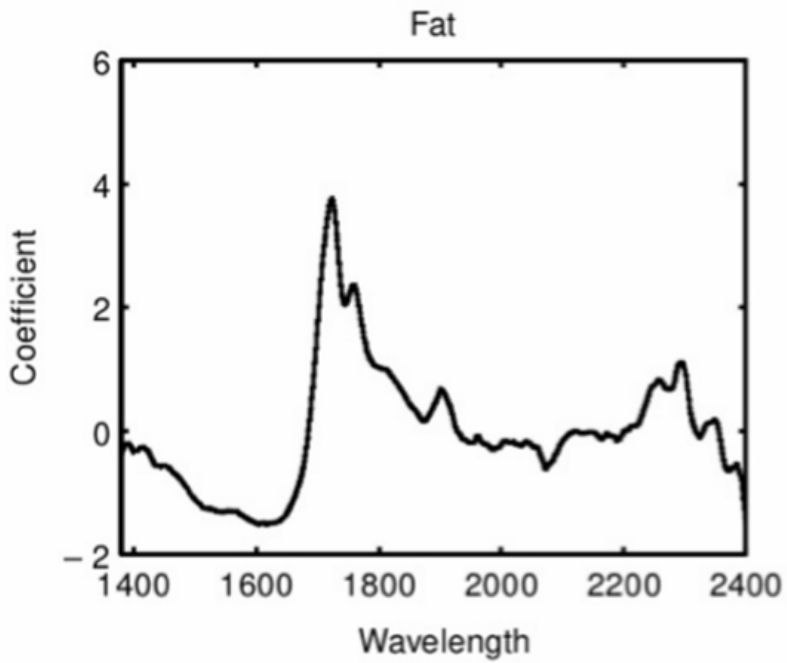
Applicazione I — Impasto di biscotti (NIRS)

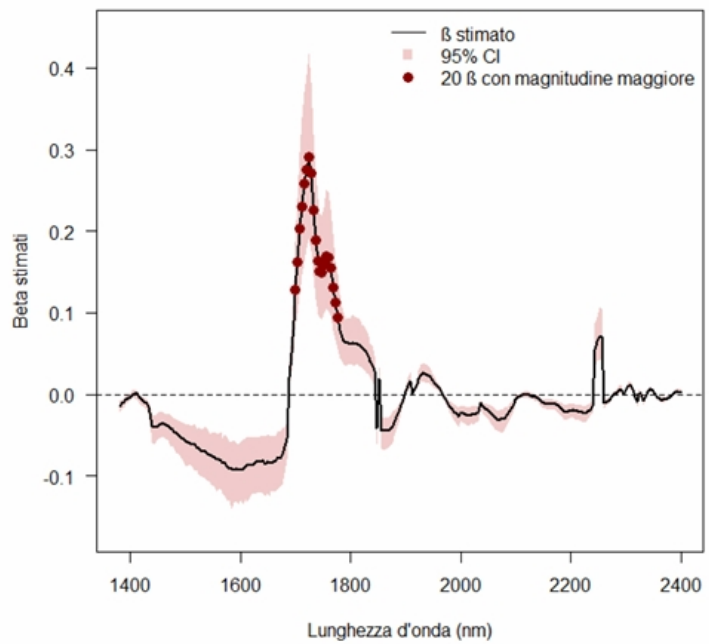
Dati NIRS raccolti su impasto di biscotti (Osborne et al., 1984). Ogni campione è descritto da 256 punti spettrali (intervallo 1380–2498 nm).

Il dataset include $n = 72$ osservazioni totali:

- **Training set:** 39 campioni
- **Test set:** 32 campioni

Variabili risposta: **grassi**, **saccarosio**, **farina**, **acqua** (in percentuale).





Coefficienti β stimati su spettri NIRS

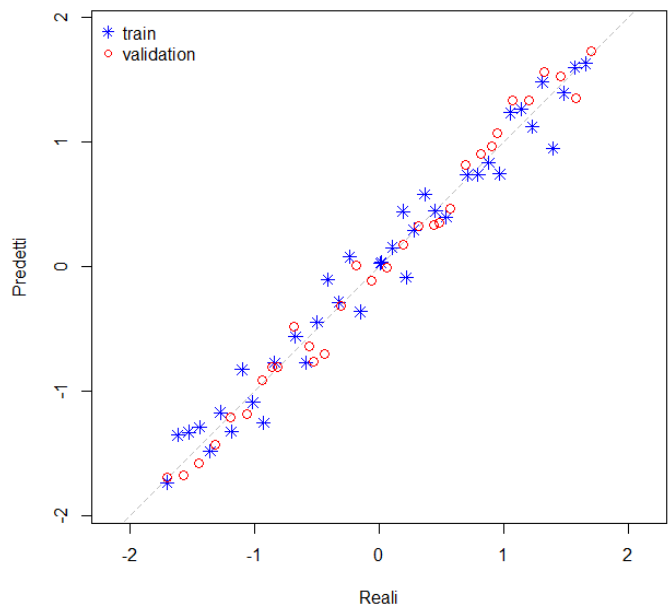
Applicazione I — Impasto di biscotti (NIRS)

Dati NIRS raccolti su impasto di biscotti (Osborne et al., 1984). Ogni campione è descritto da 256 punti spettrali (intervallo 1380–2498 nm).

Il dataset include $n = 72$ osservazioni totali:

- **Training set:** 39 campioni
- **Test set:** 32 campioni

Variabili risposta: **grassi**, **saccarosio**, **farina**, **acqua** (in percentuale).



Predizioni vs. osservati — contenuto di grassi

Confronto delle performance predittive

MSPE: Risultati dei modelli tratti da Brown et al. (2001) vs. risultati dei modelli implementati

UNIVERSITÀ DI TORINO

Metodo	Grassi	Saccarosio	Farina	Acqua
Risultati benchmark da Brown et al. (2001)				
Regressione MLR stepwise	0.044	1.188	0.722	0.221
Bayesian Decision Theory	0.076	0.566	0.265	0.176
PLS	0.151	0.583	0.375	0.105
PCR	0.160	0.614	0.388	0.106
Bayesian wavelet regression sviluppata da Brown et al. (2001)				
Media su 500 modelli	0.063	0.449	0.348	0.050
Miglior modello	0.059	0.466	0.351	**0.047**
Modelli implementati nel presente lavoro				
Ridge regression	0.020	0.059	**0.068**	0.052
S.B.I.F.M.	**0.016**	**0.024**	0.073	0.058

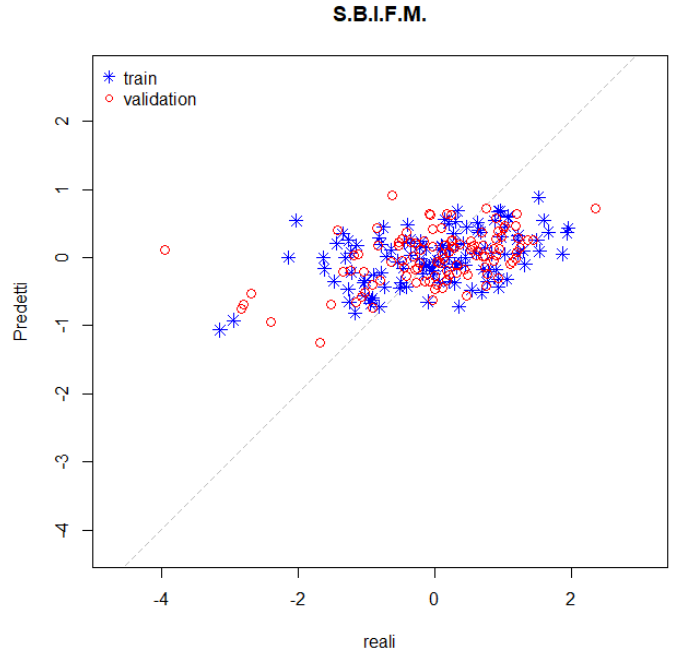
MSPE: Mean Squared Prediction Error per ciascuna variabile risposta.

Applicazione II — NIRS sul grano

Dati NIRS di assorbanza acquisiti su campioni di grano coltivati in condizioni controllate di stress idrico (Rincent et al., 2018).
$n = 223$ varietà (115 train, 108 test), $p = 1050$ punti spettrali (400–2500 nm). Variabile risposta: resa di grano.

Applicazione II — NIRS sul grano

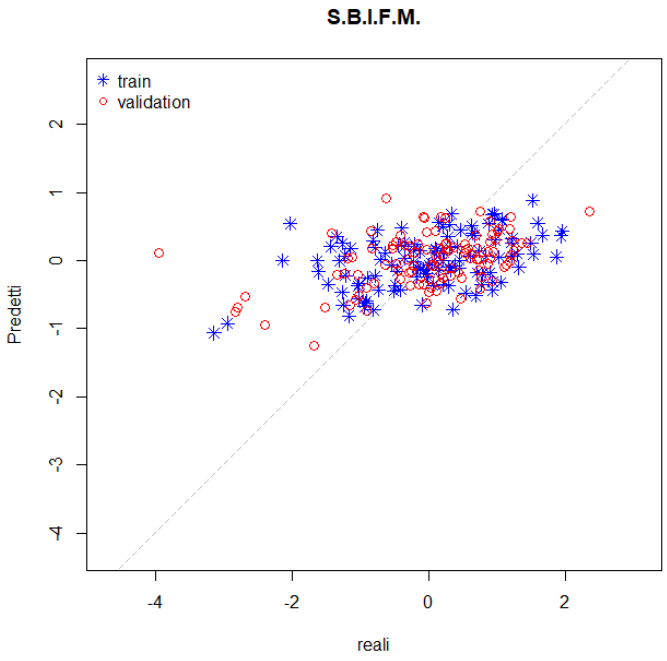
Dati NIRS di assorbanza acquisiti su campioni di grano coltivati in condizioni controllate di stress idrico (Rincent et al., 2018).
$n = 223$ varietà (115 train, 108 test), $p = 1050$ punti spettrali (400–2500 nm). Variabile risposta: resa di grano.



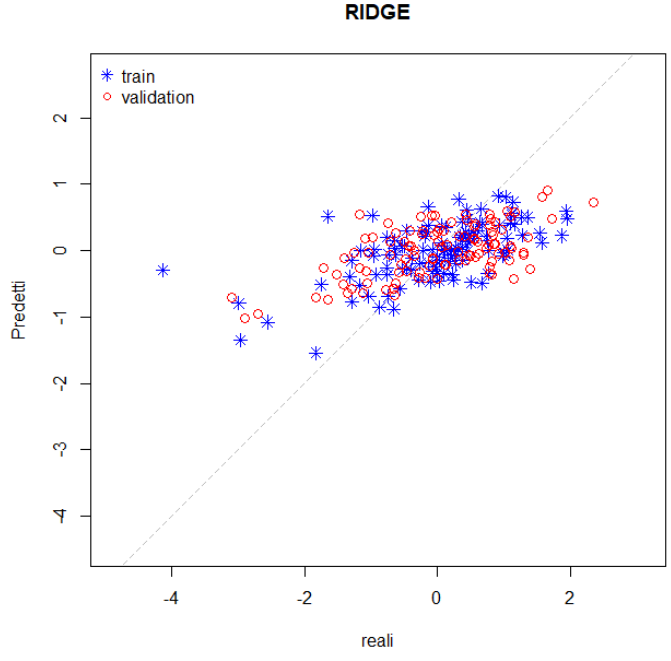
Predizioni vs. osservati SBIFM

Applicazione II — NIRS sul grano

Dati NIRS di assorbanza acquisiti su campioni di grano coltivati in condizioni controllate di stress idrico (Rincent et al., 2018).
$n = 223$ varietà (115 train, 108 test), $p = 1050$ punti spettrali (400–2500 nm). Variabile risposta: resa di grano.

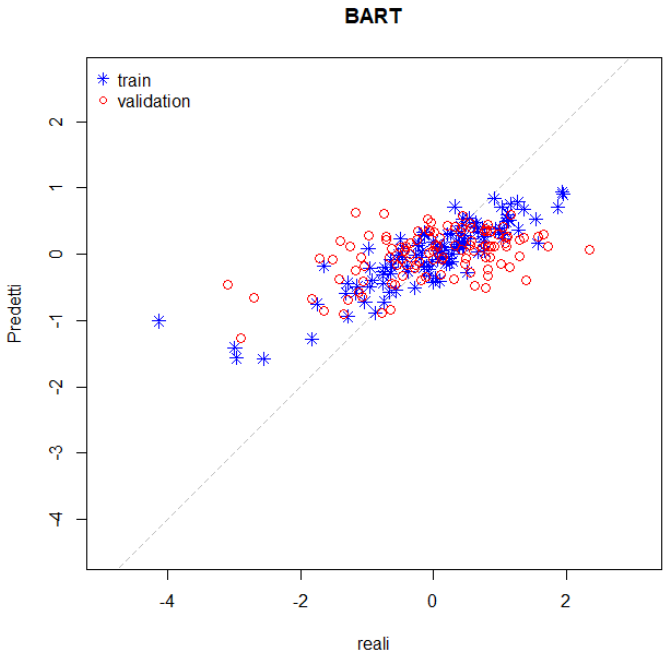


Predizioni vs. osservati SBIFM

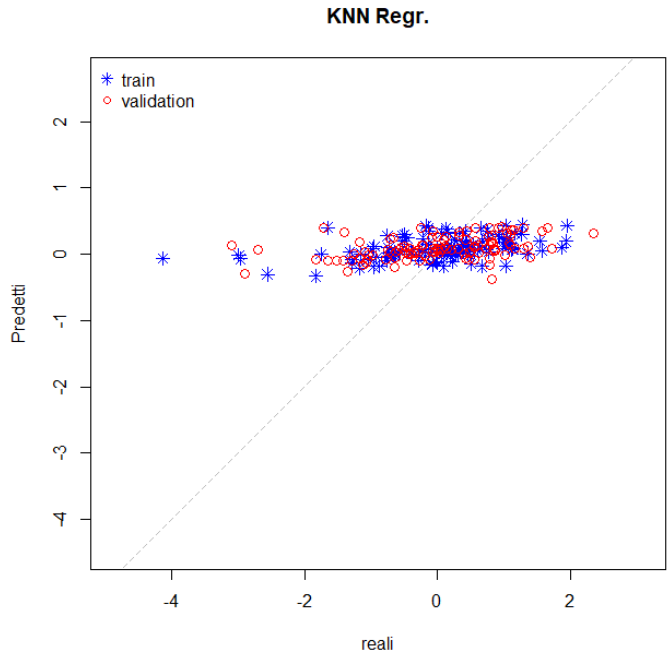


Predizioni vs. osservati Ridge

Applicazione II — NIRS sul grano

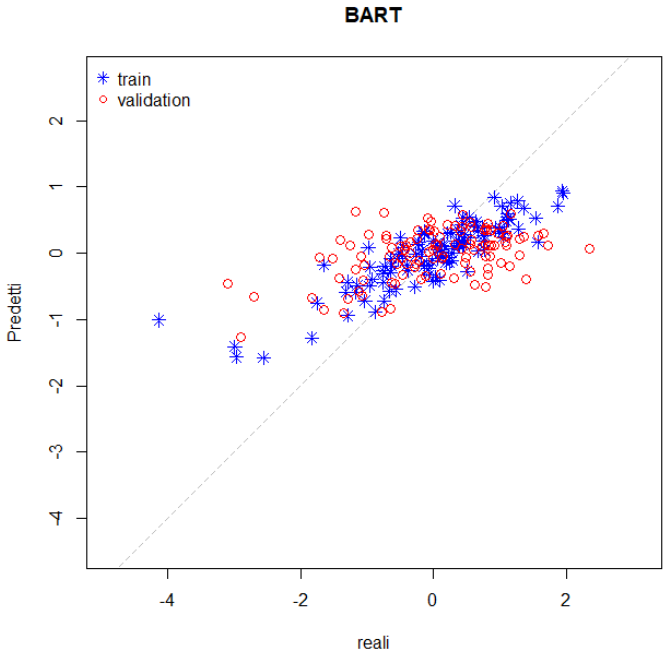


Predizioni vs. osservati BART

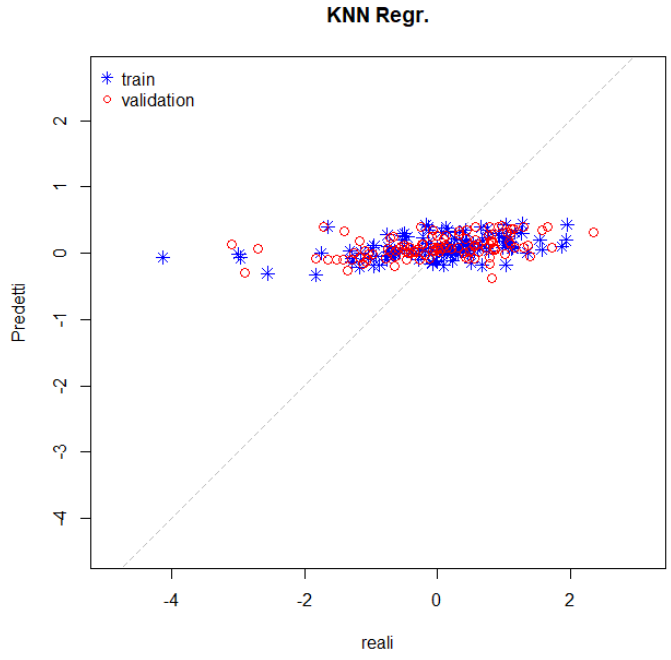


Predizioni vs. osservati KNN

Applicazione II — NIRS sul grano



Predizioni vs. osservati BART



Predizioni vs. osservati KNN

	S.B.I.F.M.	BART	KNN Regr.	Ridge
MSPE	0.81	0.76	0.89	**0.63**

Conclusioni

UNIVERSITÀ DI TORINO

- L'applicazione del modello ai dati sull'**impasto dei biscotti**, ampiamente studiati in letteratura, ha mostrato **ottime performance**, superando quasi tutti i benchmark considerati.
- L'applicazione ai dati sul **grano** ha invece fornito risultati **insoddisfacenti**. Anche altri metodi si sono dimostrati deboli su questo dataset, poco analizzato in letteratura.
- Ciò suggerisce che le difficoltà non dipendano solo dal modello: i soli dati NIRS potrebbero non essere sufficienti e potrebbe essere utile integrarli con **informazioni genetiche** delle varietà di grano.

UNIVERSITÀ
DI TORINO

Grazie per l'attenzione!