



**UNIVERSITÀ
DI TORINO**

Università degli Studi di Torino

*Corso di Laurea Magistrale in Metodi statistici ed economici
per le decisioni*

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

Tesi di Laurea

Relatore

Lanteri Alessandro

Correlatore

Kon Kam King Guillaume

Controrelatrice

Dalla Valle Luciana

Candidato

Armano Massimo

Matricola 915442

Anno Accademico 2024/2025

Indice

Introduzione	3
1 Introduzione all'Analisi Fattoriale Bayesiana	6
I L'Analisi Fattoriale	6
I.I Il modello	7
I.II Aspetti operativi	8
II Fondamenti di statistica bayesiana	9
II.I Stima puntuale, intervalli di credibilità e predizione .	11
III Metodi computazionali di Inferenza Bayesiana	13
III.I Panoramica sulle tecniche Markov Chain Monte Carlo (MCMC)	13
III.II Il Gibbs sampler	15
IV Analisi Fattoriale Bayesiana	16
IV.I Il modello	17
IV.II La stima: Gibbs sampler	20
2 Lo sparse Bayesian infinite factor model e la sua implemen- tazione	22
I Il Modello	22
I.I La stima: Gibbs sampler con troncatura adattiva . .	25
I.II Predizione tramite regressione fattoriale latente . . .	26
II Applicazione e sviluppo del metodo su dati simulati	28
3 Applicazioni a dati di spettroscopia del vicino infrarosso	40
I Applicazione ai NIRS sull'impasto dei biscotti	41
I.I Dataset, pre-processing e metodo	41
I.II Risultati	42
II Applicazione ai dati NIRS sul grano	51
II.I Dataset, pre-processing e metodo	52
II.II Risultati	54
Conclusioni	57
Appendice	62
I Codice implementato per la generazione dei dati simulati (come Bhattacharya and Dunson (2011))	62
II Codice implementato per lo studio di simulazione	62

III	Diagnosi e plot aggiuntivi	71
III.I	Applicazioni sui dati NIRS sull'impasto dei biscotti .	71
III.II	Applicazioni sui dati NIRS sul grano	73

Introduzione

Negli ultimi anni, la spettroscopia del vicino infrarosso (*Near Infrared Spectroscopy*, NIRS) ha assunto un ruolo di crescente rilievo nell'analisi di composti organici, grazie alla sua rapidità, ai costi contenuti e al carattere non invasivo della metodologia. La tecnologia NIRS consente di ottenere informazioni indirette sulla composizione chimico-fisica dei campioni, fornendo spettri di riflettanza o assorbanza che, tuttavia, non sono immediatamente informativi sulle proprietà di interesse. Per estrarre da tali dati conoscenze utili e interpretabili è quindi necessario ricorrere a metodi statistici appropriati, capaci di modellare la complessa struttura latente e l'elevata dimensionalità che caratterizzano i dati NIRS.

In particolare, questi dati si presentano in contesti ad alta dimensionalità, in cui il numero di variabili supera di gran lunga quello delle osservazioni. In situazioni di questo tipo, i metodi di analisi fattoriale assumono un ruolo centrale, in quanto permettono di ridurre la dimensionalità del problema estraendo un numero inferiore di componenti latenti (*fattori*) che catturano la struttura di dipendenza tra le variabili osservate. Tale riduzione facilita l'interpretazione dei dati e, al contempo, può migliorare la capacità predittiva dei modelli statistici applicati a questo tipo di informazioni.

L'obiettivo principale del presente lavoro è applicare e valutare un modello avanzato di analisi fattoriale bayesiana — lo *Sparse Bayesian Infinite Factor Model* — per l'analisi di dati NIRS. Si tratta di un approccio che unisce la flessibilità dell'inferenza bayesiana a una struttura capace di adattare automaticamente la dimensionalità dei fattori latenti ai dati stessi. Il metodo, proposto da Bhattacharya and Dunson (2011), rappresenta un importante sviluppo rispetto ai modelli di analisi fattoriale tradizionali e costituisce un contributo rilevante nell'ambito della modellazione statistica ad alta dimensionalità.

Un elemento di novità di questo studio risiede nell'applicazione del modello a un insieme di dati NIRS non precedentemente analizzato con tale metodologia. L'impiego dello *Sparse Bayesian Infinite Factor Model* in questo contesto consente di valutare le sue potenzialità nel trattare dati caratterizzati da elevato numero di variabili e strutture di correlazione complesse, con l'obiettivo di verificare se l'adozione di un approccio fattoriale bayesiano avanzato possa condurre a prestazioni predittive competitive o superiori rispetto a quelle ottenute mediante tecniche statistiche già affermate.

A tal fine, nella parte applicativa del lavoro, vengono confrontati diversi metodi di analisi, sia di impostazione classica sia di natura bayesiana, tut-

ti particolarmente adatti a contesti ad alta dimensionalità. Tale confronto permette di inquadrare criticamente i risultati ottenuti con il modello proposto e di valutarne la capacità di adattarsi a dati reali complessi rispetto ad approcci alternativi.

La tesi è articolata in tre capitoli principali, che guidano il lettore attraverso un percorso logico e graduale, dal contesto teorico generale fino all'applicazione empirica del modello proposto.

Nel Capitolo 1 viene introdotto il contesto concettuale e teorico necessario per comprendere l'approccio adottato. La prima parte del capitolo descrive l'analisi fattoriale classica, illustrandone gli obiettivi, le assunzioni e le caratteristiche fondamentali. Segue una sezione dedicata ai concetti di base della statistica bayesiana, con particolare attenzione ai principi dell'inferenza bayesiana e alle principali tecniche computazionali, inclusi i metodi di tipo *Markov Chain Monte Carlo* (MCMC) quali il *Gibbs sampler*. L'ultima parte del capitolo è dedicata all'analisi fattoriale bayesiana, che viene presentata nei suoi aspetti teorici, negli obiettivi e nei vantaggi rispetto all'approccio classico, fornendo così le basi concettuali necessarie per la comprensione dei modelli più avanzati presentati successivamente.

Il Capitolo 2 è dedicato allo *Sparse Bayesian Infinite Factor Model*, modello di analisi fattoriale bayesiana introdotto da Bhattacharya and Dunson (2011). Dopo una presentazione dettagliata della struttura del modello e dell'algoritmo di stima basato su *Gibbs sampling* con troncatura adattiva, viene illustrato in modo approfondito il funzionamento del procedimento inferenziale, che sarà successivamente implementato manualmente nel linguaggio R. La sezione conclusiva del capitolo illustra i risultati delle simulazioni condotte per verificare la riproducibilità delle prestazioni riportate in letteratura, confrontando le metriche di accuratezza e predizione ottenute con quelle di riferimento presentate dagli autori originali. Vengono inoltre proposte ulteriori analisi diagnostiche atte a valutare la stabilità e la coerenza del modello implementato.

Il Capitolo 3 riguarda l'applicazione del metodo ai dati reali di tipo spettroscopia del vicino infrarosso. Dopo una descrizione generale della natura di tali dati e delle operazioni di pre-processing, vengono presentate due applicazioni empiriche distinte, illustrate in modo approfondito sia nel loro funzionamento sia nelle motivazioni che ne giustificano l'impiego. Ciascuna applicazione è analizzata con attenzione, evidenziando gli aspetti metodologici e interpretativi più rilevanti ai fini della valutazione complessiva del modello proposto.

La prima applicazione riguarda i dati NIRS relativi all'impasto dei biscotti, già analizzati da da molti, tra i quali Osborne et al. (1984), Brown et al. (2001) e Bernardo et al. (2003). In questo caso, le variabili da predire corrispondono alle concentrazioni dei principali composti chimici che costituiscono l'impasto. Tali risultati, ampiamente documentati in letteratura, vengono impiegati come termine di confronto per valutare le prestazioni del modello implementato sia in termini di accuratezza predittiva, sia rispetto alla coerenza interpretativa delle lunghezze d'onda più informative nell'ambito della spettroscopia NIRS.

La seconda applicazione prende invece in esame i dati NIRS sul grano, analizzati nel contesto degli studi di Rincent et al. (2018). In questo caso, la variabile da predire è una misura fenotipica osservabile, strettamente legata al processo di miglioramento genetico delle varietà di grano. I dati utilizzati sono stati direttamente raccolti e prodotti dall'INRAE (Istituto Nazionale di Ricerca per l'Agricoltura, l'Alimentazione e l'Ambiente), ente di ricerca francese attivo nello studio e nello sviluppo di metodologie innovative in ambito agronomico. La spettroscopia nel vicino infrarosso viene qui utilizzata come possibile strumento alternativo o complementare alle analisi basate su marcatori genomici, con l'obiettivo di stimare la resa o altre caratteristiche di interesse agronomico. Il confronto dei risultati ottenuti mediante lo *Sparse Bayesian Infinite Factor Model* con quelli derivanti da diversi metodi predittivi — in particolare la *Ridge Regression*, già adottata da Rincent et al. (2018) — consente di valutare in modo sistematico le potenzialità e i limiti del modello proposto nell'ambito dell'analisi di dati NIRS complessi e ad alta dimensionalità.

Complessivamente, il lavoro mira a mostrare come i modelli di analisi fattoriale bayesiana avanzata possano essere efficacemente impiegati per l'analisi di dati NIRS in diversi contesti applicativi, con particolare riferimento ai settori alimentare e agricolo. L'obiettivo finale è valutare le potenzialità di tali strumenti nell'estrazione di informazioni predittive e interpretative da dati complessi e ad alta dimensionalità, evidenziandone i vantaggi rispetto ai metodi tradizionalmente impiegati.

Capitolo 1

Introduzione all'Analisi Fattoriale Bayesiana

Il capitolo introduce il percorso teorico e metodologico che conduce dai modelli fattoriali classici alla loro estensione in chiave bayesiana, ponendo le basi per modelli più complessi. Nella prima sezione vengono presentati i fondamenti dell'analisi fattoriale classica, necessari per comprendere la struttura dei modelli e la successiva evoluzione verso contesti più complessi. La seconda sezione illustra i principi della statistica bayesiana, chiarendo il ruolo delle distribuzioni a priori, della verosimiglianza e della distribuzione a posteriori, e motivando l'impiego di strumenti computazionali. Segue la presentazione delle tecniche Markov Chain Monte Carlo (MCMC), indispensabili per la stima della distribuzione a posteriori e per l'implementazione pratica dell'inferenza bayesiana. Infine, la quarta sezione combina questi elementi introducendo l'analisi fattoriale bayesiana, che consente di stimare fattori latenti e carichi fattoriali incorporando informazioni a priori e gestendo l'incertezza in modo naturale.

I L'Analisi Fattoriale

L'analisi fattoriale è una tecnica statistica multivariata concepita con lo scopo di ridurre la dimensionalità di un insieme di variabili osservate, cercando di individuare un numero più contenuto di variabili latenti, dette *fattori*, capaci di spiegare le correlazioni presenti tra le variabili originarie. Questo strumento metodologico si rivela particolarmente utile quando si intende semplificare modelli complessi e aumentare l'interpretabilità dei dati, senza rinunciare a una rappresentazione coerente delle relazioni sottostanti. In termini generali, i principali obiettivi dell'analisi fattoriale possono essere ricondotti a due prospettive complementari: da un lato, la necessità di sintesi dell'informazione, attraverso la condensazione di numerosi indicatori in un numero ridotto di costrutti; dall'altro, la possibilità di individuare strutture latenti, ovvero costrutti non direttamente osservabili, che permettano di mettere in evidenza relazioni nascoste tra le variabili e di migliorare, al tempo stesso, la capacità predittiva dei modelli.

I.1 Il modello

Sia

$$X = (X_1, \dots, X_p)^\top \in \mathbb{R}^p$$

un vettore di variabili osservate, caratterizzato da media

$$\mu = \mathbb{E}[X] \in \mathbb{R}^p$$

e matrice di covarianza

$$\Sigma = \text{Cov}(X) \in \mathbb{R}^{p \times p}.$$

Il modello fattoriale postula che ciascuna variabile osservata possa essere espressa come combinazione lineare di un numero ridotto di fattori comuni più una componente di errore specifico:

$$X = \mu + \Lambda f + \varepsilon, \quad (\text{I.1})$$

dove $\Lambda \in \mathbb{R}^{p \times k}$ è la matrice dei carichi fattoriali, $f \in \mathbb{R}^k$ il vettore dei fattori comuni, e $\varepsilon \in \mathbb{R}^p$ il vettore degli errori specifici. Le colonne della matrice Λ quantifica l'intensità con cui ciascun fattore influenza le variabili osservate, mentre gli errori ε rappresentano la parte di variabilità non condivisa.

Per garantire identificabilità e coerenza, il modello impone alcune condizioni. I fattori e gli errori hanno media nulla:

$$\mathbb{E}[f] = 0, \quad \mathbb{E}[\varepsilon] = 0.$$

I fattori sono standardizzati e incorrelati, ovvero

$$\text{Cov}(f) = I_k,$$

mentre errori e fattori risultano indipendenti:

$$\text{Cov}(f, \varepsilon) = 0.$$

Infine, gli errori specifici sono incorrelati tra loro e hanno varianze individuali:

$$\text{Cov}(\varepsilon) = \Psi = \text{diag}(\psi_1, \dots, \psi_p).$$

Sotto queste ipotesi, la matrice di covarianza delle variabili osservate assume la forma

$$\Sigma = \Lambda \Lambda^\top + \Psi, \quad (\text{I.2})$$

che distingue chiaramente la quota di varianza spiegata dai fattori comuni da quella specifica e indipendente.

La stima dei parametri del modello, ossia dei carichi fattoriali Λ e delle varianze specifiche Ψ , può avvenire attraverso diverse procedure. L'approccio più diffuso è quello della massima verosimiglianza (ML). Assumendo normalità multivariata,

$$X \sim \mathcal{N}_p(\mu, \Sigma),$$

la log-verosimiglianza del campione

$$\ell(\Lambda, \Psi) \propto -\frac{n}{2} [\log |\Sigma| + \text{tr}(S\Sigma^{-1})],$$

dove S è la matrice di covarianza campionaria, viene massimizzata rispetto a (Λ, Ψ) . Non essendo disponibile una soluzione in forma chiusa, si ricorre a metodi iterativi di massimizzazione o di Expectation–Maximization.

Un caso particolare è rappresentato dall’analisi delle componenti principali (PCA), che pur non essendo un’analisi fattoriale in senso stretto può essere vista come un modello in cui non si distingue tra varianza comune e specifica ($\Psi = 0$). In tal caso la decomposizione di Σ coincide con la sua decomposizione spettrale, fornendo una stima iniziale dei carichi fattoriali.

I.II Aspetti operativi

Un aspetto cruciale dell’analisi fattoriale riguarda la scelta del numero di fattori k . Poiché la stima per massima verosimiglianza tende a favorire soluzioni ad alta dimensionalità, la determinazione di k richiede criteri esterni al modello. Tra i più utilizzati si annoverano regole empiriche basate sugli autovalori della matrice di correlazione, soglie convenzionali di varianza spiegata, metodi grafici come lo scree plot, nonché procedure più rigorose fondate su test di bontà di adattamento. Nessun criterio è tuttavia univoco, e la decisione finale costituisce un compromesso tra considerazioni teoriche, proprietà empiriche e obiettivi applicativi.

Una volta fissato il numero di fattori, occorre affrontare il problema dell’interpretabilità. La soluzione fattoriale iniziale non è infatti unica, in quanto esistono infinite soluzioni equivalenti ottenibili tramite trasformazioni lineari che preservano l’adattamento del modello. Per questo motivo si ricorre a procedure di rotazione, volte a semplificare la struttura dei carichi fattoriali. Indicando con $\hat{\Lambda}$ la stima iniziale, la matrice ruotata assume la forma

$$\Lambda^* = \hat{\Lambda}T,$$

dove T è una matrice di trasformazione. Se T è ortogonale, i fattori rimangono incorrelati; le tecniche più note in questo contesto sono *Varimax*, che massimizza la varianza dei carichi quadratici, e *Quartimax*, che semplifica il contributo delle variabili. Se invece T non è vincolata a essere ortogonale, la rotazione è detta obliqua e consente di modellare fattori tra loro correlati, soluzione spesso più realistica in ambiti applicativi quali le scienze sociali.

Stabilita la struttura fattoriale, diventa rilevante la stima dei punteggi fattoriali, ossia delle realizzazioni latenti associate a ciascun individuo. Poiché i fattori non sono osservabili, essi devono essere ricostruiti indirettamente. Due metodi sono prevalenti: il *metodo della regressione*, che stima i punteggi come proiezioni lineari delle osservazioni massimizzando la correlazione con i fattori veri,

$$\hat{f}_i = \Lambda^\top \Sigma^{-1} x_i,$$

e il *metodo di Bartlett*, che invece produce stime non correlate con gli errori specifici, minimizzandone l'influenza residua,

$$\hat{f}_i = (\Lambda^\top \Psi^{-1} \Lambda)^{-1} \Lambda^\top \Psi^{-1} x_i.$$

Il primo approccio è orientato a finalità predittive, mentre il secondo privilegia un'interpretazione più accurata della struttura latente.

II Fondamenti di statistica bayesiana

Dopo aver introdotto il quadro teorico dell'analisi fattoriale classica, è utile considerare come un approccio bayesiano possa offrire una prospettiva alternativa all'inferenza sui parametri del modello. La statistica bayesiana fornisce infatti un quadro coerente e generale per affrontare l'incertezza, in cui la probabilità non si limita a descrivere la variabilità dei dati osservati, ma diviene anche uno strumento per quantificare il grado di credenza associato ai parametri incogniti. In questo contesto, le quantità sconosciute del modello — come le componenti fattoriali o le varianze specifiche — vengono trattate come variabili casuali dotate di una distribuzione a priori, che riflette l'informazione disponibile prima dell'osservazione dei dati. I principi e gli strumenti alla base di tale impostazione, che verranno introdotti nella sezione seguente, forniscono le basi teoriche necessarie per la successiva estensione dell'analisi fattoriale in chiave bayesiana.

Il fulcro dell'approccio bayesiano è il teorema di Bayes, che consente di combinare l'informazione a priori sui parametri $\theta \in \Theta \subseteq \mathbb{R}^d$ con i dati osservati $y \in \mathbb{R}^n$, ottenendo una distribuzione a posteriori che integra entrambe le fonti di informazione. In forma generale, esso si scrive come:

$$p(\theta \mid y) = \frac{p(y \mid \theta) p(\theta)}{p(y)}. \quad (\text{II.1})$$

- $p(\theta)$: distribuzione a priori (prior), che rappresenta le credenze iniziali sui parametri.
- $p(y \mid \theta)$: verosimiglianza, ovvero la probabilità di osservare i dati condizionatamente a un dato valore dei parametri.
- $p(y)$: evidenza o fattore di normalizzazione, calcolato come

$$p(y) = \int_{\Theta} p(y \mid \theta) p(\theta) d\theta,$$

il quale assicura che $p(\theta \mid y)$ sia una distribuzione di probabilità ben definita.

- $p(\theta \mid y)$: distribuzione a posteriori, che rappresenta la conoscenza aggiornata sui parametri dopo l'osservazione dei dati.

Dal punto di vista operativo, tuttavia, il calcolo esplicito del termine di normalizzazione $p(y)$ risulta spesso proibitivo, in quanto richiede l'integrazione sull'intero spazio dei parametri. Per questo motivo, in molte applicazioni pratiche, si preferisce lavorare con la forma non normalizzata della distribuzione a posteriori, ossia:

$$p(\theta | y) \propto p(y | \theta) p(\theta), \quad (\text{II.2})$$

dove il simbolo \propto indica proporzionalità fino a una costante di normalizzazione. Questa formulazione semplifica notevolmente la trattazione teorica e computazionale, soprattutto nell'ambito dei metodi numerici di inferenza, come le tecniche Monte Carlo.

Un fondamento teorico cruciale è il teorema di de Finetti. Esso afferma che una sequenza di osservazioni (Y_1, Y_2, \dots) scambiabili — cioè tale che qualsiasi permutazione degli indici non ne alteri la distribuzione congiunta — può essere rappresentata come una mistura di variabili indipendenti e identicamente distribuite, condizionate a un parametro latente θ , rispetto a una misura di probabilità P definita su Θ :

$$p(Y_1, \dots, Y_n) = \int_{\Theta} \left(\prod_{i=1}^n p(Y_i | \theta) \right) dP(\theta).$$

In altri termini, la distribuzione congiunta di una sequenza scambiabile può essere vista come una media (o mistura) delle distribuzioni condizionate $p(X_i | \theta)$, pesate secondo la distribuzione a priori $P(\theta)$.

Questo risultato giustifica la modellizzazione bayesiana: trattare i parametri come variabili casuali non è una scelta arbitraria, ma segue direttamente dalla proprietà di scambiabilità, garantendo coerenza logica al paradigma bayesiano.

È importante distinguere tra *scambiabilità finita* e *scambiabilità infinita*. Nel primo caso, la proprietà vale soltanto per un numero limitato di osservazioni, mentre nel secondo riguarda l'intera sequenza infinita $(Y_i)_{i \geq 1}$. Il teorema di de Finetti si applica a sequenze infinitamente scambiabili, che ammettono sempre una rappresentazione come mistura di distribuzioni i.i.d. Nel contesto dell'inferenza bayesiana, si assume tipicamente l'ipotesi di scambiabilità infinita, che consente di estendere il modello coerentemente a campioni di dimensione arbitraria.

Il confronto tra approccio frequentista e approccio bayesiano mette in luce differenze concettuali rilevanti e il ruolo delle distribuzioni a priori nell'inferenza. Un primo elemento distintivo tra i diversi approcci inferenziali riguarda il modo in cui vengono trattati i parametri del modello. Nel paradigma frequentista, essi sono concepiti come quantità fisse ma ignote. L'incertezza è attribuita esclusivamente ai dati, interpretati come realizzazioni di variabili casuali. L'inferenza consiste dunque nello studio della distribuzione campionaria, da cui si ricavano stimatori e procedure di test:

$$\hat{\theta} = g(Y_1, \dots, Y_n), \quad Y_i \sim F_{\theta}.$$

L'impostazione bayesiana adotta una prospettiva differente: i parametri vengono trattati come variabili casuali a tutti gli effetti. L'inferenza si basa sul calcolo della distribuzione a posteriori $p(\theta | y)$, che combina l'informazione proveniente dai dati osservati con le credenze pregresse espresse dalla distribuzione a priori. In termini formali, la probabilità che θ appartenga a un insieme A , dato il campione osservato y , è definita come

$$\mathbb{P}(\theta \in A | y) = \int_A p(\theta | y) d\theta.$$

Questa impostazione consente di rappresentare in modo coerente l'incertezza residua sui parametri, integrando dati e conoscenze pregresse all'interno di un unico schema probabilistico.

Un aspetto cruciale dell'approccio bayesiano riguarda proprio la scelta della distribuzione a priori $p(\theta)$. Essa può assumere forme diverse a seconda della disponibilità di informazione: una *prior informativa* incorpora conoscenze pregresse o evidenze esterne sul parametro, mentre una *prior debolmente informativa* o *non informativa* è costruita per influenzare il meno possibile l'inferenza, lasciando che siano i dati a determinare la forma della distribuzione a posteriori. Dal punto di vista operativo, un ruolo centrale è svolto dalle cosiddette *prior coniugate*. Esse non costituiscono una categoria autonoma, ma rappresentano scelte particolarmente convenienti per ragioni analitiche: la loro forma garantisce infatti che la distribuzione a posteriori appartenga alla stessa famiglia della prior, semplificando notevolmente i calcoli e consentendo di esprimere l'inferenza in forma chiusa. In questo modo, l'approccio bayesiano mantiene un equilibrio tra rigore teorico e praticità applicativa.

L'interpretazione bayesiana dell'inferenza risulta quindi molto naturale: la probabilità di un'ipotesi non è più solo un concetto asintotico legato alla frequenza relativa, ma una misura coerente del grado di credenza che un ricercatore attribuisce ad una certa affermazione, data l'evidenza osservata.

II.I Stima puntuale, intervalli di credibilità e predizione

La distribuzione a posteriori $p(\theta | y)$ fornisce una descrizione completa dell'incertezza relativa al parametro θ , integrando informazioni a priori e dati osservati. Tuttavia, in molte applicazioni pratiche è utile sintetizzare questa distribuzione mediante stime puntuali e intervalli di credibilità, che consentono decisioni e inferenze più agevoli.

Nel paradigma bayesiano la stima puntuale di un parametro θ si ottiene minimizzando la perdita attesa, cioè il valore medio della funzione di perdita $L(\theta, \hat{\theta})$ rispetto alla distribuzione a posteriori $p(\theta | y)$.

Nel seguito adotteremo la perdita quadratica, per la quale la stima ottimale è la *media a posteriori*:

$$\hat{\theta}_{PM} = \mathbb{E}[\theta | y] = \int_{\Theta} \theta p(\theta | y) d\theta.$$

Questa scelta minimizza l'errore quadratico medio e rappresenta un riassunto naturale della distribuzione a posteriori.

Oltre alla stima puntuale, la statistica bayesiana consente di costruire *intervalli di credibilità*, che rappresentano in modo naturale l'incertezza attorno al parametro:

$$\mathbb{P}(a \leq \theta \leq b \mid y) = 1 - \alpha, \quad (\text{II.3})$$

dove $1 - \alpha$ è il livello di credibilità desiderato.

Gli intervalli di credibilità possono essere calcolati in modi differenti, ciascuno con interpretazioni e proprietà proprie. Tra le diverse scelte, gli *intervalli di massima densità a posteriori* (Highest Posterior Density, HPD) sono frequentemente utilizzati in applicazioni pratiche poiché concentrano la massa di probabilità nella regione di più alta densità. Si definisce come l'intervallo più stretto che contiene $1 - \alpha$ della probabilità a posteriori:

$$[a, b]_{HPD} = \arg \min_{[a, b]} \{b - a : \int_a^b p(\theta \mid y) d\theta = 1 - \alpha\}.$$

Concettualmente, ogni punto all'interno dell'intervallo HPD ha densità maggiore o uguale a ogni punto esterno all'intervallo. Questo lo rende particolarmente utile per distribuzioni asimmetriche o multimodali, dove gli intervalli centrati potrebbero non rappresentare correttamente la probabilità più concentrata.

In sintesi, la stima puntuale e gli intervalli di credibilità forniscono strumenti complementari: la stima puntuale sintetizza l'informazione centrale della distribuzione a posteriori, mentre gli intervalli di credibilità quantificano formalmente l'incertezza e delimitano gli insiemi di valori più probabili per il parametro θ .

Oltre a fornire stime dei parametri, l'approccio bayesiano consente di formulare previsioni sui dati futuri, integrando l'incertezza residua dei parametri stimati. Siano: y l'insieme dei dati osservati, y^* una nuova osservazione da predire, $\theta \in \Theta$ il vettore dei parametri del modello. la distribuzione predittiva a posteriori ha la seguente formula:

$$p(y^* \mid y) = \int_{\Theta} p(y^* \mid \theta) p(\theta \mid y) d\theta, \quad (\text{II.4})$$

dove:

- $p(\theta \mid y)$ è la distribuzione a posteriori dei parametri, ottenuta tramite il teorema di Bayes (II.1) con $p(y)$ evidenza dei dati osservati;
- $p(y^* \mid \theta)$ rappresenta la distribuzione condizionata di una nuova osservazione condizionatamente al valore del parametro θ .

Concettualmente, la predizione bayesiana si basa sull'integrazione della distribuzione condizionata della nuova osservazione rispetto alla distribuzione a posteriori dei parametri. Ciò implica che, a differenza degli approcci frequentisti tradizionali, non si ricorre a un singolo stimatore puntuale di

θ , bensì si media su tutte le possibili realizzazioni compatibili con i dati osservati.

La costruzione della distribuzione predittiva procede in maniera naturale in tre passaggi concettuali. In primo luogo, si stima la distribuzione a posteriori $p(\theta | y)$, che integra le informazioni a priori $p(\theta)$ con l'evidenza dei dati $p(y | \theta)$. Successivamente, per ogni possibile valore di θ , si calcola la probabilità di osservare una nuova realizzazione y^* tramite $p(y^* | \theta)$. Infine, l'integrazione rispetto alla distribuzione a posteriori dei parametri produce la distribuzione predittiva finale: (II.4) tenendo così conto di tutte le incertezze associate ai parametri.

Dal punto di vista interpretativo, la distribuzione predittiva bayesiana risulta conservativa, in quanto incorpora simultaneamente l'incertezza sui parametri e la variabilità intrinseca dei dati futuri. Essa consente di stimare direttamente probabilità di eventi futuri, ad esempio:

$$\mathbb{P}(y^* > c | y) = \int_{y^* > c} p(y^* | y) dy^*,$$

e fornisce un quadro coerente per decisioni statistiche e predizioni, integrando conoscenze pregresse e informazioni osservate. In sintesi, la predizione bayesiana rappresenta un'estensione naturale dell'inferenza bayesiana, estendendo la quantificazione dell'incertezza dai parametri ai dati futuri in modo completo e probabilisticamente coerente.

III Metodi computazionali di Inferenza Bayesiana

III.I Panoramica sulle tecniche Markov Chain Monte Carlo (MCMC)

Nell'ambito della statistica bayesiana, l'inferenza sui parametri di un modello si basa sulla distribuzione a posteriori (II.2) dove $p(y | \theta)$ rappresenta la verosimiglianza e $p(\theta)$ la distribuzione a priori.

Nella maggior parte dei casi pratici, la distribuzione a posteriori non ammette una forma chiusa e il calcolo analitico di quantità di interesse, come medie, varianze o probabilità marginali, è generalmente impossibile. Pertanto, diventa necessario ricorrere a metodi numerici che consentano di approssimare la distribuzione a posteriori attraverso campioni simulati, piuttosto che calcolarla in modo esatto.

In questo contesto si collocano i metodi Monte Carlo e, in particolare, le tecniche *Markov Chain Monte Carlo* (MCMC). Le tecniche MCMC rappresentano oggi uno strumento imprescindibile per l'inferenza bayesiana in presenza di distribuzioni a posteriori complesse e prive di forma chiusa. L'idea chiave consiste nel generare una sequenza di campioni

$$\{\theta^{(1)}, \theta^{(2)}, \dots, \theta^{(T)}\}$$

la cui distribuzione empirica, successivamente alla fase iniziale di burn-in — ossia il periodo iniziale della catena Markoviana durante il quale la dipendenza dalle condizioni di partenza non è ancora svanita — può essere considerata un'approssimazione di $p(\theta \mid y)$, fornendo così una rappresentazione empirica della legge a posteriori.

Dal punto di vista teorico, una catena MCMC efficace deve soddisfare alcune proprietà fondamentali che ne garantiscono il corretto funzionamento. Tra queste, l'irriducibilità assicura che la catena possa raggiungere qualsiasi regione dello spazio dei parametri con probabilità positiva, permettendo un'esplorazione completa della distribuzione. L'aperiodicità evita che la catena rimanga intrappolata in cicli deterministici, garantendo la possibilità di convergere verso una distribuzione stazionaria. Infine, l'ergodicità rappresenta la condizione essenziale affinché la distribuzione della catena converga verso quella stazionaria indipendentemente dal punto di partenza. Grazie a tali proprietà, il teorema ergodico assicura che le medie campionarie delle funzioni integrabili $h(\theta)$ convergano quasi sicuramente alle corrispondenti medie a posteriori:

$$\frac{1}{T} \sum_{t=1}^T h(\theta^{(t)}) \xrightarrow{a.s.} \mathbb{E}[h(\theta) \mid y],$$

fornendo un solido fondamento teorico per l'uso dei metodi MCMC nel calcolo di quantità bayesiane complesse.

Tra gli algoritmi più diffusi si annoverano Metropolis-Hastings e Gibbs sampler. Il primo si basa su un meccanismo di proposta-accettazione e permette di campionare da distribuzioni arbitrarie, a condizione che la funzione di densità sia nota fino a una costante di normalizzazione. Il Gibbs sampler, invece, risulta particolarmente efficace quando le distribuzioni condizionali complete sono note e facilmente campionabili, consentendo un aggiornamento iterativo dei parametri uno alla volta o a blocchi. In entrambi i casi, l'efficacia dei metodi dipende dalla capacità della catena di soddisfare le proprietà teoriche di convergenza e di esplorare in maniera efficiente lo spazio dei parametri, fornendo così uno strumento flessibile per l'inferenza bayesiana anche in contesti complessi.

L'applicazione pratica degli algoritmi MCMC richiede un'attenta valutazione della qualità della catena. Un campionamento affidabile deve garantire un buon *mixing*, ossia la capacità di esplorare efficacemente lo spazio dei parametri, e un basso livello di autocorrelazione tra campioni consecutivi. Le iterazioni iniziali, infatti, risentono della scelta dei valori iniziali e riflettono una fase transitoria non rappresentativa della distribuzione stazionaria. Per questo motivo è prassi comune scartare una porzione iniziale di campioni, nota come *burn-in* o *warm-up*, considerando solo la parte della catena prossima alla distribuzione target.

La valutazione della convergenza rappresenta un passaggio cruciale. Strumenti come i *traceplot* consentono di individuare trend residui o mancanza di stabilità, mentre statistiche di confronto tra catene, come l'indice di Gelman–Rubin \hat{R} , segnalano potenziali problemi quando differiscono si-

gnificativamente da 1. L'analisi delle autocorrelazioni permette inoltre di valutare la capacità della catena di esplorare lo spazio dei parametri in modo efficiente. Poiché i campioni MCMC non sono indipendenti, il numero effettivo di campioni indipendenti (*effective sample size*, ESS) risulta inferiore al numero totale di iterazioni simulate; l'ESS fornisce un'indicazione quantitativa della qualità del campionamento e della rappresentazione della distribuzione a posteriori, con valori elevati che indicano maggiore efficienza e affidabilità delle inferenze.

III.II Il Gibbs sampler

Il *Gibbs sampler* è un algoritmo di campionamento stocastico. L'idea di base consiste nel campionare iterativamente ciascuna variabile condizionatamente ai valori correnti delle altre, utilizzando le distribuzioni condizionali complete, che spesso risultano di più agevole campionamento rispetto alla distribuzione congiunta. Si consideri un vettore di parametri

$$\theta = (\theta_1, \theta_2, \dots, \theta_d).$$

La distribuzione a posteriori congiunta $p(\theta | y)$ è spesso difficile da campionare direttamente. Tuttavia, se si conoscono le distribuzioni condizionali complete di ciascun parametro

$$p(\theta_j | \theta_{-j}, y), \quad j = 1, \dots, d,$$

dove θ_{-j} indica tutti i parametri eccetto θ_j , è possibile generare campioni iterando il seguente schema:

1. **Inizializzazione:** scegliere un valore iniziale

$$\theta^{(0)} = (\theta_1^{(0)}, \dots, \theta_d^{(0)}).$$

2. **Aggiornamento ciclico:** per ogni iterazione $t = 0, 1, 2, \dots$, aggiornare sequenzialmente ciascun parametro:

$$\begin{aligned} \theta_1^{(t+1)} &\sim p(\theta_1 | \theta_2^{(t)}, \dots, \theta_d^{(t)}, y), \\ \theta_2^{(t+1)} &\sim p(\theta_2 | \theta_1^{(t+1)}, \theta_3^{(t)}, \dots, \theta_d^{(t)}, y), \\ &\vdots \\ \theta_d^{(t+1)} &\sim p(\theta_d | \theta_1^{(t+1)}, \dots, \theta_{d-1}^{(t+1)}, y). \end{aligned}$$

Questo processo produce una catena di Markov $\{\theta^{(t)}\}$ che, sotto condizioni di irriducibilità e aperiodicità, converge in distribuzione alla posteriori:

$$\lim_{t \rightarrow \infty} \mathcal{L}(\theta^{(t)}) = p(\theta | y),$$

dove $\mathcal{L}(\theta^{(t)})$ indica la legge del vettore θ alla t -esima iterazione.

Dal punto di vista pratico, dopo un numero sufficiente di iterazioni, i campioni $\theta^{(t)}$ possono essere trattati come campioni dalla distribuzione a posteriori, permettendo di calcolare stime empiriche e intervalli di credibilità:

$$\mathbb{E}[\theta_j | y] \approx \frac{1}{T} \sum_{t=1}^T \theta_j^{(t)}, \quad \text{Var}[\theta_j | y] \approx \frac{1}{T-1} \sum_{t=1}^T \left(\theta_j^{(t)} - \bar{\theta}_j \right)^2.$$

Il Gibbs sampler si distingue per la sua semplicità concettuale e per l'efficienza che può raggiungere quando le distribuzioni condizionali complete sono note e facilmente campionabili. A differenza dell'algoritmo Metropolis-Hastings, non prevede un passo di accettazione o rifiuto, il che lo rende particolarmente efficiente da implementare. Questi aspetti ne spiegano la popolarità, ma non mancano alcune limitazioni. In presenza di una forte correlazione tra parametri, infatti, la catena può presentare una convergenza lenta, compromettendo l'efficienza complessiva del campionamento. Inoltre, quando le condizionali non hanno una forma analitica nota o non sono direttamente campionabili, il metodo diventa di difficile applicazione.

Per ovviare a tali criticità, sono state sviluppate diverse strategie di estensione. Una prima possibilità è il *blocking*, che prevede l'aggiornamento congiunto di gruppi di parametri correlati e riduce i problemi legati alla correlazione. Un approccio alternativo è la *data augmentation*, basata sull'introduzione di variabili latenti che semplificano le distribuzioni condizionali e ne facilitano il campionamento. In conclusione il Gibbs sampler si conferma uno strumento flessibile e adattabile, capace di affrontare un'ampia varietà di modelli e di scenari applicativi.

IV Analisi Fattoriale Bayesiana

L'analisi fattoriale classica costituisce uno strumento centrale per ridurre la dimensionalità dei dati e per mettere in luce strutture latenti che sottendono un insieme di variabili osservate. Nonostante la sua importanza, l'approccio tradizionale presenta alcune criticità che ne limitano l'applicabilità. In primo luogo, la matrice dei factor-loadings Λ risulta identificata solo fino a una rotazione ortogonale arbitraria: questa ambiguità impone di ricorrere a criteri post-hoc per scegliere una soluzione interpretabile, introducendo spesso un margine di soggettività. In secondo luogo, i metodi classici non offrono un meccanismo formale per incorporare informazioni preliminari o vincoli derivanti da conoscenze pregresse, da esperti o da studi precedenti.

L'approccio bayesiano (Rowe (2000)) affronta queste limitazioni adottando una prospettiva più generale, nella quale tutti i parametri del modello — inclusi i punteggi fattoriali F e il numero di fattori k — vengono trattati come variabili aleatorie dotate di distribuzioni a priori. In questo quadro, le stime ottenute risultano univoche e accompagnate da una valutazione quantitativa dell'incertezza, permettendo al tempo stesso di integrare in modo coerente conoscenze pregresse e di gestire modelli complessi senza ricorre-

re a criteri soggettivi. Ne deriva una cornice metodologica più flessibile e rigorosa, capace di superare le debolezze intrinseche dei metodi classici.

IV.I Il modello

Il modello di Analisi Fattoriale Bayesiana mantiene la stessa struttura del modello fattoriale classico (I.1), ma viene riformulato in un contesto probabilistico che consente di assegnare distribuzioni a priori ai parametri e di condurre inferenza tramite la distribuzione a posteriori.

Si considerino n osservazioni indipendenti del vettore aleatorio

$$x_i \in \mathbb{R}^p, \quad i = 1, \dots, n,$$

ciascuna delle quali è rappresentata da

$$x_i = \mu + \Lambda f_i + \varepsilon_i, \quad (\text{IV.1})$$

dove $\mu \in \mathbb{R}^p$ è il vettore delle medie, $\Lambda \in \mathbb{R}^{p \times k}$ la matrice dei carichi fattoriali, $f_i \in \mathbb{R}^k$ il vettore dei fattori latenti e $\varepsilon_i \in \mathbb{R}^p$ il vettore degli errori specifici. Come nel caso classico, si assume indipendenza tra fattori ed errori, con

$$f_i \sim \mathcal{N}(0, I_k), \quad \varepsilon_i \sim \mathcal{N}(0, \Psi),$$

dove I_k è la matrice identità $k \times k$ e Ψ è diagonale, contenente le varianze specifiche.

La differenza sostanziale rispetto al modello classico è che, nell'approccio bayesiano, anche i parametri (μ, Λ, Ψ) sono trattati come variabili aleatorie dotate di distribuzioni a priori, così che l'inferenza derivi dalla distribuzione a posteriori condizionata ai dati $X = (x_1, \dots, x_n)^\top$. Per costruire quest'ultima, occorre specificare la funzione di verosimiglianza e i prior. La verosimiglianza per una singola osservazione è data da

$$p(x_i \mid \mu, \Lambda, f_i, \Psi) = (2\pi)^{-p/2} |\Psi|^{-1/2} \exp \left\{ -\frac{1}{2} (x_i - \mu - \Lambda f_i)^\top \Psi^{-1} (x_i - \mu - \Lambda f_i) \right\},$$

e, considerando tutte le osservazioni,

$$\mathcal{L}(\mu, \Lambda, F, \Psi \mid X) = \prod_{i=1}^n p(x_i \mid \mu, \Lambda, f_i, \Psi). \quad (\text{IV.2})$$

La distribuzione a priori complessiva assume una struttura gerarchica,

$$p(\mu, \Lambda, F, \Psi, k) = p(\mu) p(\Lambda \mid \Psi, k) p(F \mid k) p(\Psi) p(k), \quad (\text{IV.3})$$

nella quale ciascun termine viene modellato in modo mirato. In particolare:

- **Prior per la media:**

$$\mu \sim \mathcal{N}_p(\mu_0, S_0).$$

Tipicamente si sceglie un prior *proprio* ma poco informativo: S_0 ‘grande’ (ad es. $S_0 = c \cdot \text{diag}(\hat{\sigma}_1^2, \dots, \hat{\sigma}_p^2)$ con $c \gg 1$) in modo da approssimare un comportamento non informativo mantenendo la proiezione di

una distribuzione propria. In applicazioni pratiche si può impostare μ_0 uguale alla media campionaria \bar{x} (scelta pragmatica che semplifica la stima) oppure un vettore nullo se le variabili sono state standardizzate.

- **Prior per la matrice dei carichi: (matrix-normal)**

$$\Lambda \mid \Psi, k \sim \text{MN}_{p \times k}(0, \Psi, G^{-1}),$$

cioè, in forma vettorializzata,

$$\text{vec}(\Lambda) \sim \mathcal{N}_{pk}(0, G^{-1} \otimes \Psi).$$

Interpretazione pratica: la covarianza di riga è Ψ (interagisce con la variabilità specifica delle variabili osservate), mentre la covarianza di colonna è G^{-1} (controlla la dipendenza fra colonne/fattori). Dal punto di vista dei singoli elementi si ha

$$\text{Var}(\lambda_{jh}) = (G^{-1})_{hh} \Psi_{jj}.$$

La scelta comune $G = g I_k$ (scalare isotropo) implica

$$\text{Var}(\lambda_{jh}) = \frac{\Psi_{jj}}{g},$$

cioè un unico iperparametro $g > 0$ che regola la *forza di shrinkage* globale dei carichi verso zero. Vantaggi di $G = g I_k$: (i) scarsa complessità iperparametrica (un solo parametro di scala), (ii) condizione di scambio (exchangeability) tra le colonne a priori, (iii) interazione naturale con Ψ che rende il prior sensibile alla scala delle variabili (utile quando si lavora con variabili non standardizzate).

- **Prior per i punteggi fattoriali:**

$$f_i \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}_k(0, I_k), \quad i = 1, \dots, n.$$

Assunzione standard che facilita la coniugazione e l'interpretabilità (fattori ortogonali con varianza unitaria).

- **Prior per la matrice degli errori specifici:**

$$\Psi \sim \mathcal{W}^{-1}(S, \nu),$$

cioè Inverse-Wishart con parametri (S, ν) . Spesso si prende S diagonale (prior indipendente sulle varianze specifiche) così da rispettare l'idea di *specificità* per ciascuna variabile. Le relazioni utili per collegare iperparametri e desiderata sulla media/varianza delle diagonal sono:

$$\mathbb{E}[\Psi] = \frac{S}{\nu - p - 1} \quad (\text{valida per } \nu > p + 1),$$

e, per un elemento diagonale j ,

$$\text{Var}(\Psi_{jj}) = \frac{2 S_{jj}^2}{(\nu - p - 1)^2(\nu - p - 3)} \quad (\text{richiede } \nu > p + 3).$$

Se desideriamo un prior con $E(\Psi_{jj}) = \bar{\psi}$ e $\text{Var}(\Psi_{jj}) = v_\psi$ (assumendo S diagonale con tutti i termini uguali per semplicità), si ottiene la scelta

$$S = (\nu - p - 1) \bar{\psi} I_p, \quad \nu = p + 3 + \frac{2\bar{\psi}^2}{v_\psi},$$

che garantisce la media e la varianza desiderate sulle diagonali; si noti la condizione necessaria $\nu > p + 3$ per varianza finita.

- **Prior sul numero di fattori k :** scelta discreta su $\{1, \dots, k_{\max}\}$. Si può prendere una distribuzione uniforme oppure una distribuzione debolmente decrescente (penalizzante i modelli più complessi).

Combinando verosimiglianza e prior tramite il teorema di Bayes (II.2), si ottiene la distribuzione a posteriori

$$p(\mu, \Lambda, F, \Psi, k \mid X) \propto p(k) p(\mu) p(\Lambda \mid \Psi, k) p(F \mid k) p(\Psi) \prod_{i=1}^n p(x_i \mid \mu, \Lambda, f_i, \Psi).$$

Da essa si ricava la marginale di interesse per il numero di fattori integrando sugli altri parametri:

$$p(k \mid X) = \int p(\mu, \Lambda, F, \Psi, k \mid X) d\mu d\Lambda dF d\Psi.$$

IV.II La stima: Gibbs sampler

Algoritmo 1 Gibbs sampler per l'Analisi Fattoriale Bayesiana

1: **Inizializzazione:** fissare valori iniziali per $\mu^{(0)}, \Lambda^{(0)}, F^{(0)}, \Psi^{(0)}$.

2: **for** $s = 1, \dots, N$ **do**

3: **Aggiornamento dei punteggi fattoriali** f_i :

$$f_i^{(s)} \sim \mathcal{N}_k \left(V^{-1} \Lambda^{(s-1)\top} \Psi^{(s-1)-1} (x_i - \mu^{(s-1)}), V^{-1} \right),$$

con $V = I_k + \Lambda^{(s-1)\top} \Psi^{(s-1)-1} \Lambda^{(s-1)}$.

4: **Aggiornamento dei loadings** λ_j :

$$\lambda_j^{(s)} \sim \mathcal{N}_k \left((G + \Psi_{jj}^{(s-1)-1} F^{(s)\top} F^{(s)})^{-1} \Psi_{jj}^{(s-1)-1} F^{(s)\top} (x_j^{(j)} - \mu_j^{(s-1)}), (G + \Psi_{jj}^{(s-1)-1} F^{(s)\top} F^{(s)})^{-1} \right).$$

5: **Aggiornamento delle varianze specifiche** ψ_j :

$$\psi_j^{(s)} \sim \text{Inv-Wishart} \left(\nu + n, S + \sum_{i=1}^n (x_{ij} - \mu_j^{(s-1)} - \lambda_j^{(s)\top} f_i^{(s)})^2 \right).$$

6: **Aggiornamento della media** μ :

$$\mu^{(s)} \sim \mathcal{N}_p \left((S_0^{-1} + n \Psi^{(s)-1})^{-1} (S_0^{-1} \mu_0 + \Psi^{(s)-1} \sum_{i=1}^n (x_i - \Lambda^{(s)} f_i^{(s)})), (S_0^{-1} + n \Psi^{(s)-1})^{-1} \right).$$

7: **end for**

8: **Output:** campioni $\{\mu^{(s)}, \Lambda^{(s)}, F^{(s)}, \Psi^{(s)}\}_{s=1}^N$, con scarto delle prime iterazioni (*burn-in*).

Le stime dei parametri si ottengono come medie a posteriori sui campioni successivi al burn-in:

$$\hat{\theta} = \frac{1}{N_{\text{post}}} \sum_{s=1}^{N_{\text{post}}} \theta^{(s)}.$$

Un criterio per la scelta del numero di fattori k è basato sulla probabilità a posteriori del modello. Lo schema operativo è il seguente:

Algoritmo 2 per la selezione di k

1: **for** $k = 1, \dots, k_{\text{max}}$ **do**

2: Eseguire l'Algoritmo 1 per ottenere campioni di $\mu, \Lambda, F, \Psi \mid k$.

3: Calcolare la probabilità a posteriori del modello:

$$p(k \mid \mu, \Lambda, F, \Psi, X) \propto p(k) p(\mu) p(\Lambda \mid \Psi, k) p(F \mid k) p(\Psi) p(X \mid \mu, \Lambda, F, \Psi, k).$$

4: **end for**

5: Selezionare il valore di k che massimizza $p(k \mid X)$.

Questo approccio consente di quantificare formalmente l'incertezza sulla dimensionalità del modello, eliminando la soggettività dei criteri ad hoc tradizionali (scree plot, varianza cumulata) e fornendo una procedura probabilisticamente coerente.

In sintesi l'Analisi Fattoriale Bayesiana offre un quadro unificato che integra conoscenze a priori, fornisce soluzioni univoche senza ricorrere a rotazioni arbitrarie e permette una valutazione formale dell'incertezza tramite la distribuzione a posteriori. Grazie all'impiego di algoritmi moderni di inferenza, questo approccio si dimostra efficiente anche in contesti complessi, estendendo l'analisi fattoriale classica verso un impianto più flessibile e probabilisticamente coerente, capace non solo di stimare i parametri, ma anche di supportare predizioni e inferenze sui dati futuri.

Capitolo 2

Lo sparse Bayesian infinite factor model e la sua implementazione

Il presente capitolo è dedicato alla descrizione del modello *sparse Bayesian infinite factor model*, con particolare attenzione ai principi teorici che ne guidano il funzionamento e alle modalità con cui esso viene implementato. Dopo aver illustrato le caratteristiche essenziali del modello e le tecniche computazionali necessarie per la sua applicazione, verrà condotto uno studio di simulazione sul codice sviluppato. Tale analisi ha l'obiettivo di verificarne l'adeguatezza operativa e di fornire evidenze preliminari sulla capacità del modello di restituire stime coerenti e affidabili nei contesti di interesse.

I Il Modello

Un importante avanzamento nell'ambito dell'Analisi Fattoriale Bayesiana è rappresentato dallo *Sparse Bayesian Infinite Factor Model* proposto da Bhattacharya and Dunson (2011). Rispetto alla convenzionale analisi fattoriale Bayesiana IV, tale modello introduce meccanismi innovativi che consentono di superare due limitazioni principali: la necessità di specificare *a priori* il numero di fattori k , e la difficoltà di ottenere strutture di carichi fattoriali Λ realmente sparse. Lo *Sparse Bayesian Infinite Factor Model* adotta infatti una formulazione gerarchica che permette di determinare in modo adattivo la dimensionalità effettiva dello spazio latente, lasciando che il numero di fattori rilevanti emerga direttamente dai dati. Parallelamente, l'utilizzo di distribuzioni a priori di tipo *shrinkage* induce una regolarizzazione automatica dei carichi fattoriali, favorendo la sparsità e riducendo il rischio di sovradattamento. Nel complesso, il modello coniuga flessibilità e parsimonia, offrendo una rappresentazione più efficiente e interpretabile delle strutture di dipendenza latente, in particolare in contesti ad alta dimensionalità.

Il punto di partenza rimane la formulazione classica del modello fattoriale, considerata qui nella versione normalizzata senza intercetta per

semplicità:

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

dove $x_i \in \mathbb{R}^p$ rappresenta l'osservazione i -esima, $f_i \sim \mathcal{N}_k(0, I_k)$ è il corrispondente vettore di fattori latenti, Λ è la matrice dei carichi di dimensione $p \times k$ e $\Psi = \text{diag}(\sigma_1^2, \dots, \sigma_p^2)$ raccoglie le varianze idiosincratiche. Il modello viene quindi esteso al caso $k \rightarrow \infty$, con una troncatura adattiva che seleziona effettivamente solo un numero finito di fattori.

Il nucleo innovativo dell'approccio risiede nella definizione della prior sui carichi fattoriali, che combina in maniera gerarchica meccanismi di *shrinkage* locale e globale. In particolare, ogni elemento λ_{jh} è specificato come

$$\lambda_{jh} \mid \phi_{jh}, \tau_h \sim \mathcal{N}(0, (\phi_{jh}\tau_h)^{-1}), \quad j = 1, \dots, p, \quad h = 1, 2, \dots \quad (\text{I.1})$$

$$\phi_{jh} \sim \text{Gamma}\left(\frac{\nu}{2}, \frac{\nu}{2}\right), \quad (\text{I.2})$$

dove ϕ_{jh} regola lo shrinkage locale, consentendo a singoli carichi di assumere valori significativi anche quando il resto della matrice tende a concentrarsi intorno allo zero. La componente globale è invece catturata da τ_h , definita come prodotto di variabili gamma indipendenti,

$$\tau_h = \prod_{\ell=1}^h \delta_\ell, \quad \delta_1 \sim \text{Gamma}(a_1, 1), \quad \delta_\ell \sim \text{Gamma}(a_2, 1) \quad (\ell \geq 2), \quad (\text{I.3})$$

cosicché per specifiche scelte di a_1 e a_2 il grado di shrinkage aumenti progressivamente al crescere di h . In tal modo, solo le prime colonne di Λ possono assumere valori rilevanti, mentre quelle di ordine superiore vengono via via regolarizzate, realizzando di fatto un numero effettivo finito di fattori pur in un modello con dimensionalità potenzialmente infinita. La scelta della distribuzione Gamma per i parametri ϕ_{jh} e δ_ℓ garantisce, oltre alla flessibilità interpretativa, una gestione computazionale agevole grazie a posteriori condizionali di forma coniugata.

Le varianze idiosincratiche σ_j^2 sono a loro volta trattate in modo probabilistico, specificando

$$\sigma_j^{-2} \sim \text{Gamma}(a_\sigma, b_\sigma), \quad j = 1, \dots, p,$$

ossia un prior di tipo Inverse-Gamma sulle varianze. Valori tipici per gli iperparametri prevedono a_σ compreso tra 1 e 2, mentre b_σ viene scelto in funzione della scala dei dati o impostato in modo da rendere la prior debolmente informativa.

Gli iperparametri che governano lo shrinkage globale (a_1, a_2, ν) svolgono un ruolo cruciale nel determinare la velocità con cui le colonne successive della matrice dei carichi vengono penalizzate. In particolare, la condizione $a_2 > 1$ — spesso rafforzata nella pratica scegliendo $a_2 > 2$ o $a_2 > 3$ — assicura che τ_h sia stocasticamente crescente al crescere di h così da garantire lo shrinkage delle colonne della matrice Λ . Configurazioni suggerite in letteratura (Bhattacharya and Dunson (2011)) comprendono valori di default come $a_1 \approx 2$, $a_2 \approx 3$ e $\nu = 3$, sebbene sia possibile assegnare a tali

parametri distribuzioni a priori, ad esempio di tipo Gamma, così da stimarli direttamente dai dati.

Il modello va quindi ad integrare una formulazione fattoriale classica con una struttura gerarchica di shrinkage locale e globale, che permette di gestire in maniera automatica sia la selezione del numero di fattori sia la sparsità della matrice dei carichi, mantenendo al contempo proprietà di interpretabilità e solidità inferenziale.

Lo Sparse Bayesian Infinite Factor Model possiede alcune proprietà teoriche e pratiche che ne giustificano l'adozione e ne spiegano l'efficacia in applicazioni reali. Una prima caratteristica è rappresentata dal cosiddetto *large support*: la prior indotta dal processo gamma moltiplicativo (MGP) sulla matrice di covarianza $\Sigma = \Lambda\Lambda^\top + \Psi$ assegna infatti probabilità positiva a qualunque intorno, per quanto piccolo, di una matrice di covarianza definita positiva. Questa ampiezza di supporto risulta cruciale per la consistenza bayesiana nella stima della covarianza. Un secondo aspetto riguarda la troncabilità con errore controllato. Se la matrice infinita dei carichi Λ viene approssimata con una versione troncata alle prime H colonne, la probabilità che tale approssimazione disti più di una quantità arbitraria ε dalla matrice reale tende rapidamente a zero con il crescere di H . Sotto condizioni adeguate sugli iperparametri, in particolare su a_2 , si ottengono infatti limiti esponenziali che garantiscono un controllo rigoroso dell'errore di troncatura. Il modello gode inoltre di invarianza rispetto all'ordine delle variabili e utilizza una parametrizzazione ridondante (*parameter expansion*) per eliminare la dipendenza dall'ordine che caratterizza altre restrizioni strutturali, come ad esempio l'imposizione di una forma triangolare inferiore per la matrice dei carichi. Questa scelta migliora le proprietà computazionali e riduce il rischio di introdurre artefatti dovuti alla specifica parametrizzazione adottata.

Infine, la combinazione di meccanismi di shrinkage locale e globale assicura una notevole sparsità effettiva nella matrice dei carichi fattoriali. Pur non imponendo a priori zeri esatti, la distribuzione a posteriori concentra gran parte delle entrate di Λ molto vicino a zero, consentendo così sia una migliore interpretabilità del modello, sia una selezione implicita dei fattori rilevanti.

I.I La stima: Gibbs sampler con troncatura adattiva

Algoritmo 3 Gibbs sampler per lo Sparse Bayesian Infinite Factor Model

- 1: **Inizializzazione:** fissare valori iniziali per $\Lambda^{(0)}, \{f_i^{(0)}\}, \Psi^{(0)}, \{\phi_{jh}^{(0)}\}, \{\delta_h^{(0)}\}$; scegliere una troncatura iniziale \tilde{k} e la soglia di troncatura $\varepsilon_{\text{zero}}$.
- 2: **for** $s = 1, \dots, N$ **do**
- 3: **Aggiornamento dei fattori latenti** f_i ($i = 1, \dots, n$):

$$f_i^{(s)} \sim \mathcal{N}_{\tilde{k}}\left(V_f \Lambda^{(s-1)\top} \Psi^{(s-1)-1} x_i, V_f\right), \quad V_f = (I_{\tilde{k}} + \Lambda^{(s-1)\top} \Psi^{(s-1)-1} \Lambda^{(s-1)})^{-1}.$$

- 4: **Aggiornamento delle righe dei loadings** λ_j ($j = 1, \dots, p$):

$$\lambda_j^{(s)} \sim \mathcal{N}_{\tilde{k}}\left((D_j^{(s-1)} + \sigma_j^{(s-1)-2} F^{(s)\top} F^{(s)})^{-1} \sigma_j^{(s-1)-2} F^{(s)\top} y_{(j)}, (D_j^{(s-1)} + \sigma_j^{(s-1)-2} F^{(s)\top} F^{(s)})^{-1}\right),$$

con $F^{(s)} = (f_1^{(s)}, \dots, f_n^{(s)})^\top$ e

$$D_j^{(s-1)} = \text{diag}(\phi_{j1}^{(s-1)} \tau_1^{(s-1)}, \dots, \phi_{j\tilde{k}}^{(s-1)} \tau_{\tilde{k}}^{(s-1)}).$$

- 5: **Aggiornamento dei parametri locali** ϕ_{jh} :

$$\phi_{jh}^{(s)} \sim \text{Gamma}\left(\frac{\nu+1}{2}, \frac{\nu + \tau_h^{(s-1)} (\lambda_{jh}^{(s)})^2}{2}\right).$$

- 6: **Aggiornamento dei parametri globali** δ_h e τ_h :

Per $h = 1$:

$$\delta_1^{(s)} \sim \text{Gamma}\left(a_1 + \frac{p\tilde{k}}{2}, 1 + \frac{1}{2} \sum_{l=1}^{\tilde{k}} \tau_l^{(s-1)} \sum_{j=1}^p \phi_{jl}^{(s)} (\lambda_{jl}^{(s)})^2\right).$$

Per $h \geq 2$:

$$\delta_h^{(s)} \sim \text{Gamma}\left(a_2 + \frac{p(\tilde{k}-h+1)}{2}, 1 + \frac{1}{2} \sum_{l=h}^{\tilde{k}} \tau_l^{(s-1)} \sum_{j=1}^p \phi_{jl}^{(s)} (\lambda_{jl}^{(s)})^2\right).$$

quindi $\tau_h^{(s)} = \prod_{\ell=1}^h \delta_\ell^{(s)}$.

- 7: **Aggiornamento delle varianze specifiche** σ_j^{-2} :

$$\sigma_j^{-2(s)} \sim \text{Gamma}\left(a_\sigma + \frac{n}{2}, b_\sigma + \frac{1}{2} \sum_{i=1}^n (x_{ij} - \lambda_j^{(s)\top} f_i^{(s)})^2\right).$$

- 8: **Adattamento della troncatura** \tilde{k} : secondo uno schema di *diminishing adaptation*, ossia con controlli via via meno frequenti al crescere di s , verificare se alcune colonne di $\Lambda^{(s)}$ hanno carichi trascurabili ($|\lambda_{jh}^{(s)}| < \varepsilon_{\text{zero}}$ per tutti j). In tal caso, eliminare le colonne ridondanti oppure, se necessario, aggiungerne di nuove campionate dal prior.

- 9: **end for**

- 10: **Output:** campioni $\{\Lambda^{(s)}, \{f_i^{(s)}\}, \Psi^{(s)}, \Phi^{(s)}, \Delta^{(s)}\}_{s=1}^N$ (dopo *burn-in* e *thinning*), con stima del numero effettivo di fattori data da k^* .
-

Dal punto di vista implementativo, alcuni accorgimenti risultano particolarmente utili per garantire l'efficienza e l'affidabilità del modello. Un primo aspetto riguarda l'aggiornamento dei carichi fattoriali: operare per blocchi

sulle righe della matrice Λ , aggiornando cioè i vettori λ_j nel loro insieme, consente di ridurre sensibilmente i costi computazionali, poiché richiede l'inversione di matrici di dimensione $k \times k$, tipicamente molto più piccole rispetto al numero di variabili p .

Un'ulteriore raccomandazione è la standardizzazione preliminare dei dati. Questa operazione semplifica infatti la calibrazione degli iperparametri relativi alle varianze specifiche σ_j^2 e rende più immediata l'interpretazione della scala dei carichi λ_{jh} , migliorando così la leggibilità dei risultati.

In ultimo luogo la troncatura della matrice dei carichi richiede di fissare una soglia $\varepsilon_{\text{zero}}$ al di sotto della quale i valori vengono considerati trascurabili. Poiché la scelta di tale soglia può influire sulla stabilità delle stime, è opportuno adottare valori molto piccoli, tipicamente compresi tra 10^{-4} e $1e^{-3}$, e verificarne la robustezza nelle applicazioni pratiche.

In conclusione il modello proposto da Bhattacharya & Dunson combina eleganza teorica (large support, troncabilità controllata) con praticità computazionale (block updates coniugati, adattamento della troncatura). La molteplicità di livelli di shrinkage (locale ϕ_{jh} e globale τ_h) garantisce la conservazione dei segnali significativi pur imponendo forte regolarizzazione sulle colonne meno rilevanti, permettendo così di lavorare in regime $p \gg n$ con buone proprietà predittive e di selezione.

I.II Predizione tramite regressione fattoriale latente

Uno dei principali vantaggi dello *Sparse Bayesian Infinite Factor Model* è la capacità di affrontare in maniera naturale il problema della predizione in contesti ad alta dimensionalità, dove le covariate possono essere numerose e fortemente correlate. L'approccio si fonda sulla possibilità di modellare congiuntamente variabili di risposta e predittori, ottenendo così distribuzioni predittive e intervalli di credibilità che rispettano pienamente la struttura del modello.

In termini generali, la predizione bayesiana può essere formulata come

$$f(z_{n+1} \mid w_{n+1}, x_1, \dots, x_n) = \int f(z_{n+1} \mid w_{n+1}, \Omega) \pi(\Omega \mid x_1, \dots, x_n) d\Omega,$$

dove Ω raccoglie i parametri del modello e $\pi(\Omega \mid x_1, \dots, x_n)$ rappresenta la distribuzione a posteriori. Questa espressione sottolinea che, a differenza di approcci frequentisti che si basano su stime puntuali dei parametri, la distribuzione predittiva bayesiana media su tutte le possibili configurazioni compatibili con i dati osservati.

Per rendere operativo questo calcolo, si adotta una rappresentazione fattoriale congiunta di risposta e covariate. Si considera quindi

$$x_i = \begin{pmatrix} z_i \\ w_i \end{pmatrix} \in \mathbb{R}^p, \quad i = 1, \dots, n, \quad (\text{I.4})$$

dove $z_i \in \mathbb{R}$ è la variabile dipendente e $w_i \in \mathbb{R}^{p-1}$ è il corrispondente vettore di covariate. La formulazione fattoriale assume allora la forma

$$x_i = \Lambda f_i + \varepsilon_i, \quad f_i \sim \mathcal{N}_{k^*}(0, I), \quad \varepsilon_i \sim \mathcal{N}_p(0, \Psi),$$

con Λ matrice dei carichi $p \times k^*$ e $\Psi = \text{diag}(\sigma_1^2, \dots, \sigma_p^2)$ matrice delle varianze specifiche. Ne segue che

$$x_i \sim \mathcal{N}_p(0, \Sigma), \quad \Sigma = \Lambda \Lambda^\top + \Psi. \quad (\text{I.5})$$

Per la variabile di risposta z_i condizionata alle covariate w_i , la matrice di covarianza Σ viene suddivisa in blocchi,

$$\Sigma = \begin{pmatrix} \Sigma_{zz} & \Sigma_{zw} \\ \Sigma_{wz} & \Sigma_{ww} \end{pmatrix},$$

da cui, per le proprietà della normalità congiunta, si ottiene

$$z_i \mid w_i \sim \mathcal{N}(w_i^\top \beta, \Sigma_{z|w}),$$

con

$$\beta = \Sigma_{ww}^{-1} \Sigma_{wz}, \quad \Sigma_{z|w} = \Sigma_{zz} - \Sigma_{zw} \Sigma_{ww}^{-1} \Sigma_{wz}.$$

Il campionamento della distribuzione predittiva avviene naturalmente nel contesto del Gibbs sampler. Ad ogni iterazione t , dalla stima campionaria della matrice di covarianza

$$\Sigma^{(t)} = \Lambda^{(t)} \Lambda^{(t)\top} + \Psi^{(t)}$$

si ricava il corrispondente coefficiente

$$\beta^{(t)} = (\Sigma_{ww}^{(t)})^{-1} \Sigma_{wz}^{(t)}.$$

Per un nuovo vettore di covariate w_{n+1} , la distribuzione predittiva della risposta si ottiene allora come

$$z_{n+1}^{(t)} \sim \mathcal{N}(w_{n+1}^\top \beta^{(t)}, \Sigma_{z|w}^{(t)}),$$

dove la variabilità riflette sia l'incertezza intrinseca del modello sia quella derivante dalla stima dei parametri.

Aggregando i campioni $\{z_{n+1}^{(t)}\}_{t=1}^N$ generati nelle iterazioni del Gibbs sampler si ottiene la distribuzione predittiva a posteriori per z_{n+1} . Da essa è possibile calcolare la media predittiva

$$\hat{z}_{n+1} = \frac{1}{N} \sum_{t=1}^N z_{n+1}^{(t)},$$

costruire intervalli di credibilità al livello desiderato e valutare misure di accuratezza predittiva, come l'errore quadratico medio o il tasso di copertura degli intervalli.

La regressione fattoriale latente costituisce un'estensione operativa della predizione bayesiana all'interno dei modelli a fattori infiniti. Essa permette di ottenere predizioni formalmente coerenti e computazionalmente efficienti, fornendo al contempo una valutazione completa dell'incertezza predittiva, caratteristica essenziale in applicazioni reali ad alta dimensionalità.

II Applicazione e sviluppo del metodo su dati simulati

Per verificare preliminarmente il corretto funzionamento dello *Sparse Bayesian Infinite Factor Model*, l'algoritmo di campionamento è stato implementato in linguaggio R (consultabile in appendice II) e applicato a dati simulati. Questa fase di replica consente di controllare la coerenza dei risultati rispetto a quanto riportato da Bhattacharya and Dunson (2011), ponendo le basi per un confronto sistematico. In particolare, sono state considerate le misure di bontà proposte dagli autori per valutare la qualità dell'inferenza e la loro riproducibilità nell'ambiente simulato, (riportate e descritte nella tabella 2.1). L'analisi è stata arricchita includendo ulteriori criteri diagnostici standard per algoritmi MCMC, come l'esame dei traceplot e la valutazione della convergenza, al fine di fornire un quadro più completo delle prestazioni del metodo.

I dati utilizzati per la replica sono stati simulati seguendo l'impostazione di Bhattacharya and Dunson (2011) (per i dettagli l'implementazione del codice per la generazione dei dati è consultabile in appendice I). Sono stati considerati diversi scenari sperimentali, caratterizzati da $(p, k) \in \{(100, 5), (500, 10), (1000, 15)\}$. In ciascun caso il numero di osservazioni è pari a $n = 200$. Per la valutazione predittiva, il dataset è stato suddiviso in due parti di uguale dimensione: 100 osservazioni utilizzate per la stima e 100 per la predizione.

Tabella 2.1: Definizioni delle metriche di valutazione utilizzate da Bhattacharya and Dunson (2011) per valutare le performance dell'implementazione

Metrica	Definizione
<i>Metriche per la matrice di varianza-covarianza Σ</i>	
Mean Squared Error (MSE, Σ)	$\text{MSE}_\Sigma = \frac{1}{p^2} \sum_{j=1}^p \sum_{l=1}^p \left(\hat{\Sigma}_{jl} - \Sigma_{jl} \right)^2$
Average Absolute Bias (AAB, Σ)	$\text{AAB}_\Sigma = \frac{1}{p^2} \sum_{j=1}^p \sum_{l=1}^p \left \hat{\Sigma}_{jl} - \Sigma_{jl} \right $
Maximum Absolute Bias (MAB, Σ)	$\text{MAB}_\Sigma = \max_{1 \leq j, l \leq p} \left \hat{\Sigma}_{jl} - \Sigma_{jl} \right $
<i>Metriche per la capacità predittiva (z)</i>	
Mean Squared Prediction Error (MSPE)	$\text{MSPE} = \frac{1}{n} \sum_{i=1}^n (\hat{z}_i - z_i)^2$
Average Absolute Prediction Error (AAPE)	$\text{AAPE} = \frac{1}{n} \sum_{i=1}^n \hat{z}_i - z_i $
Maximum Absolute Prediction Error (MAPE)	$\text{MAPE} = \max_{1 \leq i \leq n} \hat{z}_i - z_i $
<i>Metriche per i coefficienti di regressione β</i>	
Mean Squared Error (MSE, β)	$\text{MSE}_\beta = \frac{1}{p-1} \sum_{j=1}^{p-1} (\hat{\beta}_j - \beta_j)^2$
Average Absolute Bias (AAB, β)	$\text{AAB}_\beta = \frac{1}{p-1} \sum_{j=1}^{p-1} \hat{\beta}_j - \beta_j $
Maximum Absolute Bias (MAB, β)	$\text{MAB}_\beta = \max_{1 \leq j \leq p-1} \hat{\beta}_j - \beta_j $

Le tabelle seguenti presentano i valori originali (B.D.) (presenti in Bhattacharya and Dunson (2011) e lo spazio per i risultati ottenuti dall'implementazione sviluppata (Aut.).

Nel complesso, i risultati ottenuti dall'implementazione sviluppata mostrano un buon grado di coerenza con quelli riportati da Bhattacharya and Dunson (2011), confermando il corretto funzionamento del codice. Come si osserva nella Tabella 2.2, le prestazioni nella stima della matrice di varianza-covarianza risultano complessivamente in linea con quelle originali. In particolare, i *mean square error* tendono a essere lievemente più elevati rispetto ai valori di riferimento, mentre l'*average absolute bias* risulta mediamente inferiore, indicando una buona accuratezza nella stima degli elementi diagonali e off-diagonali della matrice. Le differenze osservate nel *maximum absolute bias* si mantengono contenute e mostrano un andamento analogo nei tre scenari considerati, $(p, k) = (100, 5)$, $(500, 10)$ e $(1000, 15)$. Nel complesso, tali risultati suggeriscono che la procedura implementata riproduce fedelmente la struttura di dipendenza sottostante.

Tabella 2.2: Confronto delle prestazioni nella stima della matrice di covarianza nello studio di simulazione. Sono riportati i valori medi, migliori e peggiori su 50 repliche in termini di mean square error ($\times 10^2$), average absolute bias ($\times 10^2$) e maximum absolute bias ($\times 10^2$) per i diversi scenari di (p, k) .

	(100,5)		(500,10)		(1000,15)	
	B. D.	Aut.	B. D.	Aut.	B. D.	Aut.
MSE $_{\Sigma}$						
media	0.20	0.27	0.10	0.16	0.10	0.18
min	0.10	0.10	0.02	0.09	0.02	0.10
max	0.30	0.61	0.20	0.25	0.40	0.45
AAB $_{\Sigma}$						
media	1.90	1.22	0.60	0.41	0.40	0.53
min	1.30	0.70	0.40	0.33	0.20	0.29
max	2.50	1.82	0.90	0.49	0.60	0.63
MAB $_{\Sigma}$						
media	50.90	56.00	95.40	93.14	115.00	127.00
min	38.80	38.80	50.20	63.78	52.60	67.30
max	74.10	82.50	152.00	135.56	242.00	213.00

Per quanto riguarda le prestazioni predittive, riportate in Tabella 2.3, l'implementazione proposta mostra performance superiori rispetto ai valori originali per tutte e tre le metriche considerate — *mean squared prediction error*, *average absolute prediction error* e *maximum absolute prediction error* — in ciascuno degli scenari sperimentali. In particolare, gli errori medi risultano sensibilmente inferiori, suggerendo una buona capacità del modello di riprodurre le relazioni tra fattori e variabili osservate anche in fase predittiva.

Tabella 2.3: Prestazioni predittive nello studio di simulazione. Sono riportati i valori medi, migliori e peggiori su 50 repliche in termini di mean squared prediction error (MSPE), average absolute prediction error (AAPE) e maximum absolute prediction error (MAPE) per i diversi scenari di (p, k) .

		(100,5)		(500,10)		(1000,15)	
		B. D.	Aut.	B. D.	Aut.	B. D.	Aut.
MSPE							
media		0.63	0.02	0.41	0.03	0.95	0.05
min		0.32	0.00	0.18	0.00	0.57	0.00
max		0.89	0.09	0.86	0.56	1.48	0.65
AAPE							
media		0.62	0.10	0.51	0.11	0.80	0.16
min		0.47	0.06	0.33	0.05	0.60	0.09
max		0.85	0.24	0.80	0.61	0.99	0.75
MAPE							
media		2.19	0.36	1.71	0.38	2.54	0.75
min		1.36	0.18	1.21	0.15	1.83	0.13
max		3.15	0.90	2.95	2.02	3.27	2.74

Infine, l'analisi delle prestazioni nella stima dei coefficienti di regressione (Tabella 2.4) evidenzia come, all'aumentare della dimensionalità del problema — ovvero per valori crescenti di p e k — l'implementazione sviluppata tenda a fornire stime più accurate dei parametri β in termini di MSE, risultando competitiva o addirittura migliore rispetto ai risultati di riferimento.

Tabella 2.4: Prestazioni nella stima dei coefficienti di regressione nello studio di simulazione. Sono riportati gli errori medi su 50 repliche in termini di mean squared error ($\times 10^3$), average absolute bias ($\times 10^3$) e maximum absolute bias ($\times 10^3$) per i diversi scenari di (p, k) .

		(100,5)		(500,10)		(1000,15)	
		B. D.	Aut.	B. D.	Aut.	B. D.	Aut.
MSE $_{\beta}$ (media)		1.10	1.75	0.10	0.02	0.00	0.00
AAB $_{\beta}$ (media)		10.10	10.86	1.70	0.88	0.90	0.67
MAB $_{\beta}$ (media)		176.10	274.90	172.50	79.72	102.60	85.30

In base ai risultati ottenuti, il codice implementato sembra approssimare in modo soddisfacente il numero effettivo di fattori latenti k , mostrando una leggera tendenza alla sovrastima. In particolare, il valore medio osservato sulle repliche, \bar{k} , calcolato come media dei valori finali \tilde{k} ottenuti

nelle diverse repliche del modello risulta pari a 6.44, 10.70 e 16.07 nei casi $k = 5$, $k = 10$ e $k = 15$, rispettivamente. Tale comportamento è analogo a quanto osservato da Bhattacharya and Dunson (2011), che riporta una lieve sovrastima per $k = 5$ e $k = 10$, mentre per $k = 15$ il modello proposto tende a sovrastimare lievemente dove l'approccio di Bhattacharya and Dunson (2011) mostra invece una sottostima di entità contenuta.

Le discrepanze residue rispetto ai risultati di Bhattacharya and Dunson (2011) possono essere attribuite a una combinazione di fattori. Oltre alle differenze nelle specifiche di implementazione, esse riflettono anche la componente intrinsecamente stocastica del processo di simulazione e del campionamento MCMC: i dati sintetici, pur generati secondo lo stesso schema, non coincidono perfettamente in ciascuna replica, e la variabilità dovuta alla randomicità delle catene di Markov può contribuire a leggere deviazioni nelle stime finali.

Dal punto di vista implementativo, le principali differenze rispetto alla configurazione originale riguardano: (i) l'assenza di iper-prior sulle distribuzioni Gamma dei parametri globali δ_h , per i quali Bhattacharya and Dunson (2011) introducono iper-prior Gamma con iper-iperparametri $(2, 1)$, mentre nel presente lavoro i corrispondenti iperparametri a_1 e a_2 sono stati fissati rispettivamente a 2.1 e 3.1, in accordo con quanto suggerito nel paper stesso; (ii) l'utilizzo di una soglia di troncatura $\varepsilon_{\text{zero}} = 10^{-3}$ anziché 10^{-4} , valore scelto ispirandosi all'implementazione disponibile sul repository *GitHub* associato agli autori del modello.¹

Tali differenze, unitamente alla variabilità casuale insita nei dati simulati e nel campionamento MCMC, possono spiegare le lievi discrepanze osservate, pur non alterando sostanzialmente le proprietà di convergenza e la coerenza complessiva dei risultati ottenuti.

Nel complesso, tali evidenze indicano che l'algoritmo implementato risulta stabile e coerente con quanto proposto in letteratura, garantendo una buona capacità di ricostruzione e predizione nelle diverse configurazioni sperimentali.

Al fine di supportare e rendere più immediata l'interpretazione dei risultati riportati nelle tabelle, che sintetizzano le prestazioni medie, migliori e peggiori ottenute su più repliche, sono stati prodotti anche alcuni grafici riferiti a una singola catena di simulazione. Queste rappresentazioni hanno lo scopo di fornire una verifica visiva della bontà delle stime, permettendo di apprezzare in maniera più diretta la capacità del modello di riprodurre la matrice di covarianza, di garantire prestazioni predittive accurate e di stimare correttamente i coefficienti di regressione con i relativi intervalli di credibilità. La Figura 2.1 mostra un confronto visivo tra la matrice di covarianza stimata e quella vera, riportato nel caso di una matrice 30×30 al fine di garantire una rappresentazione graficamente leggibile. Infatti, con le dimensioni utilizzate negli scenari di simulazione, la visualizzazione risulterebbe di difficile interpretazione (scenari in cui si sono già dimostrate buone

¹Il riferimento è al codice fornito dagli autori di Bhattacharya and Dunson (2011), disponibile al link: <https://github.com/rajeshbhattacharya/sparsebfa>.

performace nella tabella 2.2). Il confronto evidenzia un'elevata somiglianza tra la matrice stimata e quella di riferimento, suggerendo un'accurata capacità di riproduzione da parte del modello.

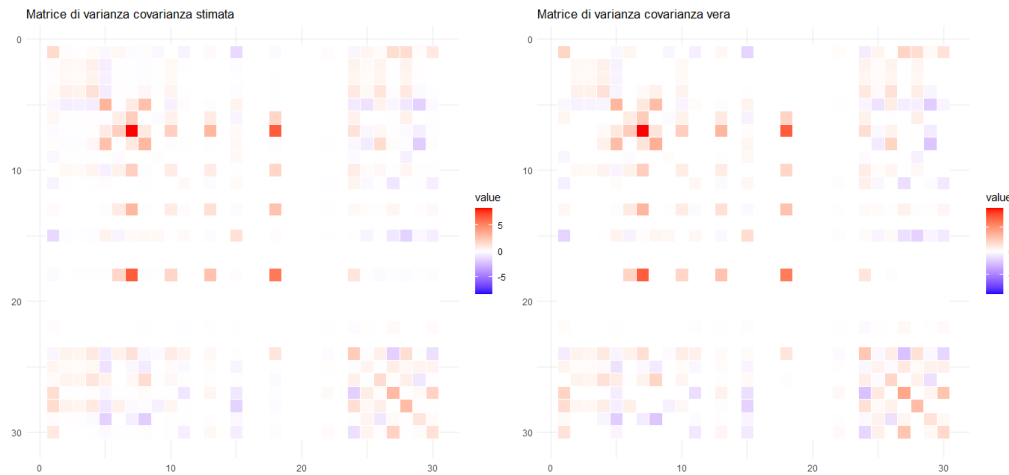


Figura 2.1: Confronto visivo tra matrice di covarianza stimata e vera e differenze in una replica del codice implementato.

La Figura 2.2 presenta la relazione tra i valori predetti e i valori reali, distinguendo tra il campione di addestramento e quello di validazione. Il grafico mette in evidenza una buona aderenza delle stime ai valori osservati, confermando l'efficacia del modello in termini di capacità predittiva.

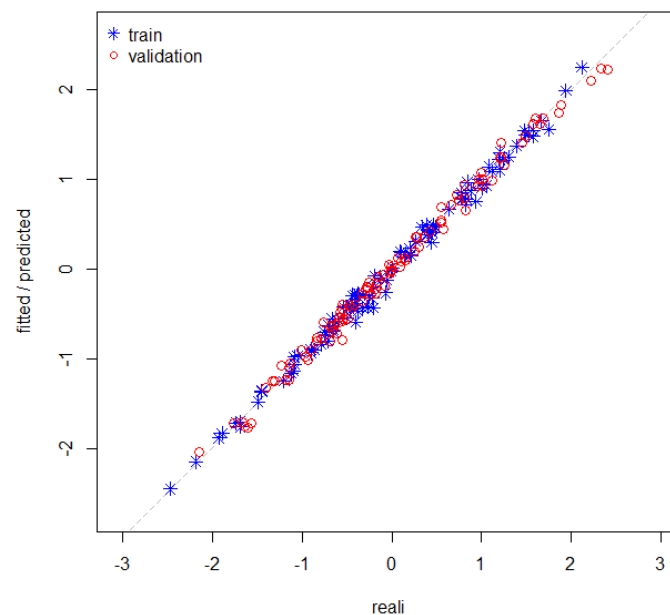


Figura 2.2: Confronto tra valori predetti e valori reali in una replica del codice implementato.

Infine, la Figura 2.3 mostra il confronto tra i coefficienti di regressione stimati e quelli veri, corredato dai rispettivi intervalli di credibilità. Anche in questo caso, i risultati appaiono complessivamente soddisfacenti, con stime che tendono a distribuirsi attorno ai valori reali e con intervalli di credibilità generalmente di ampiezza contenuta, a supporto della bontà del metodo proposto. Tuttavia, si osserva che non tutti gli intervalli di credibilità includono il valore vero dei parametri, e che diverse stime risultano prossime allo zero invece che zero. Tale comportamento risulta coerente con le proprietà del modello e con la natura del processo di stima. Come mostrato in Figura 2.1, la matrice di varianza-covarianza vera è stata generata per contenere numerosi elementi nulli (codice in appendice I). Nel processo di stima, tuttavia, l'effetto di *shrinkage* porta i corrispondenti coefficienti β a valori molto prossimi, ma non esattamente pari, a zero. Di conseguenza, per diversi parametri che nel modello vero risultano nulli, le stime tendono a essere lievemente sovrastimate. Inoltre, la soluzione della matrice di varianza-covarianza latente non è perfettamente univoca, poiché più configurazioni dei fattori possono generare strutture di dipendenza equivalenti tra le variabili osservate. In presenza di forte correlazione tra le variabili, o di un livello di rumore relativamente basso, tale non unicità si accentua, rendendo le stime dei coefficienti β più sensibili e potenzialmente instabili, fenomeno analogo a quanto si osserva nei modelli di regressione lineare con regressori fortemente correlati. Nel complesso, le stime ottenute risultano tuttavia soddisfacenti e in grado di riprodurre in modo coerente la struttura dei dati simulati. Tuttavia, si osserva che non tutti gli intervalli di credibilità includono il valore vero dei parametri e che diverse stime tendono ad assumere valori prossimi allo zero, aspetti che possono essere ricondotti alle ragioni sopra descritte.

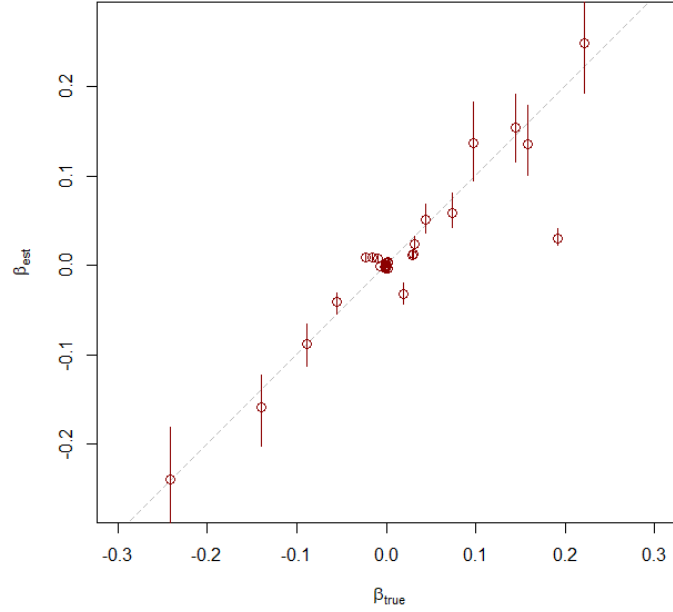


Figura 2.3: Confronto tra coefficienti stimati e coefficienti reali in una replica del codice implementato.

Diagnosi sulla convergenza Accanto alla verifica della riproducibilità dei risultati, è stata condotta un'analisi mirata alla valutazione della convergenza della catena. Tale approfondimento si rende necessario al fine di corroborare le prestazioni osservate nelle analisi precedenti, dimostrando che gli esiti ottenuti non costituiscono il prodotto di fluttuazioni casuali, bensì il risultato di una stima stabile e affidabile del modello. Per l'analisi sono state generate 20 000 iterazioni della catena MCMC, di cui le prime 5 000 sono state scartate come fase di burn-in. Al fine di valutare la stabilità e la convergenza della catena, vengono presentati i traceplot di alcune quantità diagnostiche di interesse: la log-verosimiglianza, la log-posterior e l'andamento del numero effettivo di fattori k .

La log-verosimiglianza misura il grado di compatibilità dei dati osservati y con i parametri correnti del modello. Nel caso del modello fattoriale sparso, ogni osservazione è rappresentata come

$$y \mid \Lambda, \eta, \psi \sim \mathcal{N}(\eta \Lambda^\top, \text{diag}(1/\psi)),$$

dove Λ è la matrice dei carichi, η i fattori latenti e ψ i parametri di precisione specifici. La log-verosimiglianza assume quindi la forma

$$\ell(\Lambda, \eta, \psi \mid y) = \sum_{i=1}^n \log \left[\mathcal{N}(y_i \mid \eta_i \Lambda^\top, \text{diag}(1/\psi)) \right],$$

che corrisponde all'implementazione poi effettuata nel software R. Il relativo traceplot permette di verificare se la catena esplora in modo stabile regioni

di alta probabilità: oscillazioni stazionarie attorno a un livello medio sono indice di buona convergenza, mentre trend sistematici indicano problemi di mixing o persistenza della fase transitoria.

Accanto alla verosimiglianza, la log-posterior incorpora anche i contributi dei prior sui parametri. In termini generali, la distribuzione a posteriori è

$$p(\theta \mid y) \propto p(y \mid \theta) p(\theta),$$

dove $\theta = (\Lambda, \eta, \psi, \phi, \delta)$. La log-posterior si ottiene quindi come

$$\log p(\theta \mid y) = \ell(\Lambda, \eta, \psi \mid y) + \log p(\psi) + \log p(\Lambda \mid \phi, \delta) + \log p(\eta) + \log p(\phi) + \log p(\delta).$$

Ciascun termine riflette un'informazione a priori: ad esempio, le distribuzioni Gamma sui parametri di shrinkage ϕ e δ regolano la sparsità dei carichi, mentre i prior normali sui fattori latenti η ne stabiliscono ortogonalità e scala. La funzione implementata in R riassume precisamente questa scomposizione.

Dal punto di vista diagnostico, la log-posterior è particolarmente utile poiché sintetizza l'intera struttura probabilistica del modello, combinando evidenza empirica e informazione a priori. Un traceplot stazionario della log-posterior è indicativo della capacità della catena di campionare in modo stabile dalla distribuzione target e quindi di garantire affidabilità alle inferenze bayesiane.

In sintesi, l'analisi dei traceplot della log-verosimiglianza e della log-posterior fornisce un duplice riscontro: da un lato la coerenza del modello con i dati osservati, dall'altro la stabilità della catena rispetto alla distribuzione a posteriori completa. Entrambi gli strumenti risultano pertanto indispensabili per una diagnosi accurata della convergenza.

Di seguito è riportata tale diagnosi della convergenza in uno scenario con $p = 100$ e $k = 5$. Come mostrato nella Figura 2.4, il traceplot della log-verosimiglianza evidenzia come la log-verosimiglianza (avendo scartato il burn-in) tenda a stabilizzarsi, indicando una prima forma di convergenza del campionamento.

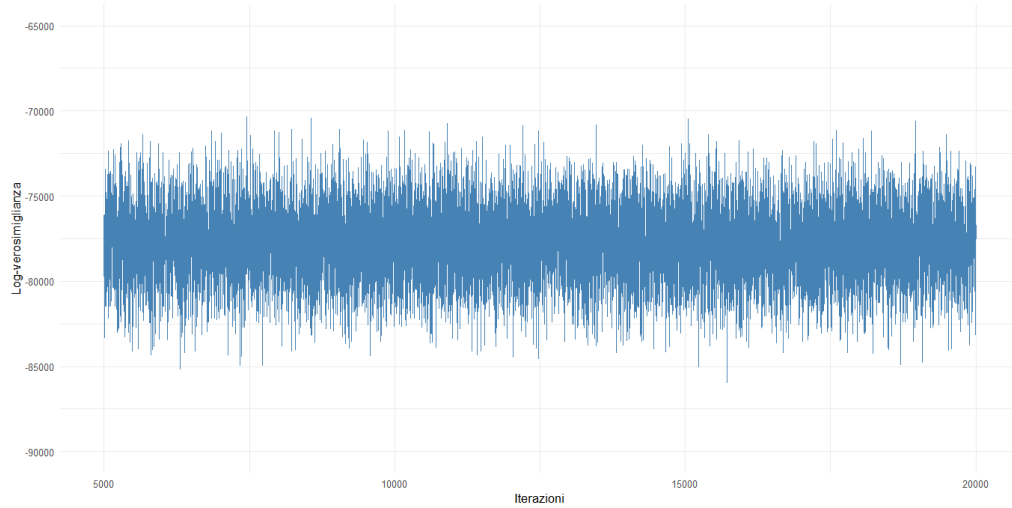


Figura 2.4: Traceplot della log-verosimiglianza lungo le iterazioni della catena MCMC di una replica delle 50 effettuate nello studio di simulazione (scenario con $p=100$ e $k=5$).

Analogamente, la Figura 2.5 riporta l'andamento della log-posterior (in forma aperta), la quale, pur mostrando anch'essa un progressivo avvicinamento alla stazionarietà, risulta maggiormente influenzata dalle variazioni del numero di fattori latenti.

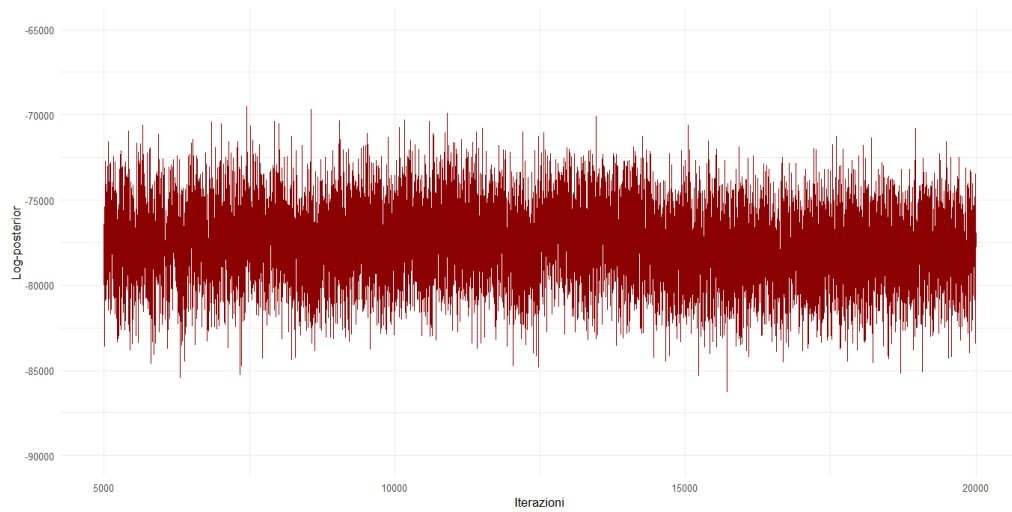


Figura 2.5: Traceplot della log-posterior (in forma aperta) lungo le iterazioni della catena MCMC di una replica delle 50 effettuate nello studio di simulazione (scenario con $p=100$ e $k=5$).

Infine, la Figura 2.6 descrive l'evoluzione di \tilde{k} , da cui si osserva l'ottima approssimazione del numero di k fattori. In secondo luogo si evince come soltanto nel momento in cui il numero di fattori stimato si stabilizza, si ottenga una chiara evidenza di convergenza effettiva della catena. L'analisi congiunta delle tre rappresentazioni consente dunque di concludere che la

stabilizzazione di k costituisce una condizione necessaria affinché si manifesti una piena convergenza sia della log-verosimiglianza sia della log-posterior.

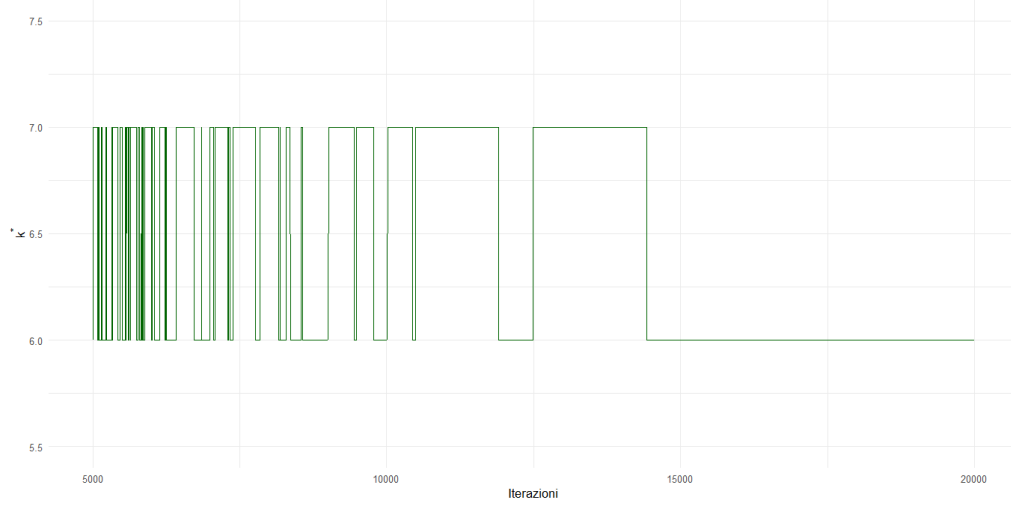


Figura 2.6: Andamento del numero di fattori k stimato lungo la catena MCMC di una replica delle 50 effettuate nello studio di simulazione (scenario con $p=100$ e $k=5$).

In aggiunta ai traceplot precedentemente presentati, - su dati generati nello stesso modo (I) e scenario ($p=100, k=5$) - è stata condotta un'analisi ulteriore con l'obiettivo di valutare in modo più diretto il comportamento del meccanismo adattivo relativo al numero effettivo di fattori k . Come discusso, nell'impostazione di Bhattacharya e Dunson la frequenza degli aggiornamenti adattivi decresce progressivamente con l'aumentare delle iterazioni, fino a stabilizzarsi su un valore asintotico. Tale proprietà, nota come *diminishing adaptation*, garantisce che l'impatto degli adattamenti sul processo stocastico si riduca nel tempo, favorendo la convergenza della catena.

Per verificare sperimentalmente la robustezza di questo meccanismo, è stata generata una catena MCMC nella quale, a metà del percorso, il processo adattivo di k è stato reimpostato. In questo modo è stato possibile osservare se, anche a seguito di una riattivazione forzata del meccanismo, la catena fosse in grado di ristabilizzare il numero effettivo di fattori in coerenza con quanto emerso nella prima parte dell'esecuzione.

Il traceplot riportato in Figura 2.7 mostra l'andamento del numero di fattori latenti \tilde{k} in seguito al riavvio del processo di adattamento con *warm start*. Dalla figura si osserva come, dopo la riattivazione dell'adattamento, l'algoritmo riprenda regolarmente la fase di esplorazione e converga nuovamente verso valori stabili, senza evidenziare comportamenti anomali o oscillazioni persistenti. Tale evidenza empirica conferma l'efficacia del meccanismo di *diminishing adaptation* e la solidità del modello nel processo di individuazione del numero di fattori rilevanti.

Da specificare che il traceplot della Figura 2.6, in cui il numero finale di fattori stimato è pari a $\tilde{k} = 6$, si riferisce a una delle cinquanta repliche

considerate nello studio di simulazione. Nel caso del riavvio con *warm start*, il valore finale stimato risulta invece pari a $\tilde{k} = 7$, differenza che è attribuibile alla naturale variabilità introdotta dalla nuova inizializzazione e al fatto che, in ciascuna replica, i semi utilizzati per la generazione dei dati sintetici e per l'avvio delle catene MCMC risultano differenti. Tale risultato risulta comunque coerente con il valore medio osservato sulle repliche, pari a $\bar{k} = 6.44$, calcolato come media dei valori finali \tilde{k} (ottenuti nelle diverse 50 repliche del modello). Questo valore si mantiene in stretta coerenza con quanto riportato da Bhattacharya and Dunson (2011), che riporta un valore medio $\bar{k} = 6.82$ in condizioni analoghe. Nel complesso, la concordanza tra le stime ottenute e i risultati di riferimento rafforza l'affidabilità del meccanismo di adattamento implementato e conferma la capacità del modello di individuare in modo stabile e realistico il numero effettivo di fattori latenti.

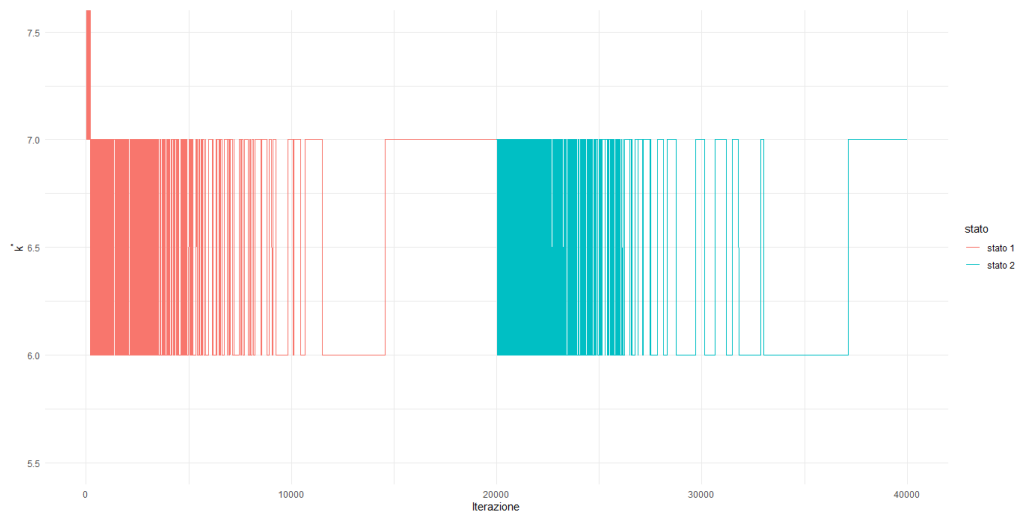


Figura 2.7: Traceplot del numero effettivo di fattori k in presenza di un warm start a metà catena (scenario con $p=100$ e $k=5$).

Capitolo 3

Applicazioni a dati di spettroscopia del vicino infrarosso

In questo capitolo vengono presentate le analisi condotte su dati reali di spettroscopia relativi a campioni di impasto di biscotti e successivamente di grano.

La spettroscopia del vicino infrarosso (Near Infrared Spectroscopy, NIRS) rappresenta una metodologia analitica rapida, non distruttiva ed economicamente vantaggiosa, sempre più utilizzata in diversi ambiti di ricerca e controllo di qualità. Essa si basa sulla misura dell'interazione della radiazione elettromagnetica con i campioni, la quale può essere osservata sia in termini di assorbanza, che quantifica la frazione di luce assorbita, sia in termini di riflettanza, che misura la quota di luce restituita dal campione. Tali grandezze sono strettamente correlate: per campioni opachi o solidi, la riflettanza può essere trasformata in un'indicazione di assorbanza mediante una relazione logaritmica inversa, permettendo una coerente interpretazione dei dati indipendentemente dalla modalità di misura adottata. Ad esempio, nei campioni di impasto la spettroscopia NIRS è stata condotta in riflettanza, mentre per i campioni di grano si è fatto ricorso alla misura diretta dell'assorbanza.

In entrambi i casi, la NIRS rileva le bande di assorbimento associate alle vibrazioni fondamentali e ai loro armonici nei legami C–H, N–H e O–H, restituendo spettri caratterizzati da un'elevata densità informativa, (Osborne and Fearn (1986)). Il risultato è un vettore continuo di dati che, pur non essendo di immediata interpretazione, risulta altamente descrittivo della composizione chimico-fisica del campione. Tali spettri, una volta sottoposti a opportuni metodi di elaborazione e modellizzazione multivariata, consentono di estrarre informazioni quantitative e qualitative di grande rilevanza, rendendo la NIRS una tecnica estremamente versatile per lo studio e la caratterizzazione di matrici complesse.

I Applicazione ai NIRS sull'impasto dei biscotti

L'impiego della spettroscopia nel vicino infrarosso trova ampia applicazione nel settore alimentare, dove la necessità di metodi rapidi, e a basso costo è particolarmente rilevante per il controllo della qualità e la caratterizzazione delle materie prime e dei prodotti trasformati. In tale contesto, l'analisi dell'impasto per prodotti da forno, come i biscotti, rappresenta un ambito di particolare interesse, poiché la composizione dell'impasto e la sua variabilità influenzano in maniera significativa le proprietà tecnologiche e sensoriali del prodotto finale. L'utilizzo della NIRS consente di acquisire spettri in grado di riflettere le differenze compositive e strutturali tra campioni, offrendo così la possibilità di monitorare parametri chimici e fisici in modo accurato ed efficiente.

I.I Dataset, pre-processing e metodo

Il dataset considerato comprende spettri di riflettanza nel vicino infrarosso (NIR), raccolti sull'impasto di biscotti (Osborne et al. (1984)). Ciascun campione è rappresentato da 700 punti spettrali, acquisiti nell'intervallo 1100–2498 nm con passo di 2 nm, sia nel training set sia nel test set. La continuità dell'informazione di riflettanza per campione è rappresentata nella figura 3.1, in questo caso gli spettri sono stati centrati. Complessivamente, il dataset include 72 osservazioni, ciascuna corrispondente a un diverso impasto, per un totale di 72×700 dati. Il training set comprende 40 osservazioni, tra cui l'osservazione 23, identificata come anomala secondo Brown et al. (2001) e quindi rimossa, mentre il test set è costituito dalle restanti 32 osservazioni. Le variabili risposta indicano le percentuali dei principali ingredienti dell'impasto (grassi, saccarosio, farina e acqua), la cui somma non raggiunge esattamente 100 a causa della presenza di componenti minori, ma risulta prossima a tale valore. Per ridurre la dimensionalità dei dati e consentire un confronto diretto con i risultati di Brown et al. (2001) è stato applicato un processo di thinning degli spettri, portando a 256 il numero di variabili spettrali. Successivamente, i dati sono stati sottoposti a normalizzazione e standardizzazione, al fine di garantire comparabilità tra i campioni e stabilità nei modelli statistici applicati.

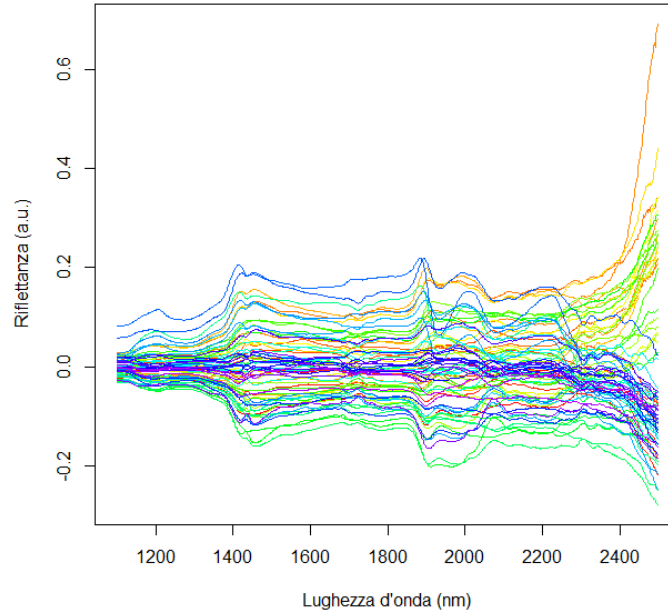


Figura 3.1: Spettri NIR di riflettanza centrati relativi a 72 campioni di impasto per biscotti.

I dati NIRS relativi ai campioni di pasta di biscotti sono stati analizzati mediante l'applicazione dello *Sparse Bayesian Infinite Factor Model* (cap. 2 sez. I). L'inferenza è stata condotta tramite campionamento MCMC (Algoritmo 3), implementato con una catena di 50mila iterazioni, di cui 30mila scartate come fase di burn-in, al fine di garantire la convergenza della catena e la stabilità delle stime.

Per la componente predittiva, è stata adottata la *regressione fattoriale latente*, la cui formulazione teorica è stata presentata nel capitolo 2 sez.I.II. Sono state misurate le performance predittive per tutte e 4 le variabili dipendenti (grasso, saccarosio, farina, acqua).

I.II Risultati

I valori di *Mean Squared Prediction Error* (MSPE) ottenuti mediante lo *Sparse Bayesian Infinite Factor Model* (indicato nella Tabella 3.1 come **S.B.I.F.M.**) sono stati confrontati con i risultati riportati da Brown et al. (2001), relativi a diversi metodi di riferimento ampiamente utilizzati in letteratura per la calibrazione di modelli predittivi su dati spettroscopici.

Nel loro studio, gli autori propongono innanzitutto la *Bayesian Wavelet Regression*, una metodologia basata su una rappresentazione a ondelette dei segnali spettrali, combinata con una media bayesiana sui modelli (*Bayes model averaging*) per integrare l'incertezza nella selezione dei coefficienti più rilevanti. L'approccio, implementato considerando i 500 modelli a maggiore probabilità a posteriori, mostra prestazioni sensibilmente superiori rispetto

alle metodologie standard, costituendo il principale contributo innovativo del lavoro di Brown et al. (2001).

A fini di confronto, gli autori includono inoltre una serie di metodi convenzionali, tra cui la *Bayesian Decision Theory* Brown et al. (1999) per la selezione congiunta di un numero ridotto di lunghezze d'onda, con l'obiettivo di ottimizzare simultaneamente la predizione dei quattro costituenti principali (grassi, saccarosio, farina e acqua). Accanto a tali approcci bayesiani, vengono riportati i risultati di metodologie deterministiche classiche, tra cui le calibrazioni derivate da Osborne et al. (1984), ottenute mediante regressione multipla con procedura stepwise per la selezione delle lunghezze d'onda più informative per ciascun costituente chimico. A questi si affiancano i metodi di regressione lineare multivariata basati sulla *Principal Component Regression* (PCR) Massy (1965) e sulla *Partial Least Squares* (PLS) Cha (1994), che rappresentano strumenti standard per la riduzione dimensionale e la modellazione di dati altamente collineari.

A fini comparativi, nel presente lavoro tali risultati sono stati affiancati a quelli ottenuti mediante una *Ridge Regression* Massy (1965), implementata ad hoc per fornire un ulteriore benchmark di tipo regolarizzato, e al modello proposto, lo *Sparse Bayesian Infinite Factor Model*. Gli MSPE relativi a tutti i metodi considerati sono riportati nella Tabella 3.1.

Nel confronto quantitativo tra i risultati ottenuti mediante lo *Sparse Bayesian Infinite Factor Model* e quelli riportati da Brown et al. (2001), emerge una chiara superiorità del modello proposto in termini di accuratezza predittiva per la maggior parte delle variabili considerate. In particolare, i valori di *Mean Square Prediction Error* (MSPE) risultano sensibilmente inferiori rispetto ai metodi di riferimento tradizionali.

Per la variabile *Grassi*, il modello bayesiano ha fornito un errore medio quadratico pari a 0.016, inferiore sia al valore ottenuto con la regressione MLR stepwise (0.044), sia a quello derivante dalla Ridge Regression (0.020), che rappresenta già un benchmark competitivo. Analogamente, per la variabile *Saccarosio* si osserva un MSPE di 0.024, nettamente più basso rispetto a tutti gli altri approcci, incluso il modello Ridge (0.059).

Per quanto riguarda invece *Farina* e *Acqua*, i valori di errore (rispettivamente 0.073 e 0.058) risultano comunque più bassi rispetto a tutti i metodi riportati da Brown, confermando la buona capacità predittiva del modello; tuttavia, la Ridge Regression mostra prestazioni leggermente superiori (0.068 per *Farina* e 0.052 per *Acqua*), evidenziando come, per tali variabili, la regolarizzazione di tipo ridge continui a offrire un vantaggio marginale in termini di accuratezza.

Tali risultati trovano conferma nelle figure 3.2, 3.3, 3.4, e 3.5 che mostrano una forte aderenza tra i valori osservati e quelli predetti dal modello, con una distribuzione dei punti prossima alla bisettrice dell'angolo primo. Ciò testimonia l'elevata capacità dello *Sparse Bayesian Infinite Factor Model* di riprodurre accuratamente le relazioni funzionali sottostanti alle diverse variabili dipendenti, fornendo una rappresentazione parsimoniosa e al tempo stesso robusta della struttura spettrale. Nel complesso, il modello proposto raggiunge prestazioni predittive superiori rispetto ai metodi classici e baye-

Tabella 3.1: Confronto del *Mean Squared Prediction Error* (MSPE) sui 39 campioni di pasta di biscotti nel set di validazione. La prima sezione riporta i risultati ottenuti dai metodi tratti da Brown et al. (2001), mentre la seconda mostra i valori relativi ai modelli implementati, incluso del modello proposto.

MSPE dei modelli tratti da Brown et al.				
Metodo	Grassi	Saccarosio	Farina	Acqua
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
B. wavelet regr. (500 mod.)	0.063	0.449	0.348	0.050
Miglior B. wavelet regr.	0.059	0.466	0.351	0.047
MSPE dei modelli implementati nel presente lavoro				
Ridge	0.020	0.059	0.068	0.052
S.B.I.F.M.	0.016	0.024	0.073	0.058

siani di riferimento, avvicinandosi ai risultati della Ridge Regression nelle variabili più difficili da modellare.

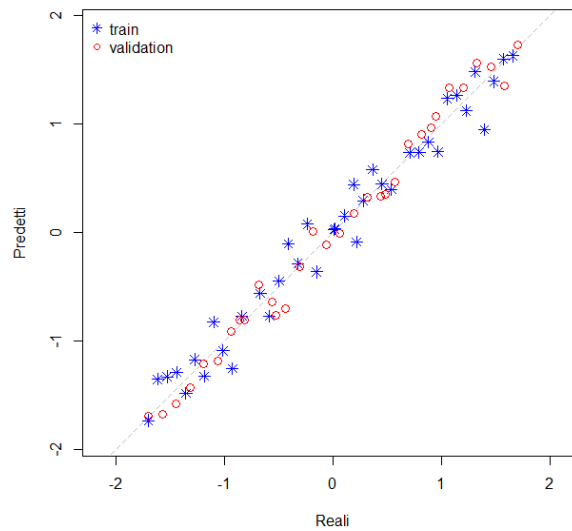


Figura 3.2: Confronto tra valori predetti e valori reali per la variabile Grassi ottenuti con il metodo *Sparse Bayesian Infinite Factor Model*, cap. 2.

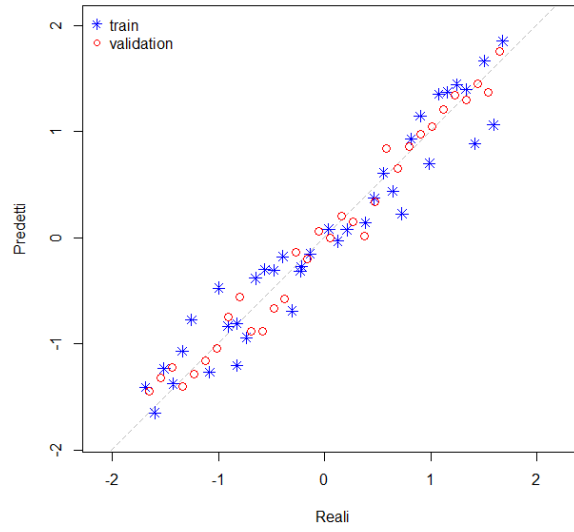


Figura 3.3: Confronto tra valori predetti e valori reali per la variabile Saccarosio ottenuti con il metodo *Sparse Bayesian Infinite Factor Model*, cap. 2.

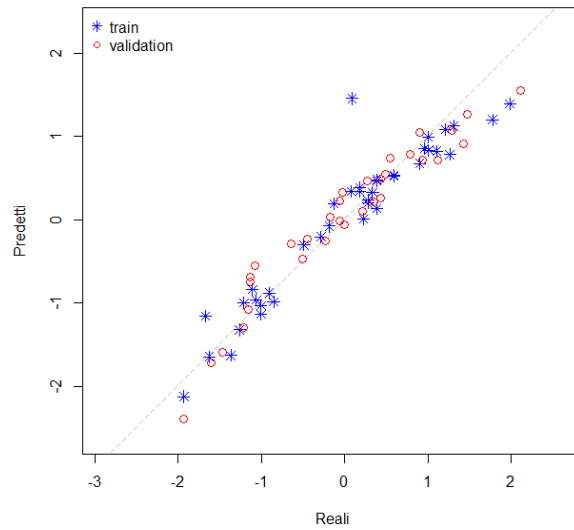


Figura 3.4: Confronto tra valori predetti e valori reali per la variabile Farina ottenuti con il metodo *Sparse Bayesian Infinite Factor Model*, cap. 2.

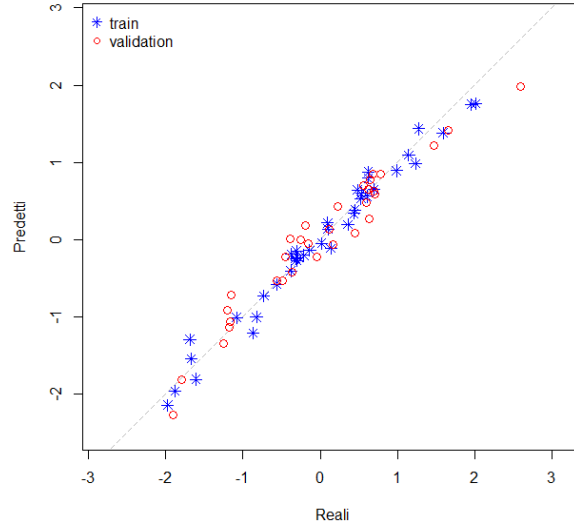


Figura 3.5: Confronto tra valori predetti e valori reali per la variabile Acqua ottenuti con il metodo *Sparse Bayesian Infinite Factor Model*, cap. 2.

Oltre al confronto in termini di prestazioni predittive, è stata condotta un'ulteriore analisi volta a esaminare la coerenza delle strutture dei coefficienti stimati. In particolare, si è verificato se le lunghezze d'onda associate ai coefficienti β di maggiore valore assoluto, quindi maggiormente informative, coincidessero con le bande rilevanti individuate nei modelli di riferimento riportati da Brown et al. (2001). A tal fine, per ogni variabile dipendente si è considerato il vettore dei coefficienti PLS presentato nello studio originale e riportato in Figura, confrontandolo con i coefficienti ottenuti mediante lo *Sparse Bayesian Infinite Factor Model*, ricavati a partire dalla matrice di varianza-covarianza stimata e calcolati come $\beta = \Sigma_{ww}^{-1} \Sigma_{wz}$. Il confronto qualitativo tra i due insiemi di coefficienti ha consentito di verificare la coerenza tra le regioni spettrali più informative.

L'analisi condotta sui coefficienti stimati ha messo in evidenza un'elevata coerenza tra le regioni spettrali più informative individuate dallo *Sparse Bayesian Infinite Factor Model* e quelle riportate nei modelli di riferimento di Brown. In particolare, per la variabile *Grassi*, le lunghezze d'onda corrispondenti ai coefficienti β di maggiore valore assoluto — sia positivi sia negativi — risultano concentrate nell'intorno dei 1700 nm, dove il modello evidenzia variazioni marcate dell'intensità spettrale. Tali andamenti, coerenti con le strutture osservate nel modello PLS di riferimento e anche con l'applicazione della SVD regression effettuata West (Bernardo et al. (2003)), indicano che le principali componenti informative vengono catturate in modo analogo anche dal modello bayesiano. Si osservano inoltre regioni con coefficienti negativi, che suggeriscono una correlazione inversa tra l'assorbanza locale e il contenuto di grassi, a conferma della capacità del modello di rappresentare relazioni complesse e bilanciate tra le variabili spettrali (figura 3.6).

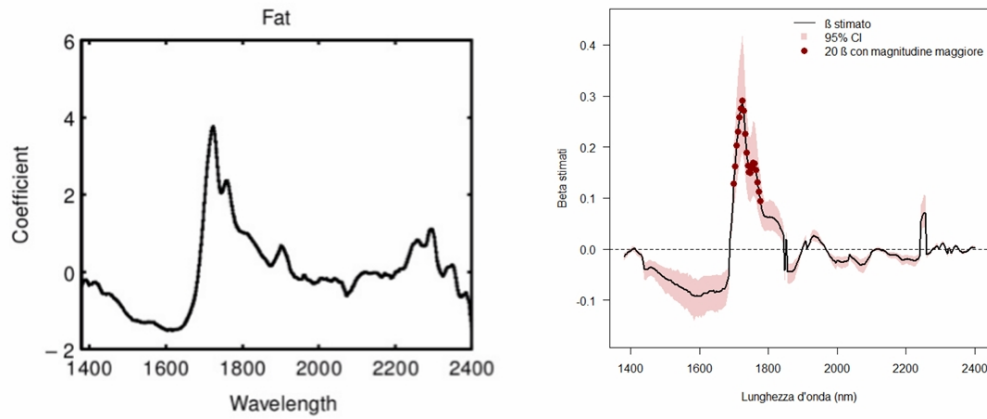


Figura 3.6: Per la variabile Grassi: coefficienti stimati con modello PLS tratti da Brown et al. (2001) (sinistra) vs coefficienti stimati mediante lo *Sparse Bayesian Infinite Factor Model*, con intervalli di credibilità al 95% (fascia rossa) (destra), lungo le lunghezze d'onda.

Per la variabile *Saccarosio*, i coefficienti stimati mostrano un picco positivo pronunciato intorno ai 2100 nm (Figura 3.7), accompagnato da una variazione negativa tra 1800 nm e 1900 nm. Questa alternanza di segni nei coefficienti evidenzia regioni spettrali a contributo opposto, ma entrambe significative nella descrizione del segnale. L'andamento complessivo risulta coerente con il profilo del modello PLS di Brown (Figura 3.7), mostrando come il modello bayesiano riesca a identificare in modo consistente le stesse porzioni di spettro rilevanti per la previsione del contenuto zuccherino.

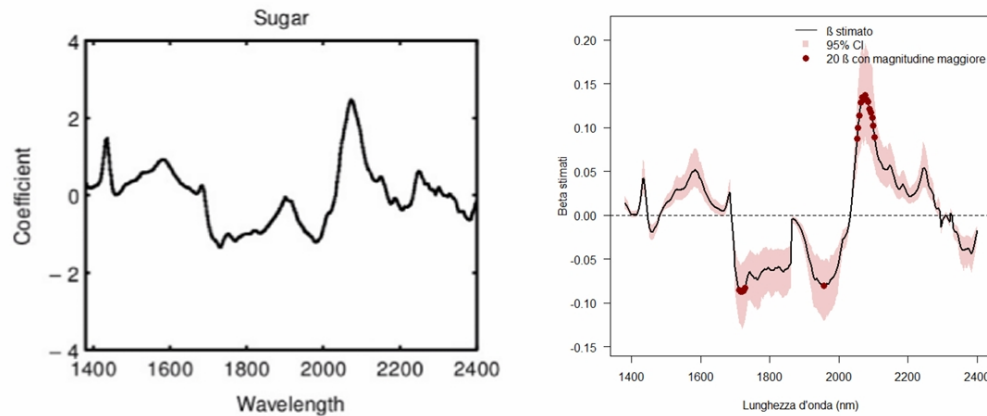


Figura 3.7: Per la variabile Saccarosio: coefficienti stimati con modello PLS tratti da Brown et al. (2001) (sinistra) vs coefficienti stimati mediante lo *Sparse Bayesian Infinite Factor Model*, con intervalli di credibilità al 95% (fascia rossa) (destra), lungo le lunghezze d'onda.

Nel caso della variabile *Farina*, i coefficienti β presentano valori positivi marcati tra 1900 nm e 2000 nm e un andamento negativo esteso oltre i 2100 nm (Figura 3.8). Le due aree contribuiscono con segno opposto alla

stima, ma risultano entrambe informative, delineando un profilo spettrale che riprende fedelmente la forma osservata nel modello PLS di riferimento (Figura 3.8). Ciò suggerisce che lo *Sparse Bayesian Infinite Factor Model* sia in grado di distinguere regioni che aumentano o riducono il contributo alla variabile di risposta, mantenendo una chiara interpretabilità qualitativa delle relazioni osservate.

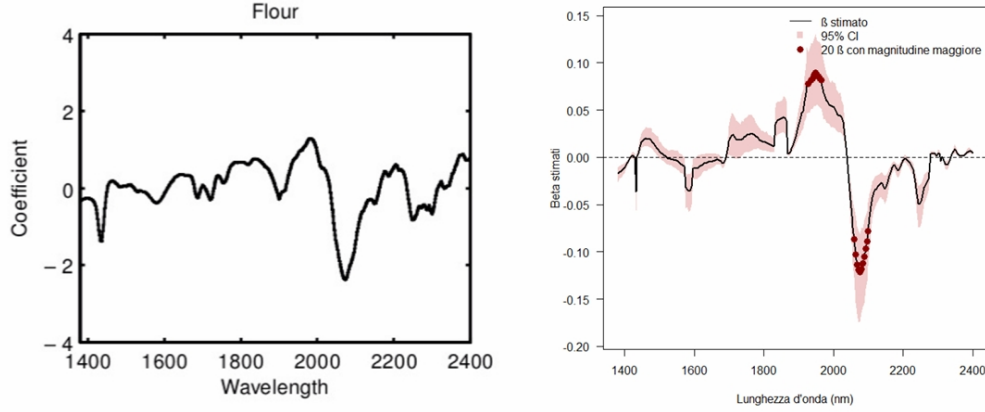


Figura 3.8: Per la variabile Farina: coefficienti stimati con modello PLS tratti da Brown et al. (2001) (sinistra) vs coefficienti stimati mediante lo *Sparse Bayesian Infinite Factor Model*, con intervalli di credibilità al 95% (fascia rossa) (destra), lungo le lunghezze d'onda.

Per la variabile *Acqua*, la struttura dei coefficienti stimati mostra un picco positivo dominante intorno ai 1450 nm e un secondo massimo più ampio presso i 1950 nm, con zone negative intermedie (Figura 3.9). Tale configurazione riproduce con buona fedeltà il profilo riportato nel modello PLS di Brown (Figura 3.9), confermando la capacità del modello bayesiano di individuare non solo le regioni di massimo contributo, ma anche quelle a effetto compensativo, che riflettono variazioni inverse nella risposta spettrale.

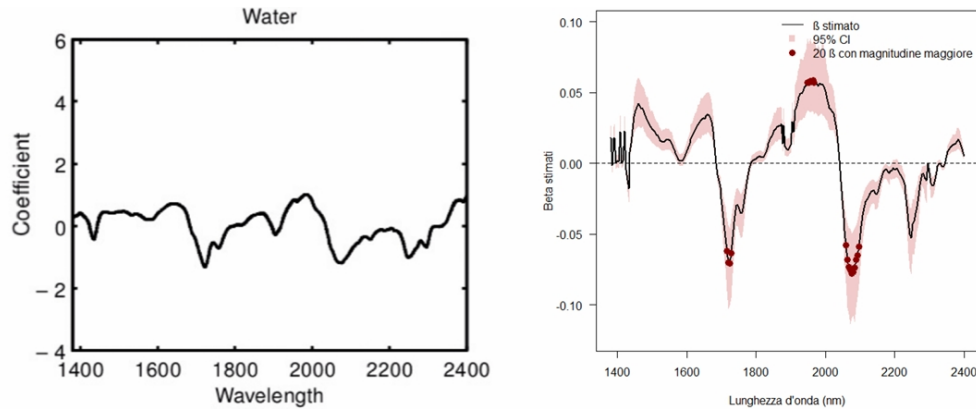


Figura 3.9: Per la variabile Acqua: coefficienti stimati con modello PLS tratti da Brown et al. (2001) (sinistra) vs coefficienti stimati mediante lo *Sparse Bayesian Infinite Factor Model*, con intervalli di credibilità al 95% (fascia rossa) (destra), lungo le lunghezze d'onda.

Nel complesso, le evidenze emerse mostrano che la struttura inferenziale proposta non solo garantisce una stima accurata dei coefficienti, ma consente anche una lettura coerente e bilanciata delle relazioni spettrali. La presenza di picchi positivi e negativi di pari rilevanza indica che il modello è in grado di cogliere le interdipendenze tra le diverse regioni dello spettro, mantenendo al contempo una chiara interpretabilità fisico-spettrale in linea con i risultati consolidati della letteratura di riferimento.

Diagnosi di convergenza Al fine di arricchire l’analisi e di fornire un riscontro empirico alla solidità dei risultati ottenuti, è stata condotta una valutazione approfondita della convergenza della catena MCMC. Analogamente a quanto realizzato nel cap. 2, sono stati esaminati l’andamento della *log-likelihood* e della *log-posterior*, in modo complementare al comportamento del numero effettivo di fattori k (descritto nella Sezione II). Tale analisi congiunta consente di verificare la stabilizzazione complessiva del processo di campionamento e di confermare la coerenza dei risultati inferenziali ottenuti.

L’analisi dei *traceplot* riportati nelle Figure 3.10–3.11, ottenuti a partire dalla catena MCMC eseguita sui dati aventi come variabile dipendente lo *Saccarosio*, evidenzia un comportamento complessivamente stabile del processo di campionamento (per ulteriori diagnosi relative alle altre variabili dipendenti e warm start consultare appendice III.I). In particolare, il *traceplot* della log-verosimiglianza mostra una rapida fase iniziale di adattamento seguita da un andamento stazionario attorno a un valore medio costante, indicativo del raggiungimento della convergenza e di una buona esplorazione dello spazio parametrico.

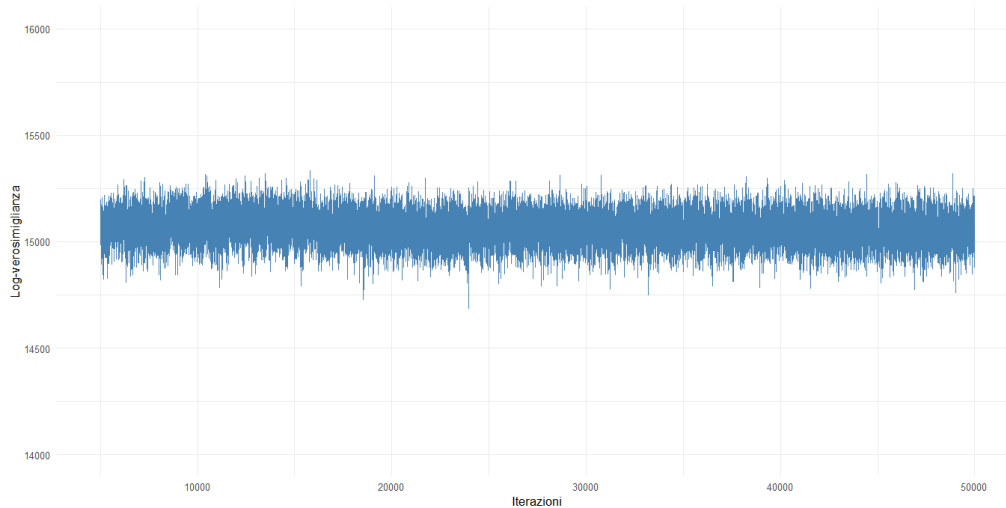


Figura 3.10: Traceplot della log-verosimiglianza lungo le iterazioni della catena MCMC con Saccarosio come variabile dipendente.

Il comportamento della log-posterior, pur risultando più sensibile alle variazioni del numero di fattori k , manifesta un’evoluzione coerente con quella

della log-verosimiglianza: le oscillazioni osservate nelle prime iterazioni riflettono infatti le fluttuazioni associate alla selezione del numero effettivo di fattori, come illustrato in Figura 3.12. Una volta stabilizzato il valore di k , anche la log-posterior tende a convergere verso un regime stazionario, confermando la stabilità del campionamento e la consistenza dell'inferenza bayesiana.

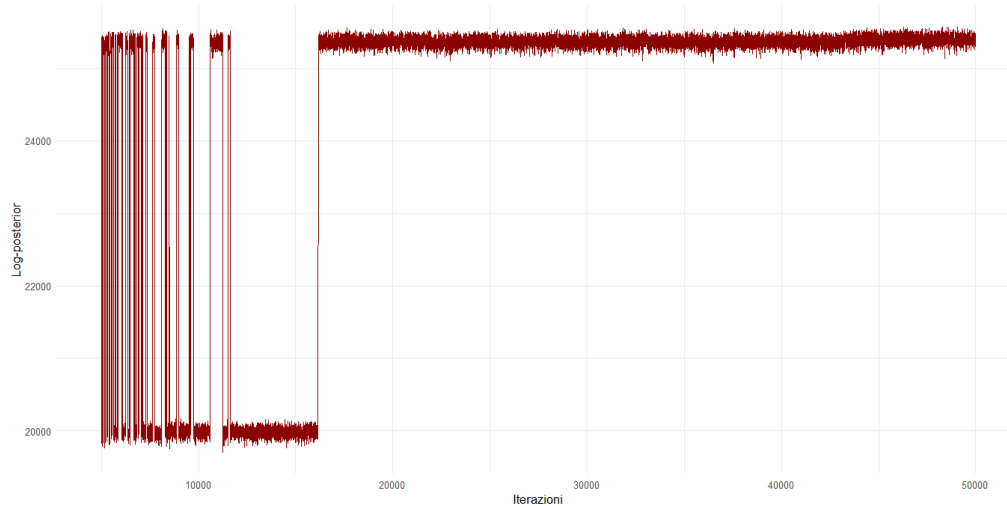


Figura 3.11: Traceplot della log-posterior (in forma aperta) lungo le iterazioni della catena MCMC con Saccarosio come variabile dipendente.

Infine, l'andamento del numero di fattori stimato mostra un iniziale periodo di variabilità seguito da una rapida stabilizzazione attorno a un valore costante, segno che il modello ha identificato in modo robusto la dimensionalità latente più appropriata per la variabile considerata. Nel complesso, l'esame congiunto dei tre *traceplot* suggerisce che la catena MCMC abbia raggiunto un soddisfacente livello di convergenza, garantendo l'affidabilità dei risultati stimati e la solidità delle inferenze effettuate.

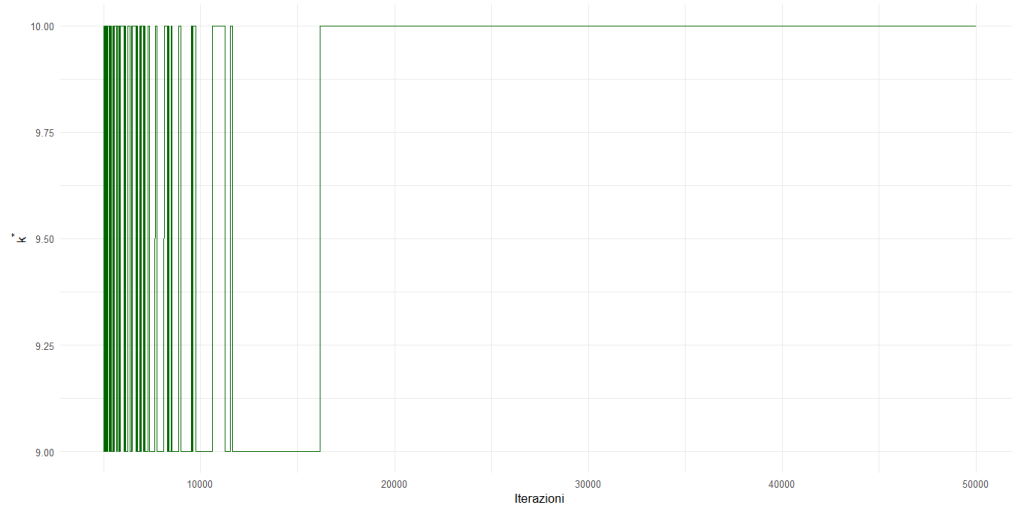


Figura 3.12: Andamento del numero di fattori k stimato lungo la catena MCMC con Saccarosio come variabile dipendente.

II Applicazione ai dati NIRS sul grano

La *genomic selection* rappresenta uno strumento di grande efficacia per il miglioramento genetico di piante e animali, poiché si fonda sulla predizione dei valori genetici tramite l'analisi di polimorfismi del DNA. Tuttavia, i costi elevati associati a questa metodologia hanno stimolato la ricerca di approcci alternativi. Tra questi si colloca la *phenomic selection*, che si basa sull'utilizzo di endofenotipi capaci di catturare le reti di regolazione che connettono genotipo e fenotipo; un modo per accedere a tali informazioni è fornito dalla spettroscopia nel vicino infrarosso (NIRS).

In termini biologici, il genotipo rappresenta l'insieme delle informazioni genetiche codificate nel DNA di un organismo, mentre il fenotipo ne descrive l'insieme delle caratteristiche osservabili, risultanti dall'interazione tra genotipo e ambiente. Gli endofenotipi si collocano in una posizione intermedia tra i due livelli: si tratta di tratti quantitativi interni, di natura biochimica, metabolica o strutturale, che riflettono più direttamente l'attività delle reti genetiche e dei processi fisiologici. L'impiego di misure endofenotipiche, come gli spettri NIRS, consente quindi di esplorare in modo indiretto ma informativo la relazione tra variabilità genetica e manifestazione fenotipica.

I dati considerati in questa sezione provengono dal lavoro di Rincet et al. (2018), in cui la *phenomic selection* è stata applicata a spettri NIRS di campioni di grano. La coltura del grano riveste un ruolo centrale per la sicurezza alimentare a livello globale, e il miglioramento genetico rappresenta uno strumento cruciale per incrementarne la resa e la resilienza in contesti agricoli sempre più complessi. L'uso della NIRS su questa specie consente di ottenere informazioni indirette ma estremamente ricche sul contenuto biochimico dei chicchi, favorendo così l'identificazione di linee varietali con caratteristiche desiderabili.

Questi dati, caratterizzati da un numero molto elevato di variabili corrispondenti alle assorbanze a differenti lunghezze d'onda, offrono un terreno ideale per valutare metodi statistici avanzati, quali lo *Sparse Bayesian Infinite Factor Model*, capaci di gestire strutture dati ad alta dimensionalità e di individuare rappresentazioni latenti parsimoniose dei segnali spettrali.

II.I Dataset, pre-processing e metodo

Il dataset considerato comprende spettri di assorbanza nel vicino infrarosso (NIRS) acquisiti su campioni di grano appartenenti alla condizione di *stress idrico controllato* (*dry*), raccolti nell'ambito di una prova sperimentale condotta a Clermont-Ferrand (Francia) durante la stagione 2015/2016 Rincet (2018). In tale trattamento, le piante sono state coltivate sotto *rain-out shelters*, strutture progettate per escludere la pioggia naturale e simulare una condizione di siccità, mentre una parcella adiacente veniva mantenuta irrigata (condizione *IRR*).

Complessivamente, il dataset includeva 228 varietà di frumento, di cui 5 osservazioni sono state rimosse a causa di anomalie tecniche nei segnali di assorbanza, portando a un pannello finale di 223 varietà.

I dati spettrali sono costituiti da 1050 predittori, corrispondenti ai valori di assorbanza misurati nell'intervallo 400–2500 nm con passo di 2 nm, centrati mostrati nella figura 3.13. Le trasformazioni preliminari sono state effettuate secondo quanto descritto in Rincet et al. (2018): in particolare, gli spettri sono stati centrati, standardizzati e trasformati mediante filtro di Savitzky–Golay, al fine di ridurre il rumore strumentale e migliorare la risoluzione delle principali bande spettrali. Per ulteriori dettagli relativi alla strumentazione e alle procedure di acquisizione, si rimanda al lavoro originale.

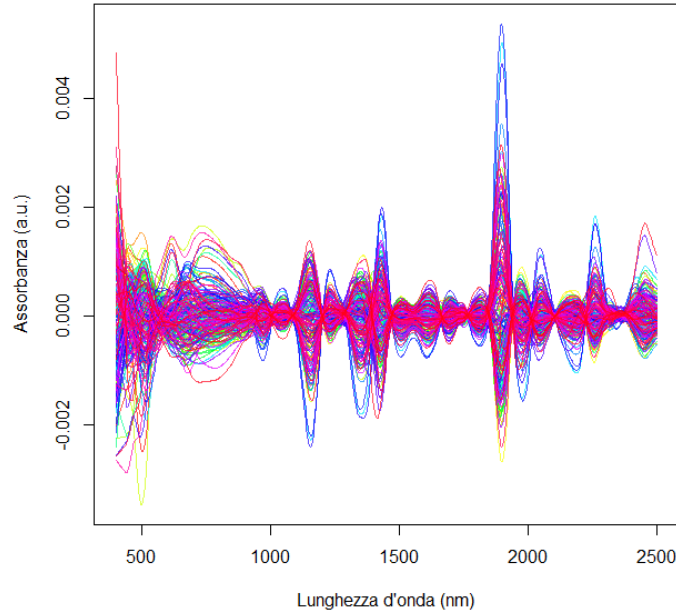


Figura 3.13: Spettri NIR di assorbanza (centrati dopo le trasformazioni originali di Rincient et al. (2018)) relativi a 223 campioni di grano sottoposti a stress idrico controllato (*dry*).

La scelta di utilizzare esclusivamente la condizione *dry* è motivata dal fatto che, come evidenziato da Rincient et al. (2018), tale trattamento rappresenta la condizione sperimentale in cui il fenotipo risulta più accuratamente spiegato dai segnali NIRS, grazie alla maggiore stabilità chimico-fisica del campione e alla ridotta influenza di fattori ambientali.

Nel presente studio, la variabile fenotipica di risposta considerata è il rendimento del grano (*grain yield*, GY), riferita ai medesimi campioni della condizione *dry*. I dati fenotipici sono stati corretti per gli effetti ambientali e spaziali secondo le procedure descritte in Rincient et al. (2018). Infine, sia i predittori spettrali sia la variabile di risposta sono stati normalizzati (centrati e standardizzati), al fine di garantire comparabilità e stabilità numerica nelle analisi successive.

L'analisi dei dati NIRS relativi ai campioni di grano è stata condotta mediante l'applicazione dello *Sparse Bayesian Infinite Factor Model*, sez. (I). L'inferenza sui parametri del modello è stata realizzata attraverso campionamento MCMC (3), eseguendo una catena di 50.000 iterazioni, delle quali le prime 30.000 sono state eliminate come fase di burn-in, così da assicurare la convergenza del processo e la stabilità delle stime a posteriori.

Per la parte predittiva è stato impiegato il framework della *regressione fattoriale latente*, descritto nella sez. I.II.

II.II Risultati

Al fine di valutare le prestazioni predittive dello *Sparse Bayesian Infinite Factor Model*, i risultati ottenuti sono stati confrontati con quelli di una regressione *ridge*, adottata come termine di riferimento in quanto impiegata da Rincent et al. (2018) nello studio originale sui medesimi dati NIRS. Per garantire una comparabilità diretta tra i due approcci, si è scelto di non replicare integralmente la procedura di suddivisione e di cross-validation adottata da Rincent et al., ma di effettuare una separazione semplice del dataset, assegnando il 52% delle osservazioni al campione di addestramento e il restante 48% a quello di test. Tale scelta ha consentito di semplificare la valutazione comparativa mantenendo la coerenza metodologica tra i due modelli.

I risultati ottenuti dallo *Sparse Bayesian Infinite Factor Model* applicato ai dati NIRS del grano hanno mostrato performance predittive complessivamente modeste. L'errore quadratico medio di predizione (*Mean Squared Prediction Error*, MSPE) risulta pari a 0.81. figura 3.14c Sebbene i valori di MSPE non siano soddisfacenti, un confronto diretto con la regressione *ridge* — adottata come modello di riferimento — evidenzia prestazioni migliori per quest'ultima (con un MSPE di 0.63, figura 3.14a), suggerendo che la maggiore complessità del modello fattoriale non si traduce, in questo caso, in un effettivo vantaggio predittivo.

Per verificare la robustezza dei risultati e tentare di migliorare la capacità predittiva del modello bayesiano, sono state sperimentate diverse strategie, nessuna delle quali ha tuttavia condotto a un sostanziale miglioramento delle performance.

In primo luogo, la catena MCMC è stata estesa da 50.000 a 100.000 iterazioni, con l'obiettivo di verificare se le prestazioni inferiori potessero essere attribuite a una lunghezza insufficiente della catena. Tuttavia, l'aumento del numero di iterazioni non ha comportato variazioni significative né nei valori delle log-verosimiglianze né nelle prestazioni predittive, suggerendo che il problema non fosse riconducibile alla durata della catena stessa.

In secondo luogo, sono state applicate differenti trasformazioni ai dati di input. Considerata la relazione tra assorbanza (A) e riflettanza (R) (Workman and Weyer (2012)),

$$A = \log_{10} \left(\frac{1}{R} \right) = -\log_{10}(R),$$

si è testata la modellizzazione diretta in termini di riflettanza, al fine di valutare se la diversa scala di misura potesse favorire una maggiore linearità nelle relazioni latenti o una migliore stabilità numerica. Anche in questo caso, tuttavia, i risultati ottenuti non hanno mostrato variazioni significative nei valori di MSPE.

Sono inoltre state sperimentate trasformazioni di normalizzazione e standardizzazione più spinte, in coerenza con l'assunzione di normalità congiunta tra la variabile dipendente \mathbf{z}_i e i predittori \mathbf{w}_i implicita nella regressione

fattoriale latente (I.II). Tale assunzione è formalmente espressa come:

$$\begin{pmatrix} \mathbf{z}_i \\ \mathbf{w}_i \end{pmatrix} \sim \mathcal{N}_p(\mathbf{0}, \mathbf{\Lambda}\mathbf{\Lambda}^\top + \mathbf{\Psi}),$$

dove \mathbf{z}_i rappresenta la variabile dipendente per i ($i = 1, \dots, n$), \mathbf{w}_i il vettore dei predittori NIRS, $\mathbf{\Lambda}$ la matrice dei carichi fattoriali, e $\mathbf{\Psi}$ la matrice diagonale delle varianze idiosincratice. Anche in questo caso, tuttavia, le diverse trasformazioni non hanno determinato miglioramenti apprezzabili nelle prestazioni predittive.

Infine, considerata la persistente difficoltà del modello fattoriale bayesiano nel catturare adeguatamente la variabilità dei dati NIRS, si è ipotizzato che la relazione tra le variabili potesse non essere di natura lineare. Poiché lo Sparse Bayesian Infinite Factor Model assume una struttura lineare tra le variabili latenti e osservate, tale caratteristica potrebbe limitare la capacità del modello di rappresentare relazioni più complesse.

Per tale motivo, è stato valutato un approccio alternativo basato su BART (Bayesian Additive Regression Trees) Chipman et al. (2010), un metodo non parametrico di tipo bayesiano in grado di modellare efficacemente relazioni non lineari e interazioni complesse tra i predittori. Tuttavia, nonostante un lieve miglioramento rispetto all'Infinite Factor Model, le prestazioni predittive ottenute si sono rivelate ancora insoddisfacenti (MSPE=0.76, figura 3.14b), risultando complessivamente inferiori a quelle fornite dal modello di riferimento lineare ridge regression.

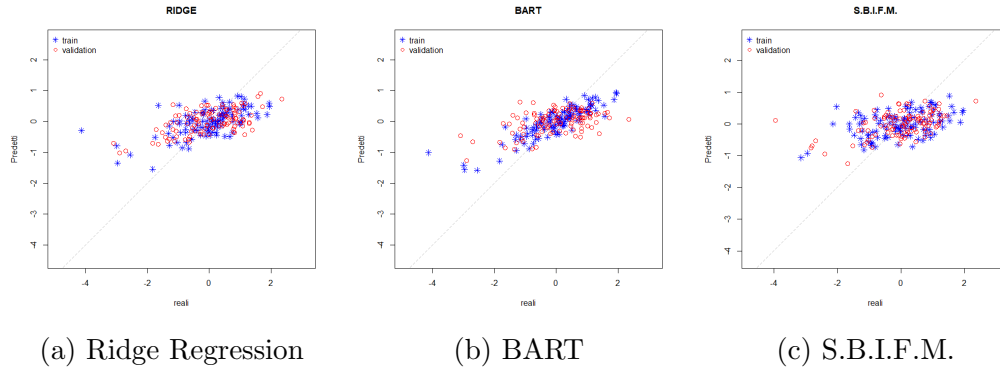


Figura 3.14: Confronto tra valori predetti e valori reali per la variabile *Grain Yield* ottenuti con i tre modelli: Ridge, BART e Sparse Bayesian Infinite Factor Model (S.B.I.F.M.)

A completamento delle analisi, è stata inoltre considerata una regressione KNN (K-Nearest Neighbors) Altman (1992), anch'essa in grado di catturare relazioni non lineari tra i predittori e la variabile risposta. Anche in questo caso, tuttavia, i risultati non sono risultati incoraggianti (MSPE=0.89): le capacità predittive si sono mantenute su livelli inferiori rispetto a quelli dell'Infinite Factor Model, e di conseguenza peggiori rispetto a tutti gli altri metodi considerati (raccolti nella tabella 3.2).

Tabella 3.2: Valori del Mean Squared Prediction Error (MSPE) ottenuti con lo *Sparse Bayesian Infinite Factor Model* e con i metodi di confronto BART, KNN Regression e Ridge Regression.

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

Analogamente a quanto discusso nel capitolo dedicato all'applicazione del modello ai dati NIRS relativi all'impasto dei biscotti (cap. 3 sez. I), è stata condotta un'analisi volta a valutare l'interpretabilità delle bande spettrali maggiormente informative rispetto alla lunghezza d'onda. Anche in questo caso si è verificato se i risultati forniti dallo *Sparse Bayesian Infinite Factor Model* e dalla *Ridge Regression* — inizializzati più volte con differenti semi casuali — mostrassero coerenza nella selezione o nella significatività di determinate lunghezze d'onda. Tuttavia, i risultati ottenuti non hanno evidenziato convergenze sostanziali tra i metodi, né indicazioni stabili circa la rilevanza di specifiche regioni spettrali, suggerendo una certa sensibilità dei modelli alle condizioni iniziali e alla complessità intrinseca dei dati (plot consultabili in appendice A.5, A.6).

Nel complesso, le analisi condotte mostrano come l'individuazione di pattern interpretativi stabili nei dati NIRS del grano rimanga una sfida aperta. Nonostante le diverse strategie esplorate, non è emersa una soluzione univoca o semplice che consenta di migliorare simultaneamente le prestazioni predittive e l'interpretabilità dei risultati. In prospettiva, ulteriori sviluppi potrebbero trarre beneficio da un'integrazione dei dati NIRS con informazioni genetiche, strategia in cui i dati spettrali non vengono utilizzati in sostituzione dei marcatori genetici, bensì in combinazione con essi per incrementare la capacità predittiva e la comprensione dei meccanismi biologici sottostanti (già accennata in Rincent et al. (2018)).

Conclusioni

Il presente lavoro ha avuto come obiettivo lo studio e l'applicazione di metodologie di analisi fattoriale in ambito bayesiano, con particolare riferimento allo *Sparse Bayesian Infinite Factor Model* proposto da Bhattacharya and Dunson (2011). Dopo aver introdotto i principi fondamentali dell'analisi fattoriale classica e le basi della statistica bayesiana, è stata posta attenzione sui vantaggi concettuali e computazionali offerti da approcci bayesiani rispetto ai metodi tradizionali. L'analisi ha successivamente condotto all'esame approfondito del modello di Bhattacharya and Dunson (2011), caratterizzato da una struttura gerarchica adattiva in grado di determinare automaticamente il numero effettivo di fattori latenti. L'impiego del *multiplicative gamma process* ha permesso di introdurre un meccanismo di regolarizzazione globale e locale che favorisce la sparsità dei carichi fattoriali, garantendo al contempo una rappresentazione parsimoniosa e interpretabile delle dipendenze tra le variabili osservate.

L'implementazione manuale del modello, sviluppata nel linguaggio R, si è dimostrata efficace ed efficiente dal punto di vista computazionale. Le analisi condotte su dati simulati hanno evidenziato come la procedura di stima proposta sia in grado di riprodurre accuratamente la matrice di varianza-covarianza, restituendo valori di errore medi contenuti e buone capacità predittive. Le diagnosi di convergenza effettuate hanno confermato la stabilità e la correttezza delle catene generate mediante l'algoritmo di *Gibbs sampling* con troncatura adattiva, attestando la bontà complessiva dell'approccio proposto.

Successivamente, il modello è stato applicato a due differenti casi di studio basati su dati reali di spettroscopia nel vicino infrarosso (NIRS), una tecnologia di analisi non invasiva, rapida ed economica, ampiamente impiegata in ambito chimico, alimentare e agricolo per la caratterizzazione qualitativa e quantitativa di campioni complessi. L'impiego della spettroscopia NIRS consente infatti di ottenere, a partire da misure spettrali, stime affidabili delle proprietà chimico-fisiche dei materiali analizzati, riducendo la necessità di analisi distruttive o costose.

In questa prospettiva, sono state considerate due applicazioni empiriche distinte, scelte al fine di valutare le prestazioni del modello su differenti tipologie di dati e contesti sperimentali. La prima applicazione riguarda campioni di impasto di biscotti, utilizzati come banco di prova classico per modelli predittivi in ambito alimentare, poiché i segnali spettrali consentono di stimare con buona precisione le concentrazioni dei principali costituenti

(grassi, zuccheri, farina e acqua). La seconda applicazione, invece, riguarda campioni di grano, per i quali la spettroscopia NIRS rappresenta uno strumento promettente a supporto del miglioramento genetico, permettendo di stimare caratteristiche fenotipiche di interesse agronomico, quali la resa o la qualità del raccolto, in modo rapido e non distruttivo.

L'analisi di questi due casi di studio consente dunque di esplorare la flessibilità del modello nel trattare dati di natura diversa — industriale e agronomica — e di valutarne la capacità di adattamento a contesti sperimentali con differenti livelli di complessità e struttura di correlazione.

La prima applicazione, relativa ai dati NIRS sull'impasto dei biscotti analizzati da Brown et al. (2001), ha avuto l'obiettivo di verificare la capacità predittiva e l'interpretabilità del modello in un contesto ad alta dimensionalità, in cui il numero di variabili spettrali supera ampiamente il numero di osservazioni. I risultati ottenuti mostrano performance predittive molto soddisfacenti, con errori medi inferiori rispetto a metodi di riferimento più classici. Dal punto di vista interpretativo, le lunghezze d'onda individuate come maggiormente informative per ciascuna variabile chimica risultano coerenti con le conclusioni ottenute da Brown et al. (2001) e, più in generale, con la letteratura di riferimento sui legami molecolari caratteristici delle bande NIRS. Tali evidenze confermano la validità dello *Sparse Bayesian Infinite Factor Model* anche in applicazioni empiriche complesse, in cui la struttura dei dati è altamente ridondante e correlata.

Nella seconda applicazione, condotta sui dati NIRS del grano analizzati da Rincet et al. (2018), l'obiettivo è stato quello di valutare la capacità del modello di prevedere una variabile fenotipica osservabile, nello specifico la resa del grano, e di identificare eventuali bande spettrali maggiormente informative. In questo caso, i risultati ottenuti non si sono rivelati pienamente soddisfacenti. Nonostante diversi tentativi di ottimizzazione — tra cui modifiche alle fasi di pre-processing, applicazione di svariati metodi e ripetizioni con differenti condizioni iniziali — le prestazioni predittive del modello sono rimaste inferiori alle attese. Tale comportamento può essere in parte attribuito alla natura stessa dei dati, che presentano una relazione debole e potenzialmente non lineare tra gli spettri NIRS e la variabile fenotipica di interesse. A differenza dei dati relativi all'impasto dei biscotti, in cui la composizione chimica determina segnali spettrali direttamente interpretabili e fortemente correlati ai costituenti analizzati, nel caso del grano le variazioni spettrali legate alla resa risultano più sottili e difficilmente catturabili. Anche l'analisi dell'interpretabilità delle lunghezze d'onda non ha evidenziato pattern stabili o facilmente riconducibili a specifiche regioni dello spettro. Confronti effettuati con altri metodi, tra cui *ridge regression*, *BART* e *k-nearest neighbors regression*, hanno confermato la difficoltà generale del problema: nessuno dei modelli testati è riuscito a ottenere performance predittive soddisfacenti, nemmeno quelli in grado di modellare relazioni non lineari complesse.

Ciononostante, il lavoro svolto rappresenta un contributo significativo alla comprensione dell'applicabilità dei modelli di analisi fattoriale bayesiana a dati NIRS. Le analisi hanno infatti messo in luce sia le potenzialità

che i limiti dello *Sparse Bayesian Infinite Factor Model* in contesti reali, indicando direzioni di sviluppo future. In particolare, un possibile ampliamento delle prospettive di ricerca riguarda l'integrazione dei dati NIRS con informazioni genomiche, come suggerito da Rincent et al. (2018). Un approccio integrato di questo tipo, in cui i dati spettrali non sostituiscono ma completano quelli genetici, potrebbe consentire un miglioramento sia delle capacità predittive sia dell'interpretabilità biologica dei risultati, senza rinunciare ai vantaggi in termini di rapidità e costo dell'analisi NIRS.

In conclusione, il lavoro conferma l'efficacia e la flessibilità dei modelli bayesiani a struttura adattiva per la riduzione dimensionale e la modellazione di dati complessi, mostrando al contempo come la loro applicazione a dati reali rappresenti un terreno di ricerca ancora aperto e ricco di prospettive di approfondimento, sia dal punto di vista metodologico sia applicativo.

Bibliografia

- Altman, N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, 46(3):175–185.
- Bernardo, J., Bayarri, M., Berger, J., Dawid, A., Heckerman, D., Smith, A., and West, M. (2003). Bayesian factor regression models in the “large p, small n” paradigm. *Bayesian statistics*, 7:733–742.
- Bhattacharya, A. and Dunson, D. B. (2011). Sparse bayesian infinite factor models. *Biometrika*, 98(2):291–306.
- Brown, B., Fearn, T., and Vannucci, M. (1999). The choice of variables in multivariate regression: a non-conjugate bayesian decision theory approach. *Biometrika*, 86(3):635–648.
- Brown, P. J., Fearn, T., and Vannucci, M. (2001). Bayesian wavelet regression on curves with application to a spectroscopic calibration problem. *Journal of the American Statistical Association*, 96(454):398–408.
- Cha, J. (1994). Partial least squares. *Adv. Methods Mark. Res.*, 407:52–78.
- Chipman, H. A., George, E. I., and McCulloch, R. E. (2010). Bart: Bayesian additive regression trees.
- Massy, W. F. (1965). Principal components regression in exploratory statistical research. *Journal of the American Statistical Association*, 60(309):234–256.
- Osborne, B. G. and Fearn, T. (1986). *Near infrared spectroscopy in food analysis*. Longman Scientific & Technical New York.
- Osborne, B. G., Fearn, T., Miller, A. R., and Douglas, S. (1984). Application of near infrared reflectance spectroscopy to the compositional analysis of biscuits and biscuit doughs. *Journal of the Science of Food and Agriculture*, 35(1):99–105.
- Rincent, R. (2018). Phenomic selection, Rincent et al.
- Rincent, R., Charpentier, J.-P., Faivre-Rampant, P., Paux, E., Le Gouis, J., Bastien, C., and Segura, V. (2018). Phenomic selection is a low-cost and high-throughput method based on indirect predictions: proof of concept on wheat and poplar. *G3: Genes, Genomes, Genetics*, 8(12):3961–3972.

Rowe, D. B. (2000). A bayesian factor analysis model with generalized prior information. Technical report.

Workman, J. and Weyer, L. (2012). Practical guide and spectral atlas for interpretive near. *Infrared Spectroscopy*, 2.

Appendice

I Codice implementato per la generazione dei dati simulati (come Bhattacharya and Dunson (2011))

```
# Parametri di simulazione
rep <- 1
n   <- 200
N   <- n * rep
p   <- 100 #500 #1000
k   <- 5 #10 #15
#loadings
Lambda <- matrix(0, nrow = p, ncol = k)
# numeff varia tra k+1 e 2*k, in ordine casuale
numeff <- k + sample.int(k, size = k)
for (h in seq_len(k)) {
  # scelgo numeff[h] righe su cui inserire valori non
  # zero
  temp <- sample.int(p, size = numeff[h])
  Lambda[temp, h] <- rnorm(numeff[h], mean = 0, sd =
    1)
}
# 1)ii)PREDIZIONE trasformo la prima rga di lambda (1
# e -1 random, il resto zero)
Lambda[1, ] <- replace(numeric(ncol(Lambda)), sample(
  ncol(Lambda), 2), c(1, -1))
# 2) Matrice di covarianza
mu <- rep(0, p)
Ot <- Lambda %*% t(Lambda) + 0.01 * diag(p)
# 3) Simulo N vettori multivariati
dat <- mvrnorm(n = N, mu = mu, Sigma = Ot)
Lam_true <- Lambda
dim(dat)
```

II Codice implementato per lo studio di simulazione

```

#mixing fubctions
log_posterior <- function(y, Lambda, eta, phi, delta,
  ps,
                        as, bs, df, ad1, bd1, ad2,
                        bd2) {
  # 1) Log-verosimiglianza:  $Y | \Lambda, \eta, \psi \sim N(\eta$ 
     $\%*\% t(\Lambda), \text{diag}(1/\psi))$ 
  mu <- eta  $\%*\% t(\Lambda)$ 
  ll <- sum(dnorm(y, mean = mu, sd = 1/sqrt(ps), log
    = TRUE))

  # 2) Prior su  $\psi$  (Gamma(as, bs))
  lp_ps <- sum(dgamma(ps, shape = as, rate = bs, log =
    TRUE))

  # 3) Prior su  $\Lambda | \phi, \delta$ 
  tauh <- cumprod(delta)
  Sd_inv <- sweep(phi, 2, tauh, "*") # precision
    per colonna
  lp_L <- sum(dnorm(Lambda, mean = 0,
    sd = 1/sqrt(Sd_inv), log = TRUE))

  # 4) Prior su  $\eta \sim N(0, I)$ 
  lp_eta <- sum(dnorm(eta, mean = 0, sd = 1, log = TRUE
    ))

  # 5) Prior su  $\phi$  (Gamma(df/2, df/2))
  lp_phi <- sum(dgamma(phi, shape = df/2, rate = df/2,
    log = TRUE))

  # 6) Prior su  $\delta[1] \sim \text{Gamma}(ad1, bd1)$ 
  lp_d1 <- dgamma(delta[1], shape = ad1, rate = bd1,
    log = TRUE)
  # e  $\delta[h>1] \sim \text{Gamma}(ad2, bd2)$ 
  lp_dh <- sum(dgamma(delta[-1], shape = ad2, rate =
    bd2, log = TRUE))

  return(ll + lp_ps + lp_L + lp_eta + lp_phi + lp_d1 +
    lp_dh)
}
log_likelihood <- function(y, Lambda, eta, ps){
  # 1) Log-verosimiglianza:  $Y | \Lambda, \eta, \psi \sim N(\eta$ 
     $\%*\% t(\Lambda), \text{diag}(1/\psi))$ 
  mu <- eta  $\%*\% t(\Lambda)$ 
  ll <- sum(dnorm(y, mean = mu, sd = 1/sqrt(ps), log
    = TRUE))
  return(ll)
}
n_iter <- 20000

```

```

thin <- 1
burn <- 5000
n_save <- (n_iter-burn)/ thin
      # scegliamo un valore di default come quello del
      paper
kinit <- rep(floor(log(p) * 3), times = rep)
#iperparametri
as <- 1
bs <- 0.3
df <- 3
ad1 <- 2.1
bd1 <- 1
ad2 <- 3.1
bd2 <- 1
alfa0 <- 1
alfa1 <- 0.0005
epsilon <- 1e-3
prop <- 1

mserep <- matrix(0, nrow = rep, ncol = 3) # mse, bias
      assoluto (medio e massimo) nella stima di matrice
      di covarianza
mse1rep <- matrix(0, nrow = rep, ncol = 3) # come sopra,
      ma su scala originale
msperep <- matrix(0, nrow = rep, ncol = 3)
mseBrep <- matrix(0, nrow = rep, ncol = 3)
logpost <- numeric(n_iter)
loglike <- numeric(n_iter)

t0 <- Sys.time()

for (g in 1:rep) {
  cat("start replicate", g, "\n")
  cat("-----\n")

  n_i <- 200
  dat_i <- dat[((g-1) * n_i+ 1):(g * n_i), ] #
      estrazione righe

  #split da mettere per ottenere le predizioni per
  replica
  prop_train <- 0.5
  n_i <- nrow(dat_i)
  n_train <- floor(prop_train * n_i)
  train_i <- sample(seq_len(n_i), size = n_train)
  y <- dat_i[train_i, , drop = FALSE]
  df_test <- dat_i[-train_i, , drop = FALSE]
  df_test <- scale(df_test, center= TRUE, scale = TRUE)

```

```

M <- colMeans(y)
VY <- apply(y, 2, var)
y <- sweep(y, 2, M, "-")
y <- sweep(y, 2, 1 / sqrt(VY), "*")
Ot1 <- Ot * (1 / sqrt( outer(VY, VY) ))

# inizializzo contatore
num <- 0

k_star <- kinit[g]

# --- 2) Inizializzazioni --- #
p <- ncol(y)
n <- nrow(y)

ps <- rgamma(p, shape = as, rate = bs)
Sigma <- diag(1/ps)

Lambda <- matrix(0, nrow = p, ncol = k_star)

eta <- matrix(rnorm(n * k_star), nrow = n, ncol =
  k_star)

phi <- matrix(rgamma(p * k_star,
  shape = df/2,
  rate = df/2),
  nrow = p, ncol = k_star)

delta <- c(rgamma(1, shape = ad1, rate = bd1),
  rgamma(k_star - 1, shape = ad2, rate =
    bd2))

tauh <- cumprod(delta)

Plam <- sweep(phi, 2, tauh, "*")

Lambda_samples <- vector("list", n_save)
sigma_samples <- matrix(NA, nrow = n_save, ncol = p)
k_star_history <- numeric(n_save)
# numero di fattori attraverso le iterazion
mseout <- matrix(0, nrow = n_save, ncol = 3) #
mse1out <- matrix(0, nrow = n_save, ncol = 3)
Omegaout <- numeric(p^2) # vettore di
lunghezza p^2
Omega1out <- numeric(p^2) # v

```

```

#GIBB SAMPLER#
save_index <- 1
for (iter in 1:n_iter) {

  #step sample eta
  Lmsg <- sweep(Lambda, 1, ps, "*")
  Veta1 <- diag(k_star) + crossprod(Lmsg, Lambda)
  U <- chol(Veta1)
  S <- backsolve(U, diag(k_star))
  Veta <- S %>% t(S)
  Meta <- y %>% Lmsg %>% Veta
  eta <- Meta + matrix(rnorm(n * k_star), nrow = n
    , ncol = k_star) %>% t(S)

  #step sample lambda via rue and held
  eta2 <- crossprod(eta)
  for (j in seq_len(p)) {
    Qlam <- diag( Plam[j, ] ) + ps[j] * eta2
    blam <- ps[j] * crossprod(eta, y[, j]) # k
    Llam <- t(chol(Qlam))
    zlam <- rnorm(ncol(eta2))
    vlam <- forwardsolve(Llam, blam) # Llam v
    = blam
    mlam <- backsolve( t(Llam), vlam) # Ll
    ylam <- backsolve( t(Llam), forwardsolve(Llam,
      zlam) )
    Lambda[j, ] <- as.numeric( mlam + ylam )
  }

  # --- 1) Update phi
  -----
  phi <- matrix(
    rgamma(p * k_star, shape = df/2 + 0.5, rate = df
      /2 + sweep(Lambda^2, 2, tauh, "*")/2 ),
    nrow = p, ncol = k_star)

  # --- Update delta e tauh
  -----

  mat <- phi * (Lambda^2) # matrice p k_star

  # --- 1) aggiorno delta[1] ---
  ad1_post <- ad1 + 0.5 * p * k_star
  bd1_post <- bd1 + 0.5 * (1 / delta[1]) * sum(tauh *
    colSums(mat))
  delta[1] <- rgamma(1, shape = ad1_post, rate = bd1_
    post)
  tauh <- cumprod(delta) # aggiorno

```

```

# --- 2) aggiornno delta[h] per h = 2..k_star ---
for (h in 2:k_star) {
  ad_h <- ad2 + 0.5 * p * (k_star - h + 1)
  bd_h <- bd2 + 0.5 * (1 / delta[h]) *
    sum( tauh[h:k_star] * colSums(mat[, h:k_star,
      drop = FALSE]) )
  delta[h] <- rgamma(1, shape = ad_h, rate = bd_h)
  tauh      <- cumprod(delta)    # riaggiorno
    ogni volta
}

# --- 3) Update Sigma (residual precision)
-----
Ytil <- y - eta %% t(Lambda)
ps    <- rgamma(p, shape = as + 0.5 * n, rate =
  bs + 0.5 * colSums(Ytil^2) )
Sigma <- diag(1 / ps)

#update precision parameter
Plam <- sweep(phi,2 , tauh ,"*")

# Calcola e salva la log-posterior
logpost[iter] <- log_posterior(
  y, Lambda, eta, phi, delta, ps,
  as, bs, df, ad1, bd1, ad2, bd2
)

# Calcola e salva la log likelihood
loglike[iter] <- log_likelihood(
  y, Lambda,eta,ps)

#parametri per k adattivo
rho_t <- 1/ exp(alfa0 + alfa1 * iter)
uu <- runif(1)
lind <- colSums(abs(Lambda) < epsilon) / p    #
  proporzione di zero per colonna
vec  <- lind >= prop    #TRUE = colonna con zeri
num  <- sum(vec)

# k adattiva
if (iter > 20 && uu < rho_t) {
  if (num == 0 && all(lind < 0.995)) {
    k_star <- k_star + 1
    Lambda <- cbind(Lambda, rep(0, p))
    eta    <- cbind(eta,    rnorm(n))
    phi    <- cbind(phi,    rgamma(p, shape = df/
      2, rate = df/2))
    delta <- c(delta, rgamma(1,ad2,bd2))
    tauh <-exp(cumsum(log (delta)))
    Plam <- sweep(phi,2 , tauh ,"*")
  }
}

```

```

    }
    else if (num > 0) {
      nonzero <- which(!vec)
      k_star <- max(length(nonzero), 1)
      Lambda <- Lambda[, nonzero, drop = FALSE]
      eta <- eta[, nonzero, drop = FALSE]
      phi <- phi[, nonzero, drop = FALSE]
      delta <- delta[ nonzero]
      tauh <- exp(cumsum(log(delta)))
      Plam <- sweep(phi, 2, tauh, "*")
    }
  }

# Salvataggio dei campioni

if (iter %% thin == 0 && iter > burn) {
  Lambda_samples[[save_index]] <- Lambda
  sigma_samples[save_index, ] <- diag(Sigma)
  k_star_history[save_index] <- k_star
  save_index <- save_index + 1
  Omega <- Lambda %*% t(Lambda) + Sigma
  Omega1 <- Omega * sqrt(outer(VY, VY))
  Omegaout <- Omegaout + as.vector(Omega) / n_
    save
  Omega1out <- Omega1out + as.vector(Omega1) / n_
    save
}

if (iter %% 1000 == 0) {
  cat("Iterazione:", iter, "k_star =", k_star, "\n"
    )
}
}

# ---- misure di sintesi specifiche per la replica
----

#1. mse on omega
{
  errcov <- Omegaout - as.vector(0t1)
  err1cov <- Omega1out - as.vector(0t)
  mserep[g, ] <- c(mean(errcov^2), mean(abs(errcov)),
    max(abs(errcov)))
  mse1rep[g, ] <- c(mean(err1cov^2), mean(abs(err1cov))
    , max(abs(err1cov)))
}

#2. mse on prediction
{
  #1. divisione in dependent e predictor

```

```

y_train <- y[,1]
x_train <- y[,-1]
y_test <- df_test[,1]
x_test <- df_test[,,-1]
#ii. calcolo dei beta
T      <- length(Lambda_samples)
step   <- 50
sel    <- seq(1, T, by = step)
Beta_samples <- matrix(NA, nrow = T/step, ncol = p
-1)

pb <- txtProgressBar(min = 0, max = T/step, style =
3)
for (i in seq_along(sel)){
  Lambda_t   <- Lambda_samples[[i]]      # p      k_t
  psi_t      <- sigma_samples[i, ]        # length p
  Omega_t    <- Lambda_t %*% t(Lambda_t) + diag(psi_t
)
  omega_xz_t <- Omega_t[2:p, 1]
  omega_xx_t <- Omega_t[-1,-1]
  Beta_samples[i, ] <- as.vector(solve(omega_xx_t)
%*% omega_xz_t)
  setTxtProgressBar(pb, i) }

close(pb)
beta_map <- colMeans(Beta_samples)

#iii.calcolo predizioni e salvataggio mspe aape
mape
pred_test <- Beta_samples %*% t(x_test)
mean_pred_test <- colMeans(pred_test)

errs <- y_test-mean_pred_test
msperep[g, ] <- c(mean(errs^2), mean(abs(errs)),
max(abs(errs)))
}

#3. mse on beta
{
  #beta TRUE
  Oxx <- Ot[-1,-1]
  Ozx <- Ot[2:p,1]
  beta_true <- solve(Oxx) %*% Ozx

  errb <- beta_map - beta_true

  mseBrep[g,] <- c(mean(errb^2),mean(abs(errb)), max(
abs(errb)))
}

```

```

cat("end replicate", g, "\n")
cat("-----\n")
}

t <- Sys.time()-t0
t

#riassunti x replica
mse <- c(mean(mserep[,1]), min(mserep[,1]),max(mserep
[,1]))*10^2
avgbias <- c(mean(mserep[,2]), min(mserep[,2]),max(
mserep[,2]))*10^2
maxbias<- c(mean(mserep[,3]), min(mserep[,3]),max(
mserep[,3]))*10^2
data.frame(mse = mse, avgbias = avgbias, maxbias =
maxbias)

mspe <- c(mean(msperep[,1]), min(msperep[,1]),max(
msperep[,1]))
aape <- c(mean(msperep[,2]), min(msperep[,2]),max(
msperep[,2]))
mape <- c(mean(msperep[,3]), min(msperep[,3]),max(
msperep[,3]))
data.frame(mspe = mspe, aape = aape, mape = mape)

mseb <- c(mean(mseBrep[,1]))*10^3
aab <-c(mean(mseBrep[,2]))*10^3
mab<-c(mean(mseBrep[,3]))*10^3
data.frame(mseb = mseb, aab = aab, mab = mab)

```

Listing 1: Codice implementato per lo sparse Bayesian infinite factor model, repliche sui dati simulati e misure di perfomance, con mixing functions

III Diagnosi e plot aggiuntivi

III.I Applicazioni sui dati NIRS sull'impasto dei biscotti

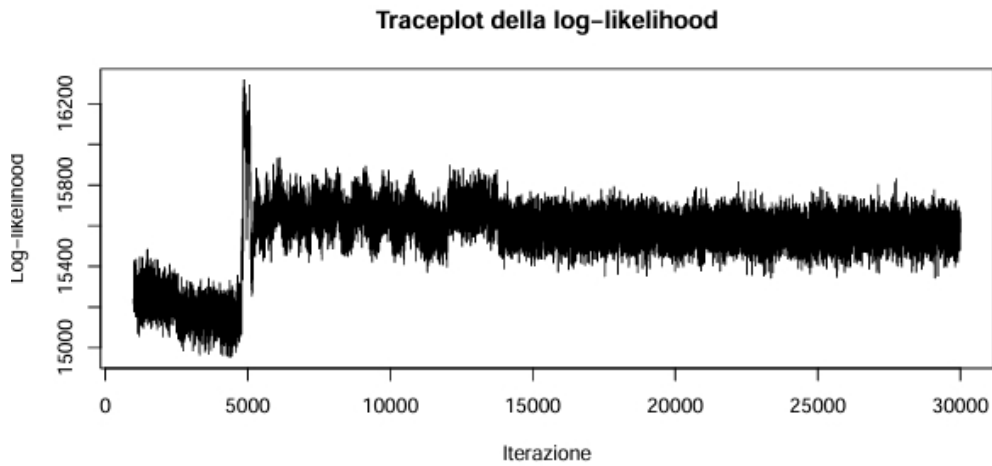


Figura A.1: Traceplot della log-verosimiglianza lungo le iterazioni della catena MCMC (30mila iterazioni) con Grassi come variabile dipendente.

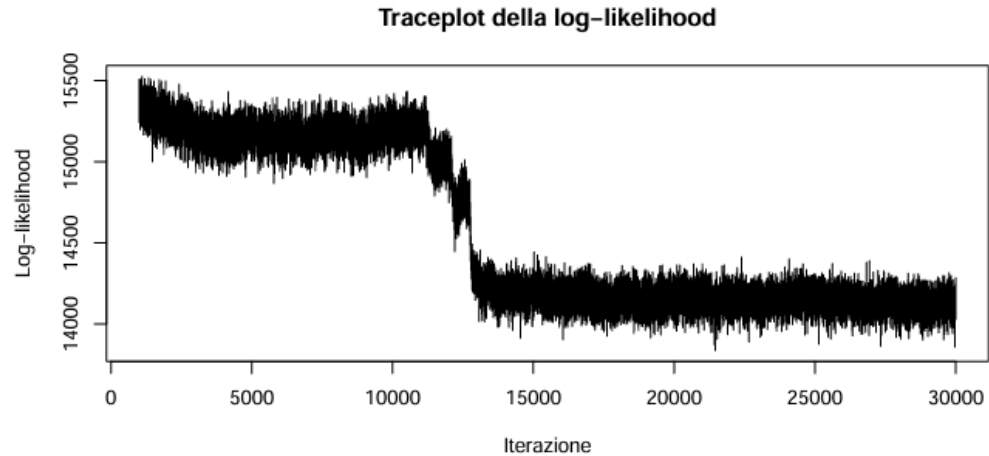


Figura A.2: Traceplot della log-verosimiglianza lungo le iterazioni della catena MCMC (30mila iterazioni) con Farina come variabile dipendente.

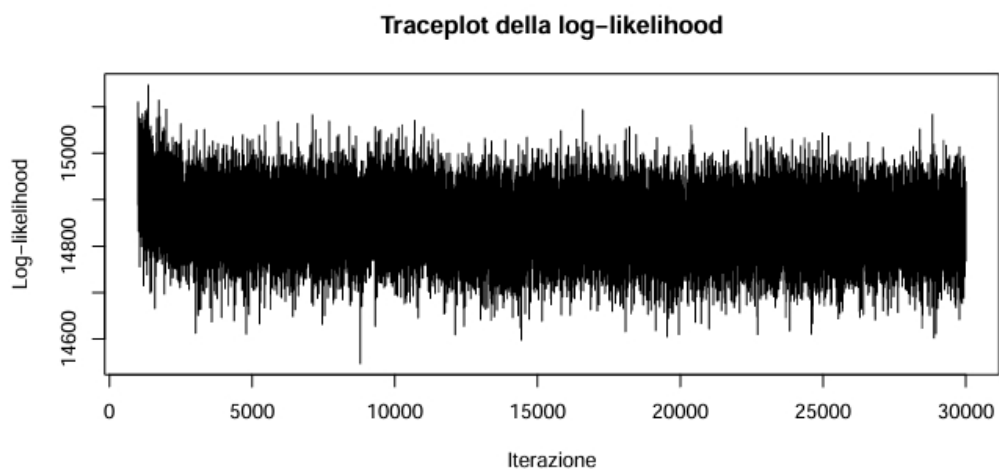


Figura A.3: Traceplot della log-verosimiglianza lungo le iterazioni della catena MCMC (30mila iterazioni) con Acqua come variabile dipendente.

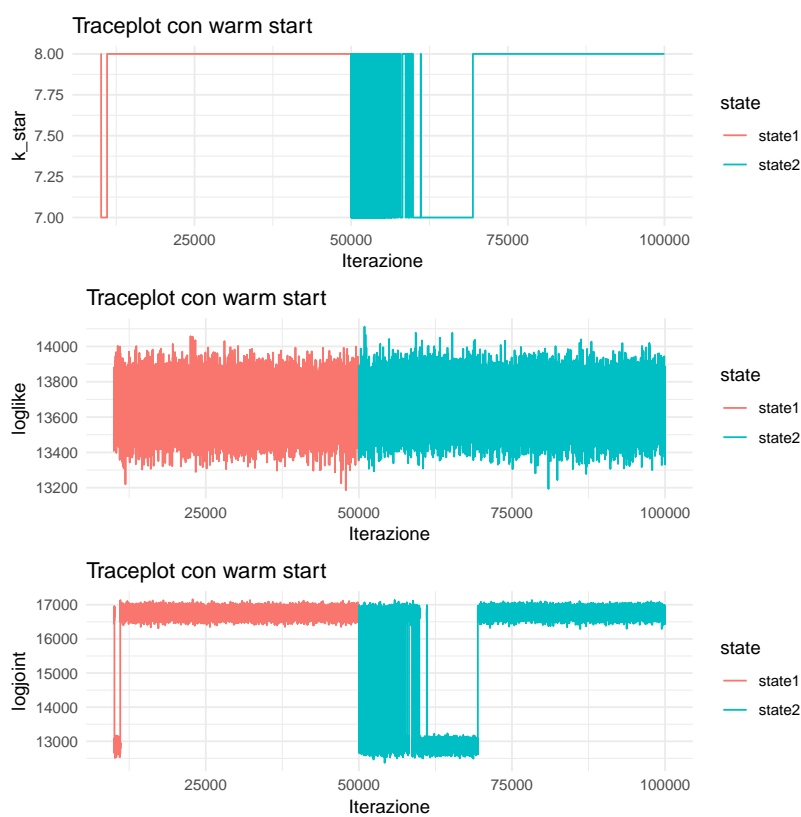


Figura A.4: Traceplot della log-verosimiglianza, la log-posterior in forma aperta, l'andamento dei k fattori lungo le iterazioni della catena MCMC in presenza di un warm start a metà catena con Acqua come variabile dipendente.

III.II Applicazioni sui dati NIRS sul grano

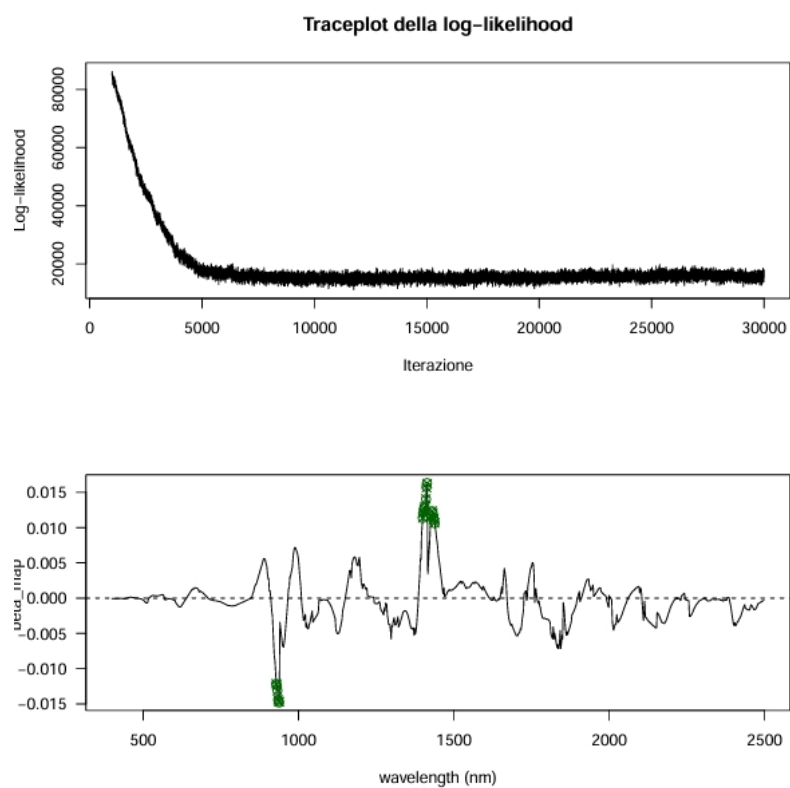


Figura A.5: Coefficienti stimati con S.B.I.F.M. per i dati NIRS sul grano

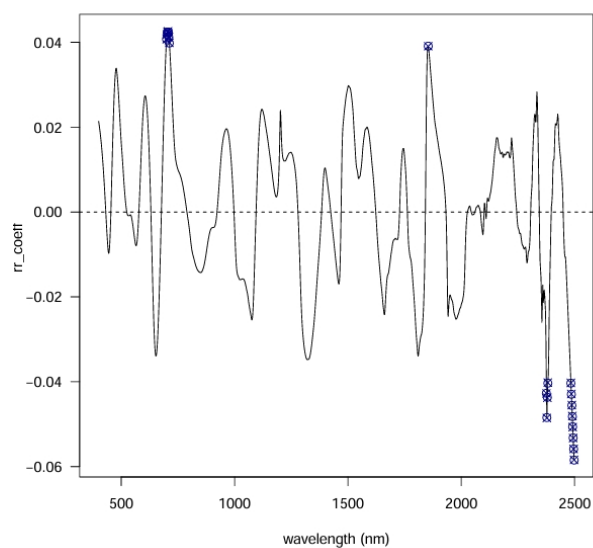


Figura A.6: Coefficienti stimati con Ridge per i dati NIRS sul grano