Università degli Studi di Padova
Dipartimento di Scienze Statistiche
Corso di Laurea Magistrale in
Scienze Statistiche

Aspetti computazionali nella selezione delle variabili nel modello di regressione multivariata

Relatrice Prof.ssa Manuela Cattelan
Dipartimento di Scienze Statistiche
Correlatore Prof. Mauro Bernardi
Dipartimento di Scienze Statistiche

Laureanda Gaia Penta
Matricola 2039075

Anno Accademico 2022/2023

Indice

Introduzione

Con l'emergere dei big data e la crescente necessità di gestire volumi di dati sempre più vasti, sono emerse due esigenze principali. La prima riguarda la selezione delle variabili, soprattutto nei casi in cui il numero di variabili esplicative (p) supera quello delle osservazioni (n). In queste circostanze, è essenziale individuare un sottoinsieme di variabili esplicative che possano spiegare in modo adeguato la variabile risposta. In secondo luogo, specialmente quando il numero di variabili è molto maggiore rispetto alle osservazioni, diventa necessario trovare metodologie che riducano il carico computazionale.

In questo lavoro, verrà trattato nello specifico il modello di regressione lineare multivariato, in cui la variabile risposta sarà quindi della forma $\mathbf{Y} \in \mathbb{R}^{n \times q}$. Il primo obiettivo sarà quello di individuare un sottoinsieme delle p variabili esplicative in grado di spiegare adeguatamente tutte le q variabili risposta. A tal fine si propone una breve introduzione relativa alla selezione delle variabili in contesti bayesiani, introducendo nello specifico il vettore di selezione p-dimensionale $\boldsymbol{\gamma} = \left(\gamma_1, \ldots, \gamma_p\right)^\intercal$, con $\gamma_j = 1$, $j = 1, \ldots, p$, se il j-esimo regressore è incluso nel modello e $\gamma_j = 0$ in caso contrario: d'interesse è individuarne la sua distribuzione a posteriori. Si utilizzerà infatti per la stima un approccio bayesiano, particolarmente adeguato in contesti ad elevata dimensionalità. Per indurre sparsità si utilizzerà una distribuzione a priori Dirac *spike and slab* sui coefficienti di regressione, marginalizzando poi la distribuzione a posteriori ottenuta rispetto al vettore $\boldsymbol{\gamma} \in \mathbb{R}^p$ precedentemente definito. Per simulare da tale distribuzione a posteriori si utilizzeranno metodi *Markov chain Monte Carlo*, in particolare il *Gibbs sampling* nella sua versione *reversible jump*, che verrà descritta nello specifico a livello teorico, valutando anche una struttura di *stochastic search*. Tale struttura risulta necessaria in quanto la dimensione dello spazio parametrico varia ad ogni iterazione. Questi aspetti verranno trattati nel Capitolo 1.

Nel Capitolo 2 si introdurranno invece gli aspetti computazionali coinvolti nella stima

del precedente modello, nonché delle tecniche utili a rendere il carico computazionale meno gravoso. I metodi proposti in questo lavoro riguardano la decomposizione QR e thin QR, dei quali verrano descritti gli aspetti prettamente algebrici. Oltre alla decomposizione QR di per sé, l'interesse è rivolto principalmente ai metodi di aggiornamento delle matrici $\mathbf{Q}$ ed $\mathbf{R}$ in seguito all'aggiunta o alla rimozione di una o più colonne dalla matrice originale di cui si è effettuata la fattorizzazione originale. A tal fine si utilizzeranno le rotazioni di Givens e le trasformazioni di Householder, anch'esse trattate nel dettaglio a livello algebrico.

Si procederà poi nel Capitolo 3 con gli studi di simulazione. Si valuterà in primo luogo l'efficienza dell'aggiornamento di generiche matrici simulate mediante decomposizione QR e thin QR, confrontato con il costo computazionale richiesto dall'effettuare una decomposizione QR della matrice originale aggiornata con il numero di colonne modificate. Sarà possibile notare il notevole risparmio in termini computazionali specialmente dell'aggiornamento thin QR. Saranno proposti in seguito due studi di simulazione, uno relativo al caso $p > n$, e il secondo relativo al caso $p < n$, applicando i modelli proposti. Si tratta di tre modelli: in primo luogo si è implementato il modello naive, che non presenta alcuna forma di aggiornamento, in seguito i modelli con forme di aggiornamento QR e thin QR. Verrà valutata sia la capacità di ciascun modello di individuare i coefficienti effettivamente non nulli, mediante l'*F1 Score*, il *True Positive Rate* e il *False Discovery Rate*, ma anche il tempo computazionale necessario al variare di p, n e p_0, ossia il numero di coefficienti non nulli. Sarà possibile apprezzare il notevole vantaggio ottenibile grazie ai metodi di aggiornamento thin QR, nonché la scarsa efficienza invece dell'aggiornamento basato sulla decomposizione QR, che risulterà essere più lento anche del modello naive. Il caso $p > n$ è risultato essere il più ostico, richiedendo l'implementazione di un metodo di ricerca stocastica del modello migliore da cui far partire l'algoritmo, accorgimento che ha consentito nei casi problematici di ottenere risultati soddisfacenti sia in termini di tempo computazionale che di accuratezza nella selezione dei coefficienti realmente non nulli.

Infine, nel Capitolo 4, si valuterà un'applicazione ad un dataset reale, nello specifico un dataset relativo a sei turbine eoliche situate nel parco eolico di Kelmarsh, nel Regno Unito, nella regione del Northamptonshire. Risulta d'interesse riuscire a prevedere la variabile risposta *Power.me*, ossia l'energia media prodotta in un intervallo di 10 minuti, essendo le rilevazioni osservate in questi frangenti di tempo, utilizzando numerose variabili esplicative relative principalmente alla velocità del vento e a dati SCADA, acronimo di *Supervisory Control And Data Acquisition*, ossia controllo di supervisione e

acquisizione dati. Essi si sostanziano in un insieme di misurazioni ambientali, operative, termiche ed elettriche, raccolte spesso per fini di manuntenzione. A causa della presenza di numerosi dati mancanti nelle rilvazioni di alcune turbine e per limiti computazionali, si sono utilizzati soltanto i dati relativi a tre turbine. Su tali turbine, si sono effettuate inizialmente alcune operazioni di *preprocessing* volte all'eliminazione e all'imputazione dei dati mancanti, proponendo poi delle analisi esplorative volte ad indagare le relazioni esistenti tra energia media prodotta e velocità del vento, tenendo in considerazione anche l'energia media teorica. Si è sviluppata poi una modellazione basata su due classi di modelli. Da un lato si sono adattati dei VAR, anche comprendendo le variabili esogene, dall'altro due modelli di regressione multivariata. Il primo è modello di regressione multivariata delineato nel Capitolo 1, stimato con l'ausilio di metodi di aggiornamento thin QR, dal momento che risulterà essere il più efficiente tra i tre proposti, confrontato con il modello MRCE definito nella Sezione 4.5.4, un modello di regressione multivariata sparsa con penalizzazioni di tipo lasso. Sono state valutate per entrambi diverse specficazioni della matrice di disegno, nonché diversi modi per effettuare le previsioni, utilizzando sia modelli MaP (Maximum a Posteriori), anche pesati, che MPM (Median Probability Model), comprendenti le variabili esplicative con probabilità di inclusione a posteriori superiore al 99% e al 95%. Si sono ottenuti poi anche spunti interpretativi, valutando quali variabili sono state incluse nel modello ritenuto migliore in termini previsivi e contestualmente di parsimonia.

Capitolo 1

Selezione delle variabili in un contesto bayesiano

1.1 Introduzione

La selezione delle variabili rappresenta il processo di identificazione di un sottoinsieme di predittori rilevanti da includere in un modello ed è stato un argomento di ricerca molto importante negli ultimi anni, soprattutto in contesti ad elevata dimensionalità. In questo lavoro il focus sarà sul modello di regressione multipla multivariata, che verrà trattato con un approccio bayesiano.

1.2 Regressione lineare multipla univariata e multivariata

Il modello di regressione lineare è il modello statistico più semplice ma anche il più utilizzato e conosciuto, rappresentando uno degli elementi fondanti della statistica e di altri modelli più complessi.

Data la variabile risposta $\mathbf{y} \in \mathbb{R}^n$ e la matrice di disegno $\mathbf{X} \in \mathbb{R}^{n \times p}$, nella sua versione univariata tale modello può essere definito come segue:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{1.1}$$

dove $\boldsymbol{\beta} \in \mathbb{R}^{p\times 1}$ è un vettore di parametri di regressione e $\boldsymbol{\varepsilon} \in \mathbb{R}^n$ è un vettore di termini di errore, con $\boldsymbol{\varepsilon} \sim \mathsf{N}_n(\mathbf{0}, \sigma\mathbf{I}_n)$. L'estensione multivariata presa in considerazione in questo lavoro prevede invece la presenza di q modelli di regressione univariata regrediti sulla medesima matrice di disegno $\mathbf{X} \in \mathbb{R}^{n\times p}$ già precedentemente definita. Per cui, data la j-esima variabile risposta $\mathbf{y}_j \in \mathbb{R}^n$, il modello di regressione lineare multivariato sarà definito come segue:

$$\mathbf{y}_j = \mathbf{X}\boldsymbol{\beta}_j + \boldsymbol{\varepsilon}_j, \qquad j = 1, 2, \ldots, q, \tag{1.2}$$

dove $\boldsymbol{\beta}_j \in \mathbb{R}^{p\times 1}$ è un vettore di parametri di regressione e $\boldsymbol{\varepsilon}_j \in \mathbb{R}^n$ è un vettore di termini di errore, con $\boldsymbol{\varepsilon}_j \sim \mathsf{N}_N(\mathbf{0}, \sigma_{j,j}\mathbf{I}_n)$. Si assume inoltre che $\mathsf{Cov}\left(\varepsilon_{i,t}, \varepsilon_{j,t}\right) = \sigma_{i,j} \neq 0$ per $i \neq j$ e $\mathsf{Cov}\left(\varepsilon_{i,t}, \varepsilon_{j,s}\right) = 0$ per ogni $i, j = 1, 2, \ldots, q$ e $t \neq s$, $t, s = 1, \ldots, n$.

Tale modello può in realtà poi essere definito anche nella sua versione matriciale. Data la matrice $\mathbf{Y} = \left(\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_q\right) \in \mathbb{R}^{n\times q}$, ottenuta aggregando le singole variabili risposta $\mathbf{y}_j$, la matrice $\mathbf{E} = \left(\boldsymbol{\varepsilon}_1, \boldsymbol{\varepsilon}_2, \ldots, \boldsymbol{\varepsilon}_q\right) \in \mathbb{R}^{n\times q}$, composta dai vettori dei termini di errore $\boldsymbol{\varepsilon}_j$ e $\mathbf{B} = \left(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_q\right) \in \mathbb{R}^{p\times q}$, il modello in equazione (1.2) in forma matriciale sarà definito come:

$$\mathbf{Y} = \mathbf{X}\mathbf{B} + \mathbf{E}, \tag{1.3}$$

dove $\mathbf{E} \sim \mathsf{N}_{n\times q}(\mathbf{0}, \boldsymbol{\Sigma}, \mathbf{I}_n)$, che indica una distribuzione normale matriciale, con media zero, matrice di varianze e covarianze $\boldsymbol{\Sigma} = [\sigma_{i,j}]_{i,j=1,\ldots,q} \in \mathbb{S}^q_{++}$ per le colonne di $\mathbf{Y}$ e $\mathbf{I}_n$, matrice identità di dimensione n, per le righe di $\mathbf{Y}$.

1.3 Inferenza in un contesto bayesiano

In un approccio bayesiano, sia la variabile risposta $\mathbf{y}$ che i parametri, definiti genericamente con $\boldsymbol{\theta}$, vengono considerati come delle variabili casuali. Su tali quantità, si assumono poi delle distribuzioni di probabilità, sia sui dati, per i quali si parla di verosimiglianza analogamente all'approccio frequentista, sia sui parametri, sui quali viene posta invece una distribuzione a priori. Tale informazione a priori poi viene aggiornata attraverso il teorema di Bayes, definito nel Teorema A.1 in Appendice, in modo da ottenere la distribuzione a posteriori.

Nel caso della regressione lineare univariata, definita in equazione (1.2), si assume

una distribuzione normale degli errori, ed equivale quindi ad assumere

$$\mathbf{y}|\mathbf{X}, \boldsymbol{\beta} \sim \mathsf{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_N), \tag{1.4}$$

mentre nel caso multivariato si ottiene

$$\mathbf{Y}|\mathbf{X}, \mathbf{B}, \boldsymbol{\Sigma} \sim \phi_{n\times q}(\mathbf{XB}, \boldsymbol{\Sigma}, \mathbf{I}_n), \tag{1.5}$$

in cui $\phi_{n\times q}$ denota una distribuzione normale matriciale. Tale densità rappresenta la funzione di verosimiglianza, che verrà indicata come $\mathcal{L}(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta})$ nel caso univariato e come $\mathcal{L}(\mathbf{Y}|\mathbf{X}, \mathbf{B}, \boldsymbol{\Sigma})$ nel caso multivariato. A questo punto è necessario definire delle distribuzioni a priori sui parametri, genericamente

$$\boldsymbol{\beta} \sim \pi(\boldsymbol{\beta}), \qquad \mathbf{B} \sim \pi(\mathbf{B}), \qquad \boldsymbol{\Sigma} \sim \pi(\boldsymbol{\Sigma}), \tag{1.6}$$

e che verranno definite nello specifico in seguito. A questo punto è possibile usare il teorema di Bayes per ottenere la distribuzione a posteriori per i parametri . Per il caso univariato si otterrà

$$\pi(\boldsymbol{\beta}|\mathbf{y}, \mathbf{X}) = \frac{\pi(\boldsymbol{\beta})\mathcal{L}(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta})}{\int_\Theta \pi(\boldsymbol{\beta})\mathcal{L}(\mathbf{Y}|\mathbf{X}, \boldsymbol{\beta})d\boldsymbol{\beta}} \tag{1.7}$$

e analogamente per il caso multivariato

$$\pi(\mathbf{B}, \boldsymbol{\Sigma}|\mathbf{Y}, \mathbf{X}) = \frac{\pi(\mathbf{B})\pi(\boldsymbol{\Sigma})\mathcal{L}(\mathbf{Y}|\mathbf{X}, \mathbf{B}, \boldsymbol{\Sigma})}{\int_{\mathbf{B}}\int_{\boldsymbol{\Sigma}} \pi(\mathbf{B})\pi(\boldsymbol{\Sigma})\mathcal{L}(\mathbf{Y}|\mathbf{X}, \mathbf{B}, \boldsymbol{\Sigma})d\mathbf{B}d\boldsymbol{\Sigma}}. \tag{1.8}$$

In quest'ultimo caso, essendoci più parametri, è possibile individuare la distribuzione marginale di ciascun parametro integrando il parametro non di interesse. Altro aspetto importante risulta essere che il denominatore è una costante di normalizzazione e di conseguenza

$$\pi(\boldsymbol{\beta}|\mathbf{Y}) \propto \pi(\boldsymbol{\beta})\mathcal{L}(\mathbf{Y}|\mathbf{X}, \boldsymbol{\beta}). \tag{1.9}$$

Si nota quindi come la densità a posteriori sia proporzionale al prodotto tra la densità a priori e la funzione di verosimiglianza (Pace et al., 2022).

1.4 Selezione delle variabili in contesti ad elevata dimensionalità

In contesti ad elevata dimensionalità il numero delle variabili esplicative può essere molto elevato e superare il numero di osservazioni presenti. Sono proprio questi i casi in cui i metodi tradizionali falliscono, richiedendo degli approcci differenti. Infatti, ad esempio, la soluzione dei minimi quadrati non risulta essere unica nei casi in cui $p > n$ e risulta necessario introdurre delle assunzioni di sparsità, che consistono nel presupporre che la maggior parte dei coefficienti di regressione siano nulli (Narisetty, 2020). Questa condizione è ovviamente desiderabile sia a fini interpretativi che di parsimonia.

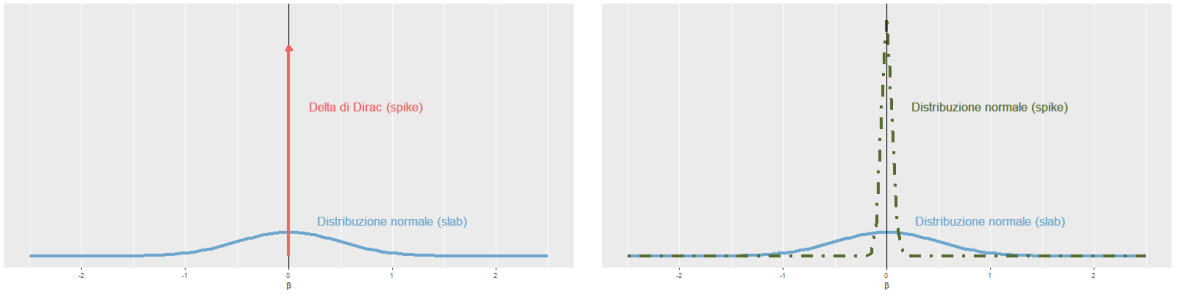
Tra i classici metodi di selezione delle variabili si annoverano il *best subset selection*, assieme ad approcci *stepwise backward* e *forward*, nonché quelli basati sui criteri di informazione, tra cui l'AIC e il BIC (Narisetty, 2020). Un altro filone della letteratura invece si basa sui metodi di penalizzazione, tra cui la regressione *ridge, lasso* e l'*elastic net*, approcci particolarmente utili in presenza di correlazione tra le variabili esplicative, altro aspetto che induce problemi nel contesto dei minimi quadrati, e che consentono di ottenere una soluzione unica del problema di minimizzazione. Si possono poi citare penalizzazioni non convesse, tra cui SCAD ed MCP.

Oltre a questi classici approcci, si inserisce anche la selezione delle variabili in un contesto bayesiano, oggetto di tutta la trattazione in seguito. A tal fine, si introduce il vettore di selezione p-dimensionale $\boldsymbol{\gamma} = \left(\gamma_1, \ldots, \gamma_p\right)^\intercal$, con $\gamma_j = 1$, $j = 1, \ldots, p$, se il j-esimo regressore è incluso nel modello e $\gamma_j = 0$ in caso contrario. L'obiettivo è individuarne la sua distribuzione a posteriori $\pi(\boldsymbol{\gamma}|\mathbf{Y}, \mathbf{X})$. Vi sono poi diversi approcci per effettuare la selezione delle variabili. In particolare, può essere utilizzato:

1. il modello *Maximum a Posteriori* (MaP), ossia il modello che massimizza la probabilità a posteriori, ovvero

$$\mathsf{arg\ max}_{\mathbf{k}} \pi(\boldsymbol{\gamma} = \mathbf{k}|\mathbf{Y}, \mathbf{X}) \tag{1.10}$$

   in cui $\mathbf{k}$ indica un modello codificato come un vettore binario, con gli uno che indicano le variabili attive e gli zero quelle inattive. Individuare il modello MaP richiederebbe il calcolo di tutti i 2^p modelli, che risulta essere molto oneroso computazionalmente con un numero di variabili esplicative grande e a tal fine si usano metodi di simulazione che saranno trattati più avanti (Narisetty, 2020).



(a) (b)

FIGURA 1.1: Distribuzione a priori *spike and slab*: a sinistra una distribuzione con *spike point mass*, a destra con *spike* continuo.

2. il *Median Probability Model*, che consiste nell'individuare l'insieme delle variabili indipendenti la cui probabilità marginale a posteriori eccede 0.5, ossia $\{j : \pi(\gamma_j = 1|\mathbf{Y}, \mathbf{X}) > 0.5\}$. In realtà, la soglia 0.5 può essere anche variata e selezionata in modo adattativo, così da individuarne il valore ottimo, ad esempio mediante criteri di informazione (Narisetty, 2020).

1.5 Distribuzione a priori spike and slab

In ambito bayesiano, è possibile indurre sparsità imponendo una distribuzione a priori *spike and slab*, ossia una distribuzione mistura di due componenti sui coefficienti di regressione $\boldsymbol{\beta}_j$ o $\mathbf{B}_j$, con $j = 1, \ldots, p$, una con un picco in 0 (lo *spike*) e l'altra più diffusa (*slab*). A tal fine infatti si introduce il vettore di selezione p-dimensionale precedentemente definito $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)^\intercal$, con $\gamma_j = 1$, $j = 1, \ldots, p$ se il j-esimo regressore è incluso nel modello e $\gamma_j = 0$ in caso contrario. L'insieme dei coefficienti inclusi nel modello saranno indicati con $\boldsymbol{\beta}_\gamma$ e $\mathbf{B}_\gamma$, quelli esclusi invece con $\boldsymbol{\beta}_{-\gamma}$ e $\mathbf{B}_{-\gamma}$, rispettivamente facendo riferimento al caso univariato e al caso multivariato.

Infatti, mentre per la distribuzione *slab* si usa solitamente una distribuzione continua, tipicamente una distribuzione normale, per la distribuzione *spike* invece sono possibili sia distribuzioni *point mass*, che attribuiscono una massa di probabilità sullo 0, che continue. In quest'ultimo caso si avrà tipicamente una mistura di due distribuzioni a priori gaussiane, una concentrata sullo zero ed una più sparsa su un insieme di valori più ampio. Un esempio di distribuzione a priori *point mass* per il caso univariato in

(1.1) risulta essere il seguente:

$$\beta_j|\sigma^2, \gamma_j \sim (1-\gamma_j)\delta(\beta_j, 0) + \gamma_j \mathsf{N}(0, h_j\sigma^2) \tag{1.11}$$

con $\delta(\beta_j, 0)$ la funzione di Dirac valutata in $\beta_j = 0$ e h_j iperparametro da definire. Quindi, $\gamma_j = 0$ esclude la j-esima variabile dal modello dal momento che la distribuzione a priori sul corrispondente coefficiente β_j è una distribuzione a massa di probabilità sullo 0, mentre $\gamma_j = 1$ determina l'inclusione del predittore nel modello determinando una priori normale su β_j (Tadesse & Vannucci, 2021). Una rappresentazione grafica è visibile in Figura 1.1(a). Un esempio di distribuzione a priori *spike and slab* continua è invece il seguente:

$$\beta_j|\sigma^2, \gamma_j = 0 \sim \mathsf{N}(0, \tau_0^2\sigma^2), \quad \beta_j|\sigma^2, \gamma_j = 1 \sim \mathsf{N}(0, \tau_1^2\sigma^2) \tag{1.12}$$

con $0 < \tau_0^2 < \tau_1^2 < \infty$ che rappresentano iperaparametri relativi alle varianze a priori che possono essere regolarizzati (Narisetty, 2020). Un esempio è visibile in Figura 1.1(b). È successivamente necessario imporre anche una distribuzione a priori su γ_j, scegliendo di solito una distribuzione Bernoulli di parametro θ o una Beta di parametri a, b (Tadesse & Vannucci, 2021). In quest'ultimo caso per una priori non informativa è comune usare dei valori $a = b = 1$. In questo caso, per il modello di regressione multivariato in (1.3) si utilizzerà il framework proposto da George & McCulloch (1997) che impone una distribuzione a priori Dirac *spike and slab*. Come detto, essa viene utilizzata per indurre sparsità nel modello, in contesti nei quali è importante individuare un sottoinsieme delle variabili esplicative utile a predire tutte le q risposte, sfruttando un vettore di selezione p-dimensionale $\boldsymbol{\gamma} = \big(\gamma_1, \dots, \gamma_p\big)^\intercal$, con $\gamma_j = 1$, $j = 1, \dots, p$ se il j-esimo regressore è incluso nel modello e $\gamma_j = 0$ in caso contrario.

La distribuzione a priori posta sugli elementi del modello definito precedentemente e su γ_j risulta essere la seguente (Brown et al., 1998a; Richardson et al., 2011; Bottolo et al., 2020):

$$\mathbf{B}|\boldsymbol{\Sigma}, \boldsymbol{\gamma} \sim \mathsf{N}_{p\times q}(\mathbf{B}_{\gamma,0}, \boldsymbol{\Sigma}, \mathbf{H}_\gamma), \qquad \boldsymbol{\Sigma}|\boldsymbol{\gamma} \sim \mathsf{IW}_q(c_0, \mathbf{C}_0), \tag{1.13}$$

dove $\mathbf{H}_\gamma = \mathbf{D}_\gamma \mathbf{R}_\gamma \mathbf{D}_\gamma$, con $\mathbf{R}_\gamma \in \mathbb{S}_{++}^q$ è una matrice di correlazione e $\mathbf{D}_\gamma = \mathsf{diag}(\upsilon_{\gamma_1}^{1/2}, \dots, \upsilon_{\gamma_p}^{1/2})$ e

$$\upsilon_{\gamma_j} = \begin{cases} \upsilon_{0,j}, & \gamma_j = 0 \\ \upsilon_{1,j}, & \gamma_j = 1, \end{cases} \tag{1.14}$$

con $\upsilon_{1,j} \gg \upsilon_{0,j}$, per $j = 1, \ldots, q$, ma in questo caso al fine effettuare selezione delle variabili si avrà $\upsilon_{0,j} = 0$ per ogni $j = 1, \ldots, p$.In questo modo $\gamma_j = 0$ indica che la j-esima riga di $\mathbf{B}$ abbia varianza pari a 0 così da avere una distribuzione degenere in 0, mentre $\gamma_j = 1$ implica che la j-esima riga di $\mathbf{B}$ ha varianza non nulla determinata da $\upsilon_{1,j}$. Si assume poi la seguente distribuzioni a priori Dirac *spike and slab*:

$$\pi\big(\mathbf{B}_\gamma|\boldsymbol{\Sigma}, \boldsymbol{\gamma}\big) = \phi_{p_\gamma \times q}(\mathbf{B}_{\gamma,0}, \boldsymbol{\Sigma}, \mathbf{H}_\gamma) \tag{1.15}$$

$$\pi\big(\boldsymbol{\Sigma}|\boldsymbol{\gamma}\big) = \varphi_{\mathsf{IW}}(c_0, \mathbf{C}_0) \tag{1.16}$$

$$\pi\big(\mathbf{B}_{-\gamma}|\boldsymbol{\gamma}\big) = \prod_{j=1}^{p} \delta\big(\mathbf{B}_j, 0\big)^{1-\gamma_j} \tag{1.17}$$

$$\pi\big(\gamma_j\big) \sim \mathrm{Ber}\big(\theta\big), \qquad j = 1, \ldots, p, \tag{1.18}$$

dove $\delta\big(\mathbf{B}_j, 0\big) = \prod_{l=1}^{q} \delta\big(\mathbf{B}_{j,l}, 0\big)$ e $\delta\big(x, 0\big) = \mathbb{1}_{\{0\}}\big(x\big)$ indica la funzione di Dirac valutata in 0, $p_\gamma = \sum_{j=1}^{p} \gamma_j$ indica il numero di covariate incluse nel modello di regressione, $\mathbf{B}_\gamma \in \mathbb{R}^{p_\gamma \times q}$ indica la sottomatrice che comprende tutte le righe di $\mathbf{B}$ per le quali $\gamma_j = 1$ per ogni $j = 1, \ldots, p$, con $\mathbf{B}_{-\gamma}$ tale che $\mathbf{B} = \mathbf{B}_\gamma \cup \mathbf{B}_{-\gamma}$ con $\mathbf{B}_\gamma \cap \mathbf{B}_{-\gamma} = \emptyset$. $\mathbf{H}_\gamma \in \mathbb{S}_{++}^{p_\gamma}$ risulta invece essere una matrice di varianza e covarianza simmetrica e definita positiva, mentre $\theta \in \big(0, 1\big)$. Come definito in precedenza, $\phi_{p_\gamma \times q}(\mathbf{B}_{\gamma,0}, \boldsymbol{\Sigma}, \mathbf{H}_\gamma)$ nell'equazione (1.15) indica la distribuzione normale matriciale con media $\mathbf{B}_{\gamma,0} \in \mathbb{R}^{p_\gamma \times q}$ e matrici di varianza e covarianza $\boldsymbol{\Sigma}$ e $\mathbf{H}_\gamma$, mentre $\varphi_{\mathsf{IW}}(c_0, \mathbf{C}_0)$ in equazione (1.16) indica una distribuzione Inverse-Wishart.

1.6 Distribuzione a posteriori

Per quanto riguarda la distribuzione a posteriori, essa si ottiene sulla base del teorema di Bayes come esplicato nel paragrafo 1.3. Quindi, dato il prodotto tra le quantità definite nelle equazioni (1.15)-(1.18) e la funzione di verosimiglianza del modello definita in (1.5), è possibile innanzitutto individuare la distribuzione a posteriori congiunta di $(\mathbf{B}_\gamma, \boldsymbol{\Sigma})$, che risulta essere definita come

$$\pi(\mathbf{B}_\gamma, \boldsymbol{\Sigma}|\mathbf{Y}, \mathbf{X}, \boldsymbol{\gamma}) = \pi(\mathbf{B}_\gamma|\boldsymbol{\Sigma}, \mathbf{Y}, \mathbf{X}, \boldsymbol{\gamma})\pi(\boldsymbol{\Sigma}|\mathbf{Y}, \mathbf{X}, \boldsymbol{\gamma}), \tag{1.19}$$

dove

$$\pi(\mathbf{B}_\gamma|\boldsymbol{\Sigma},\mathbf{Y},\mathbf{X},\boldsymbol{\gamma}) = \phi_{p_\gamma\times q}(\mathbf{B}_\gamma|\widetilde{\mathbf{B}}_\gamma,\boldsymbol{\Sigma},\widetilde{\boldsymbol{\Sigma}}_\gamma) \tag{1.20}$$

$$\pi(\boldsymbol{\Sigma}|\mathbf{Y},\mathbf{X},\boldsymbol{\gamma}) = \varphi_{\mathsf{IW}q}(c_0+n,\mathcal{Q}_\gamma), \tag{1.21}$$

dove $\widetilde{\mathbf{B}}_\gamma = \mathbf{K}_\gamma^{-1}\mathbf{M}$, $\mathbf{K}_\gamma = \mathbf{X}^\intercal\mathbf{X}+\mathbf{H}_\gamma^{-1}$, $\widetilde{\boldsymbol{\Sigma}}_\gamma = \mathbf{K}_\gamma^{-1}$, $\mathcal{Q}_\gamma = \mathbf{C}_0+\mathbf{C}+\mathbf{M}^\intercal\mathbf{K}_\gamma^{-1}\mathbf{M}$ con $\mathbf{C} = \mathbf{Y}^\intercal\mathbf{Y}+\mathbf{B}_{\gamma,0}^\intercal\mathbf{H}_\gamma^{-1}\mathbf{B}_{\gamma,0}$, $\mathbf{M} = \mathbf{X}^\intercal\mathbf{Y}-\mathbf{H}_\gamma^{-1}\mathbf{B}_{\gamma,0}$. Assumendo media zero a priori per $\mathbf{B}_\gamma$ si ha $\mathcal{Q}_\gamma = \mathbf{C}_0+\mathbf{Y}^\intercal\mathbf{Y}+\mathbf{Y}^\intercal\mathbf{X}\mathbf{K}_\gamma^{-1}\mathbf{X}^\intercal\mathbf{Y}$ e $\mathbf{M} = \mathbf{X}^\intercal\mathbf{Y}$. Si ottiene infine che la distribuzione marginale a posteriori $\pi(\mathbf{B}_\gamma|\mathbf{Y},\mathbf{X},\boldsymbol{\gamma})$, in seguito all'integrazione di $\boldsymbol{\Sigma}$, risulta essere una distribuzione t di Student matriciale

$$\pi(\mathbf{B}_\gamma|\mathbf{Y},\mathbf{X},\boldsymbol{\gamma}) = \varphi_{\mathsf{T}p_\gamma,q}(\mathbf{B}_\gamma|\widetilde{\mathbf{B}}_\gamma,\mathcal{Q}_\gamma,\widetilde{\boldsymbol{\Sigma}}_\gamma^{-1},c_0+n). \tag{1.22}$$

La distribuzione a posteriori di γ può essere invece fattorizzata come

$$\pi(\boldsymbol{\gamma}|\mathbf{Y},\mathbf{X}) = \pi(\boldsymbol{\gamma})\int \mathcal{L}(\mathbf{Y}|\mathbf{X},\boldsymbol{\gamma},\mathbf{B}_\gamma,\boldsymbol{\Sigma})\pi(\mathbf{B}_\gamma|\boldsymbol{\Sigma},\mathbf{Y},\mathbf{X},\boldsymbol{\gamma})\pi(\boldsymbol{\Sigma}|\mathbf{Y},\mathbf{X},\boldsymbol{\gamma})\,d\mathbf{B}\,d\Sigma. \tag{1.23}$$

La sua forma esplicita è visibile in Appendice in equazione A.13

1.7 Distribuzione predittiva

La distribuzione predittiva del modello risulta utile nel momento in cui si vogliano prevedere r osservazioni future $\mathbf{Y}_0 \in \mathbb{R}^{r\times q}$ data una matrice $\mathbf{X}_{0,\gamma} \in \mathbb{R}^{r\times p_\gamma}$. Dato quindi il modello definito in 1.3 e le distribuzioni a posteriori di $\mathbf{B}_\gamma$ e $\boldsymbol{\Sigma}$ definite nelle equazioni (1.20) e (1.21) rispettivamente, la distribuzione predittiva di $\mathbf{Y}_0 \in \mathbb{R}^r$ con $\mathbf{Y}_0 \sim \phi_{r\times q}(\mathbf{X}_{0,\gamma}\mathbf{B},\boldsymbol{\Sigma},\mathbf{I}_r)$, data la corrispondente matrice di disegno $\mathbf{X}_{0,\gamma} \in \mathbb{R}^{r\times p_\gamma}$ risulta essere

$$\pi\big(\mathbf{Y}_0|\mathbf{Y},\mathbf{X},\mathbf{X}_0\big) \propto \varphi_{\mathsf{T}r\times q}\big(\mathbf{X}_0\widetilde{\mathbf{B}}_\gamma,\mathcal{Q}_\gamma,\mathbf{F}_0,c0+n\big), \tag{1.24}$$

dove $\widetilde{\mathbf{B}}_\gamma = \widetilde{\boldsymbol{\Sigma}}_\gamma\mathbf{X}^\intercal\mathbf{Y}$ è la media a posteriori di $\mathbf{B}_\gamma$, $\mathcal{Q}_\gamma = \mathbf{C}_0+\mathbf{Y}^\intercal\mathbf{Y}+\mathbf{Y}^\intercal\mathbf{X}\widetilde{\boldsymbol{\Sigma}}_\gamma\mathbf{X}^\intercal\mathbf{Y}$, $\mathbf{F}_0 = \mathbf{I}_r+\mathbf{X}_0\widetilde{\boldsymbol{\Sigma}}_\gamma\mathbf{X}_0^\intercal$ e $\widetilde{\boldsymbol{\Sigma}}_\gamma = (\mathbf{K}_\gamma)^{-1} = (\mathbf{X}^\intercal\mathbf{X}+\mathbf{H}_\gamma^{-1})^{-1}$ e φ_T indica una distribuzione t di Student matriciale.

1.8 Metodi di simulazione

La distribuzione a posteriori per $\boldsymbol{\gamma}$ si ottiene direttamente dall'equazione (A.13), tuttavia l'onere computazionale risulta essere rilevante, dal momento che dovrebbe essere effettuato per tutte le 2^p combinazioni di zero e uno, cosa che non risulta essere fattibile già se il numero dei parametri supera i 25. In tal caso infatti tali combinazioni sarebbero 33.554.432, crescendo ovviamente poi in modo esponenziale all'aumentare del numero di parametri (Brown et al., 1998a). Si opta quindi per l'utilizzo di metodi numerici, che consentono di esplorare le distribuzioni a posteriori, come *Markov chain Monte Carlo* (MCMC), in particolare il *Gibbs sampling* nella sua versione *reversible jump*, che consente di effettuare anche salti transdimensionali. La marginalizzazione in (A.13), in aggiunta poi ad algoritmi basati sulla decomposizione QR per aggiornare il calcolo della verosimiglianza marginale, consente di ottenere degli schemi di simulazione più efficienti (Tadesse & Vannucci, 2021).

1.8.1 Metodi Markov chain Monte Carlo

Nell'inferenza bayesiana basata su metodi *Markov chain Monte Carlo*, risulta necessario creare una catena markoviana che abbia $\pi(\boldsymbol{\gamma}, \mathbf{B}_\gamma|\mathbf{Y})$, ossia la distribuzione a posteriori target, come distribuzione limite. Si considerano solo le catene reversibili, che soddisfano la cosiddetta *detailed balance condition*, che prevede che la probabilità di equilibrio che lo stato di una catena sia in un insieme generico A e che si muova verso un insieme generico B sia la stessa anche con A e B invertiti (Hastie & Green, 2012).

In altri termini, definendo per semplicità $x = (\boldsymbol{\gamma}, \mathbf{B}_\gamma)$, si rende necessario costruire un kernel di transizione $P(x, x')$ che sia aperiodico, irriducibile e ricorrente positivo, e che soddisfi la condizione

$$\int_A \int_B \pi(x|\mathbf{Y}, \mathbf{X}, \boldsymbol{\Sigma}) P(x, x') = \int_B \int_A \pi(x'|\mathbf{Y}, \mathbf{X}, \boldsymbol{\Sigma}) P(x', x)$$

per ogni A e B appropriato (Green, 1995). Tale condizione può essere soddisfatta proponendo un nuovo stato per la catena, accettandolo con una probabilità appropriata, ottenuta considerando la sua transizione e la sua inversa simultaneamente. In un determinato stato x, verranno quindi generati r numeri casuali u da una determinata densità g. Il nuovo stato proposto della catena x' è poi costruito sulla base di una particolare funzione deterministica h tale che $(x', u') = h(x, u)$, dove con u' si indica r' numeri

casuali generati da una densità g' necessaria per effettuare il movimento inverso da x' a x, usando la funzione inversa di h, ossia h' (Hastie & Green, 2012). Se si accetta il nuovo stato x' da x con probabilità $\alpha(x, x')$ e, coerentemente, il movimento inverso viene effettuato con probabilità $\alpha(x, x')$, la condizione *detailed-balance* può essere scritta come:

$$\int_{(x,x')\in A\times B} \pi(x)g(u)\alpha(x,x')dx,du = \int_{(x,x')\in A\times B} \pi(x')g(u')\alpha(x',x)dx',du'. \tag{1.25}$$

Se la trasformazione h da (x, u) a (x', u') e la sua inversa sono differenziabili, allora è possibile applicare la formula standard per il cambio di variabile alla parte destra dell'equazione (1.25), vedendo come l'uguaglianza risulti essere valida se

$$\pi(x)g(u)\alpha(x,x') = \pi(x')g'(u')\alpha(x',x)\left|\frac{\partial(x',u')}{\partial(x,u)}\right|. \tag{1.26}$$

In questo modo una scelta valida per α risulta essere

$$\alpha(x',x) = \mathsf{min}\left\{1, \frac{\pi(x')g'(u')}{\pi(x)g(u)}\left|\frac{\partial(x',u')}{\partial(x,u)}\right|\right\}. \tag{1.27}$$

I due metodi più utilizzati che si inseriscono in quest'ottica sono il *Gibbs sampler* e il Metropolis-Hastings. La differenza sta nel fatto che nel *Gibbs sampler*, i nuovi valori sono estratti dalle distribuzioni *full conditionals* $x_j|x_{-j}, \mathbf{Y}, \mathbf{X}$ con $j = 1, \ldots, p$, con $x_{-j} = (x_1, x_2, \ldots, x_{j-1}, x_{j+1}, \ldots, x_p)$, mentre nel Metropolis-Hastings i valori proposti x' sono estratti da una distribuzione arbitraria g (Green, 1995).

1.8.2 Reversible jump Markov chain Monte Carlo

Il MCMC *reversible jump* è un metodo per simulare dalla distribuzione a posteriori nel caso in cui la dimensione del vettore dei parametri non è fissata ed è quindi libera di variare. Si inserisce nei metodi di simulazione *across-model*, ossia in cui vi è una sola simulazione MCMC con spazio degli stati della forma $(\boldsymbol{\gamma}, \mathbf{B}_\gamma) \sim \pi(\boldsymbol{\gamma}, \mathbf{B}_\gamma|\mathbf{Y})$, distinguendosi dai metodi di simulazione *within-model* in cui invece vi sono simulazioni separate di $\mathbf{B}_\gamma \sim \pi(\mathbf{B}_\gamma|\boldsymbol{\gamma}, \mathbf{Y})$.

In questo caso quindi le dimensioni di x ed x' possono essere diverse, e si avrà $x \in \mathbb{R}^n,\ x' \in \mathbb{R}^{n'},\ u \in \mathbb{R}^r,\ u' \in \mathbb{R}^{r'}$, con le conseguenti funzioni $h : \mathbb{R}^n \times \mathbb{R}^r \to \mathbb{R}^{n'} \times \mathbb{R}^{r'}$ e $h' : \mathbb{R}^{n'} \times \mathbb{R}^{r'} \to \mathbb{R}^n \times \mathbb{R}^r$ e $(x', u') = h(x, u)$ e $(x, u) = h(x', u')$. Affinché la trasformazione sia un diffeomorfismo è necessario che $n + r = n' + r'$, in quanto in caso contrario la

funzione e la sua inversa non potrebbero essere entrambe differenziabili. In questo caso, inoltre, saranno necessari diversi tipi di movimenti per visitare l'intero spazio parametrico. Si opterà per il caso in cui i tipi di movimenti sono scelti indipendentemente per ogni ricerca dello spazio parametrico, consentendo allo stesso tempo che ogni mossa successiva dipenda dallo stato corrente.

Dati M movimenti, un particolare movimento m riferito ad una specifica coppia $(\boldsymbol{\gamma}, \boldsymbol{\gamma}')$ consiste sia nel passo in avanti da $x = (\boldsymbol{\gamma}, \mathbf{B}_\gamma)$ a $x' = (\boldsymbol{\gamma}', \mathbf{B}'_\gamma)$ che all'indietro, ossia da x' a x. Per il passo in avanti, vengono generati r_m numeri casuali u da una distribuzione nota g_m e il nuovo spazio degli stati $\mathbf{B}'_{\gamma'} \in \mathbb{R}^{n_{\gamma'}}$ viene costruito come $(\mathbf{B}'_{\gamma'}, u') = h_m(\mathbf{B}_\gamma, u)$, in cui u' sono gli r'_m numeri casuali generati dalla distribuzione congiunta g'_m necessari per il passo all'indietro, ossia per passare da $\mathbf{B}'_{\gamma'}$ a $\mathbf{B}_\gamma$ usando h'_m. Definendo $j_m(x)$ la probabilità che il movimento m sia provata allo stato x, il tipo di movimento può essere definito come

$$\int_{(x,x')\in A\times B} p(x) j_m(x) g_m(u) \alpha_m(x, x') dx du = \int_{(x,x')\in A\times B} p(x') j_m(x') g'_m(u') \alpha_m(x', x) dx du.$$

La probabilità di accettazione può essere definita come

$$\alpha_m = \min\left\{1, \frac{p(x')\ j_m(x')\ g'_m(u')}{p(x)\ j_m(x)\ g_m(u)} \left| \frac{d(\mathbf{B}'_{\gamma'}, u')}{d(\mathbf{B}_\gamma, u)} \right| \right\}.$$

Lo Jacobiano deriva dalla trasformazione da $\mathbf{B}_\gamma$ a $\mathbf{B}'_{\gamma'}$ e risulta dipendente dal tipo di movimento m effettuato (Hastie & Green, 2012).

1.8.3 Scelta della proposal

Un aspetto piuttosto importante per questo metodo risulta essere la scelta della distribuzione da cui i valori casuali vengono proposti. Il primo metodo per la scelta della proposal sono gli *order methods*. Si focalizzano principalmente su una parametrizzazione efficiente della densità della *proposal* $g(u)$, fissata la trasformazione $(\mathbf{B}'_{\gamma'}, u') = h(\mathbf{B}_\gamma, u)$ nel nuovo spazio. Ciò viene raggiunto imponendo delle specifiche restrizioni sul tasso di accettazione tra lo stato esistente $\mathbf{B}_\gamma$ del modello γ ed uno specifico *centering point* $c_{\gamma,\gamma'}(\mathbf{B}_\gamma)$ in $\boldsymbol{\gamma}'$. I vincoli che vengono posti si distinguono poi per il grado del vincolo posto. Lo *zeroth-order method* impone

$$A((\boldsymbol{\gamma}, \mathbf{B}_\gamma), (\boldsymbol{\gamma}', c_{\gamma,\gamma'}(\mathbf{B}_\gamma))) = 1$$

in cui A è il tasso di accettazione per uno specifico tipo di movimento. In questo modo, i movimenti *across-model* sono incoraggiati. Metodi *first order* (e di ordine più alto) impongono invece la derivata prima (o maggiore) del tasso di accettazione pari a 0, ad esempio

$$\nabla A((\boldsymbol{\gamma}, \mathbf{B}_\gamma), (\boldsymbol{\gamma}', c_{\gamma,\gamma'}(\mathbf{B}_\gamma))) = 0.$$

Il vantaggio di questi metodi è che la probabilità di accettazione rimane alta in una regione attorno al *centering point* (Hastie & Green, 2012).

Il secondo metodo è il *saturated space approach*. L'idea in questo caso è quella di aumentare lo spazio con delle variabili ausiliarie, in modo da assicurare che tutti i modelli condividano la stessa dimensione p_{max}, pari a quella del modello completo, ed abbiano quindi una dimensione fissata. Vengono utilizzati poi metodi MCMC per creare una catena la cui distribuzione limite sia uguale ad una distribuzione target aumentata, che combini la distribuzione target π e la distribuzione delle variabili ausiliarie (Hastie & Green, 2012).

1.8.4 Stochastic search Markov chain Monte Carlo

Come già indicato precedentemente, la scelta di priori coniugate consente di integrare i parametri non di interesse in modo da ottenere la distribuzione a posteriori marginale $\pi(\boldsymbol{\gamma}|\mathbf{Y}, \mathbf{X})$, espressa come in equazione (A.13). Tale distribuzione consente di individuare i modelli migliori, ossia quelli che hanno una probabilità a posteriori maggiore. Un algoritmo di uso frequente prevede dei movimenti di tipo *add-delete-swap*, che consentono di esplorare la distribuzione a posteriori visitando una sequenza di modelli dove ad ogni iterazione il nuovo modello differisce dal precedente per l'inclusione e/o l'esclusione di una o più variabili (Tadesse & Vannucci, 2021). In particolare, dato un vettore di partenza γ_0 scelto casualmente, ad una iterazione generica i il nuovo modello è generato a partire dal precedente scegliendo casualmente uno dei seguenti movimenti di transizione:

- aggiunta o eliminazione; consiste nella scelta casuale di uno dei p indici in γ_{i-1}, cambiandone il valore, con il risultato che una nuova variabile sarà inclusa nel modello oppure ne verrà eliminata una già inclusa.
- scambio; vengono scelti uno 0 ed un 1 a caso dal vettore γ_{i-1} e tali valori vengono scambiati. Implica quindi l'aggiunta e l'eliminazione contestuale di una variabile.

Indicando con γ_i il nuovo modello candidato, la probabilità di accettazione viene calcolata come

$$\min\left[\frac{\pi(\gamma_i|\mathbf{Y},\mathbf{X})}{\pi(\gamma_{i-1}|\mathbf{Y},\mathbf{X})},1\right],$$

con la conseguenza che se il nuovo modello proposto avrà una probabilità a posteriori maggiore, allora la catena si muoverà verso il nuovo modello (Tadesse & Vannucci, 2021). In caso contrario il nuovo modello verrà accettato solo con una certa probabilità. Si ottiene quindi una lista di modelli visitati $\boldsymbol{\gamma}_{(i)}$, con $i = 1,\dots R$ ed R il numero di iterazioni effettuate, e le corrispondenti probabilità a posteriori. La selezione delle variabili a questo punto può essere conseguita scegliendo il modello con probabilità a posteriori più alta oppure scegliendo quelle variabili che sono state incluse in una proporzione prefissata di tutti i modelli visitati (Tadesse & Vannucci, 2021).

1.8.5 Previsioni

Come già precedentemente esplicitato, la simulazione con metodi MCMC consente di ottenere R configurazioni del vettore $\boldsymbol{\gamma}$, una per ogni modello accettato all'iterazione i-esima, ottenendo quindi $\boldsymbol{\gamma}_0,\dots,\boldsymbol{\gamma}_R$. La previsione quindi può essere effettuata sia sulla base di ogni singolo modello, in particolare risulta utile utilizzare il modello MaP, ossia con probabilità a posteriori massima, ma è possibile anche ricorrere al *Bayesian model averaging* (BMA), che consiste nel mediare le previsioni relative ad un insieme di modelli più probabili a posteriori (Tadesse & Vannucci, 2021). Data quindi la distribuzione predittiva definita in Equazione (1.24), le previsioni basate sul modello MaP saranno definite come la media della distribuzione predittiva, ossia

$$\widehat{\mathbf{Y}}_{0,\mathsf{MaP}} = \mathbf{X}_{0,\gamma}\widetilde{\mathbf{B}}_{\gamma}, \tag{1.28}$$

con $\widetilde{\mathbf{B}}_{\gamma} = \widetilde{\boldsymbol{\Sigma}}_{\gamma}\mathbf{X}_{\gamma}^{\intercal}\mathbf{Y}$ e $\widetilde{\boldsymbol{\Sigma}}_{\gamma} = (\mathbf{K}_{\gamma})^{-1} = (\mathbf{X}_{\gamma}^{\intercal}\mathbf{X}_{\gamma} + \mathbf{H}_{\gamma}^{-1})^{-1}$. Con il BMA invece si calcola il valore atteso della distribuzione predittiva $\pi\big(\mathbf{Y}_0|\mathbf{Y},\mathbf{X},\mathbf{X}_0\big)$, mediato sui k modelli più probabili a posteriori, ottenendo quindi

$$\widehat{\mathbf{Y}}_{0,\mathsf{BMA}} = \sum_{j=1}^{k}\mathbf{X}_{0,\gamma_j}\widetilde{\mathbf{B}}_{\gamma_j}\pi(\boldsymbol{\gamma}_j|\mathbf{Y},\mathbf{X}). \tag{1.29}$$

Anziché una media pesata è possibile anche utilizzare la media aritmetica per mediare le previsioni dei vari modelli.

Capitolo 2

Aspetti computazionali

2.1 Introduzione

Come già accennato precedentemente, il carico computazionale derivante dalla simulazione della distribuzione a posteriori può essere importante, soprattutto quando il numero delle variabili esplicative diventa elevato. Oltre all'utilizzo del *Gibbs sampling* che consente comunque di ottenere risultati soddisfacenti pur esplorando solo parte di tutte le 2^p possibili combinazioni dei possibili vettori $\boldsymbol{\gamma}$, è possibile introdurre anche ulteriori accorgimenti (Brown et al., 1998b).

Innanzitutto, è possibile eliminare le colonne di $\mathbf{X}$ relative alle variabili che non sono selezionate dal modello, definendo $\mathbf{X}_\gamma \in \mathbb{R}^{n \times p_\gamma}$. In secondo luogo, data la distribuzione a posteriori marginalizzata in Equazione (A.13), è d'interesse focalizzarsi sul termine $\mathbf{K}_\gamma = \mathbf{X}_\gamma^\intercal \mathbf{X}_\gamma + \mathbf{H}_\gamma^{-1}$, che può essere fattorizzato come

$$\mathbf{K}_\gamma = \widetilde{\mathbf{X}}_\gamma^\intercal \widetilde{\mathbf{X}}_\gamma = \begin{pmatrix} \mathbf{X}_\gamma \\ \mathbf{H}_\gamma^{-1/2} \end{pmatrix}^\intercal \begin{pmatrix} \mathbf{X}_\gamma \\ \mathbf{H}_\gamma^{-1/2} \end{pmatrix}, \tag{2.1}$$

con $\widetilde{\mathbf{X}}_\gamma \in \mathbb{R}^{(n+p_\gamma) \times p_\gamma}$. Tale termine, si presenta nella distribuzione a posteriori nella sua forma inversa e, come noto, l'inversione di una matrice piena risulta molto oneroso soprattuto nel caso in cui il numero dei parametri è molto elevato, in particolare se $p \gg n$. A tal fine, la fattorizzazione in equazione (2.1) consente di applicare la decomposizione QR a $\widetilde{\mathbf{X}}_\gamma \in \mathbb{R}^{(n+p_\gamma) \times p_\gamma}$ e di applicare delle forme di aggiornamento più veloci.

Infatti, si definirà $\widetilde{\mathbf{X}}_\gamma = \mathbf{QR}$, con $\widetilde{\mathbf{X}}_\gamma \in \mathbb{R}^{(n+p_\gamma) \times p_\gamma}$, $\mathbf{Q} \in \mathbb{R}^{(n+p_\gamma) \times (n+p_\gamma)}$, $\mathbf{R} \in \mathbb{R}^{(n+p_\gamma) \times p_\gamma}$, che, sostituito in (2.1), consente di ottenere

$$\mathbf{K}_\gamma = \widetilde{\mathbf{X}}_\gamma^\intercal \widetilde{\mathbf{X}}_\gamma = \mathbf{R}^\intercal \mathbf{Q}^\intercal \mathbf{Q}\mathbf{R} = \mathbf{R}^\intercal \mathbf{R}. \tag{2.2}$$

Sostituendola nel termine $\mathcal{Q}_\gamma$ si ottiene

$$\mathcal{Q}_\gamma = \mathbf{C}_0 + \mathbf{Y}^\intercal \mathbf{Y} + \mathbf{Y}^\intercal \mathbf{X}_\gamma \mathbf{R}^{-1} \mathbf{R}^{-\intercal} \mathbf{X}_\gamma^\intercal \mathbf{Y} = \mathbf{C}_0 + \mathbf{Y}^\intercal (\mathbf{I} - \mathbf{V}_\gamma \mathbf{V}_\gamma^\intercal) \mathbf{Y}, \tag{2.3}$$

con $\mathbf{V} = \mathbf{X}_\gamma \mathbf{R}^{-1}$. Il vantaggio evidente risulta dal fatto che $\mathbf{R}$ è una matrice triangolare superiore ed è necessaria soltanto la sua inversione, che risulta essere meno onerosa rispetto alla medesima quantità non fattorizzata, $\mathbf{K}_\gamma$, che invece è piena.

Inoltre, possono essere utilizzate delle forme di aggiornamento della sola matrice $\mathbf{R}$ nel momento in cui vengano aggiunte o sottratte colonne dalla matrice $\mathbf{X}$ ad ogni modello proposto di diversa dimensione.

2.2 Decomposizione QR

La decomposizione QR di una generica matrice rettangolare $\mathbf{X} \in \mathbb{R}^{n \times p}$ consiste nella fattorizzazione di tale matrice nel prodotto di una matrice ortogonale $\mathbf{Q} \in \mathbb{R}^{n \times n}$ e di una matrice trapezoidale superiore $\mathbf{R} \in \mathbb{R}^{n \times p}$, tale che

$$\mathbf{X} = \mathbf{Q}\mathbf{R} = \begin{bmatrix} \mathbf{Q}_1 & \mathbf{Q}_2 \end{bmatrix} \begin{bmatrix} \mathbf{R}_1 \\ 0 \end{bmatrix} = \mathbf{Q}_1 \mathbf{R}_1, \tag{2.4}$$

con $\mathbf{Q}_1 \in \mathbb{R}^{n \times p}$, $\mathbf{Q}_2 \in \mathbb{R}^{n \times (n-p)}$ e $\mathbf{R}_1 \in \mathbb{R}^{p \times p}$. Per ottenere tale decomposizione esistono diversi metodi, che sfruttano le trasformazioni di Householder (anche a blocchi) e le rotazioni di Givens, nonché il più noto e classico algoritmo di Gram-Schmidt (Golub & Van Loan, 1996).

2.2.1 Rotazioni di Givens

Si vuole quindi calcolare la decomposizione QR di $\mathbf{X} \in \mathbb{R}^{n \times p}$, con $n > p$ e a rango pieno, applicando una matrice di trasformazione ortogonale $\mathbf{Q}^\intercal$ tale che $\mathbf{Q}^\intercal \mathbf{X} = \mathbf{R}$, dove $\mathbf{Q}$ è il prodotto di matrici ortogonali scelte in modo da trasformare $\mathbf{X}$ in una matrice triangolare superiore $\mathbf{R}$ (Hammarling & Lucas, 2008).

Il primo metodo che viene riportato per effettuare la decomposizione QR sfrutta l'utilizzo delle rotazioni di Givens, al fine di introdurre degli zeri al di sotto della diagonale principale un elemento alla volta.

Tali rotazioni di Givens sono delle correzioni di rango 2 ad una matrice identità della forma

$$\mathbf{G}(i,j) = \begin{bmatrix} 1 & \cdots & 0 & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & & \vdots & & \vdots \\ 0 & \dots & c & \dots & s & \dots & 0 \\ \vdots & & \vdots & \ddots & \vdots & & \vdots \\ 0 & \dots & -s & \dots & c & \dots & 0 \\ \vdots & & \vdots & & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & \dots & 0 & \dots & 1 \end{bmatrix} \tag{2.5}$$

dove $c = cos(\theta)$ e $s = sin(\theta)$ per un certo θ, ed è conseguentemente ortogonale. Premoltiplicare per $\mathbf{G}(i,j)^\intercal$ equivale quindi ad una rotazione in senso antiorario di θ radianti nel piano delle coordinate (i,j). Infatti, se $\boldsymbol{x} \in \mathbb{R}^n$ e $\boldsymbol{y} = \mathbf{G}(i,j)^\intercal \boldsymbol{x}$, allora

$$y_k = \begin{cases} cx_i - sx_j, & k = i \\ cx_i - sx_j, & k = j \\ x_k, & j \neq i, j. \end{cases} \tag{2.6}$$

Si può quindi forzare y_j a zero imponendo

$$c = \frac{x_i}{\sqrt{x_i^2 + x_k^2}}, \qquad s = \frac{-x_j}{\sqrt{x_i^2 + x_k^2}}, \tag{2.7}$$

Un modo più efficiente di calcolare le rotazioni di Givens è tuttavia descritto dall'Algoritmo 1, che richiede 5 *flops* ed una sola radice quadrata e non necessita dell'utilizzo di funzioni trigonometriche (Golub & Van Loan, 1996). Con *flops*, acronimo di *floating-point operations per second*, si intende una misura delle prestazioni di calcolo di un sistema informatico, espressa come il numero di operazioni in virgola mobile che può eseguire in un secondo. Quest'ultime coinvolgono calcoli matematici che includono addizioni, sottrazioni, moltiplicazioni, divisioni e altre funzioni matematiche complesse.

L'Algoritmo 2 invece consente di ottenere la fattorizzazione QR tramite le rotazioni di Givens. Per azzerare tutti gli elementi presenti sotto la diagonale principale, così da rendere $\mathbf{X}$ una matrice trapezioidale superiore, è necessaria una matrice di Givens per ogni elemento subdiagonale di $\mathbf{X}$ e applicarle con l'ordine seguente

$$\mathbf{G}(p, p+1)^\intercal \dots \mathbf{G}(n-1, n)^\intercal \mathbf{G}(1,2)^\intercal \dots \mathbf{G}(m-1, m)^\intercal \mathbf{X} = \mathbf{Q}^\intercal \mathbf{X} = \mathbf{R}.$$

Algoritmo 1 *Rotazione di Givens*

Dati gli scalari a e b, questa funzione calcola $c = cos(\theta)$ e $s = sin(\theta)$ tali che

$$\begin{bmatrix} c & s \\ -s & c \end{bmatrix}^\intercal \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} r \\ 0 \end{bmatrix}. \tag{2.8}$$

function GIVENS(a, b)
  if $b = 0$ **then**
    $c = 1;\ s = 0$
  else
    if $|b| > |a|$ **then**
      $\tau = -a/b;\ s = 1/\sqrt{1+\tau^2};\ c = s\tau$
    else
      $\tau = -b/a;\ c = 1/\sqrt{1+\tau^2};\ s = c\tau$
    end if
  end if
return c, s
end function

Algoritmo 2 *Givens QR*

Data una matrice $\mathbf{X} \in \mathbb{R}^{n\times p}$ con $n \geq p$, l'algoritmo seguente sovrascrive $\mathbf{X}$ con $\mathbf{Q}^\intercal\mathbf{X} = \mathbf{R}$, con $\mathbf{R}$ matrice triangolare superiore e $\mathbf{Q}$ ortogonale.

function GIVENSQR$(\mathbf{X})$
  $n = \text{nrow}(\mathbf{X})$, $p = \text{ncol}(\mathbf{X})$
  $\mathbf{Q} = \mathbf{I}_n$
  for $j = 1 : p$ **do**
    for $i = m : -1 : j + 1$ **do**
      $[c, s] = \text{GIVENS}(\mathbf{X}(i-1, j), \mathbf{X}(i, j))$
      $\mathbf{X}(i-1 : i, j : n) = \begin{bmatrix} c & s \\ -s & c \end{bmatrix}^\intercal \mathbf{X}(i-1 : i, j : n)$
      $\mathbf{Q}(i-1 : i, j : n) = \mathbf{Q}(i-1 : i, j : n) \begin{bmatrix} c & s \\ -s & c \end{bmatrix}$
    end for
  end for
return $\mathbf{R} = \mathbf{X}, \mathbf{Q}$
end function

2.2.2 Trasformazioni di Householder

Le rotazioni di Givens sono particolarmente utili quando si vogliono azzerare degli specifici elementi. Un approccio più efficiente per ottenere la fattorizzazione QR è quello di utilizzare le trasformazioni di Householder, delle modificazioni di rango 1 della matrice identità che introducono degli zeri in tutti gli elementi subdiagonali di una colonna contemporaneamente.

Sia $\boldsymbol{v} \in \mathbb{R}^n$ un vettore con elementi diversi da 0. Una trasformazione di Householder è una matrice $n \times n$ della forma

$$\mathbf{H} = \mathbf{I}_n - \beta \boldsymbol{v}\boldsymbol{v}^\intercal, \qquad \beta = \frac{2}{\boldsymbol{v}^\intercal \boldsymbol{v}}, \tag{2.9}$$

dove $\boldsymbol{v}$ è il vettore di Householder. Se un vettore $\boldsymbol{x}$ viene moltiplicato per $\mathbf{H}$, esso viene riflesso nell'iperpiano $\mathsf{span}\{\boldsymbol{v}\}^\perp$. Le matrici di Householder sono simmetriche e ortogonali. Supponendo di avere $0 \neq \boldsymbol{x} \in \mathbb{R}^n$ e si voglia che

$$\mathbf{H}\boldsymbol{x} = \left(\mathbf{I}_n - \frac{2\boldsymbol{v}\boldsymbol{v}^\intercal}{\boldsymbol{v}^\intercal \boldsymbol{v}}\right)\boldsymbol{x} = \boldsymbol{x} - \frac{2\boldsymbol{v}^\intercal \boldsymbol{x}}{\boldsymbol{v}^\intercal \boldsymbol{v}}\boldsymbol{v}$$

sia un multiplo di $\boldsymbol{e}_1$, corrispondente alla prima colonna di una matrice identità di dimensione $n \times n$ è possibile concludere che $\boldsymbol{v} \in \mathsf{span}\{\boldsymbol{x}, \boldsymbol{e}_1\}$. Imponendo poi $\boldsymbol{v} = \boldsymbol{x} + \alpha \boldsymbol{e}_1$, si ottiene che $\boldsymbol{v}^\intercal \boldsymbol{x} = \boldsymbol{x}^\intercal \boldsymbol{x} + \alpha \boldsymbol{x}_1$ e $\boldsymbol{v}^\intercal \boldsymbol{v} = \boldsymbol{x}^\intercal \boldsymbol{x} + 2\alpha \boldsymbol{x}_1 + \alpha^2$. In questo modo si avrà

$$\mathbf{H}\boldsymbol{x} = \left(1 - 2\frac{\boldsymbol{x}^\intercal \boldsymbol{x} + \alpha \boldsymbol{x}_1}{\boldsymbol{x}^\intercal \boldsymbol{x} + 2\alpha \boldsymbol{x}_1 + \alpha^2}\right)\boldsymbol{x} - 2\alpha \frac{\boldsymbol{v}^\intercal \boldsymbol{x}}{\boldsymbol{v}^\intercal \boldsymbol{v}}\boldsymbol{e}_1 \tag{2.10}$$

$$= \left(\frac{\alpha^2 - ||\boldsymbol{x}||_2^2}{\boldsymbol{x}^\intercal \boldsymbol{x} + 2\alpha \boldsymbol{x}_1 + \alpha^2}\right)\boldsymbol{x} - 2\alpha \frac{\boldsymbol{v}^\intercal \boldsymbol{x}}{\boldsymbol{v}^\intercal \boldsymbol{v}}\boldsymbol{e}_1. \tag{2.11}$$

Al fine di imporre il coefficiente di $\boldsymbol{x}$ pari a 0, si impone $\alpha = \pm||\boldsymbol{x}||_2$ cosicché $\boldsymbol{v} = \boldsymbol{x} \pm ||\boldsymbol{x}||_2 \boldsymbol{e}_1$ e quindi

$$\mathbf{H}\boldsymbol{x} = \left(\mathbf{I}_n - \frac{2\boldsymbol{v}\boldsymbol{v}^\intercal}{\boldsymbol{v}^\intercal \boldsymbol{v}}\right)\boldsymbol{x} = \pm||\boldsymbol{x}||_2 \boldsymbol{e}_1.$$

Ci sono poi degli importanti dettagli pratici associati al calcolo delle matrici di Householder, tra cui la determinazione del vettore di Householder e in particolare la definizione del segno di $\boldsymbol{v}$. Imponendo $\boldsymbol{v}_1 = \boldsymbol{x}_1 - ||\boldsymbol{x}||_2$ si ha che $\mathbf{H}\boldsymbol{x}$ è un multiplo positivo di $\boldsymbol{e}_1$.

Si usa poi normalizzare il vettore di Householder in modo che $\boldsymbol{v}(1) = 1$, così da poter conservare $\boldsymbol{v}(2 : m)$ (ossia la parte essenziale del vettore di Householder) dove gli zeri sono stati introdotti in $\boldsymbol{x}$, ossia $\boldsymbol{x}(2 : m)$. Il procedimento per calcolare il vettore di Householder è riportato nell'Algoritmo 3 e richiede $3m$ flops (Golub & Van Loan, 1996).

L'adattamento della decomposizione QR mediante una sequenza di trasformazioni

Algoritmo 3 *Vettore di Householder*

function HOUSEHOLDER($\boldsymbol{x}$)
  $m = \text{length}(\boldsymbol{x})$, $\sigma = \boldsymbol{x}(2:m)^\intercal \boldsymbol{x}(2:m)$, $\boldsymbol{v} = \begin{bmatrix} 1 \\ x(2:m) \end{bmatrix}$
  if $\sigma = 0$ and $\boldsymbol{x}(1) \geq 0$ **then**
    $\beta = 0$
  else if $\sigma = 0$ and $\boldsymbol{x}(1) < 0$ **then**
    $\beta = 2$
  else
    $\mu = \sqrt{\boldsymbol{x}(1)^2 + \sigma}$
    if $\boldsymbol{x}(1) \leq 0$ **then**
      $\boldsymbol{v}(1) = \boldsymbol{x}(1) - \mu$
    else
      $\boldsymbol{v}(1) = -\sigma/(\boldsymbol{x}(1) + \mu)$
    end if
    $\beta = 2\boldsymbol{v}(1)^2/(\sigma + \boldsymbol{v}(1)^2)$
    $\boldsymbol{v} = \boldsymbol{v}/\boldsymbol{v}(1)$
  end if
return $\boldsymbol{v}, \beta$
end function

di Householder applicate alle colonne di $\mathbf{X}$ è invece descritto dall'Algoritmo 4. Tale algoritmo richiede $2n^2(m - n/3)$ flops (Golub & Van Loan, 1996). In generale, per ottenere la decomposizione QR di una generica matrice $\mathbf{X}$, è necessario applicare p matrici di Householder $\mathbf{H}_j$, con ogni $\mathbf{H}_j$, con $j = 1, \ldots, p$ relative alla j-esima colonna di $\mathbf{X}$ così da ottenere

$$\mathbf{H}_p \ldots \mathbf{H}_1 \mathbf{X} = \mathbf{Q}^\intercal \mathbf{X} = \mathbf{R}$$

in cui ogni $\mathbf{H}_j$ ha vettore di Householder $\boldsymbol{v}_j = \frac{\widetilde{\boldsymbol{v}}_j}{\widetilde{\boldsymbol{v}}_j(j)}$ tale che

$$\widetilde{\boldsymbol{v}}_j = \begin{bmatrix} \mathbf{0}_{j-1} \\ \boldsymbol{x}_j(j) + \mathsf{sign}(\boldsymbol{x}_j(j)) ||\boldsymbol{x}_j(j:n)||_2 \\ \boldsymbol{x}_j(j+1:n) \end{bmatrix}. \tag{2.12}$$

In realtà, come visibile nell'algoritmo, non è necessaria la formazione esplicita della matrice di Householder. Infatti, sia la premoltiplicazione che la postmoltiplicazione di $\mathbf{H}$ ad una matrice $\mathbf{X}$ implica un aggiornamento di rango 1, rispettivamente

$$\mathbf{HX} = (\mathbf{I}_n - \beta \boldsymbol{v}\boldsymbol{v}^\intercal)\mathbf{X} = \mathbf{X} - (\beta \boldsymbol{v})(\boldsymbol{v}^\intercal \mathbf{X}), \tag{2.13}$$

$$\mathbf{XH} = \mathbf{X}(\mathbf{I}_n - \beta \boldsymbol{v}\boldsymbol{v}^\intercal) = \mathbf{X} - (\mathbf{X}\boldsymbol{v})(\beta \boldsymbol{v})^\intercal. \tag{2.14}$$

Algoritmo 4 *Householder QR*

Data una matrice $\mathbf{X} \in \mathbb{R}^{n \times p}$ *con* $n \geq p$*, l'algoritmo seguente trova le matrici di Householder* $\mathbf{H}_1, \ldots, \mathbf{H}_p$ *tali che se* $\mathbf{Q} = \mathbf{H}_1, \ldots, \mathbf{H}_p$*, allora* $\mathbf{Q}^\intercal \mathbf{X} = \mathbf{R}$ *è triangolare superiore. La parte triangolare superiore di* $\mathbf{A}$ *è sovrascritta dalla parte triangolare superiore di* $\mathbf{R}$ *e le componenti* $j+1 : n$ *del* $j-esimo$ *vettore di Householder sono immagazzinati in* $\mathbf{X}(j+1 : m, j),\ j < m$*.*

function HOUSEHOLDERQR($\mathbf{X}$)
  $\mathbf{Q} = \mathbf{I}_n$
  for $j = 1 : p$ **do**
    $[\boldsymbol{v}, \beta] =$ HOUSEHOLDER($\mathbf{X}(j : n, j)$)
    $\mathbf{X}(j : n, j : p) = (\mathbf{I} - \beta \boldsymbol{v}\boldsymbol{v}^\intercal)\mathbf{X}(j : n, j : p)$
    if $j < n$ **then**
      $\mathbf{X}(j+1 : n, j) = \boldsymbol{v}(2 : p - j + 1)$
    end if
    $\mathbf{Q}(1 : n, j : n) = (\mathbf{I} - \beta \boldsymbol{v})(\boldsymbol{v}^\intercal \mathbf{Q}(j : n, j : n))$
  end for
return $\mathbf{R} = \mathbf{X}, \mathbf{Q} = \mathbf{Q}$
end function

2.3 Algoritmi di aggiornamento QR

Nella sezione seguente verranno presentati gli algoritmi che consentono di aggiornare la decomposizione QR in seguito all'aggiunta o all'eliminazione di una o più colonne dalla matrice $\mathbf{X}$ originale.

2.3.1 Aggiunta di colonne

Nel caso si voglia aggiungere una colonna alla matrice di disegno $\mathbf{X}$ si avrà il problema di aggiornare $\mathbf{X} = \mathbf{QR}$ dopo l'aggiunta della colonna $\boldsymbol{u} \in \mathbb{R}^n$ in posizione k, con $1 \leq k \leq p+1$. Questa nuova matrice potrà quindi essere definita come

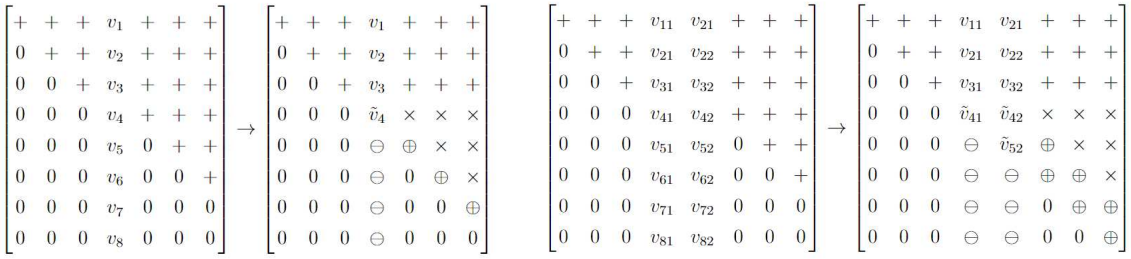
$$\widetilde{\mathbf{X}} = \begin{bmatrix} \mathbf{X}(, 1 : k-1) & \boldsymbol{u} & \mathbf{X}(, k : p) \end{bmatrix} \tag{2.15}$$

e si avrà di conseguenza

$$\mathbf{Q}^\intercal \widetilde{\mathbf{X}} = \begin{bmatrix} \mathbf{R}(, 1 : k-1) & \boldsymbol{v} & \mathbf{R}(, k : p) \end{bmatrix} \tag{2.16}$$

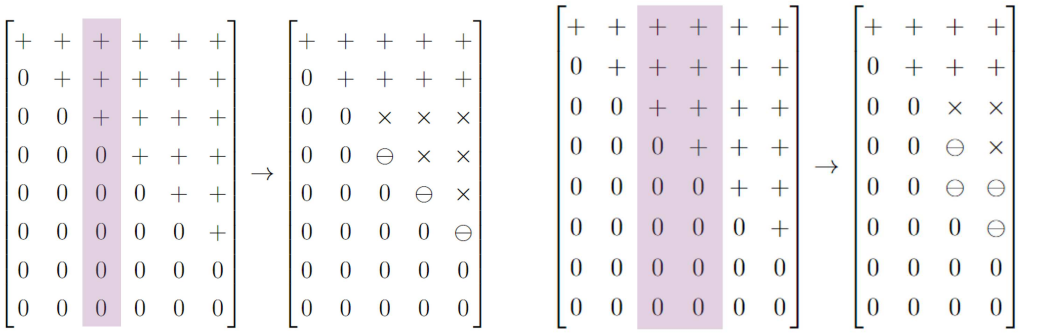
con $\boldsymbol{v} = \mathbf{Q}^\intercal \boldsymbol{u}$. Così facendo si rende necesssario azzerare gli elementi del vettore $\boldsymbol{v}$ presenti sotto la diagonale principale, ossia $\mathbf{v}(k+1 : n)$, utilizzando le rotazioni di Givens, come visibile in Figura 2.1(a). Il procedimento viene descritto nell'Algoritmo 5. Esso si sostanzia nell'applicare $n - k$ rotazioni di Givens nel seguente modo:

$$\widetilde{\mathbf{R}} = \mathbf{G}(k, k+1)^\intercal \ldots \mathbf{G}(n-1, n)^\intercal \mathbf{Q}^\intercal \widetilde{\mathbf{X}} = \widetilde{\mathbf{Q}}^\intercal \widetilde{\mathbf{X}}, \tag{2.17}$$



(a) Aggiunta di una colonna

(b) Aggiunta di $m = 2$ colonne in posizione $k = 4$



(c) Eliminazione di una colonna

(d) Eliminazione di $m = 2$ colonne in posizione $k = 3$

FIGURA 2.1: Rappresentazione grafica dell'aggiunta di una colonna 2.1(a) in posizione $k = 4$ e di $m = 2$ colonne in posizione $k = 4$ 2.1(b), dell'eliminazione di una colonna 2.1(c) in posizione $k = 3$ e di $m = 2$ colonne in posizione $k = 3$ 2.1(d). $\ominus$ indica gli elementi che devono essere azzerati, $\times$ gli elementi che vengono modificati attraverso le rotazioni di Givens e $\oplus$ gli elementi che diventano diversi da 0 in seguito all'aggiornamento ma che originariamente erano pari a 0.

con $\widetilde{\mathbf{Q}} = \mathbf{Q}\mathbf{G}(n-1, n)\dots\mathbf{G}(k, k+1)$.

Nel caso in cui si debbano invece aggiungere m colonne $\mathbf{U} \in \mathbb{R}^{n\times m}$ a partire dalla posizione k fino alla $k + m - 1$, la nuova matrice diventerà

$$\widetilde{\mathbf{X}} = \begin{bmatrix}\mathbf{X}(, 1 : k-1) & \mathbf{U} & \mathbf{X}(, k : p)\end{bmatrix} \tag{2.18}$$

e si avrà di conseguenza

$$\mathbf{Q}^\intercal\widetilde{\mathbf{X}} = \begin{bmatrix}\mathbf{R}(, 1 : k-1) & \mathbf{V} & \mathbf{R}(, k : p)\end{bmatrix} \tag{2.19}$$

con $\mathbf{V} = \mathbf{Q}^\intercal\mathbf{U}$. Anche in questo caso si rende necessario azzerare gli elementi di $\mathbf{V}$ sotto la diagonale principale ed aggiornare $\mathbf{R}(i + 1 : i + m, i)$ con $i = k, \dots, p$, utilizzando le rotazioni di Givens per azzerare determinati elementi, come visibile nell'esempio in Figura 2.1(b). Il procedimento è invece riportato nell'Algoritmo 6. Ciò equivale ad

Algoritmo 5 *Aggiunta della colonna* $\boldsymbol{u}$ *ad* $\mathbf{X}$ *in posizione* k

è

function ADDCOLQR($\mathbf{Q}, \mathbf{R}, k, \boldsymbol{u}$)
  $\boldsymbol{u} = \mathbf{Q}^\intercal \boldsymbol{u}$
  for $i = n : k+1$ **do**
    $c, s =$ GIVENS($\boldsymbol{u}(i-1), \boldsymbol{u}(i)$)
    $\boldsymbol{u}(i-1) = c\ \boldsymbol{u}(i-1) - s\ \boldsymbol{u}(i)$
    if any($\mathbf{R}(i,) \neq 0$) **then**
      if $i \leq p+1$ **then**
        $\mathbf{R}(i-1:i, i-1:p) = \begin{bmatrix} c & s \\ -s & c \end{bmatrix}^\intercal \mathbf{R}(i-1:i, i-1:p)$
      end if
    end if
    $\mathbf{Q}(1:n, i-1:i) = \mathbf{Q}(1:n, i-1:i) \begin{bmatrix} c & s \\ -s & c \end{bmatrix}$
  end for
  if $k = 1$ **then**
    $\widetilde{\mathbf{R}}$ = parte triangolare superiore di $\begin{bmatrix} \boldsymbol{u} & \mathbf{R} \end{bmatrix}$
  else if $k = p+1$ **then**
    $\widetilde{\mathbf{R}}$ = parte triangolare superiore di $\begin{bmatrix} \mathbf{R} & \boldsymbol{u} \end{bmatrix}$
  else
    $\widetilde{\mathbf{R}}$ = parte triangolare superiore di $\begin{bmatrix} \mathbf{R}(,1:k-1)\ \boldsymbol{u}\ \mathbf{R}(,k:p) \end{bmatrix}$
  end if
return $\widetilde{\mathbf{Q}} = \mathbf{Q},\ \widetilde{\mathbf{R}}$
end function

applicare le rotazioni di Givens nel seguente modo:

$$\begin{aligned} \widetilde{\mathbf{R}} &= \mathbf{G}(k+m-1, k+m)^\intercal \dots \mathbf{G}(n-1,n)^\intercal \dots \mathbf{G}(k,k+1)^\intercal \dots \mathbf{G}(n-1,n)^\intercal \mathbf{Q}^\intercal \widetilde{\mathbf{X}} \\ &= \widetilde{\mathbf{Q}}^\intercal \widetilde{\mathbf{X}}, \end{aligned} \tag{2.20}$$

con $\widetilde{\mathbf{Q}} = \mathbf{Q}\mathbf{G}(n-1,n)^\intercal \dots \mathbf{G}(k,k+1)^\intercal \dots \mathbf{G}(n-1,n)^\intercal \dots \mathbf{G}(k+m-1,k+m)^\intercal$.

2.3.2 Eliminazione di colonne

Se si vuole invece eliminare una variabile dalla matrice di disegno $\mathbf{X}$ in posizione $k,\ k \neq p$, si otterrà

$$\widetilde{\mathbf{X}} = \begin{bmatrix} \mathbf{X}(,1:k-1) & \mathbf{X}(,k+1:p) \end{bmatrix} \tag{2.21}$$

e

$$\mathbf{Q}^\intercal \widetilde{\mathbf{X}} = \begin{bmatrix} \mathbf{R}(,1:k-1) & \mathbf{R}(,k+1:p), \end{bmatrix} \tag{2.22}$$

Algoritmo 6 *Aggiunta delle colonne* $\mathbf{U}$ *ad* $\mathbf{X}$ *a partire dalla posizione* k

function ADDCOLSQR($\mathbf{Q}, \mathbf{R}, k, \mathbf{U}$)
  $\mathbf{U} = \mathbf{Q}^\intercal \mathbf{U}$
  for $j = 1 : m$ **do**
    for $j = n : k + j$ **do**
      $c, s =$ GIVENS($\mathbf{U}(i-1, j), \mathbf{U}(i, j)$)
      $\mathbf{U}(i-1) = c\ \mathbf{U}(i-1, j) - s\ \mathbf{U}(i, j)$
      if $j < m$ **then**
        $\mathbf{U}(i-1 : i, j+1 : m) = \begin{bmatrix} c & s \\ -s & c \end{bmatrix}^\intercal \mathbf{U}(i-1 : i, j+1 : m)$
      end if
      if $\mathbf{R}$ has a non-zero row **then**
        if $i \leq p + j$ **then**
          $\mathbf{R}(i-1 : i, i-j : p) = \begin{bmatrix} c & s \\ -s & c \end{bmatrix}^\intercal \mathbf{R}(i-1 : i, i-j : p)$
        end if
      end if
      $\mathbf{Q}(, i-1 : i) = \mathbf{Q}(, i-1 : i) \begin{bmatrix} c & s \\ -s & c \end{bmatrix}$
    end for
  end for
  if $k = 1$ **then**
    $\widetilde{\mathbf{R}}$ = parte triangolare superiore di $\begin{bmatrix} \mathbf{U} & \mathbf{R} \end{bmatrix}$
  else if $k = p + 1$ **then**
    $\widetilde{\mathbf{R}}$ = parte triangolare superiore di $\begin{bmatrix} \mathbf{R} & \mathbf{U} \end{bmatrix}$
  else
    $\widetilde{\mathbf{R}}$ = parte triangolare superiore di $\begin{bmatrix} \mathbf{R}(, 1 : k-1)\ \mathbf{U}\ \mathbf{R}(, k : p) \end{bmatrix}$
  end if
return $\widetilde{\mathbf{Q}} = \mathbf{Q},\ \widetilde{\mathbf{R}}$
end function

necessitando di $n - k$ matrici di Givens in modo da azzerare gli $n - k$ elementi sotto la diagonale principale. Il procedimento è espresso nell'Algoritmo 7 ed è visibile graficamente in un esempio in Figura 2.1(c). Ciò consiste nell'applicare le rotazioni di Givens nel seguente modo:

$$\widetilde{\mathbf{R}} = \mathbf{G}(p-1, p)^\intercal \dots \mathbf{G}(k, k+1)^\intercal \mathbf{Q}^\intercal \widetilde{\mathbf{X}} = \widetilde{\mathbf{Q}}^\intercal \widetilde{\mathbf{X}}, \tag{2.23}$$

con $\widetilde{\mathbf{Q}} = \mathbf{Q}\mathbf{G}(k, k+1) \dots \mathbf{G}(p-1, p)$.

Eliminando invece un blocco di colonne di dimensione m a partire dalla k-esima colonna di $\mathbf{X}$, si ottiene

$$\widetilde{\mathbf{X}} = \begin{bmatrix} \mathbf{X}(, 1 : k-1) & \mathbf{X}(, k+m : p) \end{bmatrix} \tag{2.24}$$

Algoritmo 7 *Eliminazione di una colonna da X in posizione k*

function DELCOLQR($\mathbf{Q}, \mathbf{R}, k$)
  if $k = p$ **then**
    $\widetilde{\mathbf{R}} = \mathbf{R}(, 1 : k-1)$
    $\widetilde{\mathbf{Q}} = \mathbf{Q}$
  else
    $\mathbf{R} = \mathbf{R}(, -k)$
    for $j = k : p-1$ **do**
      $c, s =$ GIVENS($\mathbf{R}(j,j), \mathbf{R}(j+1,j)$)
      $\mathbf{R}(j,j) = c\,\mathbf{R}(j,j) - s\,\mathbf{R}(j+1,j)$
      $\mathbf{R}(j : j+1, j+1 : n-1) = \begin{bmatrix} c & s \\ -s & c \end{bmatrix}^{\intercal} \mathbf{R}(j : j+1, j+1 : n-1)$
      $\mathbf{Q}(1 : m, j : j+1) = \mathbf{Q}(1 : m, j : j+1) \begin{bmatrix} c & s \\ -s & c \end{bmatrix}$
    end for
    $\widetilde{\mathbf{R}}$ = parte triangolare superiore di $\mathbf{R}(1 : n, 1 : p-1)$
    $\widetilde{\mathbf{Q}} = \mathbf{Q}$
  end if
return $\widetilde{\mathbf{Q}}$, $\widetilde{\mathbf{R}}$
end function

e si avrà

$$\mathbf{Q}^{\intercal}\widetilde{\mathbf{X}} = \begin{bmatrix} \mathbf{R}(, 1 : k-1) & \mathbf{R}(, k+m : p) \end{bmatrix}. \tag{2.25}$$

Anche in questo caso si rende necessario azzerare degli elementi sotto la diagonale principale, questa volta tuttavia mediante $n-m-k+1$ matrici di Householder $\mathbf{H}_j \in \mathbb{R}^{n \times n}$, come esplicato in Figura 2.1(d). Il procedimento, riassunto nell'Algoritmo 8, si sostanzia algebricamente nel seguente modo:

$$\widetilde{\mathbf{R}} = \mathbf{H}_{p-m}(p-m+1, p) \dots \mathbf{H}_k(k+1, k+m)\mathbf{Q}^{\intercal}\widetilde{\mathbf{X}} = \widetilde{\mathbf{Q}}^{\intercal}\widetilde{\mathbf{X}}, \tag{2.26}$$

con $\widetilde{\mathbf{Q}} = \mathbf{Q}\mathbf{H}_k(k+1, k+m)^{\intercal} \dots \mathbf{H}_{p-m}(p-m+1, p)^{\intercal}$, dove $\mathbf{H}_j(l, n)$, con $j = k, \dots p-m$, è la matrice di Householder con vettore di normalizzazione come definito in (2.12).

2.4 Thin QR

Un altro metodo utile ad aggiornare le colonne di $\widetilde{\mathbf{X}}$ sfrutta invece la decomposizione *thin QR*. Sia $\mathbf{X} \in \mathbb{R}^{n \times p}$ un matrice con rango di colonna pieno, allora la fattorizzazione *thin QR*, definita come

$$\mathbf{X} = \mathbf{Q}_1\mathbf{R}_1$$

Algoritmo 8 *Eliminazione di m colonne da X a partire dalla posizione k*

function DELCOLSQR($\mathbf{Q}, \mathbf{R}, k, m$)
  if $k = p - m + 1$ **then**
    $\widetilde{\mathbf{R}} = \mathbf{R}(, 1 : k - 1)$
    $\widetilde{\mathbf{Q}} = \mathbf{Q}$
  else
    $\mathbf{R}(1 : n, k : p - m) = \mathbf{R}(1 : n, k + m : p)$
    for $j = k : p - m$ **do**
      $\boldsymbol{v}, \beta =$ HOUSEHOLDER($\mathbf{R}(j, j), \mathbf{R}(j + 1 : j + p, j)$)
      $\mathbf{R}(j, j) = \mathbf{R}(j, j) - \beta\ \mathbf{R}(j, j) - \beta\ \boldsymbol{v}^\intercal\ \mathbf{R}(j + 1 : j + p, j)$
      if $j < p - m$ **then**
        $\mathbf{R}(j : j + m, j + 1 : p - m) = \mathbf{R}(j : j + m, j + 1 : p - m) -$
        $\beta \begin{bmatrix} 1 \\ \boldsymbol{v} \end{bmatrix} \begin{bmatrix} 1 & \boldsymbol{v}^\intercal \end{bmatrix} \mathbf{R}(j : j + m, j + 1 : p - m)$
      end if
      $\mathbf{Q}(1 : m, j : j + m) = \mathbf{Q}(1 : m, j : j + m) -$
      $\beta(j)(\mathbf{Q}(1 : m, j : j + m) \begin{bmatrix} 1 \\ \boldsymbol{v} \end{bmatrix}) \begin{bmatrix} 1 & \boldsymbol{v}^\intercal \end{bmatrix}$
    end for
    $\widetilde{\mathbf{R}}$ = parte triangolare superiore di $\mathbf{R}(1 : n, 1 : p - m)$
    $\widetilde{\mathbf{Q}} = \mathbf{Q}$
  end if
return $\widetilde{\mathbf{Q}}$, $\widetilde{\mathbf{R}}$
end function

è unica, dove $\mathbf{Q}_1 \in \mathbf{R}^{n \times p}$ ha colonne ortonormali e $\mathbf{R}_1 \in \mathbb{R}^{p \times p}$ è triangolare superiore con elementi positivi sulla diagonale. Inoltre, $\mathbf{R}_1 = \mathbf{C}^\intercal$ dove $\mathbf{C}$ è il fattore di Cholesky triangolare inferiore di $\mathbf{X}^\intercal\mathbf{X}$ (Bernardi et al., 2023).

Utilizzando tale fattorizzazione il costo computazionale può essere ridotto grazie all'applicazione di algoritmi che aggiornano soltanto la matrice $\mathbf{R}_1$, senza la necessità di salvare e aggiornare anche la matrice $\mathbf{Q}$.

2.5 Algoritmi di aggiornamento thin QR

2.5.1 Aggiunta di colonne

Si tratterà di seguito il caso in cui si vuole aggiungere una o più colonne alla matrice $\mathbf{X}$ originale, analogamente a quanto avviene negli algoritmi presentati nella Sezione 2.3.

Supponendo quindi di voler aggiungere una colonna $\boldsymbol{u} \in \mathbb{R}^n$ in posizione $k = p + 1$ ad una data matrice di disegno $\mathbf{X} \in \mathbb{R}^{n \times p}$, si avrà in questo caso $\widetilde{\mathbf{X}} = \begin{bmatrix} \mathbf{X} & \boldsymbol{u} \end{bmatrix}$, con

$\widetilde{\mathbf{X}} \in \mathbb{R}^{n\times(p+1)}$ e coerentemente

$$\mathbf{Q}^\intercal \widetilde{\mathbf{X}} = \begin{bmatrix}\mathbf{R} & \boldsymbol{v}\end{bmatrix} = \begin{bmatrix}\mathbf{R}_1 & \boldsymbol{v}_1 \\ 0 & \boldsymbol{v}_2\end{bmatrix} = \widetilde{\mathbf{R}}_1^+ \tag{2.27}$$

con $\boldsymbol{v} = \mathbf{Q}^\intercal \boldsymbol{u}$, $\boldsymbol{v}_1 = \mathbf{Q}_1^\intercal \boldsymbol{u}$ e $\boldsymbol{v}_2 = \mathbf{Q}_2^\intercal \boldsymbol{u}$, dato che $\mathbf{Q}^\intercal = \begin{bmatrix}\mathbf{Q}_1^\intercal \\ \mathbf{Q}_2^\intercal\end{bmatrix}$.

$\mathbf{R}_1^+ \in \mathbb{R}^{(p+1)\times(p+1)}$ può essere ottenuta attraverso le rotazioni di Givens come nell'Algoritmo 5, ma questo metodo richiede il calcolo della matrice $\mathbf{Q}$. A tal fine è possibile calcolare $\boldsymbol{v}_1$ risolvendo il seguente sistema lineare $\mathbf{R}_1^\intercal \boldsymbol{v}_1 = \mathbf{X}^\intercal \boldsymbol{u}$, mentre $\mathbf{R}_1^+[p+1, p+1]$ che può essere calcolato come

$$\mathbf{R}_1^+(p+1, p+1) = \sqrt{\boldsymbol{u}^\intercal \boldsymbol{u} - \sum_{i=1}^{p} \boldsymbol{v}_1^2(i)},$$

sfruttando la relazione $\widetilde{\mathbf{R}}_1^+(, p+1)^\intercal \widetilde{\mathbf{R}}_1^+(, p+1) = \boldsymbol{u}^\intercal \boldsymbol{u}$.

Un procedimento analogo avviene nel caso in cui si vogliano aggiungere m colonne $\mathbf{U} \in \mathbb{R}^{n\times m}$ alla fine della matrice $\mathbf{X}$. Si avrà in questo caso $\widetilde{\mathbf{X}} = \begin{bmatrix}\mathbf{X} & \mathbf{U}\end{bmatrix}$, con $\widetilde{\mathbf{X}} \in \mathbb{R}^{n\times(p+m)}$ e coerentemente

$$\mathbf{Q}^\intercal \widetilde{\mathbf{X}} = \begin{bmatrix}\mathbf{R} & \mathbf{V}\end{bmatrix} = \begin{bmatrix}\mathbf{R}_1 & \mathbf{V}_1 \\ 0 & \mathbf{V}_2\end{bmatrix} = \widetilde{\mathbf{R}}_1^+ \tag{2.28}$$

con $\mathbf{V} = \mathbf{Q}^\intercal \mathbf{U}$, $\mathbf{V}_1 = \mathbf{Q}_1^\intercal \mathbf{U}$ e $\mathbf{V}_1 = \mathbf{Q}_2^\intercal \mathbf{U}$. Anche in questo caso $\mathbf{R}_1^+ \in \mathbb{R}^{(p+m)\times(p+m)}$ potrebbe essere ottenuta utilizzando le rotazioni di Givens per azzerare gli elementi necessari di $\widetilde{\mathbf{R}}_1^+$, tuttavia ciò richiederebbe il calcolo di $\mathbf{Q}$ e si utilizza quindi il seguente sistema lineare $\mathbf{R}_1^\intercal \mathbf{V}_1 = \mathbf{X}^\intercal \mathbf{U}$. $\mathbf{R}_1^+[p+i, p+j]$, con $i = 1, \ldots, m$ e $j \geq 1, \ldots, m$ può essere calcolato sfruttando la relazione $\mathbf{R}_1^{+\intercal} \mathbf{R}_1^+ = \widetilde{\mathbf{X}}^\intercal \widetilde{\mathbf{X}}$ (Bernardi et al., 2023).

2.5.2 Eliminazione di colonne

Nel caso in cui si voglia eliminare una colonna da $\mathbf{X}$ in posizione k, si otterrà la nuova matrice

$$\widetilde{\mathbf{X}} = \begin{bmatrix}\mathbf{X}(, 1:k-1) & \mathbf{X}(, k+1:p)\end{bmatrix} \in \mathbb{R}^{n\times(p-1)},$$

e, coerentemente,

$$\widetilde{\mathbf{R}}_1^- = \begin{bmatrix}\mathbf{R}_1(, 1:k-1) & \mathbf{R}_1(, k+1:p)\end{bmatrix}.$$

Si rende necessaria l'applicazione del medesimo procedimento utilizzato per l'aggiornamento della QR completa come nell'Algoritmo 7, consistente quindi nell'azzerare gli elementi sotto la diagonale principale delle ultime $p-k$ colonne di $\widetilde{\mathbf{R}}_1^-$ attraverso le rotazioni di Givens, al fine di ottenere $\mathbf{R}_1^- \in \mathbb{R}^{(p-1)\times(p-1)}$. Ciò viene fatto nel modo seguente

$$\begin{bmatrix} \mathbf{R}_1^- \\ 0 \end{bmatrix} = \mathbf{G}_p(p-1,p)^\intercal \times \cdots \times \mathbf{G}_p(k,k+1)^\intercal \widetilde{\mathbf{R}}_1^-. \tag{2.29}$$

Un procedimento analogo si ha nel caso in cui si vogliano eliminare m colonne da $\mathbf{X}$ a partire dalla posizione k. Si avrà in questo caso che la matrice $\mathbf{X}$ ridotta sarà della forma

$$\widetilde{\mathbf{X}} = \begin{bmatrix} \mathbf{X}(,1:k-1) & \mathbf{X}(,k+m:p) \end{bmatrix} \in \mathbb{R}^{n\times(p-1)},$$

e, coerentemente,

$$\widetilde{\mathbf{R}}_1^- = \begin{bmatrix} \mathbf{R}_1(,1:(k-1)) & \mathbf{R}_1(,k+m:p) \end{bmatrix}.$$

Anche in questo caso si rende necessario l'azzeramento degli elementi sotto la diagonale principale di $\widetilde{\mathbf{R}}_1^-$ e ciò viene raggiunto analogamente all'Algoritmo 8 grazie alle trasformazioni di Householder, nel seguente modo:

$$\begin{bmatrix} \mathbf{R}_1^- \\ 0 \end{bmatrix} = \mathbf{H}_{p-m}(p-m+1,p)^\intercal \times \cdots \times \mathbf{H}_k(k+1,k+m)^\intercal \widetilde{\mathbf{R}}_1^-. \tag{2.30}$$

dove $\mathbf{H}_j(l,n)$, con $j = k,\ldots,p-m$, è la matrice di Householder. La matrice $\mathbf{R}_1^- \in \mathbb{R}^{(p-m)\times(p-m)}$ è la matrice triangolare superiore ricercata (Bernardi et al., 2023).

2.6 Conclusioni

In questo Capitolo si sono approfonditi da un punto di vista prettamente algebrico i metodi di decomposizione QR e thin QR, nonché i metodi di aggiornamento delle stesse, in seguito all'aggiunta o all'eliminazione di una o più colonne dalla matrice originale. In particolare, sono state analizzate le rotazioni di Givens e le trasformazioni di Householder, metodi utili a calcolare la decomposizione QR in quanto in grado di azzerare gli elementi al di sotto della diagonale di una data matrice. Tali metodi algebrici sono poi stati analizzati nell'applicazione a detti metodi di aggiornamento. Tali metodi risulteranno particolarmente utili per fattorizzare il termine $\widetilde{\mathbf{X}}_\gamma \in \mathbb{R}^{(n+p_\gamma)\times p_\gamma}$, presente nella quantità $\mathbf{K}_\gamma = \mathbf{X}_\gamma^\intercal \mathbf{X}_\gamma + \mathbf{H}_\gamma^{-1} = \begin{pmatrix} \mathbf{X}_\gamma \\ \mathbf{H}_\gamma^{-1/2} \end{pmatrix}^\intercal \begin{pmatrix} \mathbf{X}_\gamma \\ \mathbf{H}_\gamma^{-1/2} \end{pmatrix} = \widetilde{\mathbf{X}}_\gamma^\intercal \widetilde{\mathbf{X}}_\gamma$, facente riferimento

alla distribuzione a posteriori marginalizzata in Equazione (A.13), così da rendere più veloce ed efficiente la sua inversione. Tali aspetti verranno poi applicati e trattati nel seguente Capitolo.

Capitolo 3

Simulazioni

3.1 Introduzione

Nel presente Capitolo verranno introdotti gli studi di simulazione effettuati per valutare l'efficacia dei metodi proposti. Si è infatti implementato il modello di regressione lineare multivariata presentato nel Capitolo 1, utilizzando il MCMC *reversible jump* per la simulazione dalla distribuzione a posteriori di $\boldsymbol{\gamma}$. Sono stati poi implementati ed aggiunti i metodi per l'aggiornamento della matrice $\mathbf{K}_\gamma$ sia mediante la decomposizione QR che thin QR, come spiegato nel Capitolo 2, per un totale quindi di tre metodi. Infine, per trattare alcuni particolari casi che verranno poi indicati di seguito, è stato implementato anche lo *stochastic search* per individuare il modello migliore da cui far partire l'algoritmo. Tutti i modelli, assieme ai metodi di aggiornamento QR e thin QR sono stati implementati su C++, in particolare utilizzando la libreria Armadillo, utilizzando allo stesso tempo le librerie Rcpp e Rcpp Armadillo per poter caricare le varie funzioni su R.

Si effettuerà poi un primo studio di simulazione per il caso $p < n$ ed un secondo studio, maggiormente di interesse, per $p > n$. Le combinazioni del numero di osservazioni n, il numero di variabili esplicative p e di coefficienti di regressione effettivamente diversi da zero p_0 sono visibili rispettivamente nelle Tabelle 3.1 e 3.2.

Si è utilizzata una variabile risposta trivariata, quindi con $q = 3$, ottenendo quindi $\mathbf{Y} \in \mathbb{R}^{n\times 3}$ e per ciascuna combinazione dei parametri sono state effettuate 10 replicazioni. Utilizzando dati simulati, per quanto riguarda la valutazione della bontà dei modelli, si sono confrontati ad ogni iterazione il vettore $\boldsymbol{\gamma}^*$ reale, che identifica quindi la reale struttura dei dati, e il vettore $\boldsymbol{\gamma}_i$, corrispondente all'insieme di coefficienti selezionato dal modello all'iterazione i-esima. Rispetto a questi vettori reali e previsti, si

n	p	p_0
1000	100	10
1000	500	10
1000	100	20
1000	500	20
2000	100	10
2000	500	10
2000	100	20
2000	500	20

Tabella 3.1: Struttura della Simulazione 1.

n	p	p_0
100	500	10
100	1000	10
100	500	20
100	1000	20
200	500	10
200	1000	10
200	500	20
200	1000	20

Tabella 3.2: Struttura della Simulazione 2.

sono calcolati l'*F1 Score*, il *False Discovery Rate* (FDR) e il *True Positive Rate* (TPR) per ogni iterazione, effettuando poi una media, che sarà il risultato visibile nelle tabelle successive. In secondo luogo, ovviamente, è stato confrontato anche il tempo computazionale valutato in secondi. Nel corso del Capitolo i modelli verrano indicati con le sigle *RJ naive* per il modello senza alcuna forma di aggiornamento, e con il nome delle rispettive decomposizioni per quanti riguarda gli altri due modelli. Prima di mostrare i risultati di tali simulazioni, si ritiene utile mostrare anche un confronto iniziale relativo al tempo computazionale necessario ad effettuare l'aggiornamento della decomposizione QR e thin QR, basato su delle matrici create casualmente, che verrà presentato nella seguente Sezione.

3.2 Confronto dell'aggiornamento QR e thin QR

Si propone inizialmente un piccolo confronto relativo al tempo computazione richiesto per aggiornare le matrici $\mathbf{Q}$ e $\mathbf{R}$ della decomposizione QR e thin QR, confrontato con il costo computazionale richiesto dall'effettuare una decomposizione QR della matrice $\mathbf{X}$ aggiornata, ossia $\widetilde{\mathbf{X}}$, con il numero di colonne modificate. Si propongono in questo caso delle matrici generiche di partenza, generate casualmente, di dimensione 2000×500, sulle quali viene inizialmente effettuata la decomposizione QR, e poi vengono aggiunte e tolte una e più colonne (m) mediante i metodi di aggiornamento QR e thin QR, sulla base degli algoritmi descritti nel Capitolo 2. Si utilizza un valore di m pari a 3, scegliendo casualmente ad ogni replicazione il punto da cui eliminarle. L'aggiunta invece avviene sempre alla fine, essendo i metodi thin QR implementabili solo in questo modo. Per far notare il vantaggio computazionale, assieme ai metodi di aggiornamento si mostra anche il tempo necessario ad effettuare una decomposizione QR della matrice $\mathbf{X}$ a cui sono aggiunte o sottratte le rispettive colonne. Ciascuna operazione viene reiterata $r = 50$ volte. I risultati di tali simulazioni sono visibili nelle Tabelle 3.3 e 3.4.

Aggiunta di colonne

	Una colonna		M colonne	
	Media	**Relativo**	**Media**	**Relativo**
Aggiornamento thin QR	0.0058	1.0000	0.0066	1.0000
Aggiornamento QR	0.0482	8.3577	0.0653	9.8494
QR	6.7873	1177.0804	6.8350	1030.6711

TABELLA 3.3: Tempo computazionale medio e relativo (rispetto al metodo più veloce) in secondi necessario all'aggiunta di una e più (m=3) colonne alla matrice $\mathbf{R}$, confrontato con il tempo necessario a ricolare $\mathbf{R}$ dalla nuova matrice $\widetilde{\mathbf{X}}$.

Eliminazione di colonne

	Una colonna		M colonne	
	Media	**Relativo**	**Media**	**Relativo**
Aggiornamento thin QR	0.0046	1.0000	0.0269	1.0000
Aggiornamento QR	0.0265	5.7637	0.0593	2.2083
QR	7.6778	1667.5971	6.8173	253.8134

TABELLA 3.4: Tempo computazionale medio e relativo (rispetto al metodo più veloce) in secondi necessario all'eliminazione di una e più (m=3) colonne alla matrice $\mathbf{R}$, confrontato con il tempo necessario a ricolare $\mathbf{R}$ dalla nuova matrice $\widetilde{\mathbf{X}}$.

I risultati sono riportati in secondi ed oltre al tempo medio impiegato si riporta anche il tempo medio relativo, ottenuto dividendo ciascun tempo medio per il tempo minore registrato, che risulta essere in tutti i casi l'aggiornamento mediante decomposizione thin QR. Infatti, in questo caso, non viene aggiornata o calcolata la matrice $\mathbf{Q}$, al contrario degli altri due casi, cosa che come visibile risulta essere piuttosto onerosa, rendendo ciascuna operazione dalle 2 alle 11 volte più lenta, nel caso dell'aggiornamento mediante decomposizione QR. Entrambi i metodi risultano essere in ogni caso molto più efficienti rispetto al computare le matrici $\mathbf{Q}$ e $\mathbf{R}$ da capo, operazioni che risultano essere addirittura tra le 1030 e le 3000 volte più lente rispetto al metodo thin QR in tre casi. Questo divario diminuisce infatti nel solo caso di eliminazione di più colonne, in cui l'aggiornamento thin QR risulta essere soltanto il doppio più veloce rispetto all'aggiornamento QR.

3.3 Studi di simulazione

3.3.1 Caso $p < n$

Questa sezione andrà a trattare lo studio di simulazione relativo al caso $p < n$. Come già indicato sono state utilizzate le combinazioni di p, n, e p_0 visibili in Tabella 3.1 e per ciascuna combinazione sono state effettuate 10 replicazioni, per un totale quindi di

80 simulazioni. I risultati relativi alle statistiche calcolate sono visibili in Tabella 3.5 e anche graficamente in Figura 3.1. Si può notare quindi come l'*F1 Score* sia sempre piuttosto alto e si aggiri quindi attorno a livelli soddisfacenti, sempre superiore al 90% nei casi con $p = 100$.

n	p	p_0	**F1 Score medio**	*s.d.*	**FDR medio**	*s.d.*	**TPR medio**	*s.d.*
1000	100	10	0.949	0.015	0.087	0.026	0.990	0.002
1000	500	10	0.897	0.019	0.144	0.028	0.947	0.019
1000	100	20	0.975	0.012	0.037	0.022	0.990	0.004
1000	500	20	0.932	0.010	0.082	0.017	0.951	0.012
2000	100	10	0.949	0.016	0.087	0.028	0.990	0.003
2000	500	10	0.897	0.027	0.147	0.037	0.954	0.018
2000	100	20	0.974	0.012	0.037	0.024	0.989	0.004
2000	500	20	0.922	0.018	0.094	0.035	0.947	0.011

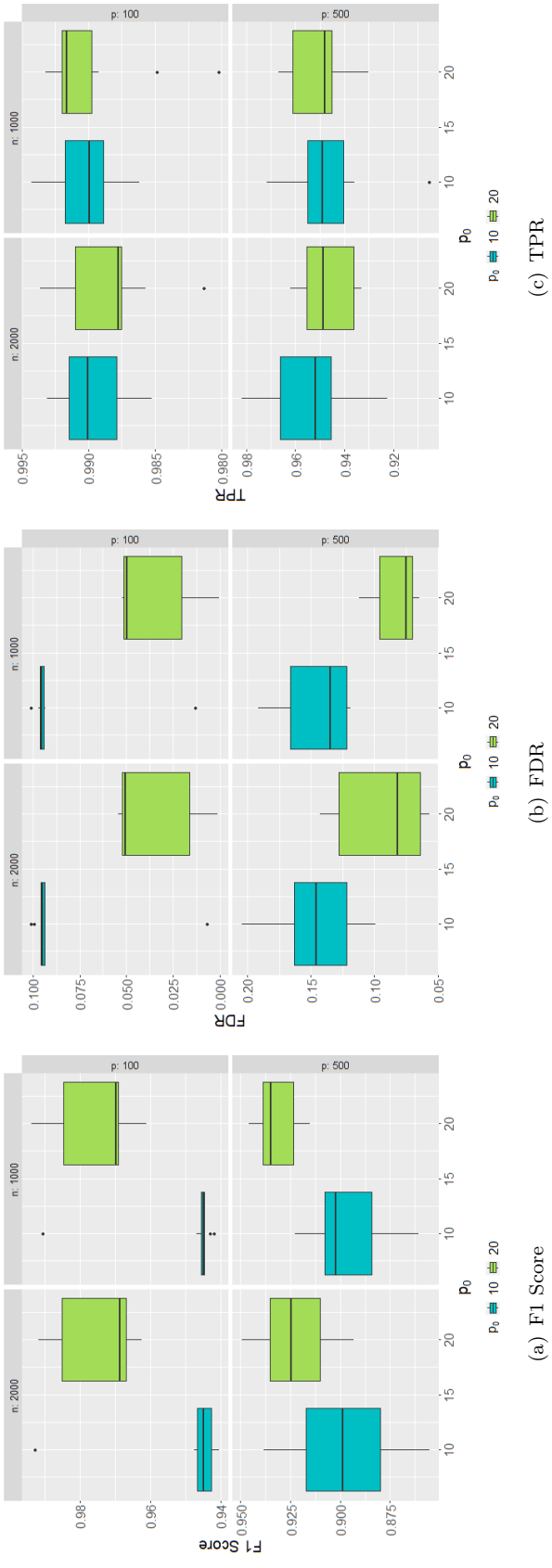
TABELLA 3.5: *F1 Score*, *False Discovery Rate* (FDR) e *True Positive Rate* (TPR) e relativa deviazione standard (s.d.) relativi alla Simulazione 1. Si riporta un solo risultato essendo il medesimo per tutti i metodi.

			RJ thin QR		**RJ naive**		**RJ QR**	
n	p	p_0	**Media**	*s.d.*	**Media**	*s.d.*	**Media**	*s.d.*
1000	100	10	1.241	0.050	1.119	0.060	401.728	0.909
1000	500	10	56.361	0.438	72.368	0.380	821.653	2.622
1000	100	20	2.501	0.118	2.111	0.072	392.183	0.456
1000	500	20	59.465	0.256	71.987	0.330	818.735	1.502
2000	100	10	2.338	0.115	1.817	0.054	1475.116	3.696
2000	500	10	61.927	0.500	74.843	0.483	2195.593	3.790
2000	100	20	3.746	0.125	3.089	0.154	1440.667	6.280
2000	500	20	62.567	0.752	75.760	0.328	2185.696	3.972

TABELLA 3.6: Tempo computazionale medio (in secondi) e relativa deviazione standard della Simulazione 1 per i tre metodi proposti.

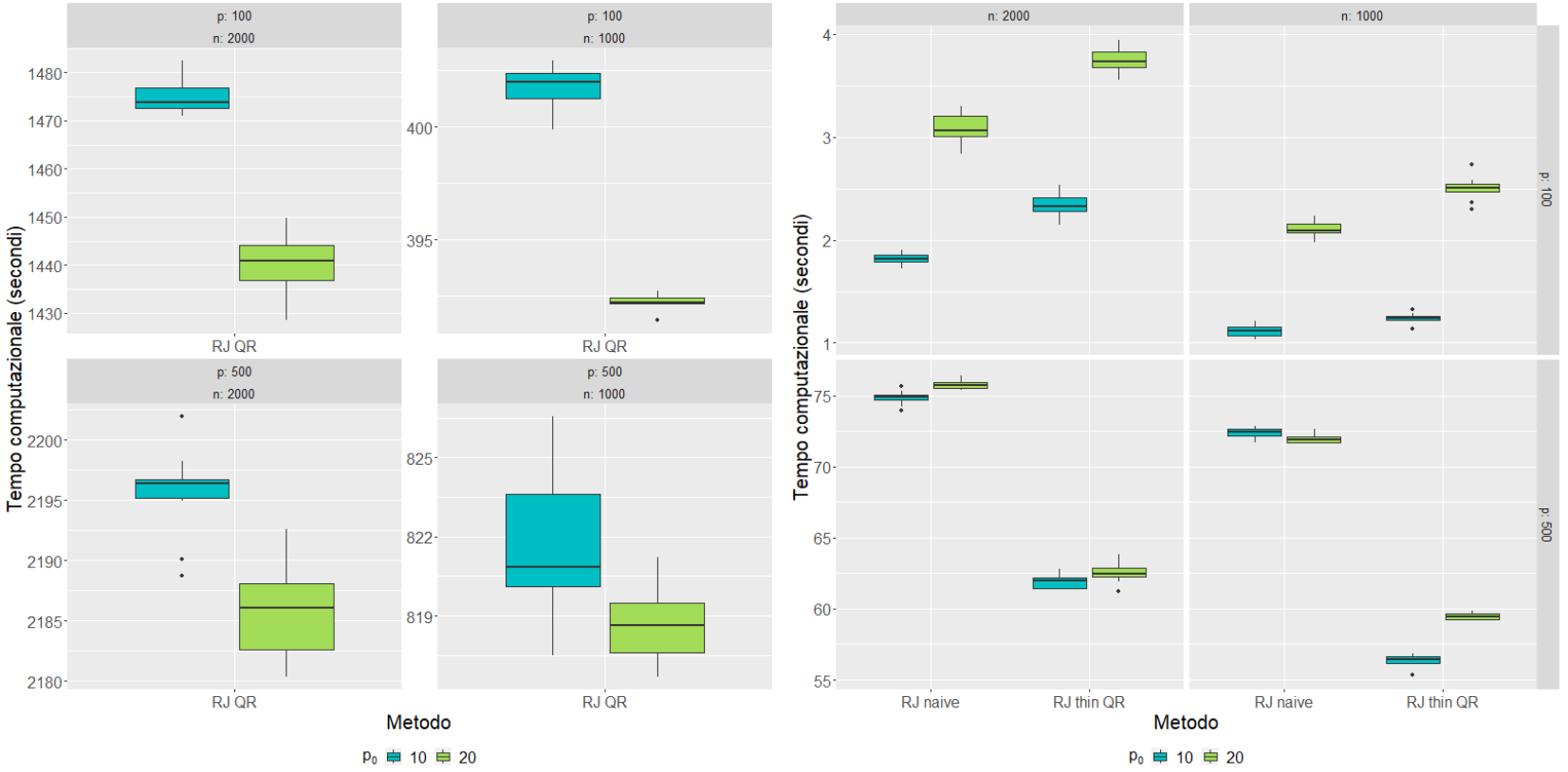
Nei casi in cui invece p è più grande, ossia $p = 500$ in questo studio di simulazione, si nota un leggera diminuzione dell'*F1 Score*. In generale poi, a prescindere dal numero di variabili esplicative presenti, si nota un *F1 Score* più alto quando p_0 è pari a 20. Il FDR segue un analogo andamento, dal momento che invece per questo indice sono auspicabili dei valori più bassi. Per quanto riguarda invece il TPR, anch'esso risulta essere più alto con un numero di variabili esplicative pari a 100.

Per quanto riguarda il tempo computazionale, i risultati sono visibili in Tabella 3.6 e graficamente in Figura 3.2(a) e 3.2(b). Per quanto riguarda il modello con decomposizione QR, i risultati sono riportati separatamente a causa della scala molto diversa dei risultati, che non consentiva una chiara rappresentazione degli stessi. Infatti, si può



(a) F1 Score (b) FDR (c) TPR

FIGURA 3.1: Boxplot delle tre metriche di valutazione adottate per la Simulazione 1. I risultati sono rappresentati in funzione dei valori di p, p_0 ed n per ciascuna delle 10 replicazioni dello studio di simulazione riportato in Tabella 3.1.



(a) RJ QR

(b) RJ naive e thin QR

FIGURA 3.2: Tempo computazionale medio in secondi relativo alle simulazioni del primo studio di simulazione, effettuate con decomposizione QR 3.2(a) e con modello RJ naive e RJ thin QR 3.2(b).

notare chiaramente come il tempo computazionale necessario per stimare il modello con la decomposizione QR sia sempre nettamente più alto rispetto agli altri due metodi. Il motivo è da ricercarsi nel fatto che l'aggiornamento tramite la decomposizione QR richiede ad ogni iterazione anche il calcolo e l'aggiornamento della matrice $\mathbf{Q}$, di dimensione $(n + p_\gamma) \times (n + p_\gamma)$, al contrario dell'aggiornamento thin QR che invece richiede solo l'aggiornamento di $\mathbf{R}$. Fissati p e p_0, infatti, si nota come il tempo computazionale aumenti notevolmente all'aumentare di n. Quanto appena esplicitato è chiaramente visibile in Figura 3.2(a), prestando attenzione al fatto che le scale in cui i grafici sono riportati sono diverse per ciascuna combinazione di p ed n. A parità di p e p_0, infatti, il tempo computazionale aumenta di circa 3 volte quando n raddoppia. Appurato quindi il fatto che tale metodo non risulta efficiente in questi casi di simulazione, si propone il confronto dei soli metodi naive e thin QR, i cui risultati sono visibili nei grafici in Figura 3.2(b). Si può notare a questo punto un risultato interessante, ossia che nelle simulazioni in cui il numero di variabili esplicative è pari a 100, il modello con decomposizione thin QR non risulta essere più veloce rispetto all'analogo modello senza alcuna decomposizione. Il modello con thin QR consente di ottenere un vantaggio computazionale invece nel caso in cui il numero di variabili esplicative sia maggiore, come visibile in Figura 3.2(b) con $p = 500$. La presenza di un numero maggiore di coefficienti non nulli determina invece un tempo computazionale più elevato.

3.3.2 Caso $p > n$

Il secondo studio di simulazione effettuato riguarda invece il caso in cui il numero di variabili esplicative risulti essere maggiore rispetto al numero di osservazioni presenti. Lo schema di simulazione è visibile in Tabella 3.2 e, come nel caso precedente, sono state effettuate 10 replicazioni, anche in questo caso quindi per un totale di 80 simulazioni.

n	p	p_0	**F1 Score medio**	*s.d.*	**FDR medio**	*s.d.*	**TPR medio**	*s.d.*
100	500	10	0.883	0.029	0.163	0.038	0.945	0.028
100	1000	10	0.921	0.021	0.141	0.032	0.996	0.006
100	500	20	0.879	0.052	0.142	0.068	0.917	0.025
100	1000	20	0.923	0.020	0.111	0.031	0.966	0.018
200	500	10	0.895	0.015	0.149	0.027	0.954	0.015
200	1000	10	0.848	0.039	0.194	0.051	0.907	0.023
200	500	20	0.921	0.018	0.091	0.030	0.942	0.015
200	1000	20	0.872	0.022	0.121	0.034	0.881	0.024

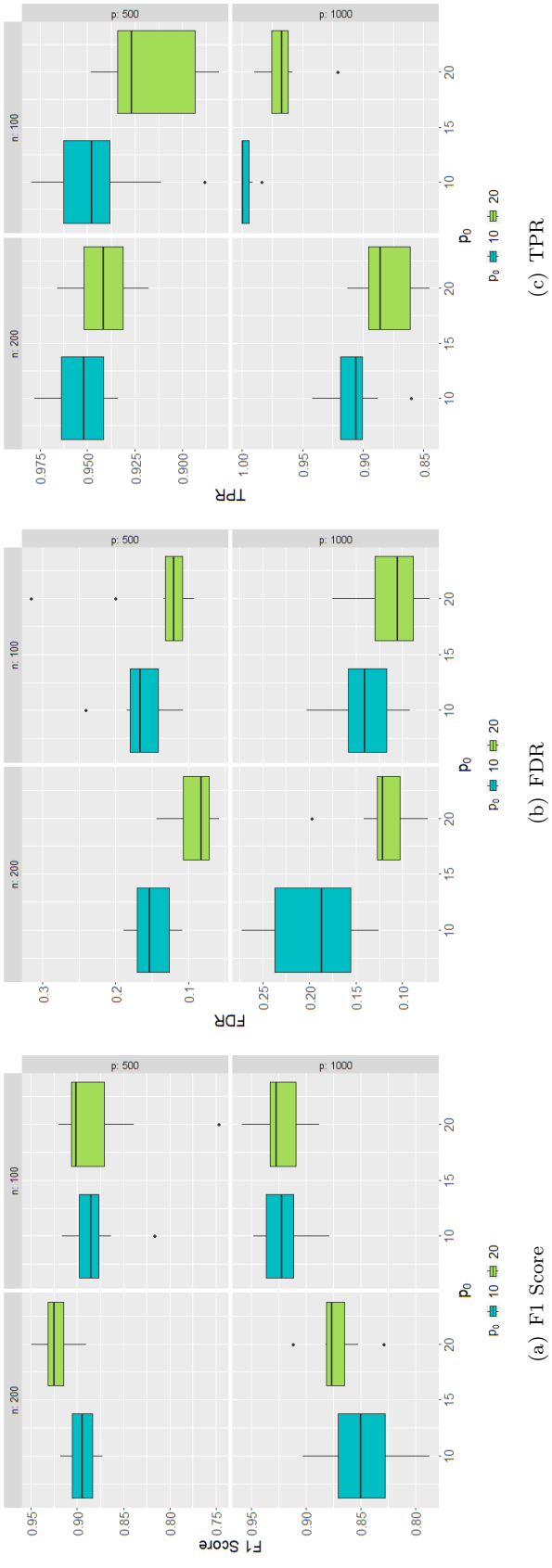
TABELLA 3.7: *F1 Score*, *False Discovery Rate* (FDR) e *True Positive Rate* (TPR) e relativa deviazione standard (s.d.) relativi alla Simulazione 2. Si riporta un solo risultato essendo il medesimo per tutti i metodi.

Le metriche di valutazione sono visibili in Tabella 3.7 e graficamente in Figura 3.3. Si notano anche in questo caso degli *F1 Score* sufficientemente alti che variano tra l'85% e il 92%, così come il TPR che si aggira in tutti i casi attorno al 90%. Il FDR si mantiene invece tra il 10% e il 15%. In generale, come visibile anche dai boxplot in Figura 3.3, si ottengono risultati migliori quando il numero di coefficienti non nulli è più alto.

			RJ thin QR		**RJ naive**		**RJ QR**		
n	p	p_0	**Media**	*s.d.*	**Media**	*s.d.*	**Media**	*s.d.*	
100	500	10	53.416	0.241	69.412	0.359	168.111	0.600	
100	1000	10	251.927	1.069	312.256	1.637	695.912	0.964	*
100	500	20	59.243	0.912	69.466	0.689	170.722	0.660	
100	1000	20	245.581	4.978	301.552	7.902	651.506	20.860	*
200	500	10	55.114	0.314	68.433	0.367	212.552	0.338	
200	1000	10	226.549	0.683	286.801	0.712	713.400	1.772	
200	500	20	58.497	0.350	69.860	0.313	214.253	0.791	
200	1000	20	229.842	0.865	293.086	0.700	714.087	0.968	

TABELLA 3.8: Tempo computazionale medio (in secondi) e relativa deviazione standard della Simulazione 2 per i tre metodi proposti.

Passando ora al tempo computazionale, si riportano anche in questo caso dei grafici separati per il modello con aggiornamento QR, in Figura 3.4(a), e per i modelli naive e con aggiornamento thin QR, in Figura 3.4(b), sempre a causa delle diverse scale. Anche in questo caso, come visibile dalla Tabella 3.8, l'aggiornamento QR non risulta essere efficiente, conseguendo dei tempi medi di simulazione circa del doppio più elevati



(a) F1 Score (b) FDR (c) TPR

FIGURA 3.3: Boxplot delle tre metriche di valutazione adottate per la Simulazione 1. I risultati sono rappresentati in funzione dei valori di p, p_0 ed n per ciascuna delle 10 replicazioni dello studio di simulazione riportato in Tabella 3.2.

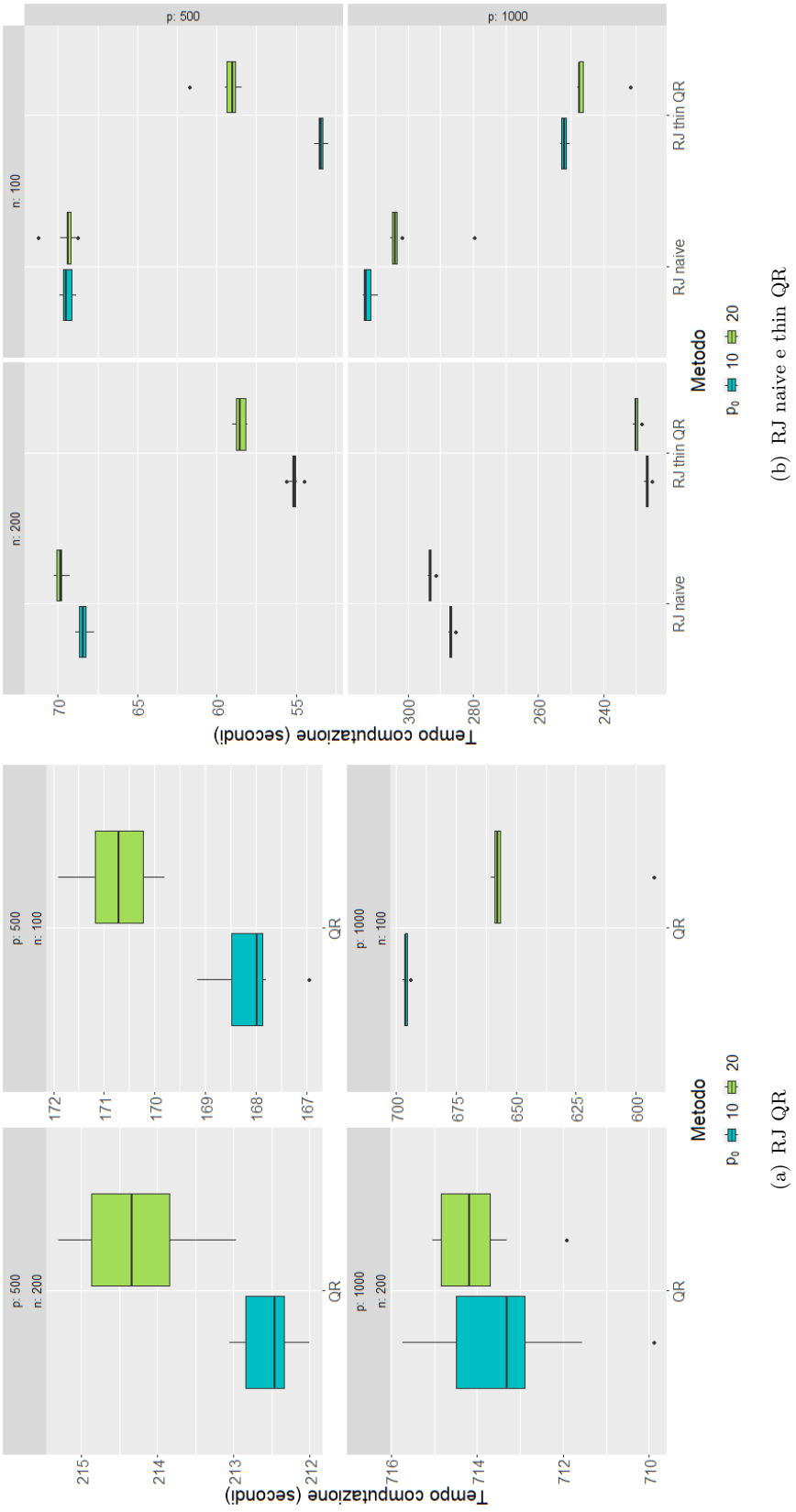
rispetto agli altri due metodi. Passando invece al confronto degli altri due metodi, come visibile in Figura 3.4(b), si può notare come si ottengano dei vantaggi abbastanza rilevanti dall'utilizzo dell'aggiornamento thin QR, ottenendo una riduzione del tempo medio computazionale di circa il 20% in sostanzialmente tutti i casi. In tutti i casi, poi, a parità di p e n, il tempo computazionale risulta essere più elevato con p_0 pari a 20 rispetto che con p_0 pari a 10. Da notarsi tuttavia che le simulazioni relative ai casi con $n = 100$ e $p = 1000$, contrassegnati con un asterisco, sono state implementate con l'ausilio dello *stochastic search*, utilizzato per individuare il modello migliore da cui far partire l'algoritmo. Infatti, in quei due casi, la partenza da un modello con molte covariate determinava un aumento dei tempi computazionali molto elevato e coerentemente anche dei risultati in termini di F1 Score molto più scarsi. L'utilizzo dello *stochastic search* ha invece consentito al modello di partire da un numero di variabili esplicative minore e più coerente con quelle realmente non nulle.

3.4 Elicitazione della distribuzione a priori

Facendo riferimento al modello descritto nella Sezione 1.5, in entrambe le simulazioni sono stati usati i seguenti valori degli iperparametri: $\mathbf{B}_{\gamma,0} = \underline{0}_{p_\gamma \times q}$, $\mathbf{H}_\gamma = v\mathbf{I}_{p_\gamma}$, con $v = 100$, $c0 = 5$, in quanto deve essere maggiore o uguale di $q + 2$ affinché la media della distribuzione Inverse-Wishart sia definita, e $\mathbf{C}_0 = k\mathbf{I}_q$, con $k = 0.2$. Il parametro θ relativo della distribuzione Bernoulli in Equazione (1.18) viene generato casualmente da una distribuzione $Beta(a, b)$ di parametri $a = 0.01$ e $b = 4.01$. Si tratta quindi di una distribuzione con molta massa di probabilità su valori prossimi allo zero, in modo da far partire l'algoritmo da un vettore $\boldsymbol{\gamma}$ sufficientemente sparso.

3.5 Conclusioni

In questo Capitolo si sono voluti analizzare sia i limiti che i punti di forza dei tre modelli implementati utilizzando due studi di simulazione. Innanzitutto, anche soltanto grazie al confronto effettuato nella Sezione 3.2, esterno ai modelli, si nota come il metodo di aggiornamento basato sulla decomposizione QR risulti sempre molto meno efficiente dell'aggiornamento thin QR, aspetto che si è palesato molto di più successivamente in entrambi gli studi di simulazione. Ciò è dovuto al fatto che ad ogni iterazione risulta necessario anche il calcolo e l'aggiornamento della matrice $\mathbf{Q}$, di dimensione $(n + p_\gamma) \times (n+p_\gamma)$, al contrario dell'aggiornamento thin QR che invece richiede solo l'aggiornamento di $\mathbf{R}$. Infatti, nella Simulazione 1, al raddoppiare di n il tempo computazionale richiesto



(b) RJ naive e thin QR

(a) RJ QR

FIGURA 3.4: Tempo computazionale medio in secondi relativo alle simulazioni del primo studio di simulazione, effettuate con decomposizione QR 3.4(a) e con modello RJ naive e RJ thin QR 3.4(b).

è triplicato. I risultati per il modello thin QR sono stati infatti notevolmente diversi, dimostrandosi più efficiente nella maggior parte dei casi considerati, ottenendo una riduzione del tempo medio computazionale di circa il 20% rispetto al modello naive. Si è registrata un'unica eccezione nella Simulazione 1 nei casi in cui il numero di variabili esplicative risultasse troppo basso, in particolare quando è stato fissato pari a 100, casi in cui il modello con decomposizione thin QR non è risultato essere più veloce rispetto al modello naive. Rispetto alla decomposizione QR invece, il tempo computazionale nella Simulazione 2 è stato circa 3 volte più basso per il modello thin QR, mentre per la Simulazione 1 i risultati sono più variabili, dal momento che i valori fissati di n erano più elevati. In questo caso il modello con decomposizione QR si è dimostrato essere fino a 600 volte più lento del modello thin QR e naive. Un aspetto problematico si è però presentato nella Simulazione 2 nei casi in cui il rapporto tra numero di variabili esplicative e il numero di osservazioni risultasse troppo elevato. In quei casi infatti, è risultato frequente che il modello di partenza fosse troppo ampio, determinando un aumento in tutti i casi del tempo computazionale e un peggioramento delle metriche di valutazione. A tal fine si è sviluppato un metodo *stochastic search* che ha consentito di partire da modello con alta probabilità a posteriori, esplorando quindi poi modelli più sensati e coerenti. In conclusione, si è quindi dimostrata l'efficacia dell'aggiornamento thin QR in tutti i casi in cui il numero di variabili esplicative fosse maggiore del numero di osservazioni, metodo che invece consente di ottenere un vantaggio minore nei casi più semplici con $p < n$, in cui il costo di effettuare l'aggiornamento risulta essere talvolta maggiore rispetto all'inversione della matrice $\mathbf{K}_\gamma$ aggiornata come avviene nel modello naive. Il metodo di aggiornamento mediante decomposizione QR non risulta mai essere idoneo ai casi in oggetto.

Capitolo 4

Applicazione

4.1 Introduzione

Dato il modello delineato nel Capitolo 1, con i successivi miglioramenti esplicati nel Capitolo 2, si è deciso di effettuare un'applicazione anche ad un dataset reale. A tal fine si è utilizzato un dataset relativo a sei turbine eoliche situate nel parco eolico di Kelmarsh, situato nel Regno Unito, nella regione del Northamptonshire. Tale parco eolico è composto da sei turbine di tipo *Senvion MM92*, la cui posizione e orientamento è visibile in Figura 4.1, ciascuna in grado di produrre 2.050 kW, per un totale quindi di 12.300 kW di energia totale producibile. Tale applicazione è stata scelta vista la crescente importanza delle forme di energia rinnovabile a livello europeo e mondiale, nonché data l'importanza che ricopre la previsione della produzione di energia eolica, a causa dell'incertezza e della stocasticità legata agli eventi atmosferici. Infatti, la generazione di energia eolica è possibile solo quando ci sono condizioni metereologiche adeguate e risulta pertanto cruciale prevedere i periodi in cui sarà necessario ricorrere ad altre fonti energetiche. Vi sono quindi diversi aspetti particolarmente interessanti: risulta infatti importante sia prevedere la velocità del vento che l'energia prodotta, essendo in realtà due quantità strettamente collegate. Lo scopo risulta quindi essere sostanzialmente triplice e riguarda:

- la selezione del luogo in cui posizionare i parchi eolici;
- sfruttare efficientemente il vento, principalmente controllando le turbine;
- integrare efficientemente la potenza generata nella rete elettrica.

In aggiunta, le previsioni sulla potenza generata possono essere effettuate anche giorni o ore in anticipo per gestire in modo più adeguato il sistema di alimentazione e il

FIGURA 4.1: Posizione delle turbine del parco eolico di Kelmarsh.

commercio di energia, consentendo di conseguenza anche una migliore gestione della manutenzione. Lo scopo principale della previsione nella produzione di energia eolica è solitamente il terzo, quindi l'integrazione efficiente dell'energia eolica prodotta nella rete elettrica (Effenberger & Ludwig, 2022).

4.2 Descrizione del dataset

Il dataset originale si compone di 1.018.494 osservazioni e 110 variabili. Alcune righe e colonne facenti riferimento alla turbina 3 sono visibili in Tabella 4.2 a scopo esemplificativo. Come già accennato, riguarda delle serie storiche facenti riferimento a periodi che vanno dal 25 settembre 2017 al 30 giugno 2021, con periodi mancanti nel mezzo. Le variabili osservate sono delle misure medie, ciascuna facente riferimento ad intervalli di 10 minuti. Per ogni variabile vengono riportate anche le relative deviazioni standard e valori massimi e minimi in quel frangente di tempo. Le variabili rilevate riguardano, oltre all'energia media prodotta, anche la velocità del vento media e i dati cosidetti *Supervisory Control And Data Acquisition* (SCADA), cioè controllo di supervisione e acquisizione dati. I dati SCADA comprendono un insieme di misurazioni ambientali, operative, termiche ed elettriche, raccolte di solito per fini manutentivi; ricadono in questa categoria infatti le rilevazioni delle temperature di varie componenti della turbina. Le variabili relative alle temperature e al tempo atmosferico in generale sono molto utili e possono aumentare l'accuratezza dei modelli: è stato notato infatti che il congelamento delle turbine induce perdite di energia fino all'80%, con il 94% delle turbine in Europa che sono afflite da questo fenomeno.

Date.time	Power.me	Power.sd	Power.min	Power.max	Pot.Power.me	Wind.speed.me
2019-01-01 00:00:00	211.6351	42.51391	117.54772	292.4185	194.24061	4.903068
2019-01-01 00:10:00	242.3192	65.76192	136.34926	359.0036	182.81341	4.800121
2019-01-01 00:20:00	402.8349	66.40950	238.96910	530.0757	367.08730	5.871437
2019-01-01 00:30:00	358.5788	46.41203	259.73447	445.1906	375.73953	5.917954
2019-01-01 00:40:00	202.4557	62.60701	92.38598	303.4200	155.84433	4.557156
2019-01-01 00:50:00	125.6818	62.26595	30.73091	236.5148	78.18429	3.786274

TABELLA 4.1: Prime righe del dataset relativo alla turbina 3.

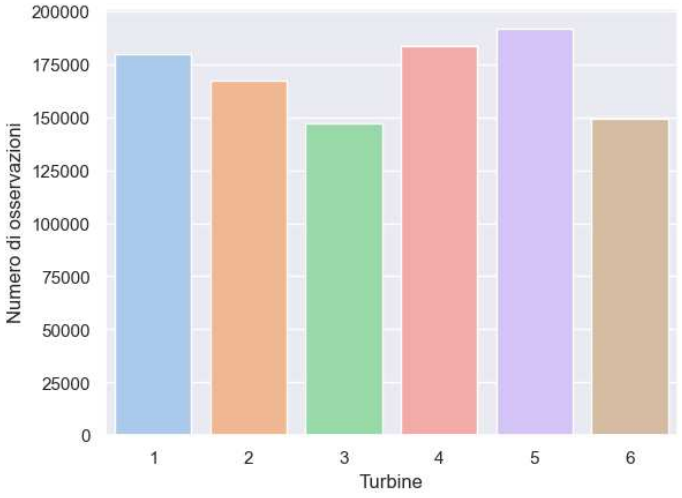


FIGURA 4.2: Numero di osservazioni presenti per ogni turbina per l'intero periodo di rilevazione.

Anche le precipitazioni e l'umidità potrebbero essere delle variabili importanti; le prime infatti possono portare all'erosione delle pale delle turbine portando ad un decremento della potenza generata (Effenberger & Ludwig, 2022), tuttavia queste informazioni non sono presenti nel dataset in oggetto.

4.3 Preprocessing

Per quanto riguarda il *preprocessing*, i dati erano già stati sottoposti ad operazioni reliminari di pulizia, atte ad identificare *outliers*, quindi osservazioni anormali o di scarsa qualità, che naturalmente possono presentarsi in dati derivanti da sensori, a causa di errori negli stessi o di condizioni metereologiche estreme. Ad una prima analisi del dataset, in ogni caso, è stato possibile notare immediatamente come non tutte le turbine avessero lo stesso numero di rilevazioni relativamente alla variabile risposta *Power.me*, con alcune turbine che presentano osservazioni mancanti per periodi anche di mesi. Ciò è visibile anche in Figura 4.2, notando come su tutto il periodo di rilevazione le turbine 3 e 6 abbiano il minor numero di rilvazioni, probabilmente a causa di alcuni periodi di inattività delle turbine.

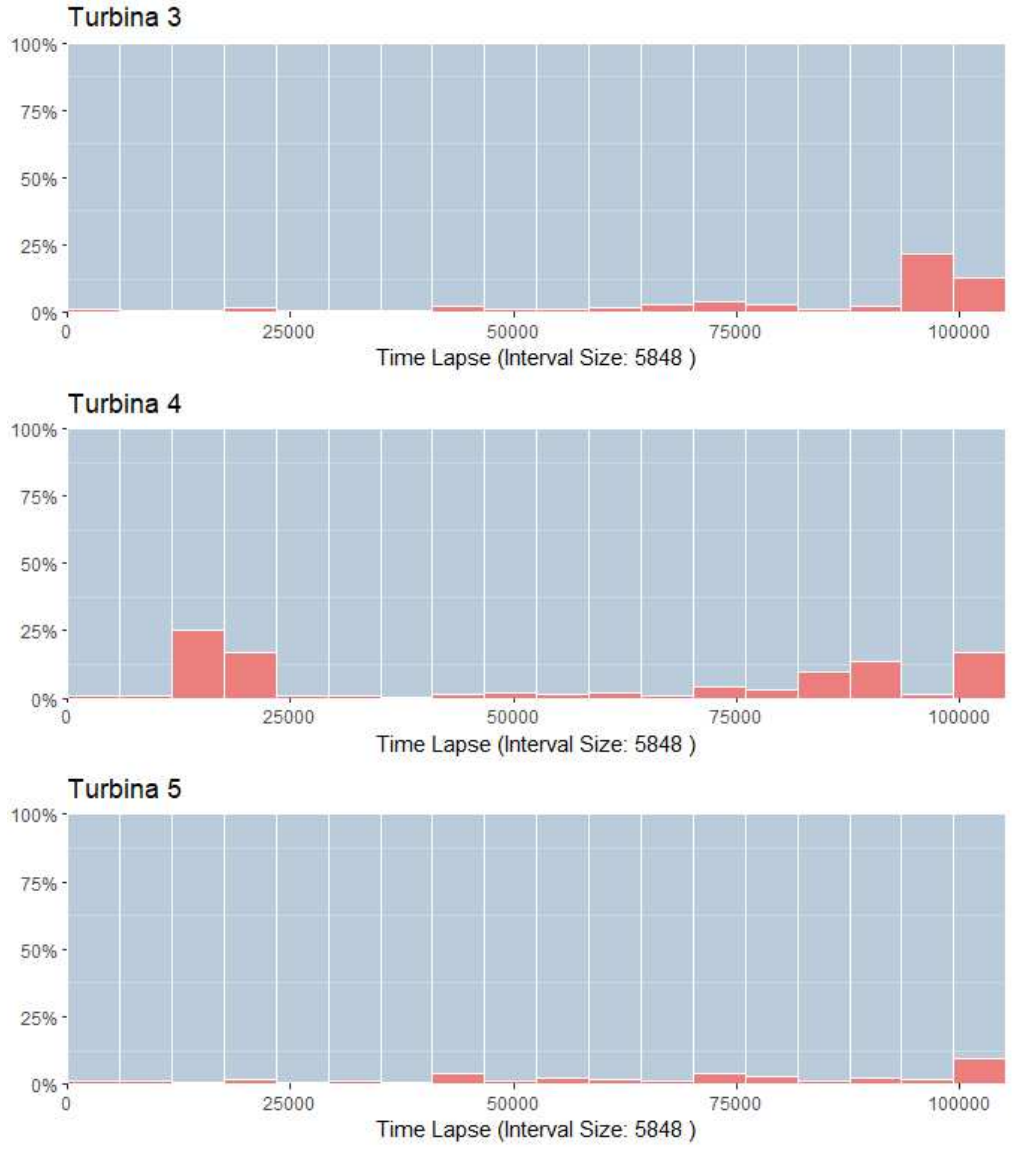


FIGURA 4.3: Percentuale degli NA (in rosso) in ciascun intervallo di tempo per ciascuna turbina.

Ai fini dell'analisi, quindi, sono state individuate tre turbine che per un periodo di due anni consecutivi avessero un numero sufficiente di osservazioni. La scelta è andata a ricadere sulle turbine 3, 4 e 5 per gli anni 2019 e 2020. A questo punto i relativi dataset sono composti rispettivamente da 101.723, 99.199 e 102.991 osservazioni. Il primo anno poi fungerà da insieme di *train* mentre il secondo anno da insieme di *test*.

Si sono dovuti quindi individuare i tempi mancanti nella variabile risposta *Power.me* in tali turbine, in quanto non erano esplicitamente indicati come NA, bensì semplicemente non erano riportati. Si ottengono quindi 105.264 osservazioni per ciascuna turbina. Si è poi andata ad analizzare la presenza di valori mancanti anche tra le covariate, la cui distribuzione in termini di frequenza relativa è visibile in Figura A.1. Si decide quindi di

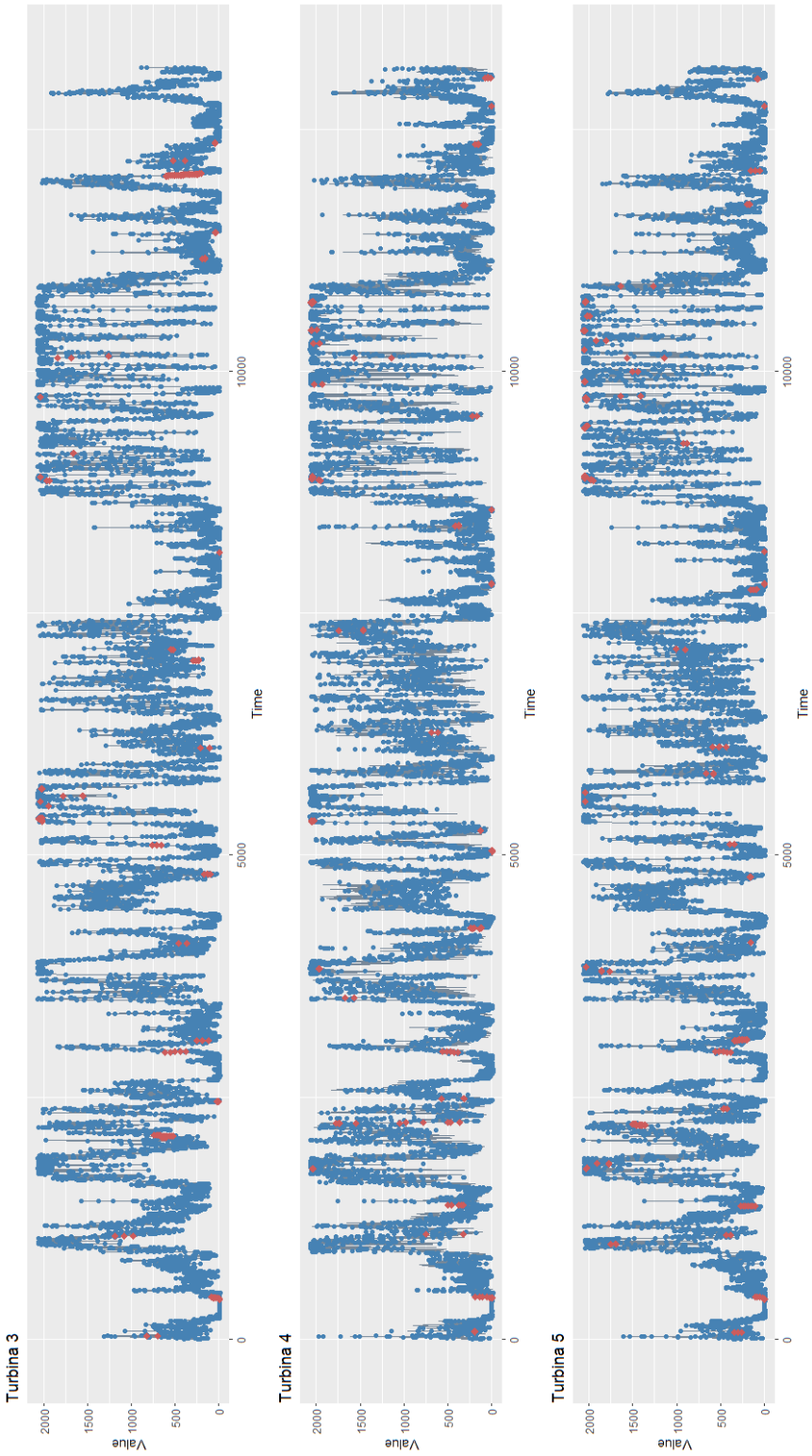


FIGURA 4.4: Imputazione degli NA per i mesi di gennaio, febbraio e marzo 2019.

eliminare tutte le variabili che presentano una frequenza di NA maggiore del 10%, eliminando quindi le medesime 54 variabili per tutte le turbine, ottenendo quindi dei dataset finali con 105.264 osservazioni e 56 variabili, comprese la variabile indicatrice relativa alla turbina e l'anno di rilevazione, che vengono quindi eliminate anch'esse. Per la successiva modellazione quindi risultano disponibili un totale di 50 variabili, di cui una la risposta *Power.me*, in quanto si eliminano le variabili relative alla deviazione standard, al massimo e al minimo della stessa. Le variabili esplicative sono quindi 49 per ciascuna turbina. A questo punto si è analizzato in quali periodi sono presenti effettivamente i valori mancanti nella variabile risposta *Power.me*. La distribuzione degli stessi è visibile in Figura 4.3. Si nota quindi per la turbina 4 dei vuoti maggiori nel periodo di aprile e maggio del 2019, mentre per le turbine 3 e 5 i valori mancanti si concentrano verso la fine del 2020. Si è reso poi necessario imputare i valori mancanti. A tal fine si è scelto di farlo utilizzando per ciascuna variabile il filtro di Kalman con la funzione *na_kalman* inclusa nel pacchetto *imputeTS*. Si rimanda per dettagli alla documentazione di tale pacchetto. A tal fine è possibile vedere in Figura 4.4 come sono stati imputati i valori mancanti per i primi tre mesi dell'insieme di stima.

4.4 Analisi esplorativa

In questa sezione saranno effettuate alcune analisi esplorative relative al dataset. Innanzitutto in Figura 4.5 si riportano le serie storiche dell'energia media prodotta (*Power.me*), la variabile risposta, e della velocità media del vento (*Wind.speed.me*) per i primi dieci giorni dell'insieme di stima, ossia tra il 1° gennaio 2019 e il 10 gennaio 2019. Si nota quindi una forte dipendenza tra le due variabili e come l'energia prodotta sia compresa in un range che va da 0 a circa 2.000 kW, data la capacità massima già statuita precedentemente di 2.050 kW. Si può osservare che vi sono periodi in cui, nonostante la velocità del vento sia positiva, la produzione di energia risulta essere nulla. Inoltre, l'energia prodotta può talvolta essere negativa, un fenomeno causato dai momenti in cui la turbina è inattiva ma continua a consumare energia.

Dall'andamento del grafico, si può notare come la generazione di energia non sia sempre completamente sfruttabile, poiché non è possibile convertire un tipo di energia in un altro senza subire delle perdite. Nel caso specifico dell'energia eolica, è possibile trasformare circa il 59% dell'energia generata dal vento; tale limite è noto come limite di Betz. Questo limite è definito da una relazione deterministica basata su leggi fisiche, che stabilisce il legame tra l'energia prodotta e la velocità del vento. Infatti l'energia

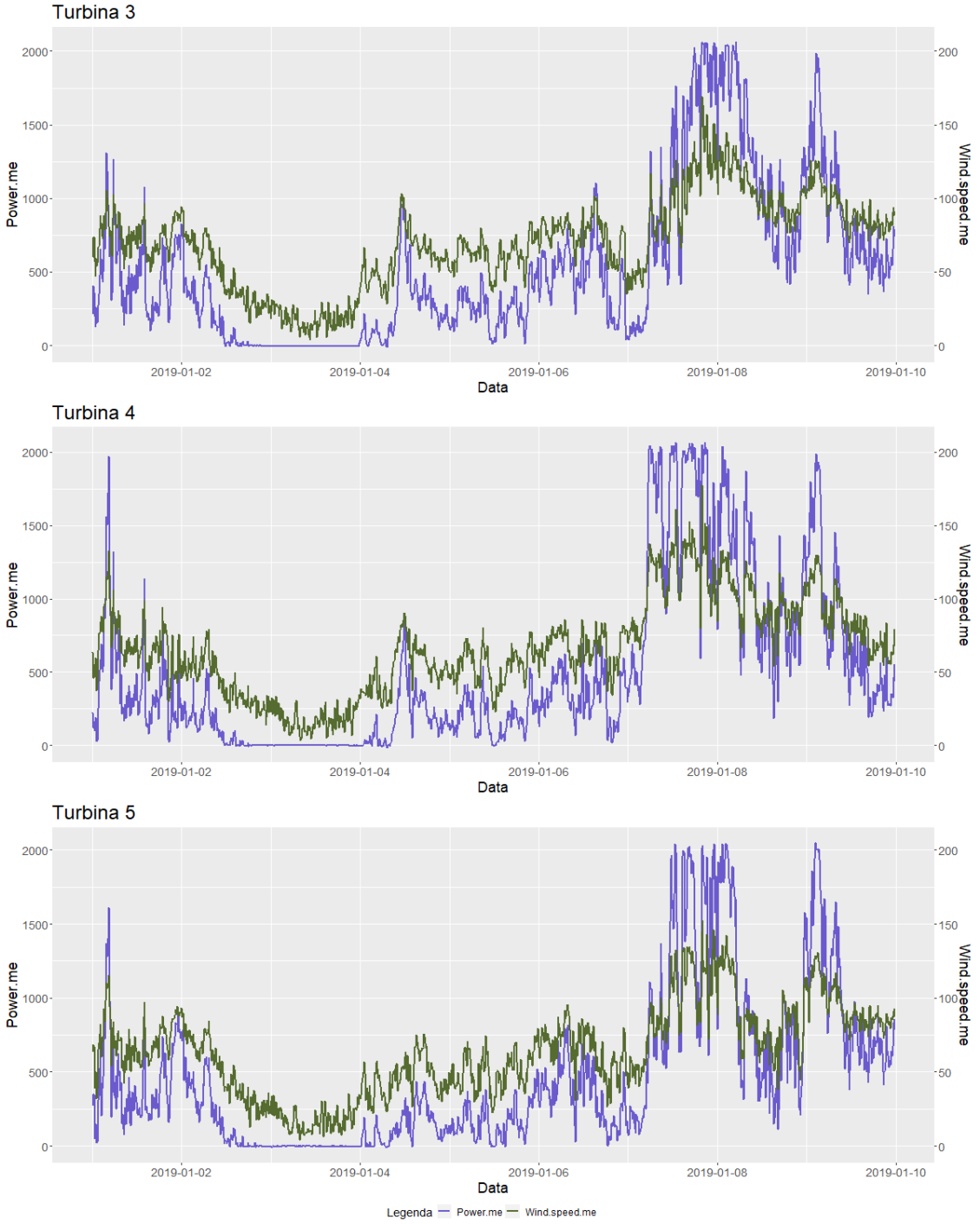


FIGURA 4.5: Energia media prodotta (*Power.me*) e velocità del vento media (*Wind.speed.me*) per il periodo che va dal 1° gennaio 2019 al 10 gennaio 2019 per le 3 turbine.

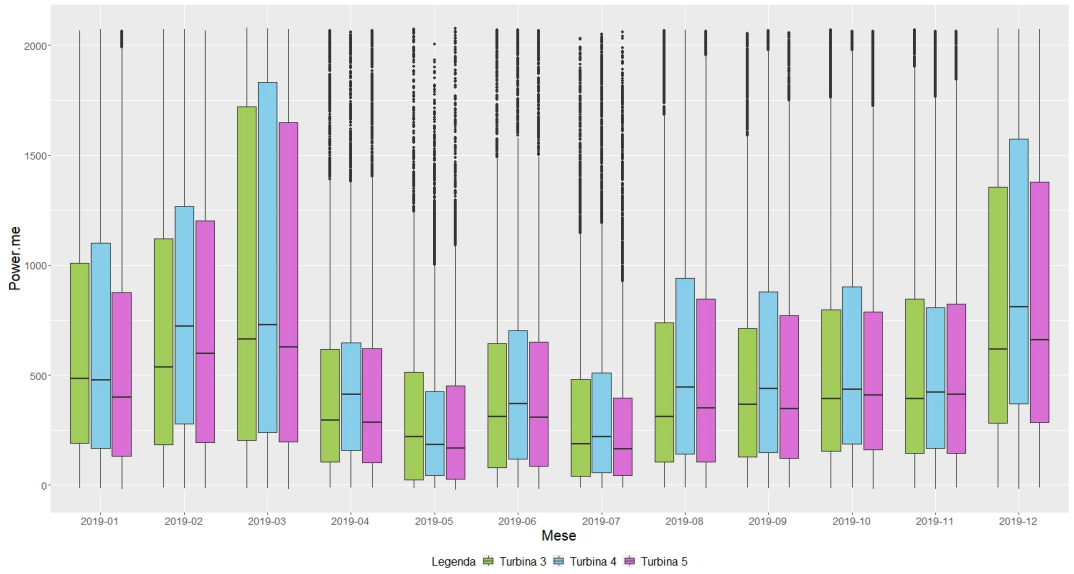


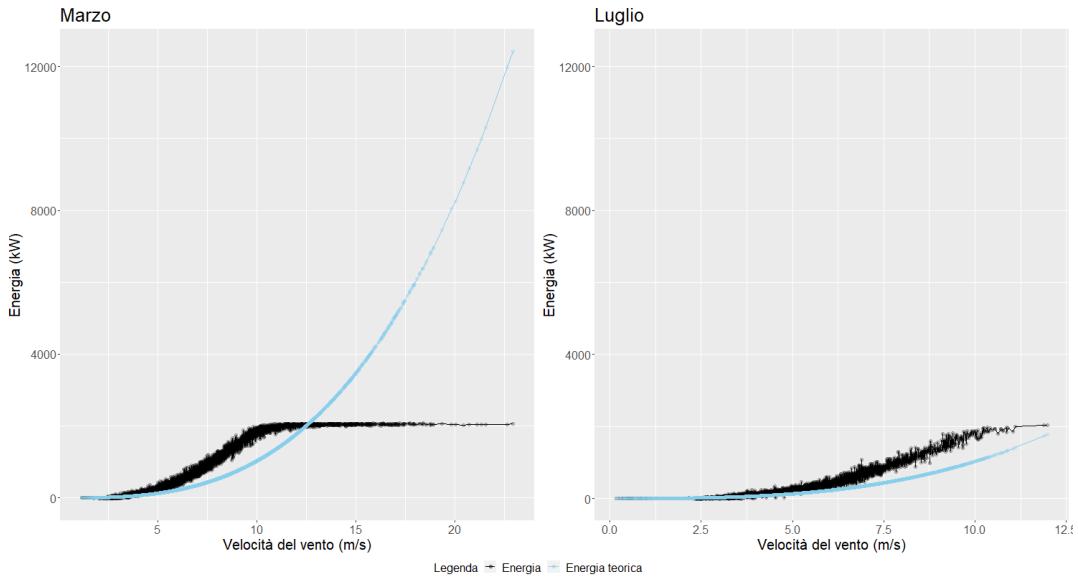
FIGURA 4.6: Energia media prodotta per l'anno 2019.

teorica producibile può essere calcolata come:

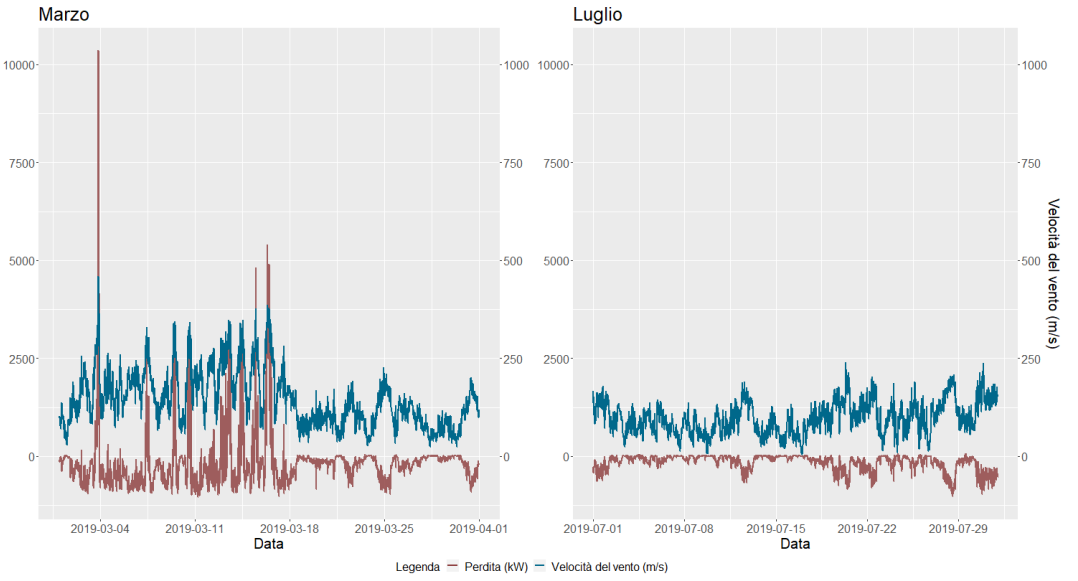
$$P_w = \frac{1}{2}\rho A v^3, \tag{4.1}$$

dove ρ indica la densità dell'aria (1.225 kg/m^3), A l'area del rotore della turbina (in questo caso 6.720/4 m^2) e v la velocità del vento in m/s (Delgado & Fahim, 2020). Essa viene raffigurata in Figura 4.7 per due mesi relativi alla turbina 3, rappresentanti dei casi estremi. Infatti, come si può notare dalla Figura 4.6, il mese di marzo ha presentato i valori più elevati di energia prodotta rispetto a tutti gli altri mesi, mentre il mese di luglio è risultato tra i meno proficui in termini di produzione energetica.

Un aspetto interessante da notare è che, nonostante l'energia teorica prevista fosse molto più alta nel mese di marzo, si raggiunge un *plateau* di produzione. Questo comportamento indica che, nonostante la potenziale energia teorica superiore, il sistema raggiunge un punto di saturazione, oltre il quale non è possibile ottenere ulteriori incrementi nella produzione di energia, data la massima energia producibile dalla turbina stessa. Si nota in ogni caso come l'energia prodotta sia una funzione cubica della velocità del vento e come vi siano delle velocità di *cut-in* e di *cut-off*, ovvero soglie al di sotto delle quali la potenza generata è nulla e al di sopra delle quali si raggiunge una *plateau* e la produzione diventa costante. Per evitare danni all'impianto infatti il funzionamento delle turbine viene interrotto, causando una perdita di energia. Questa perdita è calcolata come la differenza tra l'energia teorica che potrebbe essere generata e l'energia effettivamente prodotta. Questa situazione viene rappresentata nei grafici in



(a) Energia teorica



(b) Perdita

FIGURA 4.7: Energia media prodotta in kW (in nero) ed energia teorica (in azzurro) in funzione della velocità del vento in m/s (Figura 4.7(a)) e confronto tra velocità del vento (in blu) e perdita di energia registrata (in rosso) calcolata come differenza tra energia teorica e prodotta (Figura 4.7(b)). Tali quantità sono relative alla turbina 3.

Figura 4.7, dove la perdita è rappresentata insieme alla velocità del vento registrata. Per il mese di marzo è infatti evidente come la perdita sia notevole quando la velocità del vento supera i 500 m/s, cosa che invece non si verifica mai nel mese di luglio, non superando mai infatti i 250 m/s. Risulta poi interessante notare quali siano i mesi di maggiore e minore produzione di energia, come si può osservare nella Figura 4.6, dove emerge che i mesi compresi tra dicembre ed aprile registrano la massima produzione

di energia. Questo periodo corrisponde ai mesi invernali, che risultano essere infatti i periodi più ventosi nel Regno Unito. Si nota inoltre che la turbina 4 in generale sembra riuscire a produrre più energia, probabilmente a causa della posizione.

4.5 Modellazione

In questa Sezione, saranno sviluppati modelli al fine di analizzare la variabile di risposta rappresentata da *Power.me*, che rappresenta l'energia media prodotta, calcolata come valore medio ogni 10 minuti. Si procederà sia con modelli interpretabili che con modelli noti principalmente per la loro capacità previsiva. In particolare, verranno utilizzati i modelli autoregressivi vettoriali (VAR) come *benchmark*, ma sarà preso in considerazione anche il modello di regressione lineare multivariata sviluppato nei capitoli precedenti. I modelli verranno confrontati mediante alcune metriche, ossia l'errore quadratico medio (MSE, *Mean Squared Error*) e la sua radice (RMSE, *Root Mean Squared Error*) e l'errore percentuale assoluto medio (MAPE, *Mean Absolute Percentage Error*). Nella trattazione seguente, la variabile risposta non ritardata verrà indicata con $\mathbf{Y}_T$, mentre la sua versione ritardata di un lag con $\mathbf{Y}_{T-1}$ e saranno definite come

$$\mathbf{Y}_T = \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_T \end{bmatrix} = \begin{bmatrix} y_{1,1} & y_{2,1} & y_{3,1} \\ y_{1,2} & y_{2,2} & y_{3,2} \\ \vdots & & \\ y_{1,T} & y_{2,T} & y_{3,T} \end{bmatrix} \in \mathbb{R}^{T \times q}, \tag{4.2}$$

$$\mathbf{Y}_{T-1} = \begin{bmatrix} \mathbf{y}_0 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_{T-1} \end{bmatrix} = \begin{bmatrix} y_{1,0} & y_{2,0} & y_{3,0} \\ y_{1,1} & y_{2,1} & y_{3,1} \\ \vdots & & \\ y_{1,T-1} & y_{2,T-1} & y_{3,T-1} \end{bmatrix} \in \mathbb{R}^{T \times q}, \tag{4.3}$$

con $\mathbf{y}_t = (y_{1,t}, y_{2,t}, y_{3,t})$, con $t = 1, \ldots, T$, $T = 52.559$ e $q = 3$. Per quanto riguarda le variabili esplicative, si ha che ciascuna turbina dispone di una serie di rilevazioni proprie. Non si dispone quindi di variabili esplicative comuni. Per ogni turbina si dispone infatti di 50 variabili, di cui una la variabile risposta *Power.me*. Si avrà quindi $p_j = 49$ per ogni $j = 1, 2, 3$. Si definiscono quindi le matrici delle variabili esplicative non ritardate e ritardate di un lag, definite rispettivamente come

$$\mathbf{X}_T = \begin{bmatrix} \mathbf{X}_T^1 & \mathbf{X}_T^2 & \mathbf{X}_T^3 \end{bmatrix} \in \mathbb{R}^{T \times \sum_j p_j}, \tag{4.4}$$

$$\mathbf{X}_{T-1} = \begin{bmatrix} \mathbf{X}_{T-1}^1 & \mathbf{X}_{T-1}^2 & \mathbf{X}_{T-1}^3 \end{bmatrix} \in \mathbb{R}^{T \times \sum_j p_j}, \tag{4.5}$$

con

$$\mathbf{X}_T^j = \begin{bmatrix} \mathbf{x}_1^j \\ \mathbf{x}_2^j \\ \vdots \\ \mathbf{x}_T^j \end{bmatrix} = \begin{bmatrix} x_{1,1}^j & x_{2,1}^j & \dots & x_{p_j,1}^j \\ x_{1,2}^j & x_{2,2}^j & \dots & x_{p_j,2}^j \\ \vdots & \vdots & \ddots & \vdots \\ x_{1,T}^j & x_{2,T}^j & \dots & x_{p_j,T}^j \end{bmatrix}, \tag{4.6}$$

$$\mathbf{X}_{T-1}^j = \begin{bmatrix} \mathbf{x}_0^j \\ \mathbf{x}_1^j \\ \vdots \\ \mathbf{x}_{T-1}^j \end{bmatrix} = \begin{bmatrix} x_{1,0}^j & x_{2,0}^j & \dots & x_{p_j,0}^j \\ x_{1,1}^j & x_{2,1}^j & \dots & x_{p_j,1}^j \\ \vdots & \vdots & \ddots & \vdots \\ x_{1,T-1}^j & x_{2,T-1}^j & \dots & x_{p_j,T-1}^j \end{bmatrix}, \tag{4.7}$$

entrambe di dimensione $T \times p_j$, con $\mathbf{x}_t^j = (x_{1,t}^j, x_{2,t}^j, \dots, x_{p_j,t}^j)$, con $j = 1, 2, 3$.

4.5.1 Analisi preliminari

Prima di cominciare la modellazione si rende necessario effettuare alcune analisi preliminari riguardo alla variabile risposta $\mathbf{Y}_T$. Innanzitutto si è andata ad analizzare la funzione di autocorrelazione (ACF) e di autocorrelazione parziale (pACF), le quali vengono riportate in Figura 4.8 soltanto per la turbina 3, dal momento che risultano essere molto simili per tutte e tre le turbine.

Esse vengono riportate sia per la serie originaria che per la sua versione differenziata, quest'ultima data da $\Delta\mathbf{y}_{j,T} = \mathbf{y}_{j,T} - \mathbf{y}_{j,T-1}$, con $\mathbf{y}_{j,T} = (y_{j,1}, y_{j,2}, \dots y_{j,T})$ e $\mathbf{y}_{j,T-1} = (y_{j,0}, y_{j,1}, \dots y_{j,T-1})$ con $T = 52.559$, con $j = 1, 2, 3$. L'ACF viene riportata fino ad un lag di 25.920, ossia per 6 mesi, mentre la pACF viene riportata fino ad un lag di 1008, ossia per 7 giorni, ed è stato troncato anche l'asse delle ordinate al fine di rendere maggiormente visibile l'andamento della funzione. La funzione di autocorrelazione per $\mathbf{y}_{1,t}$ denota una forte persistenza della serie, essendo l'ACF al di fuori degli intervalli anche a lag superiori a 20.000. Dall'analisi grafica non risulta possibile individuare una stagionalità, dal momento che anche concettualmente è probabile che la stagionalità, se presente, sia a livello annuale, ma trattando un solo anno, anche per limiti computazionali, non è possibile verificarlo.

Viene analizzata poi la stazionarietà dei tre processi attraverso il test per le radici unitarie di Dickey & Fuller, che rifiuta l'ipotesi nulla di presenza di radice unitaria, confermando che i tre processi sono stazionari. Si procede quindi utilizzando le serie non differenziata.

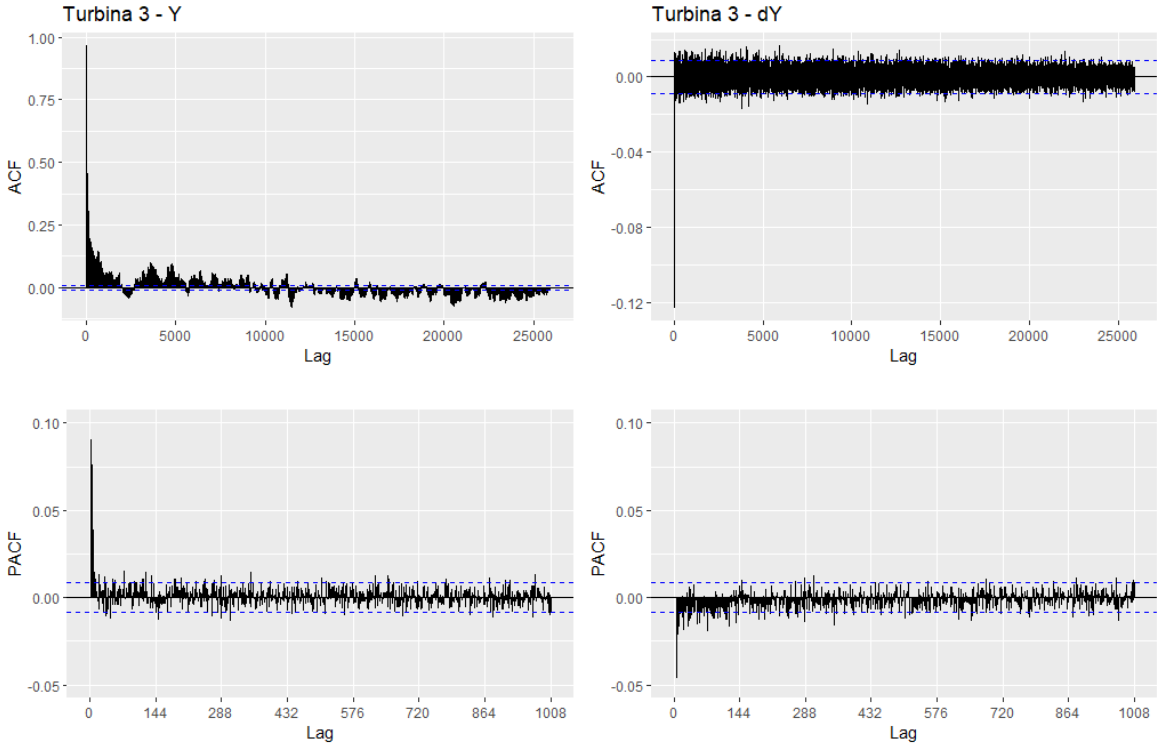


FIGURA 4.8: ACF e PACF relative alla variabile risposta Y e alla variabile risposta differenziata dY per la turbina 3.

4.5.2 Modelli autoregressivi vettoriali

Si decide di usare come modello *benchmark* il modello autoregressivo vettoriale, in inglese *Vector Autoregression* (VAR), con intercetta. Tale modello è stato stimato sia includendo le variabili esplicative standardizzate che escludendole; per la selezione del numero di ritardi da includere nel modello si decide poi di considerarne tre valori, ossia 6, 12 e 18, corrispondenti all'includere nel modello tutti i ritardi rispettivamente fino alla prima, alla seconda e alla terza ora. La variabile risposta $\mathbf{Y}_T \in \mathbb{R}^{T\times q}$, con $q = 3$ e $T = 52.559$ è definita in Equazione (4.2), mentre le variabili esplicative hanno un struttura del tipo $\mathbf{X}_T \in \mathbb{R}^{T\times\sum_j p_j}$ come in Equazione (4.4), con $\sum_j p_j = 147$, dato che $p_j = 49$ per ogni $j = 1, 2, 3$. Ciò implica l'introduzione quindi di tutte le variabili esplicative relative ad ogni turbina al tempo T, non disponendo di variabili esplicative comuni a tutte le turbine. Verranno poi considerati sia i modelli completi, ossia comprendenti tutti i parametri, che i modelli ridotti, ottenuti imponendo a 0 i parametri che risultassero non significativi. Ci si riferirà ai modelli precedenti rispettivamente con il nome VAR_{lag} e VARX_{lag}, con $lag = 6, 12, 18$, nel caso dei modelli completi, e con $\mathsf{VAR}_{lag,rid}$ e $\mathsf{VARX}_{lag,rid}$, con $lag = 6, 12, 18$ nel caso dei modelli ridotti. Il numero di parametri inclusi in ciascun modello è visibile in Tabella 4.2. Come criteri di valutazione si utilizzeranno criteri di informazione come l'AIC, il BIC e il criterio di informazione di Hannan-Quinn

Numero di parametri inclusi

	VAR_{lag}		VARX_{lag}	
Lag	**Totale**	**Selezionati**	**Totale**	**Selezionati**
6	57	43	501	310
12	111	60	555	316
18	165	79	609	327

TABELLA 4.2: Numero di parametri inclusi nel modello completo (Totale) e numero di parametri inclusi nei modelli ridotti (Selezionati).

Modelli completi senza variabili esogene VAR_{lag}

					R^2_{adj}		
Lag	**AIC**	**HQ**	**BIC**	**FPE**	**Turbina 3**	**Turbina 4**	**Turbina 5**
6	29.133	29.136	29.142	$4.492814 \times e^{12}$	0.93484	0.94177	0.93495
12	29.125	29.131	29.143	$4.455319 \times e^{12}$	0.93498	0.94201	0.93516
18	29.125	29.133	29.152	$4.453557 \times e^{12}$	0.93504	0.94204	0.93522

Modelli ridotti senza variabili esogene $\mathsf{VAR}_{lag,rid}$

					R^2_{adj}		
Lag	**AIC**	**HQ**	**BIC**	**FPE**	**Turbina 3**	**Turbina 4**	**Turbina 5**
6	29.133	29.136	29.141	$4.494384 \times e^{12}$	0.96685	0.97217	0.96662
12	29.126	29.129	29.136	$4.466172 \times e^{12}$	0.96692	0.97229	0.96673
18	29.125	29.129	29.138	$4.468420 \times e^{12}$	0.96695	0.97229	0.96675

TABELLA 4.3: Metriche relative ai modelli VAR stimati con 6, 12 e 18 ritardi senza inclusione delle variabili esogene.

(HQ), nonché il coefficiente di determinazione corretto (R^2_{adj}) e il *Final Prediction Error* (FPE). Per i modelli stimati senza variabili esogene, i risultati sono riportati in Tabella 4.3. Si nota innanzitutto come l'esclusione dal modello dei parametri non significativi consenta di ottenere un aumento del coefficiente di determinazione aggiustato di circa il 3%, mentre non si notano miglioramenti a livello di criteri di informazione. In generale, comunque, non si nota un miglioramento consistente dell'adattamento ai dati aumentando il lag. Si nota poi dalla Tabella 4.2 come il vantaggio, anche in termini di parsimonia, sia piuttosto rilevante, dal momento che i modelli ridotti non hanno mai più di 80 parametri, contro i 165 massimi del modello più ampio di ordine 12.

I risultati ottenuti in seguito all'inclusione delle variabili esogene sono invece riportati in Tabella 4.4. Rispetto ai modelli senza le variabili esplicative, si nota un consistente aumento del coefficiente di determinazione, che raggiunge il 99%, il quale rimane tale anche in seguito alla rimozione dei parametri non significativi. Il numero di parametri in generale, come visibile dalla Tabella 4.2, è significativamente più alto, superando i 500 nei modelli completi. I modelli ridotti invece includono tra i 310 e i 327 parametri.

Si prova in ogni caso ad utilizzare anche dei metodi di selezione automatica del numero di ritardi da includere nel modello, senza imporne dei valori specifici come è

Modelli completi con variabili esogene VARX_{lag}

Lag	AIC	HQ	BIC	FPE	R^2_{adj} Turbina 3	R^2_{adj} Turbina 4	R^2_{adj} Turbina 5
6	22.696	22.699	22.705	$7.192690 \times e^9$	0.99342	0.99420	0.99423
12	22.696	22.702	22.714	$7.191519 \times e^9$	0.99343	0.99421	0.99423
18	22.696	22.705	22.723	$7.190900 \times e^9$	0.99344	0.99421	0.99423

Modelli ridotti con variabili esogene $\mathsf{VARX}_{lag,rid}$

Lag	AIC	HQ	BIC	FPE	R^2_{adj} Turbina 3	R^2_{adj} Turbina 4	R^2_{adj} Turbina 5
6	22.777	22.794	22.830	$7.724039 \times e^9$	0.99664	0.99707	0.99703
12	22.778	22.794	22.831	$7.742274 \times e^9$	0.99665	0.99707	0.99703
18	22.777	22.794	22.832	$7.748559 \times e^9$	0.99665	0.99707	0.99703

Tabella 4.4: Metriche relative ai modelli VAR stimati con 6, 12 e 18 ritardi con inclusione delle variabili esogene.

	AIC	HQ	BIC	FPE
VAR	18	10	10	18
VARX	18	2	2	18

Tabella 4.5: Ritardi selezionati con criteri di selezione automatica imponendo un lag massimo pari a 18.

stato fatto in precedenza. Si è imposto un lag massimo pari a 18, valutando i medesimi criteri citati in precedenza. I risultati sono riportati in Tabella 4.5 e si nota come l'AIC e il FPE individuino un numero di lag pari a 18 in entrambi i casi, mentre il BIC e il criterio di Hannan-Quinn scelgano dei modelli più parsimoniosi.

Si sono effettuate poi le previsioni su un periodo di 14 giorni immediatamente successivi al termine dell'insieme di stima, quindi per i primi 14 giorni di gennaio 2020. Per ciascun modello si riporta il MAPE, il MSE e il RMSE in Tabella 4.6. Per quanto riguarda i modelli senza variabili esogene, si notano valori delle metriche di valutazione invariati anche in seguito alla rimozione dei parametri non significativi. Un leggero miglioramento, sia in termini di MAPE che di RMSE si nota invece nel caso dei modelli con variabili esogene. In generale si nota una performance nettamente migliore nel caso in cui le variabili esplicative vengano incluse nel modello, con RMSE minore di circa 700 kW, con anche il MAPE che diminuisce nettamente.

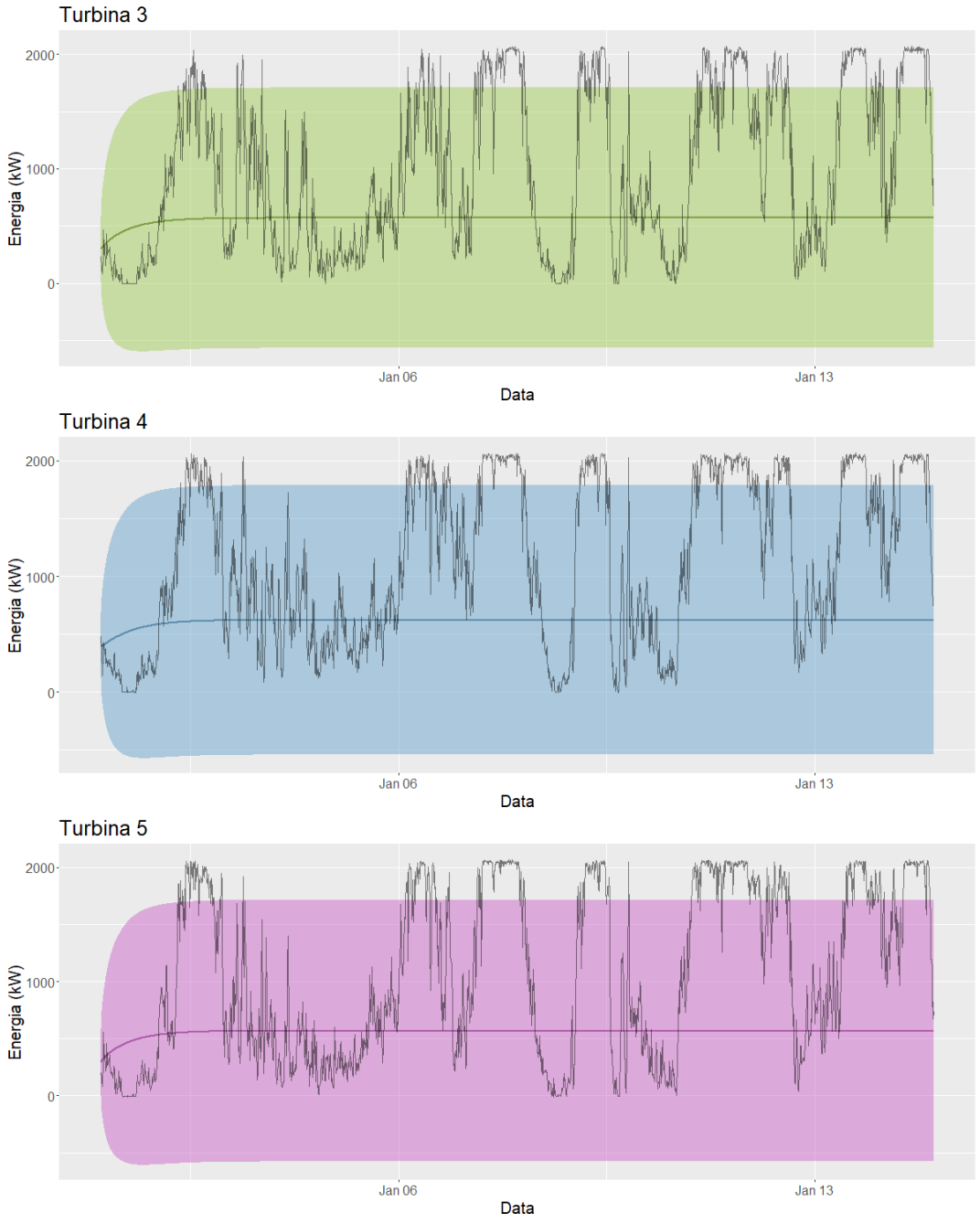


FIGURA 4.9: Previsioni (linee colorate), con relativi intervalli di confidenza, effettuate con il modello $\mathsf{VAR}_{6,rid}$ per tutte le turbine.

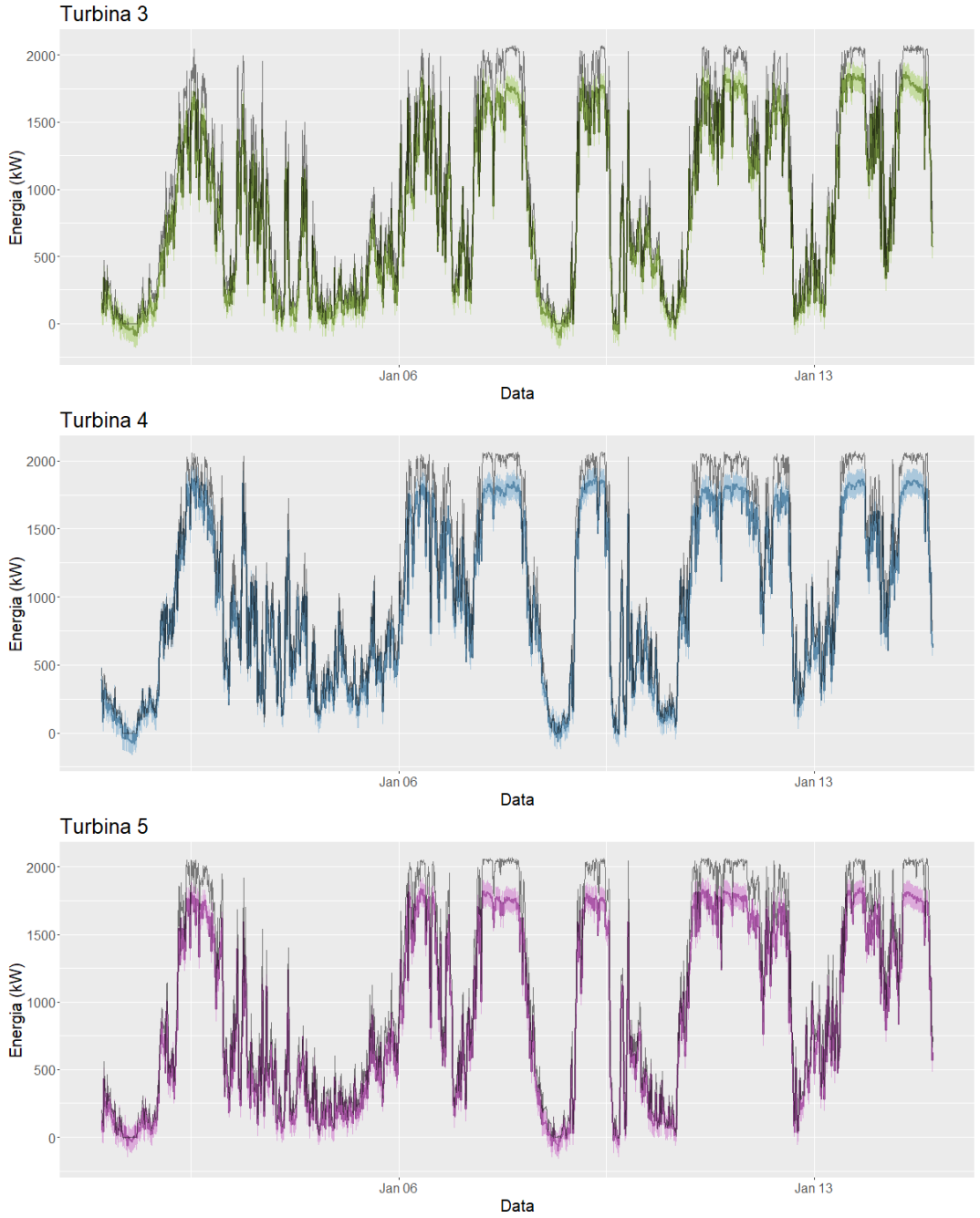


FIGURA 4.10: Previsioni (linee colorate), con relativi intervalli di confidenza, effettuate con il modello $\mathsf{VARX}_{6,rid}$ per tutte le turbine.

Modelli completi

	VAR_{lag}			VARX_{lag}		
Lag	**MAPE**	**MSE**	**RMSE**	**MAPE**	**MSE**	**RMSE**
6	1.194	733818.346	856.631	1.027	32447.468	180.134
12	1.158	734321.848	856.932	1.051	32298.476	179.721
18	1.157	734602.813	857.087	1.069	32305.813	179.743

Modelli ridotti

	$\mathsf{VAR}_{lag,rid}$			$\mathsf{VARX}_{lag,rid}$		
Lag	**MAPE**	**MSE**	**RMSE**	**MAPE**	**MSE**	**RMSE**
6	1.194	733820.257	856.634	1.041	31888.936	178.571
12	1.157	734351.522	856.946	1.043	31942.134	178.726
18	1.153	734651.138	857.127	1.066	32033.127	178.984

TABELLA 4.6: Metriche relative alle previsioni effettuate con i modelli VAR stimati con 6, 12 e 18 ritardi, sia con che senza variabili esogene.

La notevole differenza in termini di previsioni si nota anche dalle Figure 4.9 e 4.10, dove vengono riportate le previsioni e i valori osservati della potenza prodotta da ciascuna turbina. Si riportano rispettivamente in Figura 4.9 le previsioni relative al modello ridotto senza variabili esogene con 6 ritardi ($\mathsf{VAR}_{6,rid}$), e in Figura 4.10 le previsioni relative al modello con variabili esogene sempre con 6 ritardi ($\mathsf{VARX}_{6,rid}$). Si sono infatti considerati questi due i due modelli migliori nelle rispettive categorie, in quanto a parità di RMSE e di MAPE, consentono una maggiore parsimonia. Si nota come le previsioni siano nettamente più precise in seguito all'inclusione delle variabili esplicative, pur sottostimando i picchi di produzione in alcuni casi.

4.5.3 Regressione lineare multivariata

Si adatta a questo punto il modello di regressione multivariata, come definito nel Capitolo 1. Verrà utilizzato il modello con decomposizione thin QR, in quanto è stato appurato che risulta essere il più efficiente, valutando anche l'efficacia dello *stochastic*

Numero di variabili esplicative

		thin QR			thin QR SS	
Modello	p	**MaP**	**MPM 99%**	**MPM 95%**	**MaP**	**MPM 99%**	**MPM 95%**
$\mathsf{mlinreg}_{\mathsf{X}_{T-1}}$	148	77	35	64	53	52	53
$\mathsf{mlinreg}_{\mathsf{X}_{T-1,\mathsf{T}}}$	295	213	45	133	161	135	154
$\mathsf{mlinreg}_{\mathsf{X}_{T-1,\mathsf{T}},\mathsf{Y}_{T-1}}$	298	205	50	116	101	93	100

TABELLA 4.7: Numero di variabili esplicative totali (p) potenzialmente includibili nel modello e numero di variabili scelte dai modelli thin QR e thin QR con *stochastic search*, relativi alle strutture delle matrici di disegno definite nelle Equazioni (4.8)-(4.10). Si riportano i risultati relativi sia al modello MaP che MPM, utilizzando delle soglie al 99% e al 95%.

search. Verranno adattate le seguenti specificazioni della matrice di disegno:

$$\mathbf{Y}_T = \mathbf{X}_{T-1}\mathbf{B} + \mathbf{E}, \tag{4.8}$$

$$\mathbf{Y}_T = (\mathbf{X}_T, \mathbf{X}_{T-1})\mathbf{B} + \mathbf{E}, \tag{4.9}$$

$$\mathbf{Y}_T = (\mathbf{X}_T, \mathbf{X}_{T-1}, \mathbf{Y}_{T-1})\mathbf{B} + \mathbf{E}, \tag{4.10}$$

con $\mathbf{Y}_T, \mathbf{Y}_{T-1}, \mathbf{X}_T$ e $\mathbf{X}_{T-1}$ definite nelle Equazioni (4.2)-(4.5), $\mathbf{E} = \left(\boldsymbol{\varepsilon}_1, \boldsymbol{\varepsilon}_2, \ldots, \boldsymbol{\varepsilon}_q\right) \in \mathbb{R}^{T\times q}$, che rappresenta il termine di errore, e matrice dei coefficienti $\mathbf{B} = \left(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_q\right) \in \mathbb{R}^{p\times q}$. Non potendo disporre di variabili esplicative comuni per tutte e tre le turbine, si è deciso anche in questo caso infatti di regredire ciascuna variabile risposta sulle variabili esplicative di ciascuna turbina. Tutti i modelli vengono stimati con l'intercetta, con un numero variabili esplicative totali p (comprendenti anche l'intercetta) visibili in Tabella 4.7, con 10.000 iterazioni dell'algoritmo. I modelli nelle Equazioni (4.8)-(4.10) verranno indicati rispettivamente con i nomi $\mathsf{mlinreg}_{\mathsf{X}_{T-1}}$, $\mathsf{mlinreg}_{\mathsf{X}_{T-1,\mathsf{T}}}$, $\mathsf{mlinreg}_{\mathsf{X}_{T-1,\mathsf{T}},\mathsf{Y}_{T-1}}$, indicando con la sigla ss nel caso sia stato usato *stochastic search*.

Il numero di parametri inclusi in ciascun modello è visibile in Tabella 4.7, notando come *stochastic search* consenta di ottenere dei modelli più parsimoniosi. Per quanto riguarda le previsioni, vengono effettuate anche in questo caso per un periodo di 14 giorni, corrispondente a 2.016 periodi, essendo le osservazioni registrate ogni 10 minuti. Si utilizzano cinque metodi per effettuarle, i primi tre prevedono l'utilizzo dei modelli con probabilità a posteriori più alta, mentre i restanti prevedono di includere le variabili la cui probabilità di inclusione marginale a posteriori sia superiore ad una determinata soglia, in questo caso si utilizzano le soglie del 99% e del 95%. Ci si riferisce a quest'ultimi con il termine *Median Probability Model* (MPM). Per quanto riguarda la prima categoria, il primo metodo prevede l'utilizzo del modello MaP, ossia *Maximum a Posteriori*, corrispondente al modello con probabilità a posteriori massima. Gli altri due metodi invece prevedono l'individuazione dei k modelli con probabilità a posteriori più

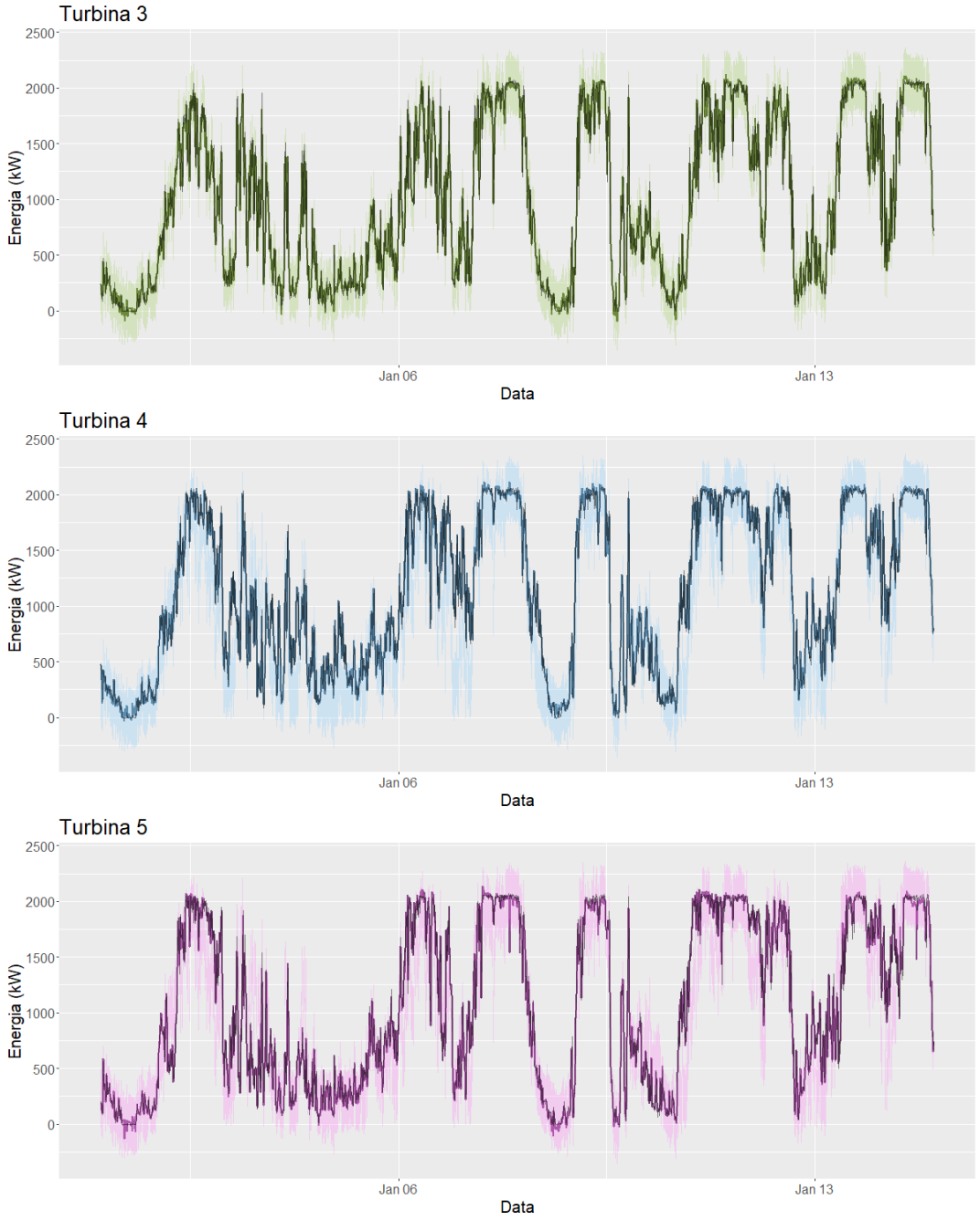


FIGURA 4.11: Previsioni (linee colorate), con relativi intervalli di confidenza, effettuate con il modello $\mathsf{mlinreg}_{\mathsf{X}_{T-1,T},\mathsf{Y}_{T-1}}$ stimato con *stochastic search*. Le previsioni sono state effettuate col modello MPM al 95%.

Modello $\mathsf{mlinreg}_{\mathsf{X}_{T-1}}$

	thin QR			thin QR SS		
	MAPE	**MSE**	**RMSE**	**MAPE**	**MSE**	**RMSE**
MaP	1.439	11486.728	107.176	1.204	8960.280	94.659
MaP media	1.473	11550.373	107.473	1.149	8718.195	93.371
MaP media pesata	1.474	11389.560	106.722	1.204	8960.280	94.659
MPM 99%	1.733	11771.799	108.498	1.173	8692.963	93.236
MPM 95%	1.558	10767.946	103.769	1.205	8960.280	94.659

Modello $\mathsf{mlinreg}_{\mathsf{X}_{T-1,\mathsf{T}}}$

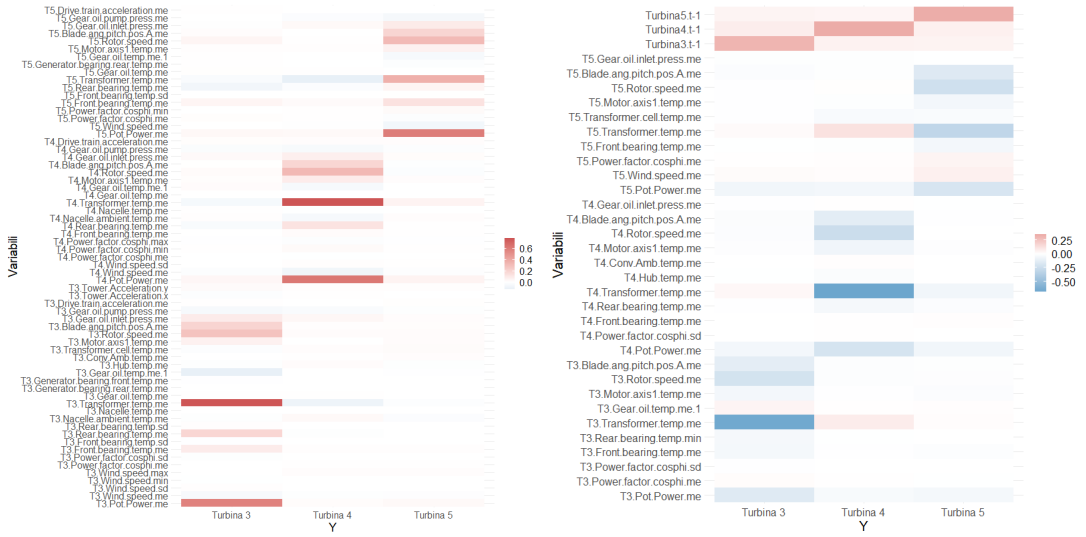
	thin QR			thin QR SS		
	MAPE	**MSE**	**RMSE**	**MAPE**	**MSE**	**RMSE**
MaP	1.087	7568.460	86.997	0.819	2912.564	53.968
MaP media	1.093	7526.360	86.755	0.813	2938.828	54.211
MaP media pesata	1.087	7510.286	86.662	0.819	2912.563	53.968
MPM 99%	1.303	8005.787	89.475	0.789	2503.802	50.038
MPM 95%	1.345	10902.004	104.413	0.805	2949.724	54.311

Modello $\mathsf{mlinreg}_{\mathsf{X}_{T-1,\mathsf{T}},\mathsf{Y}_{T-1}}$

	thin QR			thin QR SS		
	MAPE	**MSE**	**RMSE**	**MAPE**	**MSE**	**RMSE**
MaP	1.089	7232.680	85.045	0.675	2181.078	46.702
MaP media	1.100	7329.127	85.610	0.704	2195.265	46.854
MaP media pesata	1.096	7249.607	85.145	0.675	2181.078	46.702
MPM 99%	1.368	11665.602	108.007	0.742	2232.668	47.251
MPM 95%	1.017	4632.752	68.064	0.684	2181.015	46.701

TABELLA 4.8: Metriche relative alle previsioni effettuate con i modelli definiti nelle equazioni 4.8-4.10.

alta, sulle cui previsioni viene poi fatta la media, nel primo caso utilizzando una media aritmetica, nel secondo invece pesandola per le relative probabilità a posteriori. In questo caso vengono utilizzati i 10 modelli più probabili. I risultati relativi alle metriche di valutazione dell'errore di previsione sono visibili in Tabella 4.8. Si nota immediatamente come l'introduzione progressiva prima delle variabili esplicative al tempo T e poi della variabile risposta ritardata di un lag consenta progressivamente di migliorare le previsioni, con RMSE che diminuisce di quasi 50 kW dal modello $\mathsf{mlinreg}_{\mathsf{X}_{T-1}}$ al modello $\mathsf{mlinreg}_{\mathsf{X}_{T-1,\mathsf{T}},\mathsf{Y}_{T-1}}$ per quanto riguarda *stochastic search*. Analogamente anche il MAPE segue un'analoga diminuzione. La diminuzione è leggermente meno marcata nei modelli senza *stochastic search*, ma comunque rilevante. Molto spesso in ogni caso non si notano particolari differenze in termini di errore nemmeno tra previsioni effettuate con i modelli MaP e i modelli MPM, eccetto nei modelli $\mathsf{mlinreg}_{\mathsf{X}_{T-1,\mathsf{T}}}$, con *stochastic search* raggiungendo un risultato migliore, e nel modello $\mathsf{mlinreg}_{\mathsf{X}_{T-1,\mathsf{T}},\mathsf{Y}_{T-1}}$ senza *stochastic search*. In quest'ultimo caso il modello MPM al 99% ottiene un RMSE maggiore di 23 kW rispetto ai modelli MaP, mentre il modello MPM al 95% determina una diminuzione dello stesso di circa 17 kW. Si considera come modello migliore il modello MPM al 95 %



(a) Coefficienti inclusi al tempo T

(b) Coefficienti inclusi al tempo $T-1$

FIGURA 4.12: Variabili selezionate al tempo T e $T-1$ dal modello $\mathsf{mlinreg}_{\mathsf{X}_{T-1,T},\mathsf{Y}_{T-1}}$ con *stochastic search*. Viene utilizzato il modello MPM al 95%.

$\mathsf{mlinreg}_{\mathsf{X}_{T-1,T},\mathsf{Y}_{T-1}}$, che, come visibile in Tabella 4.7 prevede l'inclusione di 100 covariate. Le previsioni per ciascuna turbina, con relativi intervalli di confidenza, confrontati con i valori reali sono riportati in Figura 4.11, mentre il dettaglio delle covariate scelte ai tempi T e $T-1$ viene riportato rispettivamente nelle Figure 4.12(a)-4.12(b).

Per quanto riguarda il lato interpretativo, da queste ultime Figure si nota innanzitutto come le variabili selezionate abbiano un valore più alto per quanto riguarda le variabili proprie di ciascuna turbina, mentre assumano sostanzialmente tutte dei valori prossimi allo zero per quanto riguarda le variabili delle altre turbine. Le variabili più importanti possono sicuramente essere considerate quelle che si presentano per tutte e tre le turbine, fissato il tempo, con valori dei coefficienti più elevati. Al tempo T, ad esempio, si nota *Pot.Power.me*, l'energia potenziale media, assieme alla velocità media del vento *Wind.speed.me*.

In secondo luogo si individuano le temperature di alcune componenti della turbina, ossia *Transformer.temp.me*, *Front.Bearing.temp.me* e *Rear.bearing.temp.me*, nonché altre relative ad altre componenti, tra cui la velocità media del rotore, *Rotor.speed.me*, l'angolazione delle pale *Blade.ang.pitch.pos.A.me* e infine *Gear.oil.inlet.press.me*.

Quasi tutte si presentano con valori dei coefficienti positivi. Passando poi al tempo $T-1$ in Figura 4.12(b), si nota innanzitutto come i ritardi di ciascuna turbina siano tutti inclusi, presentino dei valori più elevati, presentando dei valori più elevati per i ritardi propri di ciascuna turbina. Tra le altre variabili, presenti per ciascuna turbina,

si annoverano sempre l'energia potenziale media *Pot.Power.me* e la velocità media del vento *Wind.speed.me*, nonché altre variabili presenti anche al tempo T, tra cui *Transformer.temp.me*, *Rotor.speed.me* e *Blade.ang.pitch.pos.A.me*. In questo caso invece la maggior parte dei coefficienti si presenta con segno negativo.

Tornando al lato previsivo e confrontando i risultati in Tabella 4.8 con quelli dei modelli VAR in Tabella 4.6, si notano risultati migliori in termini di RMSE anche rispetto ai modelli VARX, registrando una diminuzione di almeno 100 kW in tutti i modelli con *stochastic search*. Per quanto riguarda il MAPE, invece, il modello $\mathsf{mlinreg}_{\mathsf{X}_{\mathsf{T}-1}}$ risulta avere dei valori di tale metrica più elevati rispetto ai modelli VARX, mentre i modelli $\mathsf{mlinreg}_{\mathsf{X}_{\mathsf{T}-1,\mathsf{T}}}$ e $\mathsf{mlinreg}_{\mathsf{X}_{\mathsf{T}-1,\mathsf{T}},\mathsf{Y}_{\mathsf{T}-1}}$ ne registrano un miglioramento. Come visibile anche in Figura 4.11, si nota in generale un miglior adattamento del modello $\mathsf{mlinreg}_{\mathsf{X}_{\mathsf{T}-1,\mathsf{T}},\mathsf{Y}_{\mathsf{T}-1}}$ ai dati, riuscendo a prevedere in modo molto più accurato i picchi di energia, per tutte le turbine. Infatti, nel caso del $\mathsf{VARX}_{6,rid}$ in Figura 4.10, i picchi vengono in quel caso molto sottostimati.

4.5.3.1 Elicitazione della distribuzione a priori

Nei modelli precedenti, sempre facendo riferimento al modello descritto nella Sezione 1.5, sono stati usati i seguenti valori degli iperparametri: $\mathbf{B}_{\gamma,0} = \underline{0}_{p_\gamma \times q}$, $\mathbf{H}_\gamma = v\mathbf{I}_{p_\gamma}$, con $v = 100$, $c0 = 5$ e $\mathbf{C}_0 = k\mathbf{I}_q$, con $k = 0.2$. Il parametro θ relativo della distribuzione Bernoulli in Equazione 1.18 viene generato casualmente da una distribuzione $Beta(a, b)$ di parametri $a = 0.01$ e $b = 4.01$.

4.5.4 Regressione multivariata sparsa con stima della matrice di covarianza

Al fine di effettuare un'ulteriore comparazione, si è deciso di stimare anche un ulteriore modello di regressione multivariata. Si tratta del modello di regressione multivariata con stima della covarianza (*Multivariate Regression with Covariance Estimation*, MRCE) sviluppato da Rothman et al. (2010) e presente nel pacchetto *MRCE* su R. Tale modello si fonda sulla medesima struttura del modello di regressione multivariato sviluppato nel Capitolo 1, tuttavia imponendo delle penalizzazioni di tipo lasso per indurre sparsità.

Il modello, anche in questo caso, può essere espresso in forma matriciale. Sia $\mathbf{X} \in \mathbb{R}^{n \times p}$ la matrice di disegno contenente le variabili esplicative, sia $\mathbf{Y} \in \mathbb{R}^{n \times q}$ la matrice delle q variabili risposta e sia $\mathbf{E} \in \mathbb{R}^{n \times q}$ la matrice dei termini di errore, allora il modello

Modello	λ_1	λ_2	MAPE	MSE	RMSE	Variabili totali	Variabili incluse
$\mathsf{mrce}_{\mathsf{X}_{T-1}}$	1.00	0.01	0.79	30479.68	174.58	148	10
$\mathsf{mrce}_{\mathsf{X}_{T-1,T}}$	0.01	0.63	0.98	5366.11	73.25	295	40
$\mathsf{mrce}_{\mathsf{X}_{T-1,T},\mathsf{Y}_{T-1}}$	1.00	0.01	0.40	3623.85	60.20	298	38

TABELLA 4.9: Metriche relative alle previsioni effettuate con i modelli mrce, riportati assieme ai valori ottimi dei parametri scelti nonché al numero di variabili esplicative incluse e totali.

sarà definito come:

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E}, \tag{4.11}$$

con $\mathbf{E} \sim \mathsf{N}_{n\times q}(\mathbf{0}, \mathbf{\Sigma}, \mathbf{I}_n)$, che indica una distribuzione normale matriciale, con media zero, matrice di varianze e covarianze $\mathbf{\Sigma} = [\sigma_{i,j}]_{i,j=1,\dots,q} \in \mathbb{S}^q_{++}$ per le colonne di $\mathbf{Y}$ e $\mathbf{I}_n$, matrice identità di dimensione n, per le righe di $\mathbf{Y}$. Dato che $\mathbf{\Omega} = \mathbf{\Sigma}^{-1} = [\omega_{i,j}]_{i,j=1,\dots,q} \in \mathbb{S}^q_{++}$, la funzione di log-verosimiglianza negativa può essere espressa come

$$g(\mathbf{B}, \mathbf{\Sigma}) = \mathsf{tr}\left[\frac{1}{n}(\mathbf{Y} - \mathbf{XB})^\intercal(\mathbf{Y} - \mathbf{XB})\right] - \mathsf{log}|\mathbf{\Omega}|. \tag{4.12}$$

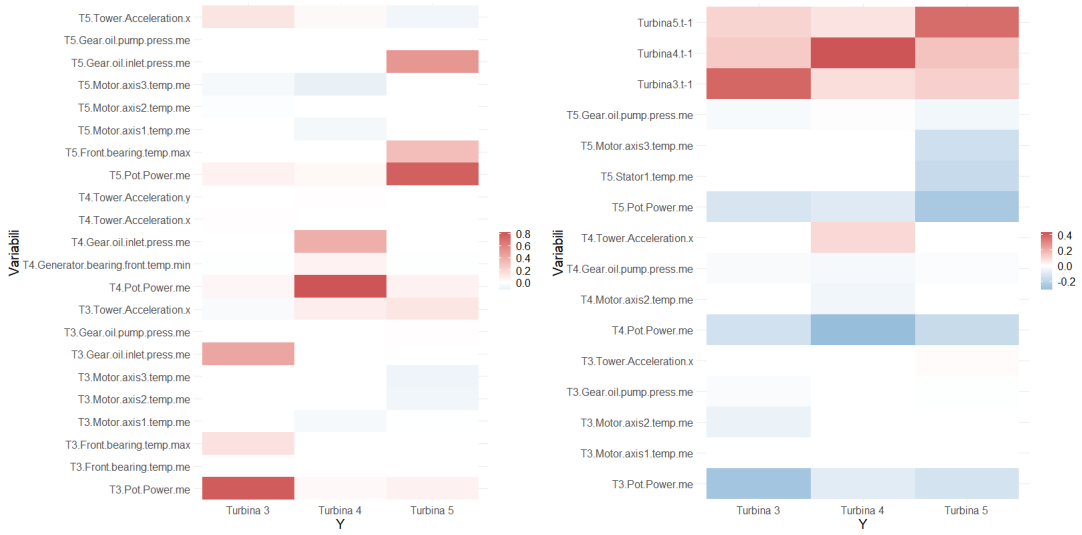
Il metodo in oggetto propone uno stimatore sparso per $\mathbf{B}$ che considera la correlazione tra gli errori penalizzando la verosimiglianza del modello. Viene aggiunta una penalità lasso sia su $\mathbf{B}$ che su $\mathbf{\Omega}$ alla funzione di log-verosimiglianza negativa del modello, ottenendo

$$(\widehat{\mathbf{B}}, \widehat{\mathbf{\Omega}}) = \mathsf{argmin}_{\mathbf{B},\mathbf{\Omega}}\left\{ g(\mathbf{B}, \mathbf{\Sigma}) + \lambda_1 \sum_{i\neq j} |\omega_{i,j}| + \lambda_2 \sum_{j=1}^{p}\sum_{k=1}^{q} |b_{j,k}| \right\}, \tag{4.13}$$

dove $b_{j,k}$ indica l'elemento in posizione j, k di $\mathbf{B}$.

La scelta dei parametri ottimi (λ_1, λ_2) può essere effettuata mediante *cross validation*. Tale modello viene quindi applicato al dataset in oggetto utilizzando la medesima specificazione delle matrici di disegno utilizzata per il modello di regressione multivariata nella Sezione precedente alle Equazioni (4.8)-(4.10), così da ottenere un confronto diretto. I rispettivi modelli saranno identificati con i nomi $\mathsf{mrce}_{\mathsf{X}_{T-1}}$, $\mathsf{mrce}_{\mathsf{X}_{T-1,T}}$ e $\mathsf{mrce}_{\mathsf{X}_{T-1,T},\mathsf{Y}_{T-1}}$.

Si effettua la scelta dei parametri ottimi (λ_1, λ_2) mediante validazione incrociata con 5 *fold*, valutando per entrambi 10 valori compresi tra 0.01 e 1, i cui valori scelti, assieme agli errori di previsione registrati, sono visibili in Tabella 4.9. Si può notare immediatamente come i valori di MAPE e RMSE siano comparabili con i valori registrati con il modello di regressione multivariata nella Sezione precedente, anche se in tutti i casi si hanno qui degli errori più alti. I risultati migliori anche in questo caso si



(a) Coefficienti inclusi al tempo T

(b) Coefficienti inclusi al tempo $T-1$

FIGURA 4.13: Variabili selezionate al tempo T e $T-1$ dal modello $\mathsf{mrce}_{\mathsf{X}_{T-1,T},\mathsf{Y}_{T-1}}$. Il colore indica il valore di ciascun coefficiente.

ottengono con il modello $\mathsf{mrce}_{\mathsf{X}_{T-1,T},\mathsf{Y}_{T-1}}$, comprendente anche il primo ritardo della variabile risposta. I grafici delle previsioni effettuate con questo modello, assieme ai veri valori della variabile risposta sono riportati in Figura 4.14, notando risultati molto simili al modello $\mathsf{mlinreg}_{\mathsf{X}_{T-1,T},\mathsf{Y}_{T-1}}$ con *stochastic search*. I coefficienti che sono stati mantenuti in tale modello sono visibili poi in Figura 4.13(a) per quanto riguarda il tempo T e in Figura 4.13(b) al tempo $T-1$.

Si può notare innanzitutto come il numero di variabili scelte da questo modello sia piuttosto inferiore rispetto al modello di regressione multivariata precedente, selezionando infatti soltanto 38 variabili, come indicato in Tabella 4.9, contro le 101 del modello $\mathsf{mlinreg}_{\mathsf{X}_{t-1,t},\mathsf{Y}_{t-1},\mathsf{SS}}$. Ad entrambi i tempi si nota tuttavia un analogo pattern relativamente ai coefficienti, che si presentano in valore assoluto più elevati per le variabili proprie di ciascuna turbina. Al tempo T si nota la presenza di *Pot.Power.me* per tutte le turbine con un valore più elevato rispetto alle restanti variabili e di *Gear.oil.inlet.press.me*. Si nota poi l'inclusione per le turbine 3 e 5 di *Front.bearing.temp.max*, (nel modello $\mathsf{mlinreg}_{\mathsf{X}_{t-1,t},\mathsf{Y}_{t-1},\mathsf{SS}}$ veniva invece inclusa *Front.bearing.temp.me*), e delle tre variabili relative a *Motor.axis.temp.me*. Anche in questo caso si presentano per la maggior parte con segno positivo. Passando al tempo $T-1$, si nota ancora una volta la presenza di *Pot.Power.me* per tutte le turbine con segno negativo, *Tower.acceleration.x* per le turbine 3 e 4, e alcune temperature medie relative a *Motor.axis.temp.me*. Si individua poi *Gear.oil.pump.press.me* per tutte e tre le turbine. Anche in questo caso tutti i ritardi delle turbine vengono inclusi nel modello, con valori più elevati per quanto riguarda i

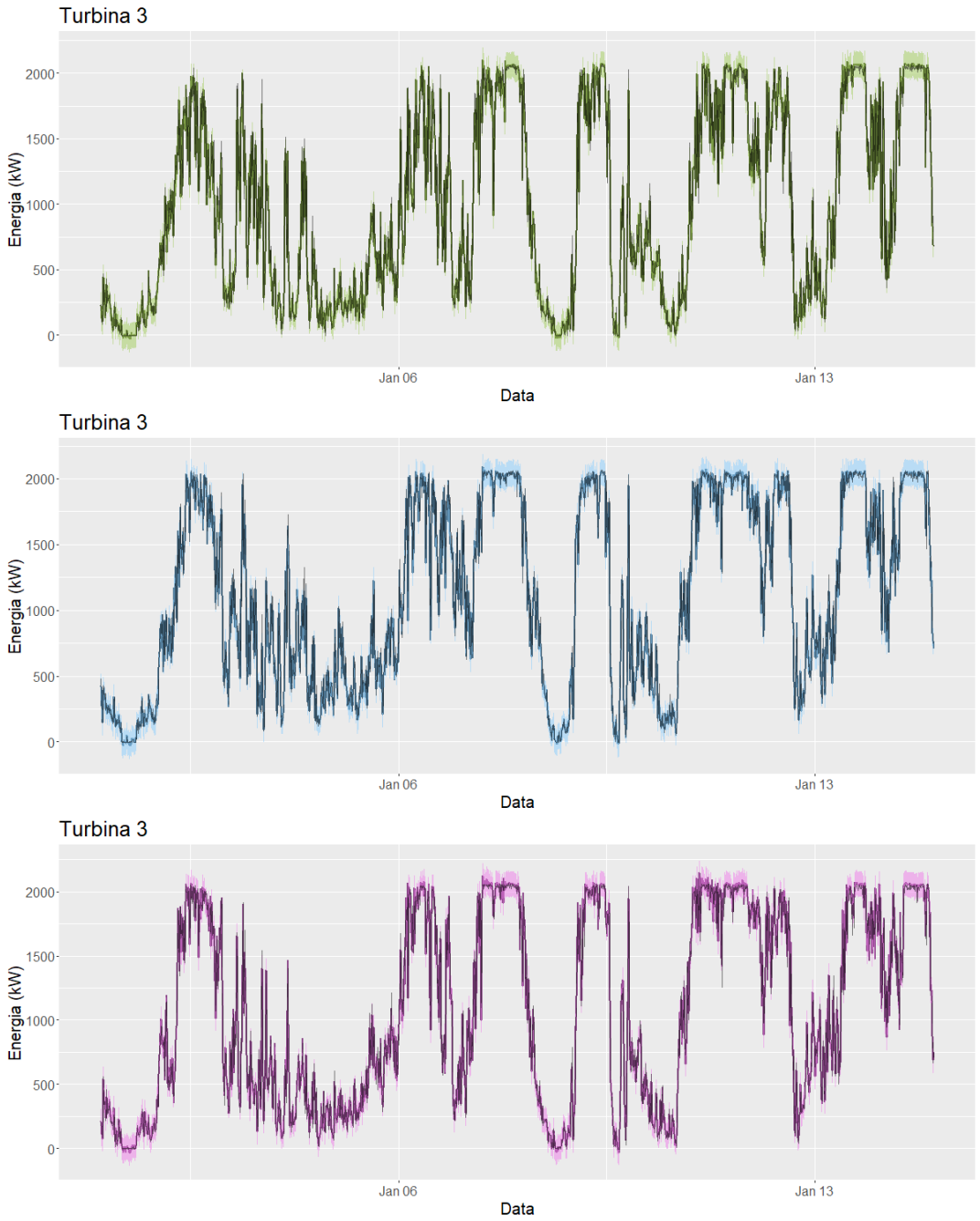


FIGURA 4.14: Previsioni (linee colorate), con relativi intervalli di confidenza, effettuate con il modello $\mathsf{mrce}_{\mathsf{X}_{T-1,T},\mathsf{Y}_{T-1}}$.

ritardi propri di ciascuna turbina.

4.6 Conclusioni

Al fine di valutare l'effettiva validità del metodo delineato in questo lavoro di tesi, si è portata a termine un'analisi ad un dataset reale relativo all'energia prodotta da alcune turbine eoliche, situate nel parco eolico di Kelmarsh, nel Regno Unito. D'interesse è risultato essere la previsione della variabile risposta *Power.me*, ossia l'energia media prodotta in intervalli di 10 minuti, sfruttando numerose variabili esplicative relative principalmente alla velocità del vento e a dati SCADA. Oltre ad alcune operazioni di preprocessing volte all'eliminazione e all'imputazione dei dati mancanti, si è sviluppata una modellazione basata su due classi di modelli. Da un lato si sono adattati dei modelli autoregressivi vettoriali, anche comprendendo le variabili esogene, dall'altro due modelli di regressione multivariata. Il primo, ovviamente, è il modello di regressione multivariata delineato nel Capitolo 1, stimato con l'ausilio di metodi di aggiornamento thin QR, confrontato con il modello MRCE definito nella Sezione 4.5.4, un modello di regressione multivariata sparsa con penalizzazioni di tipo lasso. I modelli VAR si sono rivelati essere i più carenti da un punto di vista previsivo e di adattamento ai dati. L'introduzione delle variabili esplicative ha consentito invece ai modelli VARX di ottenere dei risultati molto più soddisfacenti. Tuttavia, i vari modelli di regressione multivariata hanno consentito sostanzialmente sempre di ottenere dei risultati molto migliori sul lato previsivo rispetto alla modellazione VAR. Sono state valutate tre specificazioni della matrice di disegno, definite nelle Equazioni (4.8)-(4.10), adattate sia con il modello mlinreg che con il modello mrce. Per il modello mlinreg le previsioni sono state effettuate mediante vari modelli MaP, anche pesati, e MPM, comprendenti le variabili esplicative con probabilità di inclusione a posteriori superiore al 99% e al 95%. Per quanto riguarda il modello mrce, anch'esso ha consentito di ottenere dei risultati simili al modello mlinreg, ma non migliori. Questi due modelli hanno poi consentito anche degli spunti interpretativi valutando le variabili che sono state incluse nei rispettivi modelli migliori. Le differenze nelle previsioni tra i vari modelli sono visibili confrontando le Figure 4.9, 4.10, 4.11 e 4.14.

Conclusioni

In questo lavoro di tesi è stato trattato il modello di regressione lineare multivariato, con un focus particolare sulla stima bayesiana e l'induzione di sparsità attraverso l'uso di una distribuzione a priori *spike and slab* sui coefficienti di regressione, approccio particolarmente adatto a scenari caratterizzati da elevata dimensionalità. Si è riportato il metodo base di simulazione dalla distribuzione marginale del vettore $\boldsymbol{\gamma}$, ossia il *Gibbs sampling* nella versione *reversible jump*, esplorando poi gli aspetti computazionali legati alla sua applicazione. A tal fine sono stati introdotti dei metodi di aggiornamento basati sulla decomposizione QR e thin QR come tecniche algebriche per ridurre il carico computazionale derivante dalla stima del modello. In realtà, l'evidenza è stata che l'aggiornamento basato sulla decomposizione QR non risulta affatto efficiente, soprattutto nel momento in cui il numero di osservazioni inizia ad aumentare.

Al contrario, i procedimenti di aggiornamento fondati sulla decomposizione thin QR si sono dimostrati notevolmente più efficienti dal punto di vista del tempo computazionale, superando il modello naive in sostanzialmente tutti i casi di simulazione. L'unica eccezione si è notata quando gli esempi di simulazione avevano dimensioni di p ed n troppo esigue. Per quanto concerne l'approccio fondato sulla decomposizione QR, l'origine della sua limitata efficienza risiede nella necessità di calcolare e aggiornare anche la matrice $\mathbf{Q} \in \mathbb{R}^{n \times n}$, a differenza del metodo basato sulla decomposizione thin QR che sfrutta esclusivamente la matrice $\mathbf{R}$. Tuttavia, è emerso che i modelli esaminati presentano delle problematiche, evidenziate principalmente quando vi è una considerevole discrepanza tra il numero di variabili esplicative e il numero di osservazioni. In situazioni in cui il numero di variabili esplicative è elevato, l'algoritmo occasionalmente riscontra difficoltà nell'individuare un modello con un numero limitato di variabili da utilizzare come punto di partenza. Ciò può condurre alla creazione di modelli finali poco parsimoniosi, incapaci di rappresentare adeguatamente la struttura reale dei dati. Tuttavia, l'impiego di tecniche di *stochastic search* è risultato essere una valida soluzione al problema. Nonostante ciò, in scenari particolarmente complessi, il tempo computazionale richiesto per la stima dei modelli può comunque risultare elevato.

Appendice A

A.1 Dimostrazione della distribuzione a posteriori

Data la funzione di verosimiglianza per il modello di regressione multivariata per l'equazione (1.3) definita come

$$\mathcal{L}(\mathbf{Y}|\mathbf{X},\boldsymbol{\gamma},\mathbf{B}_\gamma,\boldsymbol{\Sigma}) = \frac{\exp\left\{-\frac{1}{2}\mathsf{tr}\left[\boldsymbol{\Sigma}^{-1}(\mathbf{Y}-\mathbf{X}\mathbf{B}_\gamma)^\intercal(\mathbf{Y}-\mathbf{X}\mathbf{B}_\gamma)\right]\right\}}{\mathcal{C}_{\mathsf{gauss}}(\boldsymbol{\Sigma},\mathbf{I}_n,n,q)}, \tag{A.1}$$

dove $\mathcal{C}_{\mathsf{gauss}}(\boldsymbol{\Sigma},\mathbf{I}_n,n,q) = (2\pi)^{nq/2}\ |\boldsymbol{\Sigma}|^{n/2}$ è una costante di normalizzazione, e date le distribuzione a priori in (1.13) per $\mathbf{B}$ e $\boldsymbol{\Sigma}$, esprimibili come

$$\pi(\mathbf{B}_\gamma|\mathbf{X},\boldsymbol{\gamma},\boldsymbol{\Sigma}) = \frac{\exp\left\{-\frac{1}{2}\mathsf{tr}\left[\boldsymbol{\Sigma}^{-1}(\mathbf{B}_\gamma-\mathbf{B}_{\gamma,0})^\intercal\mathbf{H}_\gamma^{-1}(\mathbf{B}_\gamma-\mathbf{B}_{\gamma,0})\right]\right\}}{\mathcal{C}_{\mathsf{gauss}}(\boldsymbol{\Sigma},\mathbf{H}_\gamma,q,p_\gamma)} \tag{A.2}$$

$$\pi(\boldsymbol{\Sigma}|c_0,\mathbf{C}_0) = \frac{|\boldsymbol{\Sigma}|^{-(c_0+q+1)/2}\exp\{-\frac{1}{2}\mathsf{tr}(\boldsymbol{\Sigma}^{-1}\mathbf{C}_0)\}}{\mathcal{C}_{\mathsf{IW}}(c_0,\mathbf{C}_0,q)}, \tag{A.3}$$

dove

$$\mathcal{C}_{\mathsf{gauss}}(\boldsymbol{\Sigma},\mathbf{H}_\gamma,q,p_\gamma) = (2\pi)^{p_\gamma q/2}|\boldsymbol{\Sigma}|^{p_\gamma/2}|\mathbf{H}_\gamma|^{q/2}$$

$$\mathcal{C}_{\mathsf{IW}}(c_0,\mathbf{C}_0,q) = 2^{c_0q/2}\pi^{(q(q-1))/4}|\mathbf{C}_0|^{-c_0/2}\prod_{j=1}^{q}\Gamma\big(\frac{c_0+1-j}{2}\big),$$

e $\Gamma(a) = \int_{\mathbb{R}}^{+} x^{a-1}e^{-x}dx$. Si ha che la distribuzione a posteriori congiunta di $(\mathbf{Y}, \mathbf{B}_\gamma, \boldsymbol{\Sigma})$ può essere fattorizzata come:

$$\begin{aligned}
\pi(\mathbf{Y}, \mathbf{B}_\gamma, \boldsymbol{\Sigma}|\mathbf{X}, \boldsymbol{\gamma}) &= \frac{\exp\left\{-\frac{1}{2}\mathsf{tr}\left[\boldsymbol{\Sigma}^{-1}(\mathbf{Y}-\mathbf{X}\mathbf{B}_\gamma)^\intercal(\mathbf{Y}-\mathbf{X}\mathbf{B}_\gamma)\right]\right\}}{\mathcal{C}_{\mathsf{gauss}}(\boldsymbol{\Sigma}, \mathbf{I}_n, n, q)} \\
&\quad \times \frac{\exp\left\{-\frac{1}{2}\mathsf{tr}\left[\boldsymbol{\Sigma}^{-1}(\mathbf{B}_\gamma-\mathbf{B}_{\gamma,0})^\intercal\mathbf{H}_\gamma^{-1}(\mathbf{B}_\gamma-\mathbf{B}_{\gamma,0})\right]\right\}}{\mathcal{C}_{\mathsf{gauss}}(\boldsymbol{\Sigma}, \mathbf{H}_\gamma, q, p_\gamma)} \\
&\quad \times \frac{|\boldsymbol{\Sigma}|^{-(c_0+q+1)/2}\exp\{-\frac{1}{2}\mathsf{tr}(\boldsymbol{\Sigma}^{-1}\mathbf{C}_0)\}}{\mathcal{C}_{\mathsf{IW}}(c_0, \mathbf{C}_0, q)} \\
&= \exp\left\{-\frac{1}{2}\mathsf{tr}\left[\boldsymbol{\Sigma}^{-1}\left(\left(\mathbf{B}_\gamma-\mathbf{K}_\gamma^{-1}\mathbf{M}\right)^\intercal\mathbf{K}_\gamma\left(\mathbf{B}_\gamma-\mathbf{K}_\gamma^{-1}\mathbf{M}\right)\right)\right]\right\} \\
&\quad \times \exp\left\{-\frac{1}{2}\mathsf{tr}\left[\boldsymbol{\Sigma}^{-1}\mathcal{Q}_\gamma\right]\right\} \\
&\quad \times \frac{|\boldsymbol{\Sigma}|^{-(c_0+q+1)/2}}{\mathcal{C}_{\mathsf{IW}}(c_0, \mathbf{C}_0, q)\mathcal{C}_{\mathsf{gauss}}(\boldsymbol{\Sigma}, \mathbf{H}_\gamma, q, p_\gamma)\mathcal{C}_{\mathsf{gauss}}(\boldsymbol{\Sigma}, \mathbf{I}_n, n, q)}, \qquad \text{(A.4)}
\end{aligned}$$

dove $\mathcal{Q}_\gamma = \mathbf{C}_0+\mathbf{C}+\mathbf{M}^\intercal\mathbf{K}_\gamma^{-1}\mathbf{M}$ con $\mathbf{C} = \mathbf{Y}^\intercal\mathbf{Y}+\mathbf{B}_{\gamma,0}^\intercal\mathbf{H}_\gamma^{-1}\mathbf{B}_{\gamma,0}$, $\mathbf{M} = \mathbf{X}^\intercal\mathbf{Y}-\mathbf{H}_\gamma^{-1}\mathbf{B}_{\gamma,0}$, $\mathbf{K}_\gamma = \mathbf{X}^\intercal\mathbf{X} + \mathbf{H}_\gamma^{-1}$. Il risultato in (A.4) si ottiene dal completamento del quadrato in $\mathbf{B}_\gamma$ nell'espressione

$$(\mathbf{Y}-\mathbf{X}\mathbf{B}_\gamma)^\intercal(\mathbf{Y}-\mathbf{X}\mathbf{B}_\gamma) + (\mathbf{B}_\gamma-\mathbf{B}_{\gamma,0})^\intercal\mathbf{H}_\gamma^{-1}(\mathbf{B}_\gamma-\mathbf{B}_{\gamma,0}).$$

Questa può essere sviluppata, ottenendo

$$\begin{aligned}
&\mathbf{Y}^\intercal\mathbf{Y} + \mathbf{B}_{\gamma,0}^\intercal\mathbf{H}_\gamma^{-1}\mathbf{B}_{\gamma,0} + \mathbf{B}_\gamma^\intercal\mathbf{X}^\intercal\mathbf{X}\mathbf{B}_\gamma + \mathbf{B}_\gamma^\intercal\mathbf{H}_\gamma^{-1}\mathbf{B}_\gamma - 2\mathbf{B}_\gamma^\intercal\mathbf{X}^\intercal\mathbf{Y} - 2\mathbf{B}_\gamma^\intercal\mathbf{H}_\gamma^{-1}\mathbf{B}_{\gamma,0} \\
&\qquad = \mathbf{C} + \mathbf{B}_\gamma^\intercal\mathbf{K}_\gamma\mathbf{B}_\gamma - 2\mathbf{M}^\intercal\mathbf{B}_\gamma. \qquad \text{(A.5)}
\end{aligned}$$

Aggiungendo e sottraendo poi la quantità $\mathbf{M}^\intercal\mathbf{K}_\gamma^{-1}\mathbf{M}$ in (A.5), si ottiene

$$\begin{aligned}
&\mathbf{C} + \mathbf{B}_\gamma^\intercal\mathbf{K}_\gamma\mathbf{B}_\gamma - 2\mathbf{M}^\intercal\mathbf{B}_\gamma + \mathbf{M}^\intercal\mathbf{K}_\gamma^{-1}\mathbf{M} - \mathbf{M}^\intercal\mathbf{K}_\gamma^{-1}\mathbf{M} \\
&\qquad = \mathbf{C} + (\mathbf{B}_\gamma-\mathbf{K}_\gamma^{-1}\mathbf{M})^\intercal\mathbf{K}_\gamma(\mathbf{B}_\gamma-\mathbf{K}_\gamma^{-1}\mathbf{M}) - \mathbf{M}^\intercal\mathbf{K}_\gamma^{-1}\mathbf{M}. \qquad \text{(A.6)}
\end{aligned}$$

Analizzando l'equazione (A.4), si nota che la distribuzione a posteriori congiunta di $(\mathbf{B}_\gamma, \boldsymbol{\Sigma})$ è proporzionale a

$$\pi(\mathbf{B}_\gamma, \boldsymbol{\Sigma}|\mathbf{Y}, \mathbf{X}, \boldsymbol{\gamma}) = \pi(\mathbf{B}_\gamma|\boldsymbol{\Sigma}, \mathbf{Y}, \mathbf{X}, \boldsymbol{\gamma})\pi(\boldsymbol{\Sigma}|\mathbf{Y}, \mathbf{X}, \boldsymbol{\gamma}), \qquad \text{(A.7)}$$

dove

$$\pi(\mathbf{B}_\gamma|\boldsymbol{\Sigma},\mathbf{Y},\mathbf{X},\boldsymbol{\gamma}) = \phi_{p_\gamma\times q}(\mathbf{B}_\gamma|\widetilde{\mathbf{B}}_\gamma,\boldsymbol{\Sigma},\widetilde{\boldsymbol{\Sigma}}_\gamma) \tag{A.8}$$

$$\pi(\boldsymbol{\Sigma}|\mathbf{Y},\mathbf{X},\boldsymbol{\gamma}) = \varphi_{\mathsf{IW}}(c_0+n,\mathcal{Q}_\gamma), \tag{A.9}$$

dove $\widetilde{\mathbf{B}}_\gamma = \mathbf{K}_\gamma^{-1}\mathbf{M},\ \mathbf{K}_\gamma = \left(\mathbf{X}^\intercal\mathbf{X}+\mathbf{H}_\gamma^{-1}\right),\ \widetilde{\boldsymbol{\Sigma}}_\gamma = \mathbf{K}_\gamma^{-1}$ e $\mathcal{Q}_\gamma$ sono definiti come in precedenza. Considerando poi il corollario A.4, si ha la distribuzione a posteriori marginale di $\pi(\mathbf{B}_\gamma|\mathbf{Y},\mathbf{X},\boldsymbol{\gamma})$ dove $\boldsymbol{\Sigma}$ viene integrata

$$\pi(\mathbf{B}_\gamma|\mathbf{Y},\mathbf{X},\boldsymbol{\gamma}) = \varphi_{\mathsf{T}p_\gamma,q}(\mathbf{B}_\gamma|\widetilde{\mathbf{B}}_\gamma,\mathcal{Q}_\gamma,\widetilde{\boldsymbol{\Sigma}}_\gamma^{-1},c_0+n). \tag{A.10}$$

La distribuzione a posteriori marginale per $\boldsymbol{\gamma}$ si ottiene integrando $(\mathbf{B}_\gamma,\boldsymbol{\Sigma})$ dalla distribuzione a posteriori non normalizzata in (A.4). La quantità ottenuta poi verrà moltiplicata per la priori di $\boldsymbol{\gamma}$, che si traduce nel moltiplicare l'equazione (A.4) per la seguente quantità

$$\frac{\mathcal{C}_{\mathsf{IW}}(c_0+n,\mathcal{Q}_\gamma,q)\mathcal{C}_{\mathsf{gauss}}(\boldsymbol{\Sigma},\widetilde{\boldsymbol{\Sigma}}_\gamma,q,p_\gamma)}{\mathcal{C}_{\mathsf{IW}}(c_0+n,\mathcal{Q}_\gamma,q)\mathcal{C}_{\mathsf{gauss}}(\boldsymbol{\Sigma},\widetilde{\boldsymbol{\Sigma}}_\gamma,q,p_\gamma)}\ \pi(\boldsymbol{\gamma}). \tag{A.11}$$

Si ottiene quindi

$$\begin{aligned}\pi(\mathbf{Y},\mathbf{B}_\gamma,\boldsymbol{\Sigma}|\mathbf{X},\boldsymbol{\gamma}) = \exp\left\{-\frac{1}{2}\mathsf{tr}\left[\boldsymbol{\Sigma}^{-1}\mathcal{Q}_1(\mathbf{B}_\gamma)\right]\right\}\exp\left\{-\frac{1}{2}\mathsf{tr}\left[\boldsymbol{\Sigma}^{-1}\big(\mathbf{C}_0+\mathcal{Q}_2(\mathbf{B}_{\gamma,0})\big)\right]\right\}\\ \times\frac{|\boldsymbol{\Sigma}|^{-(c_0+q+1)/2}}{\mathcal{C}_{\mathsf{IW}}(c_0,\mathbf{C}_0,q)\mathcal{C}_{\mathsf{gauss}}(\boldsymbol{\Sigma},\mathbf{H}_\gamma,q,p_\gamma)\mathcal{C}_{\mathsf{gauss}}(\boldsymbol{\Sigma},\mathbf{I}_n,n,q)}\\ \times\frac{\mathcal{C}_{\mathsf{IW}}(c_0+n,\mathcal{Q}\gamma,q)\mathcal{C}_{\mathsf{gauss}}(\boldsymbol{\Sigma},\widetilde{\boldsymbol{\Sigma}}_\gamma,q,p_\gamma)}{\mathcal{C}_{\mathsf{IW}}(c_0+n,\mathcal{Q}_\gamma,q)\mathcal{C}_{\mathsf{gauss}}(\boldsymbol{\Sigma},\widetilde{\boldsymbol{\Sigma}}_\gamma,q,p_\gamma)}\ \pi(\boldsymbol{\gamma}),\end{aligned} \tag{A.12}$$

dove $\mathcal{Q}_1(\mathbf{B}_\gamma) = \left(\mathbf{B}_\gamma-\mathbf{K}_\gamma^{-1}\mathbf{M}\right)^\intercal\mathbf{K}_\gamma\left(\mathbf{B}_\gamma-\mathbf{K}_\gamma^{-1}\mathbf{M}\right)$ e $\mathcal{Q}_2(\mathbf{B}_{\gamma,0}) = \mathbf{C}+\mathbf{M}^\intercal\mathbf{K}_\gamma^{-1}\mathbf{M}$.

Integrando poi rispetto a $(\mathbf{B}_\gamma, \boldsymbol{\Sigma})$, si ottiene

$$\begin{aligned}
\pi(\boldsymbol{\gamma}|\mathbf{Y},\mathbf{X}) &= \frac{|\boldsymbol{\Sigma}|^{-(c_0+q+1)/2}}{\mathcal{C}_{\mathsf{IW}}(c_0,\mathbf{C}_0,q)\mathcal{C}_{\mathsf{gauss}}(\boldsymbol{\Sigma},\mathbf{H}_\gamma,q,p_\gamma)\mathcal{C}_{\mathsf{gauss}}(\boldsymbol{\Sigma},\mathbf{I}_n,n,q)} \\
&\quad \times \frac{\mathcal{C}_{\mathsf{IW}}(c_0+n,\mathcal{Q}\gamma,q)\mathcal{C}_{\mathsf{gauss}}(\boldsymbol{\Sigma},\widetilde{\boldsymbol{\Sigma}}_\gamma,q,p_\gamma)}{\mathcal{C}_{\mathsf{IW}}(c_0+n,\mathcal{Q}_\gamma,q)\mathcal{C}_{\mathsf{gauss}}(\boldsymbol{\Sigma},\widetilde{\boldsymbol{\Sigma}}_\gamma,q,p_\gamma)}\pi(\boldsymbol{\gamma}) \\
&= \frac{|\boldsymbol{\Sigma}|^{-(c_0+q+1)/2}}{2^{c_0q/2}\pi^{q(q-1)/4}|C_0|^{-c_0/2}\Gamma_q(c_0)}\frac{1}{(2\pi)^{nq/2}|\boldsymbol{\Sigma}|^{n/2}} \\
&\quad \times \frac{1}{(2\pi)^{p_\gamma q/2}|\boldsymbol{\Sigma}|^{p_\gamma/2}|\mathbf{H}_\gamma|^{q/2}}\frac{|\boldsymbol{\Sigma}|^{p_\gamma/2}|\widetilde{\boldsymbol{\Sigma}}|^{q/2}}{|\boldsymbol{\Sigma}|^{p_\gamma/2}|\widetilde{\boldsymbol{\Sigma}}|^{q/2}}\frac{(2\pi)^{p_\gamma q/2}}{(2\pi)^{p_\gamma q/2}} \\
&\quad \times \frac{2^{q(c_0+n)/2}\pi^{q(q-1)/4}|\mathcal{Q}_\gamma|^{-(c_0+n)/2}\Gamma_q(c_0+n)}{2^{q(c_0+n)/2}\pi^{q(q-1)/4}|\mathcal{Q}_\gamma|^{-(c_0+n)/2}\Gamma_q(c_0+n)}\pi(\boldsymbol{\gamma}) \\
&= \frac{|\widetilde{\boldsymbol{\Sigma}}_\gamma|^{q/2}}{|\mathbf{H}_\gamma|^{q/2}}\frac{\pi^{-nq/2}\ \Gamma_q(c_0+n)}{\Gamma_q(c_0)\ |\mathbf{C}_0|^{-c_0/2}}|\mathcal{Q}_\gamma|^{-(c_0+n)/2}\ \pi(\boldsymbol{\gamma}). \qquad \text{(A.13)}
\end{aligned}$$

A.2 Dimostrazione della distribuzione predittiva

La distribuzione predittiva del modello di regressione multivariata definita in Equazione (1.3) viene definita come

$$\begin{aligned}
\pi\big(\mathbf{Y}_0|\mathbf{Y},\mathbf{X},\mathbf{X}_0\big) &\propto \int\int \phi_{r\times q}(\mathbf{Y}_0|\mathbf{X}_0\mathbf{B},\boldsymbol{\Sigma},\mathbf{I}_r)\pi\big(\mathbf{B}_\gamma,\boldsymbol{\Sigma}|\mathbf{Y},\mathbf{X}\big)d\mathbf{B}_{\boldsymbol{\gamma}}d\boldsymbol{\Sigma} \\
&\propto \int\int \mathcal{L}(\mathbf{Y}_0|\mathbf{X}_0,\boldsymbol{\gamma},\mathbf{B}_\gamma,\boldsymbol{\Sigma})\pi\big(\mathbf{B}_\gamma|\boldsymbol{\Sigma},\mathbf{Y},\mathbf{X}\big)\pi\big(\boldsymbol{\Sigma}|\mathbf{Y},\mathbf{X}\big)d\mathbf{B}d\boldsymbol{\Sigma}
\end{aligned} \qquad \text{(A.14)}$$

dove $\pi\big(\mathbf{B}_\gamma,\boldsymbol{\Sigma}|\mathbf{Y},\mathbf{X}\big) = \pi\big(\mathbf{B}_\gamma|\boldsymbol{\Sigma},\mathbf{Y},\mathbf{X}\big)\pi\big(\boldsymbol{\Sigma}|\mathbf{Y},\mathbf{X}\big)$ è la distribuzione a posteriori congiunta definita in Equazione (A.16), con le rispettive distribuzioni a posteriori definite come in (1.20) e (1.21) rispettivamente. Si ha quindi che

$$\mathcal{L}(\mathbf{Y}_0|\mathbf{X}_0,\boldsymbol{\gamma},\mathbf{B}_\gamma,\boldsymbol{\Sigma}) = \frac{\exp\left\{-\frac{1}{2}\mathsf{tr}\left[\boldsymbol{\Sigma}^{-1}(\mathbf{Y}_0-\mathbf{X}_0\mathbf{B}_\gamma)^\intercal(\mathbf{Y}_0-\mathbf{X}_0\mathbf{B}_\gamma)\right]\right\}}{\mathcal{C}_{\mathsf{gauss}}(\boldsymbol{\Sigma},\mathbf{I}_r,r,q)}, \qquad \text{(A.15)}$$

dove $\mathcal{C}_{\mathsf{gauss}}(\boldsymbol{\Sigma},\mathbf{I}_r,r,q) = (2\pi)^{rq/2}\ |\boldsymbol{\Sigma}|^{r/2}$ è una costante di normalizzazione e

$$\pi(\mathbf{B}_\gamma|\boldsymbol{\Sigma},\mathbf{Y},\mathbf{X},\boldsymbol{\gamma}) = \frac{\exp\left\{-\frac{1}{2}\mathsf{tr}\left[\boldsymbol{\Sigma}^{-1}(\mathbf{B}_\gamma-\widetilde{\mathbf{B}}_\gamma)^\intercal\widetilde{\boldsymbol{\Sigma}}_\gamma(\mathbf{B}_\gamma-\widetilde{\mathbf{B}}_\gamma)\right]\right\}}{\mathcal{C}_{\mathsf{gauss}}(\boldsymbol{\Sigma},\widetilde{\boldsymbol{\Sigma}}_\gamma,q,p_\gamma)} \qquad \text{(A.16)}$$

$$\pi(\boldsymbol{\Sigma}|\mathbf{Y},\mathbf{X}) = \frac{|\boldsymbol{\Sigma}|^{-(c_0+n+q+1)/2}\exp\{-\frac{1}{2}\mathsf{tr}(\boldsymbol{\Sigma}^{-1}\mathcal{Q}_\gamma)\}}{\mathcal{C}_{\mathsf{IW}}(c_0+n,\mathcal{Q}_\gamma,q)}, \qquad \text{(A.17)}$$

dove

$$\mathcal{C}_{\mathsf{gauss}}(\boldsymbol{\Sigma}, \widetilde{\boldsymbol{\Sigma}}_\gamma, q, p_\gamma) = (2\pi)^{p_\gamma q/2}|\boldsymbol{\Sigma}|^{p_\gamma/2}|\widetilde{\boldsymbol{\Sigma}}_\gamma|^{q/2},$$
$$\mathcal{C}_{\mathsf{IW}}(c_0+n, \mathcal{Q}_\gamma, q) = 2^{(c_0+n)q/2}\pi^{(q(q-1))/4}|\mathcal{Q}_\gamma|^{-(c_0+n)/2}\prod_{j=1}^{q}\Gamma\big(\frac{c_0+n+1-j}{2}\big).$$

Per prima cosa si vuole integrare $\mathbf{B}_\gamma$. Ciò può essere fatto notando che la sua distribuzione a posteriori aggiornata dopo aver osservato il nuovo insieme di osservazioni $\big(\mathbf{Y}_0, \mathbf{X}_0\big)$ diventa

$$\begin{aligned}
\pi\big(\mathbf{B}_\gamma|\mathbf{Y}_0, \mathbf{X}_0, \mathbf{Y}, \mathbf{X}, \boldsymbol{\Sigma}\big) &\propto \phi_{r\times q}(\mathbf{Y}_0|\mathbf{X}_0\mathbf{B}, \boldsymbol{\Sigma}, \mathbf{I}_r)\phi_{p_\gamma\times q}(\mathbf{B}_\gamma|\widetilde{\mathbf{B}}_\gamma, \boldsymbol{\Sigma}, \widetilde{\boldsymbol{\Sigma}}_\gamma) \\
&\propto |\boldsymbol{\Sigma}|^{-r/2}\exp\left\{-\frac{1}{2}\mathsf{tr}\left[\boldsymbol{\Sigma}^{-1}(\mathbf{Y}_0-\mathbf{X}_0\mathbf{B}_\gamma)^\intercal(\mathbf{Y}_0-\mathbf{X}_0\mathbf{B}_\gamma)\right]\right\} \\
&\times |\boldsymbol{\Sigma}|^{-p_\gamma/2}|\widetilde{\boldsymbol{\Sigma}}_\gamma|^{-q/2}\exp\left\{-\frac{1}{2}\mathsf{tr}\left[\boldsymbol{\Sigma}^{-1}(\mathbf{B}_\gamma-\widetilde{\mathbf{B}}_\gamma)^\intercal\widetilde{\boldsymbol{\Sigma}}_\gamma(\mathbf{B}_\gamma-\widetilde{\mathbf{B}}_\gamma)\right]\right\} \\
&\propto \phi_{p_\gamma\times q}\big(\widehat{\mathbf{B}}_0, \boldsymbol{\Sigma}, \widehat{\boldsymbol{\Sigma}}_0\big), \qquad \text{(A.18)}
\end{aligned}$$

con

$$\begin{aligned}
\widehat{\mathbf{B}}_0 &= \widehat{\boldsymbol{\Sigma}}_0\big(\mathbf{X}^\intercal\mathbf{Y} + \mathbf{X}_0^\intercal\mathbf{Y}_0\big) && \text{(A.19)}\\
\widehat{\boldsymbol{\Sigma}}_0 &= \big(\widetilde{\boldsymbol{\Sigma}}^{-1} + \mathbf{X}_0^\intercal\mathbf{X}_0\big)^{-1} && \text{(A.20)}\\
&= \widetilde{\boldsymbol{\Sigma}} - \widetilde{\boldsymbol{\Sigma}}\mathbf{X}_0^\intercal\big(\mathbf{I}_r + \mathbf{X}_0\widetilde{\boldsymbol{\Sigma}}\mathbf{X}_0^\intercal\big)^{-1}\mathbf{X}_0\widetilde{\boldsymbol{\Sigma}} && \text{(A.21)}\\
&= \widetilde{\boldsymbol{\Sigma}} - \widetilde{\boldsymbol{\Sigma}}\mathbf{X}_0^\intercal\mathbf{F}_0^{-1}\mathbf{X}_0\widetilde{\boldsymbol{\Sigma}} && \text{(A.22)}\\
&= \big(\mathbf{I}_p - \mathbf{K}_0\mathbf{X}_0\big)\widetilde{\boldsymbol{\Sigma}} && \text{(A.23)}\\
\mathbf{K}_0 &= \widetilde{\boldsymbol{\Sigma}}\mathbf{X}_0^\intercal\mathbf{F}_0^{-1} && \text{(A.24)}\\
\mathbf{F}_0 &= \mathbf{I}_r + \mathbf{X}_0\widetilde{\boldsymbol{\Sigma}}\mathbf{X}_0^\intercal. && \text{(A.25)}
\end{aligned}$$

Inoltre, per $\widehat{\mathbf{B}}_0$ si ha che

$$\begin{aligned}
\widehat{\mathbf{B}}_0 &= \widehat{\boldsymbol{\Sigma}}_0\big(\mathbf{X}^\intercal\mathbf{Y} + \mathbf{X}_0^\intercal\mathbf{Y}_0\big) && \text{(A.26)}\\
&= \big(\mathbf{I}_p - \mathbf{K}_0\mathbf{X}_0\big)\widetilde{\boldsymbol{\Sigma}}\big(\mathbf{X}^\intercal\mathbf{Y} + \mathbf{X}_0^\intercal\mathbf{Y}_0\big) && \text{(A.27)}\\
&= \widetilde{\boldsymbol{\Sigma}}\mathbf{X}^\intercal\mathbf{y} - \mathbf{K}_0\mathbf{X}_0\widetilde{\boldsymbol{\Sigma}}\mathbf{X}^\intercal\mathbf{Y} + \widetilde{\boldsymbol{\Sigma}}\mathbf{X}_0^\intercal\mathbf{Y}_0 - \mathbf{K}_0\mathbf{X}_0\widetilde{\boldsymbol{\Sigma}}\mathbf{X}_0^\intercal\mathbf{Y}_0 && \text{(A.28)}\\
&= \widetilde{\boldsymbol{\Sigma}}\mathbf{X}^\intercal\mathbf{Y} - \mathbf{K}_0\mathbf{X}_0\widetilde{\boldsymbol{\Sigma}}\mathbf{X}^\intercal\mathbf{Y} + \mathbf{K}_0\mathbf{Y}_0 && \text{(A.29)}\\
&= \widetilde{\mathbf{B}}_\gamma + \mathbf{K}_0(\mathbf{Y}_0 - \mathbf{X}_0\widetilde{\mathbf{B}}_\gamma), && \text{(A.30)}
\end{aligned}$$

e $\widetilde{\mathbf{B}} = \widetilde{\boldsymbol{\Sigma}}\mathbf{X}^\intercal\mathbf{Y}$, dove l'ultimo risultato si ottiene utilizzando le equazioni (A.24)-(A.25).

Infatti, dall'equazione (A.25) si ha $\mathbf{X}_0\widetilde{\mathbf{\Sigma}}\mathbf{X}_0^\intercal = \mathbf{F}_0 - \mathbf{I}_r$ e quindi

$$-\mathbf{K}_0\mathbf{X}_0\widetilde{\mathbf{\Sigma}}\mathbf{X}_0^\intercal\mathbf{Y}_0 = \mathbf{K}_0\mathbf{Y}_0 - \mathbf{K}_0\mathbf{F}_0\mathbf{Y}_0 = \mathbf{K}_0\mathbf{Y}_0 - \widetilde{\mathbf{\Sigma}}\mathbf{X}_0^\intercal\mathbf{F}_0^{-1}\mathbf{F}_0\mathbf{Y}_0, \tag{A.31}$$

applicando l'equazione (A.25). La distribuzione di $\mathbf{B}_\gamma$ diventa quindi

$$\begin{aligned}\pi\big(\mathbf{B}_\gamma|\mathbf{\Sigma},\mathbf{Y},\mathbf{X},\mathbf{Y}_0,\mathbf{X}_0\big) \propto \exp\left\{-\frac{1}{2}\mathsf{tr}\left[\mathbf{\Sigma}^{-1}\left(\left(\mathbf{B}_\gamma-\widehat{\mathbf{B}}_0\right)^\intercal\widehat{\mathbf{\Sigma}}_0^{-1}\left(\mathbf{B}_\gamma-\widehat{\mathbf{B}}_0\right)\right)\right]\right\}\\ \times\exp\left\{-\frac{1}{2}\mathsf{tr}\left[\mathbf{\Sigma}^{-1}\widehat{\mathbf{Q}}_\gamma\right]\right\}|\mathbf{\Sigma}|^{-(r+p_\gamma)/2}|\widetilde{\mathbf{\Sigma}}_\gamma|^{-q/2},\end{aligned}$$

con $\widehat{\mathbf{Q}}_\gamma = \mathbf{Y}_0^\intercal\mathbf{Y}_0 - \widehat{\mathbf{B}}_0^\intercal\widehat{\mathbf{\Sigma}}_0^{-1}\widehat{\mathbf{B}}_0 - (\mathbf{X}_0^\intercal\mathbf{Y}_0 - \mathbf{X}^\intercal\mathbf{Y})^\intercal\widehat{\mathbf{\Sigma}}_0(\mathbf{X}_0^\intercal\mathbf{Y}_0 - \mathbf{X}^\intercal\mathbf{Y})$.

Integrando $\mathbf{B}$ dall'espressione precedente si ottiene

$$\begin{aligned}\pi\big(\mathbf{Y},\mathbf{X},\mathbf{Y}_0,\mathbf{X}_0|\mathbf{\Sigma}\big) &\propto |\widehat{\mathbf{\Sigma}}_0|^{p_\gamma/2}|\widetilde{\mathbf{\Sigma}}_\gamma|^{-q/2}|\mathbf{\Sigma}|^{-r/2}\\ &\quad\times\exp\left\{-\frac{1}{2}\mathsf{tr}\left[\mathbf{\Sigma}^{-1}\left(\mathbf{Y}_0^\intercal\mathbf{Y}_0+\mathbf{Y}^\intercal\mathbf{X}\widetilde{\mathbf{\Sigma}}_\gamma\mathbf{X}^\intercal\mathbf{Y}+\widehat{\mathbf{B}}_0^\intercal\widehat{\mathbf{\Sigma}}_0^{-1}\widehat{\mathbf{B}}_0\right)\right]\right\}\\ &\propto |\widehat{\mathbf{\Sigma}}_0|^{p_\gamma/2}|\widetilde{\mathbf{\Sigma}}_\gamma|^{-q/2}|\mathbf{\Sigma}|^{-r/2}\\ &\quad\times\exp\left\{-\frac{1}{2}\mathsf{tr}\left[\mathbf{\Sigma}^{-1}\left((\mathbf{Y}_0-\mathbf{X}_0\widehat{\mathbf{B}}_0)^\intercal\mathbf{F}_0^{-1}(\mathbf{Y}_0-\mathbf{X}_0\widehat{\mathbf{B}}_0)\right)\right]\right\}\end{aligned}$$

Per capire il risultato precedente si noti che

$$\begin{aligned}-\mathbf{Y}_0^\intercal\mathbf{Y}_0+\mathbf{Y}_0^\intercal\mathbf{X}_0\widehat{\mathbf{\Sigma}}_0\mathbf{X}_0^\intercal\mathbf{Y}_0 &= -\mathbf{Y}_0^\intercal\big(\mathbf{I}_r-\mathbf{X}_0\widehat{\mathbf{\Sigma}}_0\mathbf{X}_0^\intercal\big)\mathbf{Y}_0\\ &= -\mathbf{Y}_0^\intercal\big(\mathbf{I}_r-\mathbf{X}_0\big(\widetilde{\mathbf{\Sigma}}_\gamma^{-1}+\mathbf{X}_0^\intercal\mathbf{X}_0\big)^{-1}\mathbf{X}_0^\intercal\big)\mathbf{Y}_0\\ &= -\mathbf{Y}_0^\intercal\big(\mathbf{I}_r+\mathbf{X}_0\widetilde{\mathbf{\Sigma}}_\gamma\mathbf{X}_0^\intercal\big)^{-1}\mathbf{Y}_0\\ &= -\mathbf{Y}_0^\intercal\mathbf{F}_0^{-1}\mathbf{Y}_0,\end{aligned} \tag{A.32}$$

e

$$\begin{aligned}\mathbf{Y}^\intercal\mathbf{X}\widehat{\mathbf{\Sigma}}_0\mathbf{X}^\intercal\mathbf{y}-\mathbf{Y}^\intercal\mathbf{X}\widetilde{\mathbf{\Sigma}}_\gamma\mathbf{X}^\intercal\mathbf{y} &= -\mathbf{y}^\intercal\mathbf{X}\widetilde{\mathbf{\Sigma}}_\gamma\mathbf{X}^\intercal\mathbf{y}+\mathbf{y}^\intercal\mathbf{X}\widetilde{\mathbf{\Sigma}}_\gamma\mathbf{X}^\intercal\mathbf{y}\\ &\quad-\mathbf{y}^\intercal\mathbf{X}\widetilde{\mathbf{\Sigma}}_\gamma\mathbf{X}_0^\intercal\mathbf{F}_0^{-1}\mathbf{X}_0\widetilde{\mathbf{\Sigma}}_\gamma\mathbf{X}^\intercal\mathbf{y}\\ &= -\mathbf{y}^\intercal\mathbf{X}\widetilde{\mathbf{\Sigma}}_\gamma\mathbf{X}_0^\intercal\mathbf{F}_0^{-1}\mathbf{X}_0\widetilde{\mathbf{\Sigma}}_\gamma\mathbf{X}^\intercal\mathbf{y}\\ &= -\widehat{\mathbf{B}}_0^\intercal\mathbf{X}_0^\intercal\mathbf{F}_0^{-1}\mathbf{X}_0\widehat{\mathbf{B}}_0,\end{aligned} \tag{A.33}$$

e che

$$\begin{aligned}\mathbf{Y}^\intercal\mathbf{X}\widehat{\boldsymbol{\Sigma}}_0\mathbf{X}_0^\intercal\mathbf{Y}_0 &= \mathbf{Y}^\intercal\mathbf{X}\widetilde{\boldsymbol{\Sigma}}_\gamma\mathbf{X}_0^\intercal\mathbf{Y}_0 - \mathbf{Y}^\intercal\mathbf{X}\widetilde{\boldsymbol{\Sigma}}_\gamma\mathbf{X}_0^\intercal\mathbf{F}_0^{-1}\mathbf{X}_0\widetilde{\boldsymbol{\Sigma}}_\gamma\mathbf{X}_0^\intercal\mathbf{Y}_0 \\ &= \widehat{\mathbf{B}}_0^\intercal\mathbf{X}_0^\intercal\mathbf{y}_0 - \widehat{\mathbf{B}}_0^\intercal\mathbf{X}_0^\intercal\mathbf{F}_0^{-1}\mathbf{X}_0\widetilde{\boldsymbol{\Sigma}}_\gamma\mathbf{X}_0^\intercal\mathbf{Y}_0 \\ &= \widehat{\mathbf{B}}_0^\intercal\mathbf{X}_0^\intercal\mathbf{F}_0^{-1}\big(\mathbf{F}_0 - \mathbf{X}_0\widetilde{\boldsymbol{\Sigma}}_\gamma\mathbf{X}_0^\intercal\big)\mathbf{Y}_0 = \widehat{\mathbf{B}}_0^\intercal\mathbf{X}_0^\intercal\mathbf{F}_0^{-1}\mathbf{Y}_0. \end{aligned} \tag{A.34}$$

Si ha quindi che $\pi\big(\mathbf{Y}_0, \mathbf{X}_0|\mathbf{Y}, \mathbf{X}, \boldsymbol{\Sigma}\big) \propto \phi_{r\times q}(\mathbf{X}_0\widehat{\mathbf{B}}_0, \boldsymbol{\Sigma}, \mathbf{F}_0^{-1})$ e dato che $\pi(\boldsymbol{\Sigma}|\mathbf{Y}, \mathbf{X}, \boldsymbol{\gamma}) = \varphi_{\mathsf{IW}}(c_0 + n, \mathcal{Q}_\gamma)$ è possibile utilizzare il Teorema A.5, ottenendo

$$\begin{aligned}\pi\big(\mathbf{Y}_0|\mathbf{Y}, \mathbf{X}\big) &\propto \int \pi\big(\mathbf{Y}_0, \mathbf{X}_0|\mathbf{Y}, \mathbf{X}, \boldsymbol{\Sigma}\big)\pi\big(\boldsymbol{\Sigma}|\mathbf{Y}, \mathbf{X}\big)d\boldsymbol{\Sigma} \\ &\propto \varphi_{\mathsf{T}r\times q}\big(\mathbf{X}_0\widetilde{\mathbf{B}}_\gamma, \mathcal{Q}_\gamma, \mathbf{F}_0, c0 + n\big), \end{aligned} \tag{A.35}$$

che completa la dimostrazione.

A.3 Definizioni

Teorema A.1 (Teorema di Bayes)**.** *Sia A_i, con $i \in I \subseteq \mathbb{N}$, una partizione dello spazio S in eventi non trascurabili. Allora per ogni $E \subseteq S$ con $P(E) > 0$ si ha*

$$P(A_i) = \frac{P(A_i \cap E)}{P(E)} = \frac{P(E|A_i)P(A_i)}{\sum_{j\in I} P(E|A_j)P(A_j)}. \tag{A.36}$$

Definizione A.2 (Distribuzione t di Student matriciale)**.** Sia $\mathbf{Y} \in \mathbb{R}^{p\times q}$ una matrice stocastica di dimensioni $(p \times q)$. Allora $\mathbf{Y}$ avrà distribuzione t di Student matrice, ossia $\mathbf{Y} \sim \mathsf{MT}_{p\times q}(\mathbf{M}, \boldsymbol{\Xi}, \boldsymbol{\Psi}, \nu)$, dove $\mathbf{M} \in \mathbb{R}^{p\times q}$ è la matrice dei parametri di posizione e $\boldsymbol{\Psi} \in \mathbb{S}_{++}^p$, $\boldsymbol{\Xi} \in \mathbb{S}_{++}^q$ sono matrici di varianza e covarianza simmetriche e definite di dimensione p e q, rispettivamente, e $\nu > q + 1$ è il parametro relativo ai gradi di libertà se la funzione di densità di $\mathbf{Y}$ ha la forma seguente:

$$\begin{aligned} f_{\mathbf{Y}}(\mathbf{Y}|\mathbf{M}, \boldsymbol{\Psi}, \boldsymbol{\Xi}, \nu) &= \frac{|\boldsymbol{\Xi} + (\mathbf{Y} - \mathbf{M})^\intercal\boldsymbol{\Psi}(\mathbf{Y} - \mathbf{M})|^{-(\nu+p)/2}}{\mathcal{C}_T(\boldsymbol{\Psi}, \boldsymbol{\Xi}, p, q)}\mathbb{1}_{\mathbb{R}^{p\times q}}(\mathbf{Y}) \\ &= \frac{|\mathbf{I}_q + \boldsymbol{\Xi}^{-1}(\mathbf{Y} - \mathbf{M})^\intercal\boldsymbol{\Psi}(\mathbf{Y} - \mathbf{M})|^{-(\nu+p)/2}}{\mathcal{C}_T^\star(\boldsymbol{\Psi}, \boldsymbol{\Xi}, p, q)}\mathbb{1}_{\mathbb{R}^{p\times q}}(\mathbf{Y}), \end{aligned} \tag{A.37}$$

con

$$\mathcal{C}_T(\mathbf{\Psi}, \mathbf{\Xi}, \nu, p, q) = \pi^{pq/2}|\mathbf{\Psi}|^{-q/2}|\mathbf{\Xi}|^{-\nu/2}\prod_{j=1}^{q}\frac{\Gamma\left(\frac{\nu+1-j}{2}\right)}{\Gamma\left(\frac{\nu+p+1-j}{2}\right)}$$
$$\mathcal{C}_T^{\star}(\mathbf{\Psi}, \mathbf{\Xi}, \nu, p, q) = |\mathbf{\Xi}|^{-(\nu+p)/2}\mathcal{C}_T(\mathbf{\Psi}, \mathbf{\Xi}, \nu, p, q),$$

e $\Gamma(a) = \int_{\mathbb{R}^+} x^{a-1}e^{-x}dx$.

Definizione A.3 (Distribuzione Inverse-Wishart)**.** Sia $\mathbf{Y} \in \mathbb{S}{+}{+}^q$ una matrice stocastica simmetrica definita positiva. La matrice $\mathbf{Y}$ si dice avere una distribuzione Inverse-Wishart, indicata con $\mathbf{Y} \sim \mathsf{IW}q(\mathbf{S}, \nu)$, dove $\mathbf{S} \in \mathbb{S}_{++}^q$ è una matrice di varianza-covarianza definita positiva e $\nu > q + 1$ è il parametro di gradi di libertà, se la sua funzione di densità di probabilità ha la seguente forma:

$$f_{\mathbf{Y}}(\mathbf{Y}|\mathbf{S}, \nu) = \frac{|\mathbf{Y}|^{-(\nu+q+1)/2}\exp\{-\frac{1}{2}\mathsf{tr}(\mathbf{Y}^{-1}\mathbf{S})\}}{\mathcal{C}_{IW}(\mathbf{S}, \nu, q)}\mathbb{1}_{\mathbb{S}_{++}^q}(\mathbf{Y}), \tag{A.38}$$

con $\mathcal{C}_{IW}(\mathbf{S}, \nu, q) = 2^{\nu q/2}\pi^{(q(q-1))/4}|\mathbf{S}|^{-\nu/2}\prod_{j=1}^{q}\Gamma\left(\frac{\nu+1-j}{2}\right)$ e $\Gamma(a) = \int_{\mathbb{R}^+} x^{a-1}e^{-x}dx$.

Corollario A.4 (Full conditional della distribuzione matriciale t di Student)**.** *Sia*

$$\begin{aligned}\mathbf{Y}|\mathbf{\Sigma} &\sim \mathsf{Mn}_{p\times q}(\mathbf{M}, \mathbf{\Sigma}, \mathbf{\Psi}^{-1})\\ \mathbf{\Sigma} &\sim \mathsf{IW}_q(\mathbf{\Xi}, \nu),\end{aligned} \tag{A.39}$$

allora

$$\begin{aligned}\mathbf{Y} &\sim \mathsf{MT}_{p\times q}(\mathbf{M}, \mathbf{\Xi}, \mathbf{\Psi}, \nu)\\ \mathbf{\Sigma}|\mathbf{Y} &\sim \mathsf{IW}_q(\mathbf{\Xi} + (\mathbf{Y} - \mathbf{M})^{\intercal}\mathbf{\Psi}(\mathbf{Y} - \mathbf{M}), \nu + p),\end{aligned} \tag{A.40}$$

dove $\mathbf{M} \in \mathbb{R}^{p\times q}$ *è la matrice dei parametri di locazione e* $\mathbf{\Psi} \in \mathbb{S}_{++}^p$, $\mathbf{\Xi} \in \mathbb{S}_{++}^q$ *sono delle matrici di varianza-covarianza simmetriche e definite positive rispettivamente di dimensione* p *e* q*, mentre* $\nu > q + 1$ *indica i gradi di libertà.*

Teorema A.5 (Rappresentazione stocastica della distribuzione t di Student matriciale)**.** *Sia* $\mathbf{Y} \in \mathbb{R}^{p\times q}$ *una matrice stocastica con una distribuzione matriciale* t *di Student, ad esempio* $\mathbf{Y} \sim \mathsf{MT}_{p\times q}(\mathbf{M}, \mathbf{\Xi}, \mathbf{\Psi}, \nu)$*, dove* $\mathbf{M} \in \mathbb{R}^{p\times q}$ *è la matrice dei parametri di posizione e* $\mathbf{\Psi} \in \mathbb{S}_{++}^p$, $\mathbf{\Xi} \in \mathbb{S}_{++}^q$ *sono matrici di varianza e covarianza simmetriche e definite positive di dimensione* p *e* q*, rispettivamente, e* $\nu > q+1$ *sono i gradi di libertà. Allora*

$$p(\mathbf{Y}|\mathbf{M}, \mathbf{\Psi}, \mathbf{\Xi}, \nu) = \int_{\mathbb{S}_{++}^q} p(\mathbf{Y}|\mathbf{M}, \mathbf{\Sigma}, \mathbf{\Psi}^{-1})p(\mathbf{\Sigma}|\mathbf{\Xi}, \nu)d\mathbf{\Sigma}, \tag{A.41}$$

dove $p(\mathbf{Y}|\mathbf{M}, \mathbf{\Sigma}, \mathbf{\Psi}^{-1})$ *rappresenta la funzione densità di una distribuzione Gaussiana matriciale, definita nell'equazione* (**??**)*, con* $\mathbf{Y} \sim \mathsf{MN}_{p\times q}(\mathbf{M}, \mathbf{\Sigma}, \mathbf{\Psi}^{-1})$*, e* $p(\mathbf{\Sigma}|\mathbf{\Xi}, \nu)$

rappresenta la funzione densità di una distribuzione Inverse-Wishart, come definita nell'equazione (A.38)*, con* $\boldsymbol{\Sigma} \sim \mathsf{IW}_q(\boldsymbol{\Xi}, \nu)$.

Dimostrazione. Al fine di provare la Proposizione A.5, si noti che $(\mathbf{Y}, \boldsymbol{\Sigma})$ ha la seguente distribuzione a posteriori congiunta:

$$\begin{aligned} f_{\mathbf{Y},\boldsymbol{\Sigma}}(\mathbf{Y}, \boldsymbol{\Sigma}) &= \frac{\exp\left\{-\frac{1}{2}\mathsf{tr}\left[\boldsymbol{\Sigma}^{-1}(\mathbf{Y}-\mathbf{M})^\intercal \boldsymbol{\Psi}(\mathbf{Y}-\mathbf{M})\right]\right\}}{\mathcal{C}_N(\boldsymbol{\Psi}^{-1}, \boldsymbol{\Sigma}, p, q)} \\ &\quad \times \frac{|\boldsymbol{\Sigma}|^{-(\nu+q+1)/2}\exp\{-\frac{1}{2}\mathsf{tr}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Xi})\}}{\mathcal{C}_{IW}(\boldsymbol{\Xi}, \nu, q)} \mathbb{1}_{\mathbb{R}^{p\times q}}(\mathbf{Y}) \times \mathbb{1}_{\mathbb{S}^q_{++}}(\boldsymbol{\Sigma}) \\ &= \frac{\exp\left\{-\frac{1}{2}\mathsf{tr}\left[\boldsymbol{\Sigma}^{-1}\big(\boldsymbol{\Xi} + (\mathbf{Y}-\mathbf{M})^\intercal \boldsymbol{\Psi}(\mathbf{Y}-\mathbf{M})\big)\right]\right\}}{\mathcal{C}_N(\boldsymbol{\Psi}^{-1}, \boldsymbol{\Sigma}, p, q)\mathcal{C}_{IW}(\boldsymbol{\Xi}, \nu, q)} \\ &\quad \times |\boldsymbol{\Sigma}|^{p/2}|\boldsymbol{\Sigma}|^{-(\nu+p+q+1)/2} \mathbb{1}_{\mathbb{R}^{p\times q}}(\mathbf{Y}) \times \mathbb{1}_{\mathbb{S}^q_{++}}(\boldsymbol{\Sigma}). \end{aligned} \tag{A.42}$$

Analizzando la funzione densità nell'equazione (A.42) come una funzione di $\boldsymbol{\Sigma}$ dato $\mathbf{Y}$, riconosciamo che è proporzionale a una distribuzione Inverse-Wishart:

$$\begin{aligned} f_{\boldsymbol{\Sigma},\mathbf{Y}}(\boldsymbol{\Sigma}, \mathbf{Y}) &= \frac{|\boldsymbol{\Sigma}|^{p/2}\mathcal{C}_{IW}(\mathbf{S}, \nu+p, q)}{\mathcal{C}_N(\boldsymbol{\Psi}^{-1}, \boldsymbol{\Sigma}, p, q)\mathcal{C}_{IW}(\boldsymbol{\Xi}, \nu, q)} \\ &\quad \times \frac{|\boldsymbol{\Sigma}|^{-(v+p+q+1)/2}\exp\left\{-\frac{1}{2}\mathsf{tr}\left[\boldsymbol{\Sigma}^{-1}\mathbf{S}\right]\right\}}{\mathcal{C}_{IW}(\mathbf{S}, \nu+p, q)} \mathbb{1}_{\mathbb{R}^{p\times q}}(\mathbf{Y}) \times \mathbb{1}_{\mathbb{S}^q_{++}}(\boldsymbol{\Sigma}), \end{aligned} \tag{A.43}$$

con $\mathbf{S} = \boldsymbol{\Xi} + (\mathbf{Y}-\mathbf{M})^\intercal \boldsymbol{\Psi}(\mathbf{Y}-\mathbf{M})$. Integrando l'equazione (A.43) rispetto a Σ, si ottiene:

$$\begin{aligned} f_{\mathbf{Y}}(\mathbf{Y}) &= \frac{|\boldsymbol{\Sigma}|^{p/2}\mathcal{C}_{IW}(\mathbf{S}, \nu+p, q)}{\mathcal{C}_N(\boldsymbol{\Psi}^{-1}, \boldsymbol{\Sigma}, p, q)\mathcal{C}_{IW}(\boldsymbol{\Xi}, \nu, q)} \mathbb{1}_{\mathbb{R}^{p\times q}}(\mathbf{Y}) \\ &= \frac{|\boldsymbol{\Sigma}|^{p/2}2^{(\nu+p)q/2}\pi^{(q(q-1))/4}|\mathbf{S}|^{-(\nu+p)/2}\prod_{j=1}^q \Gamma\left(\frac{\nu+p+1-j}{2}\right)}{(2\pi)^{pq/2}|\boldsymbol{\Psi}|^{-q/2}|\boldsymbol{\Sigma}|^{p/2}2^{\nu q/2}\pi^{(q(q-1))/4}|\boldsymbol{\Xi}|^{-\nu/2}\prod_{j=1}^q \Gamma\left(\frac{\nu+1-j}{2}\right)} \mathbb{1}_{\mathbb{R}^{p\times q}}(\mathbf{Y}), \end{aligned} \tag{A.44}$$

e riordinando i termini si ha

$$\begin{aligned} f_{\mathbf{Y}}(\mathbf{Y}) &= \frac{|\mathbf{S}|^{-(\nu+p)/2}\prod_{j=1}^q \Gamma\left(\frac{\nu+p+1-j}{2}\right)}{\pi^{pq/2}|\boldsymbol{\Psi}|^{-q/2}|\boldsymbol{\Xi}|^{-\nu/2}\prod_{j=1}^q \Gamma\left(\frac{\nu+1-j}{2}\right)} \mathbb{1}_{\mathbb{R}^{p\times q}}(\mathbf{Y}) \\ &= \frac{|\boldsymbol{\Xi} + (\mathbf{Y}-\mathbf{M})^\intercal \boldsymbol{\Psi}(\mathbf{Y}-\mathbf{M})|^{-(\nu+p)/2}}{\mathcal{C}_T(\boldsymbol{\Psi}, \boldsymbol{\Xi}, \nu, p, q)} \mathbb{1}_{\mathbb{R}^{p\times q}}(\mathbf{Y}), \end{aligned} \tag{A.45}$$

con $\mathcal{C}_T(\boldsymbol{\Psi}, \boldsymbol{\Xi}, \nu, p, q) = \pi^{pq/2}|\boldsymbol{\Psi}|^{-q/2}|\boldsymbol{\Xi}|^{-\nu/2}\frac{\prod_{j=1}^q \Gamma\left(\frac{\nu+1-j}{2}\right)}{\prod_{j=1}^q \Gamma\left(\frac{\nu+p+1-j}{2}\right)}$, che completa la dimostrazione. □

A.4 Grafici

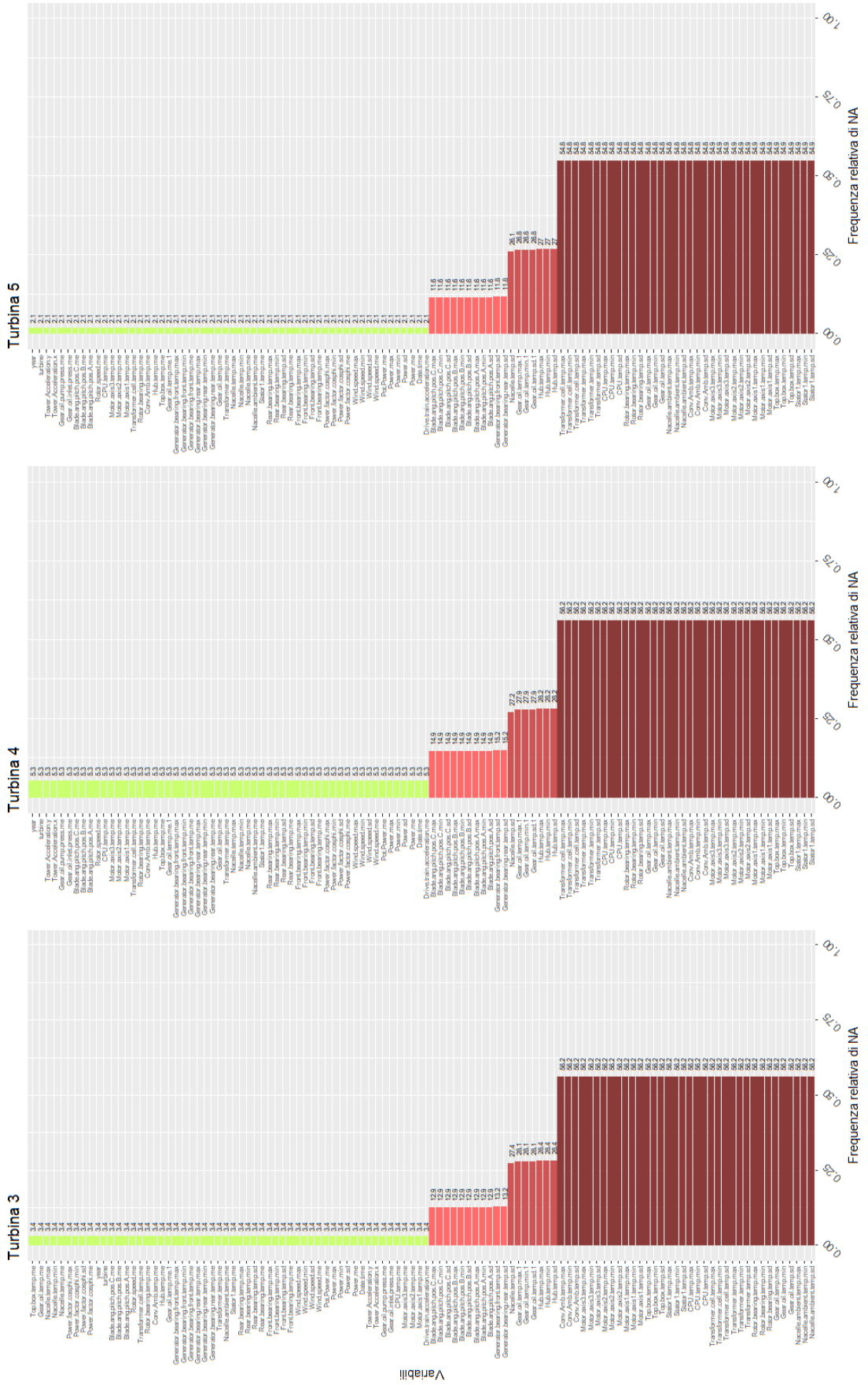


FIGURA A.1: Frequenza relativa dei valori mancanti in ciascuna turbina.

Bibliografia

Bernardi, M., Busatto, C. & Cattelan, M. (2023). Fast QR methods for statistical applications. [*Manoscritto non pubblicato*].

Bottolo, L., Banterle, M., Richardson, S., Ala-Korpela, M., Järvelin, M.-R. & Lewin, A. (2020). A computationally efficient bayesian seemingly unrelated regressions model for high-dimensional quantitative trait loci discovery. *Journal of the Royal Statistical Society. Series C, Applied statistics vol. 70,4* .

Brown, P. J., Vannucci, M. & Fearn, T. (1998a). Bayesian wavelength selection in multicomponent analysis. *Journal of Chemometrics* **12**, 173–182.

Brown, P. J., Vannucci, M. & Fearn, T. (1998b). Multivariate bayesian variable selection and prediction. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **60**, 627–641.

Delgado, I. & Fahim, M. (2020). Wind Turbine Data Analysis and LSTM-Based Prediction in SCADA System. *Energies* **14**, 1–21.

Effenberger, N. & Ludwig, N. (2022). A collection and categorization of open-source wind and wind power datasets. *Wind Energy* **25**, 1659–1683.

George, E. I. & McCulloch, R. E. (1997). Approaches for bayesian variable selection. *Statistica Sinica* **7**, 339–373.

Golub, G. H. & Van Loan, C. F. (1996). *Matrix Computations*. The Johns Hopkins University Press, 3rd ed.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82**, 711–732.

Hammarling, S. & Lucas, C. (2008). Updating the QR factorization and the least squares problem. *University of Manchester* .

Hastie, D. & Green, P. (2012). Model choice using reversible jump Markov chain Monte Carlo. *Statistica Neerlandica* **66**.

Narisetty, N. N. (2020). Bayesian model selection for high-dimensional data. In *Principles and Methods for Data Science*, A. S. Srinivasa Rao & C. Rao, eds., vol. 43 of *Handbook of Statistics*. Elsevier, pp. 207–248.

Pace, L., Salvan, A. & Sartori, N. (2022). Statistical Inference: Theory and Methods.

Richardson, S., Bottolo, L. & Rosenthal, J. S. (2011). Bayesian models for sparse regression analysis of high dimensional data. In *Bayesian statistics 9*. Oxford Univ. Press, Oxford, pp. 539–568.

Rothman, A. J., Levina, E. & Zhu, J. (2010). Sparse multivariate regression with covariance estimation. *Journal of Computational and Graphical Statistics* **19**, 947–962.

Tadesse, M. & Vannucci, M. (2021). *Handbook of Bayesian Variable Selection*. Chapman & Hall/CRC Handbooks of Modern Statistical Methods. CRC Press.