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

Accuratezza frequentista dell'inferenza bayesiana sulle componenti di varianza nei GLMM

Relatore Prof. Nicola Sartori
Dipartimento di Scienze Statistiche

Laureando Luca Riotto
Matricola 2086462

Anno Accademico 2023/2024

Indice

Introduzione

Nell'ambito della modellazione statistica un modo per trattare dati non indipendenti tra loro è attraverso i modelli lineari generalizzati a effetti misti (GLMM). In questa ampia classe di modelli l'assunzione di base è che condizionatamente a un vettore non osservato di effetti casuali, le osservazioni siano tra loro indipendenti, con valore atteso condizionato modellato attraverso un predittore lineare e varianza condizionata definita da una funzione di varianza. Gli effetti casuali sono tipicamente assunti distribuiti normalmente con valore atteso nullo e matrice di varianza dipendente da un insieme di parametri comunemente noti in questo contesto come componenti di varianza.

Tradizionalmente, la stima delle componenti di varianza è stata affrontata utilizzando metodi frequentisti, i quali forniscono stime puntuali e intervalli di confidenza basati su distribuzioni asintotiche dei corrispondenti stimatori o di opportune statistiche basate sulla verosimiglianza. Tuttavia, negli ultimi decenni, l'approccio di inferenza bayesiana ha guadagnato una crescente attenzione grazie alla sua flessibilità e capacità di incorporare informazioni a priori. In particolare, gli intervalli di credibilità bayesiani rappresentano un'alternativa agli intervalli di confidenza frequentisti, offrendo una visione probabilistica delle stime ottenute. In letteratura è molto comune l'utilizzo di approcci di inferenza bayesiana nei modelli lineari generalizzati a effetti misti (GLMM), specialmente nell'ultimo decennio grazie allo sviluppo di strumenti automatici per ottenere approssimazioni delle distribuzioni a posteriori dei parametri dei modelli.

Tuttavia, ad oggi non esiste uno studio che dimostri se e quanto gli intervalli di credibilità bayesiani possiedano buone proprietà frequentiste. L'obiettivo principale di questo lavoro è di valutare, attraverso un esteso studio di simulazione, l'accuratezza frequentista delle procedure inferenziali bayesiane e in particolare la copertura degli intervalli di credibilità bayesiani nel contesto delle stime delle componenti di varianza nei modelli a effetti casuali. La copertura, in questo contesto, è un concetto proprio dell'inferenza frequentista che si riferisce alla frequenza con cui gli intervalli di credibilità

contengono il vero valore delle componenti di varianza. Tale copertura sarà anche confrontata con quella di altri metodi frequentisti comunemente utilizzati in pratica.

Il Capitolo 1 introduce i modelli a effetti misti, e in particolare i GLMM. Nel Capitolo 2 viene fatta una rassegna dei metodi di stima frequentisti per le componenti di varianza nei GLMM. Nel Capitolo 3 vengono introdotte alcune tecniche di inferenza bayesiana e infine nel Capitolo 4 vengono presenetati i risultati dello studio di simulazione.

Capitolo 1

Modelli a effetti casuali

I dati gerarchici, noti anche come dati multilivello, rappresentano una vasta categoria di dati in cui le osservazioni sono organizzate su più livelli. In questo tipo di struttura, le unità di livello inferiore sono raggruppate all'interno di unità di livello superiore, formando così dei cluster di osservazioni. Un esempio classico è uno studio sulle competenze scolastiche dei bambini in una determinata area geografica, dove i dati possono essere organizzati su tre livelli distinti: i bambini all'interno delle classi, le classi all'interno delle scuole e le scuole all'interno dell'area geografica di riferimento.

I dati gerarchici rappresentano una generalizzazione di diverse categorie di dati, tra cui i dati longitudinali, noti nel contesto economico come dati panel. Questi consistono in misurazioni ripetute della stessa variabile su un insieme di individui, o più in generale unità, nel tempo. In tale contesto, il livello inferiore è costituito dalle osservazioni in momenti temporali differenti, mentre il livello superiore corrisponde agli individui. Uno studio longitudinale presenta in genere le seguenti caratteristiche (Wu, 2009):

- sono disponibili più misurazioni per ciascun individuo, con il numero e la tempistica delle misurazioni che possono variare tra gli individui, rendendo i dati spesso sbilanciati;
- le misurazioni ripetute per lo stesso individuo tendono a essere correlate, mentre quelle tra individui differenti sono generalmente considerate indipendenti;
- può esservi ampia variabilità sia tra le misurazioni ripetute per lo stesso individuo, sia tra quelle effettuate su individui diversi;
- i dati osservati sono spesso complessi o incompleti, con possibili mancanze, errori di misurazione, censure e *outlier*.

In questo tipo di dati, emergono due fonti principali di variabilità: la variabilità *within*, ossia quella tra le osservazioni dello stesso individuo, e la variabilità *between*, ovvero quella tra le osservazioni di individui differenti.

Un'altra forma di dati gerarchici, strettamente legata ai dati longitudinali, è costituita dai dati con misurazioni ripetute. In questo caso, vengono raccolte più osservazioni per ciascun individuo, senza che sia necessariamente previsto un intervallo temporale tra le misurazioni. Ad esempio, è possibile raccogliere dati relativi a variabili meteorologiche in diverse zone di una città nello stesso momento temporale.

Un elemento comune nei dati gerarchici è l'assunzione che le osservazioni all'interno di uno stesso cluster siano correlate, mentre quelle appartenenti a cluster differenti siano indipendenti. Questa assunzione giustifica l'uso di strumenti statistici adeguati alla modellazione di dati correlati, come i modelli a effetti misti. Nei modelli a effetti misti, viene introdotto un effetto casuale per ciascun individuo o cluster, al fine di indurre una correlazione tra le osservazioni appartenenti allo stesso gruppo; in questo modo, le osservazioni che condividono lo stesso effetto casuale risulteranno correlate tra loro.

1.1 Modello lineare normale a effetti misti

I Modelli Lineari a Effetti Misti (*Linear Mixed Effects Models*, LMM) costituiscono un'estensione dei modelli lineari classici, in cui il predittore lineare include, oltre agli effetti fissi (i parametri del modello), anche degli effetti casuali. Gli effetti fissi sono impiegati per descrivere la media della variabile risposta, mentre gli effetti casuali ne definiscono la struttura di varianza-covarianza. L'introduzione degli effetti casuali consente di semplificare la specificazione della matrice di covarianza della variabile risposta, riducendone la complessità in termini di parametri da stimare. Questi modelli trovano ampia applicazione nell'analisi di dati gerarchici, poiché permettono di modellare la correlazione tra le osservazioni appartenenti allo stesso gruppo.

1.1.1 Specificazione del modello

Sia $\mathbf{y}$ un vettore $n \times 1$ dimensionale, un modello lineare normale a effetti casuali può essere definito come segue

$$\begin{aligned} \mathbf{y} &= X\boldsymbol{\beta} + Z\boldsymbol{u} + \boldsymbol{\varepsilon}, \\ \boldsymbol{\varepsilon} &\sim N(0, R), \\ \boldsymbol{u} &\sim N(0, D(\boldsymbol{\theta})), \\ \boldsymbol{\varepsilon} &\perp\!\!\!\perp \boldsymbol{u}, \end{aligned} \tag{1.1}$$

con X matrice di regressori di dimensione $n \times p$, $\boldsymbol{\beta}$ vettore dei parametri di dimensione $p \times 1$, Z matrice del disegno per gli effetti casuali di dimensione $n \times q$, e $\boldsymbol{u}$ vettore degli effetti casuali di dimensione $q \times 1$.

Spesso si assume che $R = \sigma^2 I_n$, dove I_n è la matrice identità $n \times n$, ovvero, condizionatamente all'effetto casuale $\boldsymbol{u}$, le osservazioni siano indipendenti e con varianza costante. Condizionatamente all'effetto casuale $\boldsymbol{u}$ si ha

$$\begin{aligned} \mathbb{E}[\mathbf{y} \mid \boldsymbol{u}] &= X\boldsymbol{\beta} + Z\boldsymbol{u}, \\ \mathbb{V}(\mathbf{y} \mid \boldsymbol{u}) &= R, \end{aligned}$$

mentre marginalmente

$$\begin{aligned} \mathbb{E}[\mathbf{y}] &= X\boldsymbol{\beta}, \\ \mathbb{V}(\mathbf{y}) &= ZD(\boldsymbol{\theta})Z^\top + R. \end{aligned}$$

Si osservi che, gli effetti fissi modellano la media marginale della variabile risposta, mentre la componente casuale ne modella la varianza marginale. Inoltre, grazie all'assunzione di indipendenza tra l'effetto casuale $\mathbf{u}$ e il termine di errore $\boldsymbol{\varepsilon}$, la varianza marginale di $\mathbf{y}$ risulta dalla somma della variabilità indotta dalla componente casuale e quella del termine di errore.

A partire dal modello in (1.1) è possibile specificare un modello per dati gerarchici che risalta la presenza di gruppi nei dati. Una formulazione generale di un modello gerarchico, Datta & Lahiri (2000), è la seguente

$$\mathbf{y}_i = X_i\boldsymbol{\beta} + Z_i\boldsymbol{u}_i + \boldsymbol{\varepsilon}_i, \quad i = 1, \ldots, K, \tag{1.2}$$

dove $\mathbf{y}_i$ è il vettore di risposta di dimensione $n_i \times 1$ associato all'i-esimo individuo, X_i e Z_i sono rispettivamente le matrici del disegno degli effetti fissi e casuali, $\boldsymbol{\beta}$ è un vettore di parametri, $\boldsymbol{u}_i$ il vettore degli effetti casuali e $\boldsymbol{\varepsilon_i}$ il vettore degli errori. Si assume che $\boldsymbol{u}_i$ e $\boldsymbol{\varepsilon}_i$ siano indipendenti, e che $(\boldsymbol{u}_i, \boldsymbol{\varepsilon}_i)$ sia indipendente da $(\boldsymbol{u}_{i'}, \boldsymbol{\varepsilon}_{i'})$ con $i \neq i'$, $i, i' = 1, \ldots, K$ e con

$$\boldsymbol{u}_i \sim N(0, D_i(\boldsymbol{\theta})), \quad \boldsymbol{\varepsilon}_i \sim N(0, R_i), \tag{1.3}$$

Tipicamente la matrice di varianza della componente di errore del modello, $\boldsymbol{\varepsilon}$, è del tipo $R = diag(R_1, \ldots, R_K) = \sigma^2 I_n$ con $n = \sum_{i=1}^{K} n_i$, tuttavia è possibile specificare R al fine di permettere la presenza di correlazione tra errori delle osservazioni di uno stesso gruppo. Ad esempio con dati longitudinali si può assumere una struttura di correlazione autoregressiva con $\text{Cor}(y_{ij}, y_{ih}) = \rho^{|j-h|}$, oppure una struttura più generale come quella di Toeplitz con $\text{Cor}(y_{ij}, y_{ih}) = \rho_{|j-h|}$, dove ogni elemento lungo una diagonale parallela alla principale è governata da un solo parametro.

La matrice di covarianza $D(\boldsymbol{\theta}) = \text{diag}(D_1(\boldsymbol{\theta}), \ldots, D_K(\boldsymbol{\theta}))$ dipende dalla specificazione degli effetti casuali. È possibile distinguere tre semplici specificazioni per la componente casuale del modello ampiamente utilizzate nel contesto dei dati gerarchici, e di particolare interesse in questa tesi, quali:

- modello a intercetta casuale, con $Z_i = \mathbf{1}_{n_i}$ e $D(\boldsymbol{\theta}) = \text{diag}(D_1(\boldsymbol{\theta}), \ldots, D_k(\boldsymbol{\theta})) = \sigma_u^2 I_k$ dove $\mathbf{1}_{n_i} = [1 \ldots 1]^\top$ di dimensione $n_i \times 1$,
- modello a pendenza casuale, con $Z_i = x_i$ e $D(\boldsymbol{\theta}) = \text{diag}(D_1(\boldsymbol{\theta}), \ldots, D_k(\boldsymbol{\theta})) = \sigma_u^2 I_k$ dove x_i è una covariata per il modello,
- modello a intercetta e pendenza casuale, con $Z_i = [\mathbf{1}_{n_i} x_i]$ e $D_i(\boldsymbol{\theta}) = \begin{bmatrix} \sigma_1^2 & \rho \\ \rho & \sigma_2^2 \end{bmatrix}$.

Le tre diverse specificazioni sono illustrate graficamente in Figura 1.1

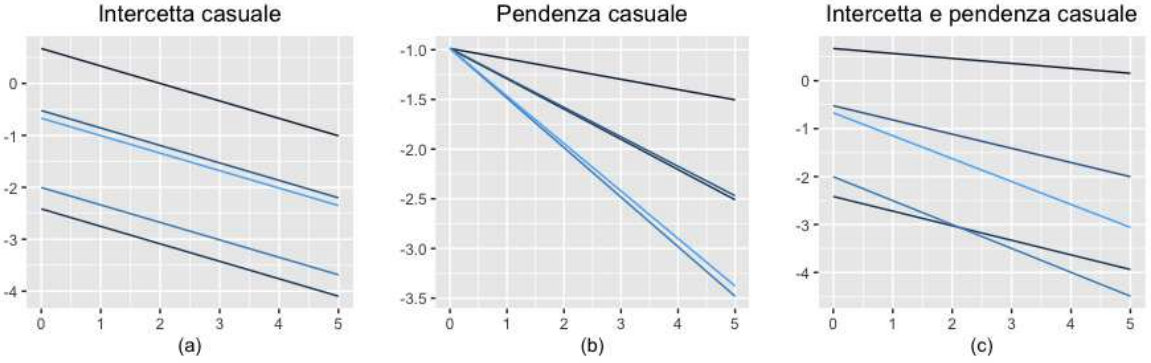

FIGURA 1.1: Rappresentazione grafica di tre diverse specificazioni della componente casuale di un LMM.

1.2 Modello lineare generalizzato a effetti misti

I modelli lineari generalizzati a effetti misti (GLMM) rappresentano un'estensione del modello lineare normale a effetti misti, che consente di modellare variabili risposta distribuite secondo una qualsiasi distribuzione della famiglia di dispersione esponenziale.

Una formulazione generale dei GLMM è la seguente (Hui, 2021)

$$\begin{aligned} \boldsymbol{y} \mid \mathbf{u} &\sim p(\boldsymbol{y} \mid \mathbf{u}; \boldsymbol{\beta}, \phi) \\ g(\boldsymbol{\mu}) = \boldsymbol{\eta} &= X\boldsymbol{\beta} + Z\mathbf{u} \\ \mathbf{u} &\sim N(\mathbf{0}, D(\boldsymbol{\theta})) \end{aligned} \tag{1.4}$$

dove $\boldsymbol{y}$ è variabile risposta di dimensione $n \times 1$, X e Z denotano le matrici di covariate degli effetti fissi e casuali, di dimensioni $n \times p$ e $n \times q$ rispettivamente, mentre $\boldsymbol{\beta}$ e $\mathbf{u}$ indicano rispettivamente gli effetti fissi e quelli casuali. Infine $\boldsymbol{\theta}$ è un vettore di parametri che definisce la matrice di varianza e covarianza dell'effetto casuale $\mathbf{u}$.

Condizionatamente agli effetti casuali $\mathbf{u}$, le osservazioni $\mathbf{y}$ sono assunte indipendenti e distribuite secondo una famiglia esponenziale $p(\mathbf{y} \mid \mathbf{u}; \boldsymbol{\beta}, \phi)$ con valore atteso $E(\mathbf{y} \mid \mathbf{u}) = \boldsymbol{\mu}$ e varianza condizionata $\mathrm{Var}(\mathbf{y} \mid \mathbf{u}) = \phi V(\boldsymbol{\mu})$, dove ϕ è il parametro di dispersione della famiglia di dispersione esponenziale. Il valore atteso $\boldsymbol{\mu}$ è legato al predittore lineare tramite una funzione di legame nota $g(\cdot)$ detta *link function*. Due casi di particolare interesse in questo lavoro sono:

- GLMM binomiali dove $p(\mathbf{y} \mid \mathbf{u}; \boldsymbol{\beta}, \phi)$ è la distribuzione binomiale, $g(\mu) = \log\{\mu/(1-\mu)\}$ è la funzione logistica di legame (canonico), $\phi = 1$ e la funzione di varianza è $V(\mu) = \mu(1-\mu)$,
- GLMM Poisson dove $p(\mathbf{y} \mid \mathbf{u}; \boldsymbol{\beta}, \phi)$ è la distribuzione di Poisson, $g(\mu) = \log(\mu)$ è la funzione logaritmo di legame (canonico), $\phi = 1$ e la funzione di varianza è $V(\mu) = \mu$.

Analogamente al modello lineare normale a effetti misti, anche il modello lineare generalizzato a effetti misti può essere riformulato per il caso particolare in cui si disponga di dati con una struttura gerarchica.

Capitolo 2

Metodi per la stima delle componenti di varianza

I metodi maggiormente impiegati per la stima nei modelli a effetti misti sono la stima di massima verosimiglianza (ML) e la stima di massima verosimiglianza ristretta (REML). La stima tramite massima verosimiglianza non ristretta tende a produrre stimatori delle componenti di varianza distorti verso lo zero, poiché non considera i gradi di libertà persi nella stima dei coefficienti degli effetti fissi. La stima REML, invece, presenta il vantaggio di correggere tale distorsione, tenendo conto dei gradi di libertà persi, e conduce a stimatori meno distorti delle componenti di varianza. Negli ultimi decenni, la stima REML nei modelli lineari a effetti misti (LMM) è stata oggetto di ampia trattazione nella letteratura. Tuttavia, come osservato da Maestrini et al. (2024), la definizione dello stimatore REML al di fuori di questa classe di modelli non è completamente chiara.

In questo capitolo, seguendo il lavoro di Maestrini et al. (2024), saranno presentati alcuni metodi per la stima delle componenti di varianza nei modelli a effetti misti, con particolare attenzione ai modelli GLMM.

2.1 Stima delle componenti di varianza nei LMM

Si consideri il modello in (1.4). La funzione di log verosimiglianza marginale per tale modello è definita come

$$\ell_M(\boldsymbol{\beta}, \phi, \boldsymbol{\theta}; \mathbf{y}) = \log \left\{ \int p(\mathbf{y} \mid \mathbf{u}; \boldsymbol{\beta}, \phi) p(\mathbf{u}; \boldsymbol{\theta}) d\mathbf{u} \right\}. \tag{2.1}$$

Nei modelli lineari normali a effetti misti, dove $p(\mathbf{y} \mid \mathbf{u}; \boldsymbol{\beta}, \phi)$ è la distribuzione normale, si ottiene lo stimatore di massima verosimiglianza non ristretta massimizzando

la funzione di log verosimiglianza marginale in (2.1) che in questo caso ha forma chiusa, infatti marginalmente la variabile risposta ha distribuzione normale.

Tuttavia, come menzionato in precedenza, lo stimatore ML non tiene conto dei gradi di libertà persi nella stima dei parametri fissi $\boldsymbol{\beta}$. Per correggere le stime distorte delle componenti di varianza nei modelli lineari misti, si utilizza la stima di massima verosimiglianza ristretta (REML), ottenuta massimizzando la funzione di log-verosimiglianza ristretta. La funzione obiettivo REML può essere ottenuta in diversi modi, in questa sede vediamo la derivazione condizionale di Verbyla (1990) per un modello lineare normale a effetti misti del tipo

$$\boldsymbol{y} = X\boldsymbol{\beta} + Z\mathbf{u} + \boldsymbol{\varepsilon}, \quad \mathbf{u} \sim N(0, \sigma^2 G), \quad \boldsymbol{\varepsilon} \sim N(0, \sigma^2 \rho R), \tag{2.2}$$

dove $\boldsymbol{\varepsilon}$ rappresenta un vettore di errori casuali.

Si noti che il modello in (2.2) è un caso particolare del GLMM in (1.4), con $D(\boldsymbol{\theta}) = \sigma^2 G$, $\boldsymbol{\theta} = (\sigma^2, \text{vech}(G)^\top)^\top$ e $\phi = \sigma^2 \rho$, dove vech$(\cdot)$ denota l'operatore di vettorizzazione sulla triangolare inferiore (o superiore). La funzione di log-verosimiglianza basata sulla distribuzione congiunta di $\boldsymbol{y}$ e $\mathbf{u}$ per il modello in (2.2) è data da

$$\begin{aligned}\ell_{LMM,J}(\boldsymbol{\beta}, \rho, \boldsymbol{\theta}, \text{vech}(R); \mathbf{y}, \mathbf{u}) = &-\frac{1}{2}\left\{\log\det(\sigma^2 G) + \log\det(\sigma^2 \rho R)\right\} \\ &- \frac{1}{2\sigma^2}\mathbf{u}^\top G^{-1}\mathbf{u} \\ &- \frac{1}{2\sigma^2\rho}(\mathbf{y} - X\boldsymbol{\beta} - Z\mathbf{u})^\top R^{-1}(\mathbf{y} - X\boldsymbol{\beta} - Z\mathbf{u}),\end{aligned}$$

Si supponga che la matrice di varianza $\sigma^2 V = \phi R + \sigma^2 ZGZ^\top$ esista e che non dipenda dai parametri $\boldsymbol{\beta}$. Denotiamo i parametri di varianza in V con $\boldsymbol{\psi}$. L'obiettivo è quello di costruire la verosimiglianza marginale per σ^2 e i parametri $\boldsymbol{\psi}$ eliminando i parametri di disturbo $\boldsymbol{\beta}$. Possiamo scrivere $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\zeta}$, dove $\boldsymbol{\zeta} \sim \mathcal{N}_n(0, \sigma^2 V)$. Se i parametri $\boldsymbol{\psi}$ e di conseguenza V fossero noti, lo stimatore di massima verosimiglianza di $\boldsymbol{\beta}$ sarebbe

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{\psi}} = (X^\top V^{-1} X)^{-1} X^\top V^{-1}\mathbf{y} = \boldsymbol{\beta} + (X^\top V^{-1} X)^{-1} X^\top V^{-1}\boldsymbol{\zeta},$$

la cui distribuzione è $\mathcal{N}_p(\boldsymbol{\beta}, \sigma^2(X^\top V^{-1} X)^{-1})$. Sia $H = X(X^\top X)^{-1} X^\top$ la matrice $n \times n$ che soddisfa $HX = X$. Un punto di partenza per costruire la verosimiglianza marginale è il vettore dei residui $(I_n - H)\mathbf{y}$, la cui distribuzione non dipende da $\boldsymbol{\beta}$. Tuttavia, poiché $I_n - H$ ha rango $n - p$, questa distribuzione è degenere, si prendono quindi solo $n - p$ residui linearmente indipendenti (Davison, 2003, Capitolo 12, pag.658). A tal fine, si consideri la trasformazione $(\mathbf{y}_1^\top, \mathbf{y}_2^\top)^\top = L\mathbf{y}$ dove $L = [L_1, L_2]$ è una matrice non singolare.

Le matrici L_1 e L_2 sono rispettivamente di dimensione $n \times p$ e $n \times (n-p)$ e soddisfano $L_1^\top X = I_p$, $L_2^\top X = O_{n-p,p}$, con $O_{n-p,p}$ matrice $(n-p) \times p$ di zeri, $L_2 L_2^\top = I_n - H$ e $L_2^\top L_2 = I_{n-p}$. Si ha che $\mathbf{y}_2 = L_2^\top \mathbf{y} = L_2^\top L_2 L_2^\top \mathbf{y} = L_2^\top (I_n - H)\mathbf{y}$ è una combinazione lineare dei residui.

La distribuzione di $L\mathbf{y}$ è data da

$$\begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{pmatrix} \sim N\left(\begin{pmatrix} \boldsymbol{\beta} \\ \mathbf{0} \end{pmatrix}, \sigma^2 \begin{pmatrix} L_1^\top V L_1 & L_1^\top V L_2 \\ L_2^\top V L_1 & L_2^\top V L_2 \end{pmatrix} \right),$$

Utilizzando le proprietà della distribuzione normale multivariata, la funzione di log-verosimiglianza congiunta può essere espressa come la somma di due componenti

$$\begin{aligned} \ell_{LMM,J}(\boldsymbol{\beta}, \rho, \boldsymbol{\theta}, \mathrm{vech}(R); \mathbf{y}, \mathbf{u}) = &\log p(\mathbf{y}_1 \mid \mathbf{y}_2; \boldsymbol{\beta}, \rho, \boldsymbol{\theta}, \mathrm{vech}(R)) \\ &+ \log p(\mathbf{y}_2; \rho, \boldsymbol{\theta}, \mathrm{vech}(R)). \end{aligned}$$

La distribuzione marginale di $\mathbf{y}_2$ non dipende dal vettore dei parametri degli effetti fissi $\boldsymbol{\beta}$, ed è su questa distribuzione che si costruisce la funzione di log-verosimiglianza ristretta

$$\begin{aligned} \ell_{LMM,R}(\rho, \boldsymbol{\theta}, \mathrm{vech}(R); \mathbf{y}) = &-\frac{1}{2}\left\{(n-p)\log(\sigma^2) + \log\det(L_2^\top V L_2)\right\} \\ &- \frac{1}{2\sigma^2}\mathbf{y}_2^\top (L_2^\top V L_2)^{-1}\mathbf{y}_2. \end{aligned} \tag{2.3}$$

Lo stimatore REML della matrice di varianza degli effetti casuali e del parametro di dispersione si ottiene massimizzando $\ell_{LMM,R}(\rho, \boldsymbol{\theta}, \mathrm{vech}(R); \mathbf{y})$.

Le variabili casuali Y_2 e $\hat{\boldsymbol{\beta}}_{\boldsymbol{\psi}}$ sono distribuite normalmente, con covarianza

$$\mathbb{E}\{Y_2(\hat{\boldsymbol{\beta}}_{\boldsymbol{\psi}} - \boldsymbol{\beta})\} = L_2^\top \mathbb{E}\{\boldsymbol{\zeta}^\top \boldsymbol{\zeta}^\top\} V^{-1} X (X^\top V^{-1} X)^{-1} = \sigma^2 L_2^\top V V^{-1} X (X^\top V^{-1} X)^{-1} = 0,$$

poiché $L_2^\top X = 0$. Quindi Y_2 e $\hat{\boldsymbol{\beta}}_{\boldsymbol{\psi}}$ sono indipendenti, e pertanto

$$f(\mathbf{y}_2; \boldsymbol{\psi}) = \frac{f(\mathbf{y}_2; \boldsymbol{\psi}) f(\hat{\boldsymbol{\beta}}_\psi; \boldsymbol{\beta}, \boldsymbol{\psi})}{f(\hat{\boldsymbol{\beta}}_\psi; \boldsymbol{\beta}, \psi)} = \frac{f(\mathbf{y}_2, \hat{\boldsymbol{\beta}}_\psi; \boldsymbol{\beta}, \boldsymbol{\psi})}{f(\hat{\boldsymbol{\beta}}_\psi; \boldsymbol{\beta}, \boldsymbol{\psi})} = \frac{f(\mathbf{y}; \boldsymbol{\beta}, \boldsymbol{\psi})}{f(\hat{\boldsymbol{\beta}}_\psi; \boldsymbol{\beta}, \boldsymbol{\psi})} \left| \frac{\partial \mathbf{y}}{\partial(\mathbf{y}_2, \hat{\boldsymbol{\beta}}_\psi)} \right|, \tag{2.4}$$

e lo jacobiano per il cambiamento di variabile da $(\mathbf{y}_2, \hat{\boldsymbol{\beta}}_\psi)$ a $\mathbf{y}$ è

$$\begin{aligned}\left|\frac{\partial(\mathbf{y}_2, \hat{\boldsymbol{\beta}}_\psi)}{\partial \mathbf{y}}\right| &= \left|L_2\, X(X^\top X)^{-1}\right| \\ &= \left|\begin{pmatrix} L_2^\top L_2 & L_2^\top X(X^\top X)^{-1} \\ (X^\top X)^{-1}X^\top L_2 & (X^\top X)^{-1} \end{pmatrix}\right|^{1/2} \\ &= \left|\begin{pmatrix} I_{n-p} & 0 \\ 0 & (X^\top X)^{-1} \end{pmatrix}\right|^{1/2} = \left|X^\top X\right|^{-1/2}.\end{aligned}$$

Sostituendo quest'ultimo e le densità normali di $\mathbf{y}$ e $\hat{\boldsymbol{\beta}}_{\boldsymbol{\psi}}$ in (2.4), si ottiene

$$f(\mathbf{y}_2; \boldsymbol{\psi}) = \frac{|X^\top X|^{1/2}|V^{-1}|^{1/2}}{(2\pi\sigma^2)^{(n-p)/2}|X^\top V^{-1} X|^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{y} - X\hat{\boldsymbol{\beta}}_\psi)^\top V^{-1}(\mathbf{y} - X\hat{\boldsymbol{\beta}}_\psi)\right\} \quad (2.5)$$

in cui L_2 non compare. Quindi l'equazione (2.5) non dipende dalla scelta di L_2.

L'espressione (2.5) è un'espressione alternativa alla (2.3) su cui si basa l'inferenza per $\boldsymbol{\psi}$, ed è nota come verosimiglianza ristretta perché il suo spazio dei parametri coinvolge solo σ^2 e $\boldsymbol{\psi}$. Tipicamente viene massimizzata la log verosimiglianza ristretta, ovvero il logaritmo della verosimiglianza ristretta in equazione (2.5), rispetto a $\boldsymbol{\psi}$, mediante una procedura di Newton-Raphson, oppure utilizzando l'algoritmo EM.

Si può dimostrare che lo stimatore REML ha le usuali proprietà dello stimatore di massima verosimiglianza, ossia le proprietà di normalità asintotica e matrice di varianza asintotica ottenuta attraverso l'hessiano di $\ell_{LMM,R}(\rho, \boldsymbol{\theta}, \text{vech}(R); \mathbf{y})$ (si veda ad esempio Verbyla, 1990; Cressie & Lahiri, 1993) .

2.2 Stima delle componenti di varianza nei GLMM

Analogamente al modello lineare normale a effetti misti, la stima di massima verosimiglianza per il modello lineare generalizzato a effetti misti (GLMM) si ottiene massimizzando la log-verosimiglianza marginale definita nell'equazione (2.1). Tuttavia, per i GLMM con distribuzioni non normali, l'integrale presente in (2.1) non ha una soluzione esplicita, pertanto deve essere approssimato con tecniche come l'approssimazione di Laplace, la quadratura di Gauss, le approssimazioni variazionali o i metodi Monte Carlo. Per ulteriori dettagli su queste tecniche si rimanda ai lavori di Breslow & Clayton (1993), McCulloch (1997), Hall et al. (2020), Hui (2021), Ormerod & Wand (2012), e Shun & McCullagh (1995).

Nel caso dei modelli lineari normali a effetti misti, la derivazione della log-verosimiglianza ristretta non è univoca; tuttavia, diverse formulazioni conducono alla stessa funzione obiettivo riportata in (2.3) (Maestrini et al., 2024). Questa proprietà di equivalenza tra le diverse formulazioni non si estende ai GLMM con distribuzioni non normali. Infatti, ciascuna derivazione della log-verosimiglianza ristretta può portare a una funzione obiettivo diversa nei GLMM.

In questo paragrafo, sulla base del lavoro di Maestrini et al. (2024), si presenterà una panoramica dei principali metodi di stima REML che sono stati proposti per i GLMM. Questi metodi possono essere classificati in quattro principali approcci REML: linearizzazione approssimata, verosimiglianza integrata, verosimiglianza profilo modificata e metodi basati sulla correzione della distorsione nella funzione punteggio.

2.2.1 Linearizzazione approssimata

Gran parte della letteratura sulla stima REML per i GLMM utilizza la tecnica che prende il nome di linearizzazione approssimata. Tale tecnica consiste nell'approssimare un modello lineare generalizzato a effetti misti con un modello lineare a effetti misti al fine di poter sfruttare la letteratura esistente per lo stimatore REML nei LMM.

Si vedrà in seguito la tecnica della linearizzazione approssimata per ottenere uno stimatore REML in tre differenti approcci:

- REML attraverso la tecnica della variabile risposta modificata (Schall, 1991; Wolfinger & O'connell, 1993);
- REML mediante il calcolo del miglior predittore lineare non distorto (BLUP) attraverso una procedura di calcolo iterativa (McGilchrist, 1994);
- un approccio di quasi-verosimiglianza per la stima dei GLMM (Breslow & Clayton, 1993; Engel & Keen, 1994).

Linearizzazione ed il metodo della variabile risposta modificata

Si consideri un GLMM con predittore lineare legato alla media della variabile risposta nel modo seguente (Schall, 1991)

$$g(\boldsymbol{\mu}) = X\boldsymbol{\beta} + \sum_{k=1}^{K} Z_k \mathbf{u}_k$$

con $\text{Cov}(\mathbf{u}) = D(\boldsymbol{\theta}) = \text{diag}(\sigma_1^2 I_{q_1}, \ldots, \sigma_K^2 I_{q_K})$, dove $\boldsymbol{\theta} = (\sigma_1^2, \ldots, \sigma_K^2)^\top$ sono le componenti di varianza della componente casuale del modello. Si osservi che tale modello è un

caso particolare del modello (1.4), dove $Z = \text{diag}(Z_1, \ldots, Z_K)$ di dimensione $n \times q$, con $q = \sum_{k=1}^{K} q_k$ e $\mathbf{u} = (\mathbf{u}_1^\top, \ldots, \mathbf{u}_K^\top)^\top$.

Il metodo di Schall (1991) utilizza una linearizzazione approssimata della funzione legame

$$g(y_i) \approx \eta_i + g'(\mu_i)(y_i - \mu_i) = \xi_i,$$

dove $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_n)^\top$ è un vettore di variabili dipendenti corrette.

A partire da tale variabile dipendente corretta, si può costruire un GLMM linearizzato approssimato

$$\boldsymbol{\xi} = X\boldsymbol{\beta} + Z\mathbf{u} + \mathbf{e}, \quad \text{E}(\mathbf{e}) = 0, \quad \text{Cov}(\mathbf{e}) = W^{-1}, \tag{2.6}$$

dove W^{-1} è una matrice diagonale con elementi $\text{Var}\{g'(\mu_i)(y_i - \mu_i)|\mathbf{u}\} = \phi V(\mu_i)\{g'(\mu_i)\}^2$. Se si considera l'equazione (2.6) come un LMM con valore atteso e varianza marginali $\mathbb{E}[\boldsymbol{\xi}] = X\boldsymbol{\beta}$ e $V(\boldsymbol{\xi}) = W^{-1} + ZD(\boldsymbol{\theta})Z^\top$, possiamo adottare lo stimatore REML per i LMM.

Schall (1991) propone una procedura iterativa in due fasi:

1. dati i valori correnti dei parametri $(\boldsymbol{\beta}^\top, \boldsymbol{\theta}^\top, \phi)^\top$ e degli effetti casuali $\mathbf{u}$, si aggiorna la variabile dipendente corretta $\boldsymbol{\xi}^\dagger = X\boldsymbol{\beta} + Z\mathbf{u}$, e calcola $D(\boldsymbol{\theta})$ e W;
2. si calcolano $\boldsymbol{\beta}$ e $\mathbf{u}$ risolvendo le equazioni del modello a effetti casuali di Henderson (1963),

$$A \begin{bmatrix} \boldsymbol{\beta} \\ \mathbf{u} \end{bmatrix} = \begin{bmatrix} X^\top W X & X^\top W Z \\ Z^\top W X & Z^\top W Z + D(\boldsymbol{\theta})^{-1} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta} \\ \mathbf{u} \end{bmatrix} = \begin{bmatrix} X^\top W \boldsymbol{\xi}^\dagger \\ Z^\top W \boldsymbol{\xi}^\dagger \end{bmatrix}. \tag{2.7}$$

Infine, si aggiornano le stime REML delle componenti di varianza.

Sia $T = \{Z^\top W Z - Z^\top W X (X^\top W X)^{-1} X^\top W Z + D(\boldsymbol{\theta})^{-1}\}^{-1}$ la matrice formata dalle ultime q colonne e q righe di A^{-1}. È possibile partizionare la matrice T in blocchi $T_{kk'}$ per $k, k' = 1, \ldots, K$ coerentemente con la struttura diagonale a blocchi di $D(\boldsymbol{\theta})$, e aggiornare gli elementi di $\boldsymbol{\theta}$ tramite la formula $\sigma_k^2 \leftarrow (\mathbf{u}_k^\top \mathbf{u}_k)/\{q_k - \sigma_k^{-2}\text{trace}(T_{kk})\}$ dove $\text{trace}(\cdot)$ è l'operatore traccia di una matrice. Se anche il parametro di dispersione ϕ deve essere stimato, si esegue un ulteriore passaggio seguendo uno schema di stima di tipo REML (si veda Fellner, 1986).

Di particolare rilevanza è il lavoro di Wolfinger & O'connell (1993), in cui si è considerata una specificazione del modello più generale come in (1.4). Nel dettaglio Wolfinger & O'connell (1993) assumono $\boldsymbol{\xi} \sim N(X\boldsymbol{\beta} + Z\mathbf{u}, W^{-1})$ e quindi, analogamente a quanto visto per il modello lineare normale a effetti misti, ottengono per $\boldsymbol{\xi}$ la funzione

di log-verosimiglianza ristretta (profilo rispetto a ϕ)

$$\ell_{\text{lin},R}(\boldsymbol{\theta};\mathbf{y}) = -\frac{1}{2}\log\det(V) - \frac{n-p}{2}\log r^\top V^{-1} r \\ -\frac{1}{2}\log\det(X^\top V^{-1} X),$$

dove $V = W^{-1} + ZD(\boldsymbol{\theta})Z^\top$, $r = \boldsymbol{\xi} - X(X^\top V^{-1}X)^{-1}X^\top V^{-1}\boldsymbol{\xi}$. Si osservi inoltre che la matrice V qui è anche una funzione di $\boldsymbol{\beta}$ e $\mathbf{u}$ attraverso W. Wolfinger & O'connell (1993) hanno poi seguito una procedura iterativa in due passi; prima aggiornando i parametri fissi $\boldsymbol{\beta}$ e casuali $\boldsymbol{u}$ mediante l'equazione (2.7) e ϕ attraverso una stima di massima verosimiglianza profilo, poi massimizzando $\ell_{\text{lin},R}(\boldsymbol{\theta})$ per aggiornare $\boldsymbol{\theta}$.

Linearizzazione e l'approccio BLUP

Si consideri il modello definito tramite $g(\mu) = X\boldsymbol{\beta} + \sum_{k=1}^{K} Z_k \mathbf{u_k}$ e $\text{Cov}(u) = D(\boldsymbol{\theta}) = \text{diag}(\sigma_1^2 A_1, \ldots, \sigma_K^2 A_K)$, simile al modello di Schall (1991), con l'aggiunta di matrici note $A_1, \ldots, A_K$. La log-verosimiglianza congiunta basata sulla distribuzione congiunta di $\boldsymbol{y}$ e $\mathbf{u}$ è

$$\begin{aligned}\ell_J(\boldsymbol{\theta}, \boldsymbol{\beta}, \phi; \mathbf{y}, \mathbf{u}) &= \log\{p(\mathbf{y}|\mathbf{u}; \boldsymbol{\beta}, \phi)\} - \frac{1}{2}\sum_{k=1}^{K}\left\{q_k \log(\sigma_k^2) + \frac{1}{\sigma_k^2}\mathbf{u_k}^\top A_k^{-1}\mathbf{u_k}\right\} \\ &= \ell_1(\boldsymbol{\beta}, \phi, \mathbf{u}) + \ell_2(\boldsymbol{\theta}, \mathbf{u})\end{aligned} \tag{2.8}$$

con $\mathbf{u_k}$ un vettore di dimensione $q_k \times 1$.

Sia $H = -\nabla^2_{\boldsymbol{\beta},\mathbf{u}}\ell_1(\boldsymbol{\beta}, \phi, \mathbf{u})$ la matrice di informazione osservata per $\boldsymbol{\beta}$ e $\mathbf{u}$. McGilchrist (1994) propone di sostituire $\ell_1(\boldsymbol{\beta}, \phi, \mathbf{u})$ con un'approssimazione quadratica in un intorno di $(\tilde{\boldsymbol{\beta}}^\top, \tilde{\mathbf{u}}^\top) = \arg\max_{\boldsymbol{\beta},\mathbf{u}} \ell_1(\boldsymbol{\beta}, \phi, \mathbf{u})$,

$$\ell_1^*(\boldsymbol{\beta}, \phi, \mathbf{u}) = -\frac{1}{2}\begin{bmatrix}\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}} \\ \mathbf{u} - \tilde{\mathbf{u}}\end{bmatrix}^\top H \begin{bmatrix}\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}} \\ \mathbf{u} - \tilde{\mathbf{u}}\end{bmatrix} = -\frac{1}{2}(\mathbf{y}^* - X\boldsymbol{\beta} - Z\mathbf{u})^\top W(\mathbf{y}^* - X\boldsymbol{\beta} - Z\mathbf{u}),$$

dove $\mathbf{y}^* = X\tilde{\boldsymbol{\beta}} + Z\tilde{\mathbf{u}}$ è la risposta stimata e W che è la matrice diagonale definita nel modello in (2.6). Si osservi che l'approssimazione quadratica è costruita solo da ℓ_1, di conseguenza i parametri fissi e casuali, $\boldsymbol{\beta}$ e $\mathbf{u}$, tipicamente non possono essere identificati poiché la matrice $[X, Z]$ non è a rango pieno. Tuttavia, dato che la forma finale di ℓ_1^* richiede solo la risposta stimata $\mathbf{y}^*$ ($\boldsymbol{\beta}$ e $\mathbf{u}$ non sono identificabili ma $\mathbf{y}^*$ è unica), e W è

una matrice diagonale che può essere invertita, l'algoritmo di stima basato su questa approssimazione quadratica non incontra problemi.

Possiamo considerare l'approssimazione della funzione di log-verosimiglianza congiunta $\ell_J^*(\boldsymbol{\theta}, \boldsymbol{\beta}, \phi; \mathbf{y}, \mathbf{u}) = \ell_1^* + \ell_2$ come una log-verosimiglianza linearizzata approssimata, e a partire da essa, McGilchrist (1994) propone di costruire una procedura standard per la stima di un LMM come segue. Dati $\boldsymbol{\theta}$ e ϕ, calcoliamo gli stimatori BLUP di $\boldsymbol{\beta}$ e $\mathbf{u}$ massimizzando $\ell_J^*(\boldsymbol{\theta}, \boldsymbol{\beta}, \phi; \mathbf{y}, \mathbf{u})$. Si osservi che, come sottolineato in Maestrini et al. (2024), ciò equivale a risolvere l'equazione in (2.7) con $\xi^\dagger$ sostituito da $\mathbf{y}^*$. Dati $\boldsymbol{\beta}$ e $\mathbf{u}$, si ottengono gli aggiornamenti per le stime REML di σ_k^2 in modo analogo a quelli in Schall (1991). Per $k = 1, \ldots, K$, si utilizza la formula di aggiornamento $\sigma_k^2 \leftarrow (\mathbf{u}_k^\top A_k^{-1} \mathbf{u}_k)/(q_k - \sigma_k^{-2}\text{trace}(A_k^{-1} T_{kk}))$. Se anche ϕ deve essere stimato, può essere utilizzato uno stimatore di tipo REML.

McGilchrist (1994) puntualizza che sotto la stessa specificazione del modello tale metodo di linearizzazione approssimata è equivalente e produce le stesse stime REML del metodo proposto da Schall (1991).

Linearizzazione e stime di quasi verosimiglianza

Fino a ora ci si è focalizzati su metodi di stima basati sulla funzione di verosimiglianza. Per costruire una funzione di verosimiglianza è generalmente necessario specificare un meccanismo probabilistico generatore dei dati osservati. Spesso non vi è teoria disponibile sul meccanismo casuale mediante il quale i dati sono stati generati. In tale situazione è comune utilizzare delle tecniche che prendono il nome di quasi-verosimiglianza. I metodi di quasi-verosimiglianza sfruttano la relazione tra la media e la varianza della variabile risposta per costruire una funzione obiettivo, senza richiedere la completa specificazione di una distribuzione probabilistica del modello. Per approfondimenti circa i metodi di quasi-verosimiglianza si veda ad esempio McCullagh & Nelder (1989, Capitolo 9).

Nel contesto dei GLMM Breslow & Clayton (1993) hanno proposto due metodi di quasi-verosimiglianza e successivamente sviluppato la stima REML per l'adattamento dei GLMM. Si consideri il modello in (1.4) su cui vengono fatte solo assunzioni del secondo ordine e non assunzioni distributive per $\mathbf{y}$. Si assume quindi

$$g(\mathbb{E}(y_i|\mathbf{u})) = g(\mu_i) = \mathbf{x}_i^\top \boldsymbol{\beta} + \mathbf{z}_i^\top \mathbf{u} \quad \text{e} \quad \text{Var}(y_i|\mathbf{u}) = \phi a_i V(\mu_i) \quad \text{per } i = 1, \ldots, n,$$

dove a_i è una opportuna costante e $V(\mu)$ è una funzione di varianza nota. Se gli effetti casuali sono distribuiti normalmente con media zero e matrice di covarianza $D(\boldsymbol{\theta})$, allora

possiamo definire la funzione di log quasi-verosimiglianza per il modello come

$$\ell_{\text{quasi},M}(\boldsymbol{\beta},\phi,\boldsymbol{\theta};\mathbf{y}) = -\frac{1}{2}\log\det\{D(\boldsymbol{\theta})\} + \log\left[\int \exp\{-\kappa(\boldsymbol{\beta},\phi,\boldsymbol{\theta};\mathbf{y},\mathbf{u})\}d\mathbf{u}\right], \qquad (2.9)$$

dove $\kappa(\boldsymbol{\beta},\phi,\boldsymbol{\theta};\mathbf{y},\mathbf{u}) = \frac{1}{2}\sum_{i=1}^{n} d_i(y_i,\mu_i)/\phi + \frac{1}{2}\mathbf{u}^\top D(\boldsymbol{\theta})^{-1}\mathbf{u}$, e $d_i(y_i,\mu_i) = -2\int_{y_i}^{\mu_i}\frac{(y_i-t)}{a_i V(t)}dt$. Si osservi che d_i rappresenta la funzione di quasi-verosimiglianza per μ_i. Per alcuni esempi di funzioni di quasi-verosimiglianza per diverse specificazioni di $V(\mu)$ di veda McCullagh & Nelder (1989, Capitolo 9, Tabella 9.1).

Per risolvere l'integrale in (2.9) si utilizza un'approssimazione di Laplace basata su un'espansione di Taylor al secondo ordine in un intorno del minimo, $\tilde{\mathbf{u}}$, di $\kappa(\boldsymbol{\beta},\phi,\boldsymbol{\theta};\mathbf{y},\mathbf{u})$. Siano κ' e κ'' le derivate parziali prima e seconda di κ rispetto a $\boldsymbol{u}$. L'approssimazione di Laplace ci fornisce la seguente funzione di log quasi-verosimiglianza

$$\begin{aligned}\ell_{\text{quasi},M}(\boldsymbol{\beta},\boldsymbol{\theta},\phi) \approx &-\frac{1}{2}\log\det\{D(\boldsymbol{\theta})\} - \frac{1}{2}\log\mid \kappa''(\boldsymbol{\beta},\phi,\boldsymbol{\theta};\mathbf{y},\tilde{\boldsymbol{u}}) \mid \\ &- \kappa(\boldsymbol{\beta},\phi,\boldsymbol{\theta};\mathbf{y},\tilde{\mathbf{u}}))\end{aligned}$$

dove $\tilde{\mathbf{u}} = \tilde{\mathbf{u}}(\boldsymbol{\beta},\boldsymbol{\theta})$ denota la soluzione di

$$\kappa'(\boldsymbol{\beta},\phi,\boldsymbol{\theta};\mathbf{y},\mathbf{u}) = -\sum_{i=1}^{n}\frac{(y_i-\mu_i)\mathbf{z}_i}{\phi a_i V(\mu_i)g'(\mu_i)} + D^{-1}\mathbf{u} = 0$$

che minimizza $\kappa(\boldsymbol{\beta},\phi,\boldsymbol{\theta};\mathbf{y},\mathbf{u})$ rispetto a $\mathbf{u}$. Differenziando nuovamente rispetto a $\mathbf{u}$, si ottiene

$$\begin{aligned}\kappa''(\boldsymbol{\beta},\phi,\boldsymbol{\theta};\mathbf{y},\mathbf{u}) &= \sum_{i=1}^{n}\frac{\mathbf{z}_i\mathbf{z}_i^\top}{\phi a_i V(\mu_i)[g'(\mu_i)]^2} + D^{-1} + Q \\ &\approx Z^\top W Z + D^{-1},\end{aligned}$$

dove W^{-1} è la matrice definita per il modello in (2.6) con la differenza che gli elementi diagonali sono moltiplicati per a_i e $Q = -\sum_{i=1}^{n}(y_i-\mu_i)\mathbf{z}_i\frac{\partial}{\partial\mathbf{u}}\left[\frac{1}{\phi a_i V(\mu_i)g'(\mu_i)}\right]$ che ha valore atteso nullo. Sotto l'assunzione che gli elementi di W varino lentamente (o non varino proprio) al variare del valore atteso della risposta e omettendo il logaritmo del determinante di D, Breslow & Clayton (1993) giungono alla definizione della funzione di log quasi-verosimiglianza penalizzata (PQL)

$$\ell_{PQL,M}(\boldsymbol{\beta},\phi;\mathbf{y},\mathbf{u}) = -\frac{1}{2\phi}\sum_{i=1}^{n} d(y_i,\mu_i) - \frac{1}{2}\mathbf{u}^\top D(\boldsymbol{\theta})^{-1}\mathbf{u}, \qquad (2.10)$$

in cui si cercano $(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{u}}) = (\hat{\boldsymbol{\beta}}(\boldsymbol{\theta}),\hat{\mathbf{u}}(\boldsymbol{\theta}))$, con $\hat{\mathbf{u}}(\boldsymbol{\theta}) = \tilde{\mathbf{u}}(\boldsymbol{\beta}(\boldsymbol{\theta}))$, e ϕ che congiuntamente la massimizzino.

Dati ϕ e $\boldsymbol{\theta}$ massimizzare la (2.10) equivale a risolvere l'equazione (2.7). Tuttavia,

Breslow & Clayton (1993) propongono di utilizzare un algoritmo di tipo Fisher *scoring* per ottimizzare la funzione di log quasi-verosimiglianza penalizzata. Per aggiornare i parametri di varianza in $\boldsymbol{\theta}$, Breslow & Clayton (1993) utilizzano una linearizzazione approssimata per costruire $\xi_i = \eta_i + g'(\mu_i)(y_i - \mu_i)$. Adattando lo sviluppo di Harville (1977) per i LMM, dati i valori di $\boldsymbol{\beta}$, ϕ e $\mathbf{u}$, la funzione di quasi verosimiglianza profilo REML è definita come

$$\begin{aligned}\ell_{\text{PQL,R}}(\boldsymbol{\theta}; \mathbf{y}) = &-\frac{1}{2}\log\det(V) - \frac{1}{2}(\boldsymbol{\xi}^\dagger - X\boldsymbol{\beta})^\top V^{-1}(\boldsymbol{\xi}^\dagger - X\boldsymbol{\beta}) \\ &-\frac{1}{2}\log\det(X^\top V^{-1} X),\end{aligned} \tag{2.11}$$

dove $\boldsymbol{\xi}^\dagger = X\boldsymbol{\beta} + Z\mathbf{u}$, $V = W^{-1} + ZD(\boldsymbol{\theta})Z^\top$. Iterando tra le equazioni (2.10) e (2.11) si ottengono le stime PQL per il modello.

Come affermato da Maestrini et al. (2024) è possibile notare la stretta somiglianza tra la stima PQL e il metodo della pseudo-verosimiglianza di Wolfinger & O'connell (1993). Infatti, i due approcci producono in molte situazioni le stesse stime REML delle componenti di varianza. McGilchrist (1994) osserva che l'approccio PQL condivide anche alcuni elementi con il metodo BLUP. Inoltre anche il lavoro di Schall (1991) può essere riformulato all'interno del metodo PQL (Maestrini et al., 2024).

2.2.2 Verosimiglianza integrata

Un altro approccio per costruire stimatori REML nei GLMM è motivato da una formulazione bayesiana del modello, in cui gli effetti fissi vengono considerati variabili casuali, a cui viene assegnata una distribuzione a priori non informativa Harville (1977). Si consideri il modello in (1.4) in cui si specifica una *flat prior* per i parametri degli effetti fissi $\boldsymbol{\beta}$. Si ottiene l'approccio di verosimiglianza integrata marginalizzando la funzione di verosimiglianza congiunta sia rispetto agli effetti fissi che agli effetti casuali del modello, ottenendo

$$\begin{aligned}\ell_{\text{int, R}}(\boldsymbol{\theta}, \phi; \mathbf{y}) &= \log\left\{\int \exp[\ell_J(\boldsymbol{\theta}, \boldsymbol{\beta}, \phi; \mathbf{y}, \mathbf{u},)]p(\boldsymbol{\theta})d\boldsymbol{\beta}d\mathbf{u}\right\} \\ &= \log\left\{\int p(\mathbf{y} \mid \mathbf{u}; \boldsymbol{\beta}, \boldsymbol{\theta})p(\mathbf{u} \mid \boldsymbol{\theta})d\boldsymbol{\beta}d\mathbf{u}\right\}.\end{aligned} \tag{2.12}$$

La motivazione che sta dietro all'uso dell'approccio di verosimiglianza integrata come forma di stima REML deriva dal fatto che, per i modelli lineari a effetti misti (LMM), integrare rispetto a $\boldsymbol{\beta}$ cui si è assegnata una a priori costante, porta esattamente alla log-verosimiglianza ristretta in equazione (2.3). Tale equivalenza è stata mostrata per la prima volta da Laird & Ware (1982).

Stiratelli et al. (1984) hanno considerato un GLMM per dati longitudinali e proposto un processo iterativo a due fasi. In primo luogo, dati i parametri $\boldsymbol{\theta}$, si aggiornano $(\boldsymbol{\beta}, \mathbf{u})$ massimizzandone la distribuzione a posteriori. Si osservi che di fatto, avendo specificato una distribuzione a priori costante per i parametri degli effetti fissi, la distribuzione a posteriori per i parametri $(\boldsymbol{\beta}, \mathbf{u})$ coincide con la funzione di log-verosimiglianza congiunta in equazione (2.8)

$$\begin{aligned} p(\boldsymbol{\beta}, \mathbf{u} \mid \phi, \boldsymbol{\theta}; \mathbf{y}) &\propto \exp\{\ell_J(\boldsymbol{\beta}, \phi, \boldsymbol{\theta}; \mathbf{y}, \mathbf{u})\} p(\boldsymbol{\beta}) \\ &= \exp[\log\{p(\mathbf{y} \mid \mathbf{u}; \boldsymbol{\beta}, \phi)\} + \log\{p(\mathbf{u}; \boldsymbol{\theta})\}]. \end{aligned}$$

La massimizzazione di tale funzione equivale a con risolvere l'equazione in (2.7) e massimizzare la PQL in (2.10). Ottenute le stime

$$(\hat{\boldsymbol{\beta}}^\top, \hat{\mathbf{u}}^\top)^\top = \underset{(\boldsymbol{\beta}^\top, \mathbf{u}^\top)^\top}{\arg\max}\, p(\boldsymbol{\beta}, \mathbf{u} \mid \phi, \boldsymbol{\theta}; \mathbf{y})$$

si calcolano le stime REML per i parametri di varianza in $\boldsymbol{\theta}$ massimizzando la funzione di log-verosimiglianza integrata in (2.12). Stiratelli et al. (1984) propongono di massimizzare la (2.12) attraverso l'algoritmo EM, tuttavia in questo contesto il passo di *Expectation* risulta essere computazionalmente intrattabile (Maestrini et al., 2024). È possibile approssimare la distribuzione di $(\hat{\boldsymbol{\beta}}^\top, \hat{\mathbf{u}}^\top)^\top$ condizionata a $\mathbf{y}$ con una distribuzione normale multivariata centrata nelle stime correnti della moda a posteriori e con matrice di varianza data dall'inversa della matrice di informazione osservata $\boldsymbol{\Omega}(\boldsymbol{\beta}, \phi, \boldsymbol{\theta}; \mathbf{y}, \mathbf{u}) = \left\{-\nabla^2_{\boldsymbol{\beta},\mathbf{u}} \ell_J(\boldsymbol{\beta}, \phi, \boldsymbol{\theta}; \mathbf{y}, \mathbf{u})\right\}^{-1}$, ovvero

$$(\boldsymbol{\beta}, \mathbf{u} \mid \phi, \boldsymbol{\theta}; \mathbf{y}) \dot{\sim} N\left(\begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}} \end{bmatrix}, \Omega(\hat{\boldsymbol{\beta}}, \phi, \boldsymbol{\theta}, \hat{\mathbf{u}}; \mathbf{y})\right). \tag{2.13}$$

Tale approssimazione facilita notevolmente l'ottimizzazione della log-verosimiglianza integrata, in quanto il passo di *Expectation* si riduce al calcolo dell'integrale $\int \mathbf{u}\mathbf{u}^\top p(\boldsymbol{\beta}, \mathbf{u} \mid \phi, \boldsymbol{\theta}; \mathbf{y}) d\mathbf{u}$, il quale, utilizzando la distribuzione approssimata in (2.13), ha forma chiusa.

Sebbene l'algoritmo EM sia ampiamente utilizzato per la massimizzazione della log-verosimiglianza integrata in (2.1), quando la numerosità campionaria o la dimensione dei parametri diventa elevata, tale approccio risulta essere computazionalmente intensivo.

Un approccio più moderno alla stima REML dei parametri di varianza di un GLMM è la massimizzazione della verosimiglianza integrata approssimata tramite approssimazione di Laplace al primo ordine. La log-verosimiglianza integrata approssimata mediante Laplace al primo ordine è data da

$$\begin{aligned}\ell_{\text{int,R}}(\boldsymbol{\theta};\mathbf{y}) \approx &\quad \ell_J(\hat{\boldsymbol{\beta}},\hat{\phi},\boldsymbol{\theta},\hat{\mathbf{u}};\mathbf{y}) - \frac{1}{2}\log\det\{\boldsymbol{\Omega}(\hat{\boldsymbol{\beta}},\hat{\phi},\boldsymbol{\theta},\hat{\mathbf{u}};\mathbf{y})\},\\ \approx &-\frac{1}{2}\log\det\{D(\boldsymbol{\theta})\} - \frac{1}{2}\hat{\mathbf{u}}^\top D(\boldsymbol{\theta})^{-1}\hat{\mathbf{u}}\\ &-\frac{1}{2}\log\det\left[Z^\top\left(\hat{W} - \hat{W}X\left(X^\top\hat{W}X\right)^{-1}X^\top\hat{W}\right)Z + D(\boldsymbol{\theta})^{-1}\right]\\ &+\text{c},\end{aligned} \tag{2.14}$$

dove, dato $\boldsymbol{\theta}$, $(\hat{\boldsymbol{\beta}}^\top, \hat{\phi}^\top, \hat{\mathbf{u}}^\top)^\top = \arg\max_{\boldsymbol{\beta},\boldsymbol{\phi},\mathbf{u}} \ell_J(\boldsymbol{\beta},\phi,\boldsymbol{\theta};\mathbf{y},\mathbf{u})$, $\hat{W}$ è l'inversa della matrice di pesi definita per il modello in (2.6) valutata in $(\hat{\boldsymbol{\beta}}^\top, \hat{\phi}^\top, \hat{\mathbf{u}}^\top)^\top$, e c denota termini costanti rispetto a $\boldsymbol{\theta}$.

2.2.3 Verosimiglianza profilo modificata

Per i modelli lineari a effetti misti è possibile dimostrare (Bellhouse, 1990; Cox & Reid, 1992) che gli approcci di verosimiglianza profilo modificata (Severini, 2000, Capitlo 9) e verosimiglianza condizionata approssimata (Cox & Reid, 1987) producono stimatori REML delle componenti di varianza.

Bellio & Brazzale (2011) hanno proposto un approccio REML per la stima di modelli lineari generalizzati a effetti misti che consiste nel formulare una funzione di verosimiglianza profilo modificata, che nel caso dei LMM porta alla funzione di log-verosimiglianza ristretta in (2.3). Tale funzione è definita come

$$\begin{aligned}\ell_{\text{MPL},R}(\boldsymbol{\theta};\mathbf{y}) =& \ell_P(\boldsymbol{\theta};\mathbf{y}) + \frac{1}{2}\log\det\left\{\frac{\partial^2}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^\top}\ell_M(\boldsymbol{\beta},\boldsymbol{\theta};\mathbf{y})\right\}\\ &-\log\det\{C(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\theta}};\hat{\boldsymbol{\beta}}_{\boldsymbol{\theta}},\boldsymbol{\theta})\},\end{aligned} \tag{2.15}$$

dove $\ell_P(\boldsymbol{\theta};\mathbf{y}) = \ell_M(\hat{\boldsymbol{\beta}}_{\boldsymbol{\theta}},\boldsymbol{\theta};\mathbf{y})$ è la funzione di log-verosimiglianza profilo ottenuta sostituendo i coefficienti degli effetti fissi con $\hat{\boldsymbol{\beta}}_{\boldsymbol{\theta}}$, ovvero le stime di massima verosimiglianza di $\boldsymbol{\beta}$ dati i parametri di varianza $\boldsymbol{\theta}$, mentre $(\hat{\boldsymbol{\beta}}^\top, \hat{\boldsymbol{\theta}}^\top)^\top$ sono le stime di massima verosimiglianza ottenute dalla funzione di log-verosimiglianza marginale in equazione (2.1), infine $C(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\theta}},\hat{\boldsymbol{\beta}}_{\boldsymbol{\theta}},\boldsymbol{\theta})$ è un opportuno termine di correzione.

Per il termine di correzione è possibile adottare diverse specificazioni asintoticamente equivalenti. Bellio & Brazzale (2011) usano

$$C(\boldsymbol{\beta_0},\boldsymbol{\theta_0};\boldsymbol{\beta_1},\boldsymbol{\theta_1}) = \text{Cov}_{\boldsymbol{\beta_0},\boldsymbol{\theta_0}}\left(\frac{\partial}{\partial\boldsymbol{\beta}}\ell_M(\boldsymbol{\beta_0},\boldsymbol{\theta_0};\mathbf{y}), \frac{\partial}{\partial\boldsymbol{\beta}}\ell_M(\boldsymbol{\beta_1},\boldsymbol{\theta}_1;\mathbf{y})\right),$$

dove $\frac{\partial}{\partial \boldsymbol{\beta}} \ell_M(\boldsymbol{\beta_0}, \boldsymbol{\theta_0}; \mathbf{y})$ è la derivata, rispetto agli effetti fissi $\boldsymbol{\beta}$, della funzione di log-verosimiglianza marginale, calcolata in un generico $\boldsymbol{\beta_0}$ e $\boldsymbol{\theta_0}$, e $\frac{\partial}{\partial \boldsymbol{\beta}} \ell_M(\boldsymbol{\beta_1}, \boldsymbol{\theta_1}; \mathbf{y})$ è definita in modo analogo. Per il caso di GLMM con gruppi indipendenti e ϕ noto, il termine di correzione può essere approssimato empiricamente da

$$\hat{C}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}}; \hat{\boldsymbol{\beta}}_{\boldsymbol{\theta}}, \boldsymbol{\theta}) = \sum_{i=1}^{K} \left\{ \frac{\partial}{\partial \boldsymbol{\beta}} \ell_M(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}}; \mathbf{y_i}) \right\} \left\{ \frac{\partial}{\partial \beta} \ell_M(\hat{\boldsymbol{\beta}}_{\boldsymbol{\theta}}, \boldsymbol{\theta}; \mathbf{y}_i) \right\}^{\top},$$

dove $\ell_M(\boldsymbol{\beta}, \boldsymbol{\theta}; \mathbf{y}_i) = \log \int p(\mathbf{y}_i \mid \mathbf{u}_i; \boldsymbol{\beta}) p(\mathbf{u}_i; \boldsymbol{\theta}) d\mathbf{u}_i$ è la funzione di log-verosimiglianza marginale per l'i-esimo cluster con variabile risposta $\mathbf{y}_i = (y_{i1}, \ldots, y_{in_i})^{\top}$.

Una limitazione dell'approccio della verosimiglianza profilo modificata per la stima REML è la sua applicazione a configurazioni complesse (Maestrini et al., 2024). Ad esempio, nei GLMM non nested (modelli per cui un effetto casuale può essere condiviso su più gruppi) oltre al problema computazionale dato dal calcolo di integrali di dimensione maggiore, l'approssimazione empirica $\hat{C}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}}; \hat{\boldsymbol{\beta}}_{\boldsymbol{\theta}}, \boldsymbol{\theta})$ non è facilmente generalizzabile a causa della dipendenza tra le osservazioni nei diversi gruppi.

2.2.4 Correzione della distorsione nella *score* function

L'ultima classe di metodi di stima REML che vediamo in questo capitolo è stata proposta da Liao & Lipsitz (2002) e si basa sulla soluzione di un insieme di equazioni di stima derivate dalla correzione della distorsione nella funzione punteggio profilo delle componenti di varianza dei GLMM.

L'idea alla base della costruzione di una funzione *score* profilo corretta è da attribuirsi a McCullagh & Tibshirani (1990), attraverso la quale hanno costruito uno stimatore REML per un modello lineare a effetti misti.

Si assuma che ϕ sia noto, e che $(\hat{\boldsymbol{\beta}}^{\top}, \hat{\boldsymbol{\theta}}^{\top})^{\top}$ denotino gli stimatori di massima verosimiglianza standard ottenuti ottimizzando la log-verosimiglianza marginale in (2.1). Si osserva che, dalla funzione di log-verosimiglianza congiunta $\ell_J(\boldsymbol{\beta}, \phi, \boldsymbol{\theta}; \mathbf{y}, \mathbf{u}) = \log\{p(\mathbf{y} \mid \mathbf{u}; \boldsymbol{\beta}, \phi)\} + \log\{p(\mathbf{u}; \boldsymbol{\theta})\}$, fissato $\boldsymbol{\beta}$, la componente corrispondente ai parametri di varianza $\boldsymbol{\theta}$ è data da

$$\ell_{\mathrm{P}}(\boldsymbol{\theta}) = -\log\det\{D(\boldsymbol{\theta})\} - \mathrm{tr}\{[D(\boldsymbol{\theta})]^{-1}\mathbf{u}\mathbf{u}^{\top}\}$$

da cui otteniamo lo stimatore di massima verosimiglianza profilo per $\boldsymbol{\theta}$, $\hat{\boldsymbol{\theta}}_{\beta}$.

L'approccio di Liao & Lipsitz (2002) confronta l'equazione che eguaglia a zero la funzione score per $\hat{\boldsymbol{\theta}}$ con quella basata sulla log verosmiglianza profilo, per $\hat{\boldsymbol{\theta}}_{\beta}$, con l'obiettivo di individuare un termine di correzione al fine di derivare uno stimatore

REML che presenti una minore distorsione. Si consideri la funzione

$$h(\boldsymbol{\theta}; S) = -\frac{\partial}{\partial \boldsymbol{\theta}} \left(\log |D(\boldsymbol{\theta})| + \text{tr}\{[D(\boldsymbol{\theta})]^{-1} S\} \right),$$

dove S è una matrice con le stesse dimensioni di $D(\boldsymbol{\theta})$ ed è costante rispetto a $\boldsymbol{\theta}$. È possibile dimostrare che l'elemento r-esimo di $h(\boldsymbol{\theta}, S)$ è dato da

$$h(\boldsymbol{\theta}, S)_r = \text{tr}\left[\{D(\boldsymbol{\theta})\}^{-1} \{S - D(\boldsymbol{\theta})\} \{D(\boldsymbol{\theta})\}^{-1} \partial D(\boldsymbol{\theta}) \partial \boldsymbol{\theta}_r \right].$$

Da un'identità generale per l'algoritmo EM (McLachlan & Krishnan, 2007, Capitolo 3, pag. 95), l'equazione che eguaglia a zero la funzione score per $\hat{\boldsymbol{\theta}}_\beta$ è data da

$$\frac{\partial}{\partial \boldsymbol{\theta}} \ell_M(\boldsymbol{\beta}, \boldsymbol{\theta}; \mathbf{y}) = h(\boldsymbol{\theta}; \mathbb{E}_{\mathbf{u}|\boldsymbol{\beta},\boldsymbol{\theta},y}(\mathbf{u}\mathbf{u}^\top \mid \boldsymbol{\beta}, \boldsymbol{\theta}; \mathbf{y})) = 0, \tag{2.16}$$

dove il valore atteso di $\mathbf{u}\mathbf{u}^\top$ è calcolato rispetto alla distribuzione condizionata $p(\mathbf{u}|\mathbf{y}, \boldsymbol{\theta}) \propto p(\mathbf{y}|\mathbf{u}; \beta) p(\mathbf{u}; \boldsymbol{\theta})$. Ricordando che $\hat{\boldsymbol{\theta}}$ massimizza la funzione di log-verosimiglianza profilo $\ell_P(\boldsymbol{\theta}; \mathbf{y}) = \ell_M(\hat{\boldsymbol{\beta}}_{\boldsymbol{\theta}}, \boldsymbol{\theta}; \mathbf{y})$ segue dalla (2.16) che l'equazione dello score profilo per $\hat{\boldsymbol{\theta}}$ è data da

$$\frac{\partial}{\partial \boldsymbol{\theta}} \ell_P(\boldsymbol{\theta}; \mathbf{y}) = h\left(\boldsymbol{\theta}; \mathbb{E}_{\mathbf{u}|\hat{\boldsymbol{\beta}},\boldsymbol{\theta},\mathbf{y}} \left(\mathbf{u}\mathbf{u}^\top \mid \hat{\boldsymbol{\beta}}_{\boldsymbol{\theta}}, \boldsymbol{\theta}; \mathbf{y} \right) \right) = 0. \tag{2.17}$$

Confrontando le equazioni (2.16) e (2.17), e sfruttando il fatto che $\mathbb{E}_{\mathbf{y}|\boldsymbol{\beta},\boldsymbol{\theta}} \left\{ \mathbb{E}_{\mathbf{u}|\boldsymbol{\beta},\boldsymbol{\theta},\mathbf{y}}(\mathbf{u}\mathbf{u}^\top \mid \boldsymbol{\beta}, \boldsymbol{\theta}; \mathbf{y}) \right\} = D(\boldsymbol{\theta})$, dove il valore atteso esterno è calcolata rispetto alla distribuzione marginale di $\mathbf{y}$ per $\boldsymbol{\theta}$ e $\boldsymbol{\beta}$ noti, segue che l'equazione dello score per $\hat{\boldsymbol{\theta}}_\beta$ è non distorta (poiché $h(\boldsymbol{\theta}, D(\boldsymbol{\theta})) = 0$), mentre l'equazione dello score profilo per $\hat{\boldsymbol{\theta}}$ è distorta a causa della stima di $\hat{\boldsymbol{\beta}}_{\boldsymbol{\theta}}$. È possibile quindi definire il termine di distorsione come

$$\text{bias}(\boldsymbol{\beta}, \boldsymbol{\theta}) = \mathbb{E}_{\mathbf{y}|\boldsymbol{\beta},\boldsymbol{\theta}} \left\{ \mathbb{E}_{\mathbf{u}|\boldsymbol{\beta},\boldsymbol{\theta};\mathbf{y}}(\mathbf{u}\mathbf{u}^\top \mid \hat{\boldsymbol{\beta}}_{\boldsymbol{\theta}}, \boldsymbol{\theta}; \mathbf{y}) \right\} - \mathbb{E}_{\mathbf{y}|\boldsymbol{\beta},\boldsymbol{\theta}} \left\{ \mathbb{E}_{\mathbf{u}|\boldsymbol{\beta},\boldsymbol{\theta};\mathbf{y}}(\mathbf{u}\mathbf{u}^\top \mid \boldsymbol{\beta}, \boldsymbol{\theta}; \mathbf{y}) \right\}$$

e costruire poi l'equazione di stima corretta per la distorsione

$$h(\boldsymbol{\theta}; \mathbb{E}(\mathbf{u}|\mathbf{u}|^\top \mid \hat{\boldsymbol{\beta}}_{\boldsymbol{\theta}}, \boldsymbol{\theta}; \mathbf{y}) - \text{bias}(\boldsymbol{\beta}, \boldsymbol{\theta})) = 0.$$

Infine, Liao & Lipsitz (2002) suggeriscono di sostituire $\boldsymbol{\beta}$ con $\hat{\boldsymbol{\beta}}_{\boldsymbol{\theta}}$, e quindi definire lo stimatore REML di $\boldsymbol{\theta}$ come la soluzione all'equazione di stima

$$h\{\boldsymbol{\theta}; \mathbb{E}(\mathbf{u}\mathbf{u}^\top \mid \hat{\boldsymbol{\beta}}_\theta, \boldsymbol{\theta}; \mathbf{y}) - \text{bias}(\hat{\boldsymbol{\beta}}_\theta, \boldsymbol{\theta})\} = 0. \tag{2.18}$$

Una stima degli effetti fissi può essere quindi ottenuta calcolando $\hat{\boldsymbol{\beta}}_\theta$ alla stima REML

di $\boldsymbol{\theta}$. Nonostante l'equazione di stima (2.18) non è più esattamente non distorta con la sostituzione di $\hat{\boldsymbol{\beta}}_\theta$, Liao & Lipsitz (2002) affermano che la dipendenza del termine di distorsione, bias$(\boldsymbol{\beta}, \boldsymbol{\theta})$, da $\boldsymbol{\beta}$ è trascurabile e quindi la differenza tra bias$(\boldsymbol{\beta}, \boldsymbol{\theta})$ e bias$(\hat{\boldsymbol{\beta}}_{\boldsymbol{\theta}}, \boldsymbol{\theta})$ dovrebbe essere trascurabile.

Liao & Lipsitz (2002) hanno calcolato gli stimatori di massima verosimiglianza $(\hat{\boldsymbol{\beta}}^\top, \hat{\boldsymbol{\theta}}^\top)$ e successivamente le stime REML tramite l'algoritmo EM, valutando il termine di distorsione in equazione (2.18) mediante tecniche Monte Carlo.

È importante tener presente che lo stimatore REML basato sulla correzione della distorsione sull'equazione dello *score* è probabilmente la classe computazionalmente più onerosa di stimatori REML tra quelli visti in questo capitolo (Maestrini et al., 2024).

Capitolo 3

Stima bayesiana delle componenti di varianza

Un approccio inferenziale per i modelli lineari generalizzati a effetti misti alternativo a quello frequentista è rappresentato dal paradigma bayesiano. L'inferenza bayesiana consiste nel considerare anche i parametri del modello statistico come variabili casuali e nel sintetizzare i risultati mediante la costruzione di una distribuzione di probabilità a posteriori sui parametri del modello e su quantità non osservate, come ad esempio le previsioni per nuove osservazioni (Gelman et al., 1995). Il processo di analisi dei dati secondo l'approccio bayesiano può essere idealizzato attraverso i seguenti tre passaggi fondamentali (Gelman et al., 1995):

1. specificazione un modello probabilistico completo: definizione di una distribuzione di probabilità congiunta che includa tutte le quantità osservabili e non osservabili rilevanti per il problema in esame;
2. condizionamento ai dati osservati: calcolo e interpretazione della distribuzione a posteriori dei parametri del modello, ovvero la distribuzione di probabilità delle quantità non osservate di interesse, condizionata ai dati osservati;
3. valutazione dell'adeguatezza del modello: esame della bontà del modello e delle implicazioni della distribuzione a posteriori risultante.

Nel presente capitolo verrà fornita un'introduzione alle tecniche inferenziali bayesiane applicate ai GLMM, con particolare attenzione ai metodi di calcolo della distribuzione a posteriori, che, come verrà illustrato in seguito, rappresentano una delle principali sfide nell'inferenza bayesiana.

Uno dei motivi principali che giustificano l'adozione del paradigma bayesiano risiede nella maggiore facilità di interpretazione delle conclusioni statistiche che esso offre (Gelman et al., 1995). Ad esempio, un intervallo di credibilità bayesiano per una quantità

di interesse (come un parametro) può essere interpretato direttamente come l'intervallo entro cui vi è un'elevata probabilità che si trovi il valore sconosciuto di tale quantità, a differenza dell'intervallo di confidenza frequentista, il quale deve essere interpretato rigorosamente alla luce del principio del campionamento ripetuto. Inoltre, nei modelli a effetti misti, qualora il numero di gruppi sia ridotto o il modello risulti complesso (con numerose intercette variabili, pendenze e componenti non annidate), potrebbe non esserci sufficiente informazione per stimare con precisione i parametri di varianza. In tali contesti, l'approccio bayesiano consente di ottenere un'inferenza più ragionevole poiché le distribuzioni a priori contribuiscono ad aumentare l'informazione disponibile. Inoltre tale approccio tiene conto in modo naturale dell'incertezza associata a tutti i parametri del modello (Gelman & Hill, 2007, Capitolo 16 pag. 345).

Si consideri un modello statistico $\mathcal{F} = \{p(\mathbf{y};\boldsymbol{\omega}), \mathbf{y} \in \mathcal{Y}, \boldsymbol{\omega} \in \Omega \in \mathbb{R}^p\}$, dove $\mathbf{y}$ denota i dati osservati e $\boldsymbol{\omega}$ un vettore di parametri di interesse. Il modello bayesiano, rappresentato dalla coppia $\pi(\boldsymbol{\omega})$ e $\mathcal{F}$, formalizza il processo generatore dei dati come un modello gerarchico, articolato in due stadi

$$\boldsymbol{\omega} \sim \pi(\boldsymbol{\omega})$$

$$Y|\boldsymbol{\omega} \sim p(\mathbf{y}|\boldsymbol{\omega})$$

dove $\pi(\boldsymbol{\omega})$ rappresenta la distribuzione a priori sui parametri del modello. La distribuzione a priori riflette lo stato di conoscenza o incertezza sui parametri stessi prima dell'osservazione dei dati (si veda Box & Tiao, 2011, Capitolo 1).

Il teorema di Bayes consente di ottenere la distribuzione a posteriori di $\boldsymbol{\omega}$ condizionatamente al campione osservato, $\mathbf{y}$, la cui densità è data da

$$\pi(\boldsymbol{\omega}|\mathbf{y}) = \frac{\pi(\boldsymbol{\omega})p(\mathbf{y}|\boldsymbol{\omega})}{\int_\omega \pi(\boldsymbol{\omega})p(\mathbf{y}|\boldsymbol{\omega})\, d\boldsymbol{\omega}}. \tag{3.1}$$

L'obiettivo principale dell'inferenza bayesiana consiste nell'ottenere tale distribuzione a posteriori. La quantità $\int_\Theta \pi(\boldsymbol{\omega})p(\mathbf{y}|\boldsymbol{\omega})\, d\boldsymbol{\omega}$ rappresenta la distribuzione marginale di Y, $p(\mathbf{y})$, ed è anche la costante di normalizzazione per la densità a posteriori (3.1). Tuttavia, per modelli complessi con elevato numero di parametri (p molto grande), il calcolo di $p(\mathbf{y})$ può risultare computazionalmente oneroso. Una soluzione comunemente adottata consiste nell'inferire $\boldsymbol{\omega}$ mediante un campione simulato dalla distribuzione a posteriori $\pi(\boldsymbol{\omega}|\mathbf{y})$.

Esiste una classe di metodi computazionali, nota come Markov Chain Monte Carlo (MCMC), che consente di simulare un campione dalla distribuzione a posteriori $\pi(\boldsymbol{\omega}|\mathbf{y})$ utilizzando unicamente la conoscenza del nucleo di $p(\boldsymbol{\omega}|\mathbf{y})$, ossia di $\pi(\boldsymbol{\omega})p(\mathbf{y}|\boldsymbol{\omega})$.

3.1 Campionamento Markov Chain Monte Carlo

Markov Chain Monte Carlo (MCMC) è una classe di algoritmi utilizzata per campionare da distribuzioni di probabilità complesse. La caratteristica distintiva degli algoritmi MCMC risiede nella capacità di campionare da distribuzioni note a meno della costante di normalizzazione, caratteristica che li ha resi strumenti essenziali per la diffusione pratica dell'inferenza bayesiana.

3.1.1 Nozioni di base

Si consideri un processo stocastico, ovvero una sequenza di variabili casuali,

$$X_0, X_1, X_2, ..., X_i, ...$$

con $X_i \in \mathcal{X}$. In questo contesto, i un indice discreto e $\mathcal{X} \subseteq \mathbb{R}$, il che definisce un processo stocastico a tempi discreti e a stati continui.

Il processo stocastico $X_0, X_1, X_2, ..., X_i, ...$ è una catena di Markov se è soddisfatta la seguente condizione

$$\begin{aligned} \mathbb{P}\left(X_i \in A \mid X_0 = x_0, X_1 = x_1, \ldots, X_{i-1} = x_{i-1}\right) = \\ \mathbb{P}\left(X_i \in A \mid X_{i-1} = x_{i-1}\right), \end{aligned} \tag{3.2}$$

ovvero la probabilità che la variabile casuale X_i assuma valori in A condizionatamente alla storia passata del processo dipende esclusivamente dal valore di X_{i-1}, rendendo così il processo dotato della proprietà markoviana.

È possibile descrivere il comportamento di una catena di Markov tramite un *kernel* di transizione, ovvero una funzione che determina la transizione dallo stato generico i allo stato successivo $i+1$. La densità di un *kernel* di transizione è funzione definita su $\mathcal{X}$ tale che $\forall x \in \mathcal{X}, K(x, \cdot)$ è una funzione di densità. L'equazione (3.2), espressa in termini di densità del *kernel* di transizione diventa

$$P(X_{i+1} \in A | X_0 = x_0, ..., X_k = x_i) = P(X_{i+1} \in A | X_i = x_i) = \int_A K(x_i, x)dx$$

La catena markoviana è detta omogenea nel tempo se la distribuzione di $(X_{t_1}, \ldots, X_{t_k})$ condizionata a X_{t_0} coincide con la distribuzione di $(X_{t_1-t_0}, X_{t_2-t_0}, \ldots, X_{t_k-t_0})$ condizionata a X_0, per ogni k e per ogni $t_0 \leq t_1 \leq \ldots \leq t_k$. In tal caso, se lo stato iniziale è noto, la costruzione della catena di Markov (X_n) è completamente determinata dal kernel di transizione, che definisce la distribuzione di X_n condizionatamente a X_{n-1}.

Una proprietà fondamentale delle catene markoviane è quella di invarianza o stazionarietà della distribuzione.

Definizione 1. Sia $\{X_i\}$ una catena di Markov, con *kernel* di transizione $K(\cdot,\cdot)$. Una misura di probabilità π è detta essere invariante rispetto a $K(\cdot,\cdot)$ e quindi rispetto alla catena, se

$$\pi(t) = \int_{\mathcal{X}} \pi(x)K(x,t)dx.$$

Questa proprietà garantisce l'esistenza di una distribuzione di probabilità π tale che $X_{n+1} \sim \pi$ se $X_n \sim \pi$. Gli algoritmi MCMC si fondano sul fatto che tale condizione possa essere soddisfatta. Una caratteristica della distribuzione invariante è che, sotto opportune condizioni, è anche la distribuzione limite della catena. I metodi MCMC mirano a costruire catene di Markov con distribuzione invariante π, che, nell'ambito dell'inferenza bayesiana, corrisponde alla distribuzione a posteriori $\pi(\boldsymbol{\omega}|\mathbf{y})$. Quindi, generando dalla catena di Markov si generano campioni casuali dalla distribuzione invariante. Per ulteriori dettagli sulle proprietà che le catene di Markov devono soddisfare per la costruzione degli algoritmi MCMC, si rimanda a Robert & Casella (1999).

3.1.2 Metropolis-Hastings

L'algoritmo di Metropolis-Hastings, (Metropolis et al., 1953; Hastings, 1970), rappresenta una tecnica comunemente utilizzata per generare campioni da una distribuzione di probabilità desiderata $\pi(x)$.

Sia $\pi(\mathbf{x}) : \mathbb{R}^p \to \mathbb{R}$ una funzione di densità *target* e $q(\mathbf{x}^*|\mathbf{x})$ una funzione di densità condizionata. L'algoritmo di Metropolis-Hastings, che genera una catena markoviana $\{X_i\}_{i=1}^N$ è definito nell' Algoritmo 1.

Algoritmo 1 Metropolis-Hastings

1. Inizializza $\mathbf{x}_0$
2. Per i da 1 a N, ripeti:
   a) Simulare $\mathbf{x}^*$ dalla densità condizionata $q(\cdot|\mathbf{x}_{i-1})$
   b) Calcolare la probabilità di accettazione della catena

$$\alpha = \alpha(\mathbf{x}_{i-1}, \mathbf{x}^*) = \min\left\{\frac{\pi(\mathbf{x}^*)q(\mathbf{x}_{i-1}|\mathbf{x}^*)}{\pi(\mathbf{x}_{i-1})q(\mathbf{x}^*|\mathbf{x}_{i-1})}, 1\right\}$$

   c) Si pone

$$\mathbf{x}_i = \begin{cases} \mathbf{x}^* & \text{con probabilità } \alpha \\ \mathbf{x}_{i-1} & \text{con probabilità } 1-\alpha \end{cases}$$

La distribuzione $q(\cdot|\mathbf{x}_{i-1})$ è comunemente definita come distribuzione strumentale, mentre $\alpha(\mathbf{x}_{i-1}, \mathbf{x}^*)$ rappresenta la probabilità di accettazione di Metropolis-Hastings. Si osservi che $q(\cdot|\mathbf{x}_{i-1})$ non corrisponde al *kernel* di transizione della catena, poiché la transizione da uno stato al successivo dipende anche dalla probabilità di accettazione α. Per ulteriori approfondimenti riguardanti la scelta efficiente della distribuzione strumentale $q(\cdot|\mathbf{x}_{i-1})$, si rimanda a Gelman et al. (1996). È possibile dimostrare che la catena markoviana $\{X_i\}_{i=1}^N$ generata attraverso l'Algoritmo 1 ha distribuzione invariante π (si veda Robert & Casella, 1999, Cap. 7 p. 272).

Esiste una variante dell'algoritmo di Metropolis-Hastings appena introdotto, in cui le componenti vengono aggiornate una per volta. Questa versione dell'algoritmo è particolarmente utile quando la dimensione del vettore $\mathbf{x}$ è elevata o quando è difficile definire una *proposal* che si muova in modo efficiente sull'intero spazio $\mathbb{R}^p$. In questa variante, per ogni passo dell'algoritmo si aggiorna una componente del vettore $\mathbf{x}$ alla volta, mantenendo le altre fissate ai loro valori correnti. Questa versione dell'algoritmo di Metropolis-Hastings è definita nell'Algoritmo 2. La densità $q_j(\cdot|\mathbf{x}_{i+1})$ è la distribuzione strumentale per la componente j-esima, e α_j rappresenta la probabilità di accettazione per quella componente.

Algoritmo 2 Metropolis-Hastings con aggiornamento componente per componente

1. Inizializza $\mathbf{x}_0$
2. Per i da 0 a $N-1$, ripeti:
   a) Si pone $\mathbf{x}_{i+1} = \mathbf{x}_i$
   b) Per ogni componente $j = 1, \ldots, p$:
      i) Simulare x_j^* da una densità di transizione unidimensionale $q_j(\cdot|\mathbf{x}_{i+1})$ e si pone $\mathbf{x}^* = \mathbf{x}_{i+1}$ con $x_{i+1,j}$ sostituito da x_j^*
      ii) Calcolare la probabilità di accettazione per la componente j:

$$\alpha_j = \alpha(x_{i,j}, x_j^*) = \min\left\{\frac{\pi(\mathbf{x}^*)q_j(\mathbf{x}_{i+1}|x_j^*)}{\pi(\mathbf{x}_{i+1})q_j(x_j^*|\mathbf{x}_{i+1})}, 1\right\}$$

      iii) Si pone

$$x_{i+1,j} = \begin{cases} x_j^* & \text{con probabilità } \alpha_j \\ x_{i,j} & \text{con probabilità } 1-\alpha_j \end{cases}$$

3.1.3 Gibbs-sampling

Il Gibbs-sampling (Geman & Geman, 1984) si può vedere come un caso particolare dell'algoritmo di Metropolis-Hastings, specificamente per il caso multivariato, in cui

le distribuzioni strumentali sono scelte in modo tale che la probabilità di accettazione risulti pari a uno.

Si consideri il problema di generare un campione da una distribuzione *target* multivariata π, le cui distribuzioni condizionate $\pi(x_j|\mathbf{x}_{(j)})$ sono note, dove $\mathbf{x}_{(j)}$ indica il vettore $(x_1, \ldots, x_p)$ che esclude la componente j-esima.

È possibile dimostrare che, se si scelgono come distribuzioni strumentali le distribuzioni condizionate $\pi(x_j|\mathbf{x}_{(j)})$ la probabilità di accettazione Metropolis-Hastings risulta uguale a uno.

La variante Gibbs dell'algoritmo di Metropolis-Hastings, comunemente nota come Gibbs-sampling, è descritta nell'Algoritmo 3.

Algoritmo 3 Gibbs-sampling

1. Inizializza $\mathbf{x}_0$
2. Per i da 1 a N, ripeti:
   a) Porre $\mathbf{x}^* = \mathbf{x}_{i-1}$
   b) Per j da 1 a p, ripeti:
      - Simulare $x_{i,j}$ dalla densità condizionata $q(x_j|\mathbf{x}^*_{(j)})$
      - Porre $x^*_j = x_{i,j}$
   c) Porre $\mathbf{x}_i = \mathbf{x}^*$

Gli algoritmi di Metropolis-Hastings e Gibbs-sampling hanno contribuito all'enorme diffusione dei metodi bayesiani nelle applicazioni a partire dai primi anni '90 del secolo scorso. Tuttavia, nonostante la loro generalità, non è sempre banale produrre un algoritmo MCMC che esplori in modo efficiente lo spazio dei valori della distribuzione stazionaria.

3.1.4 Hamiltonian Monte Carlo (HMC)

Il campionamento Hamiltonian Monte Carlo (HMC), introdotto in letteratura da Duane et al. (1987), è una tecnica appartenente alla famiglia dei metodi Markov Chain Monte Carlo (MCMC) che sfrutta la dinamica Hamiltoniana per esplorare lo spazio delle probabilità in modo più efficiente rispetto ai metodi MCMC tradizionali.

Dinamica Hamiltoniana

La dinamica Hamiltoniana ha origine nella fisica classica e descrive l'evoluzione temporale di un sistema fisico. Lo stato del sistema è determinato da un vettore di posizione $\boldsymbol{q}$, e dal *momentum* $\boldsymbol{p}$ (definito come il prodotto della massa e della velocità).

Per semplicità, consideriamo un sistema bidimensionale che rappresenta la dinamica di scivolamento senza attrito di un oggetto su una superficie con altezza variabile. In questo

contesto, sia $\boldsymbol{q}$ che $\boldsymbol{p}$ assumono valori in $\mathbb{R}^2$. L'energia potenziale del sistema, $U(\boldsymbol{q})$, è proporzionale all'altezza della superficie, mentre l'energia cinetica, $K(\boldsymbol{p})$, è proporzionale a $\boldsymbol{p}^T M^{-1}\boldsymbol{p}/2$, dove M è una matrice 2×2 simmetrica e definita positiva..

Grazie all'energia cinetica, l'oggetto può muoversi in salita e aumentare la sua altezza sulla superficie, fino a quando il suo *momentum* e di conseguenza la sua energia cinetica diventano nulli. A quel punto, il moto cambia direzione, iniziando una discesa che riduce l'energia potenziale e aumenta quella cinetica.

Nelle applicazioni MCMC della dinamica Hamiltoniana, la posizione $\boldsymbol{q}$ corrisponde alle variabili di interesse da cui si vuole campionare, l'energia potenziale è pari al logaritmo della densità congiunta delle variabili cambiata di segno, e il vettore $\boldsymbol{p}$ (il *momentum*) rappresenta un insieme di variabili aggiuntive introdotte artificialmente.

Questa dinamica può essere estesa al caso d dimensionale, dove sia $\boldsymbol{p}$ che $\boldsymbol{q}$ assumono valori in $\mathbb{R}^d$ e M è una matrice $d \times d$ simmetrica e definita positiva, solitamente proporzionale alla matrice identità.

Il sistema fisico è descritto da una funzione di $\boldsymbol{p}$ e $\boldsymbol{q}$ nota come funzione di Hamiltoniana. Nell'ambito dell'HMC, si utilizza generalmente una funzione Hamiltoniana della forma

$$H(\boldsymbol{q}, \boldsymbol{p}) = U(\boldsymbol{q}) + K(\boldsymbol{p}) \tag{3.3}$$

che rappresenta la somma dell'energia cinetica e potenziale del sistema. Le derivate parziali dell'Hamiltoniana, chiamate equazioni di Hamilton, determinano come variano nel tempo, t, le variabili $\boldsymbol{p}$ e $\boldsymbol{q}$

$$\frac{dq_i}{dt} = \frac{\partial H}{\partial p_i}, \tag{3.4}$$

$$\frac{dp_i}{dt} = -\frac{\partial H}{\partial q_i}, \tag{3.5}$$

per $i = 1, \ldots, d$. Data la forma di H in (3.3) e di $K(\boldsymbol{p}) = \boldsymbol{p}^T M^{-1}\boldsymbol{p}/2$ possiamo riscrivere le equazioni (3.4) e (3.5) come

$$\frac{dq_i}{dt} = [M^{-1}\boldsymbol{p}]_i,$$

$$\frac{dp_i}{dt} = -\frac{\partial U}{\partial q_i},$$

Soluzioni alle equazioni di Hamilton

Le equazioni di Hamilton devono essere approssimate utilizzando una discretizzazione del tempo, t, utilizzando un parametro ε, che deve essere scelto in modo opportuno.

Partendo dal tempo $t = 0$, gli stati del sistema vengono calcolati iterativamente ai tempi $\varepsilon, 2\varepsilon, 3\varepsilon$ e così via.

Assumiamo, per semplicità, che la matrice M sia diagonale, con elementi $m_1, m_2, \ldots, m_d$, così che l'energia cinetica possa essere espressa come

$$K(p) = \sum_{i=1}^{d} \frac{p_i^2}{2m_i}$$

L'algoritmo più comunemente usato in questo contesto per approssimare gli stati del sistema è l'algoritmo di *leapfrog*, che aggiorna le variabili del sistema come segue

$$p_i\left(t + \varepsilon/2\right) = p_i(t) - (\varepsilon/2)\,\frac{\partial U}{\partial q_i}\{q(t)\}, \tag{3.6}$$

$$q_i(t + \varepsilon) = q_i(t) + \varepsilon\frac{p_i\left(t + \varepsilon/2\right)}{m_i}, \tag{3.7}$$

$$p_i(t + \varepsilon) = p_i\left(t + \varepsilon/2\right) - (\varepsilon/2)\,\frac{\partial U}{\partial q_i}\{q(t + \varepsilon)\}. \tag{3.8}$$

Nel primo passo (3.6) le variabili di quantità di moto vengono aggiornate di un passo temporale pari a $\varepsilon/2$. Successivamente, si utilizza questo aggiornamento per fare un passo di ε unità di tempo sulle variabili di posizione (3.7), e infine si effettua un ulteriore passo di aggiornamento di $\varepsilon/2$ sulle variabili di quantità di moto, utilizzando i nuovi valori di posizione (3.8). Questo schema può essere generalizzato a qualsiasi funzione di energia cinetica sostituendo $\partial K/\partial p_i$ con p_i/m_i (Neal, 2011, Capitolo 5).

L'algoritmo HMC

La dinamica Hamiltoniana introdotta viene utilizzata per costruire un algoritmo Markov Chain Monte Carlo (MCMC). In questo contesto, si genera una catena di Markov in cui, a ogni iterazione, viene simulato un vettore di quantità di moto, $\boldsymbol{p}$, che viene impiegato per costruire una proposta tramite la dinamica Hamiltoniana.

Consideriamo una funzione di energia $E(x)$ per lo stato x, di un sistema fisico. La distribuzione canonica sugli stati (Neal, 2011, Capitolo 5) ha una funzione di densità della forma

$$P(x) = \frac{1}{Z}\exp\left(-E(x)\right).$$

dove Z è una costante di normalizzazione.

Se siamo interessati a una certa funzione di densità $P(x)$, possiamo ottenerla come distribuzione canonica ponendo $E(x) = -\log P(x) - \log Z$. L'Hamiltoniana è una

funzione di energia per lo stato congiunto di posizione, $\boldsymbol{q}$, e *momentum* $\boldsymbol{p}$, e definisce una distribuzione di probabilità congiunta per esse come

$$P(\boldsymbol{q}, \boldsymbol{p}) = \frac{1}{Z} \exp\left(-H(\boldsymbol{q}, \boldsymbol{p})\right).$$

Utilizzando l'equazione (3.3) possiamo riscrivere la funzione di densità sugli stati come

$$P(q, p) = \frac{1}{Z} \exp\left(-U(q)\right) \exp\left(-K(p)\right), \tag{3.9}$$

che mostra che $\boldsymbol{q}$ e $\boldsymbol{p}$ sono indipendenti, con funzioni di energia rispettivamente $U(\boldsymbol{q})$ e $K(\boldsymbol{p})$.

Nell'inferenza Bayesiana, la distribuzione *target* da cui si vuole campionare è la distribuzione a posteriori del modello, dove i parametri del modello fungono da variabili di posizione $\boldsymbol{q}$. L'energia potenziale può essere espressa in termini della distribuzione a posteriori come

$$U(\boldsymbol{q}) = -\log[\pi(\boldsymbol{q}) p(y|\boldsymbol{q})]$$

Si osservi che, per come è stata definita l'energia cinetica, $K(\boldsymbol{p})$, la distribuzione canonica sugli stati $\boldsymbol{p}$ è una gaussiana centrata in zero con matrice di varianza e covarianza M. L'algoritmo HMC campiona $\boldsymbol{p}$ e $\boldsymbol{q}$ dalla distribuzione canonica definita in (3.9), in cui $\boldsymbol{q}$ ha la distribuzione di interesse (la distribuzione a posteriori) e il vettore *momentum*, $\boldsymbol{p}$, è distribuito secondo una gaussiana multivariata.

Ogni iterazione dell'algoritmo HMC può essere suddivisa nei seguenti passi:

1. Si campiona un valore per le variabili *momentum* dalla distribuzione gaussiana.
2. Si effettua un passo di aggiornamento Metropolis utilizzando la dinamica Hamiltoniana per proporre un nuovo stato per le variabili $\boldsymbol{q}$. A partire dallo stato corrente $(\boldsymbol{p}, \boldsymbol{q})$, si simulano L passi di ampiezza ε tramite l'algoritmo di *leapfrog*. I parametri L e ε sono parametri dell'algoritmo e devono essere scelti in modo opportuno. Alla fine, il vettore $\boldsymbol{p}$ viene cambiato di segno per ottenere una nuova proposta $(\boldsymbol{p}^*, \boldsymbol{q}^*)$. Il nuovo stato proposto, $(\boldsymbol{p}^*, \boldsymbol{q}^*)$, viene accettato con probabilità di accettazione Metropolis definita come

$$\begin{aligned} \alpha &= \min\left[1, \exp(-H(\boldsymbol{q}^*, \boldsymbol{p}^*) + H(\boldsymbol{q}, \boldsymbol{p}))\right] = \\ &= \min\left[1,\ \exp(-U(\boldsymbol{q}^*) + U(\boldsymbol{q}) - K(\boldsymbol{p}^*) + K(\boldsymbol{p}))\right]. \end{aligned}$$

Il cambiamento di segno del vettore $\boldsymbol{p}$ rende la proposta simmetrica, garantendo la validità della probabilità di accettazione Metropolis.

L'implementazione dell'algoritmo HMC è sintetizzata nell'Algoritmo 4.

Algoritmo 4 HMC

1. Inizializza M, $\boldsymbol{q}_1, \boldsymbol{p}_1, \epsilon, L, N$
2. Per i da 2 a N, ripeti:
   a) Campionare $\boldsymbol{p}' \sim \mathcal{N}_d(0, M)$
   b) Porre $(\boldsymbol{q}^*, \boldsymbol{p}^*) \leftarrow \text{Leapfrog}((\boldsymbol{q}_{i-1}, \boldsymbol{p}'), U(\boldsymbol{q}), M, \epsilon, L)$ †
   c) Calcolare
   $$\alpha = \min\left[1, \exp(-H(\boldsymbol{q}^*, \boldsymbol{p}^*) + H(\boldsymbol{q}_{i-1}, \boldsymbol{p}_{i-1}))\right]$$
   d) Porre
   $$(\boldsymbol{q}_i, \boldsymbol{p}_i) = \begin{cases} (\boldsymbol{q}^*, \boldsymbol{p}^*) & \text{con probabilità } \alpha \\ (\boldsymbol{q}_{i-1}, \boldsymbol{p}_{i-1}) & \text{con probabilità } 1-\alpha \end{cases}$$

† dove con *Leapfrog* si intende la simulazione di un sistema Hamiltoniano mediante l'algoritmo di *leapfrog*.

Per approfondimenti circa le proprietà di convergenza dell'algoritmo si veda Neal (2011).

3.1.5 Diagnostiche di convergenza

Uno dei principali problemi nell'impiego delle tecniche di campionamento MCMC per l'inferenza bayesiana riguarda la verifica della convergenza del campione simulato alla distribuzione *target*. Due aspetti centrali nello sviluppo degli algoritmi MCMC, cruciali per garantire il successo della convergenza alla distribuzione stazionaria, consistono nella scelta del punto di partenza e nella definizione del criterio di arresto dell'algoritmo. Questi aspetti sono strettamente legati alla convergenza della catena di Markov alla distribuzione stazionaria e alla convergenza degli stimatori Monte Carlo ai valori attesi della popolazione (Roy, 2020). Come affermato nel paragrafo 3.1.2, sotto alcune semplici condizioni, la distribuzione di X_n converge alla distribuzione stazionaria della catena per $n \to \infty$, indipendentemente dal valore iniziale. Poiché X_0 generalmente non segue la distribuzione π e l'aggiornamento di X_n dipende da X_{n-1}, generando così campioni correlati, quanto più X_0 è distante da π, tanto maggiore sarà il tempo necessario affinché X_n si avvicini alla distribuzione *target*. In particolare, se il valore iniziale X_0 non appartiene a una regione ad alta densità di π, i campioni generati nelle prime iterazioni potrebbero risultare lontani dalla distribuzione desiderata. In tali circostanze, è prassi comune scartare le prime realizzazioni della catena, iniziando a considerare i campioni solo dopo che l'effetto del valore iniziale si è (idealmente) esaurito. Questa tecnica, denominata *burn-in* o *warm-up*, mira a utilizzare esclusivamente i campioni generati quando la catena di Markov è sufficientemente vicina alla distribuzione stazionaria. Per

ottenere un campione il più possibile rappresentativo della distribuzione stazionaria, sarebbe preferibile inizializzare l'algoritmo in una regione ad alta densità di π. Tuttavia, poiché in molti casi è complesso identificare con precisione un valore iniziale che soddisfi tale requisito, può essere necessario scartare una parte iniziale dei campioni generati e avviare la raccolta solo dopo un determinato numero di iterazioni n'.

Una volta stabilito il valore iniziale, è necessario determinare il momento in cui interrompere la simulazione. In tal senso la letteratura esistente si concentra principalmente su due approcci distinti (Cowles & Carlin, 1996). Il primo approccio è di natura teorica e si basa sullo studio del *kernel* di transizione della catena di Markov. Tale studio mira a determinare il numero di transizioni necessarie affinché sia garantita la convergenza alla distribuzione *target*. Tuttavia, questo approccio richiede spesso l'impiego di strumenti matematici avanzati e calcoli complessi che devono essere ripetuti per ogni modello considerato. Di conseguenza risulta difficilmente applicabile nella pratica e viene utilizzato raramente in contesti applicativi. La maggior parte dei lavori pratici, invece, si basa su un secondo approccio, che consiste nell'applicazione di strumenti diagnostici all'output prodotto dall'algoritmo MCMC. Questi strumenti consentono di valutare empiricamente se la catena ha raggiunto la distribuzione stazionaria di interesse, fornendo indicazioni utili su quando interrompere il campionamento. D'altra parte, gli strumenti di diagnostica di convergenza delle catene di Markov sono stati oggetto di critiche da parte di studiosi teorici, i quali sottolineano come tali metodi possano talvolta generare falsi segnali di convergenza. Nonostante queste limitazioni, molti statistici continuano a fare ampio affidamento su tali strumenti, ritenendo che, sebbene non privi di difetti, uno strumento di diagnostica anche imperfetto sia preferibile all'assenza di qualsiasi valutazione della convergenza (Cowles & Carlin, 1996).

A partire dai primi anni '90, con l'uso crescente degli algoritmi MCMC, vi è stato un grande sforzo da parte della comunità scientifica al fine di sviluppare strumenti di diagnostica di convergenza. Tali metodi possono essere classificati in diverse categorie. Ad esempio, alcuni di questi strumenti sono progettati per valutare la convergenza della catena di Markov in relazione alla distribuzione stazionaria, mentre altri verificano la convergenza di statistiche riassuntive, come la media campionaria e quantili campionari, rispetto ai corrispondenti valori attesi di popolazione (Roy, 2020). Gli strumenti di diagnostica di convergenza per gli algoritmi MCMC possono essere inoltre classificati secondo diversi criteri, ad esempio secondo il livello di fondamento teorico, alla loro capacità di diagnosticare la convergenza congiunta di più variabili, alla possibilità di utilizzarli con più catene parallele, e al loro supporto di strumenti di visualizzazione.

Nel seguito vengono descritti alcuni strumenti diagnostici per MCMC che possono

essere impiegati per determinare la convergenza della catena di Markov o per interrompere il campionamento MCMC.

Diagnostica di Gelman-Rubin

Lo strumento diagnostico di Gelman e Rubin (GR) è uno tra i metodi più popolari per analizzare i campioni ottenuti dagli algoritmi MCMC. Lo strumento diagnostico GR utilizza più catene in parallelo $\{X_{i0}, X_{i1}, \ldots, X_{in-1}\}$ per $i = 1, \ldots, m$, che vengono inizializzate da punti iniziali provenienti da una densità sovradispersa rispetto alla densità *target* π. Gelman & Rubin (1992) propongono dei metodi per creare una distribuzione iniziale sovradispersa, tuttavia nella pratica l'individuazione di tale distribuzione dipende dal problema che si affronta. Mediante l'utilizzo di catene parallele, gli autori propongono di costruire due stimatori della varianza di X, dove $X \sim \pi$, quali

$$W = \sum_{i=1}^{m} \sum_{j=0}^{n-1} \frac{(X_{ij} - \bar{X}_i)^2}{m(n-1)},$$

che rappresenta una stima della varianza *within* delle catene e

$$\hat{V} = \frac{(n-1)}{n} W + \frac{B}{n},$$

che rappresenta una stima della varianza totale, dove $\frac{B}{n} = \frac{1}{m-1} \sum_{i=1}^{m} \left(\bar{X}_i - \bar{X}\right)^2$ è una stima della varianza *between* tra le catene e $\bar{X}_i$ e $\bar{X}$ sono la media della i-esima catena e la media complessiva rispettivamente, con $i = 1, 2, \ldots, m$.

Infine, gli autori definiscono il *potential scale reduction factor* (PSRF) come

$$\hat{R} = \frac{\hat{V}}{W} \tag{3.10}$$

Poiché le catene hanno valori iniziali generati da una distribuzione sovradispersa, il numeratore in (3.10) sovrastima la variabilità *target* mentre il denominatore la sottostima, rendendo $\hat{R}$ maggiore di 1. Per tale motivo la simulazione dovrebbe essere arrestata quando $\hat{R}$ è sufficientemente vicino a 1. Generalmente viene utilizzato il valore soglia di 1.1, come raccomandato da Gelman et al. (2013). Si osservi che la definizione di $\hat{R}$ in (3.10) differisce leggermente da quella proposta originariamente da Gelman & Rubin (1992), tuttavia si è scelto di riportare tale definizione in quanto ampiamente utilizzata nelle applicazioni pratiche.

Questa metodologia diagnostica proposta consente di valutare se i campioni simulati per quantità unidimensionali abbiano raggiunto la distribuzione stazionaria π. Tuttavia,

non offre una visione complessiva sulla convergenza di tutte le quantità campionate dall'algoritmo MCMC. Al fine di valutare la convergenza congiunta di tutte le quantità di interesse, Brooks & Gelman (1998) propongono il PSRF multivariato (MPSRF), che permette di diagnosticare la convergenza dei campioni nel caso multivariato. Tale fattore è definito come

$$\hat{R}_p = \max_{\mathbf{a}} \frac{\mathbf{a}^T \hat{V}^* \mathbf{a}}{\mathbf{a}^T W^* \mathbf{a}} = \frac{n-1}{n} + \left(1 + \frac{1}{m}\right) \lambda_1,$$

dove $\hat{V}^*$ è la matrice di covarianza totale, W^* è la matrice di covarianza *within* e B^* è la matrice di covarianza *between* tra le catene, e λ_1 è il più grande autovalore della matrice $(W^{*-1}B^*)/n$. Come nel caso univariato, la simulazione si interrompe quando $\hat{R}_p \approx 1$.

Gelman & Rubin (1992) suggeriscono di campionare m catene parallele, ciascuna delle quali di lunghezza $2n$. Successivamente, propongono di scartare le prime n simulazioni per ciascuna delle m catene, e di calcolare l'$\hat{R}$ sulle ultime n iterazioni. Tuttavia, questa procedura comporta l'esclusione di un elevato numero di osservazioni per ciascuna catena. Per questo motivo, tale approccio non è universalmente accettato dalla comunità scientifica. Ad esempio, Roy (2020) suggeriscono di non adottare questa metodologia.

Effective sample size

Anche se raggiunta la condizione di stazionarietà, vi è una chiara differenza nell'utilizzo della media empirica rispetto alla stima standard di Monte Carlo basata su un campione di osservazioni indipendenti ed identicamente distribuite (i.i.d.). Infatti, usando un campione ottenuto mediante un algoritmo MCMC, non è possibile associare alla media empirica $S_n = \frac{1}{n}\sum_{t=1}^{n} h(\omega^{(t)})$ come stima di

$$\int h(\omega) f(\omega) d\omega,$$

il classico stimatore della varianza

$$\hat{\nu}_n = \frac{1}{n^2} \sum_{t=1}^{n} \left(h(\omega^{(t)}) - S_n\right)^2$$

a causa della correlazione tra i valori $\omega^{(t)}$. Infatti, tale stima della varianza tende a sottostimare la vera varianza dello stimatore S_n. Una soluzione, seppur approssimativa, consiste nell'usare l'*effective sample size* (ESS), $\hat{n}^S$, che rappresenta il numero di osservazioni indipendenti equivalenti a un campione con osservazioni correlate, in cui l'equivalenza è definita in termini di *standar error*. Tale misura fornisce un'indicazione della perdita di efficienza dovuta all'uso di una catena di Markov rispetto a un campione

i.i.d. Questo valore è calcolato come

$$\hat{n}^S = \frac{n}{1 + 2\sum_{i=1}^{\infty} \text{Corr}_\pi(h(\omega_0), h(\omega_i))} = \frac{n}{\kappa(h)},$$

dove $\kappa(h)$ rappresenta il tempo di autocorrelazione associato alla sequenza $h(\omega^{(t)})$, Sostituendo n con $\hat{n}^S$ si ottiene una stima più affidabile della varianza di S_n (Robert & Casella, 1999),

$$\hat{\nu}_n = \frac{1}{n \times \hat{n}^S} \sum_{t=1}^{n} \left(h(\omega^{(n)}) - S_n \right)^2.$$

L'ESS viene comunemente utilizzato per determinare la numerosità delle catene di Markov al fine di raggiungere un livello di variabilità desiderato per le statistica di interesse. Può essere inoltre impiegato nel confronto di diversi algoritmi MCMC (che hanno la stessa distribuzione stazionaria) sia in termini di efficienza computazionale che statistica. L'ESS è implementato in diversi pacchetti R, tra cui `coda` (Plummer et al., 2006) e `mcmcse` (Flegal et al., 2012).

Strumenti grafici

Assieme ai metodi di diagnostica quantitativi già presentati, è possibile ricorrere anche a strumenti grafici per valutare il comportamento delle catene di Markov. Questi strumenti offrono un approccio qualitativo per analizzare la dinamica delle catene e individuare eventuali problemi legati alla convergenza o all'elevata autocorrelazione.

Lo strumento grafico di diagnostica della convergenza più comune è il *trace plot*. Il *trace plot* permette di visualizzare l'andamento della catena nello spazio degli stati. Questo metodo viene utilizzato per osservare come la catena di Markov esplora il supporto della distribuzione *target*; più velocemente la catena riesce a esplorare il supporto della distribuzione *target*, tanto migliore sarà il campionamento risultante. Se la catena MCMC rimane bloccata in una regione dello spazio degli stati, il *trace plot* mostra andamenti piatti, sintomo di una lenta convergenza alla distribuzione stazionaria. Questo fenomeno può verificarsi, ad esempio, quando la distribuzione *proposal* ha una bassa probabilità media di accettazione e quando molte proposte vengono rifiutate consecutivamente. D'altra parte, se il probabilità media di accettazione è troppo elevata, la catena potrebbe accettare troppe proposte consecutive, muovendosi lentamente senza esplorare in modo efficace l'intero spazio degli stati. Inoltre, tendenze o variazioni evidenti nel *trace plot* suggeriscono che la stazionarietà non è ancora stata raggiunta. Come mostrato in Figura 3.1 (a), un *trace plot* con tendenze visibili evidenzia problemi di convergenza, mentre la Figura 3.1 (b) fornisce un esempio di buon andamento (*mixing*) della catena.

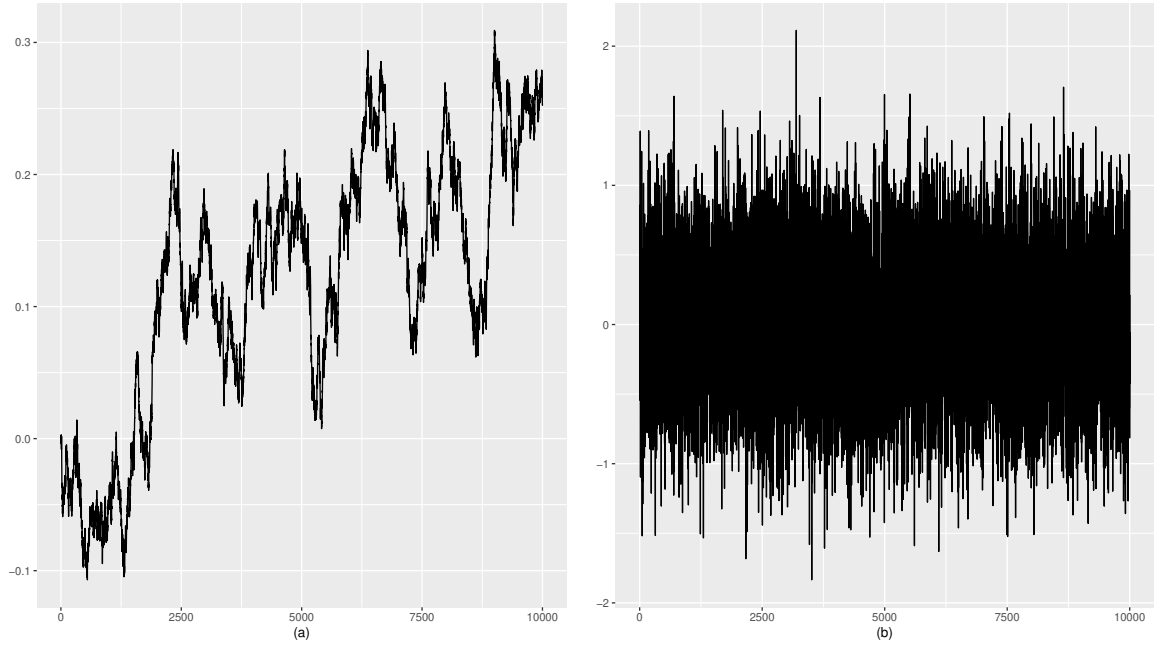

FIGURA 3.1: Esempio di *trace plot* di una catena non convergente (a) e di una catena convergente (b)

A differenza del campionamento i.i.d., gli algoritmi MCMC generano campioni correlati. Quanto maggiore è la correlazione tra le osservazioni del campione generato, tanto più lentamente la catena riuscirà a esplorare il supporto della distribuzione *target*. Uno strumento di diagnostica che permette di valutare l'autocorrelazione della catena è il grafico dell'autocorrelazione (ACF). Questo grafico mostra i valori della funzione di autocorrelazione ai *lag* k per valori crescenti di k. Una catena di Markov costruita in modo opportuno dovrebbe avere valori di autocorrelazione che scendono vicino a zero il più rapidamente possibile all'aumentare di k. Chiaramente un ACF che mostra un'alta autocorrelazione a *lag* elevati implica un basso ESS.

Un altro metodo grafico utilizzato in pratica è il grafico che mostra come varia media mobile al crescere del numero di iterazioni della catena. In questo grafico la media mobile dovrebbe auspicabilmente stabilizzarsi su un valore fisso al crescere del numero di iterazioni, e la mancata convergenza ad un valore stabile è un indicatore della mancata convergenza della catena alla distribuzione stazionaria. Tale grafico viene comunemente utilizzato per individuare il numero di iterazioni necessarie per arrestare l'algoritmo.

Nel caso multivariato, i singoli grafici discussi in questa sezione vengono costruiti sulla base di ciascuna catena marginale, sperando che un buon comportamento delle marginali implichi anche un buon comportamento della distribuzione congiunta. È importante sottolineare che, sebbene sia gli strumenti diagnostici quantitativi che quelli grafici siano molto utili per individuare i casi in cui la convergenza non è stata raggiunta,

essi difficilmente permettono di affermare con assoluta certezza che la convergenza sia stata ottenuta.

3.2 MCMC nei GLMM

La maggior parte degli algoritmi di campionamento per i modelli lineari generalizzati (GLM) si basa sull'introduzione di variabili latenti (*data augmentation*), che riconducono il GLM a un modello lineare. Questo approccio consente l'applicazione di algoritmi noti per i modelli lineari con errori gaussiani. Tale idea è stata introdotta da Albert & Chib (1993) per la regressione probit e successivamente estesa alla regressione logistica da Held & Holmes (2006). Ulteriori approcci sono stati proposti da Frühwirth-Schnatter et al. (2009) e Polson et al. (2013). In particolare, Polson et al. (2013) hanno sviluppato uno schema di data augmentation per i modelli di regressione logistica, basato sulla distribuzione di Pólya-Gamma.

Una volta ricondotto il modello lineare generalizzato a un modello lineare normale, è possibile utilizzare schemi di campionamento adatti a quest'ultimo. Si veda, ad esempio, Wang et al. (1993), che hanno implementato uno schema di campionamento Gibbs per modelli lineari a effetti misti.

È importante sottolineare che nel contesto dell'inferenza bayesiana non vi è una così netta distinzione tra modelli a effetti misti e modelli a effetti fissi come avviene nell'ambito dell'inferenza frequentista. Questo perchè nell'inferenza bayesiana tutti i parametri sono casuali e ciò che distingue le due tipologie di modelli è il livello di gerarchia adottato nella specificazione delle distribuzioni a priori. Infatti in tale contesto i modelli a effetti casuali vengono comunemente chiamati modelli gerarchici. Di conseguenza, gli schemi di campionamento per i modelli a effetti fissi possono essere facilmente adattati anche ai modelli a effetti misti.

Sono disponibili diverse implementazioni di algoritmi di campionamento di tipo Gibbs (o Metropolis within Gibbs), come ad esempio il pacchetto `MCMCglmm` per R (Hadfield, 2010) o il software di analisi bayesiana JAGS (Hornik et al., 2003).

Un'alternativa agli schemi di campionamento Gibbs è rappresentata dal campionamento Hamiltoniano Monte Carlo (HMC), descritto nel paragrafo 3.1.4. Un'implementazione di HMC per l'analisi bayesiana è disponibile nel linguaggio di programmazione probabilistica Stan (Stan Development Team, 2018a), il quale offre anche un'interfaccia per R (Stan Development Team, 2018b). Per ulteriori approfondimenti, si veda Carpenter et al. (2017).

L'uso dell'Hamiltonian Monte Carlo risulta particolarmente vantaggioso per la stima bayesiana dei GLMM, in quanto sfrutta una *proposal* che tiene conto della forma della distribuzione a posteriori del modello. Ciò risulta fondamentale poiché gli schemi di tipo Gibbs per questa classe di modelli tendono a generare campioni con elevata autocorrelazione. Inoltre, un grosso vantaggio di Stan è che ha un modo automatico per il *tuning* dei parametri dell'algoritmo. In questo modo, l'utente finale può dedicarsi maggiormente alla scelta del modello statistico piuttosto che agli aspetti algoritmici e computazionali per la sua stima. Un altro vantaggio di Stan è che compila in c++ e quindi tipicamente è più veloce di altro codice scritto esclusivamente in R.

3.2.1 Specificazione del modello

Analogamente al GLMM frequentista in (1.4) possiamo specificare la sua controparte bayesiana come segue

$$
\begin{aligned}
&\mathbf{y}_i \mid \boldsymbol{\beta}, \mathbf{u}_i, \phi, \boldsymbol{\theta} \sim p(\mathbf{y}_i; \boldsymbol{\beta}, \mathbf{u}_i, \phi, \boldsymbol{\theta}),\\
&g(\boldsymbol{\mu}_i) = \boldsymbol{\eta}_i = X_i^\top \boldsymbol{\beta} + Z_i^\top \mathbf{u}_i,\\
&\mathbf{u}_i \mid \boldsymbol{\theta} \sim N(\mathbf{0}, D(\boldsymbol{\theta})),\\
&\boldsymbol{\beta} \sim \pi_{\boldsymbol{\beta}}(\boldsymbol{\beta}),\\
&\boldsymbol{\theta} \sim \pi_{\boldsymbol{\theta}}(\boldsymbol{\theta}),\\
&\phi \sim \pi_{\phi}(\phi),
\end{aligned}
$$

dove $\pi_{\boldsymbol{\beta}}$, $\pi_{\boldsymbol{\theta}}$ e π_{ϕ} sono le distribuzioni a priori per $\boldsymbol{\beta}$, $\boldsymbol{\theta}$ e ϕ rispettivamente, $\mathbf{y}_i$ è la variabile risposta per l'i-esimo gruppo con $i = 1, \ldots, K$ e θ sono i parametri di varianza (e covarianza) del vettore casuale $\mathbf{u}_i$.

Per quanto concerne la distribuzione a priori per $\boldsymbol{\beta}$, è comune specificare una distribuzione normale o una distribuzione non informativa costante ($\pi_{\boldsymbol{\beta}}(\boldsymbol{\beta}) \propto 1$). Ad esempio, il pacchetto `brms` utilizza di *default* per i parametri fissi $\boldsymbol{\beta}$ la distribuzione non informativa costante $\pi_{\boldsymbol{\beta}}(\boldsymbol{\beta}) \propto 1$.

3.2.2 Distribuzione a priori per le componenti di varianza

La specificazione delle distribuzioni a priori per le componenti di varianza dei GLMM rappresenta un aspetto delicato, poiché tali parametri sono tra i più complessi da stimare. In alternativa alla specificazione di una distribuzione a priori per i singoli parametri $\boldsymbol{\theta}$, è possibile definire una distribuzione a priori direttamente sulla matrice di varianza e covarianza degli effetti casuali $D(\boldsymbol{\theta})$. Gelman & Hill (2007) suggerisce l'uso di una

distribuzione Wishart inversa per $D(\boldsymbol{\theta})$, poiché nel modello lineare normale è coniugata alla funzione di verosimiglianza, facilitando il campionamento di tipo Gibbs.

In Stan, non è necessario utilizzare distribuzioni a priori che rendano note le *full-conditional* del modello. Infatti, Stan Development Team (2012) raccomandano di decomporre la matrice di varianza e covarianza degli effetti casuali come segue

$$D(\boldsymbol{\theta}) = \text{diag}(\boldsymbol{\tau})\Phi\text{diag}(\boldsymbol{\tau}),$$

dove Φ è una matrice di correlazione e $\boldsymbol{\tau}$ è il vettore dei parametri di scala. Le componenti del vettore $\boldsymbol{\tau}$ e della matrice Φ possono essere ottenute come segue

$$\tau_k = \sqrt{D(\boldsymbol{\theta})_{k,k}}$$
$$\Phi i, j = \frac{D(\boldsymbol{\theta})_{i,j}}{\tau_i \tau_j}.$$

Per il vettore dei parametri di scala $\boldsymbol{\tau}$ si possono specificare distribuzioni a priori appropriate. Stan Development Team (2012) raccomandano una distribuzione a priori poco informativa, come una distribuzione half-Cauchy (Cauchy troncata) con parametro di scala piccolo, ad esempio

$$\tau_k \sim \text{Cauchy}(0, 2.5) \quad \text{per } k \in 1 : K \text{ e con il vincolo } \tau_k > 0.$$

Nel pacchetto `brms` (Bürkner, 2017), si utilizza invece una distribuzione half-t come suggerito da Gelman (2006). Inoltre, Stan Development Team (2012) raccomandano l'uso della distribuzione LKJ per la matrice di correlazione Φ (Lewandowski et al., 2009)

$$\Phi \sim \text{LKJCorr}(\eta).$$

La distribuzione LKJ per la matrice di correlazione è definita come

$$\text{LkjCorr}(\Phi \mid \eta) \propto \det(\Phi)^{\eta-1}.$$

con $\eta > 0$.

Il comportamento di base della distribuzione di correlazione LKJ è simile a quello di una distribuzione beta. Per $\eta = 1$, il risultato è una distribuzione uniforme. Per $\eta > 1$, la densità concentra sempre più massa intorno alla matrice identità, cioè favorendo meno correlazione. Per $\eta < 1$, la densità concentra sempre più massa nella direzione opposta, cioè favorendo una maggiore correlazione. Per una rappresentazione grafica della distribuzione LKJ per il parametro di correlazione in una matrice di correlazione

2×2 al variare del parametro η si veda la Figura 3.2. In generale si suggerisce di utilizzare $\eta \geq 1$. Per una discussione più approfondita sulla decomposizione della matrice di covarianza al fine di specificare distribuzioni a priori sulla matrice di correlazione e sui parametri di scala, si veda Barnard et al. (2000).

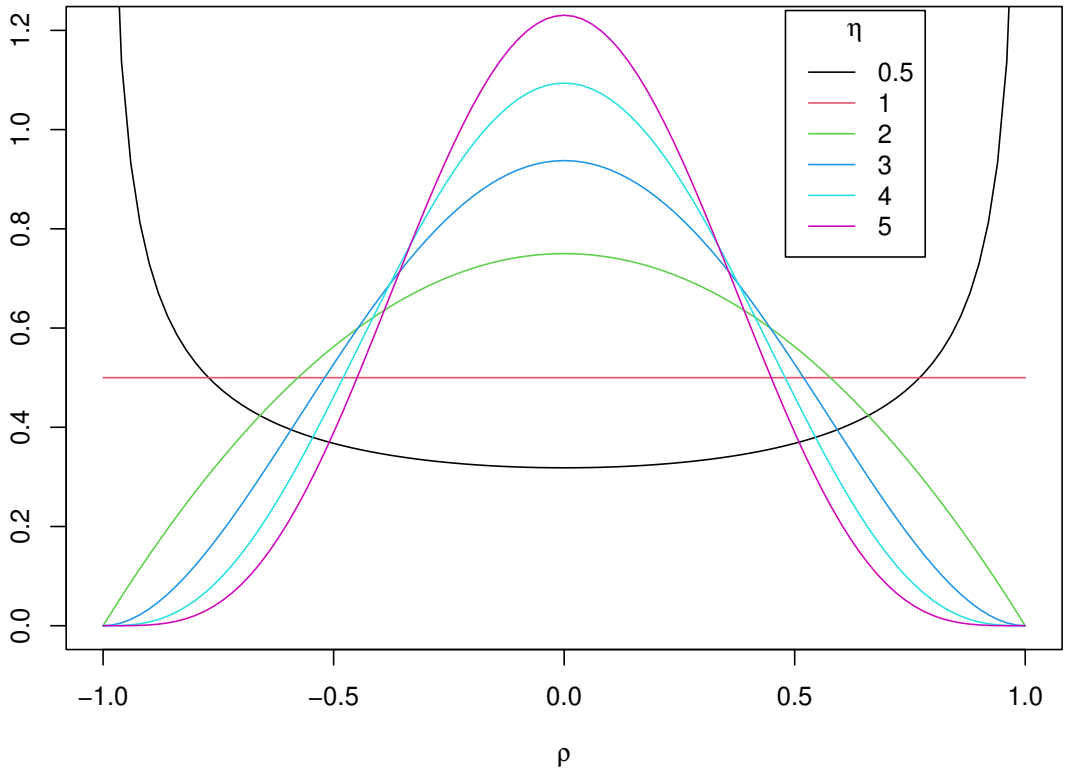

FIGURA 3.2: Distribuzione LKJ per il coefficiente di correlazione.

Capitolo 4

Studi di simulazione

4.1 Introduzione

L'obiettivo di questo capitolo è valutare, attraverso studi di simulazione, le proprietà frequentiste delle tecniche inferenziali bayesiane nel contesto dei modelli lineari generalizzati a effetti misti (GLMM). In particolare, si intende confrontare la distorsione degli stimatori e la copertura degli intervalli di credibilità bayesiani rispetto a quelli frequentisti, in particolare quelli descritti nel Capitolo 2.

Per quanto riguarda le stime frequentiste, si è deciso di includere nello studio di simulazione alcuni dei metodi utilizzati in Maestrini et al. (2024). Questi includono: 1) linearizzazione approssimata tramite la stima di quasi-verosimiglianza, ottenuta modificando la funzione `glmmPQL` della libreria `MASS` di R (Ripley, 2002) per implementare le equazioni (2.10) e (2.11); 2) la stima di verosimiglianza integrata utilizzando la libreria `glmmTMB`, con la componente REML, che sfrutta l'approssimazione di Laplace descritta nell'equazione (2.14); 3) l'approccio basato sulla verosimiglianza profilo modificata, descritto dall'equazione (2.15). Analogamente a quanto fatto da Maestrini et al. (2024) per ciascuno di questi metodi, è stata inoltre stimata la versione con la massima verosimiglianza non ristretta. Utilizzando l'opzione ML predefinita sia per `glmmPQL` che per `glmmTMB`, ed il codice disponibile per la stima della massima verosimiglianza non ristretta fornito da Bellio & Brazzale (2011). Per motivi legati ai costi computazionali, non è stata adottata la tecnica di stima basata sulla correzione della distorsione nella funzione punteggio. Per i risultati relativi a questa specifica metodologia, si rimanda a Maestrini et al. (2024).

Nel presentare i risultati della simulazione, sono stati utilizzati gli acronimi "PQL", "TMB", "MPL" per indicare i metodi di stima di massima verosimiglianza non ristretta e "PQL REML", "TMB REML", "MPL REML" per indicare i tre metodi REML.

Per l'inferenza bayesiana è stato impiegato il pacchetto `rstan` di R, un'interfaccia R con il software di programmazione probabilistica Stan (Carpenter et al., 2017). La scelta di utilizzare Stan è motivata dal suo ampio utilizzo nell'ultimo decennio, il che lo ha reso un metodo di riferimento nella comunità scientifica. Inoltre, Stan offre la possibilità di campionare automaticamente da distribuzioni complesse, richiedendo un'interazione minima da parte dell'utente, il che lo rende particolarmente efficiente e versatile per la stima di modelli complessi quali i GLMM. Si veda l'Appendice A.1 per un esempio di codice stan per la modellazione di dati binari attraverso un modello logistico a effetti casuali e l'Appendice A.2 per un esempio di codice per la modellazione di dati di conteggio attraverso un modello di Poisson a effetti casuali.

Per validare le stime ottenute tramite campionamento HMC in termini di convergenza delle catene di Markov alla distribuzione a posteriori, si sono utilizzati strumenti standard per la diagnostica di convergenza. In particolare, sono stati adottati i metodi descritti nel paragrafo 3.1.5, mediante campionamento di quattro catene parallele per ciascuna iterazione della simulazione, ognuna composta da 10000 iterazioni, delle quali le prime 5000 sono state scartate come fase di *burn-in*. Si è deciso di considerare un campione affidabile se la statistica di Gelman-Rubin (PSRF) soddisfa il criterio $\hat{R} < 1.01$ e se l'effective sample size è maggiore di 1000.

4.2 Struttura dello Studio di Simulazione

In linea con quanto proposto da Maestrini et al. (2024), è stato sviluppato uno studio di simulazione articolato in due sezioni: una prima dedicata ai modelli GLMM per dati binari e una seconda rivolta ai modelli GLMM Poisson per dati di conteggio.

4.2.1 Dati di binari

Lo studio di simulazione per dati di binari si articola di quattro scenari. Per ciascun scenario, sono stati generati 500 campioni su cui è stata effettuata la stima dei parametri mediante i metodi di stima precedentemente citati. La generazione dei dati binari si basa sul seguente modello logistico a effetti misti:

$$Y_{ij}|\mathbf{u}_i \sim \text{Bernoulli}\left(\frac{1}{1+\exp\left(-(\boldsymbol{x_{ij}}^T\boldsymbol{\beta}+\boldsymbol{z_{ij}}^T\boldsymbol{u_i})\right)}\right),$$

con $\boldsymbol{u_i} \sim \mathcal{N}(0,\Sigma)$, per $i = 1,\ldots,50$ *cluster* e $j = 1,\ldots,10$ osservazioni per ciascun *cluster*.

Per questo studio di simulazione sono stati considerati quattro scenari distinti:

- Nel primo scenario sono stati considerati quattro effetti fissi definiti dal vettore $\beta = (\beta_0, \beta_1, \beta_2, \beta_3)^T = (0.5, 1, -1, -0.5)^T$, con associate covariate $x_{ij} = (1, x_{ij,1}, x_{ij,2}, x_{ij,3})^T$ tali che: $x_{ij,1} = (j-5)/4$, $x_{ij,2} = 0$ per $i = 1, \ldots, 25$ e $x_{ij,2} = 1$ per $i = 26, \ldots, 50$, e $x_{ij,3} = x_{ij,1}x_{ij,2}$. Le covariate degli effetti casuali sono $z_{ij} = (1, x_{ij,1})^T$, con matrice di covarianza diagonale Σ in cui $\Sigma_{11} = \Sigma_{22} = 0.5$.
- Il secondo scenario è simile al primo, ma con l'aggiunta di quattro effetti fissi ($\beta_4 = \beta_5 = \beta_6 = \beta_7 = 0$) cui le corrispondenti covariate $x_{ij,l} = x_{i,l}$ per $l = 4, \ldots, 7$ sono generate, indipendentemente dalla variabile risposta, dalla distribuzione normale standard.
- Il terzo scenario è molto simile al secondo, ma con l'aggiunta di dieci covariate generate indipendentemente dalla risposta, anziché quattro.
- Il quarto scenario è simile al primo, con la differenza che la matrice di covarianza degli effetti casuali ha elementi non diagonali $\Sigma_{12} = \Sigma_{21} = 0.25$.

Il terzo scenario risulta il più complesso per la stima delle componenti di varianza mediante metodi di massima verosimiglianza non ristretta a causa del maggior numero di effetti fissi che devono essere stimati

Per l'inferenza bayesiana sono state adottate distribuzioni a priori non informative per i parametri fissi del modello ($\pi_{\boldsymbol{\beta}}(\boldsymbol{\beta}) \propto 1$). Per i parametri di varianza, è stata utilizzata una distribuzione half-t con tre gradi di libertà, si è posto il parametro di scala $\sigma = 2.5$, valore di default nel pacchetto `brms` (Bürkner, 2017) per i modelli considerati. Sono stati considerati anche altri valori per il parametro di scala σ, come $\sigma = 5$ e $\sigma = 3$; tuttavia, poiché non hanno prodotto risultati particolarmente diversi rispetto al caso con $\sigma = 2.5$, si è deciso di non riportare tali risultati nella presente analisi. Per il parametro di covarianza tra gli effetti casuali si è riparametrizzato il modello e specificata la distribuzione a priori LKJ sulla matrice di correlazione come nel paragrafo 3.2.2, con iper parametro $\eta = 1$.

4.2.2 Dati di conteggio

Lo studio di simulazione per dati di conteggio si articola di quattro scenari e si basa sul seguente modello di Poisson a effetti misti:

$$Y_{ij}|\mathbf{u}_i \sim \text{Poisson}\left(\exp(\boldsymbol{x_{ij}}^T\boldsymbol{\beta} + \boldsymbol{u_i})\right),$$

dove $\boldsymbol{u_i} \sim \mathcal{N}(0, \sigma^2)$, considerando anche in questo caso $i = 1, \ldots, 50$ *cluster* e $j = 1, \ldots, 10$ osservazioni per ciascun *cluster*. Maestrini et al. (2024) puntualizzano che in questa configurazione di simulazione è stata considerata esclusivamente l'intercetta

casuale, poiché i codici disponibili pubblicamente per MPL e MPL-REML sono stati sviluppati solo per il caso con intercetta casuale nei GLMM di tipo Poisson.

Lo studio di simulazione per dati di Poisson si articola di tre scenari del tutto analoghi ai primi tre scenari definiti in Sezione 4.2.1, con la sola differenza che la matrice di covarianza degli effetti casuali Σ è sostituita da una varianza $\sigma^2 = 0.25$. Per ciascuno dei tre scenari, sono stati generati 500 campioni.

Per quanto riguarda le distribuzioni a priori del modello bayesiano, si sono scelte le stesse specificazioni del modello logistico nel paragrafo 4.2.1

4.3 Risultati

I risultati relativi agli studi di simulazione sono stati analizzati in termini di distorsione delle stime dei parametri e copertura degli intervalli di confidenza e credibilità. Tali misure sono state calcolate per tutti gli scenari descritti nei paragrafi 4.2.1 e 4.2.2, e confrontate tra i metodi frequentisti e bayesiani.

4.3.1 Dati di binari

I metodi di stima frequentisti PQL e PQL-REML hanno generato errori in alcuni dei campioni simulati. In particolare, il metodo PQL non ha prodotto stime convergenti in 53 campioni nello scenario 1, 60 campioni nello scenario 2, 109 campioni nello scenario 3 e 102 campioni nello scenario 4. Analogamente, il metodo PQL-REML ha mostrato problemi di convergenza in 45 campioni nello scenario 1 e in 71 campioni nello scenario 4. Per quanto riguarda l'inferenza bayesiana, le catene di Markov campionate hanno soddisfatto in tutti i casi la condizione di convergenza della statistica di Gelman e Rubin, con $\hat{R} < 1.01$, e l'effective sample size non è mai risultato inferiore a 1300.

Le Figure 4.1, 4.2 e 4.3 mostrano i boxplot delle differenze tra le stime e i corrispondenti valori veri dei parametri di varianza Σ_{11}, Σ_{22} e Σ_{12}, rispettivamente. I tre metodi basati sulla verosimiglianza REML dimostrano una notevole efficacia nel ridurre il bias nelle stime di Σ_{11}, Σ_{22} e Σ_{12} rispetto alle versioni di verosimiglianza non ristretta. Tale riduzione del bias risulta più marcata nel secondo e terzo scenario, rispetto al primo e al quarto. Tuttavia, si osserva che la riduzione del bias comporta un aumento della variabilità delle stime. Tra tutti i metodi di stima, MPL REML e PQL REML si dimostrano i migliori. Tuttavia, è importante sottolineare che i risultati ottenuti con il metodo PQL REML devono essere interpretati con cautela negli scenari 1 e 4 a causa della presenza di numerosi campioni per i quali le stime non convergono. Per quanto riguarda il parametro Σ_{11}, le stime bayesiane risultano leggermente distorte verso l'alto

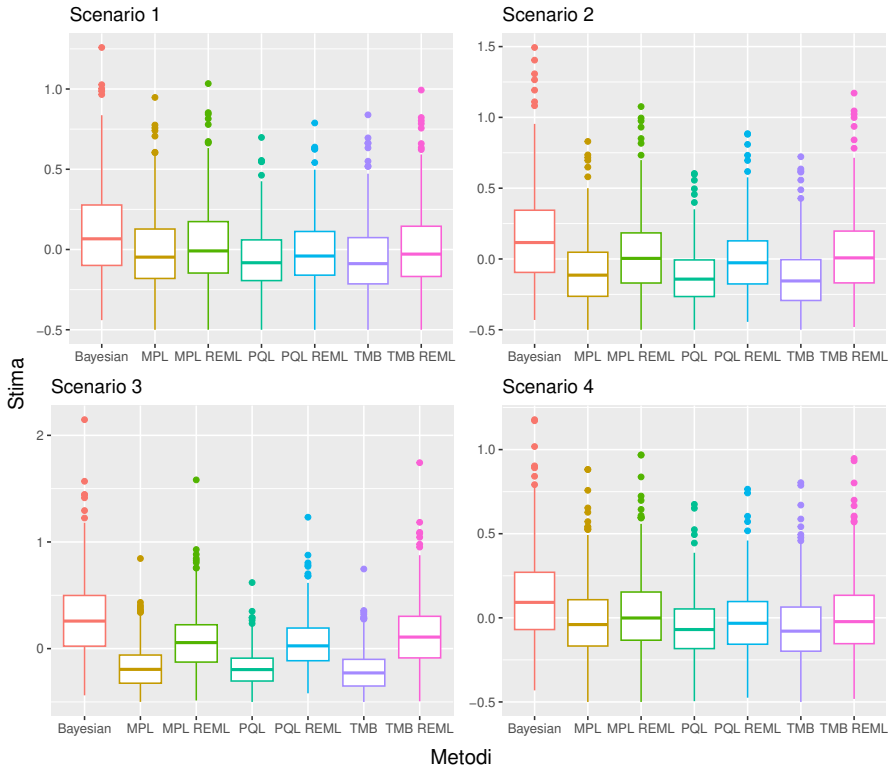

FIGURA 4.1: Boxplot della distorsione delle stime stima di Σ_{11} nei quattro scenari per dati binari.

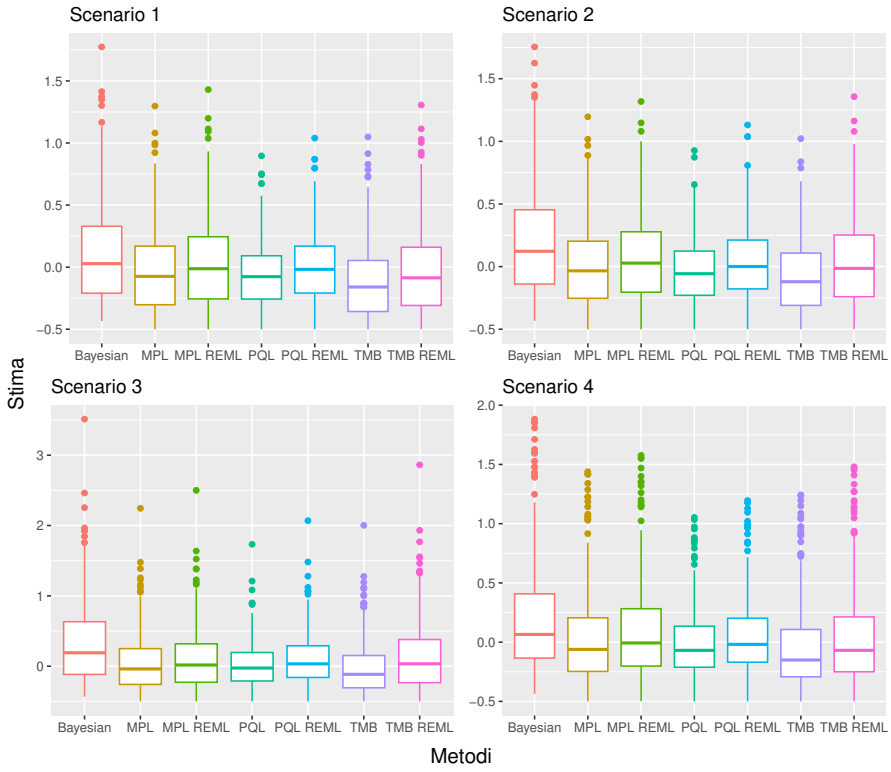

FIGURA 4.2: Boxplot della distorsione delle stime stima di Σ_{22} nei quattro scenari per dati binari.

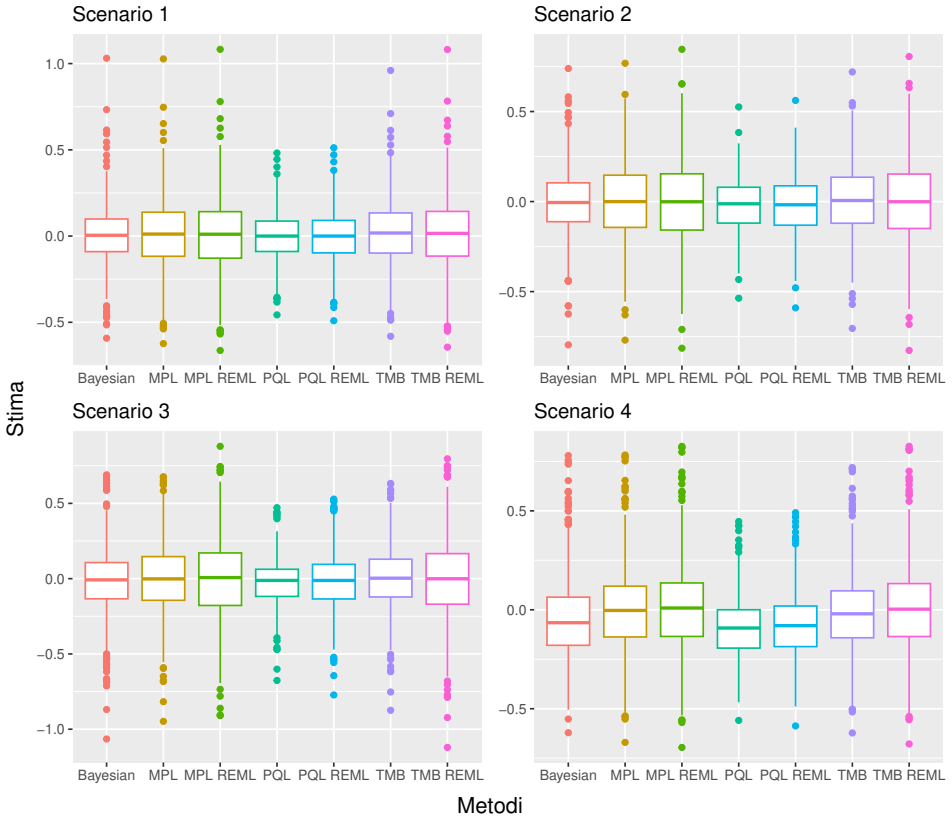

FIGURA 4.3: Boxplot della distorsione delle stime stima di Σ_{12} nei quattro scenari per dati binari.

nel scenario 1, e la distorsione tende ad aumentare con il crescere del numero di parametri stimati nel modello. Il scenario in cui le stime bayesiane di Σ_{11} presentano maggiori problematiche è il terzo, ovvero quello in cui il numero di parametri fissi, $\boldsymbol{\beta}$, è maggiore. In questo caso, circa il 75% delle stime risulta superiore al vero valore del parametro. I risultati delle stime bayesiane di Σ_{22} sono molto simili a quelli di Σ_{11}, con la sola differenza che, per Σ_{22}, le stime risultano meno distorte in tutti gli scenari. Sia per Σ_{11} che per Σ_{22} le stime bayesiane mostrano una variabilità superiore rispetto alle stime di massima verosimiglianza e una variabilità comparabile a quella delle stime di massima verosimiglianza ristretta. Per le stime di Σ_{12}, tutti i metodi proposti producono risultati comparabili in termini di distorsione nel primo, secondo e terzo scenario. I metodi PQL, PQL REML e di inferenza bayesiana presentano una minore variabilità di stima rispetto agli altri metodi. Nel quarto scenario, l'unico in cui $\Sigma_{12} = 0.25$, i metodi PQL, PQL REML e di inferenza bayesiana conducono a stime distorte verso il basso, a differenza degli altri approcci di stima, i cui risultati sono comparabili a quelli degli altri scenari.

Sono state inoltre analizzate le coperture degli intervalli di credibilità bayesiani e degli intervalli di confidenza frequentisti. Per l'inferenza bayesiana sono stati stimati gli intervalli di credibilità *equi-tailed* e *highest posterior density* che verranno in seguito

indicati rispettivamente con le sigle EQUI e HPD. Per l'inferenza frequentista è stato scelto di stimare gli intervalli di confidenza esclusivamente per i metodi proposti da Bellio & Brazzale (2011), in quanto tra tutti i metodi frequentisti risultano avere il miglior compromesso tra convergenza e accuratezza delle stime puntuali.

Per tutti e tre i parametri Σ_{11}, Σ_{12} e Σ_{22} sono stati stimati intervalli alla Wald, che verranno indicati con le sigle "MPL" e "MPL REML". Tuttavia, per i parametri Σ_{11} e Σ_{22} non è garantito che tali intervalli rispettino i vincoli imposti dallo spazio parametrico. Per tale motivo, per questi due parametri, si è scelto di costruire gli intervalli alla Wald in scala logaritmica e trasformarli poi in scala originale. Questa operazione non solo permette di ottenere intervalli che rispettino lo spazio parametrico, ma dovrebbe portare anche a proprietà migliori, poiché la distribuzione degli stimatori delle due componenti di varianza dovrebbe risultare più vicina alla distribuzione gaussiana. Gli intervalli in scala logaritmica verranno indicati in seguito con le sigle "MPL log" e "MPL REML log".

Sia per i metodi bayesiani che per quelli frequentisti, sono stati stimati intervalli al 99%, 95% e 90%.

	HPD	EQUI	MPL log	MPL REML log	MPL	MPL REML
Scenario 1	98.4	98.6	98.53	97.72	93.42	95.10
Scenario 2	98.0	98.2	99.78	91.24	88.33	88.59
Scenario 3	99.2	97.2	99.56	79.82	81.53	79.83
Scenario 4	99.2	99.0	99.36	97.47	96.64	96.04

Tabella 4.1: Copertura degli intervalli al 99% per Σ_{11}

	HPD	EQUI	MPL log	MPL REML log	MPL	MPL REML
Scenario 1	93.4	93.2	97.05	94.81	89.30	90.2
Scenario 2	92.6	92.0	98.68	86.75	81.25	83.2
Scenario 3	95.0	89.8	99.12	74.60	70.91	71.8
Scenario 4	95.8	95.2	97.86	95.15	92.65	92.5

Tabella 4.2: Copertura degli intervalli al 95% per Σ_{11}

	HPD	EQUI	MPL log	MPL REML log	MPL	MPL REML
Scenario 1	86.4	86.4	93.89	90.87	85.80	87.35
Scenario 2	86.4	86.0	97.36	82.91	75.00	78.42
Scenario 3	88.2	84.8	98.45	70.07	63.48	69.20
Scenario 4	91.0	91.0	96.37	93.04	88.45	89.38

Tabella 4.3: Copertura degli intervalli al 90% per Σ_{11}

Nelle Tabelle 4.1, 4.2 e 4.3 sono riportate le percentuali di copertura degli intervalli rispettivamente al 99%, 95% e 90%.

Gli intervalli di confidenza stimati in scala originale mostrano, per tutti e quattro gli scenari e per tutti e tre i livelli di copertura nominale, una copertura empirica inferiore a quella nominale. Inoltre, all'aumentare del numero di parametri stimati nel modello, la percentuale di intervalli che contiene il vero valore del parametro diminuisce drasticamente. Ad esempio, per gli intervalli al 99%, la copertura del metodo MPL REML passa dal 95% nel primo scenario al 79% nel terzo scenario; si ricorda che nel terzo scenario vengono stimati dieci parametri fissi in più rispetto al primo scenario. Nel primo scenario, gli intervalli MPL REML log presentano una copertura molto prossima a quella nominale, tuttavia, all'aumentare del numero di parametri stimati, la copertura empirica di tali intervalli peggiora, arrivando, ad esempio, al 79% nel terzo scenario per gli intervalli al 99% (Tabella 4.1).

Si osserva che gli intervalli stimati mediante il metodo di verosimiglianza non ristretta presentano una copertura empirica sempre superiore a quella nominale, specialmente nel terzo scenario per gli intervalli al 95% e al 90% (Tabelle 4.2, 4.3).

Gli intervalli di credibilità bayesiani presentano coperture molto prossime a quelle nominali rispetto agli intervalli frequentisti, con coperture empiriche molto vicine a quelle attese, specialmente nel quarto scenario, per tutti e tre i livelli di copertura. Mentre per gli intervalli di confidenza frequentisti si osserva un deterioramento delle loro proprietà con l'aumentare del numero di parametri fissi stimati, questo fenomeno non sembra verificarsi per gli intervalli bayesiani, rendendo questi ultimi più affidabili anche nei casi in cui si stima un alto numero di parametri fissi.

	HPD	EQUI	MPL log	MPL REML log	MPL	MPL REML
Scenario 1	99.4	99.8	99.37	95.64	89.94	88.87
Scenario 2	99.8	99.8	99.78	92.52	93.29	88.22
Scenario 3	99.0	97.6	98.01	78.46	91.89	73.19
Scenario 4	99.0	98.2	97.22	94.51	92.32	92.58

Tabella 4.4: Copertura degli intervalli al 99% per Σ_{22}

	HPD	EQUI	MPL log	MPL REML log	MPL	MPL REML
Scenario 1	94.2	95.4	97.26	91.08	87.27	85.83
Scenario 2	96.2	95.6	96.92	87.18	90.57	84.71
Scenario 3	94.2	92.0	94.48	70.52	89.40	68.94
Scenario 4	94.2	95.0	94.87	89.87	88.17	90.31

Tabella 4.5: Copertura degli intervalli al 95% per Σ_{22}

Nelle Tabelle 4.4, 4.5 e 4.6 sono riportate le coperture degli intervalli ai livelli nominali del 99%, 95% e 90% per il parametro Σ_{22}. I risultati ottenuti per Σ_{22} sono molto

	HPD	EQUI	MPL log	MPL REML log	MPL	MPL REML
Scenario 1	86.8	89.8	94.53	87.55	84.39	82.59
Scenario 2	90.6	89.6	94.29	81.20	87.42	81.20
Scenario 3	85.6	85.8	92.27	66.44	85.86	65.96
Scenario 4	85.6	89.6	91.03	85.65	84.02	85.36

TABELLA 4.6: Copertura degli intervalli al 90% per Σ_{22}

simili a quelli relativi a Σ_{11}. Infatti, anche in questo caso le coperture empiriche degli intervalli di tipo Wald per il parametro in scala originale sono notevolmente inferiori a quelle nominali. Tra i diversi intervalli di confidenza frequentisti, il metodo basato sulla massima verosimiglianza non ristretta con intervalli costruiti per il parametro in scala logaritmica mostra coperture empiriche più vicine a quelle nominali. I metodi bayesiani HPD ed EQUI stimano intervalli di credibilità con coperture molto prossime a quelle nominali per i livelli di confidenza al 99% e 95%, mentre tendono a sottostimare leggermente la copertura nel caso del 90%, specialmente per il metodo HPD.

Infine nelle Tabelle 4.7, 4.8 e 4.9 sono riportate le coperture degli intervalli ai livelli nominali del 99%, 95% e 90% per il parametro Σ_{12}. Si osservi che per tale parametro gli intervalli sono stati calcolati solo in scala originale, in quanto lo spazio parametrico non presenta alcun vincolo. Ancora una volta il metodo di stima MPL REML sottostima le coperture attese e, all'aumentare del numero di effetti fissi stimati, le percentuali di intervalli che contengono il vero valore del parametro decrescono. Il metodo MPL porta a intervalli con una copertura empirica molto prossima alla nominale in tutti gli scenari a eccezione del terzo per gli intervalli al 90%, in cui sottostima la copertura attesa. Gli intervalli di credibilità bayesiani, sia EQUI che HPD, sono troppo ampi, portando di conseguenza a coperture empiriche molto elevate. Si veda ad esempio la Tabella 4.8, in cui le coperture per gli intervalli al 95% per il metodo HPD sono tutte superiori al 99%.

	HPD	EQUI	MPL	MPL REML
Scenario 1	99.8	99.8	98.74	95.64
Scenario 2	100.0	99.8	98.46	88.68
Scenario 3	100.0	100.0	98.90	72.79
Scenario 4	100.0	99.2	98.72	96.41

TABELLA 4.7: Copertura degli intervalli al 99% per Σ_{12}

4.3.2 Dati di conteggio

Nello studio di simulazione per dati di conteggio il metodo di stima MPL REML ha generato stime irrealistiche per alcuni dei campioni simulati. In particolare, per 23

	HPD	EQUI	MPL	MPL REML
Scenario 1	99.6	98.6	96.63	92.74
Scenario 2	99.2	98.0	95.60	83.76
Scenario 3	99.4	97.8	93.82	68.03
Scenario 4	99.4	94.6	95.09	93.46

TABELLA 4.8: Copertura degli intervalli al 95% per Σ_{12}

	HPD	EQUI	MPL	MPL REML
Scenario 1	98.6	96.0	91.58	87.97
Scenario 2	98.2	96.6	90.99	79.70
Scenario 3	98.2	94.8	85.43	62.59
Scenario 4	98.2	87.0	90.60	87.34

TABELLA 4.9: Copertura degli intervalli al 90% per Σ_{12}

campioni nel primo scenario, 69 campioni nel secondo scenario e 93 campioni nel terzo scenario il parametro di varianza è stato stimato con valori superiori al 20 (quando il vero valore è 0.25). I risultati dello studio riportati in seguito omettono tali casi. Per quanto riguarda l'inferenza bayesiana, le catene di Markov campionate hanno soddisfatto in tutti i casi la condizione di convergenza della statistica di Gelman e Rubin, con $\hat{R} < 1.01$, e l'effective sample size non è mai risultato inferiore a 3570.

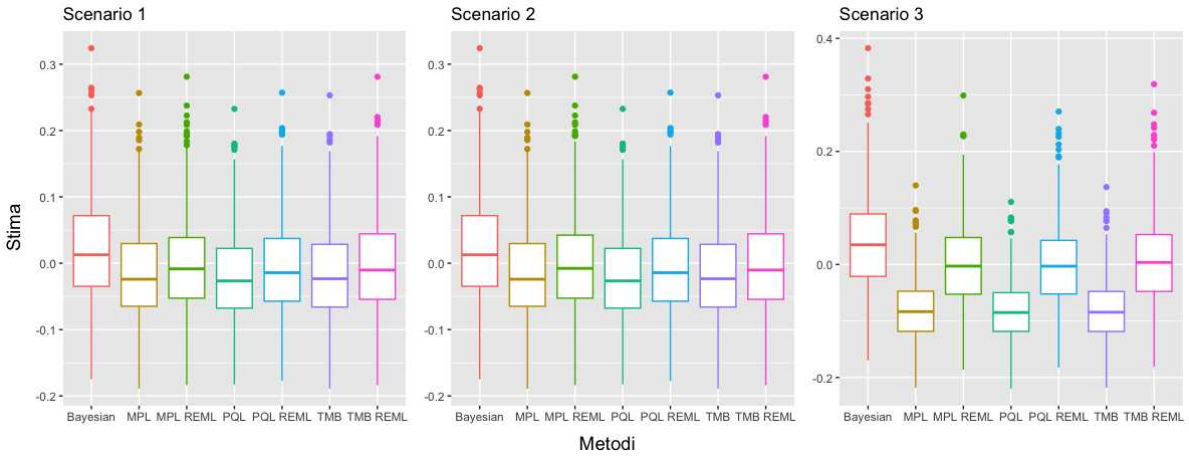

FIGURA 4.4: Boxplot della distorsione delle stime stima di σ^2 nei tre scenari per dati di conteggio.

La Figura 4.4 mostra i boxplot delle differenze tra le stime e il vero valore del parametro di varianza σ^2. Anche per dati di conteggio, i tre metodi basati sulla verosimiglianza REML si dimostrano efficaci nel ridurre il bias delle stime di σ^2 rispetto alle versioni di verosimiglianza non ristretta. Tale riduzione del bias risulta più marcata nel terzo scenario rispetto al primo e al secondo, ovvero il scenario in cui viene stimato il maggior numero di effetti fissi. Si osserva che nei primi due scenari la variabilità delle stime è

molto simile tra i metodi frequentisti, mentre nel terzo scenario i metodi di massima verosimiglianza ristretta presentano una maggiore variabilità. Si ricorda che i risultati per il metodo di stima MPL REML devono essere interpretati con cautela, in quanto in tutti e tre i casi sono state omesse alcune stime poiché irrealistiche.

Lo stimatore bayesiano risulta essere leggermente distorto verso l'alto nel primo e secondo scenario, mentre tale distorsione è maggiore nel terzo scenario, in cui viene stimato un numero più elevato di parametri fissi. Si osserva, inoltre, che la variabilità di stima è molto simile a quella degli stimatori di massima verosimiglianza ristretta.

Analogamente a quanto visto nel caso di dati binari, sono state analizzate le coperture degli intervalli di credibilità bayesiani e degli intervalli di confidenza frequentisti. Per l'inferenza bayesiana sono stati stimati gli intervalli di credibilità EQUI e HPD. Per l'inferenza frequentista si è scelto di stimare gli intervalli di confidenza utilizzando il metodo PQL REML (la versione REML del metodo di quasi-verosimiglianza penalizzata) ed il metodo di massima verosimiglianza profilo, indicato precedentemente come MPL. Rispetto al caso dei dati binari, si è deciso di stimare gli intervalli di confidenza con il metodo PQL REML, poiché si è rivelato migliore in termini di convergenza delle stime rispetto al metodo MPL REML in questo contesto. Anche in questo caso sono stati stimati gli intervalli di confidenza frequentisti sia in scala originale che in scala logaritmica.

	HPD	EQUI	MPL log	PQL REML log	MPL	PQL REML
Scenario 1	99.0	98.6	98.8	99.2	93.6	95.2
Scenario 2	99.2	98.8	96.2	99.4	84.6	95.2
Scenario 3	99.0	98.6	91.6	99.4	70.4	96.8

Tabella 4.10: Copertura degli intervalli al 99% per σ^2

	HPD	EQUI	MPL log	PQL REML log	MPL	PQL REML
Scenario 1	93.6	93.0	91.0	93.2	88.2	90.8
Scenario 2	94.8	95.0	86.0	95.2	75.4	88.8
Scenario 3	96.2	95.4	72.4	96.0	58.4	94.2

Tabella 4.11: Copertura degli intervalli al 95% per σ^2

	HPD	EQUI	MPL log	PQL REML log	MPL	PQL REML
Scenario 1	88.6	88.0	85.8	88.4	82.8	86.6
Scenario 2	89.4	87.8	78.0	90.2	69.8	85.0
Scenario 3	92.6	89.6	61.8	91.4	49.6	90.0

Tabella 4.12: Copertura degli intervalli al 90% per σ^2

Nelle Tabelle 4.10, 4.11 e 4.12 sono riportate le percentuali di copertura degli intervalli rispettivamente al 99%, 95% e 90% per i tre scenari presi in considerazione.

Gli intervalli di confidenza stimati in scala originale mostrano, per tutti e tre gli scenari e per tutti e tre i livelli di copertura nominale, una copertura empirica inferiore a quella attesa, specialmente per quelli stimati mediante il metodo MPL. All'aumentare del numero di effetti fissi nel modello, le coperture empiriche degli intervalli per il metodo MPL, sia in scala logaritmica che originale, diminuiscono drasticamente. Tuttavia, lo stesso non si verifica per gli intervalli stimati mediante il metodo di massima verosimiglianza penalizzata ristretta. Infatti, il metodo PQL REML, con intervalli stimati in scala logaritmica, presenta coperture empiriche molto prossime a quelle attese in tutti e tre gli scenari e per tutti e tre i livelli di confidenza nominale considerati.

Gli intervalli di credibilità bayesiani portano a risultati molto simili a quelli ottenuti con il metodo frequentista PQL REML, con coperture empiriche molto prossime a quelle attese in tutti e tre gli scenari e per tutti e tre i livelli di confidenza nominale. Inoltre, tali intervalli non perdono copertura empirica all'aumentare del numero di effetti fissi stimati nel modello.

Conclusioni

In questa tesi è stato affrontato il problema della stima puntuale e intervallare delle componenti di varianza nei Modelli Lineari Generalizzati Misti (GLMM). L'obiettivo principale del lavoro è stato quello di confrontare e valutare l'accuratezza delle stime ottenute tramite inferenza bayesiana rispetto a quelle fornite da metodi di inferenza frequentista.

Il confronto tra i due approcci inferenziali è stato realizzato mediante uno studio di simulazione applicato a due modelli di regressione: il modello logistico e il modello di Poisson. Dai risultati delle simulazioni emerge che lo stimatore bayesiano delle componenti di varianza presenta una distorsione positiva non trascurabile, in particolare per il parametro di varianza relativo all'intercetta casuale nel modello logistico. Inoltre, questa distorsione tende ad aumentare con l'incremento del numero di effetti fissi stimati nel modello.

D'altro canto, i metodi di stima frequentisti basati sulla massima verosimiglianza non ristretta, mostrano una tendenza opposta, le stime sono distorte verso il basso, con una distorsione crescente al crescere del numero di effetti fissi presenti nel modello. I metodi basati sulla massima verosimiglianza ristretta si sono rivelati i più efficaci nel produrre stime puntuali meno distorte. Tuttavia è importante notare che, mentre gli stimatori bayesiani hanno fornito sempre stime coerenti, i metodi frequentisti hanno spesso portato a risultati meno affidabili con valori in alcuni casi molto elevati. In particolare, nei casi di dati di conteggio e dati binari, i metodi frequentisti PQL, PQL REML e MPL REML hanno riscontrato problemi di stima, compromettendo l'affidabilità dei risultati. Tra i metodi frequentisti i metodi TMB e TMB REML sono risultati essere i più affidabili in termini di convergenza delle stime, in quanto non hanno riscontrato problemi di convergenza in nessuo dei campioni simulati.

Per quanto riguarda le stime intervallari, i metodi bayesiani hanno prodotto intervalli di credibilità con coperture più vicine ai valori nominali rispetto agli intervalli di confidenza dei metodi frequentisti per i parametri di varianza dell'intercetta e pendenza casuale. Gli intervalli di credibilità bayesiani hanno dimostrato buona stabilità anche con un

numero crescente di parametri fissi nel modello, mentre i metodi frequentisti hanno evidenziato una tendenza a portare a coperture empiriche decrescenti all'aumentare del numero di effetti fissi. Tuttavia gli intervalli di credibilità del parametro di covarianza tra intercetta e pendenza casuale sono risultati essere troppo ampi. Intervalli troppo ampi per tale parametro potrebbero portare a risultati fuorvianti, poiché potrebbero indurre erroneamente a concludere che vi sia covarianza nulla qualora includessero lo zero, mentre il vero valore del parametro fosse diverso da zero.

Le stime bayesiane non hanno evidenziato problemi di convergenza, tuttavia questo risultato è influenzato dalla parametrizzazione adottata per il modello. Quando si specifica il GLMM, il campionamento tramite Hamiltonian Monte Carlo può generare catene caratterizzate da elevata correlazione e lenta convergenza verso la distribuzione stazionaria. Per ottenere catene di Markov con una bassa autocorrelazione, è stata necessaria una riparametrizzazione del modello mediante la standardizzazione degli effetti casuali e la specificazione di una distribuzione a priori normale standard per questi ultimi.

In questa tesi sono state analizzate le proprietà degli stimatori per una numerosità campionaria fissata. Un possibile sviluppo futuro potrebbe consistere nell'ampliare lo studio di simulazione esplorando scenari con diverse configurazioni di *cluster*, variando sia il numero di *cluster* sia la numerosità campionaria all'interno di ciascuno. Inoltre, sarebbe interessante valutare il comportamento dei diversi metodi inferenziali in presenza di un numero molto elevato di effetti fissi. Questa estensione consentirebbe di analizzare in modo più approfondito l'impatto della complessità del modello sulla bontà delle stime, offrendo indicazioni utili per l'applicazione dei GLMM in contesti di elevata dimensionalità. Un ulteriore sviluppo di ricerca potrebbe riguardare lo studio delle proprietà teoriche frequentiste delle stime bayesiane, esaminandone le proprietà asintotiche al crescere della numerosità campionaria, sia numero di *cluster* che numerosità campionaria all'interno del *cluster*, e del numero di effetti fissi del modello.

Appendice A

Codice stan

A.1 Dati binari

```
functions {
 /* arg of the function
  *   z: matriz of standardized random effects
  *   SD: standard deviation parameters of random effect
  *   L: cholesky correlation matrix of random effect
  */
  matrix scale_r_cor(matrix z, vector SD, matrix L) {
    // return the random effects non standardized
    return transpose(diag_pre_multiply(SD, L) * z);
  }
}

data {
  int<lower=1> N;  // total number of observations
  array[N] int Y;  // response variable
  int<lower=1> K;  // number of population-level effects
  matrix[N, K] X;  // population-level design matrix
  int<lower=1> Nid;  // number of group levels
  array[N] int<lower=1> idNum;  // grouping indicator per observation
  vector[N] Z1;     // 1_n vector for random intercept
  vector[N] Z2;     // observation for random angular coefficient
}
```

```
transformed data {
  matrix[N, K] Xc;  // centered version of X
  vector[K] means_X;
  for (i in 1:K) {
    means_X[i] = mean(X[, i]);
    Xc[, i] = X[, i] - means_X[i];
  }
}

parameters {
  vector[K] beta;  // regression coefficients
  real Intercept;  // intercept for centered predictors
  vector<lower=0>[2] sd_1;  // group-level standard deviations
  matrix[2, Nid] z_1;  // standardized group-level effects
  cholesky_factor_corr[2] L;  // cholesky factor of correlation matrix
}

transformed parameters {
  matrix[Nid, 2] r_1;    // actual group-level effects
  vector[Nid] r_1_1;     // group level effect for intercept
  vector[Nid] r_1_2;     // group level effect for slope
  real lprior = 0;  // prior contributions to the log posterior
  // compute actual group-level effects
  r_1 = scale_r_cor(z_1, sd_1, L);
  r_1_1 = r_1[, 1];
  r_1_2 = r_1[, 2];
  lprior += student_t_lpdf(sd_1 | 3, 0, 2.5)
    - 2*student_t_lccdf(0 | 3, 0, 2.5);
  lprior += lkj_corr_cholesky_lpdf(L | 1);
}

model {
  // initialize linear predictor term
    vector[N] mu = rep_vector(0.0, N);
    mu += Intercept;
    for (n in 1:N) {
```

```
      // add more terms to the linear predictor
      mu[n] += r_1_1[idNum[n]] * Z1[n] + r_1_2[idNum[n]] * Z2[n];
    }
    target += bernoulli_logit_glm_lpmf(Y | Xc, mu, beta);

  // priors including constants
  target += lprior;
  target += std_normal_lpdf(to_vector(z_1));
}



generated quantities {
  // compute group-level correlations
  corr_matrix[2] Cor_1 = multiply_lower_tri_self_transpose(L);
  real <lower=0> sigma1;
  real <lower=0> sigma2;
  real sigma12;
  sigma1 = sd_1[1]^2;
  sigma2 = sd_1[2]^2;
  sigma12 = Cor_1[1, 2]*sqrt(sigma1*sigma2);
}
```

A.2 Dati di conteggio

```
data {
  int<lower=1> N;  // total number of observations
  array[N] int Y;  // response variable
  int<lower=1> K;  // number of population-level effects
  matrix[N, K] X;  // population-level design matrix
  int<lower=1> Nid;  // number of levels
  array[N] int<lower=1> idNum;  // grouping indicator per observation
  vector[N] Z1;     // matrix for random component (only intercept)
}

transformed data {
```

```
  matrix[N, K] Xc;  // centered version of X
  vector[K] means_X;
  for (i in 1:K) {
    means_X[i] = mean(X[, i]);
    Xc[, i] = X[, i] - means_X[i];
  }
}

parameters {
  vector[K] beta;  // regression coefficients
  real Intercept;  // intercept for centered predictors
  real<lower=0> sd_1;  // group-level standard deviations
  vector[Nid] z_1;  // standardized group-level effects
}

transformed parameters {
  vector[Nid] r_1;     // group level effect for intercept
  real lprior = 0;  // prior contributions to the log posterior
  // compute actual group-level effects
  r_1 = (sd_1*(z_1));
  // prior for parameters
  lprior += student_t_lpdf(sd_1 | 3, 0, 2.5)
    - 1*student_t_lccdf(0 | 3, 0, 2.5);
}

model {
    // initialize linear predictor term
    vector[N] mu = rep_vector(0.0, N);
    mu += Intercept;
    for (n in 1:N) {
      // add more terms to the linear predictor
      mu[n] += r_1[idNum[n]]*Z1[n];
    }
    target += poisson_log_glm_lpmf(Y | Xc, mu, beta);

  // priors including constants
```

```
  target += lprior;
  target += std_normal_lpdf(z_1);
}

generated quantities {
  real <lower=0> sigma;
  sigma = sd_1^2;
}
```

Bibliografia

Albert, J. H. & Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American statistical Association* **88**, 669–679.

Barnard, J., McCulloch, R. & Meng, X.-L. (2000). Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage. *Statistica Sinica* **10**, 1281–1311.

Bellhouse, D. (1990). On the equivalence of marginal and approximate conditional likelihoods for correlation parameters under a normal model. *Biometrika* **77**, 743–746.

Bellio, R. & Brazzale, A. R. (2011). Restricted likelihood inference for generalized linear mixed models. *Statistics and Computing* **21**, 173–183.

Box, G. E. & Tiao, G. C. (2011). *Bayesian Inference in Statistical Analysis*. John Wiley & Sons.

Breslow, N. E. & Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American statistical Association* **88**, 9–25.

Brooks, S. P. & Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics* **7**, 434–455.

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software* **80**, 1–28.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P. & Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of statistical software* **76**, 1–32.

Cowles, M. K. & Carlin, B. P. (1996). Markov Chain Monte Carlo convergence diagnostics: a comparative review. *Journal of the American statistical Association* **91**, 883–904.

Cox, D. & Reid, N. (1992). A note on the difference between profile and modified profile likelihood. *Biometrika* **79**, 408–411.

Cox, D. R. & Reid, N. (1987). Parameter orthogonality and approximate conditional inference. *Journal of the Royal Statistical Society: Series B (Methodological)* **49**, 1–18.

Cressie, N. & Lahiri, S. N. (1993). The asymptotic distribution of REML estimators. *Journal of Multivariate Analysis* **45**, 217–233.

Datta, G. S. & Lahiri, P. (2000). A unified measure of uncertainty of estimated best linear unbiased predictors in small area estimation problems. *Statistica Sinica* **10**, 613–627.

Davison, A. C. (2003). *Statistical Models*, vol. 11. Cambridge University press.

Duane, S., Kennedy, A. D., Pendleton, B. J. & Roweth, D. (1987). Hybrid Monte Carlo. *Physics letters B* **195**, 216–222.

Engel, B. & Keen, A. (1994). A simple approach for the analysis of generalizea linear mixed models. *Statistica neerlandica* **48**, 1–22.

Fellner, W. H. (1986). Robust estimation of variance components. *Technometrics* **28**, 51–60.

Flegal, J., Hughes, J., Vats, D. & Dai, N. (2012). mcmcse: Monte Carlo standard errors for mcmc r package version 1.0-1.

Frühwirth-Schnatter, S., Frühwirth, R., Held, L. & Rue, H. (2009). Improved auxiliary mixture sampling for hierarchical models of non-Gaussian data. *Statistics and Computing* **19**, 479–492.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis* **1**, 515 – 534.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A. & Rubin, D. B. (2013). *Bayesian Data Analysis*. Chapman Hall/CRC Texts in Statistical Science. Milton: Chapman and Hall/CRC, third edition. ed.

Gelman, A., Carlin, J. B., Stern, H. S. & Rubin, D. B. (1995). *Bayesian Data Analysis*. Chapman and Hall/CRC.

Gelman, A. & Hill, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press.

Gelman, A., Roberts, G. O. & Gilks, W. R. (1996). Efficient Metropolis jumping rules. *Bayesian statistics* **5**, 599–608.

Gelman, A. & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical science* **7**, 457–472.

Geman, S. & Geman, D. (1984). Stochastic relaxation, gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6**, 721–741.

Hadfield, J. D. (2010). MCMC methods for multi-response generalized linear mixed models: The MCMCglmm R package. *Journal of Statistical Software* **33**, 1–22.

Hall, P., Johnstone, I., Ormerod, J., Wand, M. & Yu, J. (2020). Fast and accurate binary response mixed model analysis via expectation propagation. *Journal of the American Statistical Association* **115**, 1902–1916.

Harville, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American statistical association* **72**, 320–338.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov Chains and their applications. *Biometrika* **57**, 597–109.

Held, L. & Holmes, C. C. (2006). Bayesian auxiliary variable models for binary and multinomial regression. *Bayesian Analysis.* **1**, 145–168.

Henderson, C. R. (1963). Selection index and expected genetic advance. *Statisitical Genetics and Plant Breeding* , 141–163.

Hornik, K., Leisch, F., Zeileis, A. & Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In *Proceedings of the 3rd international workshop on distributed statistical computing*, vol. 124. Vienna, Austria.

Hui, F. K. (2021). On the use of a penalized quasilikelihood information criterion for generalized linear mixed models. *Biometrika* **108**, 353–365.

Laird, N. M. & Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics* **38**, 963–974.

Lewandowski, D., Kurowicka, D. & Joe, H. (2009). Generating random correlation matrices based on vines and extended onion method. *Journal of Multivariate Analysis* **100**, 1989–2001.

Liao, J. & Lipsitz, S. R. (2002). A type of restricted maximum likelihood estimator of variance components in generalised linear mixed models. *Biometrika* **89**, 401–409.

Maestrini, L., Hui, F. K. & Welsh, A. H. (2024). Restricted maximum likelihood estimation in generalized linear mixed models. *arXiv preprint arXiv:2402.12719* .

McCullagh, P. & Nelder, J. (1989). *Generalized Linear Models, Second Edition*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis.

McCullagh, P. & Tibshirani, R. (1990). A simple method for the adjustment of profile likelihoods. *Journal of the Royal Statistical Society: Series B (Methodological)* **52**, 325–344.

McCulloch, C. E. (1997). Maximum likelihood algorithms for generalized linear mixed models. *Journal of the American statistical Association* **92**, 162–170.

McGilchrist, C. (1994). Estimation in generalized mixed models. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **56**, 61–69.

McLachlan, G. J. & Krishnan, T. (2007). *The EM algorithm and extensions*. John Wiley & Sons.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. & Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics* **21**, 1087–1092.

Neal, R. M. (2011). MCMC using Hamiltonian dynamics. In *Handbook of Markov Chain Monte Carlo*, S. Brooks, A. Gelman, G. Jones & X.-L. Meng, eds., chap. 5. Boca Raton: Chapman and Hall/CRC, pp. 113–162.

Ormerod, J. T. & Wand, M. P. (2012). Gaussian variational approximate inference for generalized linear mixed models. *Journal of Computational and Graphical Statistics* **21**, 2–17.

Plummer, M., Best, N., Cowles, K., Vines, K. et al. (2006). Coda: convergence diagnosis and output analysis for mcmc. *R news* **6**, 7–11.

Polson, N. G., Scott, J. G. & Windle, J. (2013). Bayesian inference for logistic models using Pólya–gamma latent variables. *Journal of the American statistical Association* **108**, 1339–1349.

Ripley, B. D. (2002). *Modern Applied Statistics with S*. Springer.

Robert, C. P. & Casella, G. (1999). *Monte Carlo Statistical Methods*, vol. 2. Springer.

Roy, V. (2020). Convergence diagnostics for Markov Chain Monte Carlo. *Annual Review of Statistics and Its Application* **7**, 387–412.

Schall, R. (1991). Estimation in generalized linear models with random effects. *Biometrika* **78**, 719–727.

Severini, T. A. (2000). *Likelihood Methods in Statistics*. Oxford University Press.

Shun, Z. & McCullagh, P. (1995). Laplace approximation of high dimensional integrals. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **57**, 749–760.

Stan Development Team (2012). *Stan Modeling Language User's Guide and Reference Manual, Version 1.0*.

Stan Development Team (2018a). The Stan Core Library. Version 2.18.0.

Stan Development Team (2018b). RStan: the R interface to Stan. R package version 2.17.3.

Stiratelli, R., Laird, N. & Ware, J. H. (1984). Random-effects models for serial observations with binary response. *Biometrics* **40**, 961–971.

Verbyla, A. P. (1990). A conditional derivation of residual maximum likelihood. *Australian and New Zeland Journal of Statistics* **32**, 227–230.

Wang, C., Rutledge, J. & Gianola, D. (1993). Marginal inferences about variance components in a mixed linear model using gibbs sampling. *Genetics Selection Evolution* **25**, 41–62.

Wolfinger, R. & O'connell, M. (1993). Generalized linear mixed models a pseudo-likelihood approach. *Journal of Statistical Computation and Simulation* **48**, 233–243.

Wu, L. (2009). *Mixed Effects Models for Complex Data*. Chapman and Hall/CRC.