Politecnico di Milano

DIPARTIMENTO DI MATEMATICA
SCUOLA DI INGENGERIA DEI SISTEMI

Tesi di laurea specialistica

Bayesian Variable selection for logit models with random intercept: application to STEMI dataset

Laureando:
Francesco MAURI
Matricola 739231

Relatore:
Prof.ssa Alessandra GUGLIELMI

Anno Accademico 2010–2011

a mia madre
Maria Pia

Like dreams, statistics are a form of wish fulfillment.
(Le statistiche sono una forma di realizzazione del desiderio, proprio come i sogni)
cit. Jean Baudrillard

You just think lovely wonderful thoughts- Peter explained -and they lift you up in the air.
(Fai grandi sogni, saranno loro a portarti in alto)
cit. J. M. Barrie in
Peter Pan, or The Boy Who Wouldn't Grow Up

Ringraziamenti

Un grandissimo grazie va, senza ombra di dubbio, alla Prof.ssa Alessandra Guglielmi, la quale mi ha dato la possibilità di utilizzare ancora una volta la statistica in campo medico. Grazie veramente perché mi ha dato tanto: la voglia di scoprire ambiti statistici che non conosco ancora, la voglia di ricercare *attrezzature* adatte a rispondere con lo studio alle domande che un buono studente (e futuro lavoratore) si deve fare davanti ai problemi che si presentano. Grazie per avermi dato consigli, anche in ambiti non di sua competenza (come la scelta sul dopo laurea). Grazie per avermi portato, sopportato e supportato alla stesura di questa Tesi.

Ringrazio inoltre la Prof.ssa Helga Wagner che grazie alla sua gentilezza, ha dato a noi la posibilità di utilizzare un metodo (e soprattuto il suo codice Matlab) per lo studio fatto in questo elaborato. Grazie per aver sempre risposto con velocità a tutti i nostri questiti e sempre con l'aggiunta di materiale bibliografico adatto.

Voglio inoltre dire un grazie al Dott. Jacopo Soriano per il suo interessamento e per i suoi consigli nei momenti di *buio statistico* e alla Dott.ssa Francesca Ieva che ha sempre risposto a tutte le mie domande sul dataset. Mi siete serviti tanto, solo uno sciocco e uno sconsiderato pensa di poter andare avanti senza chiedere mai aiuto.

In questi anni universitari sono successe tante cose (forse troppe); un parsona come me non sarebbe riuscita a rimanere in piedi senza l'aiuto solido di una famiglia come la mia. Ringrazio i miei fratelli Luca e Matteo che con le loro discussioni e litigate mi hanno aiutato a staccare il cervello dall'università, mia sorella Chiara che mi ha insegnato a gettarmi su strade non conosciute e che a prima vista fanno paura. Un grazie a Franca, entrata da poco nella nostra famiglia, ma già una presenza importante.

Un grazie e mia nonna Livia che per i miei esami ha esaurito tutte le candele della chiesa, a mia nonna Francesca (Franca) che era ed è una insegnate favolosa di matematica. Un grazie a tutti i miei zii e miei cugini che durante i momenti importanti erano sempre presenti (anche con lunghe

telefonate serali).

Dopo sei anni, riesco a chiudere questa *avventura universitaria*. Tanti sono stati coloro che mi hanno aiutato in questa esperienza. Fondamentali sono stati (e lo sono ancora) i miei *colleghi* di studio, con i quali ho passato molte ore in Stiva e nelle varie aule. Stefania grazie alla tua saggezza, alla tua voglia di studiare ed impegnarti sempre con il sorriso mi hai dato un grandissimo aiuto e un grandissimo sostegno; Luca, senza la tua precisione e la tua voglia di aiutare in tutti i modi chiunque ti chiedesse un consiglio, questi anni sarebbero stati molto confusionari; Elisabetta, i due anni della specialistica senza di te sono stati lunghissimi, ma i nostri comuni amici (quello giallo con la cravattina e quello rosa con i bermuda) mi hanno tirato su di morale facendomi pensare alle risate e ai momenti piacevoli passati sui banchi della *Nave*.

Un enorme grazie va ai miei amici *di terzo livello*, in ordine alfabetico: Andrea, Beppe, Emanuele, Gianluca, Luca, Marco, Mauro e Sonia. Senza di voi le serate sarebbero senza senso, senza di voi le uscite sarebbero noiose. Grazie a te, amica di *quarto livello* Michela, per tutto tutto tutto.

Un grazie a mio padre Roberto, che con la tua immancabile frase -*Quanto hai preso?... Ecco potevi prendere di più*- mi hai insegnato a ricercare sempre il meglio di me stesso (dirti grazie per tutti i soldi spesi in questi anni sarebbe banale e scontato).

E poi ci sei tu: a causa tua ho fatto il test d'ingresso per il Politecnico, a causa tua ho scelto Ingegneria Matematica, a causa tua e a causa della tua vita mi sono appassionato alla statistica applicata alla *medicina*, alla *farmacologia* e alla *vita umana*. Lo hai fatto involontariamente e anzi mi hai sempre detto che avrei dovuto sceglie da solo la mia strada, ma è a causa tua che l'ho trovata. Grazie per essere stata questa *causa*.

Indice

Elenco delle figure

Elenco delle tabelle

Sommario

In questo elaborato di tesi abbiamo cercato di identificare quali possano essere le relazioni, statisticamente significative, tra la variabile risposta sopravvivenza e le 13 covariate presenti nel dataset MOMI^2 (*MOnth MOnitoring Myocardical Infraction in MIlan*) del progetto STEMI della Regione Lombardia.

Per poterlo studiare, abbiamo considerato un GLMM di tipo logit con un solo effetto casuale, anche detto intercetta aleatoria, con un approccio bayesiano gerarchico in cui l'effetto di raggruppamento (primo livello) è dato dalle strutture ospedaliere dove il paziente con infarto miocardico acuto si è recato al riscontrare dei sintomi.

Per l'implementazione computazionale si è utilizzato un algoritmo MCMC di tipo Gibbs Sampler descritto in Wagner and Duller (2010), grazie al quale si è riusciti a fare selezione bayesiana di variabili diminuendo la dimensionalità del problema di regressione a sole 3 covariate fisse e trovando la varianza aggiuntiva causata dalla struttura ospedaliera considerata (l'effetto aleatorio).

`Keywords:`Bayesian Variable Selection, Spike and Slab smoothing priors, Ridge regression, Lasso, Bayesian Lasso, SSVS, Random Intercept Model, MOMI2 dataset, STEMI patients

Introduzione

In questo elaborato di tesi abbiamo studiato una classe di modelli bayesiani di tipo *GLMM* (logit) con un solo effetto casuale per la scelta delle covariate (effetti fissi e intercetta aleatoria). Abbiamo utilizzato un algortimo presentato in Wagner and Duller (2010), modificando ove necessario il codice fornitoci dalle autrici stesse, applicandolo ad un *dataset* d'interesse. I dati provengono dal progetto MOMI^2 (*MOnth MOnitoring Myocardial Infarction in MIlan*) della Regione Lombardia sui pazienti a cui è stato diagnosticato un infarto miocardico con *ST*-elevato. In molti studi medici si è interessati nell'identificare quali siano fattori che hanno un effetto su una risposta ad una malattia o ad un trattamento medico.

Nel dataset MOMI^2 sono registrati, per ogni paziente infartuato, una serie di covariate (ad esempio la struttura ospedaliera, l'automezzo con cui si è presentato in pronto soccorso, l'età, la gravità dell'infarto, etc.). L'obiettivo di questo progetto è quello di rilevare quali possano essere le relazioni tra i processi di *health-care* e i decessi dei pazienti infartuati con il fine di poter prevedere l'esito dell'evento ma soprattuto per poter migliorare quei fattori strutturali (come ad esempio la presenza di corsie preferenziali all'interno dell'unità di pronto soccorso) che influenzano la probabilità di decesso. Per poter raggiungere questo obiettivo si è utilizzata una tecnica di *scelta del modello di regressione*, ovvero *variale selecion*, che a partire da un insieme di dimensione elevata di covariate, ne estrae un sottoinsieme di dimensione inferiore che riesca a descrivere al meglio la variabile risposta, migliorando in questo modo la interpretazione *fisica* del problema. Il dataset MOMI è stato studiato in in precedenza da Guglielmi et al. (2010), tramite un analisi multilivello di dati *raggruppati*, in cui gli autori affermano che le covariate *Killip*, *Età* e il valore logaritmico de *Onset to Baloon* sono quelle che possono avere un collegamento con la sopravvivenza del paziente.

Le conclusioni di questa tesi si discosteranno di poco da quanto detto nello studio precedente: le covariate di interesse saranno il *Killip*, l'*Età* e la covariata *modo* che riporta il tipo di automezzo utilizzato per arrivare al pronto soccorso e non più l'*Onset to Baloon*.

In entrambi gli studi, in ogni caso, si è giunti alla conclusione che esiste un fattore di raggruppamento determinato dalle diverse strutture ospedaliere.

Nel primo capitolo sono stati trattati alcuni metodi e modelli per la selezione di variabile del problema di regressione lineare. Partendo da una veloce rilettura della soluzione ai *minimi quadrati ordinari* si è deciso di illustrare brevemente la regressione lineare penalizzata L^2, ovvero la *Ridge Regression*. Questa è una tecnica inizialmente introdotta con lo scopo di ovviare al problema di singolarità (o quasi-singolarità) della matrice disegno, che al posto di minimizzare la somma degli scarti quadratici minimizza la somma della norma-l^2 degli scarti e della norma-l^2 dei coefficienti dei regressori. Come si vedrà meglio nella sezione dedicata, questa tecnica riduce (*shrink*) il valore e il peso di ogni coefficiente ma non li pone mai nulli. Successivamente è stato trattato il metodo *Lasso*, che differisce dalla Ridge regressione perché penalizza la funzione degli scarti quadratici medi con la norma-l^1 dei parametri di regressione. Il Lasso, rispetto alla penalizzazione L^2 ha la proprietà di aggiungere all'effetto di *shrink*, l'effetto di *variable selection*. Nello stesso capitolo viene poi presentata una rilettura bayesiana di queste due tecniche di regressione lineare il *Bayesian Lasso* e la *Bayesian Ridge Regression*. Da questa visione bayesiana dei due modelli si arriva, alla fine del capitolo, alla descrizione dei modelli di *Stochastic Search Variable Selection* e alla descrizione delle prior *spike-and-slab* per modelli di regressione lineare.

Il secondo capitolo richiama brevemente i modelli lineari di regressione, i modelli lineari generalizzati e i modelli lineari a effetti misti, nei quali oltre agli effetti fissi si aggiungono gli effetti aleatori, nel capitolo stesso vengono definiti cosa sono questi effetti. Infine sono stati brevemente descritti i *Generalized Linear Mixed-Effects Models* focalizzandosi su un caso particolare: il modello logit a intercetta aleatoria. La struttura dei questo capitolo è rappresentata nella Figura (1), nella quale si può notare come i modelli lineari si possano generalizzare sia nei *GLM* sia nei modelli lineari a effetti misti. Dalla fusione di questi si possono definire i *GLMM*. Un caso particolare di *GLMM* è quello dei modelli generalizzati a intercetta aleatoria dove gli effetti casuali si riducono ad una sola variabile. Come sotto-caso si arriva al modello descritto nel capitolo 3 utilizzato nell'analisi di questa tesi.

Nel terzo capitolo ho descritto un particolare modello GLMM, cioè un modello logit a intercetta aleatoria per fare selezione bayesiana di variabili tramite l'elicitazione di prior di *smoothing e shrinkage* dette*Spike-Slab*. Dopo una descrizione della verosimiglianza e delle prior per il vattore di tutti i parametri, compresi gli effetti fissi e aleatori, ho brevemente illustrato l'algritmo di tipo *Gibbs Sampling* utilizzato per la selezione delle variabili su un dataset a risposta dicotomica con un solo fattore di raggruppamento.

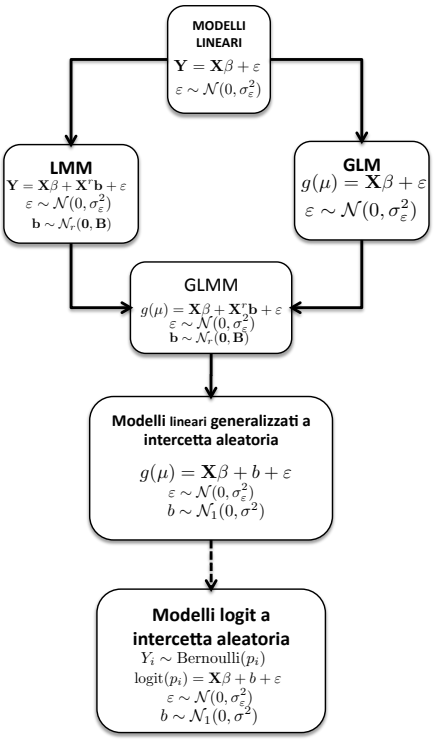


Figura 1: Dai modelli lineari al modello Logit a intercetta aleatoria, passaggi teorici

Nell'ultimo capitolo ho applicato il modello descritto nel capitolo 3 ad un dataset reale (MOMI del progetto STEMI della Regione Lombardia con il 118) che si riferisce a pazienti colpiti da un infarto miocardico acuto. L'obiettivo è stabilire quali siano le coavariate influenti per la descrizione del caso clinico e se vi possa essere una differenza tra le diverse strutture di pronto soccorso degli ospedali Lombardi.

Francesco Mauri

Capitolo 1

Metodi di Shrinkage

In questo capitolo, presenteremo due tecniche per la selezione di variabile nel caso di modelli lineari: Ridge Regression e Lasso. Dopo un primo ripasso della soluzione ai minimi quadrati dei modelli lineari, si passa alla definizione delle soluzioni penalizzate L^2 e L^1 mettendole a confronto e mostrandone una rilettura bayesiana tramite la definizione di modelli gerarchici. Infine il capitolo si conclude con dei cenni alla *Stochastic Search Variable Selection*. Per ulteriori dettagli sui metodi di *shrinkage* si rimanda al Capitolo 3 del libro Hastie et al. (2009) che tratta in modo esaustivo Lasso e Ridge regression, più altri metodi molto utili che in questa tesi non vengono trattati.

1.1 Modelli lineari

Sia dato il seguente modello lineare:

$$\mathbf{Y} = \beta_0 \mathbf{1}_n + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{1.1}$$

tale che $\mathbb{E}[\boldsymbol{\varepsilon}] = \mathbf{0}_n$, $Cov[\boldsymbol{\varepsilon}] = \mathbb{E}[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'] = \sigma^2 \mathbf{I}_n$, con $\mathbf{X}$ matrice disegno di dimensioni $n \times p$, $\mathbf{Y}$ vettore risposte di lunghezza n; $\boldsymbol{\varepsilon}$, $\boldsymbol{\beta}$ e β_0 sono parametri incogniti da stimare.

Lo stimatore (classico) ai *minimi quadrati* è:

$$\left(\hat{\beta}_0^{ols}, \hat{\boldsymbol{\beta}}^{ols}\right) = \underset{\boldsymbol{\beta}\in\mathbb{R}, \beta_0 \in \mathbb{R}}{\arg\min} \left\{ (\mathbf{Y} - \beta_0 \mathbf{1} - \mathbf{X}\boldsymbol{\beta})' (\mathbf{Y} - \beta_0 \mathbf{1} - \mathbf{X}\boldsymbol{\beta}) \right\} \tag{1.2}$$

Nel caso in cui le colenne della matrice disegno siano centrate in $\mathbf{0}$ (cioè $\sum_j x_{ij} = 0$ per ogni $i = 1, ..., p$), con semplici passaggi, si può calcolare $\beta_0^{ols} = \overline{\mathbf{y}} = \frac{1}{N} \sum y_i$ (media campionaria delle risposte). Possiamo quindi, senza perdita di generalità, sottrarre ad entrambi i membri di (1.1) la quantità

$\beta_0 \mathbf{1}_n$ e ridefinendo la variabile risposta $\mathbf{Y}$ come $\mathbf{Y} - \beta_0 \mathbf{1}_n$. La formula (1.2) può essere quindi semplificata:

$$\begin{aligned} \hat{\boldsymbol{\beta}}^{ols} &= \arg\min_{\boldsymbol{\beta}\in\mathbb{R}} \left\{ (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})' (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \right\} \\ \hat{\boldsymbol{\beta}}^{ols} &= \arg\min_{\boldsymbol{\beta}\in\mathbb{R}} \ \mathrm{RSS}(\boldsymbol{\beta}) \end{aligned} \tag{1.3}$$

avendo posto $\mathrm{RSS}(\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})' (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$ il *Residual Sum of Square*. Notiamo come RSS sia una funzione quadratica nel vetore $\boldsymbol{\beta}$:

$$RSS(\boldsymbol{\beta}) = \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta} - \mathbf{y}'\mathbf{X}\boldsymbol{\beta} - \boldsymbol{\beta}\mathbf{X}\mathbf{y} + \mathbf{y}'\mathbf{y}.$$

Per trovare il minimo basta quindi trovare il punto stazionario della funzione calcolando le derivate prime (ponendole uguali a zero) e seconde (imponendo che siano positive):

$$\begin{cases} \dfrac{\partial \mathrm{RSS}}{\partial \boldsymbol{\beta}} = -2\mathbf{X}'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \\ \dfrac{\partial^2 \mathrm{RSS}}{\partial \boldsymbol{\beta}^2} = 2\mathbf{X}'\mathbf{X} \end{cases} \tag{1.4}$$

Se si assume che $\mathbf{X}$ sia a rango pieno sulle colonne, allora $\mathbf{X}'\mathbf{X}$ è matrice definita positiva, la derivata seconda della funzione RSS è quindi positiva (qualsiasi sia $\boldsymbol{\beta}$). Per trovare il minimo basta annullare la derivata prima:

$$\hat{\boldsymbol{\beta}}^{ols} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y} \tag{1.5}$$

Questo stimatore è non distorto e la sua matrice di varianza-covarianza è $Cov[\boldsymbol{\beta}^{ols}] = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$. Tutti i risultati non variano per traslazioni fatte sulle colonne della matrice disegno. Nel caso in cui le sue colonne non siano centrate gli stimatori del vettore $\boldsymbol{\beta}$ non cambiano, ma lo stimatore dell'intercetta deve essere riscritto come:

$$\hat{\beta}_0^{ols} = \overline{y} - \frac{1}{n} \sum_j \sum_i x_{ij} \beta_j.$$

Avere un gran numero di predittori è uno svantaggio nella interpretazione *fisica-reale* del modello in esame. È preferibile avere il minor numero di predittori tale che la descrizione del modello lineare sia la più esaustiva possibile. Si apre quindi il campo ai metodi di *selezione di variabili*, tramite i quali si cerca di restringere l'insieme delle covariate considerate nel modello di regressione al fine di migliorarne sia la interpretazione sia l'accuratezza di

predizione. Un primo approcio al problema di *variable selection* è dato dai metodi di *Best-Subset selection* che per ogni $d = 1, ..., p$ selezionano tra tutti i possibili sottoinsiemi di regressori quello descrive meglio il dataset con sole d covariate. Alla fine si avranno p modelli di dimensione crescente, dal quale il ricercatore potrà scegliere ad esempio in base a tecniche di *bootstrap*. Il *Best Subset Selection* non è però ottimale nella scelta delle covariate, infatti partendo da un problema di dimensione p si dovranno effettuare 2^p confronti tra sottoinsiemi, un'altra problematica è che questo questo metodo si basa su un processo discreto di scelta della dimensione del sottoinsieme e ci si potrebbe trovare difronte al caso in cui nei sottoinsiemi di dimensione k e $k+2$ una variabile non venga selezionata mentre in quello di dimensione $k+1$ sì. Per una descrizione di questi metodi e dei suoi punti deboli si consulti il Capitolo 3 del libro Hastie et al. (2009).

Un secondo approccio è dato dai metodi di *shrinkage*: ridge regression, lasso regression. Come descritto nei paragrafi successivi, questi due metodi possono avere una rilettura bayesiana che se generalizzata porta alla descrizione dei modelli di *bayesian variable selection*.

1.2 Ridge regression

Quando vi sono molte variabili correlate in un modello lineare, i coefficienti β_j potrebbero essere male determinati e presentare una grande varianza. Ci si potrebbe trovare in un caso in cui un coefficiente positivo di una variabile sia bilanciato nel modello da una sua variabile correlata ma con coefficiente negativo i cui effetti annullano quelli della prima. Imponendo un vincolo sul valore e sulla dimensione dei β_j $(j = 1, ..., p)$ è possibile diminuire questo effetto. La ridge regression (conosciuta anche come *Tikhonov regularization* con matrice $\Gamma = I_p$) impone una *penalità* sul valore dei regressori tramite un vincolo non lineare sulla norma euclidea del vettore $\boldsymbol{\beta}$, che ne diminuisce il peso e ne contrae il loro valore :

$$\begin{cases} (\hat{\boldsymbol{\beta}}, \hat{\beta}_0)^{ridge} = \underset{\boldsymbol{\beta}, \beta_0}{\arg\min} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum x_{ij}\beta_j \right)^2 \\ s.t. \quad \sum_{j=1}^{p} \beta_j^2 \leq t \end{cases} \tag{1.6}$$

Da notare β_0 non è stato inserito nel termine di penalizzazione.Bisogna fare molta attenzione in quanto la soluzione ridge non rimane inalterata sotto riscalamenti della matrice disegno: le soluzioni con i dati grezzi e con i

dati standardizzati potrebbero essere differenti. Normalmente si preferisce standardizzare i dati prima di ricercare la soluzione L^2-penalizzata. Quello definito in (1.6) è un problema di programmazione non lineare vincolata e grazie alla teoria dei moltiplicatori di Lagrange (di Ottimizzazione vincolata) sappiamo che esiste un unico $\lambda \in [0, +\infty)$ in corrispondenza *univoca* con t tale che il problema primario formulato in (1.6) e il problema sottostante hanno ugual soluzione:

$$(\hat{\boldsymbol{\beta}}, \hat{\beta}_0)^{ridge} = \arg\min_{\boldsymbol{\beta}, \beta_0} \left\{ \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}. \quad (1.7)$$

Poiché β_0 non viene considerato nel vincolo è facilmente calcolabile:

$$\hat{\beta}_0^{ridge} = \frac{1}{N}\sum_{i=1}^{N} y_i - \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{p} x_{ij}\beta_j = \overline{y} - \sum_{j=1}^{p} \overline{x}_j \beta_j.$$

Supponendo quindi che la matrice disegno abbia colonne centrate in $\mathbf{0}$ lo stimatore della intercetta è $\hat{\beta}_0^{ridge} = \overline{y}$ media campionaria delle risposte. Nel caso generale in cui le colonne non siano centrate è facile notare come traslando tutto il problema con $\mathbf{y}^c = \mathbf{y} - \overline{\mathbf{X}}\boldsymbol{\beta}$ e $\mathbf{X}^c = \mathbf{X} - \overline{\mathbf{X}}$ le soluzioni dei β_j non cambino, dove si è posto $\overline{\mathbf{X}}$ la matrice la cui colonna $\overline{\mathbf{x}}_j = \overline{x}_j \mathbf{1}_N$ è la media campionaria della $j-$esima variabile $j = 1, ..., p$.

Supponiamo per semplicità di calcoli che le colonne della matrice disegno siano già state centrate, posso ridurre la dimensione del problema a p (invece che $p+1$) e ridefinire la funzione *Residual sum of squares* aggiungendo un termine di penalizzazione:

$$RSS(\boldsymbol{\beta}, \lambda) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda \boldsymbol{\beta}'\boldsymbol{\beta}.$$

Per trovare il minimo di questa funzione, calcolo le derivate prime e seconde vettoriali:

$$\begin{cases} \dfrac{\partial \mathrm{RSS}}{\partial \boldsymbol{\beta}} = -2\mathbf{X}'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + 2\lambda I \boldsymbol{\beta}' \\ \dfrac{\partial^2 \mathrm{RSS}}{\partial \boldsymbol{\beta}^2} = 2\mathbf{X}'\mathbf{X} + 2\lambda I > 0 \end{cases} \quad (1.8)$$

La funzione *RSS* è quadratica e definita positiva, per trovarne il minimo basta calcolare il punto stazionario e imporre l'annullamento della derivata prima. La formula (chiusa) dello stimatore ridge è quindi:

$$\hat{\boldsymbol{\beta}}^{ridge} = (\mathbf{X}'\mathbf{X} + \lambda I)^{-1}\mathbf{X}\mathbf{y}. \quad (1.9)$$

Notiamo come nel caso ortonormale in cui $\mathbf{X}'\mathbf{X} = I$ lo stimatore ridge possa essere riscritto come un riscalamento dello stimatore ordinario: $\hat{\boldsymbol{\beta}}^{ridge} = \hat{\boldsymbol{\beta}}^{ols}/(1+\lambda)$.

1.2.1 Bayesian ridge regression

Lo stimatore ridge regression può avere una formulazione bayesiana. Infatti imponendo il seguente modello:

$$\begin{aligned} &\mathbf{Y}|\boldsymbol{\beta}, \tau_i^2, \sim \mathcal{N}_n(\mu + \mathbf{x}_i^T\boldsymbol{\beta}, \sigma^2\mathbf{I}_n) \\ &\boldsymbol{\beta} \sim \mathcal{N}_p(0, A) \qquad A = diag(\tau_1^2, ..., \tau_p^2) \\ &(\tau_1^2, ..., \tau_p^2), \sigma^2 > 0 \end{aligned} \tag{1.10}$$

e supponendo che $\tau^2 = \tau_i^2 \; \forall i = 1, ..., p$ con τ^2 e σ^2 valori noti, la distribuzione a posteriori del vettore $\boldsymbol{\beta}$ è:

$$\boldsymbol{\beta}|\mathbf{y} \sim \mathcal{N}_p((\mathbf{X}^T\mathbf{X} + \sigma^2/\tau^2\mathbf{I}_n)^{-1}\mathbf{X}^T\mathbf{y}, \sigma^2(\mathbf{X}^T\mathbf{X} + \sigma^2/\tau^2\mathbf{I}_n)) \tag{1.11}$$

la cui moda (che in questo caso risulta essere uguale alla media) a posteriori dei $\boldsymbol{\beta}$ è proprio l'espressione (1.9) del regressore Ridge

$$\boldsymbol{\beta} = \text{argmax}\left\{log(\pi(\boldsymbol{\beta}|\mathbf{y}))\right\} = \text{arg min}\left\{\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_2^2\right\}$$

con $\lambda = \sigma^2/\tau^2$.

1.3 Il Lasso

La *Ridge* regression è un metodo stabile che contrae il valore dei coefficienti, ma è molto carente come metodo di *selezione di Variabili*, in quanto non pone (se non in casi rari) alcun β_j uguale a zero. Per una semplice spiegazione si prenda ad esempio il caso ortonormale la cui soluzione è $\hat{\beta}_j^{ridge} = \hat{\beta}_j^{ols}/(1+\lambda)$, come si può notare dal grafico a sinistra della figura 1.2 vi è uno *shrink* proporzionale al valore del β_j che non porta mai il coefficiente ad annullarsi. In Tibshirani (1996) vienne presentata una nuova tecnica di shrinkage *Least absolute shrinkage and selection operator* (LASSO), in cui si propone di ridurre sia il numero di regressori sia di diminuire il valore dei coefficienti β_j non nulli. Al posto del vincolo quadratico della *Ridge regression* l'autore dell'articolo inserisce un vincolo non lineare sulla norma l^1 del vettore $\boldsymbol{\beta}$.

Siano dati y_i e $\mathbf{x}_i$ le variabili risposte e i predittori (supponiamoli standardizzati) dell'i-esima unità statistica ($i = 1, ..., n$). Il modello lineare sarà

$\mathbf{Y} = \beta_0 + X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. Lo stimatore lasso è soluzione del sistema (di ricerca operativa):

$$\begin{cases} \arg\min\limits_{\beta_0,\boldsymbol{\beta}} \left\{ \sum\limits_{i=1}^{N} \left(Y_i - \beta_0 - \sum\limits_j \beta_j x_{ij} \right)^2 \right\} \\ \text{s.t. } \sum\limits_{j=1}^{p} |\beta_j| \leq t \end{cases} \tag{1.12}$$

Come fatto precedentemente si può calcolare lo stimatore dell'intercetta: $\hat{\beta}_0^{\,lasso} = \overline{y} - \sum \overline{x}_j \beta_j$. Possiamo centrare le colonne della matrice disegno e avere $\hat{\beta}_0 = \overline{y}$. Per semplicità di calcoli (e senza perdita di generalità) supponiamo fin da subito che le colonne $\mathbf{x}_j$ siano già centrate riducendo il problema di una dimensione: da $p+1$ a p.

Il parametro t, limite superiore della norma-l^1, controlla il maggiore e minore effetto dello shrink. Sia $\hat{\boldsymbol{\beta}}^{ols}$ la soluzione della regressione lineare con il metodo ai minimi quadrati ordinari e poniamo $t_{ols} = \sum |\hat{\boldsymbol{\beta}}_j^{ols}|$ il valore della sua norma-l^1:

- per valori $t \geq t_{ols}$ non si avrà shrinkage, infatti il minimo senza vincoli (OLS) si troverebbe all'interno dell'iper-rombo imposto nel sistema (1.12)
- per valori $t \leq t_{ols}$ si avrà una diminuizione dei singoli valori β_i e, in alcuni casi, alcune delle componenti del vettore $\boldsymbol{\beta}$ verrebbero poste uguali a 0 portando una selezione di variabili.

Utilizzando la teoria dei moltiplicatori di Lagrange è possibile riscrivere il problema in questa forma:

$$\hat{\boldsymbol{\beta}}^{lasso} = \arg\min_{\boldsymbol{\beta}} \left\{ (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} |\beta_j| \right\} \tag{1.13}$$

Caso matrice disegno Ortonormale

Proposizione 1.1. *Sia dato il problema di ottimizzazione* (1.12) *e supponiamo che la matrice disegno* $\mathbf{X} \in \mathbb{R}^{(n \times p)}$ *sia tale che* $\mathbf{X}^T\mathbf{X} = I$ *allora la soluzione del sistema* (1.12) *è:*

$$\hat{\beta}_j^{lasso} = sign(\hat{\beta}_j^{ols})(|\hat{\beta}_j^{ols}| - \gamma)^+ \tag{1.14}$$

con γ *determinato con la condizione* $\sum |\hat{\beta}_j| = t$.

Dimostrazione.

$$\begin{cases} \dfrac{\partial RSS}{\partial \boldsymbol{\beta}} = -2\mathbf{X}^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda \dfrac{\partial}{\partial \boldsymbol{\beta}} \displaystyle\sum_{j=1}^{p} |\boldsymbol{\beta}_j| \\ \dfrac{\partial^2 RSS}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} = 2\mathbf{X}^T\mathbf{X} \end{cases}$$

Ricordo che

$$\left(\frac{\partial RSS}{\partial \boldsymbol{\beta}}\right)_j = \frac{\partial RSS}{\partial \boldsymbol{\beta}_j}$$

$$\begin{aligned} \frac{\partial RSS}{\partial \boldsymbol{\beta}_i} &= \frac{\partial}{\partial \boldsymbol{\beta}_i}\left[\mathbf{y}^T\mathbf{y} - 2\mathbf{y}^T\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} + \lambda\sum_{j=1}^{p}|\boldsymbol{\beta}_j|\right] \\ &= \frac{\partial}{\partial \boldsymbol{\beta}_i}\left[-2\hat{\boldsymbol{\beta}}_{ols}^T\boldsymbol{\beta} + \boldsymbol{\beta} + \lambda\sum_{j=1}^{p}|\boldsymbol{\beta}_j|\right] \\ &= \left[-2\hat{\boldsymbol{\beta}}_i^{ols} + 2\boldsymbol{\beta}_i + \lambda \text{sign}(\boldsymbol{\beta}_i)\right] \end{aligned}$$

caso 1 $\boldsymbol{\beta}_i > 0$

$$\frac{\partial RSS}{\partial \boldsymbol{\beta}_i} = -2\hat{\boldsymbol{\beta}}_i^{ols} + 2\boldsymbol{\beta}_i + \lambda = 0 \rightarrow \hat{\boldsymbol{\beta}}_i^{lasso} = (\hat{\boldsymbol{\beta}}_i^{ols} - \lambda/2)^+$$

caso 2 $\boldsymbol{\beta}_i < 0$

$$\begin{aligned} \frac{\partial RSS}{\partial \boldsymbol{\beta}_i} = -2\hat{\boldsymbol{\beta}}_i^{ols} + 2\boldsymbol{\beta}_i - \lambda = 0 \rightarrow \hat{\boldsymbol{\beta}}_i^{lasso} &= (\hat{\boldsymbol{\beta}}_i^{ols} + \lambda/2)^- \\ &= -(-\hat{\boldsymbol{\beta}}_i^{ols} - \lambda/2)^+ \end{aligned}$$

da cui la tesi con $\gamma = \lambda/2$. □

Caso $p = 1$ **e** $p = 2$

Studiamo ora due casi semplici di applicazione de *Lasso*: caso con una sola variabile e il caso con due variabili.

Per $p = 1$ il problema di ottimizzazione è:

$$\begin{cases} \arg\min_{\beta_1} \|\mathbf{y} - \beta_1\mathbf{x}\|^2 \\ s.t. \quad |\beta_1| \leq t \end{cases}$$

Come è facile notare, se la soluzione ai minimi quadrati ordinari è $\hat{\beta}_1^{ols} \leq t$, allora il vincolo è rispettato e si avrà $\hat{\beta}_1^{ols} = \hat{\beta}_1^{lasso}$; se così non fosse allora $\hat{\beta}_1^{lasso} = t$, il minimo vincolato si troverebbe esattamente sul bordo del segmento $[-t, t]$.

Nel caso $p = 2$ supponiamo di essere nel caso in cui entrambi gli stimatori *OLS* siano positivi, supponiamo inoltre di essere nel caso ortonormale (in cui la matrice disegno è tale che $\mathbf{X}^T\mathbf{X} = \mathbf{I}_n$). Grazie alla proposizione 1.1 sappiamo che

$$\hat{\beta}_j^{lasso} = (\hat{\beta}_j^{ols} - \gamma)^+ \quad j = 1, 2$$

dove γ deve essere calcolata tramite la supposizione che il vincolo sia attivo $\hat{\beta}_1^{lasso} + \hat{\beta}_2^{lasso} = t$ (se il vincolo non fosse attivo significherebbe che entrambi gli stimatori ai minimi quadrati ordinari si trovano dentro l'iperrombo) . Risolvendo si arriva ad avere le formule:

$$\hat{\beta}_1^{lasso} = \left(\frac{t}{2} + \frac{\hat{\beta}_1^{ols} - \hat{\beta}_2^{ols}}{2}\right)^+, \qquad \hat{\beta}_2^{lasso} = \left(\frac{t}{2} - \frac{\hat{\beta}_1^{ols} - \hat{\beta}_2^{ols}}{2}\right)^+$$

Geometria del Lasso

La funzione da minimizzare nel sistema (1.12) è uguale ad una funzione quadratica traslata

$$\sum_{i=1}^{n}(y_i - \sum_j \beta_j x_{ij})^2 = (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^{ols})^T\mathbf{X}^T\mathbf{X}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^{ols}) + const$$

la *RSS* nel caso Lasso è una funzione avente curve di livello ellittiche centrate nella soluzione ai minimi quadrati ordinari. La soluzione lasso può essere vista geometricamente come il punto di tangenza tra il contorno dell'iperrombo e una delle curve di livello della funzione RSS($\boldsymbol{\beta}$).

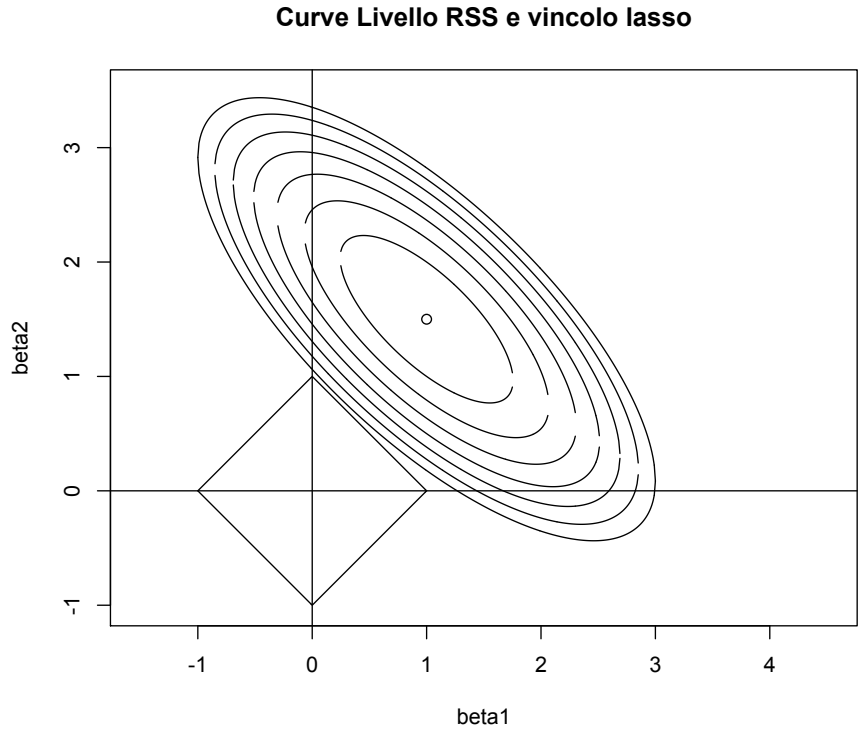


Figura 1.1: Curve di Livello RSS e regione di vincolo caso $p = 2$

1.3.1 Il Bayesian Lasso

Come già accennato in Tibshirani (1996) e ripreso in Park and Casella (2008) lo stimatore Lasso può avere una rilettura bayesiana tramite la rappresentazione:

$$\begin{aligned}
&\mathbf{Y}|\mu, \mathbf{X}, \boldsymbol{\beta}, \sigma^2 \sim \mathcal{N}(\mu\mathbf{I}_N + \mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_N) \\
&\boldsymbol{\beta}|\sigma^2, \tau_1^2, ..., \tau_p^2 \sim \mathcal{N}(\mathbf{0}_p, \sigma^2\mathbf{D}_\tau) \\
&\mathbf{D}_\tau = diag(\tau_1^2, ..., \tau_p^2) \\
&\sigma^2 \sim \pi(\sigma^2) \\
&\tau^2, ..., \tau_p^2 \ i.i.d. \sim \mathcal{E}(\lambda^2/2)
\end{aligned} \tag{1.15}$$

Integrando su τ_i^2 e ricordando che la distribuzione Laplaciana può essere rappresentata tramite la seguente mistura:

$$\frac{a}{2}e^{-a|z|} = \int_0^\infty \frac{1}{\sqrt{2\pi s}} e^{-z^2/(2s)} \frac{a^2}{2} e^{-a^2 s/2} ds \tag{1.16}$$

si arriva ad avere una prior Laplaciana (o esponenziale doppia) per $\boldsymbol{\beta}$:

$$\boldsymbol{\beta}|\sigma^2 \sim \prod_{j=1}^{p} \frac{\lambda}{2\sqrt{\sigma^2}} e^{-\lambda|\boldsymbol{\beta}_j|/\sqrt{\sigma^2}} \tag{1.17}$$

Proposizione 1.2. *Sia dato il modello* (1.15)*, allora la moda a posteriori del vettore* $\boldsymbol{\beta}$ *è lo stimatore* Lasso*.*

Dimostrazione. Imponendo una prior non informativa su σ^2: $\pi(\sigma^2) = 1/\sigma^2$ la *log-posterior* di $(\boldsymbol{\beta}, \sigma^2)$ è proporzionale a:

$$\ln(\pi(\sigma^2)) - \frac{n+p-1}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 - \frac{\lambda}{\sqrt{\sigma^2}}\|\boldsymbol{\beta}\|_1$$

con un cambio di variabile $\phi = \boldsymbol{\beta}/\sqrt{\sigma^2}$ e $\rho = 1/\sqrt{\sigma^2}$ si ha

$$\ln(\pi(1/\rho^2)) + (n+p-1)\ln(\rho) - \frac{1}{2}\|\rho\mathbf{y} - \mathbf{X}\phi\|_2^2 - \lambda\|\phi\|_1 \tag{1.18}$$

Trovare la moda della posterior significa calcolare il massimo su ϕ della formula (1.18) la cui espressione è uguale a quella della definizione del *Lasso.* □

Per trovare lo stimatore Lasso, posso utilizzare un algoritmo di Gibbs Sampler in grado di simulare dalle posterior e quindi di ricavare la moda a posteriori. Lo schema ha bisogno della elicitazione delle full-conditional:

$$\begin{aligned}
&\boldsymbol{\beta}|\mathbf{y}, \sigma^2, \tau_1^2, ..., \tau_p^2 \sim \mathcal{N}_p([\mathbf{X}^T\mathbf{X} + \mathbf{D}_\tau^{-1}]^{-1}\mathbf{X}^T\mathbf{y}, \sigma^2[\mathbf{X}^T\mathbf{X} + \mathbf{D}_\tau^{-1}]^{-1}) \\
&\sigma^2|\boldsymbol{\beta}, \mathbf{y}, \tau_1^2, ..., \tau_p^2 \sim \text{Inv-Gamma}\left(\frac{n+p-1}{2}; \frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \frac{1}{2}\boldsymbol{\beta}^T\mathbf{D}_\tau^{-1}\boldsymbol{\beta}\right) \\
&1/\tau_j^2 \sim \text{Inv-Gaussian}\left(\sqrt{\frac{\lambda^2\sigma^2}{\boldsymbol{\beta}_j^2}}, \lambda^2\right)
\end{aligned} \tag{1.19}$$

Come è possibile notare, l'iperparametro λ regola la regione di vincolo (l'iperrombo) del Lasso. Per ogni diverso valore di λ si avranno differenti valori per β_j^{lasso}. Nella visione frequentista, per trovare il valore ottimale dell'iperparametro tale che il modello lineare stimato descriva in modo efficiente i dati raccolti, è possibile utilizzare la tecnica della *cross-validation.*

Nel caso bayesiano invece è possibile ricavare l'iperparametro λ come massimo calcolato dalla sua distribuzione marginale tramite due schemi diversi:

- Tramite un algoritmo di Monte Carlo di tipo *Expectation Massimization* (EM) che grazie ai parametri generati nella iterazione precedente, stima il valore di λ. Come suggerito nell'articolo Park and Casella (2008) la catena può essere inizializzata con il valore:

$$\lambda^{(0)} = \frac{p\sqrt{\hat{\sigma^2_{ols}}}}{\sum_{j=1}^{p} |\hat{\boldsymbol{\beta}}_j^{ols}|}.$$

  Ogni passo iterativo il parametro viene aggiornato tramite la seguente formula:

$$\lambda^{(k)} = \sqrt{\frac{2p}{\sum_{j=1}^{p} \mathbb{E}_{\lambda^{(k-1)}}[\tau_j^2|\mathbf{y}]}}.$$

  Per maggiori dettagli si veda l'*appendice C* di Park and Casella (2008).

- Esplicitando una iperprior per il parametro λ^2 e non per λ:

$$\pi(\lambda^2) = \frac{\delta^r}{\Gamma(r)}(\lambda^2)^{r-1}e^{-\delta\lambda^2} \qquad \lambda^2 > 0, r > 0, \delta > 0$$

  $\lambda^2 \sim$ Gamma(r, δ). La full-conditional è anch'essa una distribuzione Gamma di parametri $p + r$ e $\sum_{j=1}^{p} \tau_j^2/2 + \delta$

1.4 Confronto tra Ridge e Lasso

Ponendoci nel caso di matrice disegno $\mathbf{X}$ ortonormale, mettiamo a confronto i due metodi di *Variable Selection* tramite *Shrink* presentati in questo capitolo.

Metodo	Formula
Ridge	$\hat{\boldsymbol{\beta}}_j^{ols}/(1+\lambda)$
Lasso	$\text{sign}(\hat{\boldsymbol{\beta}}_j^{ols})(\lvert\hat{\boldsymbol{\beta}}_j^{ols}\rvert - \lambda)^+$

Tabella 1.1: BSS Ridge e Lasso a confronto nel caso Ortonormale

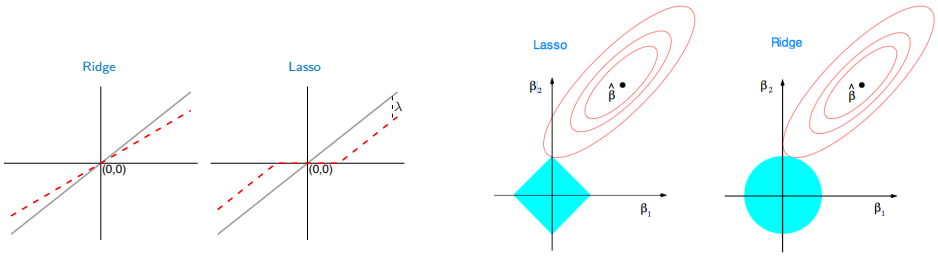


Figura 1.2: Confronto tra Lasso e Ridge

Come si nota dalla figura (1.2) sinistra, il metodo *Ridge* ha uno shrink proporzionale allo stimatore *OLS* mentre il metodo *Lasso* trasla ogni coefficiente di una quantità λ e vicino all'origine tronca a 0. Nella figura di destra è rappresentato un caso non ortonormale e bidimensionale. Il punto segnato come $\hat{\boldsymbol{\beta}}$ è la soluzione ai minimi quadrati ordinari della regressione lineare, gli ellissoidi rappresentano le curve di livello della funzione *somma dei residui quadratici*. Grazie alla figura, è semplice notare come entrambi i metodi abbiano come soluzione i punti in cui le curve di livello ellittiche toccano il bordo della regione di vincolo.

Un altro effetto importante che differenzia i due tipi di *shrinkage* è il comportamento degli stimatori in presenza di correlazione tra le diverse variabili della matrice disegno. Nel Lasso al crescere della correlazione tra le variabili, non cambiano i valori degli stimatori dei regressori mentre, come descritto in un esempio in Tibshirani (1996), gli stimatori Ridge cambiano i loro valori al variare della correlazione.

1.5 Variable Selection in campo bayesiano

Mettendo a confronto il Lasso con il Bayesian Lasso e lo stimatore Ridge con lo stimatore Bayesian Ridge è possibile notare come grazie all'approccio bayesiano sia possibile fare *selezione di variabili* in un modello lineare tramite una opportuna scelta della prior da assegnare ai β_j con $j = 1, ..., p$. Nel *Bayesian lasso*, infatti, la prior assegnata ai β_j è una *doppia-esponenziale*, che oltre a fare *shrinkage*, tronca a 0 quei β_j relativi alle covariate X_j non utili nella descrizione lineare della variabile dipendente Y. Esistono diverse classi di prior che perettono la scelta del modello. Per esempio in Mitchell and Beauchamp (1988) gli autori presentano una tecnica di selezione delle variabili basata sulla definizione *ad hoc* di prior per i β_j. Essi propongono di utilizzare una distribuzione a priori mistura di una massa concentrata in 0 e una distribuzione uniforme distribuita nell'intervallo $[-f_j, f_j]$ (con f_j molto grande in modo da poter spalmare su più valori possibili la prior). Questo

tipo di prior viene definita dagli stessi autori come *spike and slab prior*. La massa concentrata in 0 tenderà ad annullare i coefficienti regressivi relativi a quelle covariate che non dovranno essere scelte per la descrizione finale del modello lineare, mentre la componente uniforme è utile nel caso opposto in cui i predittori sono iportanti.

In George and McCulloch (1993) viene introdotta una prior mistura di due diverse componenti; la tecnica di selezione delle cariabili corrispondente si chiama *Stochastic Search Variable Selection*, SSVS. Le due componenti sono entrambe gaussiane ma con varianze diverse:

$$\beta_j|\delta_j \sim (1-\delta_j)\mathcal{N}(0,\tau_j^2) + \delta_j\mathcal{N}(0,c_j^2\tau_j^2)$$

la variabile δ_j è ua variabile latente dicotomica t.c. $\mathbb{P}(\delta_j = 1) = p_j$. Il parametro τ_j è molto piccolo e c_j al contrario è molto grande in modo tale da avere una mistura di due gaussiane, una che concentra tutta la sua massa vicino a 0 e una che spazia su un intervallo più ampio. In questo modello p_j può essere interpretato come la probabilità a priori che il β_j relativo non venga assunto nullo, o equivalentemente che la variabile X_j debba essere considerata nel modello finale. Notiamo che in questo caso, a differenza della scelta di prior *spike-and-slab* fatta in Mitchell and Beauchamp (1988), si ha a che fare con distribuzioni continue e definite su tutto lo spazio e non solamamente su un intervallo finito. Poiché la densità a posteriori dei parametri di regressione non ha una espressione analitica semplice da calcolare, gli autori dell'articolo propongono un algoritmo di tipo *Gibbs Sampler* per la generazione della sequenza del vettore dei parametri ed in particolare dei δ_j che rappresentano l'inclusione o meno della relativa covariata nel modello regressivo. Per avere una breve descrzione di che cosa sia il *Gibbs Sampling* si veda l'Appendice C di questa tesi.

Un utile riferimento che descrive diverse metodi per fare selezione di variabile nell'approccio bayesiano è George and McCulloch (1997) che inserisce il *SSVS* come un caso particolare di un gruppo di tecniche chiamate *Bayesian Variable Selection*.

Infine, in Ishwaran and Rao (2005) gli autori unificano la visione fatta negli articoli precedenti e ridefiniscono le prior di tipo *spike-and-slab* tramite un modello bayesiano

$$\begin{aligned}
\mathbf{Y}|\mathbf{X},\boldsymbol{\beta},\sigma^2 &\sim \mathcal{N}_n(\mathbf{X}\boldsymbol{\beta},\sigma^2\mathbf{I}_n),\\
\boldsymbol{\beta}|\boldsymbol{\gamma} &\sim \mathcal{N}_p(\mathbf{0},\boldsymbol{\Gamma}) \quad \boldsymbol{\Gamma} = diag(\gamma_1,...,\gamma_K),\\
\boldsymbol{\gamma} &\sim \pi(d\boldsymbol{\gamma}) \quad t.c.\ \mathbb{P}(\gamma_k > 0) = 1 \quad \forall k = 1,...,K\\
\sigma^2 &\sim \pi(d\sigma^2),
\end{aligned} \tag{1.20}$$

dove è la misura di probabilità $\pi(\boldsymbol{\gamma})$ a portare alla elicitazione delle due componenti slab e spike. Nel caso descritto in George and McCulloch (1993) si ha:

$$\begin{aligned}\gamma_k|c_k, \tau_k^2, \mathcal{I}_k &\sim (1-\mathcal{I}_k)\Delta_{\tau_k^2}(\gamma_k) + \mathcal{I}_k\Delta_{c_k\tau_k^2}(\gamma_k)\\ \mathcal{I}_k|w_k &\sim \text{Be}(w_k)\end{aligned}$$

Anche il Bayesian Lasso può essere riletto tramite lo schema (1.20) in cui la prior di $\pi(\boldsymbol{\gamma})$ è esponenziale.

Capitolo 2

I Modelli Lineari e lineari generalizzati nell'approccio Bayesiano

In questo capitolo, partendo dalla descrizione dei modelli lineari già trattati nel capitolo precedente, si passerà alla descrizione dei *GLM* e dei modelli a effetti misti. Entrambi sono una generalizzazione dei modelli lineari: il primo passa da una descrizione lineare tra i coefficienti di regressione e la variabile risposta ad una descrizione della variabile risposta tramite una trasformazione monotona dei regressori; il secondo introduce in aggiunta agli effetti regressivi fissi del modello, gli effetti aleatori. Dall'unione di queste due generalizzazioni si arriva a definire i *GLMM*, modelli lineari generalizzati ad effetti sia fissi che aleatori.

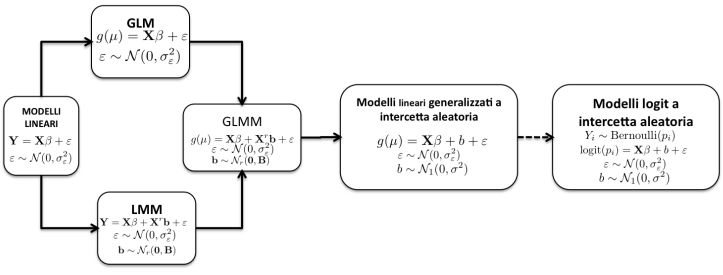


Figura 2.1: Dai Modelli Lineari al Modello Logit a Intercetta Aleatoria

2.1 Modelli lineari generalizzati (GLM)

I modelli lineari sono di facile interpretazione e possono essere applicati ad un gran numero di problemi statistici. Quando, però, la relazione lineare tra la matrice disegno $\mathbf{X}$ e $\mathbb{E}[Y|\mathbf{X}]$ (rispettivamente di dimensioni $(n \times p+1)$ e n) o l'ipotesi di normalità delle variabili risposta decadono, sarebbe utile poter avere un diverso modello per lo studio di regressione. Nell'articolo Nelder and Wedderburn (1972) fu introdotta una visione d'insieme per descrivere tutte le famiglie di modelli utilizzati per l'analisi di regressione con variabili risposta non gaussiane: i Modelli lineari generalizzati (*Generalized Linear Models*). Le tre ipotesi iniziali per la definizione di un GLM sono:

- la variabile risposta Y, condizionatamente a $\mathbf{X}$, ha media μ e deviazione standard σ^2;
- Sia $\eta = \mathbf{X}\boldsymbol{\beta}$, dove $\mathbf{X}$ è una matrice di dimensioni $(n \times p + 1)$ e $\boldsymbol{\beta} = [\beta_0, ..., \beta_p]^T$ il vettore dei regressori, η è detto *parametro canonico*;
- sia $g(\cdot)$ una funzione invertibile detta *link function* tale che $\mu = g^{-1}(\eta) = g^{-1}(\mathbf{X}\boldsymbol{\beta})$, dove $\mu = \mathbb{E}[Y|\mathbf{X}]$.

In particolare, nel caso di variabili risposta binarie, la media μ assume il significato di probabilità p di successo e la *link function* $h = g^{-1}$, dato che $p = h(\mathbf{X}\boldsymbol{\beta}) \in [0, 1]$, solitamente si assume pari ad una funzione di ripartizione F.

2.1.1 Variabile latente per modelli GLM con risposta binaria

Se la risposta è dicotomica, le variabili aleatorie $Y_1, ..., Y_n$ che rappresentano i dati si assumono condizionatamente indipendenti e

$$Y_i|p_i \sim \text{Be}(p_i), \qquad i = 1, ..., n \tag{2.1}$$

In questo caso, come presentato in Albert and Chib (1993), risulta utile (a fini computazionali) introdurre una variabile aleatoria latente Z con distribuzione $F_Z(z - (\mathbf{X}\boldsymbol{\beta}))$.

In particolare si parla di modello probit quando:

$$\begin{aligned} &Y_i = \begin{cases} 1 & , se\ Z_i \geq 0 \\ 0 & , se\ Z_i < 0 \end{cases} \\ &Z_i \sim \mathcal{N}((\mathbf{X}\boldsymbol{\beta})_i, 1) \\ &p_i = \mathbb{P}(Y_i = 1) = \mathbb{P}(Z_i \geq 0) = 1 - \Phi(-(\mathbf{X}\boldsymbol{\beta})_i) = \Phi((\mathbf{X}\boldsymbol{\beta})_i). \end{aligned} \tag{2.2}$$

In modo analogo si introduce il modello Logit nel caso in cui debba utilizzare una regressione logistica . È possibile modificare il modello probit sopra proposto utilizzando una variabile latente con distribuzione logistica

$$\begin{aligned} Y_i &= \begin{cases} 1 & , se\ Z_i \geq 0 \\ 0 & , se\ Z_i < 0 \end{cases} \\ Z_i &\sim \text{Logistic}((\mathbf{X}\boldsymbol{\beta})_i, 1) \\ p_i &= \mathbb{P}(Y_i = 1) = \mathbb{P}(Z_i \geq 0) = 1 - \frac{1}{1 + e^{(\mathbf{X}\boldsymbol{\beta})_i}} = \frac{e^{(\mathbf{X}\boldsymbol{\beta})_i}}{1 + e^{(\mathbf{X}\boldsymbol{\beta})_i}} \end{aligned} \tag{2.3}$$

che può essere riscritto come $\text{logit}(p_i) = \log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \sum \beta_j x_{ij}$. Questo secondo modello è più oneroso computazionalmente rispetto a quello Probit. Si noti che è possibile approssimare la distribuzone logistica tramite una mistura di sei gaussiane (con pesi prestabiliti e con varianze diverse). La descrizione di questa approssimazione la si può trovare alla Sezione 3.1.3 di questa tesi. Nella visione bayesiana il vettore $\boldsymbol{\beta}$ è una variabile aleatoria con una propria distribuzione a priori, la scelta più comune (che si può ritrovare in Dellaportas and Smith (1993)) è $\boldsymbol{\beta} \sim \mathcal{N}_p(b_0, \Sigma)$ con b_0 e Σ iperparametri noti.

Per altri esempi notevoli di *GLM*, non utili in questo elaborato, si rimanda al capitolo 16 del libro Gelman et al. (2003) oppure ai capitolo 1-3 del libro Dey et al. (2000).

2.2 Random Effects Models

In questa sezione descriviamo brevemente i modelli a effetti aleatori, anche detti *variance components model*. In generale il dataset da dover analizzare statisticamente consiste di unità statistiche provenienti da un numero finito di gruppi i quali fanno parte di un insieme (di dimensione maggiore) di tutti i possibili gruppi, i dati sono cioè *raggruppati*. Ad esempio le misure ripetute su uno stesso pazientein istanti diversi o misure fatte su pazienti provenienti da strutture ospedaliere diverse. Se si utilizzassero i modelli lineari e i modelli lineari generalizati con soli effetti *fissi* si avrebbero risultati affetti da *bias*. Non si terrebbe in considerazione il cosiddetto effetto prodotto dal *raggruppamento*. Bisogna quindi introdurre un vettore di *parametri aleatori* che rappresentano questo effetto che influenza la misurazione dei dati di ogni gruppo (paziente nel primo caso o ospedale nel secondo). I modelli a effetti aleatori sono dei modelli a multilivello o gerarchici in cui al primo livello vi sono le unità mentre al secondo livello le misure ripeture. Utilizzare modelli

gerarchici bayesiani aggiunge la possibilità di fare inferenza, ad esempio, su ospedali nei quali non vi è la presenza di pazienti nello studio. Solitamente il vettore degli effetti fissi $\boldsymbol{\beta}$ e degli effetti aleatori $\mathbf{b}$ sono assunti a priori indipendenti.

In questo elaborato di tesi gli effetti fissi verranno descritti dal vettore $\boldsymbol{\beta}$ di lunghezza p (o $p+1$ se nel vettore è presente β_0 l'intercetta) mentre gli effetti aleatori dal vettore $\mathbf{b}$ di lunghezza r.

2.2.1 LMM e GLMM

Un modello lineare a effetti misti, fu introdotto nel lavoro Laird and Ware (1982) per lo studio dei dati *longitudinali*

$$\begin{aligned} &Y_{it} = (\mathbf{X}\boldsymbol{\beta})_{it} + (\mathbf{Zb})_{it} + \varepsilon_{it} \quad \varepsilon_{it} \sim \mathcal{N}(0, \sigma_i^2) \\ &b_i|\mathbf{Q} \sim \mathcal{N}_r(\mathbf{0}, \mathbf{Q}) \end{aligned} \tag{2.4}$$

con r la dimensione non nota dei coefficienti degli effetti aleatori, Z_t e X_t di dimensioni rispettivamente (n, r) e (n, p) matrici disegno per gli effetti aleatori e gli effetti misti infine b_i e ε_{it} sono indipendenti. Dalla (2.4) si può scrivere la distribuzione dei dati risposta come

$$Y_{it}|\boldsymbol{\beta}, b_i \sim \mathcal{N}\left((\mathbf{X}\boldsymbol{\beta})_{it} + (\mathbf{Zb})_{it}, \sigma_i^2\right)$$

oppure

$$Y_{it}|\boldsymbol{\beta}, \sigma_i^2, \mathbf{Q} \sim \mathcal{N}\left((\mathbf{X}\boldsymbol{\beta})_{it}, \mathbf{Z}_t\mathbf{Q}\mathbf{Z}_t^T\sigma_i^2\right).$$

Poiché le unità statistiche possono essere viste come un campionamento finito senza rimpiazzo da una popolazione più ampia di unità, diventa essenziale l'assunzione di *scambiabilità* del vettore $\mathbf{b} = (b_1, ..., b_r)$.

Definizione 2.1 (Scambiabilità)**.** *Una sequenza infinita $\{b_i\}_{i=1}^{\infty}$ di variabili aleatorie è detta scambiabile se per ogni $n \in \mathbb{N}$ e per ogni coppia di permutazioni $\pi(i)$ e $\pi'(i)$ i due vettori aleatori di dimensione finita*

$$(b_{\pi(1)}, ..., b_{\pi(r)}) \quad e \quad (b_{\pi'(1)}, ..., b_{\pi'(r)})$$

hanno la stessa distribuzione di probabilità.

Grazie al *Teorema di Rappresentazione di DeFinetti* la scambiabilità della sequenza di variabili aleatorie b porta all'indipendenza tra le stesse condizionatamente alla conoscenza del parametro di varianza $\mathbf{Q}$.

Come nel caso di modelli lineari si era passati a definire i Modelli Lineari Generalizzati, così è possibile definire i *Generalized Mixed-Effects Linear*

Models (GLMM) come una estensione dei modelli lineari a effetti misti. Consideriamo l'effetto aleatorio $\mathbf{b}|\mathbf{Q} \sim \mathcal{N}_r(\mathbf{0}, \mathbf{Q})$, la *link function* in questo caso sarà

$$h(\eta_{it}) = \mu_{it} = (\mathbf{X}\boldsymbol{\beta})_{it} + (\mathbf{Zb})_{it}$$

Le usuali prior, come detto in Zeger and Karim (1991) per i vettori parametri degli effetti sono:

$$\boldsymbol{\beta} \sim \mathcal{N}_p(\nu_0, \Sigma_0)$$
$$b_1, ..., b_r|\mathbf{Q} \overset{iid}{\sim} \mathcal{N}(0, q)$$

2.3 Variable selection bayesiana tramite modelli a intercetta aleatoria

I modelli a intercetta aleatoria sono un caso particolare dei modelli a effetti misti in cui la dimensione del vettore $\mathbf{b}$ è unitaria: $r = 1$. Ad esempio in questo lavoro di tesi, le unità statistiche sono delle strutture ospedaliere (primo livello del modello gerarchico) mentre i pazienti infartuati (secondo livello) sono le misure ripetute che influiscono sulla intercetta: b è quindi l'intercetta aleatoria che andrà a sommarsi all'intercetta fissa β_0.

Früwirth-Shnatter and Wagner (2010) tramite l'uso di prior unimodali non-Gaussiane (esempio Bayesian Lasso) o di prior *Spike-and-Slab* degli effetti aleatori presentano metodi di *variable selection bayesiana* applicata a modelli sia lineari che generalizzati a effetti misti.

2.3.1 Caso con dati risposta Gaussiani

Siano Y_{it} misure ripetute delle variabili risposta di J unità statistiche ($i = 1, ..., J$) nei tempi $t = 1, ..., T_i$. Supponiamo di avere un modello lineare a effetti misti con un'unico *random effect* $r = 1$:

$$Y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + b_i + \varepsilon_{it}, \qquad \varepsilon_{it} \sim \mathcal{N}(0, \sigma^2_\epsilon) \tag{2.5}$$

con matrice disegno formata da vettori colonna $\mathbf{x}_{it}$ di dimensione $p + 1$ ed infine il vettore dei coefficienti di regressione (ignoti) $\boldsymbol{\beta} = (\beta_0, ..., \beta_p)$ dove la prima colonna della matrice disegno è una colonna di soli 1 e il primo regressore del vettore β_0 è l'intecetta. Per ogni unità statistica è presente un b_i variabile aleatoria che si aggiunge all'intercetta generale ($\boldsymbol{\beta}_0$) con una specifica deviazione standard. Assumiamo, come detto precedentemente, che

$b_1, ..., b_J|\theta$ sia un vettore scambiabile, cioè condizionatamente all'iperparametro $\theta \sim p(\theta)$, b_i e b_j sono indipendenti per ogni i e j. Una scelta molto tipica è:

$$b_i|Q \sim \mathcal{N}(0, Q), \qquad Q \sim \text{Inv-Gamma}(c_0, C_0) \tag{2.6}$$

dove in questo caso Q è la varianza degli effetti fissi.

La *variable selection* per il modello a intercetta aleatoria può essere considerata come un problema di selezione della varianza. L'esistenza di un effetto misto implica l'aggiungersi, in questo caso, di una variabilità nella intercetta: se $Q = 0$ l'effetto aleatorio non c'è, se $Q \neq 0$ allora l'intercetta sarà descritta da $\beta_0 + b_i$ con b_i v.a. la cui varianza deve essere stimata per poter fare inferenza. Essendo in approccio bayesiano si dovrà assegnare una prior $\pi(b_i|\theta)$ all'intercetta aleatoria.

Si potrebbe pensare di non tenere conto della defferenza tra le due tipologie di intercetta e riscrivere il modello (2.5) con un solo vettore di parametri, $\alpha = (\boldsymbol{\beta}, b_1, ..., b_J)$ e quindi:

$$Y_{it} = \mathbf{x}_{it}\alpha + \varepsilon_{it} \qquad \varepsilon_{it} \sim \mathcal{N}(0, \sigma^2_\varepsilon) \tag{2.7}$$

e quindi fare selezione di variabili sul vettore α. Di solito, a priori si assume che $\boldsymbol{\beta}$ e $\mathbf{b} = (b_1, ..., b_J)$ siano indipendenti:

$$\pi(\alpha) = \pi(\boldsymbol{\beta})\pi(b_1, ..., b_j).$$

Essendo $\mathbf{b}$ un vettore scambiabile, grazie al *teorema di DeFinetti* rappresentaremo ogni singola componente del vettore come:

$$b_i|\psi_i \sim \mathcal{N}(0, \psi_i) \qquad \psi_i|\theta \sim \pi(\psi_i|\theta) \tag{2.8}$$

dove ψ_i è un iperparametro di supporto avente una prior propria. Più avanti si capirà l'utilità dell'aggiunta di questo parametro. La distribuzione marginale dell'effetto aleatorio sarà quindi:

$$b_i|\theta \sim \int \pi(b_i|\psi_i)\pi(\psi_i|\theta). \tag{2.9}$$

Notiamo che $(b_i|\psi_i)$ è stocasticamente indipendente da $(b_j|\psi_j)$e la distribuzione marginale $\pi(b_i|\theta)$ potrebbe non essere una legge gaussiana. Ad esempio, sia $\boldsymbol{\theta} = (\nu, Q)$ e sia $\psi_i|\boldsymbol{\theta} \sim \text{Inv-Gamma}(\nu, 1/Q)$ allora avremo che l'intercetta aleatoria ha distribuzione marginale:

$$b_i|\boldsymbol{\theta} \sim t(0, 1/Q\nu, 2\nu).$$

Questo tipo di prior viene detta *di Shrinkage non Gaussiana* che incoraggia a porre uguale a 0 gli effetti aleatori non significativi nel modello, ma potrebbe allontanare in modo significativo da 0 gli altri. Un altro esempio di *shrinkage non-Gaussian prior* è $\psi_i|Q \sim \mathcal{E}(2Q)$ che descrive un modello a intercetta aleatoria con distribuzione Laplaciana:

$$b_i|Q \sim \text{Laplace}(\sqrt{Q})$$

che è un *Bayesian Lasso* applicato ad un modello a intercetta aleatoria. La scelta della iperprior del parametro ψ_i è quindi importante nel caso in cui si voglia fare selezione di variabile. In contesti di selezione di variabile in modelli di regressione è possibile utilizzare anche le *prior spike-and-slab*. Sono distribuzioni a priori mistura finita di due componenti: l'una detta *spike* con piccola varianza e che concentra tutta la massa vicino a 0 e l'altra detta *slab* che spalma la sua massa su un intervallo molto più ampio:

$$\pi(b_i|\omega, \boldsymbol{\theta}) = (1-\omega)\pi_{spike}(b_i|\boldsymbol{\theta}) + \omega\pi_{slab}(b_i|\boldsymbol{\theta}).$$

ricordiamo, dati ω e $\boldsymbol{\theta}$, i b_i sono indipendenti a priori (ipotesi di scambiabilità). Questo tipo di prior può essere riscritta in una versione differente aggiungendo una variabile γ_i tale che:

$$\begin{cases} \mathbb{P}(\gamma_i = 1|\omega) = \omega \\ \pi(b_i|\gamma_i, \boldsymbol{\theta}) = (1-\gamma_i)\pi_{spike}(b_i|\boldsymbol{\theta}) + \gamma_i\pi_{slab}(b_i|\boldsymbol{\theta}) \end{cases}.$$

Capitolo 3

Selezione bayesiana di variabili per modelli logit a intercetta aleatoria

In questo capitolo presenteremo un modello di regressione di tipo logit per *Bayesian variable selection* tramite l'uso di prior di tipo *spike-and-slab*. Ne successivo capitolo verrà utilizzato per analizzare il dataset *MOMI* del progetto per lo studio su *pazienti con infarto miocardio con tratto ST elevato* (STEMI) della Regione Lombardia . È un GLMM con un unico effetto aleatorio che si somma all'intercetta fissa. Il modello e l'algoritmo Gibbs-sampler sono introdotti in Wagner and Duller (2010).

3.1 Il modello

Supponiamo di avere n pazienti e di essere interessati se un certo evento clinico si presenta o meno (ad esempio la ricaduta nella malattia o il decesso a causa della stessa). Dunque Y_i è una variabile aleatoria binaria pari ad 1 se l'evento accade oppure 0 in caso contrario. Sia p_i la frequenza relativa con cui l'evento si presenta. In questa tesi considereremo un modello logit:

$$\mathbb{E}[y_i] = \frac{e^{\beta_0 + \mathbf{x}_i\boldsymbol{\beta} + b_{c(i)}}}{1 + e^{\mu + \mathbf{x}_i\beta + b_{c(i)}}} \tag{3.1}$$

$$\log\left(\frac{p_i}{1-p_i}\right) = \text{logit}(p_i) = \mu + \mathbf{x}_i\beta + b_{c(i)}, \tag{3.2}$$

dove $\mathbf{x}_i$ con $i = 1, ..., n$ è il vettore risposta di dimensione p del i-esimo paziente mentre $\boldsymbol{\beta}$ è il vettore degli effetti fissi di regressione di dimensione

p. Il $b_{c(i)}$ rappresenta l'intercetta aleatoria relativa alla clinica $c(i) = 1, ..., J$ presso la quale l'i-esimo paziente si è presentato per farsi curare, J numero di strutture ospedaliere considerate nello studio.

Supponiamo che:

1. $\mathbf{x}_i$ con $i = 1, ..., N$ siano tutte centrate con vettore $\mathbf{0}$ come media;
2. β_0 sia la media generale tra i vari pazienti;

Prior

La prior dell'effetto aleatorio è $b_{c(i)} \sim \mathcal{N}(0, \sigma_c^2)$ e può essere vista come una deviazione specifica dalla media generale tra le popolazioni dei diversi ospedali. Come proposto in Tüchler (2008) è più semplice computazionalmente utilizzare il modello con parametrizzazione non centrata, ponendo cioè $\theta = \pm\sqrt{\sigma_c^2}$. Il modello si riformula come segue:

$$\text{logit}(p_i) = \mu + \mathbf{x}_i\boldsymbol{\beta} + \theta\widetilde{b}_{c(i)} \qquad \widetilde{b}_{c(i)} \sim \mathcal{N}(0, 1) \tag{3.3}$$

l'estensione di θ a tutto l'asse reale, da la possibilità di definire la sua prior $\pi(\theta)$ nello stesso modo degli effetti fissi $\boldsymbol{\beta}$. Assumiamo, come detto nel capitolo precedente, che a priori i tre parametri β_0, $\boldsymbol{\beta}$ e θ siano indipendenti $\pi(\mu, \boldsymbol{\beta}, \theta) = \pi(\mu)\pi(\boldsymbol{\beta})\pi(\theta)$. La distribuzione a priori della media comune tra gli ospedali (intercetta fissa) è:

$$\beta_0 \sim \mathcal{N}(m_0, M_0) \qquad M_0 > 0; \tag{3.4}$$

per gli altri due parametri (il vettore $\boldsymbol{\beta}$ e θ) scegliamo una prior di tipo *spike and slab*.

3.1.1 Spike e Slab

La distribuzione di probabilità di tipo *spike-and-slab*, già descritta nel Capitolo 1 è una mistura di due differenti componenti. La componente *spike* concentra la sua massa intorno al valore zero, mentre la componente *slab* distribuisce la massa su un più ampio intervallo di possibili valori.

$$\begin{aligned} \pi(\beta_j) &= (1 - \omega_\delta)\pi_{spike}(\beta_j) + \omega_\delta\pi_{slab}(\beta_j) \\ \pi(\theta) &= (1 - \omega_\gamma)\pi_{spike}(\theta) + \omega_\gamma\pi_{slab}(\theta) \end{aligned} \tag{3.5}$$

Inoltre assumiamo che a priori, condizionatamente alla variabile ω_δ, le componenti del vettore $\boldsymbol{\beta}$ siano indipendenti. Introducendo le variabili *indicatori* δ_j con $j = 1, ..., d$ bernoulliane tali che assumono valore unitario se la

corrispettiva β_j è allocata alla compenente slab della prior (allo stesso modo si introduce la variabile γ su θ) si ha che le prior in (3.5) possono essere riformulate come segue:

$$\begin{aligned} &\mathbb{P}(\delta_j = 1|\omega_\delta) = \omega_\delta \\ &\pi(\beta_j|\delta_j) = (1-\delta_j)\pi_{spike}(\beta_j) + \delta_j \pi_{slab}(\beta_j) \end{aligned} \tag{3.6}$$

$$\begin{aligned} &\mathbb{P}(\gamma = 1|\omega_\gamma) = \omega_\gamma \\ &\pi(\theta|\gamma) = (1-\gamma)\pi_{spike}(\theta) + \gamma \pi_{slab}(\theta) \end{aligned} \tag{3.7}$$

Di solito, due sono i tipi di componenti *spike* assunte in questi modelli:

1. spike assolutamente continue alla componente slab;

2. spike Delta di Dirac.

D'ora in poi tutte le considerazioni verranno fatte sulle variabili β_j, per la variabile θ non vi è alcuna differenza.

Spike assolutamente continue

In questo caso, sia la componente spike che la componente slab appartengono alla stessa femiglia di distribuzioni e sono tali che il rapporto tra le due varianze sia al di sotto del valore 1:

$$r = \frac{Var_{spike}(\beta_j|\tau)}{Var_{slab}(\beta_j|\tau)} \ll 1 \tag{3.8}$$

Assumiamo quindi che:

$$\begin{aligned} &\beta_j|\delta_j, \psi_j \sim \mathcal{N}(0, r(\delta_j)\psi_j) \quad \psi_j|\tau \sim \pi(\psi_j|\tau) \quad j = 1, ..., p \\ &r(\delta_j) = r(1-\delta_j) + \delta_j \end{aligned} \tag{3.9}$$

Di seguito vengono presentati alcuni casi di prior utilizzati in campo medico e considerati in letteratura.

La mistura di spike-slab *gaussiane*, le quali possono essere ricavate da (3.9) imponendo $\psi_j = V$ costante e non aleatoria:

$$\beta_j|\omega_\delta \sim (1-\omega_\delta)\mathcal{N}(0, rV) + \omega_\delta \mathcal{N}(0, V)$$

Dimostrazione. Bisogna integrare sulle variabili δ_j e ψ_j. Poiché in questo caso la variabile ψ_j è una costante, basta solo integrare sulla δ_j

$$\begin{aligned}\pi(\beta_j|\omega_\delta) &= \pi(\beta_j|\delta_j)\pi(\delta_j|\omega_\delta) = \\ &= \mathcal{N}(0, r(\delta_j)V)[(1-\omega_\delta)\Delta_0(\delta_j) + \omega_\delta\Delta_1\delta_j] = \\ &= (1-\omega_\delta)\mathcal{N}(0, r(\delta_j=0)V) + \omega_\delta\mathcal{N}(0, r(\delta_j=1)V),\end{aligned}$$

dove con $\Delta_a(x)$ si è indicata la funzione Delta di Dirac centrata in a con argomento x. □

Nel caso si abbia a che fare con dati di tipo *tempo di sopravvivenza* la variabile aggintiva avrà distribuzione $\psi_j \sim$ Inv-Gamma(ν, Q) tale che $\mathbb{E}[\psi_j] = \nu/Q$; la prior marginalizzata su ψ_j di β_j è una mistura di due componenti *t-Student* a tre parametri.

$$\beta_j|\omega_\delta \sim (1-\omega_\delta)t_{2\nu}(0, r/(\nu Q)) + \omega_\delta t_{2\nu}(0, 1/(\nu Q))$$

Dimostrazione. Poiché

$$f(\beta|\theta) = \int f(\beta|\psi)f(\psi|\theta)d\psi,$$

dalla integrazione di δ_j (come nel caso precedente) ricaviamo le due compenenti della prior (spike e slab), mentre integrando su ψ_i si arriva a tesi. Ricordiamo che prese due variabili aleatorie indipendenti Z normale standard e V distribuita secondo una χ^2 con ν gradi di libertà, allora:

$$\frac{Z}{\sqrt{V/\nu}} \sim \mathrm{t}(\nu)$$

□

Infine si ha un caso che può essere visto come estensione del *Bayesian Lasso* nella visione *spike-slab*, assumiamo $\psi_j \sim \mathcal{E}((\lambda^2/2))$:

$$\beta_j|\omega_\delta \sim (1-\omega_\delta)\mathrm{Laplace}(\lambda\sqrt{r}) + \omega_\delta\mathrm{Laplace}(\lambda)$$

Notiamo infatti che ogni singola componente della mistura è una distribuzione *esponenziale doppia*; la ricerca del relativo massimo a posteriori è quindi il *bayesian lasso* per ciascuna componente.

Spike delta di Dirac

In questo caso, la componente spike viene vista (come nel caso della *SSVS* in George and McCulloch (1993)) come una massa discretizzata su 0:

$$\pi(\beta_j|\delta_j = 0) = \pi_{spike}(\beta_j) = \Delta_0(\beta_j). \tag{3.10}$$

Questo caso può essere riletto come un caso particolare di spike assolutamente continua in cui si fa tendere il rapporto r delle varianze a 0. È possibile esprimere la componente spike della prior congiunta di β_j e ψ_j come :

$$\pi(\beta_j, \psi_j|\delta = 0) = \Delta_0(\beta_j)\pi(\psi_j)$$

3.1.2 Prior per ω_δ e ω_γ

Assumiamo a priori l'indipendenza dei due pesi di $\boldsymbol{\beta}$ e di θ:

$$\omega_{\delta_j} \sim \mathcal{B}eta(a_{j,0}, b_{j,0}) \quad j = 1, ..., d \tag{3.11}$$
$$\omega_\gamma \sim \mathcal{B}eta(a_{\gamma,0}, b_{\gamma,0}) \tag{3.12}$$

3.1.3 Data Augmentation

Come proposto inizialmente nell'articolo Albert and Chib (1993) i modelli *Logit* per risposte dicotomiche (Y_i) possono essere riparametrizzate tramite una variabile latente Z_i distribuita come una Logistica:

$$Y_i = \begin{cases} 1 & Z_i > 0 \\ 0 & Z_i \leq 0 \end{cases} \tag{3.13}$$

$$Z_i = \beta_0 + \mathbf{x}_i\beta + \varepsilon_i \qquad \varepsilon \sim \text{Logistic}(0, 1). \tag{3.14}$$

Come descritto nel paragrafo (2.4) di Frühwirth-Schnatter and Frühwirth (2010), la densità logistica standard è ben approssimata da una mistura di sei componenti gaussiane di medie 0.

$$\pi(\varepsilon) = \frac{e^\varepsilon}{(1 + e^\varepsilon)^2} \approx g(\varepsilon) = \sum_{r=1}^{6} w_r \mathcal{N}_\varepsilon(0, s_r^2)$$

i cui valori dei pesi e delle varianze sono in tabella (3.1).

r	s_r^2	$100w_r$
1	0.68159	1.8446
2	1.2419	17.268
3	2.2388	37.393
4	4.0724	31.697
5	7.4371	10.89
6	13.772	0.90745

Tabella 3.1: Pesi e deviazioni standard delle sei componenti gaussiane che approssimano una logistica standard secondo Monahan e Stefanski

Nel nostro caso è quindi possibile riscrivere il modello come segue:

$$
\begin{aligned}
Y_i &= \begin{cases} 1 \ , Z \geq 0 \\ 0 \ , Z < 0 \end{cases} \\
Z_i &= \beta_0 + \sum_{j=1}^{d} \delta_j x_{ij} \beta_j + \gamma \widetilde{b}_{c(i)} \theta + \varepsilon_{r_i} \\
\varepsilon_{r_i} &\sim \mathcal{N}(0, s^2_{r_i}) \\
\widetilde{b}_{c(i)} &\sim \mathcal{N}(0,1)
\end{aligned}
\tag{3.15}
$$

Il modello è completato assumendo come prior per $\boldsymbol{\beta}$ e θ le distribuzioni specificate nei paragrafi [3.1.1] e [3.1.2].

3.2 MCMC

Per poter fare inferenza statistica è necessario ricorrere ad un algoritmo di tipo MCMC per simulare dalla distribuzione multivariata a posteriori. Lo schema è descritto in Wagner and Duller (2010) e in Frühwirth-Schnatter and Frühwirth (2010) e utilizza un algoritmo di tipo di *Gibbs-Sampler* di cui devo elicitare le full-conditionals.

1. Si campionano il vettore delle variabili latenti $\mathbf{Z}$ e il vettore degli indicatori $\mathbf{r}$ condizionatamente a $\beta_0, \boldsymbol{\beta}, \widetilde{b}$:

   (a) ponendo $\lambda_i = \exp(\beta_0 + \mathbf{x}_i\beta + \theta\widetilde{b}_{c(i)})$ genero da $\pi(Z_i|\beta_0, \beta, \theta, \widetilde{b}_{c(i)}, y_i)$:

   $$Z_i = \log(\lambda_i V_i + y_i) - \log(1 - V_i + \lambda_i(1 - y_i)), \qquad V_i \sim \mathcal{U}[0,1]$$

(b) genero dalla densità discreta $\pi(r_i|\beta_0, \beta, \theta, \widetilde{b}_{c(i)}, z_i)$:

$$\mathbb{P}(r_i = j|\beta_0, \boldsymbol{\beta}, \theta, \widetilde{b}_{c(i)}, z_i) \propto \frac{w_j}{s_j} \exp\left[-\frac{1}{2}\left(\frac{z_i - \log\lambda_i}{s_j}\right)^2\right]$$

dove i parametri w_j e s_j sono tabulati in (3.1);

2. genero i pesi $\omega_{\boldsymbol{\delta}}$ e ω_γ dalle full-conditional:

$$\begin{aligned} \omega_\delta &\sim \mathcal{B}eta(a_{\boldsymbol{\delta},0} + d_1, b_{\boldsymbol{\delta},0} + d - d_1) \\ \omega_\gamma &\sim \mathcal{B}eta(a_{\gamma,0} + \gamma, b_{\gamma,0} + 1 - \gamma) \end{aligned} \tag{3.16}$$

dove $d_1 = \sum \delta_j$;

3. genero gli indicatori $\boldsymbol{\delta}$ e γ, dati $\omega_{\boldsymbol{\delta}}$ e ω_γ ed infine genero il vettore $(\beta_0, \boldsymbol{\beta}, \theta)$ (se c'è anche ψ). In base al tipo di spike utilizata il metodo di campionamento cambia, per conoscere entrambe le procedure si veda Wagner and Duller (2010). Di seguito verrà presentato il metodo con la spike delta di dirac. Sia $\zeta = (\beta_0, \beta, \theta)$ campiono dalla densità a posteriori

$$\pi(\boldsymbol{\delta}, \gamma, \zeta|\mathbf{Z}, \mathbf{r}, , \omega_{\boldsymbol{\delta}}, \omega_\gamma, \psi) \tag{3.17}$$

(a) campiono ogni elemento di $\boldsymbol{\delta}$ da $\pi(\delta_j|\boldsymbol{\delta}_{-j}, \gamma, ...)$ (questo aggiornamento viene fatto tramite una permutazione) ed infine campiono $\gamma \sim \pi(\gamma|\boldsymbol{\delta}, ...)$

(b) se $\delta_j = 0$ pongo $\beta_j = 0$ ed in modo analogo se $\gamma = 0$ allora $\theta = 0$, altrimenti si definisce il vettore $\zeta^* = (\beta_0, \boldsymbol{\beta}^\delta, \theta)$ dove $\boldsymbol{\beta}^\delta$ è il vettore dei soli regressori diversi da 0 e genero dal modello di regressione lineare:

$$\mathbf{Z} = \mathbf{U}^*\zeta^* + \varepsilon \qquad \varepsilon \sim \mathcal{N}(0, \Sigma) \tag{3.18}$$

con la matrice disegno $\mathbf{U}^* = [\mathbf{1}, \mathbf{X}^\delta, \widetilde{\mathbf{b}}]$ (se il corrispettivo $\gamma = 0$ tolgo l'ultima colonna di $\mathbf{U}^*$ e l'ultimo elemento di ζ^*). Impongo come prior di $\zeta^*\mathcal{N}(0, A_0)$ con matrice di covarianza A_0 diagonale in cui gli elementi non nulli sono l'M_0 varianza di β_0 e gli ψ_j provenienti dagli β_j e ξ rispetto a θ. Come sappiamo dalla teoria bayesiana dei modelli lineari, la densità a posteriori è $U^* \sim \mathcal{N}_n(a_N, A_N)$ con

$$\begin{aligned} a_N &= (A_0^{-1} + X^{\delta\prime}X^\delta)^{-1}X^\delta\mathbf{y} \\ A_N &= (A_0^{-1} + X^{\delta\prime}X^\delta) \end{aligned}$$

Infine se necessario campione, per $j =, ..., d$ $\psi_j \sim \pi(\psi_j|\delta_j, \beta_j)$ e $\xi \sim \pi(\xi|\gamma, \theta)$ (nel caso in esame sono due valori fissati).

4. campiono gli effetti aleatori $(\widetilde{\mathbf{b}}|\beta_0, \boldsymbol{\beta}, \theta, \mathbf{r}, \mathbf{z})$ dal modello lineare:

$$\widetilde{\mathbf{y}} = \mathbf{Z} - \beta_0 - \mathbf{X}\boldsymbol{\beta} = \theta\mathbf{H}\widetilde{\mathbf{b}} + \boldsymbol{\varepsilon} \qquad \varepsilon \sim \mathcal{N}(\mathbf{0}, \Sigma) \tag{3.19}$$

dove la matrice H selezione per ogni i-esimo paziente il j-esimo ospedale (o clinica). La sua full condizional è una normale multivariata:

$$\begin{aligned} &\widetilde{\mathbf{b}} \sim \mathcal{N}(\mathbf{b}_N, B_N) \\ &B_N^{-1} = \theta^2 H'\Sigma^{-1}H + I \\ &\mathbf{b} = \theta B H'\Sigma^{-1}\tilde{\mathbf{y}} \end{aligned} \tag{3.20}$$

5. ricordando come era stato parametrizzato inizialmente il modello $\theta = \pm\sqrt{\sigma_c^2}$ dovrò cambiare il segno sia di θ che di $\widetilde{\beta}$ in modo causale con probabilità 0.5.

Capitolo 4

Caso applicativo: la scelta delle covariate nel dataset MOMI^2 per pazienti infartuati

In questo capitolo studieremo un data-set di pazienti infartuati MOMI^2 della Regione Lombardia su pazienti affetti da STEMI (*ST-Elevation Myocardial Infraction*). L'obiettivo è, tramite un modello di *variable selection* di approccio bayesiano descritto nel capitolo precedente, di stabilire quali siano le covariate influenti per la descrizione del caso clinico e se vi possa essere una differenza tra le diverse strutture di pronto soccorso degli ospedali Lombardi. Per poter successivamente intervenire laddove le variabili influenti siano causati da fattori strutturali dei singoli ospedali o dei mezzi di pronto soccorso.

4.1 Descrizione del Dataset

L'intero dataset è formato da sei collezioni di dati raccolti su periodi mensili/bimensili nell'area urbana di Milano a partire dal 2001. Dell'intero progetto MOMI^2, abbiamo tenuto in considerazione solamente la terza e la quarta collezione (dal 1 Gugno al 31 Luglio 2007 e dal 15 Novembre al 15 dicembre 2007) con pazienti il cui ospedale di registrazione era presente nel database del Servizio Sanitario Nazionale. Le variabili presenti nel data set sono:

STATO è la variabile risposta che assume due soli valori 1 se il paziente è sopravvissuto dopo l'infarto oppure 0 se è deceduto;

ETA' variabile numerica recante l'età del paziente;

OB variabile numerica Onset to Baloon time (durante lo studio si prenderà in esame il suo valore logaritmico);

Killip variabile di classificazione a quattro livelli utilizzata sugl'individui colpiti da *infarto miocardico acuto*, in questo elaborato i quattro livelli sono stati ridefiniti solamente su due livelli 0 con infarto meno grave (equivalente alle classi 1 e 2) e 1 per quelli più gravi (classi 3 e 4);

Sex variabile dicotomica (1-2) per il sesso del paziente;

ECGtime variabile numerica nella quale è riportato il tempo di *Symptom Onset to Door*;

festivo assume valore 1 se il paziente ha avuto l'infarto durante un giorno feriale (LUN-VEN h 6-18) o 2 altrimenti;

NumRicPrec variabile numerica in cui è registrato il numero dei ricoveri precedenti;

modo variabile categorica con la quale si registra la modalità con la quale il paziente è entrato nella struttura ospedaliera:

1. MSA tramite mezzo di soccorso avanzato (auto medica con a bordo il dottore)
2. MSA+teleECG mezzo di soccorso avanzato con apparecchiatura per teletrasmettere l'esame ECG
3. MSB mezzo di soccorso di base (ambulanza con a bordo volontari ma non medici)
4. SPONTANEO autopresentato con mezzi propri
5. TRASFERITO da un'altra struttura ospedaliera;

sintomo variabile categorica riportante i sintomi: ACC (arresto cardio-circolatorio), ADDOMINALGIA, DISPNEA, DOLORETOR (dolore toracico), SINCOPE, ALTRO ;

sede variabile categorica sede dell'infarto: ANTLAT (antero-laterale), BBS (Blocco di Branca Sinistra), INFOPOST (infero-posteriore);

PTCA variabile categorica sulla modalità di Angioplastica Coronarica Transluminale Percutanea (pratica chirurgica): ELETTIVA, RESCUE, PRIMARIA, NO ;

fast.track variabile categorica che riporta la presenza o meno di una corsia preferenziale per velocizzare il percorso assistenziale del paziente: EMO (corsia presente per il unità operativa di Emodinamica), UTIC (corsia presente per l'unità coronarica), NO (corsia preferenziale non presente nel prontosoccorso);

trombolisi variabile categorica: PREH (pre-ospedaliera), SI (effettuata in ospedale), NO;

centro variabile numerica che identifica la struttura sanitaria presso la quale il paziente si è recato all'insorgere della patologia, nel nostro studio è la variabile *random effect*.

Il data-set finale comprende 240 pazienti (unità statistiche), distribuiti in 17 ospedali dell'area metropolitana milanese, dopo un evento STEMI ed è fortemente sbilanciato (il 95% dei pazienti è sopravvissuto). In ogni ospedale del dataset vi sono un numero di pazienti che varia da un minimo di una unità ad un massimo di 32 (con media sugli ospedali di circa 14.12) con un tasso di sopravvivenza che varia da 75% a 100%. Nel dataset erano presenti diversi dati mancanti (NA), che sono stati sostituiti da valori campionati dalle distribuzioni empiriche marginali di ogni singola covariata.

Il data-set è già stato analizzato in Guglielmi et al. (2010), nel quale tramite modelli GLMM bayesiani, gli autori sono riusciti a selezionare tre variabili (l'età, il Killip e il logaritmo dell'Onset to Baloon time) delle quattordici iniziali. Poiché siamo a conoscenza della precedente selezione di tre variabili, inizieremo con lo studio del dataset ridotto a questi tre *effetti fissi*. Successivamente lo amplieremoa tutte quelle covariate che possano ricondursi a una variabile numerica (cioè le variabili numeriche e quelle dicotomiche). Infine verrà preso in considerazione tutto il dataset (con l'aggiunta delle variabili categoriche nominali) analizzate con l'aiuto della tecnica delle *Dummy*.

4.2 Modello di base

In questa sezione cominceremo a studiare i dati MOMI utilizzando unicamente le variabili ritenute significative dal Guglielmi et al. (2010): l'ETA dei pazienti, il valore logaritmico del tempo Onset to Baloon e la gravità dell'infarto KILLIP. Utilizzando un modello GLMM per la variabile risposta Y_i, che assume valore 1 se il paziente è sopravvissuto e 0 altrimenti. Dunque

$$
\begin{aligned}
&Y_i|p_i \sim \text{Be}(p_i) \qquad i = 1, ..., 240; \\
&\text{logit}(p_i) = \beta_0 + \beta_1 ETA' + \beta_2 log(OB) + \beta_3 KILLIP + \theta\widetilde{b}_{c(i)}, \\
&\widetilde{b}_{c(i)} \sim \mathcal{N}(0, 1), \\
&\beta_0 \sim \mathcal{N}(m_0, M_0), \\
&\beta_j \sim (1 - \omega_{\delta_j})\Delta_0(\beta_j) + \omega_{\delta_j}\mathcal{N}(0, A_{0,jj}) \quad j = 1,\ 2,\ 3; \\
&\theta \sim (1 - \omega_\gamma)\Delta_0(\gamma) + \omega_\gamma\mathcal{N}(0, A_{0,44}), \\
&\omega_{\delta_j} \sim \mathcal{B}eta(a_0, b_0) \quad j = 1, 2, 3; \\
&\omega_\gamma \sim \mathcal{B}eta(a_0, b_0).
\end{aligned}
\tag{4.1}
$$

Per generare le MCMC del Gibbs Sampler ci siamo avvalsi dell'aiuto del programma `Matlab` della prof.ssa *Helga Wagner*. Si è deciso di far campionare 70.000 passi della catena di Markov e di impostare un burn-in di 20.000. Per poter inizializzare gli iperparametri della MCMC, per i primi 10.000 passi della catena si è deciso di imporre che i β_j e θ venissero sempre assegnati alla componente slab.

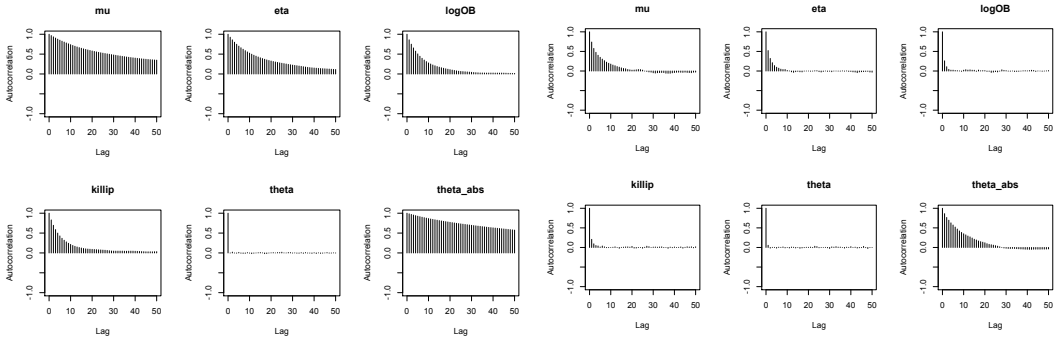
4.2.1 Studio iniziale

Inizialmente presenteremo i risultati ottenuti impostando come parametri di ingresso le varianze delle componenti slab dei β_i e del θ (intercetta aleatoria) al valore 5, una varianza a priori di β_0 (intercetta) pari a 100 mentre gli iperparametri dei pesi ω_δ e ω_γ sono stati posti uguali a $a_0 = b_0 = 1$ (per dare una distribuzione a priori uniforme sull'intervallo $[0, 1]$).

Come si può notare dai grafici di autocorrelazione in Figura (4.1), le catene sono fortemete autocorrelate (eccetto il θ, ma si può ricondurre la sua non autocorrelazione al Punto 5 dello schema di campionamento della MCMC descritto nel precedente Capitolo). I grafici della funzione di autocorrelazione sono stati calcolati con il comando `codamenu()` del pacchetto, di `R`, `coda`. Come si può notare dal grafico (4.1b), già con un thinning di 10 si ha un'autocorrelazione bassa dopo pochi passi, unica eccezione l'intercetta $\beta_0 = \mu$. Decido quindi, da ora in poi, di fare sempre un thinning a 10 per passare da un campione originale di 50.000 a uno con 5.000 passi della catena del Gibbs Sampler.

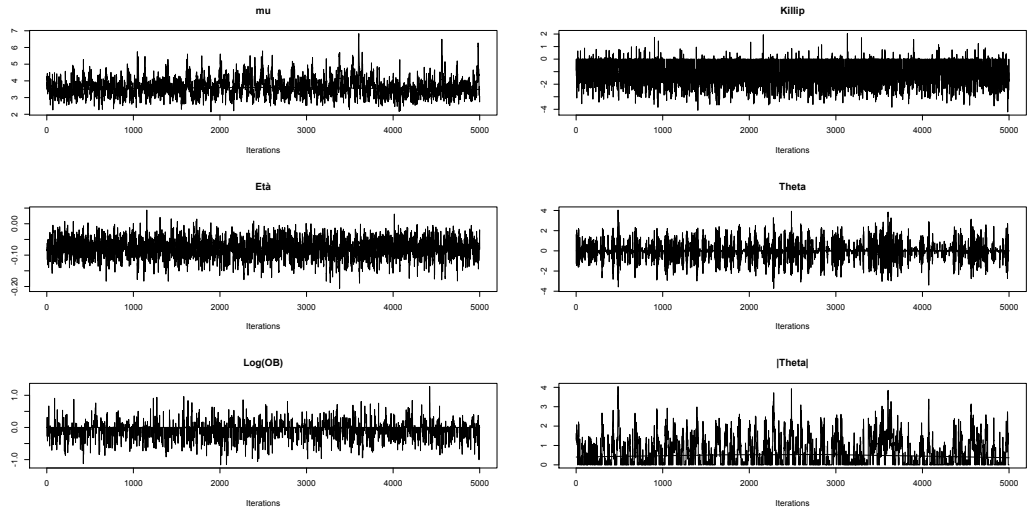
Prima di continuare con l'analisi infernziale dei risultati è opportuno uno studio di diagnostica di convergenza delle catene. Per questo tipo di analisi si è deciso di utilizzare quattro diverse strade: lo studio qualitativo delle *trace* delle catene generate, il test di *Geweke*, i due test di *stazionarietà*

di Heidelberger e Welch e il test di stazionarietà dei quantili di *Raftery e Lewis*. Per una breve descrizione di questi test si rimanda all'appendice B. Ad eccezione del β_2 (LogOB) e del θ tutti i test confermano la stazionarietà delle MCMC.



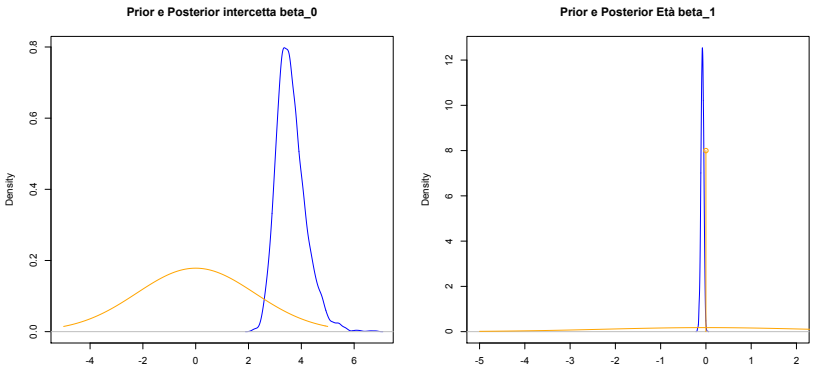
(a) Senza Thinning (b) Con Thinning

Figura 4.1: Autocorrelazione del modello di base

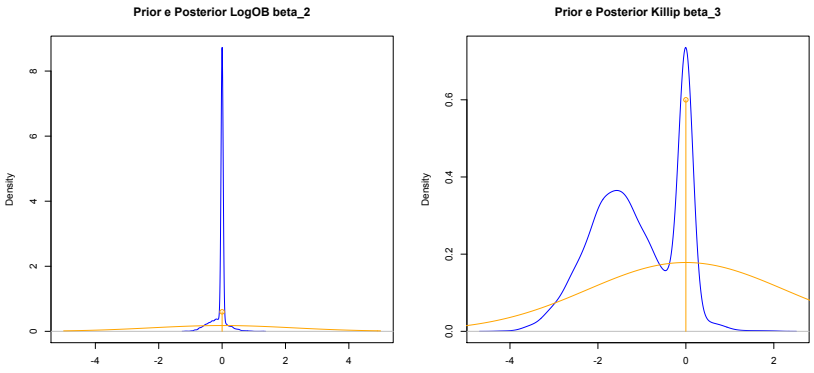


(a) Tracce β_0, β_1, β_2 (b) Tracce β_3, θ, $|\theta|$

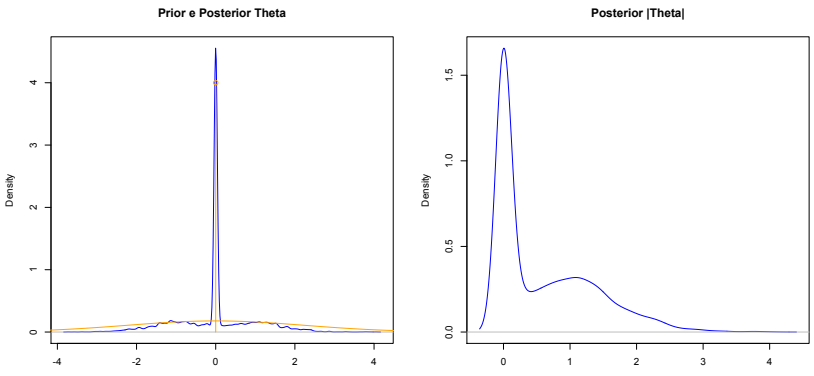
Figura 4.2: Trace dei β_j di θ e di $|\theta|$



(a) Intercetta (b) β_1 Età



(c) β_2 LogOB (d) β_3 Killip



(e) θ (f) $|\theta|$

Figura 4.3: Prior (arancione) e Posterior (blu) dei regressori ad effetti fissi e variabili

Come si può notare dai grafici delle densità a posteriori empiriche (linee blu) in Figura (4.3) e dalla tabella degli intervalli HPD, il β_2 associato alla variabile *Log(OB)* è concentrato molto sullo zero. Mentre β_1 della variabile Età si concentra su un valore diverso da zero. La prima impressione sarebbe quella di escludere dal modello finale la covariata LogOB e di accettare l'Età. Se consideriamo anche la Tabella (4.2) nella quale sono presenti i valori della probabilità a posteriori di avere allocato il β_j alla componente slab, si può affermare con decisione che la variabile LogOB non è da tenere in considerazione, le variabili Età e Killip invece hanno un peso rilevante nel modello finale.

	Mean	SD
β_0 intercetta	3.613594	0.56045
β_1 =Età	-0.078672	0.03138
β_2 = log(OB)	-0.045561	0.20167
β_3 =Killip	-1.098332	0.97515
θ	-0.003937	0.93570
$\|\theta$	0.583631	0.73133

(a) Medie e dev.std

	lower	upper
$\beta_0 = \mu$	2.65910	4.779900
β_1=Età	-0.14055	-0.016517
β_2 =Log(OB)	-0.62913	0.310670
β_3 =Killip	-3.05700	0.012674
θ	-2.02010	2.034300
$\|\theta\|$	0.00000	2.029500

(b) HPD

Tabella 4.1: Tabelle di media , dev.std. e Highest Posterior Density intervals

δ_1(Età)	1.0000
δ_2(logOB)	0.2984
δ_3(Killip)	0.7332
$\theta = \|\theta\|$	0.5276

Tabella 4.2: Probabilità a posteriori di essere nella componente slab

Caso ambiguo è il θ, la varianza della intercetta aleatoria (l'effetto aleatorio della struttura ospedaliera), la sua densità a posteriori è stata generata in modo differente dai quella dei β_j. Ricordando infatti il Punto 5 dell'Algoritmo, dopo aver campionato il θ gli si cambia il segno algebrico con probabilità $1/2$ generando così una densità a posteriori simmetrica rispetto all'asse 0. Si è deciso di prendere in considerazione il suo valore assoluto. Analizzando il dato in Tabella (4.2), la componente slab e la componente spike a posteriori di θ hanno circa lo stesso *peso* empirico, che ci pone in un caso di indecisione. Abbiamo comunque scelto di non assumere $\theta = 0$ supponendo, grazie agli studi precedentemente fatti in Guglielmi et al. (2010), che il suo valore sia effettivamente diverso da 0.

4.2.2 Robustezza del modello rispetto alla varianza a priori

Proseguendo nello studio del modello utilizzato, ci si è posti la questione sulla robustezza delle simulazioni al variare degli iperparametri di varianza a priori di $\boldsymbol{\beta}$ e di θ. Si è pensato quindi di fare una serie di simulazioni con diversi valori di $A_{0,jj} \in 10^{\{-1.2,-1,-0.8,\ldots,+3.8,+4\}}$.

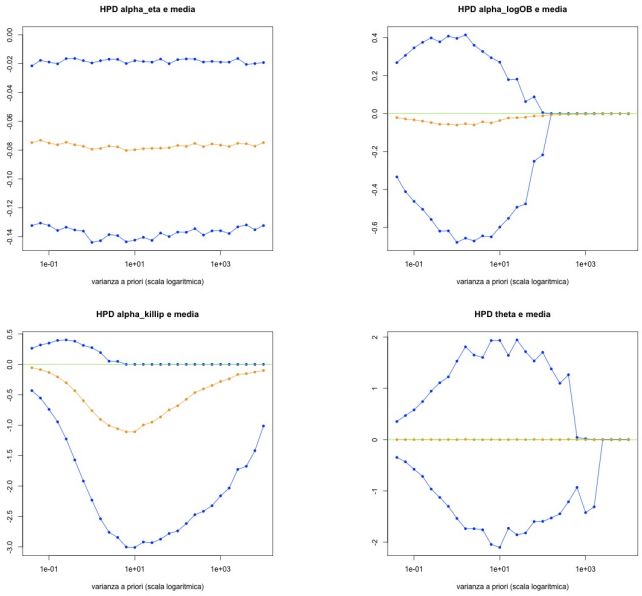


Figura 4.4: Robustezza HPD e medie rispetto a $A_{0,jj}$ varianza a priori

La variabile *Età* non cambia sostanzialmente i suoi valori a posteriori al variare del parametro A_0; *Log(OB)* sembra inizialmente aumentare il suo intervallo di credibilità (ma la sua media rimane molto vicina allo zero) tendendo poi ad un collassamento verso la delta di Dirac; anche per la covariata *Killip* abbiamo un iniziale aumento dell'HPD che dopo il valore $A_0 \approx 10$ sembra decrescere.

Per la variabile θ il discorso deve essere fatto con più attenzione per via del Punto 5 dell'Algoritmo già precedentemente citato.

Riportiamo anche i grafici della probabilità di venire assegnati alla componente *slab* (cioè la probabilità di essere diversi da 0).

Tralasciando il risultato del regresore della covariata *Età*, che rimane inalterato per ogni valore di A_0, è possibile notare come gli altri tre valori dei β_j (relativi alle covariate *Log(OB)* e *Killip*) e di θ (varianza dell'intercetta

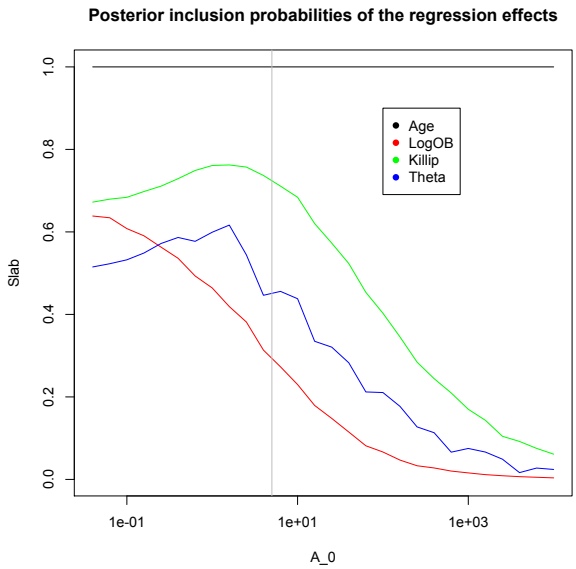


Figura 4.5: Probabilità di essere nella componente SLAB rispetto dell'iperparametro $A_{0,jj}$ $(j = 1, .., 4)$. Grafico in scala logaritmica. Linea verticale in $A_{0,jj} = 5$

aloatoria) decrescano velocemente al crescere della varianza a priori. Inoltre si può notare l'esistenza di un livello di saturazione nei δ_3 del *Killip* e in γ. La domanda che ci siamo posti è stata quella di capire cosa provocasse questa veloce discesa in funzione della varianza quando quest'ultima è più grande del valore 5.

Questa situazione ha una spiegazione. È uno dei problemi, già riscontrati in altri articoli, sull'utilizzo del metodo *Dirac spike-slab prior* con varianze troppo ampie. E' noto in letteratura come *Paradosso di Lindley*. E' una situazione, resa celebre in Lindley (1957), nella quale vi è un disaccordo tra i risultati dei test d'ipotesi nella teoria frequentista da quelli bayesiani.

La differenza tra risultati delle due teorie nasce dalla scelta non opportuna della prior. Supponiamo infatti di avere un test statistico con ipotesi nulla *non accetto la variabile* e alternativa *l'accetto*, se la prior assegna un picco sull'ipotesi nulla e un'ampia distribuzione sul resto dei valori questo può portare a risultati non buoni. L'ipotesi nulla non verrà quasi mai rigettata.

Nel nostro caso, avendo ipotesi nulla $\beta_i = 0$; più si aumenta la varianza della componente slab più il picco in zero (dato dalla componente Dirac) avrà un forte peso sul test statistico, non facendo mai rifiutare l'ipotesi nulla. Una distribuzione *spike-slab* con grande varianza nella componente *slab*, concentra

grande massa in tre punti: 0, $+\infty$ e $-\infty$. Questo porta il modello bayesiano a scegliere sempre il valore nullo per i β_i. Perciò come suggerito da Malsiner-Walli and Wagner (2011), bisogna trovare un giusto equilibrio nella scelta della varianza a priori delle componenti *slab*, magari scegliendo una varianza dello stesso ordine di grandezza del valore della soluzione ai minimi quadrati rdinari. Dal grafico in Figura 4.5 si è scelto, in questo progetto di tesi, di utilizzare una varianza a priori pari a 5.

4.2.3 Robustezza rispetto ai parametri $a_0 = b_0$ e $A_{0,jj}$

Come fatto precedentemente, abbiamo cercato di studiare la robustezza delle stime fatte, facendo variare i valori degli iperparametri delle variabili di supporto ω_δ e ω_γ su 4 valori

$$a_0 = b_0 = \left\{\frac{1}{2}, 1, 2, 5\right\}$$

e per ogni coppia di valori di a_0 e b_0, tenendo in considerazione i risultati della sezione precedente, abbiamo scelto:

$$A_{0,jj} = \{5, 6, ..., 24, 25\} \quad j = 1, ..., 4.$$

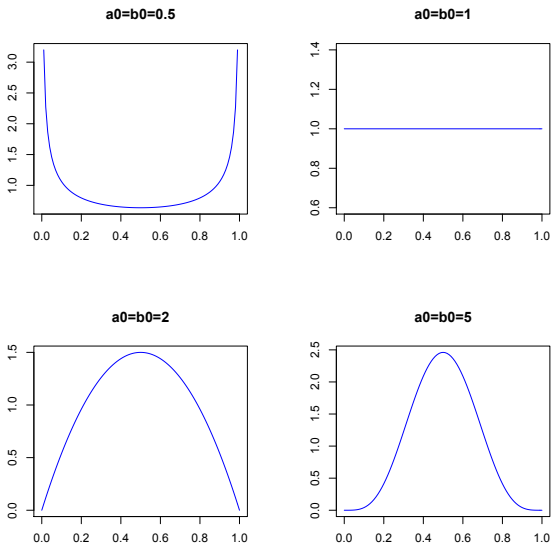
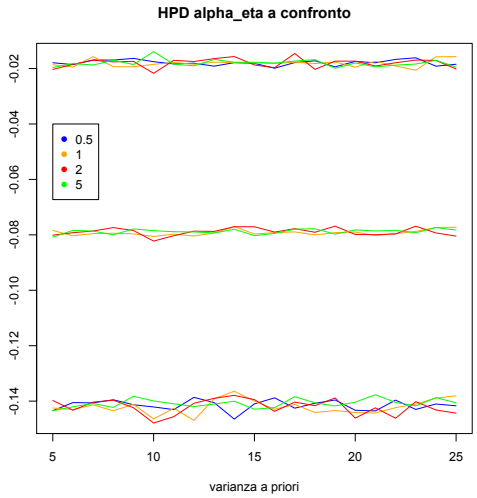
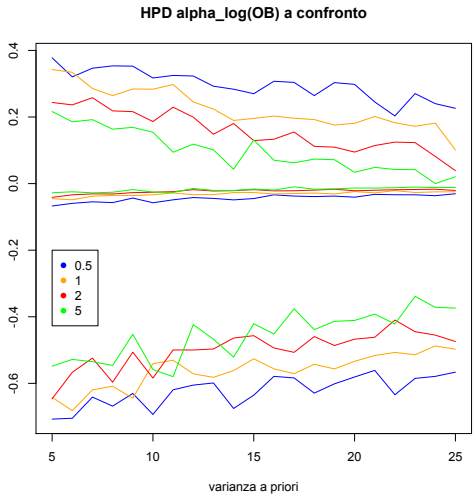


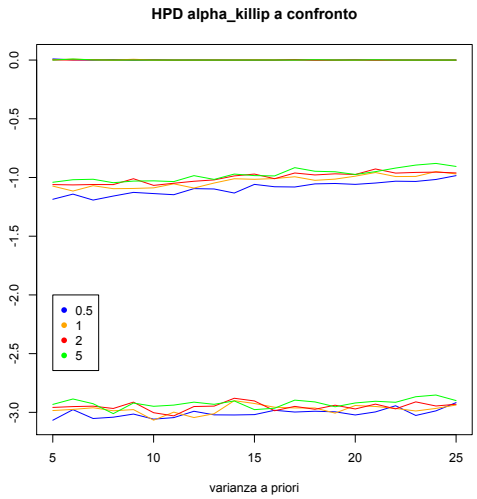
Figura 4.6: Prior Beta dei pesi ω_δ e ω_γ al variare di a_0 e b_0



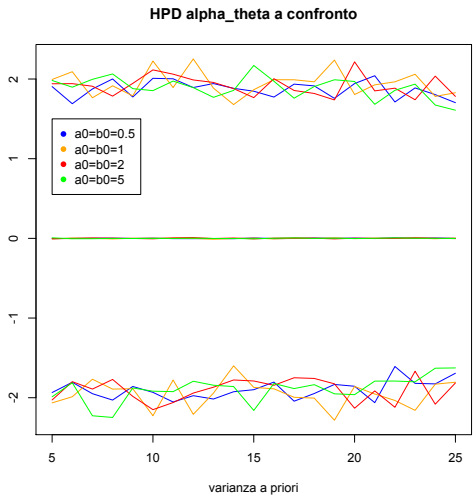
(a) β_1 regressore dell'Età



(b) β_2 regressore LogOB



(c) β_3 di Killip



(d) θ

Figura 4.7: HPD e medie a confronto al variare di A_0 (asse delle ascisse) e di a_0 e b_0 (nei diversi colori)

Nelle Figure (4.7) e (4.8) è possibile vedere i risultati ottenuti. La variazione degli iperparametri dei pesi ω_δ e ω_γ non sembra avere effetti rilevanti sull'inferenza statistica né degli intervalli di credibilità a più alta densità a posteriori né sulla media e nemmeno sulla probabilità di essere allocati alla componente slab della posterior. Come già precedentemente notato, si hanno

variazioni significative quando si fa variare l'iperparametro A_0. Abbiamo deciso di assegnare ad a_0 e b_0 il valore 1 per tutte le simulazioni successivamente.

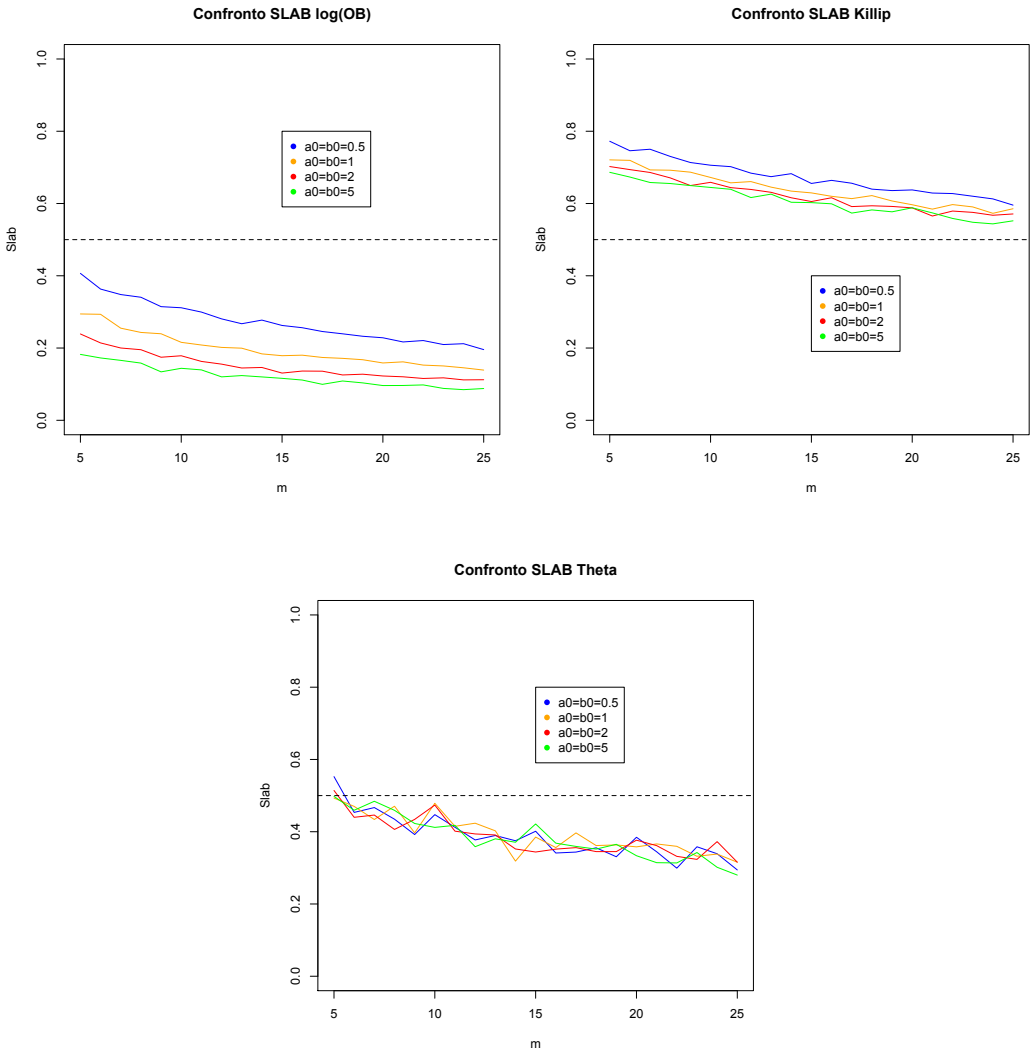


Figura 4.8: Probabilità di essere nella compenente SLAB al variare di A_0(asse delle ascisse), e al variare di a_0 e b_0 (nei diversi colori) a confronto

4.3 Modello con sole covariate numeriche

Dopo aver studiato i risultati ottenuti prendendo in considerazione le sole variabili a effetti fissi Età, LogOB e Killip e la variabile ospedale come effetto aleatorio, aumentiamo la visione del dataset prendendo in considerazione tutte quelle variabili che possono essere espresse con un valore numerico. Il modello considerato è stato:

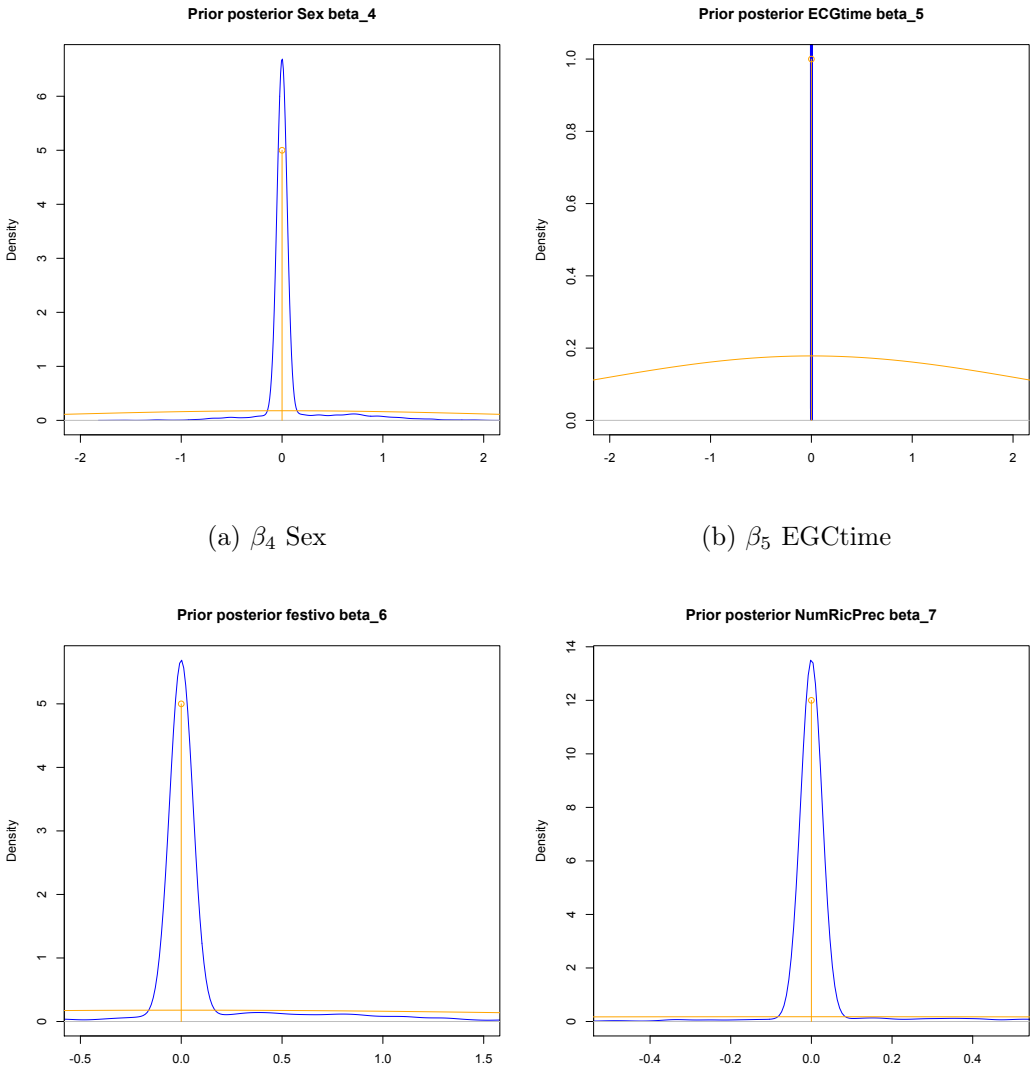
$$\begin{aligned}
&Y_i|p_i \sim \text{Be}(p_i) \qquad i = 1, ..., 240; \\
&\text{logit}(p_i) = \beta_0 + \beta_1 ETA + \beta_2 LOG(OB) + \beta_3 KILLIP + \\
&\qquad + \beta_4 SEX + \beta_5 ECGtime + \beta_6 FESTIVO + \\
&\qquad + \beta_7 NUMricPREC + \theta \widetilde{b}_{c(i)}, \\
&\widetilde{b}_{c(i)} \sim \mathcal{N}(0, 1), \quad c(i) = 1, ..., 17; \\
&\beta_0 \sim \mathcal{N}(m_0, M_0), \\
&\beta_j \sim (1 - \omega_{\delta_j})\Delta_0(\beta_j) + \omega_{\delta_j}\mathcal{N}(0, A_{0,jj}) \quad j = 1, ..., 7; \\
&\theta \sim (1 - \omega_\gamma)\Delta_0(\gamma) + \omega_\gamma \mathcal{N}(0, A_{0,88}), \\
&\omega_{\delta_j} \sim \mathcal{B}eta(a_0, b_0) \quad j = 1, ..., 7; \\
&\omega_\gamma \sim \mathcal{B}eta(a_0, b_0).
\end{aligned} \tag{4.2}$$

Come nel caso a sole tre covariate, abbiamo generato catene con 70.000 passi, un burn-in di 20.000 e un thinning tale da avere un campione di 5.000 dati. I test di diagnostica di convergenza non hanno rilevato problemi e quindi si è passati all'analisi inferenziale. Non essendo presente alcuna differenza notevole tra i risultati presentati nella precedente sezione sulle variabili Età, LogOB e Killip ci soffermiamo nello studiare solamente i β_j relativi alle covariate aggiunte.

Come si può notare dai grafici delle posterior in Figura (4.9), tutte le densità sono concentrate intorno allo zero. Il regressore della variabile *ECGtime* invece è totalmente allocato alla delta di Dirac la cui probabilità a posteriori di essere allocato alla componente slab è circa zero. I valori della tabella degli intervalli HPD conferma l'idea che nessuna di questa variabili aggiuntive debba essere inserita al modello finale e quindi non rifiutiamo l'ipotesi di $\beta_j = 0$.

Casi interessanti sono però il β_3 relativo alla covariata Killip e il θ. Dalla Tabella 4.4 si può notare come i valori dei relativi δ_3 e γ siano passati al di sotto di 0.5. Si potrebbe pensare di non accettare in questo caso le relative covariate nel modello finale. Ricordiamo però che, avendo rifiutato tutte le variabili aggiuntive si ricadrebbe nel modello già studiato precedentemente. Questa forte differenza tra i risultati ottenuti nei due casi, ci porta a pensare

che vi siano delle covariate correlate e non indipendenti che, se aggiunte al modello, ne cambiano i risultati.



(a) β_4 Sex

(b) β_5 EGCtime

(c) β_6 Festivo

(d) β_7 NUMricPREC

Figura 4.9: Prior (arancione) e Posterior (blu) dei coefficienti di regressione delle sole variabili numerice

	Mean	SD
$\beta_0 = \mu$	3.445382	0.7116593
β_1 Età	-0.077899376	0.0311643912
β_2 logOB	-0.01334277	0.1124870116
β_3 Killip	-0.7105335	0.9394166243
β_4 Sex	0.6187	0.3016234852
β_5 ECGtime	-0.00000211	0.0003476845
β_6 festivo	0.09649337	0.3335218848
β_7 NumRicPrec	0.02529395	0.1626753549
θ	-0.02250	0.7826447214
$\lvert\theta\rvert$	0.4694079	0.6397885904

(a) Medie e dev.std.

lower	upper
1.94250	5.002400
-0.13904	-0.016666
-0.34764	0.112070
-2.67930	0.000000
-0.29044	1.056500
0.00000	0.000000
-0.22737	1.171900
-0.12520	0.440740
-1.69620	1.774200
0.00000	1.751500

(b) HPD intervals

Tabella 4.3: Medie dev.std e intervalli HPD con variabili numeriche

δ_{eta}	1
δ_{logOB}	0.0996
δ_{killip}	0.4836
δ_{sex}	0.1758
$\delta_{ECGtime}$	0.0036
$\delta_{festivo}$	0.1918
$\delta_{numricPrec}$	0.1078
γ	0.4744

Tabella 4.4: Probabilità di essere nella componente slab

4.4 Modello con covariate numeriche e categoriche

Dopo aver studiato il modello con tutte le variabili numeriche e avendo notato quanto l'aggiunta di nuove covariate possa influenzare i risultati, consideriamo ora il modello generalizzato con tutte le covariate presenti nel dataset iniziale del progetto MOMI della Regione Lombardia.

Avendo aggiunto al modello covariate categoriche, non ordinali, ognuna delle quali ha diverse etichette, si è deciso di utilizzare le *dummy variables*.

Supponiamo di avere una variabile categorica con k livelli e ne scegliamo uno di riferimento. Nel modello di regressione si aggiungeranno $k-1$ variabili dummy relativi ad ognuno dei restanti livelli. Sia d_l con $l = 1, .., k-1$ la colonna relativa all'l-esima etichetta, alla posizione i vi sarà il valore 1

se l'unità statistica i-esima ha l'eticheta l-esima e 0 altrimenti. Per ogni variabile categorica infine bisogna aggiungere un'intercetta che si aggiungerà all'intercetta generale β_0.

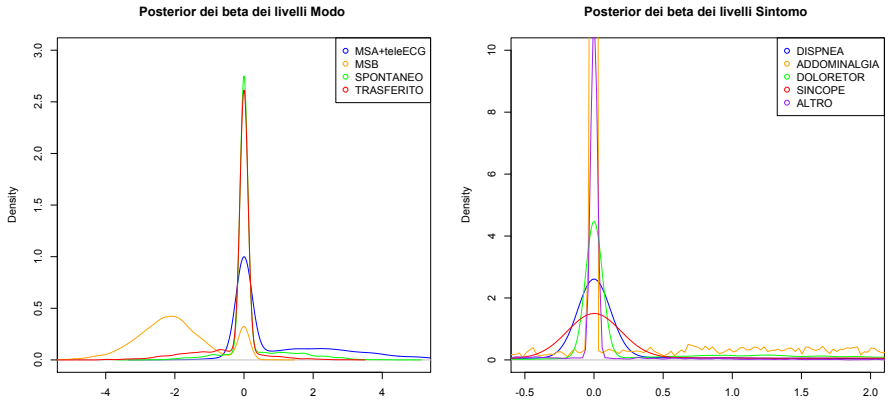
Il nostro modello dovrà quindi essere scritto come segue:

$$
\begin{aligned}
Y_i|p_i &\sim \text{Be}(p_i) \qquad i = 1, ..., 24; 0 \\
\text{logit}(p_i) &= \beta_0 + \beta_1 ETA + \beta_2 LOG(OB) + \\
&+ \beta_3 KILLIP + \beta_4 SEX + \\
&+ \beta_6 FESTIVO + \beta_7 NUMricPREC + \\
&+ (\beta_8 MSAteleECG + \beta_9 MSB + \beta_{10} SPONTANEO + \\
&\quad + \beta_{11} TRASFERITO)_{modo} \\
&+ (\beta_{12} ADDOMINALGIA + \beta_{13} ALTRO + \beta_{14} DISPNEA + \\
&\quad + \beta_{15} DOLORETOR + \beta_{16} SINCOPE)_{sintomo} + \\
&+ (\beta_{17} BBS + \beta_{18} INFPOST)_{sede} + \\
&+ (\beta_{19} PS + \beta_{20} UTIC)_{fast.track} + \\
&+ (\beta_{21} NO + \beta_{22} PRIMARIA + \beta_{23} RESCUE)_{ptca} + \\
&+ (\beta_{24} PREH + \beta_{25} SI)_{trombolisi} + \theta \widetilde{b}_{c(i)}, \\
\widetilde{b}_{c(i)} &\sim \mathcal{N}(0, 1), \quad c(i) = 1, ..., 17; \\
\beta_0 &\sim \mathcal{N}(m_0, M_0), \\
\beta_j &\sim (1 - \omega_{\delta_j})\Delta_0(\beta_j) + \omega_{\delta_j}\mathcal{N}(0, A_{0,jj}) \quad j = 1, ..., 31; \\
\theta &\sim (1 - \omega_\gamma)\Delta_0(\gamma) + \omega_\gamma \mathcal{N}(0, A_{0,32-32}), \\
\omega_{\delta_j} &\sim \mathcal{B}eta(a_0, b_0) \quad j = 1, ..., 31; \\
\omega_\gamma &\sim \mathcal{B}eta(a_0, b_0),
\end{aligned}
\tag{4.3}
$$

dove si è deciso di assegnare come variabili di riferimento: il livello MSA per la covariata *modo*, ACC per *sintomo*, ANTLAT per *sede*, EMO per *fast.track*, ELETTIVA per *ptca* e NO per *trombolisi*.

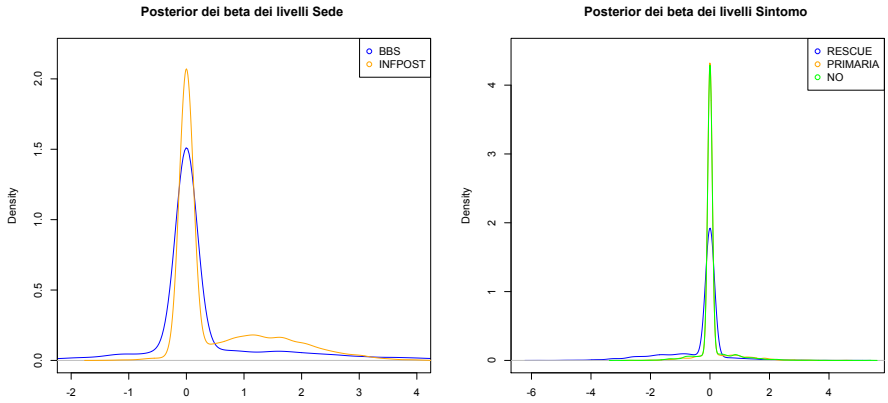
$\delta_{Eta'}$	1.0000000
δ_{Killip}	0.5732853
$\delta_{modo.MSAteleECG}$	0.4441112
$\delta_{modo.MSB}$	0.8698260
γ	0.6114777

Tabella 4.5: Tabella con le sole probabilità al di sopra del 0.4



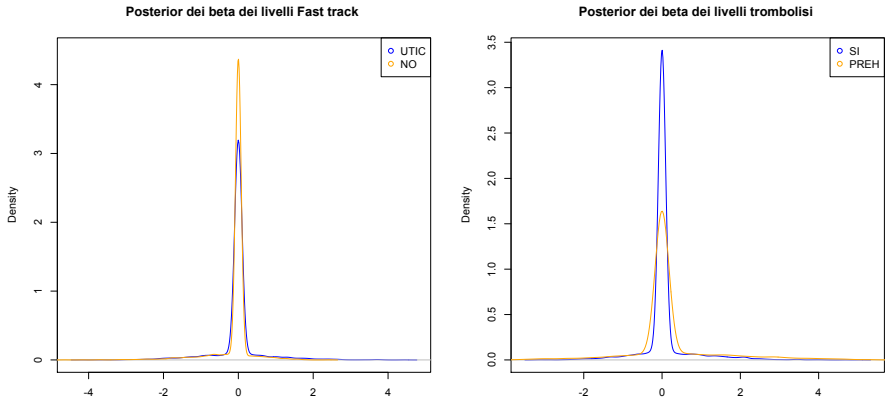
(a) β Modo

(b) β Sintomo



(c) β Sede

(d) β PTCA



(e) β fast track

(f) β Trombolisi

Figura 4.10: Posterior delle etichette per ogni variabile categorica

Dalla Tabella 4.5 si può notare che, come detto nei paragrafi precedenti, le variabili prese in considerazione sono l'*Età* il *Killip* e il θ varianza della intercetta aleatoria. Tra tutte le variabili categoriche l'unica che possa essere presa in considerazione è la variabile *Modo*. Osservando anche le densità a posteriori dei livelli di ogni variabile categorica si può notare come tutte si concentrino sul valore zero. I grafici potrebbero ingannare in quanto sembra siano delle gaussiane centrate in 0, ma ciò deriva dal fatto che si è utilizzato un comando di `R` che interpola con un kernel density di tipo gaussiano.

4.5 Conclusioni

Dopo questo studio si può affermare che le variabili selezionate tramite questo modello sono: l'Età, il Killip, la variabile Modo ed infine vi è un effetto causale dato dai centri ospedalieri di varianza θ^2.

$$\begin{aligned} \text{logit}(p_i) &= 3.697105 - 0.07536544 * Eta - 1.081306 * Killip + \\ &\quad + 1.000195 * Modo(MSA + teleECG) - 1.969107 * Modo(MSB) + b_{c(i)} \\ b_{c(i)} &\sim \mathcal{N}(0, (0.004750056)^2) \end{aligned} \tag{4.4}$$

Dove per Età è da interndersi: una persona con valore 0 è 64 enne , un 80-enne avrà invece un valore $80 - 64 = 16$.

Dall'analisi dei risultati ottenuti, è possibile osservare che la probabilità di rimanere in vita diminuisce all'aumentare dell'età e della gravità dell'infarto. Inoltre le diverse modalità con cui il paziente entra in ospedale hanno sul tasso di sopravvivenza effetti differenti: poter essere trasportati da un mezzo avanzato con un apparecchio per teletrasmettere l'esame ECG aumenta la possibilità di sopravvivenza mentre il caso peggiore coincide con il trasporto in un'ambulanza in cui non è prevista la presenza di personale medico o paramedico. Infine è stata rilevata una dipendenza delle risposte dalla struttura ospedaliera considerata. Come già detto all'inizio del capitolo, durante la descrizione del dataset, si può notare come l'intercetta *fissa* sia molto alta. Nel caso di un paziente 64 enne con un *Killip* basso, arrivato con un automezzo proprio la probabilità di sopravvivere è:

$$p_i \simeq \frac{e^{3.7}}{1 + e^{3.7}} \simeq 0.97.$$

Il solo fattore incidente nella sopravvivenza dei pazienti che possa essere modificato, quindi migliorato, è quello rappresenato dalla variabile *modo*. Se

si riuscisse ad informare i cittadini quanto sia fondamentale la descrizione dei sintomi durante una chiamata all'unità di emergenza, l'operatore del *118* o delle unità di *Emergenza* sarebbe consapevole della gravità dell'evento di infarto e potrebbe richiedere l'automezzo più idoneo al caso.

Nella Sezione 4.3 si è accennato a come il modello cambi i propri valori stimati (specialmente i pesi ω_{δ_j} di essere allocati alla componente *slab*) se si aggiungono o meno delle variabili. Rileggendo le covariate si può supporre come alcune di queste siano ridondanti. L'esistenza o meno di una *fast track* è un servizio proprio di pronto soccorso, gli effetti che questa variabile apporta nel modello saranno quindi già interamente descritti dalla intercetta alatoria che differenzia gli ospedali.

Possiamo affermare, grazie alla rilettura bayesiana fatta nella Sezione 1.5, che il modello *GLMM* bayesiano utilizzato per analizzare il dataset *MOMI* in questo elaborato di tesi, è una estensione della *Ridge regression*. Dove la Ridge è presente nella componente slab della distribuzionea a priori e si ha una delta di Dirac nella componente Spike.

Appendice A

Notazione

In questa appendice vengono elencate ed esplicitate tutte le notazioni e tutte le funzioni di distribuzione di probabilità utilizzate nel presente elaborato.

A.1 T-Student a 3 parametri

Una t-Student a tre parametri con μ parametro di locazione e λ parametro di scala e ν gradi di libertà è distribuita secondo la seguente legge:

$$\begin{aligned} & X \sim t(\mu, \lambda, \nu) \qquad , \nu > 2 \\ & f_X(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)} \left(\frac{\lambda}{\pi\nu}\right)^{1/2} \left[1 + \frac{\lambda(x-\mu)^2}{\nu}\right]^{-\frac{\nu+1}{2}} \end{aligned} \tag{A.1}$$

con media μ e varianza $\frac{1}{\lambda}\frac{\nu}{\nu-2}$

A.2 Esponenziale e Gamma

Le distribuzioni esponenziale e gamma utilizzate in questo elaborato hanno la seguente parametrizzazione

$$\begin{aligned} & X \sim \mathcal{E}(\lambda) \qquad f_X(x) = \lambda e^{-\lambda x} \mathbb{I}_{\{x \geq 0\}} \\ & \mathbb{E}[X] = \frac{1}{\lambda} \qquad Var[X] = \frac{1}{\lambda^2} \end{aligned} \tag{A.2}$$

e

$$\begin{aligned} & X \sim \text{Gamma}(\alpha, \beta) \qquad f_X(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} \mathbb{I}_{\{x \geq 0\}} \\ & \mathbb{E}[X] = \frac{\alpha}{\beta} \qquad Var[X] = \frac{\alpha}{\beta^2} \end{aligned} \tag{A.3}$$

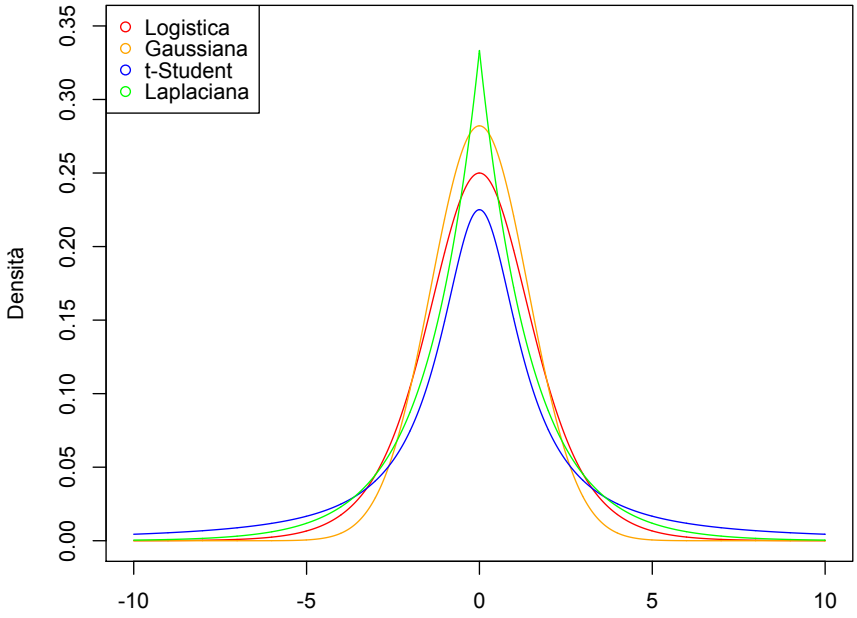


Figura A.1: Confronto tra le distribuzioni

A.3 Gaussiana inversa e Gamma inversa

La distribuzione gaussiana inversa a due parametri μ e λ ha la seguente distribuzione di probabilità:

$$\begin{aligned} &X \sim \text{Inv-Gaussian}(\mu, \lambda) \\ &f_X(x) = \sqrt{\frac{\lambda}{2\pi}} x^{-3/2} \exp\left\{ \frac{-\lambda(x-\mu)^2}{2\mu^2 x} \right\} \mathbb{I}_{\{x>0\}} \end{aligned} \tag{A.4}$$

media μ e varianza μ^3/λ.
La Gamma inversa di parametri α e β ha come *pdf*:

$$\begin{aligned} &X \sim \text{Inv-Gamma}(\alpha, \beta) \\ &f_X(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{-(\alpha+1)} e^{-\frac{\beta}{x}} \mathbb{I}_{\{x>0\}} \end{aligned} \tag{A.5}$$

con media $\frac{\beta}{\alpha-1}$ e varianza $\frac{\beta}{(\alpha-1)^2(\alpha-2)}$.

A.4 Doppia Esponenziale o Laplaciana

La distribuzione doppia esponenziale di parametri μ e b ha distribuzione di probabilità:

$$\begin{aligned} &X \sim \text{Laplace}(\mu, b) \\ &f_X(x) = \frac{1}{2b} e^{-\frac{|x-\mu|}{b}} \end{aligned} \tag{A.6}$$

com media μ e varianza $2b^2$

A.5 Distribuzione Logistica

Una variabile aleatoria X con media μ e varianza $\frac{\pi^2}{3}s^2$ ha distribuzione logistica se:

$$\begin{aligned} &X \sim \text{Logistic}(\mu, s) \\ &f_X(x) = q \\ &F_X(x) = \mathbb{P}(X \leq x) = \frac{1}{1 + e^{-\frac{x-\mu}{s}}} \end{aligned} \tag{A.7}$$

A.6 Distribuzione Delta di Dirac

Distribuzione che assegna tutta la massa di probabilità su un singolo valore:

$$\Delta_a(x) = \begin{cases} 1, & x = a \\ 0, & altrimenti \end{cases} \tag{A.8}$$

è la delta di Dirac sulla variabile x che vale 1 se $x = 1$ e zero altrimenti.

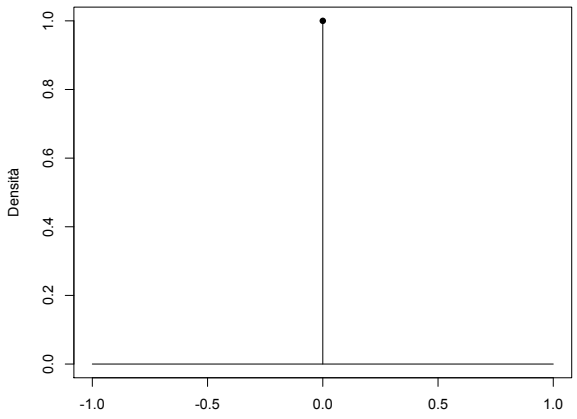


Figura A.2: Delta di Dirac centrata in 0

Appendice B

Diagnostica di Convergenza

La diagnostica di convergenza è un'analisi da effettuare ogni qual volta si ha a che fare con algoritmi di tipo *Markov Chains Monte Carlo* (MCMC), o più in generale con catene di Markov. Questo tipo di studio serve ad avere la conferma statistica che le iterazioni, ovviamente in numero finito, considerate portino a conclusioni corrette raggiungibili solo con infinite iterazioni. In questa appendice verranno presentati 2 diversi test.

B.1 Test Geweke

Sia data una MCMC $\{X_j\}$ di lunghezza n e avente un burn-in di lunghezza m (in totale si avrà quindi un numero di passi della catena pari a $n+m$). Estraggo da questa catena n_b passi iniziali e n_a passi finali e calcolo la media di queste due finestre della catena markoviana.

$$\overline{X}_b = \frac{1}{n_b} \sum_{j=m+1}^{m+n_b} X_j$$

$$\overline{X}_a = \frac{1}{n_a} \sum_{j=m++n-n_a+1}^{m+n} X_j$$

Se la catena ha comportamento ergodico, ci si aspetta che al crescere di n (lasciando però invariate le frazioni n_a/n e n_b/n) il comportamento delle medie ergodiche sopra definite sia uguale:

$$Z_G = \frac{\overline{X}_b - \overline{X}_a}{\sqrt{\hat{Var}(X_a) + \hat{Var}(X_b)}} \xrightarrow[n\to\infty]{Legge} \mathcal{N}(0,1) \tag{B.1}$$

Di seguito riporto i grafici fatti con il comando `codamenu()` del pacchetto `coda` implementato in `R`.

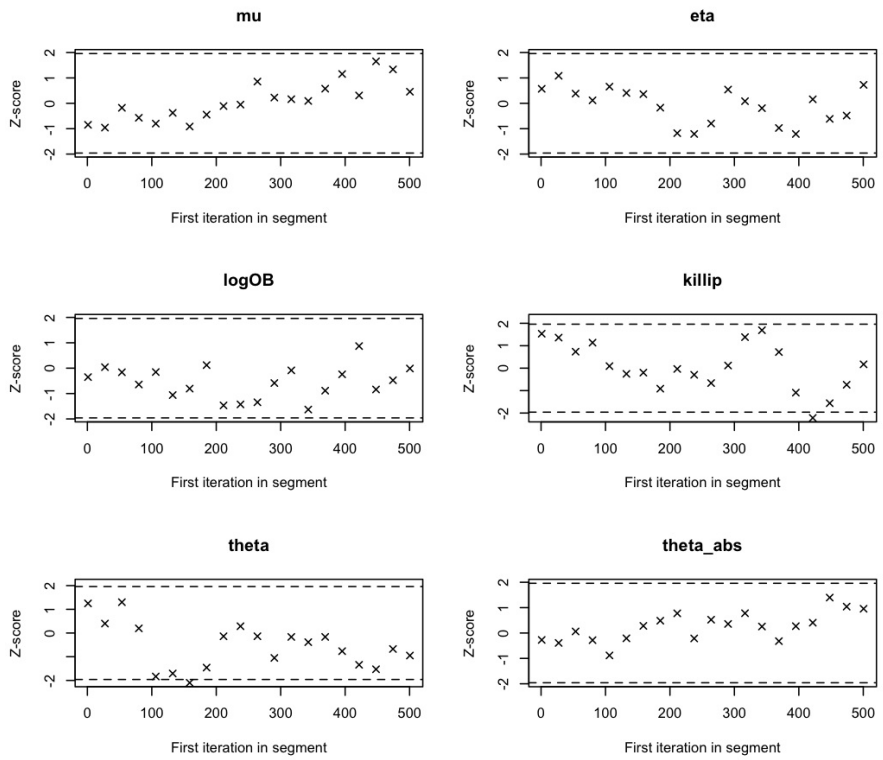


Figura B.1: Diagnostica di Convergenza delle catene con Geweke del modello ristretto

Nel modello con le sole variabili Età, Log(OB) e Killip non rifiuto mai l'ipotesi di convergenza delle catene tramite il *test di Geweke*. I relativi β_j e θ sono giusti a convergenza con 5000 iterazioni.

B.2 Heidelberg and Welch Diagnostic

Il test è composto da due parti: una prima di stazionarietà e una seconda detta *Half-Width test* che controlla se la lunghezza della catena di Markov generata sia sufficiente per garantire una stima accurata degli stimatori. Sia $\{X_j\}_{j=1}^n$ la catena di Markov, definito $S_n = \sum_{j=1}^n X_j$ e $\overline{X} = \frac{1}{n}S_n$ e $\hat{p(0)}$ la stima della densità spettrale alla frequenza 0 , costruisco la sequenza:

$$B_n(s) = \frac{S_{[ns]} - [ns]\overline{X}}{\sqrt{n\hat{p}(0)}} \quad s \in \left\{\frac{1}{n}, \frac{2}{n}, ..., 1\right\} \tag{B.2}$$

dove con $[\cdot]$ è l'operatore parte intera. Per n grande, B_n converge in distribuzione a un *ponte browniano*. La statistica utilizzata in questo test è

quella definita nella procedura di *Cramer-Von Mises*. Calcolo lo stimatore utilizzando tutta la catena, se l'ipotesi nulla di stazionarietà viene rigettata a favore dell'ipotesi alternativa, si escude dalla catena il primo 10% dei passi generati e si esegue nuovamente il test. Se anche a questo punto l'ipotesi alternativa viene rifiutata si continua ad escludere una parte della catena. Si ripete ciò fino a quando il test non rifiuta l'ipotesi alternativa oppure fino a quando non è stato escluso dal test il 50% dei passi generati.

Se la catena *passa* il primo stadio del test, la parte della catena che è stata identificata come stazionaria, quindi ciò che ne resta dopo l'esclusione dei primi passi, vine sottoposta al *Half-Width test*. In questo test si calcola il cosidetto *relative half-width* (RHW) dell'intervallo di credibilità di livello $(1-\alpha)$. Per maggiori informazioni si consulti Heidelberger and Welch (1981).

Di seguito riportiamo i risultati del test applicato al dataset formato dalle sole covariate *numeriche*

```
HEIDELBERGER AND WELCH STATIONARITY AND INTERVAL HALFWIDTH TESTS
================================================================

Iterations used = 1:5000
Thinning interval = 1
Sample size per chain = 5000

Precision of halfwidth test = 0.1

$chain1

           Stationarity start     p-value
           test         iteration
mu         passed           1     0.806989
eta        passed           1     0.689746
logOB      passed        1502     0.260765
killip     passed           1     0.566482
sex        passed           1     0.544180
ECGtime    passed        2002     0.134649
festivo    passed           1     0.848878
numricPrec passed           1     0.937906
theta      failed          NA     0.000457

           Halfwidth Mean      Halfwidth
           test
mu         passed     3.45e+00 3.72e-02
```

```
eta         passed    -7.79e-02 1.58e-03
logOB       failed    -9.11e-03 3.78e-03
killip      passed    -7.25e-01 3.22e-02
sex         failed     6.20e-02 1.06e-02
ECGtime     failed     9.19e-06 1.13e-05
festivo     failed     9.65e-02 1.45e-02
numricPrec failed      2.52e-02 5.85e-03
theta       <NA>             NA       NA
```

La prima parte del test, come possiamo notare la passano tutte le variabili con eccezion fatta dal θ: le catene sono giunte a stazionarietà dopo 5000 iterazioni. La seconda parte del test, quella detta *Half-Width test* invece viene passata solamente da quelle variabili che alla fine dello studio verranno prese in coniderazione del modello finale: β_0 intercetta fissa, β_1 relativa alla covariata Età e β_3 della covariata Killip.

B.3 Raftery and Lewis Diagnostic

Supponiamo di essere interessati ad un qualsiasi quantile q di una catena di Markov $\{X_t\}_t$. Definiamo ϵ il parametro di tolleranza e s la probabilità di essere nell'intervallo di tolleranza $(q_\alpha - \epsilon; q_\alpha + \epsilon)$. Il test di *Raftery and Lewis* calcola la lunghezza N della catena di Markov il *burn-in* M necessari alla soddisfazione delle condizioni: ε e s. Tipicamente si prendono in considerazione i quantili di ordine 0.025 e 0.975 con $s = 0.95$.

Il test genera, per ogni iterazione t, una sequenza Z_t binaria di $(1, 0)$:

$$Z_t = \begin{cases} 1, & se\ X_t < q_\alpha \\ 0, & altrimenti \end{cases} \tag{B.3}$$

tale sequenza è derivata da una catena di Markov, ma lei stessa non lo è. Come detto in Raftery and Lewis (1992) è ragionevole supporre che la dipendenza di questa sequenza binaria decresca velocemente e che il processo sia ergodico, definiamo quindi un nuovo processo $\left\{Z_t^{(k)}\right\}$ come $Z_t^{(k)} = Z_{1+(t-1)k}$.In prima approssimazione con un k sufficientemente grande è possibile assumere questa nuova sequenza come una catena di Markov a due stati a tempo discreto e avente matrice di transizione del tipo:

$$P = \begin{pmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{pmatrix}, \tag{B.4}$$

e $\pi = (\pi_0, \pi_1) = (\alpha + \beta)^{-1}(\alpha, \beta)$ distribuzione di equilibrio (stazionaria). Quello a cui siamo interessati è determinare il numero di iterazioni necessarie al *burn-in*:M. La matrice di transizione dell'l-esimo passo è:

$$P^l = \begin{pmatrix} \pi_0 & \pi_1 \\ \pi_0 & \pi_1 \end{pmatrix} + \frac{(\lambda)^l}{\alpha + \beta} \begin{pmatrix} \alpha & -\alpha \\ -\beta & \beta \end{pmatrix},$$

dove $\lambda = 1 - \alpha - \beta$. Supponiamo di richiedere che $\mathbb{P}(Z_m^{(k)} = i | Z_0^{(k)} = j)$ sia all'interno dell'intervallo di tolleranza di π_i. Dovremo avere che:

$$\lambda^l \leq \frac{\epsilon(\alpha + \beta)}{\max(\alpha, \beta)} = h$$

che è vera se $m = m^* = \frac{\log(h)}{\log \lambda}$. Quindi il *burn-in* sarà $M = m^* k$. Sia $\overline{Z}_n^{(k)} = \frac{1}{n} \sum_{t=1}^{n} Z_t^{(k)}$ che per n grande tende ad una distribuzione gaussiana di media q e varianza $\frac{1}{n} \frac{\alpha\beta(2-\alpha beta)}{(\alpha+\beta)^3}$:

$$\mathbb{P}\left(q - r \leq \overline{Z}_n^{(k)} \leq q + r\right) = s \Longleftrightarrow n = n^* = \frac{\frac{\alpha\beta(2-\alpha-\beta)}{(\alpha+\beta)^3}}{\left[\frac{r}{\Phi((1+s)/2)}\right]^2} \tag{B.5}$$

dove con Φ si intende la distribuzione cumulativa gaussiana. Abbiamo quindi $N = n^* k$. Bisogna ora determinare k dalla serie $\left\{Z_t^{(k)}\right\}$ con $k = 1, 2, ...$ Come detto nell'articolo, vi sono diverse strade per calcolarsi il k, come ad esempio usando il criterio *BIC*. Per determinare il numero minimo di iterazioni necessarie si ponga $M = 0$ e $k = 1$

$$N_{min} = \left[\Phi^{-1}\left(\frac{s+1}{2}\right) \frac{\sqrt{q_\alpha(1 - q_\alpha)}}{r}\right]^2 \tag{B.6}$$

è il numero di iterazioni necessarie quando si suppone che l'autocorrelazione sia nulla.

Infine il test viene accompagnato da un *dependence factor* $I = \frac{M+N}{N_{min}}$

Il test di diagnostica *R-L* avrà differenti risultati in base alla scelta del quantile preso in considerazione ed è un test applicabile singolarmente ad ogni variabile. Si nota anche che il test è piuttosto conservativo, cioè tende a suggerire un numero di iterazioni più elevato del necessario.

Di seguito portiamo alcuni risultati ottenuti durante lo studio delle catene generate durante la preparazione di questa tesi con i dati provenienti dal dataset completo sui quantili di ordine 0.025 e 0.975.

```
RAFTERY AND LEWIS CONVERGENCE DIAGNOSTIC
========================================

Iterations used = 1:5000
Thinning interval = 1
Sample size per chain = 5000

$chain1

Quantile (q) = 0.025
Accuracy (r) = +/- 0.005
Probability (s) = 0.95
```

	Burn-in (M)	Total (N)	Lower bound (Nmin)	Dependence factor (I)
mu	4	5146	3746	1.370
eta	10	11468	3746	3.060
logOB	4	4896	3746	1.310
killip	4	4919	3746	1.310
sex	3	4363	3746	1.160
ECGtime	6	775	3746	0.207
festivo	2	4020	3746	1.070
numricPrec	2	4020	3746	1.070
modo.MSAteleECG	3	4154	3746	1.110
modo.MSB	5	5503	3746	1.470
modo.SPONTANEO	4	5232	3746	1.400
modo.TRASFERITO	10	9566	3746	2.550
sintomo.ADDOMINALGIA	3	4154	3746	1.110
sintomo.ALTRO	2	3706	3746	0.989
sintomo.DISPNEA	2	3955	3746	1.060
sintomo.DOLORETOR	2	3955	3746	1.060
sintomo.SINCOPE	2	3891	3746	1.040
sede.BBS	2	3955	3746	1.060
sede.INFPOST	20	240504	3746	64.200
fast.track.PS	16	16234	3746	4.330
fast.track.UTIC	7	7680	3746	2.050
PTCA.NO	4	4737	3746	1.260
PTCA.PRIMARIA	3	4292	3746	1.150
PTCA.RESCUE	3	4508	3746	1.200
trombilisi.PREH	3	4087	3746	1.090
trombolisi.SI	4	5410	3746	1.440

```
 theta                20       27644  3746          7.380

RAFTERY AND LEWIS CONVERGENCE DIAGNOSTIC
========================================

Iterations used = 1:5000
Thinning interval = 1
Sample size per chain = 5000

$chain1

Quantile (q) = 0.975
Accuracy (r) = +/- 0.005
Probability (s) = 0.95

                     Burn-in  Total Lower bound  Dependence
                     (M)      (N)   (Nmin)       factor (I)
 mu                  10       11006 3746         2.940
 eta                 4        4791  3746         1.280
 logOB               5        5576  3746         1.490
 killip              2        2092  3746         0.558
 sex                 5        6077  3746         1.620
 ECGtime             6        775   3746         0.207
 festivo             10       12264 3746         3.270
 numricPrec          14       15628 3746         4.170
 modo.MSAteleECG     12       13538 3746         3.610
 modo.MSB            2        399   3746         0.107
 modo.SPONTANEO      10       13606 3746         3.630
 modo.TRASFERITO     3        4409  3746         1.180
 sintomo.ADDOMINALGIA 4       5123  3746         1.370
 sintomo.ALTRO       1        3739  3746         0.998
 sintomo.DISPNEA     4        4712  3746         1.260
 sintomo.DOLORETOR   8        9534  3746         2.550
 sintomo.SINCOPE     3        4409  3746         1.180
 sede.BBS            2        3679  3746         0.982
 sede.INFPOST        8        10464 3746         2.790
 fast.track.PS       5        5870  3746         1.570
 fast.track.UTIC     7        7533  3746         2.010
 PTCA.NO             6        6293  3746         1.680
 PTCA.PRIMARIA       8        10200 3746         2.720
 PTCA.RESCUE         2        3865  3746         1.030
```

`trombilisi.PREH`	5	5482	3746	1.460
`trombolisi.SI`	6	6755	3746	1.800
`theta`	20	19712	3746	5.260

Come si può notare, per entrambi i quantili, il numero minimo di iterazioni necessarie alla convergenza è 3746, se l'ipotesi di autocorrelazione nulla decade questo numero tende a crescere tranne che in rari casi. Come si può notare, eccetto che per i β_j relativi all'intercetta fissa al *MODO MSA+teleECG* e d θ varianza dell'intercetta aleatoria, le catene generate dalle variabili prese in considerazione nel modello finale hanno tutte un numero minore di 5000, iterazioni prese da noi in considerazione durante lo studio.

Appendice C

Il Gibbs Sampler o Gibbs Sampling

Il *Gibbs Sampler* è una metodo per generare una catena di Markov irriducibile e aperiodica tale che abbia una distribuzione stazionaria. Di solito nelle applicazione bayesiane la distribuzione stazionaria è la posterior, che è generalmente una legge su uno spazio di grandi dimensioni. Come si vedrà successivamente, la grande utilità di questa tecnica consiste nel fatto che è sufficiente, ad ogni passo, campionare da distribuzioni univariate.

Per la sua descrizione ci poniamo nel caso bivariato: (X, Y) è un vettore aleatorio con distribuzione congiunta $\pi(x, y)$ (distribuzione). Per ogni x sia $Y|X = x \sim F(\cdot, x)$ la distribuzione condizionata della componente Y , dato $X = x$. In modo analogo sia $X|Y = y \sim G(\cdot, y)$ la distribuzione condizionata di X, dato $Y = y$. Si noti che F e G sono funzioni di ripartizione univariate. Il Gibbs Sampler genera una catena di Markov $\{Z_k = (X_k, Y_k),\ k = 0, 1, ...\}$ a partire da un $Z_0 = (x_0, y_0)$ iniziale. Al primo passo verrà generato

$$Y_0|X_0 = x_0 \sim F_{Y_0}(y, X_0 = x_0)$$

e successivamente, noto il valore di $Y_0 = y_0$ è possibile campionare un nuovo valore dalla prima componente, cioè:

$$X_1|Y_0 = y_0 \sim G_{X_1}(x, Y_0 = y_0).$$

L'algoritmo prosegue in questo modo e al passo k-esimo si avrà:

$$\begin{aligned} &Z_{k-1} = (X_{k-1}, Y_{k-1}) \\ &\quad Y_k|X_{k-1} = x_{k-1} \sim F_{Y_k}(y, X_{k-1} = x_{k-1}) \\ &\quad X_k|Y_k = y_k \sim G_{X_k}(x, Y_k = y_k) \\ &Z_k = (X_k, Y_k) \end{aligned} \tag{C.1}$$

Questo algoritmo è facilmente generalizzabile al caso multivariato: sia dato un vettore aleatorio $\mathbf{X} = (X_1, ..., X_p)$ di dimensione p; sia $\mathbf{x} = (x_1, ..., x_p)$ un qualsiasi vettore di dimensione p, definisco $x_{-i} = (x_1, ..., x_{i-1}, x_{i+1}, ..., x_p)$ con $i = 1, ..., p$ il vettore di dimensione $p-1$ pari al vettore $\mathbf{x}$ decurtato del suo i-esimo elemento. Definiamo la disribuzione condizionata monovariata della componente i-esima del vettore aleatorio $\mathbf{X}$ dato $\mathbf{X}_{-i} = \mathbf{x}_{-i}$:

$$X_i|\mathbf{X}_{-i} = \mathbf{x}_{-i} \sim \pi_i(x|\mathbf{x}_{-i}).$$

Inizializziamo la catena con un valore per ogni componente del vettore aleatorio $\mathbf{X}$ con:

$$\mathbf{X}^0 = (x_1^0, ..., x_p^0)$$

e seuqenzialmente generiamo tutte le componenti del vettore, alla k-esima iterazione si avrà:

$$\begin{aligned}
&\mathbf{X}^{k-1} = (X_1^{k-1}, ..., X_p^{k-1}) \\
&\quad X_1^k|\mathbf{X}_{-1}^{k-1} = \mathbf{x}_{-1}^{k-1} \sim \pi_1(x_1|x_2^{k-1}, x_3^{k-1}, ..., x_p^{k-1}) \\
&\quad X_2^k|X_1^k = x_1^k, \mathbf{X}_{-1,-2}^{k-1} \sim \pi_2(x_2|x_1^k, x_3^{k-1}, ..., x_p^{k-1}) \\
&\quad \vdots \\
&\quad X_j^k|\mathbf{X}_{1,...,j-1}^k = \mathbf{x}_{1,...,j-1}^k, \mathbf{X}_{j+1,...,p}^{k-1} = \mathbf{x}_{j+1,...,p}^{k-1} \\
&\qquad\qquad \sim \pi_j(x_j|x_1^k, ..., x_{j-1}^k, x_{j+1}^{k-1}, ..., x_p^{k-1}) \\
&\quad \vdots \\
&\quad X_p^k|\mathbf{X}_{-p}^k = \mathbf{x}_{-p}^k \sim \pi_p(x_p|\mathbf{x}_{-p}^k) \\
&\mathbf{X}^k = (X_1^k, ..., X_p^k)
\end{aligned} \tag{C.2}$$

Le distribuzioni π_i prendono il nome di *full-conditionals*. È possibile dimostrare che quest'algortimo genera una catena markoviana ergodica. Per maggiori dettagli si veda Ghosh et al. (2006) o Robert and Casella (2005).

L'algoritmo *Gibbs Sampler* può essere implementato anche in problemi di *Variable selection*. Un utile riferimento è George and McCulloch (1993) nel quale viene presentata una procedura detta *Stochastic Search Variable Selection* (SSVS) che seleziona in modo efficiente il sottoinsieme di Variabili necessarie alla descrizione ottimale di un modello lineare.

Bibliografia

Albert, J. H. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88(422):669–679.

Casella, G. and George, E. I. (1992). Explaining the gibbs sampler. *The American Statistician*, 46(3):pp. 167–174.

Dey, D., Ghosh, S. K., and Mallick, B. K. (2000). *Generalized linear models: a Bayesian perspective*. Marcel Dekker, New York.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics*, 32(2):407–451.

Fahrmeir, L., Kneib, T., and Konrath, S. (2010). Bayesian regularisation in structured additive regression: a unifying perspective on shrinkage, smoothing and predictor selection. *Statistics and Computing*, 20(2):203–219.

Frühwirth-Schnatter, S. and Frühwirth, R. (2010). Data augmentation and mcmc for binary and multinomial logit models. *Statistical Modelling and Regression Structures*, pages 111–132.

Früwirth-Shnatter, S. and Wagner, H. (2010). Bayesian variable selection for random intercept modeling of gaussian and non-gaussian data. In *Bayesian Statistic 9*.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2003). *Bayesian Data Analysis, Second Edition (Texts in Statistical Science)*. Chapman and Hall/CRC, 2 edition.

George, E. and McCulloch, R. (1997). Approaches for bayesian variable selection. *Statistica Sinica*, 7:339–374.

George, E. I. and McCulloch, R. E. (1993). Variable selection via gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889.

Ghosh, J., Delampady, M., and Samanta, T. (2006). *An introduction to Bayesian analysis: theory and methods*. Springer Verlag.

Guglielmi, A., Ieva, F., Paganoni, A. M., and Ruggeri, F. (2010). A bayesian random-effect model for survival probabilities after acute myocardial infarction. Technical report, Politecnico di Milano and IMATI-CNR.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer Verlag.

Heidelberger, P. and Welch, P. (1981). A spectral method for confidence interval generation and run length control in simulations. *Communications of the ACM*, 24(4):233–245.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.

Ishwaran, H. and Rao, J. S. (2005). Spike and slab variable selection: Frequentist and bayesian strategies. *The Annals of Statistics*, 33(2):730–773.

Laird, N. M. and Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, 38(4):pp. 963–974.

Lindley, D. V. (1957). A statistical paradox. *Biometrika*, 44(1/2):pp. 187–192.

Malsiner-Walli, G. and Wagner, H. (2011). Comparing spike and slab priors for bayesian variable selection. Technical report, Johannes Kepler Universität Linz, Department of Applied Statistics.

Mitchell, T. J. and Beauchamp, J. J. (1988). Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, 83(404):pp. 1023–1032.

Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)*, 135(3):pp. 370–384.

Park, T. and Casella, G. (2008). The Bayesian Lasso. *Journal of the American Statistical Association*, 103(482):681–686.

Raftery, A. and Lewis, S. (1992). How many iterations in the gibbs sampler. *Bayesian statistics*, 4(2):763–773.

Robert, C. P. and Casella, G. (2005). *Monte Carlo Statistical Methods (Springer Texts in Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288.

Tüchler, R. (2008). Bayesian variable selection for logistic models using auxiliary mixture sampling. *Journal of Computational and Graphical Statistics*, 17(1):76–94.

Wagner, H. and Duller, C. (2010). Bayesian model selection for logit random intercept models. Technical report.

Yuan, M. and Lin, Y. (2005). Efficient empirical bayes variable selection and estimation in linear models. *Journal of the American Statistical Association*, 100(472):pp. 1215–1225.

Zeger, S. L. and Karim, M. R. (1991). Generalized linear models with random effects; a gibbs sampling approach. *Journal of the American Statistical Association*, 86(413):pp. 79–86.