UNIVERSITÀ DEGLI STUDI DI MILANO-BICOCCA

Scuola di Economia e Statistica

Corso di laurea Magistrale in Scienze Statistiche ed Economiche

PREVISIONE DELLA SOPRAVVIVENZA DI PAZIENTI AFFETTI DA SARCOMI TRAMITE UN MODELLO BAYESIANO PER DATI AD ELEVATA DIMENSIONALITÀ

Relatore: Prof. Tommaso Rigon

Correlatrice: Prof.ssa Rosalba Miceli

Tesi di Laurea di:

Gabriele Tinè

Matr. N. 826904

Anno Accademico 2022/2023

Indice

Ci sono più combinazioni

su una scacchiera

che atomi nell ' universo.

(Claude Shannon, 1950)

1 Introduzione

La statistica è una disciplina caratterizzata da una forte impronta quantitativa-metodologica attraverso cui sintetizzare la realtà. Una definizione sufficientemente generale ma che coglie in modo accurato la natura della disciplina statistica è fornita da Wallis and Roberts (2014): *"Statistics is a body of methods for making wise decisions in the face of uncertainty"*. La statistica è un insieme di metodi che consentono di quantificare l'incertezza e prendere decisioni in condizioni di incertezza. Si osservi che l'incertezza è proprio ciò che la statistica tenta di minimizzare, al fine di ottimizzare la decisione. Si osservi inoltre che l'incertezza caratterizza numerosi fenomeni.

Nel presente lavoro di tesi verrà sviluppata una metodologia statistica volta a modellare l'incertezza associata a dati di sopravvivenza su un campione di pazienti affetti da sarcoma agli arti.

Il sarcoma è un particolare tipo di tumore. Sebbene i processi alla base dell'oncogenesi non siano stati ancora del tutto compresi ed enucleati, la ricerca ha individuato un fattore biologico che influisce su tali processi. In particolare, parte

del processo dell'oncogenesi sembra potersi ascrivere al processo di duplicazione del DNA. Al fine della generazione di nuove cellule il DNA è sottoposto a un processo di replicazione che coinvolge numerosi complessi proteici. Non è questa la sede in cui si desidera enucleare ogni singolo passaggio della replicazione ma, dal punto di vista statistico vi è un evento, durante tale processo, di particolare interesse: la riparazione degli errori di accoppiamento tra basi (*mismatch*) tramite la DNA polimerasi (complesso proteico di riparazione), per dettagli si veda Lodish et al. (2008). Alla fine del processo di replicazione la polimerasi di riparazione esegue una scansione dell'intero filamento di DNA al fine di riparare eventuali errori di accoppiamento tra basi commessi in fase di duplicazione. Tale processo evidenzia come il meccanismo di replicazione cellulare sia un fenomeno soggetto ad errori. Questo fenomeno biologico è pertanto caratterizzato da grande aleatorietà, che può essere gestita e compresa tramite opportune analisi statistiche.

Inoltre, per quanto la statistica cerchi di controllare l'errore, rimarrà sempre un errore residuo irriducibile, caratterizzante l'aleatorietà del fenomeno. Infatti durante il processo di riparazione degli errori di accoppiamento non tutti gli errori vengono individuati e corretti: si consideri che durante il processo di replicazione il tasso di errore è di circa uno ogni 100 mila coppie di basi, il che potrebbe apparire come un errore abbastanza contenuto; tuttavia, se si considera il numero complessivo di coppie di basi che costituiscono il DNA, circa 6 miliardi di coppie, e il numero medio di errori commessi durante la replicazione di una singola cellula, questo sarebbe nell'ordine di 120 mila per cellula. Tramite correzioni dirette, durante il processo di replicazione, tali errori vengono ridotti del 99% circa, il che riduce il numero medio di errori a circa 1200 per ciascuna replicazione; di questi 1200 errori residui deve occuparsi la polimerasi di riparazione la quale,

mediamente, presenta un tasso di errore compreso tra un errore ogni 100 e uno ogni 1000, si veda Johnson et al. (2000). In realtà il tasso è influenzato anche da altri fattori quali, a titolo di esempio, l'età; in particolare all'aumentare dell'età i medesimi autori evidenziano un tasso di errore più elevato, in quanto le cellule perdono progressivamente capacità di riparazione.

Si osservi che, nonostante il numero medio di mutazioni sembri elevato, se si considera il numero complessivo di cellule nell'organismo, non tutte le mutazioni proliferano: le cellule sono dotate di altri meccanismi per bloccare la proliferazione delle mutazioni come, a titolo di esempio, l'apoptosi (morte cellulare programmata). Inoltre, la maggior parte delle mutazioni che si originano non sono dannose per l'organismo. Tuttavia alcune mutazioni, in corrispondenza di alcuni specifici geni, possono portare alla formazione di cellule cosiddette cancerogene e dare inizio al processo di oncogenesi.

I fattori ambientali con cui l'organismo interagisce, inoltre, sembra possano innalzare in modo significativo la probabilità di mutazioni a livello del DNA, secondo meccanismi non ancora ben noti e, conseguentemente, possono innalzare la probabilità che tali mutazioni si traducano in un processo di cancerogenesi. Ne consegue che l'ambiente interagisce con i meccanismi biologici alla base dell'oncogenesi, aggiungendo all'errore irriducibile una fonte di errore esogena aggiuntiva. Questo spiega in parte perché, allo stato attuale, si registrino circa 19 milioni di nuovi casi di tumore ogni anno. Negli Stati Uniti la probabilità di sviluppare un qualsiasi tipo di tumore maligno nel corso della propria vita risulta pari a circa il 40%; nel Regno Unito a circa il 50% per la popolazione maschile e al 45% per quella femminile. In Italia circa il 50% della popolazione maschile e il 35% di quella femminile, svilupperà un tumore maligno nel corso della propria vita,

stando alle ultime proiezioni di Sung et al. (2021). In generale, dalla medesima fonte è possibile constatare come, non sorprendentemente, il tumore che mostra una maggior incidenza a livello globale (14.3%) nella popolazione maschile sia il tumore al polmone, mentre per la popolazione femminile quello alla mammella (11.7%); seguiti dal tumore alla prostata per la popolazione maschile (14.1%) e da quello al polmone per la popolazione femminile (11.4%).

Tra i tumori maligni più rari e al contempo presentanti un tasso di letalità estremamente elevato, i cui dati epidemiologici disaggregati, infatti, non vengono riportati nella fonte citata, vi è il sarcoma. Il sarcoma presenta caratteristiche peculiari che lo rendono, a seconda dell'istologia e della sede in cui si sviluppa, estremamente aggressivo, propenso a dar luogo a metastasi, spesso farmaco-resistente e avente, con riferimento all'età e tenuto conto dell'istologia, un'incidenza atipica, rispetto alla maggior parte dei tumori. In particolare risulta più frequente nelle popolazioni più giovani. Per ulteriori approfondimenti circa le caratteristiche dei sarcomi si veda Sinha and Peach (2010).
Proprio per la bassa incidenza risulta estremamente difficile trovare studi validi ed approfonditi riguardanti tali tumori, inoltre tale lavoro viene svolto prevalentemente da strutture specializzate e dedicate quasi esclusivamente alla ricerca e alla cura di patologie rare. In Italia ne sono un esempio l'Istituto Europeo di Oncologia e l'Istituto Nazionale dei Tumori di Milano.

In particolare quest'ultimo ha sviluppato un indice prognostico, sottoposto a numerose validazioni esterne, in grado di predire con una buona accuratezza la probabilità di sopravvivenza a 5 e 10 anni per pazienti affetti da diversi tipi di sarcomi, tenendo conto dell'istologia. Tali informazioni sono estremamente utili all'oncologo per pianificare il trattamento, per la scelta stessa del trattamento,

per valutare l'evolversi della prognosi e, dunque, della malattia stessa nel tempo e, infine, per coadiuvarlo a decidere quando sia il momento più appropriato per attivare l'assistenza del reparto di cure palliative.

L'Istituto Nazionale dei Tumori di Milano (INT) è inoltre l'IRCCS che ha messo a disposizione i dati per il presente lavoro di ricerca, che ha l'obiettivo di verificare se l'indice prognostico denominato *Sarculator*, costruito dall'INT, possa essere reso più accurato grazie all'aggiunta di variabili di natura radiomica.

Pertanto, dopo aver fornito un breve inquadramento epidemiologico sul sarcoma, si procederà illustrando quali sono le variabili che caratterizzano il *Sarculator*, si condurrà una breve analisi descrittiva dei dati a disposizione, e si cercherà di capire come modellare la moltitudine di variabili radiomiche fornite dai radiologi di concerto all'indice prognostico già sviluppato, per verificare se vi sia la possibilità di migliorare l'accuratezza prognostica del *Sarculator*. In altri termini, si costruirà un metodo ad hoc per tenere conto dell'informazione aggiuntiva derivante dalle variabili radiomiche cercando, soprattutto, di quantificare adeguatamente l'incertezza modellistica in un contesto in cui il numero di variabili risulta di molto superiore al numero delle osservazioni.

1.1 Sarcomi agli arti: epidemiologia

I sarcomi sono tumori, come già evidenziato, estremamente rari, rappresentanti meno dell'1% di tutti i tumori maligni negli adulti ma circa il 20% di tutti i tumori pediatrici ($0 - 14$ anni), Stiller et al. (2013). Sono di origine mesenchimale (il mesenchima è il tessuto embrionale da cui derivano i tessuti connettivi, muscolari e alcune cellule endocrine, ovvero cellule secernenti ormoni) e possono originare

dall'osso o da tessuti molli quali muscoli, strutture tendinee, vascolari e nervose. Proprio per il loro particolare meccanismo di origine risultano, solitamente, abbastanza aggressivi, in quanto tendono a diffondersi in parti anatomiche in cui la chirurgia può risultare spesso non praticabile o praticabile solo parzialmente, con conseguente non eradicazione completa della malattia. Si pensi, a titolo di esempio, ai sarcomi che intaccano le strutture vascolari o nervose, essi possono risultare chirurgicamente non approcciabili a causa del rischio della compromissione delle funzionalità vitali, cognitive, motorie e sensoriali.

In generale, per il sistema di stadiazione dei sarcomi si veda Amin et al. (2017), se il sarcoma risulta di *grading* 1 non presenta metastasi ed è trattabile chirurgicamente senza elevata probabilità di incorrere in complicanze, si propende per la chirurgia radicale. Diversamente, l'oncologo valuta di caso in caso quale sia l'iter terapeutico più appropriato. Sovente viene effettuata la chemioterapia neo-adiuvante e/o post-adiuvante la chirurgia. Con chemioterapia neo-adiuvante si intende una terapia effettuata prima della chirurgia per ridurre le dimensioni della lesione tumorale; similmente la chemioterapia post-adiuvante è una terapia effettuata dopo l'intervento chirurgico al fine di eliminare eventuali cellule tumorali residue. Il trattamento radioterapico è un ulteriore opzione terapeutica sebbene venga adottato con maggior cautela nel caso di sarcomi che originano dall'osso e caratterizzati da certi profili istologici, dal momento che in tali casi la radioterapia incrementa notevolmente la probabilità di sviluppare un sarcoma secondario da radiazioni. L'immunoterapia rappresenta un'ulteriore opzione terapeutica anche se, solitamente, non è un trattamento che viene considerato di prima istanza per i sarcomi. L'immunoterapia sfrutta il sistema immunitario del paziente stesso inducendolo a produrre autoanticorpi, ovvero anticorpi diretti contro cellule prodotte

dall'organismo. Solitamente il sistema immunitario non sviluppa autoanticorpi, ma anticorpi volti a contrastare agenti patogeni esogeni come batteri e virus. Infatti la produzione spontanea di autoanticorpi si verifica in caso di patologie chiamate appunto autoimmunitarie. Sono condizioni anomale in quanto il corpo produce anticorpi diretti contro cellule sane dell'organismo. La sclerosi multipla è un classico esempio di malattia autoimmune, per dettagli si veda Dobson and Giovannoni (2019). La produzione di specifici autoanticorpi può essere tuttavia indotta anche farmacologicamente, proprio tramite l'immunoterapia. In questo caso i trattamenti immunoterapici cercano di stimolare autoanticorpi PD1 e/o PDL-1. Diversi studi clinici hanno infatti mostrato come tali autoanticorpi riescano a inibire i *checkpoints* immunitari (regolatori dei principali processi di risposta del sistema immunitario ad agenti esogeni). Si osservi che l'inibizione indotta dagli autoanticorpi compromette la normale risposta immunitaria ad agenti patogeni tuttavia può inibire, al contempo, l'accrescimento tumorale. Infatti il tumore sembra sfruttare proprio i *checkpoints* immunitari per poter proliferare. Inibendo tali *checkpoints* si può dunque inibire la proliferazione del tumore stesso sino ad ottenere una stabilizzazione della malattia o, in alcuni casi, anche la remissione. Per approfondimenti circa il meccanismo di azione e l'efficacia dell'immunoterapia si veda Gandini et al. (2016).

Sebbene l'immunoterapia rappresenti una delle principali aree di ricerca in oncologia e i risultati ad oggi disponibili sembrino promettenti, per i sarcomi agli arti le ricerche condotte sembrano concordi nell'indicare una scarsa efficacia in termini di blocco dell'accrescimento, remissione o stabilizzazione della malattia. Fanno eccezione alcuni particolari tipi di sarcomi per i quali l'immunoterapia in combinazione con la chemioterapia si è dimostrata statisticamente più efficace

della sola chemioterapia, si veda Fazel et al. (2023).

Esistono più di ottanta tipi molecolari caratterizzanti sarcomi differenti (il termine tipo molecolare in oncologia identifica il tipo di alterazione molecolare a livello cellulare), per la classificazione dei tipi molecolari si veda Jain et al. (2010). Da un punto di vista clinico, esiste una prima suddivisione sulla base del distretto anatomico interessato dal sarcoma: si parla infatti di sarcomi retroperitoneali se il sarcoma si sviluppa nello spazio retroperitoneale (regione addominale situata posteriormente al peritoneo, cioè alla membrana che riveste la cavità addominale e in parte quella pelvica) e di sarcoma agli arti se l'origine del sarcoma è diversa rispetto a quella retroperitoneale, in quanto, come detto, i sarcomi sono di origine mesenchimale e dunque, sovente, anche se originano da un muscolo o da una struttura nervosa si innestano molto velocemente sull'arto stesso. Clinicamente si distinguono i sarcomi in queste due macro-categorie in quanto presentano una prognosi e un approccio terapeutico differente. Per ulteriori dettagli epidemiologici e riguardanti la classificazione degli istotipi si veda Burningham et al. (2012).

I sarcomi agli arti sono particolarmente rari e al contempo aggressivi. Si consideri che per quanto riguarda la popolazione italiana i nuovi casi di sarcomi agli arti ogni anno ammontano, in media, a circa 2600 rappresentando il 75% del totale dei sarcomi. La sopravvivenza a 5 anni dalla diagnosi risulta pari al 57%. Quasi un paziente su due affetto da sarcomi agli arti muore entro i 5 anni dalla diagnosi a causa del tumore stesso, come si può evincere dal lavoro di Trama et al. (2019), che combina diversi dati di registro, e fa un'analisi ben dettagliata e distinta per istotipo sulla sopravvivenza di tali pazienti.

Non sorprende che la sopravvivenza sia così bassa: essendo una patologia rara risulta estremamente difficile poter costruire un campione di dimensioni

adeguate per effettuare studi di qualsiasi tipologia, che siano di tipo osservazionale, retrospettivo, caso-controllo o randomizzato.

Inoltre l'iter diagnostico è spesso difficile e si basa anzitutto sull'intuito del medico di medicina generale, il quale, dopo aver escluso le cause più comuni per i sintomi riferiti e aver messo in atto i principi della diagnosi differenziale, invia il paziente all'oncologo, il quale ne valuta l'obiettività e prescrive esami diagnostici quali radiografie, PET-CT (*Positron Emission Tomography - Computerized Tomography*), RM (Risonanza Magnetica), biopsie CT guidate ed esame istologico (esame che permette di identificare il tipo molecolare) prima di giungere a una diagnosi definitiva. Per ulteriori dettagli circa l'iter diagnostico si consiglia l'articolo di Cormier and Pollock (2004).

Infine, i pazienti affetti da patologie rare, tipicamente, giungono in centri di ricerca specializzati ma solo per la diagnosi e l'impostazione della terapia. Dopodiché vengono preferite cliniche locali, ubicate vicino alla propria residenza, per motivi di natura logistica. Le suddette complicazioni generano numerosi persi al *follow-up* negli studi ed il sarcoma, in quanto patologia rara, non fa eccezione.

Una bassa incidenza, un iter diagnostico non sempre semplice e veloce, un'elevata letalità e un alto numero di persi al *follow-up* rende difficile poter effettuare ricerca e miglioramenti sulle possibili cure per i sarcomi.

1.2 Il *Sarculator* come indice prognostico

Visti i problemi per condurre studi con numerosità campionarie adeguate, per i quali si rimanda al paragrafo 1.1, centri con una elevata concentrazione di risorse disponibili per la ricerca hanno avviato uno studio retrospettivo tramite il quale

poter costruire un indice prognostico che predicesse la sopravvivenza per pazienti affetti da sarcomi agli arti e che fungesse da predittore di rischio clinico per orientare e coadiuvare l'oncologo nell'impostazione terapeutica anche in centri non specializzati in tali patologie. In particolare l'INT ha sviluppato un nomogramma su un campione di 1452 pazienti affetti da sarcomi agli arti, raccolto dal 1994 al 2013, tenendo in considerazione come variabili cliniche: l'età, il *grading*, l'istologia tumorale e il diametro della lesione.

L'indice prognostico, chiamato dagli autori Sarculator, è il risultato dell'applicazione di un modello di Cox sul campione di 1452 pazienti tenendo in considerazione variabili di rilevanza clinica quali quelle sopracitate.

Si osservi che il modello di Cox è un modello semiparametrico, la componente non parametrica deriva dalla funzione di rischio su cui non viene specificata alcuna distribuzione. L'assunzione che viene fatta è che l'effetto delle covariate sia proporzionale al rischio nel tempo. Si osservi che la nozione di funzione di rischio e tutte le nozioni riguardanti l'analisi di sopravvivenza verranno enucleate nel Capitolo 2, per una breve disamina sul modello di Cox si rimanda invece al Capitolo 4.

Il modello di Cox stima la probabilità di sopravvivenza nel tempo. Diversamente dagli usuali modelli di regressione che predicono una singola quantità per ciascun individuo, i modelli di sopravvivenza prevedono la probabilità di sopravvivenza individuale nel tempo. Per essere più precisi, a ciascun individuo vengono associate diverse probabilità di sopravvivenza: una in corrispondenza di ciascun istante di tempo considerato. Si osservi che, implicitamente, si è già fornita una informazione rilevante circa la natura della variabile risposta in un modello di sopravvivenza. Infatti, un modello di sopravvivenza ha come variabile

risposta una coppia di variabili per ciascun individuo: ciò che viene definito lo stato dell'individuo, e il tempo trascorso dall'inizio dell'osservazione dell'individuo al tempo di rilevazione dello stato. Si osservi che i modelli di sopravvivenza devono il loro nome all'ampio utilizzo che ne viene fatto in ambito oncologico. In realtà, tali modelli si dovrebbero più correttamente definire modelli per variabili di tipo tempo all'evento. Infatti, lo stato, che compare nella variabile risposta, può essere un qualunque evento di interesse.

Tornando alla costruzione del Sarculator, la variabile dipendente considerata è data dal tempo intercorso tra la data di prima chirurgia e la data di decesso o, per i pazienti vivi, la data disponibile all'ultimo *follow-up*. In un simile contesto l'evento di interesse è il decesso; un individuo in vita rappresenta invece l'evento di censura, si veda il paragrafo 2.1 per dettagli. La variabile risposta risulta pertanto composta, per l'i-simo individuo, dalla coppia (δ_i, t_i), dove δ_i può assumere due valori: 1 se il paziente è deceduto, 0 se il paziente è vivo e dove t_i rappresenta il tempo intercorso dalla data di prima chirurgia all'osservazione dello stato δ_i. Il Sarculator è stato quindi ottenuto regredendo le variabili di interesse sulla risposta formata dalle coppie (δ_i, t_i) ed estraendo il predittore lineare del modello. All'aumentare del predittore lineare le probabilità di sopravvivenza risultano più basse.

Sono stati quindi estratti i coefficienti stimati associati a ciascuna variabile inserita in modo tale da disporre di un insieme di coefficienti da utilizzare su nuovi dati per effettuare previsioni.

Costruito il modello prognostico, lo studio ha beneficiato di dati raccolti nello stesso intervallo temporale da altri centri di ricerca, quali: l'Istituto francese Gustave Roussy (420 pazienti), il Royal Marsden Hospital di Londra (444 pazienti)

e il Mount Sinai Hospital di Toronto (1436 pazienti), su cui validare l'indicatore prognostico costruito. La validazione multicentrica ha mostrato risultati in linea con il Sarculator sotto il profilo dell'accuratezza prognostica e della calibrazione. L'accuratezza è stata valutata in termini di C-*index* mentre la calibrazione è stata valutata in termini di *Brier-Score* a 5 e 10 anni. In particolare, considerando i 3 valori di C-index ottenuti in fase di validazione, la media dei 3 valori è risultata pari a 0.76. Il valore del C-index registrato dal Sarculator sui dati su cui è stato costruito è risultato invece pari a 0.77. La validazione esterna ha quindi permesso di verificare l'assenza di sovrastima e di sottostima. Per ulteriori dettagli circa la costruzione e la validazione dell'indice prognostico si veda Callegaro et al. (2016) e Callegaro et al. (2019).

Data l'accuratezza dell'indicatore prognostico garantita dalla validazione multicentrica, l'Istituto Nazionale dei Tumori di Milano ha proceduto brevettando e sviluppando, grazie al supporto di Digital Forest SRL, un'applicazione denominata **Sarculator** gratuitamente scaricabile Miceli et al. (2022).

In tal modo anche i centri meno specializzati che si imbattono in questo tipo di patologie possono avere un valido strumento prognostico di supporto in modo da assicurare al paziente una gestione più efficace e un approccio terapeutico quanto più possibile adeguato.

Poco prima del rilascio dell'indice prognostico è stata effettuata un'altra validazione esterna su una casistica più ampia composta da 9738 pazienti estratta dal National Cancer Data Base (NCDB) americano Bilimoria et al. (2008). I risultati della validazione si sono rivelati in linea con i precedenti sia in termini di accuratezza sia in termini di calibrazione, per dettagli si veda Voss et al. (2022).

Da quanto appena descritto si può comprendere quanto uno strumento pro-

gnostico debba essere validato prima di poter essere effettivamente utilizzato. Tuttavia il processo di validazione esterna risulta necessario per poter ottenere garanzie circa le reali capacità prognostiche dello strumento. Inoltre, ogni miglioramento apportabile a tale indice in termini di accuratezza, se valido e generalizzabile, avrebbe conseguenze dirette sul processo terapeutico del paziente stesso. Si intuisce quindi come l'obiettivo del seguente lavoro di ricerca trovi un rilevante risvolto pratico da un punto di vista clinico-oncologico. Nel caso oggetto d'analisi si hanno a disposizione tutte le variabili cliniche considerate dal Sarculator, pertanto, al fine di creare l'indice prognostico corrispondente al Sarculator è sufficiente considerare i coefficienti stimati del Sarculator e moltiplicarli per le variabili corrispondenti presenti nell'insieme di dati a disposizione. Si sommano quindi le nuove variabili ottenute per giungere ad un'unica variabile che rappresenta l'indice prognostico costruito con il Sarculator sul campione oggetto d'analisi. In Tabella 1 si riportano i coefficienti stimati del Sarculator con cui si è creato l'indice di prognosi per i dati oggetto d'analisi.

Si osservi che il Sarculator utilizza solo variabili di natura clinica. I macroraggruppamenti di istologia sono stati effettuati in fase di costruzione del Sarculator sulla base di indicazioni cliniche come riportato da Callegaro et al. (2016). Infine, utilizza delle trasformate delle variabili età e diametro, in particolare utilizza delle *Restricted Cubic Splines* per la cui trattazione si rimanda al paragrafo 2.7. Il predittore lineare ottenuto a partire dalle variabili cliniche presenti nel dataset e dai coefficienti forniti dal Sarculator (Tabella 1) sarà un elemento rilevante ai fini del lavoro di tesi. Infatti, calcolato tale predittore, a partire dal campione a disposizione si elaborerà una strategia per verificare se le variabili radiomiche possano essere aggiunte al predittore lineare, e dunque al Sarculator stesso, per

Variabile	$\hat{\beta}_{Sarc}$
Intercetta	−2.492
Età	0.007
*Restricted Cubic Splines** Età	0.014
Diametro (cm)	0.224
*Restricted Cubic Splines** Diametro (cm)	−0.179
Grading: 2	0.987
Grading: 3	1.447
Istologia: Liposarcoma Dedifferenziato/pleomorfico	−0.527
Istologia: Liposarcoma Myxoide	−0.918
Istologia: Tumore Maligno della guaina dei nervi periferici	−0.281
Istologia: Myxofibrosarcoma	−0.426
Istologia: Sarcoma Sinoviale	0.077
Istologia: Sarcoma Pleomorfico non differenziato	−0.680
Istologia: Sarcoma vascolare	0.842
Istologia: Altri	−0.230

I nodi per le RCS (*Restricted Cubic Splines*) vengono posti pari a 30, 54, 75 per l'età e 2, 6, 15 per il diametro.

* Componente non lineare delle RCS

Valori di riferimento variabili categoriche, Grading: 1; Istologia: Leyomiosarcoma

Tabella 1: Coefficienti del predittore lineare del Sarculator

migliorarne l'accuratezza.

1.3 Dati Radiomici

I dati di natura radiomica derivano dal processo di estrazioni di variabili caratterizzanti differenti aspetti delle immagini radiologiche. In linea teorica l'estrazione di variabili radiomiche può essere effettuata sulla base di qualsiasi tipo di immagine ricavata dagli attuali sistemi di *imaging*: dalle radiografie ottenute tramite l'emissione di raggi X, dalle scansioni di una Tomografia Assiale Computerizzata (TAC), dalla PET, in cui la sorgente di emissioni non è esterna al paziente ma è il paziente medesimo a cui viene somministrato un radiofarmaco, alle imma-

gini di Risonanza Magnetica. Le immagini sono composte da pixel che possono assumere differenti valori numerici corrispondenti a diversi valori della scala di grigio. All'aumentare della densità di pixel dell'immagine aumenta la quantità di informazione in essa contenuta. Si osservi, tuttavia, che non sempre un'elevata densità di pixel coincide con maggior informazione utile. Se da un lato, quindi, all'aumentare della densità dei pixel aumenta l'informazione, dall'altro non tutta l'informazione risulta necessaria per caratterizzare le lesioni tumorali. Infatti l'informazione contenuta all'interno di un'immagine con un livello di dettaglio più elevato può costituire una notevole fonte di rumore. Da quanto premesso, essendo la RM la metodica radiologica che garantisce l'acquisizione di immagini con la maggior densità di pixel, essa è anche la metodica che tende produrre numerose variabili non informative per la caratterizzazione della lesione. Infatti è ragionevole attendersi un tasso di variabili di rumore elevato, per cui molte variabili non saranno in grado di cogliere le differenze utili a identificare e profilare correttamente la lesione.

Le variabili radiomiche estratte sul campione oggetto d'analisi derivano dall'estrazione di diverse caratteristiche da immagini di RM funzionale. Per approfondimenti e per una trattazione dettagliata su come vengono acquisite le immagini di RM si rimanda Coriasco et al. (2014).

Per il campione in oggetto sono state considerate sequenze DWI (*Diffusion Weighted Imaging*) in grado di fornire informazioni, di tipo funzionale, riguardo la cellularità e l'integrità delle membrane cellulari dei tessuti analizzati: dal punto di vista delle immagini risultanti, l'intensità di ogni pixel è dipendente dall'entità della riduzione di segnale (le aree dove è ridotto il movimento delle molecole d'acqua sono iperintense). Le variabili radiomiche sono quindi state estratte a

partire dalle mappe ADC (*Apparent Diffusion Coefficient*) dove l'intensità di ogni pixel è proporzionale al valore assoluto del coefficiente di diffusione apparente misurato (le aree a più basso coefficiente di diffusione risultano ipointense). Per l'acquisizione delle DWI sono stati usati gradienti emittenti un campo di 1.5 Tesla. Per dettagli sull'acquisizione di sequenze DWI e sulle mappe ADC si veda Wedeen et al. (2005).

L'estrazione delle variabili radiomiche è stata effettuata sulla base delle *First Order Statistics* (FOS), sulla base della *Gray Level Co-occurrence Matrix* (GLCM) e della *Gray Level Run Length Matrix* (GLRLM) per le caratteristiche di *texture*, e sulla base di funzioni ad hoc costruite per estrarre variabili descriventi maggiormente la morfologia, per ulteriori dettagli circa l'estrazione di variabili si veda Bologna et al. (2023). In totale 2144 variabili radiomiche sono state estratte per ciascun paziente sulla base di sequenze DWI.

1.3.1 Radiomica e prognosi

Per la discriminazione di lesioni benigne da lesioni maligne o per la differenziazione del *grading* tumorale, l'utilizzo delle variabili radiomiche, di concerto con quelle cliniche, è ampiamente diffuso in letteratura, a riguardo si veda Corino et al. (2018), e per una *review* approfondita riguardante in particolare i sarcomi si rimanda a Fanciullo et al. (2022). Al contrario la letteratura in merito alle variabili radiomiche utilizzate per la costruzione di un indice prognostico risulta molto più scarsa. Tra i pochi studi riguardanti sarcomi agli arti, prognosi e radiomica, quello di Spraker et al. (2019) considera sequenze RM T1 pesate e presenta come maggiore limitazione l'esiguo numero di variabili radiomiche

considerate (30). Il limite non risiede nel numero di variabili considerate ma dalle variabili che non sono state considerate. Infatti, non viene enucleato in modo chiaro le motivazioni che hanno indotto gli autori ad estrarre un sottocampione di variabili. Non considerando altre variabili radiomiche caratterizzanti immagini di risonanza potrebbe essere stato selezionato un sottoinsieme di variabili, tra le 30 considerate, non ottimale. Un altro limite sembra risiedere nella procedura utilizzata per la selezione. Infatti le variabili vengono selezionate tramite un modello di Cox con penalizzazione LASSO. La regressione LASSO, tuttavia, non fornisce garanzie inferenziali e la procedura potrebbe soffrire del *bias* di selezione, ovvero dell'incertezza associata al processo di selezione stessa.

Un altro tentativo di costruzione di un indice prognostico che tenesse conto anche di variabili radiomiche è stato effettuato nel lavoro di Peeken et al. (2019b), nel quale vengono considerate congiuntamente sequenze RM T1 e T2 pesate, con un numero totale di variabili radiomiche estratte pari a 1394. In questo caso il numero di variabili sembra adeguato per delle immagini di RM. Il lavoro degli autori tuttavia sembra presentare la medesima limitazione dello studio precedentemente citato. Infatti anche in questo caso la selezione viene effettuata tramite l'applicazione di un Cox-LASSO, il quale tuttavia non fornisce garanzie inferenziali sui predittori selezionati, in quanto non vi è il controllo per il *false discovery rate*. In un precedente studio Peeken et al. (2019a) implementano algoritmi di *Machine Learning* su immagini derivanti, in questo caso, da CT. Anche in questo caso tuttavia, leggendo attentamente il lavoro, sembra non essere presente il controllo per il *false discovery rate*. Pare configurarsi nuovamente il problema della mancanza di garanzie inferenziali circa i predittori selezionati. Inoltre, i risultati riportati non evidenziano una differenza statisticamente significativa tra

l'accuratezza prognostica dell'indice costruito sulla sola base di variabili cliniche e l'accuratezza prognostica dell'indice costruito sulla base di variabili cliniche e radiomiche congiuntamente. Tale risultato porterebbe a concludere che le variabili radiomiche derivate dalla CT non aggiungono potere prognostico a quelle cliniche, tuttavia non essendo presenti garanzie inferenziali circa la selezione la conclusione potrebbe essere fuorviante.

Un ulteriore studio di Zhao et al. (2019) si basa sull'estrazione di 103 variabili radiomiche da sequenze RM DWI, sembra presentare le medesime limitazioni degli altri lavori: viene applicato un Cox-LASSO che non fornisce garanzie inferenziali sui predittori selezionati e il numero di variabili radiomiche risulta esiguo. Anche in questo caso sembra non venir fornita una spiegazione del perché si sia deciso di estrarre solo le variabili in oggetto, come nel lavoro di Spraker et al. (2019).

Tutti gli studi appena menzionati, oltre a presentare una numerosità campionaria simile, sembrano fare uso di un Cox-LASSO in modo statisticamente non ortodosso. Le analisi vengono eseguite sul medesimo campione in due fasi, in altri termini la regressione Cox-LASSO viene utilizzata impropriamente. Nella prima fase la regressione viene impiegata per selezionare i predittori, nella seconda vengono considerati i predittori selezionati per costruire un nuovo modello di Cox non penalizzato al fine di poter estrarre i p-value relativi a ciascuna variabile. Questo introduce due *bias*: un *bias* riguardante i test di ipotesi sui coefficienti e i rispettivi p-value, infatti i p-value non sarebbero ricavabili giacché il LASSO, per costruzione, rende appositamente distorti gli stimatori e quindi i coefficienti. Ne discende che i test effettuati sui coefficienti del nuovo modello soffrono tutti del medesimo *bias* e quindi non sarebbero validi. Un secondo *bias* riguarda il processo stesso di selezione: i dati vengono utilizzati due volte: la prima per selezionare le

variabili, la seconda per stimare il modello. Ciò induce in via naturale un errore dovuto al fatto che il modello viene stimato sugli stessi dati su cui le variabili vengono selezionate. Infine, non sembra essere presente un controllo per il *false discovery rate* e ciò rende la selezione priva di garanzie inferenziali. Sembra cioè che la selezione manchi di informazioni circa l'affidabilità della selezione stessa.

Gli studi attualmente condotti in ambito radiomico al fine di costruire un indicatore prognostico valido e generalizzabile sembra possano soffrire di mancanza di generalizzabilità e di un *bias* dovuto all'utilizzo degli stessi dati per selezionare le variabili e per poi stimare i coefficienti del modello sul sottoinsieme di variabili selezionate. Tutti gli studi mostrano infine una bassa numerosità campionaria, ma questo è comprensibile vista la bassa incidenza della patologia.

Inoltre, nessuno studio sinora condotto ha cercato di integrare le variabili radiomiche ad un indice prognostico costruito su una ben più ampia casistica e validato su quattro distinte coorti di pazienti, di cui due superanti i mille individui, provenienti da quattro diversi centri di ricerca di quattro distinti paesi e da due diversi continenti.

La ricerca e l'obiettivo di questo lavoro è proprio quello di utilizzare tecniche che non soffrano della mancanza di generalizzabilità, di cui sembrano affetti gli studi citati, per integrare le informazioni contenute nelle variabili radiomiche al Sarculator al fine di migliorarne l'accuratezza prognostica. In questo contesto si ha il vantaggio di poter utilizzare l'intero campione concentrandosi sulle variabili radiomiche, senza dover perdere gradi di libertà per l'inclusione di variabili cliniche, dal momento che l'indicatore clinico di prognosi è proprio il Sarculator stesso.

1.3.2 Descrizione del campione

Verranno ora riportati i principali indici di sintesi per descrivere il campione oggetto d'analisi, con particolare riferimento alle variabili di interesse da un punto di vista modellistico. Per quanto riguarda il tempo di osservazione, esso è stato ricavato tenendo conto delle censure ai diversi tempi, considerando come evento proprio la censura, in particolare considerando la censura come evento è possibile ricavare la curva riportata in Figura 1 esprimente, per ciascun istante temporale, la relazione sussistente tra tempo e probabilità di non osservare la censura.

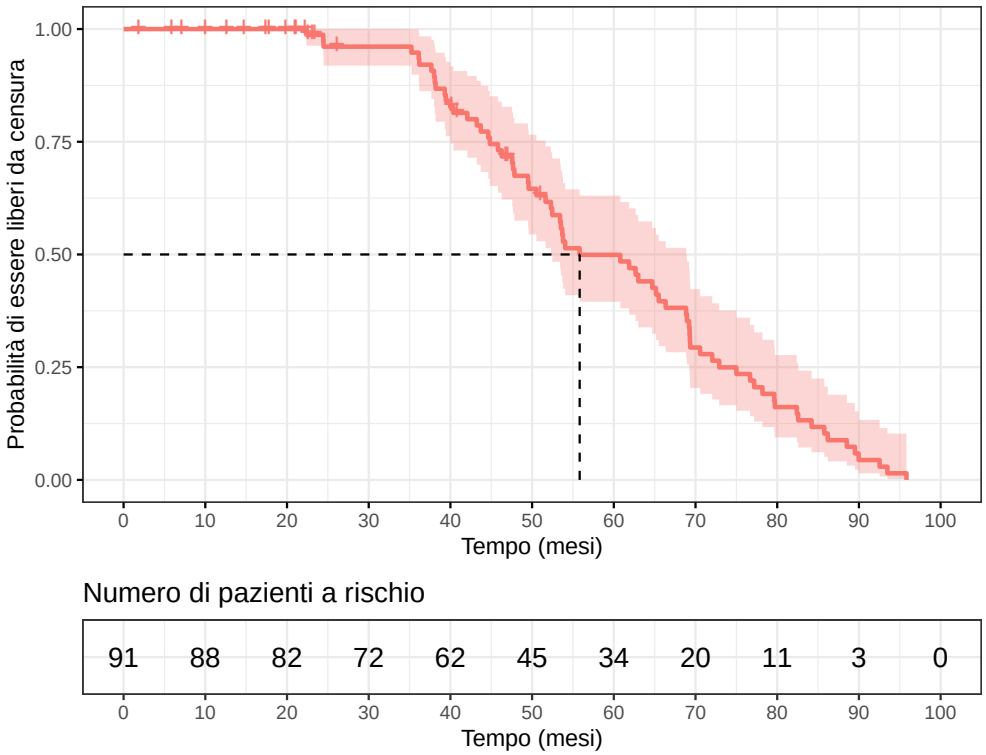


Figura 1: Curva utile per il calcolo del tempo mediano di osservazione, che tiene conto delle censure temporali. Le linee verticali, in questo caso, denotano gli eventi. La linea tratteggiata denota il tempo mediano di osservazione

In Tabella 2 vengono riportati i principali indici di sintesi (mediana, primo e terzo quartile, minimo e massimo) calcolati sul campione oggetto d'analisi per le

variabili quantitative di interesse.

Variabile	Mediana	$(1°, 3°$quartile$)$	Minimo	Massimo
Età (anni)	59	$(46, 71)$	21	85
Diametro (cm)	8.0	$(5.0, 10.0)$	3.0	22.0
Tempo di osservazione (mesi)	55.8	$(44.8, 72.9)$	22.4	95.8

Il minimo tempo di osservazione fa riferimento non già al primo evento ma alla prima censura.

Tabella 2: Statistiche descrittive delle variabili numeriche di interesse

In Tabella 3 vengono riportate le frequenze assolute e relative (in percentuale) calcolate sul campione per le principali variabili categoriche di interesse.

Variabile	n	(%)
Grading: 1	22	(24.2%)
Grading: 2	21	(23.1%)
Grading: 3	48	(52.7%)
Istologia: Leyomiosarcoma	8	(8.8%)
Istologia: Liposarcoma Dedifferenziato/pleomorfico	8	(8.8%)
Istologia: Liposarcoma Myxoide	17	(18.7%)
Istologia: Tumore Maligno della guaina dei nervi periferici	3	(3.3%)
Istologia: Myxofibrosarcoma	20	(22.0%)
Istologia: Sarcoma Sinoviale	4	(4.4%)
Istologia: Sarcoma Pleomorfico non differenziato	24	(26.4%)
Istologia: Sarcoma vascolare	1	(1.1%)
Istologia: Altri	6	(6.5%)

Tabella 3: Statistiche descrittive delle variabili categoriche di interesse

A scopo descrittivo risulta altresì utile fornire le stime della curva di sopravvivenza ottenute tramite lo stimatore non parametrico di Kaplan-Meier senza considerare alcuna covariata, si veda la Figura 2. Per ulteriori dettagli circa lo stimatore di Kaplan-Meier si rimanda a Kaplan and Meier (1958).

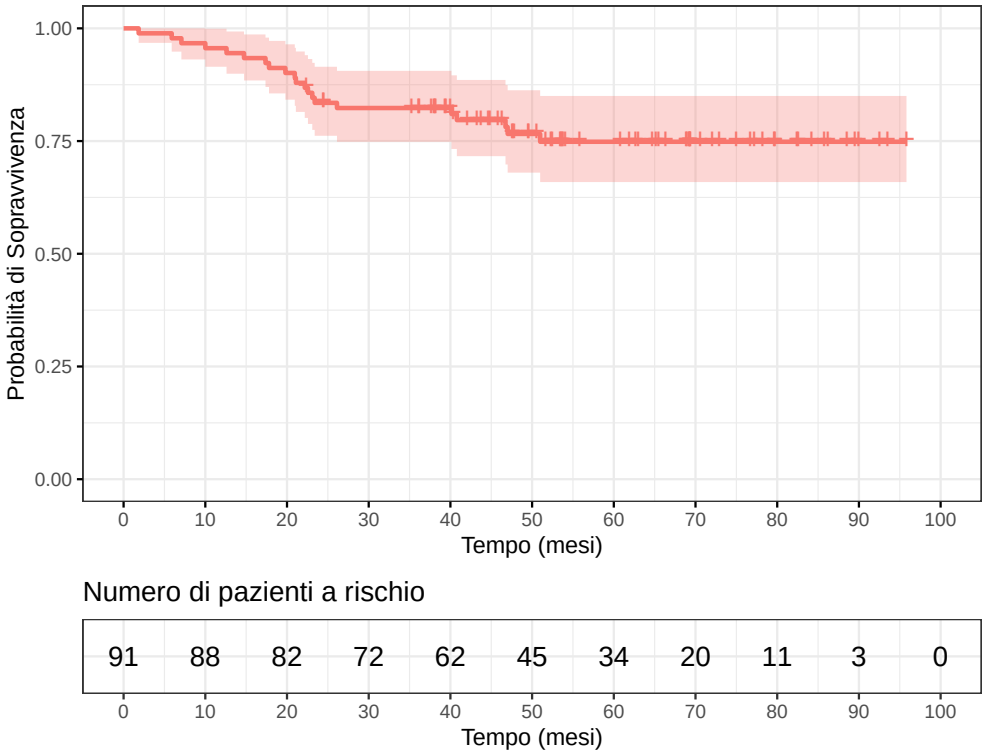


Figura 2: Curva di sopravvivenza stimata con metodo di Kaplan-Meier. Le linee verticali denotano le censure.

Sempre a scopo descrittivo, per valutare se l'andamento marginale della variabile *grading* rispetto alla probabilità di sopravvivenza viene confermato dalla casistica in oggetto si è proceduto con la stima (KM) delle curve stratificate per *grading*, si veda la Figura 3. Si osservi che in generale a un *grading* più elevato corrisponde una prognosi maggiormente sfavorevole e, conseguentemente, una probabilità di sopravvivenza inferiore.

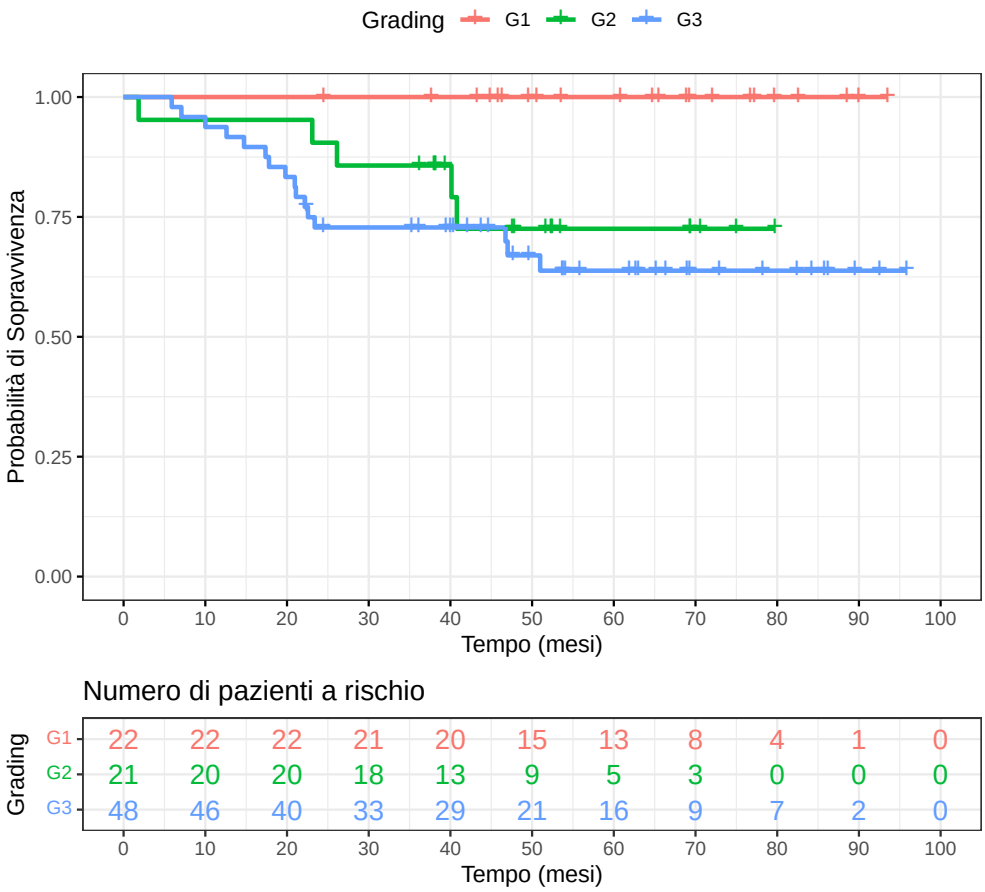


Figura 3: Esempio di curve di sopravvivenza stratificate per Grading tumorale, con probabilità stimata con approccio non parametrico: stimatore di Kaplan-Meier. Le linee verticali denotano le censure.

Facendo riferimento alla Figura 3 è possibile osservare un andamento peggiore, dal punto di vista prognostico, all'aumentare del *grading*. In particolare, si osservi che sulla casistica oggetto d'analisi, i pazienti G1 presentano solo censure, i pazienti G3 sono invece i pazienti presentanti il maggior tasso di eventi. Inoltre, si osservi che tutte le morti avvengono entro i primi 60 mesi (5 anni), ciò trova ampia conferma con quanto riportato in letteratura circa l'andamento della sopravvivenza per i sarcomi agli arti, come largamente discusso nel paragrafo 1.1.

Infine, il campione contiene, per ogni paziente, 2144 variabili di natura radiomica estratte da immagini di Risonanza Magnetica funzionale acquisite

con sequenze DWI, come enucleato nel paragrafo 1.3. Data la natura delle variabili in oggetto, esse assumono range di variazione differente e appaiono a gruppi sufficientemente correlate: la correlazione è oscillante tra lo 0.2 e lo 0.75 per variabili appartenenti alla medesima classe. Con variabili appartenenti alla medesima classe si intende variabili estratte dalla medesima matrice di proiezione (per dettagli si rimanda al paragrafo 1.3, in cui viene spiegato come vengono estratte le variabili). Si procederà pertanto alla standardizzazione delle medesime prima di iniziare con la fase di modellizzazione.

2 Dati di sopravvivenza e loro modellizzazione

Dal momento che il problema oggetto d'analisi impiegherà ampiamente metodi derivanti dall'analisi di sopravvivenza risulta doveroso fornire una breve panoramica sui concetti essenziali, utili alla comprensione di tali metodi e delle principali tecniche che possono essere utilizzate per modellare dati di sopravvivenza. Per una disamina più esaustiva si veda Cox and Oakes (1984).

I dati di sopravvivenza si presentano ogni qualvolta l'obiettivo sia studiare il tempo trascorso da un particolare istante iniziale (*starting point*) al verificarsi di un evento di interesse (*end point*). In particolare il tempo all'evento è una variabile casuale, usualmente chiamata sopravvivenza o *failure time*.

Si osservi che, a differenza degli usuali modelli di regressione, un modello di sopravvivenza non modella una risposta di tipo continuo, discreto o categorico

ma la coppia: (T, δ) con δ indicatore così definito:

$$\delta = \begin{cases} 1 & \text{Se l'evento è stato osservato} \\ 0 & \text{altrimenti} \end{cases}$$

In altri termini quando si modellano dati di sopravvivenza non si è interessati solo a capire se l'evento di interesse si verificherà ma anche quando si verificherà fissato un istante temporale di partenza.

Si osservi altresì che, solitamente, uno studio clinico si estende per un arco temporale definito, pertanto, endemicamente, non sempre è possibile conoscere l'istante esatto in cui si verifica l'evento, si può solo supporre che l'evento possa verificarsi dopo l'arco temporale entro il quale lo studio è stato condotto. In altri termini, se un individuo non esperisce l'evento entro l'arco temporale di osservazione dell'individuo stesso, allora si considera il dato censurato.

Risulta dunque necessario definire cosa si intende esattamente per censura e che tipi di censura esistono.

2.1 Censura temporale

I tempi all'evento sono noti esattamente solo se si verificano all'interno di una certa finestra temporale. I tempi degli individui che non esperiscono l'evento entro tale finestra temporale sono detti censurati, ovvero non influiscono sulla probabilità di sopravvivenza se non tramite l'informazione derivante dal fatto che dall'istante iniziale in cui è partita l'osservazione dell'individuo, esso non ha esperito l'evento e, dunque, è rimasto permanentemente nell'insieme degli individui detti a rischio.

Esistono in realtà diversi tipi di censura negli studi clinici:

1. **Censura a sinistra**: (poco comune negli studi clinici) l'evento di interesse è già avvenuto per l'individuo prima che venga osservato nello studio.
2. **Censura intervallare**: l'evento di interesse si verifica solo entro un intervallo predefinito.
3. **Censura a destra**: (comune negli studi clinici) soggetti senza evento entro il termine dello studio. Formalmente, ogni individuo ha un tempo di sopravvivenza T e un tempo di censura T_C, ma ciò che è osservabile in uno studio è solo il $\min(T, T_C)$.

Nell'analisi di sopravvivenza, fissato un tempo t, si definisce insieme di pazienti a rischio, detto anche insieme a rischio, l'insieme dei pazienti che a tempo t non hanno ancora esperito l'evento e che non sono stati soggetti a censura.

Si osservi inoltre che gli individui possono sperimentare un evento che impedisce di osservare l'evento di interesse, che non si sostanzia in un evento competitivo ma come un vero e proprio evento preclusivo (si pensi, a titolo di esempio, alla perdita di un individuo al *follow-up*). Tale eventualità, di per sè, non rappresenta un problema, semplicemente l'individuo risulterebbe "censurato" ed uscirebbe dall'insieme a rischio, purché non vi sia una relazione tra il tempo di censura e il tempo di sopravvivenza. In tal caso infatti la censura fungerebbe da fattore confondente. Non a caso l'analisi di sopravvivenza pone a monte tale ipotesi, ovverosia per la maggior parte delle procedure statistiche nell'ambito di sopravvivenza si pone l'assunzione di indipendenza tra il tempo di censura e il tempo di sopravvivenza. Se tale assunzione viene violata è evidente che il tempo di censura pone in essere un fenomeno di autoselezione degli individui, il quale deve essere tenuto in debita considerazione in fase di analisi.

2.2 Trattamento della variabile temporale in uno studio

Un altro aspetto di fondamentale importanza nell'analisi di sopravvivenza è come trattare operativamente individui che entrano nello studio in istanti diversi. Infatti se non si considera tale elemento tutti i risultati derivati con qualsiasi metodica sono viziati dal tempo d'ingresso. Risulta infatti evidente che, in uno studio osservazionale, in cui si sia interessati a modellare il tempo all'evento, sia fondamentale tenere conto che il tempo di osservazione degli individui varia al variare dell'istante di ingresso nello studio del periodo stesso.

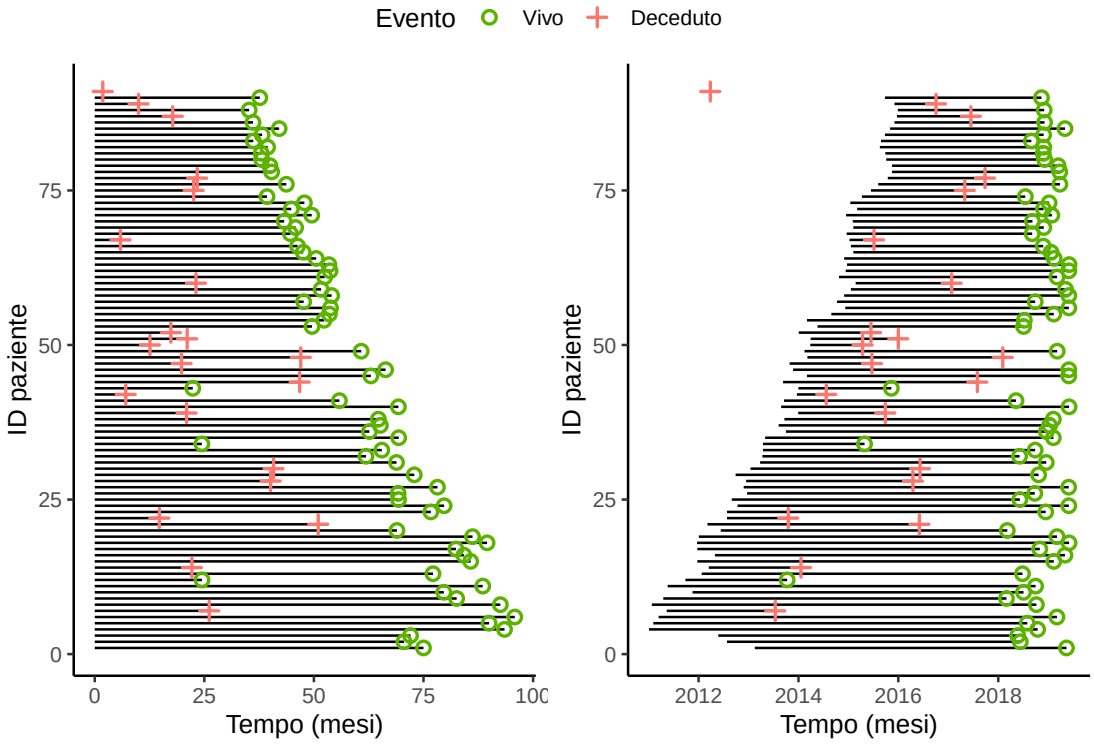


Figura 4: Trattamento della variabile temporale nel contesto dell'analisi di sopravvivenza. A sinistra una rappresentazione della traslazione temporale, a destra la rappresentazione dei tempi non traslata. I tempi e gli eventi fanno riferimento ai dati oggetto d'analisi

Tipicamente, quindi, come variabile temporale viene considerata la differenza tra l'istante di ingresso e l'istante in cui l'individuo esperisce l'evento o l'istante in cui si smette di osservare l'individuo e non già la differenza tra quest'istante e quello di inizio dello studio. In altri termini si trasla il tempo in modo da considerare la durata effettiva di osservazione dell'individuo.

Per comprendere meglio questo aspetto si faccia riferimento alla Figura 4. A sinistra è illustrata la traslazione della variabile temporale, per tenere in considerazione l'effettivo tempo di osservazione, tempo calcolato tramite la differenza tra l'ultima data di *follow-up* e la data di chirurgia, rappresentante il cosiddetto *starting point* individuale. A destra viene riportato l'esempio, invece, di come non tenendo in considerazione l'effettivo tempo di osservazione possa essere fortemente confondente da un punto di vista dell'analisi del dato stesso; infatti le linee rappresentano il periodo di effettiva osservazione, ma in questo caso si considera l'intervallo temporale che va dall'inizio dello studio (che coincide con la data del primo paziente sottoposto a chirurgia) all'ultima data di *follow-up*. Ora, è chiaro che in questo caso non si tiene in considerazione l'effettivo periodo di osservazione dell'individuo, in altri termini non si tiene in considerazione quando l'individuo entra nello studio, cioè dello *starting point* individuale e questo porta inevitabilmente a viziare le analisi. Infatti è diverso sopravvivere sino all'ultima data di *follow-up* essendo entrati quasi all'inizio dello studio piuttosto che sopravvivere sino all'ultima data di *follow-up* essendo entrati quasi alla fine. Sempre con riferimento alla Figura 4 è evidente che, se per un individuo che entra dopo anni dall'inizio dello studio rimane vero che alla fine dello studio risulta sopravvissuto in entrambi i sistemi di riferimento, è altresì vero che per esso è trascorso meno tempo rispetto a un individuo entrato a ridosso della data di inizio

dello studio. Dunque, per l'individuo entrato molto dopo si ha a disposizione meno tempo per osservare l'evento e questo deve essere tenuto in considerazione nelle analisi.

2.3 Quantità di interesse inferenziale

Sia T il tempo all'evento, nel caso in oggetto, il tempo al quale si verifica il decesso. Nell'ambito dell'analisi di sopravvivenza è possibile caratterizzare in modo equivalente la distribuzione di probabilità di una variabile di sopravvivenza aleatoria tramite diverse quantità di interesse inferenziale.

1. Funzione di rischio: descrive l'andamento nel tempo del tasso di eventi che occorrono all'istante t.

$$h(t) = \lim_{\Delta t \to 0} \frac{\mathbb{P}(t < T \leq t + \Delta t \mid T > t)}{\Delta t}. \tag{1}$$

   L'interpretazione della funzione di rischio è immediata: rappresenta la probabilità istantanea di sviluppare l'evento; detto in termini più espliciti: la funzione di rischio valuta la probabilità che un individuo esperisca l'evento l'istante infinitesimo successivo al tempo t, dato che, in t, l'individuo non ha sperimentato l'evento.

2. Funzione di rischio cumulativa:

$$\Lambda(t) = \int_0^t h(z)\mathrm{d}z,$$

   descrive l'accumulazione del rischio nel tempo.

3. Funzione di densità:

$$f(t) = \lim_{\Delta t \to 0} \frac{\mathbb{P}(t < T \leq t + \Delta t)}{\Delta t}$$

4. Funzione di densità cumulativa:

$$F(t) = \mathbb{P}(T \leq t) = \int_0^t f(z)\mathrm{d}z,$$

descrive la probabilità di esperire l'evento entro il tempo t.

5. Funzione di sopravvivenza:

$$S(t) = \mathbb{P}(T > t) = 1 - F(t) = \int_t^{+\infty} f(z)\mathrm{d}z, \tag{2}$$

descrive la probabilità di esperire l'evento dopo il tempo t.

Si tratta di funzioni che caratterizzano in modo equivalente una variabile casuale di sopravvivenza e, infatti, sono strettamente legate.

Se per la funzione di sopravvivenza è facile mostrare come è stato fatto nell'equazione (2), la sua relazione con la funzione di densità cumulativa, e quindi anche con la funzione di densità, meno intuitiva risulta la relazione tra funzione di sopravvivenza e funzione di rischio.

Per mostrare come le due funzioni sono legate si noti che:

$$\mathbb{P}\left[(t < T \leq t + \Delta t) \cap \mathbb{P}(T > t)\right] = \mathbb{P}(t < T \leq t + \Delta t)$$

e quindi, in virtù dell'equazione (1) si ha:

$$
\begin{aligned}
h(t) &= \lim_{\Delta t \to 0} \frac{\mathbb{P}(t < T \leq t + \Delta t \mid T > t)}{\Delta t} \\
&= \lim_{\Delta t \to 0} \frac{\mathbb{P}(t < T \leq t + \Delta t) \cap \mathbb{P}(T > t)}{\Delta t \ \mathbb{P}(T > t)} \\
&= \lim_{\Delta t \to 0} \frac{\mathbb{P}(t < T \leq t + \Delta t)}{\Delta t \ S(t)} \\
&= \lim_{\Delta t \to 0} \frac{\mathbb{P}(T \leq t + \Delta t) - \mathbb{P}(T \leq t)}{\Delta t \ S(t)} \\
&= \lim_{\Delta t \to 0} \frac{F(t + \Delta t) - F(t)}{\Delta t} \frac{1}{S(t)} \\
&= \frac{d}{dt} F(t) \frac{1}{S(t)} = \frac{f(t)}{S(t)}.
\end{aligned}
$$

Di conseguenza:

$$
f(t) = h(t) S(t) \tag{3}
$$

Si osservi che vale analogamente:

$$
h(t) = -\frac{d}{dt} F(t) \frac{1}{S(t)} = \frac{1}{S(t)} - \frac{d}{dt} \left[1 - S(t)\right] = -\frac{1}{S(t)} \frac{d}{dt} S(t),
$$

da cui discende che:

$$
h(t) = -\frac{d}{dt} \log \left[S(t)\right] \Longrightarrow S(t) = \int_t^{+\infty} e^{-h(t)} \mathrm{d}t.
$$

Relazione che risulterà utile nella fase di scelta e specificazione della parametrizzazione del modello.

2.4 Principali distribuzioni per la funzione di rischio

Definite le quantità di interesse inferenziale si riportano, brevemente, i principali modelli parametrici utilizzati nell'analisi di sopravvivenza. Si osservi, a tal proposito, che in letteratura è possibile trovare altre parametrizzazioni del tutto equivalenti dei modelli che verranno presi in considerazione, si veda Tjorve (2009), le cui funzioni di rischio e di sopravvivenza vengono riportate in Tabella 4.

Modello	Rischio	Sopravvivenza
Esponenziale	$h(t) = \gamma$	$S(t) = \exp(-\gamma t)$
Weibull	$h(t) = \gamma\alpha t^{\alpha-1}$	$S(t) = \exp(-\gamma t^{\alpha})$
Log-logistic	$h(t) = \frac{\gamma\alpha(\gamma t)^{\alpha-1}}{1+(\gamma)^{\alpha}}$	$S(t) = \frac{1}{1+(\gamma)^{\alpha}}$

Tabella 4: Principali modelli di sopravvivenza e relative funzioni di rischio e sopravvivenza

Si osservi che, il modello esponenziale è un caso particolare del modello di Weibull quando viene posto il parametro di forma $\alpha = 1$. In Figura 5 vengono rappresentate le funzioni di sopravvivenza e di rischio di ciascun modello presentato. Si noti che tali funzioni sono state stimate a partire dai dati oggetto d'analisi.

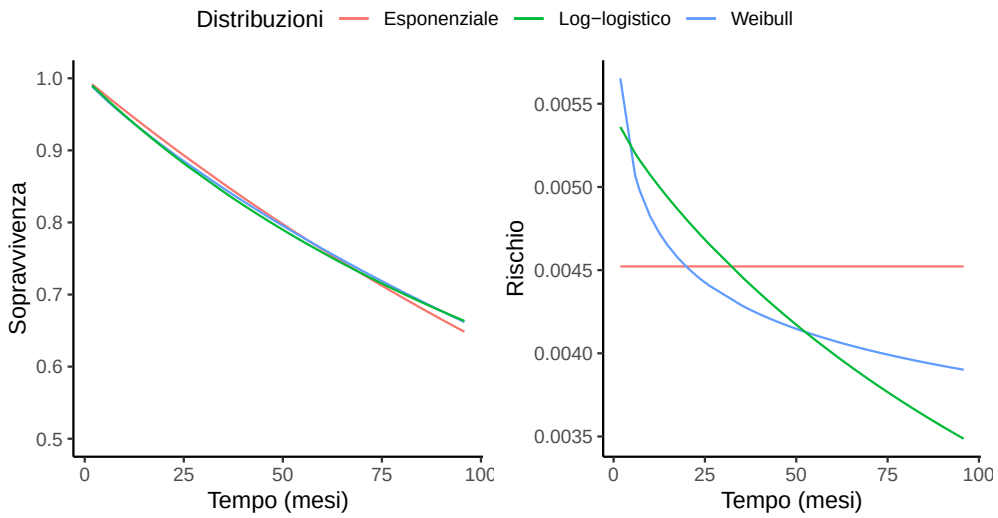


Figura 5: Esempio, sui dati oggetto d'analisi, dell'impatto sul rischio e sulla sopravvivenza, indotto dalla diversa specificazione della distribuzione per la funzione di rischio, per i principali modelli considerati

Rilevante osservare che per il modello esponenziale il rischio risulta costante nel tempo.

2.5 Specificazione del modello

Nel paragrafo 2.4 sono state illustrate le principali distribuzioni utilizzate nell'analisi di sopravvivenza per specificare la forma funzionale del rischio. In generale, tuttavia, qualsiasi sia la distribuzione scelta, bisogna tenere conto delle censure. A tal fine si consideri c l'evento indicante la censura, e sia

$$\delta_i = \begin{cases} 1 & \text{se} \quad t_i \leq c \\ 0 & \text{se} \quad t_i > c \end{cases}$$

ovvero un indicatore che è pari a uno se al tempo t_i il paziente non ha esperito la censura, zero altrimenti. Sia $\mathbf{X}$ una matrice di dimensioni $n \times p$, dove n rappresenta il numero di osservazioni e p il numero di covariate, sia $\boldsymbol{t} = (t_1, \dots t_n)$

il vettore dei tempi all'evento e sia $\boldsymbol{\delta}$ il vettore contenete gli indici di evento e di censura. Risulta ora possibile definire, a prescindere dal modello, come poter ricavare la funzione di verosimiglianza:

$$\mathscr{L}(\boldsymbol{\vartheta}, \mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta}) = \prod_{i=1}^{n} h(t_i \mid \cdot)^{\delta_i} S(t_i \mid \cdot).$$

L'interpretazione diviene dunque la seguente: avendo introdotto l'indicatore della presenza di censura, la funzione di rischio viene attivata, per ciascun individuo, soltanto se esso, in corrispondenza dell'istante temporale t_i, non ha esperito la censura.

2.6 Principali indici di accuratezza prognostica

Nell'ambito dell'analisi di sopravvivenza, gli indici per valutare l'accuratezza del modello differiscono rispetto agli usuali indici che si è abituati a considerare per problemi di regressione e classificazione. In particolare, specificando un modello per il tempo all'evento le previsioni di tale modello forniscono il rischio stimato nel tempo per ciascun individuo a partire dal profilo di covariate che esso presenta. Considerando il rischio stimato, che verrà denotato con il simbolo $\hat{\eta}$, è possibile definire tre indici che quantifichino l'accuratezza prognostica del modello, in particolare: il *Concordance Index* (C-index) di Harrell Jr et al. (1996), il *Brier Score*, per i cui dettagli dimostrativi si consiglia la lettura del testo di Brier (1950), e l'*Area Under the Curve* (AUC) tempo-dipendente.

1. **Il C-index** di Harrel si basa sull'idea che: se il modello costruito è valido allora agli individui che hanno registrato tempi all'evento più brevi dovrebbero essere associati rischi stimati più alti rispetto ai pazienti che

registrano un tempo all'evento più lungo. Considerando una coppia di individui, l'individuo con il rischio stimato associato più elevato dovrebbe aver esperito l'evento prima dell'altro individuo.

Si osservi che il C-index tiene conto anche della censura, infatti per ogni coppia di individui i, j con $i \neq j$ si considerino i rispettivi rischi stimati e i tempi all'evento, o di censura T_i, T_j.

- Se sia l'individuo i che l'individuo j, esperiscono l'evento allora T_i e T_j rappresentano il tempo in cui si verifica l'evento per i due individui rispettivamente. In tal caso si dice che la coppia (i, j) è una coppia concordante se $\hat{\eta}_i > \hat{\eta}_j$ e $T_i < T_j$, ed è una coppia discordante se $\hat{\eta}_i > \hat{\eta}_j$ e $T_i > T_j$. Infatti la coerenza vorrebbe che $\hat{\eta}_i > \hat{\eta}_j$ allora l'individuo i esperisca l'evento prima dell'individuo j.
- Se sia l'individuo i che l'individuo j, non esperiscono l'evento allora T_i e T_j rappresentano il tempo in cui si verifica la censura per i due individui rispettivamente. In tal caso, non potendo sapere chi tra i due individui ha esperito per primo l'evento, la coppia (i, j) non viene considerata ai fini del calcolo del C-index.
- Infine, se uno tra i due individui i, j, non esperisce l'evento allora si osserva un solo evento. Ora, se è l'individuo i a presentare l'evento a tempo T_i allora T_j è il tempo di censura dell'individuo j e se è l'individuo j a presentare l'evento a tempo T_j allora T_i risulta il tempo di censura per l'individuo i.

  Ora, si supponga che sia proprio l'individuo i a esperire l'evento (vale il medesimo ragionamento, con segni invertiti qualora fosse l'individuo j a presentare l'evento).

Se $T_j < T_i$, allora non è possibile sapere con certezza chi ha esperito l'evento per primo, dal momento che T_j è censorizzato, quindi non si considera tale coppia nel calcolo del C-index.

Se, al contrario, $T_j > T_i$, allora è possibile saper con certezza che l'individuo i è ha esperito l'evento prima dell'individuo j. Quindi, (i, j) è una coppia concordante se $\hat{\eta}_i > \hat{\eta}_j$, ed è una coppia discordante se $\hat{\eta}_i < \hat{\eta}_j$.

Ne discende che il C-Index può essere dunque calcolato come il rapporto tra il numero di coppie concordanti e la somma tra il numero di coppie concordanti e quelle discordanti. In altri termini, rappresenta la quota di coppie concordanti predette dal modello:

$$\text{C-index} = \frac{\sum_{i \neq j} \mathbb{I}(\hat{\eta}_i > \hat{\eta}_j)\mathbb{I}(T_i < T_j)\delta_i}{\sum_{i \neq j} \mathbb{I}(T_i < T_j)\delta_i},$$

dove, indicando con c il tempo di censura, si ha:

$$\delta_i = \begin{cases} 1 & \text{se} \quad t_i \leq c \\ 0 & \text{se} \quad t_i > c \end{cases} \qquad \mathbb{I}(\hat{\eta}_i > \hat{\eta}_j) = \begin{cases} 1 & \text{se} \quad \hat{\eta}_i > \hat{\eta}_j \\ 0 & \text{se} \quad \hat{\eta}_i \leq \hat{\eta}_j \end{cases}$$

$$\text{e} \quad \mathbb{I}(T_i < T_j) = \begin{cases} 1 & \text{se} \quad T_i < T_j \\ 0 & \text{se} \quad T_i \geq T_j \end{cases}$$

L'indice può variare tra zero e uno, tanto è più vicino a uno tanto più è alta la proporzione di coppie concordanti, tanto migliore è il modello specificato.

2. **Il Brier Score** (BS) è intuitivamente molto simile al *Mean Squared Error*

(MSE); l'idea principale alla base degli indici di accuratezza prognostica è quella di verificare e quantificare lo scarto sussistente tra rischio stimato dal modello, rischio atteso, e il rischio osservato. Contrariamente a un problema di natura, ad esempio, di classificazione, in cui è chiaro quale sia la risposta osservata, in ambito di sopravvivenza è meno chiaro che cosa sia il rischio osservato. La strategia più naturale per definire il rischio osservato è quella di stabilire degli *end-point* temporali, all'interno dell'intervallo di tempo a disposizione, e verificare l'incidenza di eventi all'interno di ciascun sottointervallo specificato.

Più dettagliatamente, se si stabilisce un *end-point* a tempo $t = t_e$, si suddivide il rischio in sottointervalli (tipicamente decili, quintili, quartili) e si calcola, all'interno di ciascun sottointervallo di rischio, la differenza al quadrato tra il rischio stimato e l'incidenza dell'evento osservata. Si calcola quindi il valor medio di tali differenze ottenendo il Brier Score per l'*end-point* di interesse.

Formalmente, sia Q il numero di suddivisioni dell'intervallo di rischio per il generico *end-point* t_e, sia o_i l'incidenza di eventi osservata nell'intervallo i e $\hat{\eta}_i$ la stima del rischio per l'intervallo i-simo fornita dal modello. Il Brier Score per l'*end-point* t_e risulta:

$$\mathrm{BS}_{t_e} = \frac{1}{Q} \sum_{i=1}^{Q} \left(\hat{\eta}_i - o_i\right)^2 .$$

Si noti che il BS può essere calcolato anche fissando come *end-point* l'ultimo istante temporale a disposizione; in questo caso si starebbe valutando la calibrazione dell'intero modello, si ricordi tuttavia che BS_{t_e} è una misura

d'errore utile a scopo comparativo.

L'elemento chiave di quest'indice risiede proprio nel fatto che, fissati diversi *end-point*, esso può fornire informazioni su come varia l'errore nel tempo e, dunque, se fidarsi maggiormente del predittore di rischio a tempi relativamente vicini o se il predittore funziona meglio nel lungo periodo.

Dunque il BS oltre a fornire indicazioni circa l'affidabilità in ottica temporale del previsore di rischio, svolge un ruolo di fondamentale importanza nel confronto tra modelli e, infine, risulta l'indice d'elezione per valutare la calibrazione: si potrebbe infatti presentare uno scenario in cui, sebbene il modello sembri registrare *performance* elevate (ovvero presenti un AUC, in corrispondenza di un *end-point*, molto elevato) in realtà non risulta ben calibrato. In altri termini il modello può apparire, sotto il profilo dell'AUC, uno stimatore accurato del rischio, tuttavia tale misura risulta viziata se si guarda la calibrazione, ovvero un AUC alto può essere ottenuto sottostimando sistematicamente rischi elevati e contemporaneamente sovrastimando rischi bassi.

Infine una versione più generale del BS e meno influenzata dalla scelta del numero di quantili in cui suddividere l'intervallo, si può ottenere considerando lo scarto quadratico medio tra ciascun rischio stimato all'*end-point* considerato e l'indicatore di evento. In altri termini, se si considera t_e, e o_i con $o_i = \delta_i$ con δ_i pari a 0 se non si verifica l'evento e pari a 1 se si verifica l'evento, allora il BS complessivo per l'*endpoint* t_e è dato da:

$$\mathrm{BS}_{t_e} = \frac{1}{n} \sum_{i=1}^{n} \left(\hat{\eta}_i - o_i\right)^2,$$

con $\hat{\eta}_i$ stima del rischio per l'i-esimo individuo a tempo t_e

3. **L'AUC tempo-dipendente** non differisce molto rispetto alla sua versione standard. Di nuovo, similmente al *Brier Score*, si stabilisce un *end-point* e, qui vi è l'elemento di differenza, si valuta se per ogni coppia di valori di rischio stimato, la probabilità che il rischio dell'i-simo individuo sia maggiore del rischio del j-simo individuo, dato il vero stato di ciascun individuo (δ_i pari a 0 se non si verifica l'evento e pari a 1 se si verifica l'evento). Di nuovo, si considera il paragone tra rischi stimati e rischi osservati, o più specificamente, tra rischi stimati e i veri stati dell'individuo. Il calcolo dell'AUC segue quindi la stessa logica del caso binario, l'unica differenza è che bisogna specificare l'*end-point* di interesse. Anche in questo caso l'*end-point* può essere fissato all'ultimo istante disponibile per avere una misura complessiva dell'AUC del modello. Formalmente:

$$\text{AUC} = \frac{1}{\text{card}\left\{i : i \neq j\right\}} \sum_{i \neq j} \mathbb{P}\left(\hat{\eta}_i > \hat{\eta}_j \mid \delta_i = 1; \delta_j = 0\right)$$

Si ricordi che, solitamente, gli indici di cui al punto 1,2,3 vengono considerati congiuntamente, proprio per valutare sia l'aspetto dell'accuratezza sia quello della calibrazione di un indice prognostico.

2.7 *Restricted Cubic Spline* nell'analisi di sopravvivenza

Dal momento che in Tabella 1 i coefficienti del Sarculator contengono due coefficienti associati alla componente non lineare delle *Restricted Cubic Splines* e vi è

un solo coefficiente associato al termine non lineare è opportuno chiarire a che cosa corrisponda una trasformata di variabile con RCS. A tal fine si consideri la seguente

Definizione 2.7.1 *Una espansione di base è una funzione f tale che:*

$$f(x) = \sum_{j=1}^{M} \alpha_j B_j(x),$$

dove $\alpha_j \in \mathbb{R} \quad \forall j = 1, \ldots, M$ e $B_j(x)$ sono funzioni note per ogni $j = 1, ..., M$, dette funzioni di base.

Dalla Definizione 2.7.1 è possibile fornire la seguente

Definizione 2.7.2 *Una B-spline di grado M è una funzione di base composta da M funzioni di base e da nodi $\xi_1, \ldots, \xi_K$, avente derivate continue di ordine $0, \ldots, M-1$ in ciascuno nodo.*

In sintesi l'idea alla base delle B-spline è quella di considerare funzioni di base che siano locali, cioè diverse da zero per una piccola porzione dell'intervallo della covariata e che siano delimitate superiormente.

Per ulteriori dettagli relativi alla regressione tramite l'utilizzo di spline si veda il Cap. 4 di Azzalini and Scarpa (2012). Dal momento che le *splines* cubiche sono spesso instabili sulle code, cioè prima del primo nodo e dopo l'ultimo nodo, si è soliti porre come vincolo la linearità sulle code. Una B-spline di grado 3 soggetta a vincolo di linearità sulle code è detta *Natural Cubic Spline* (NCS) o *Restricted Cubic Splines*. In analisi di sopravvivenza si è più soliti utilizzare il termine *Restricted Cubic Splines* perché ci si riferisce a una rappresentazione differente delle NCS rispetto a quella usuale. Siano $\xi_1, \ldots, \xi_k$ k nodi, una NCS a k nodi per x si può scrivere come:

$$f(x) = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_{k-1} x_{k-1}, \tag{4}$$

dove $x_1 = x$ e per $j = 1, \ldots, k-2$,

$$x_{j+1} = (x - \xi_j)^3_+ - \frac{(x - \xi_j)^3_+ (\xi_k - \xi_j)}{(\xi_k - \xi_{k-1})} + \frac{(x - \xi_k)^3_+ (\xi_{k-1} - \xi_j)}{(\xi_k - \xi_{k-1})}, \tag{5}$$

con

$$(x - \xi_j)^3_+ = \begin{cases} (x - \xi_j)^3 & \text{se} \quad x > 0 \\ 0 & \text{se} \quad x \leq 0 \end{cases},$$

j-sima funzione di base. Infatti, una volta stimati $\alpha_0, \ldots \alpha_{k-1}$ si può riscrivere la NCS nella canonica forma:

$$f(x) = \alpha_0 + \alpha_1 x + \alpha_2 (x - \xi_1)^3_+ + \alpha_3 (x - \xi_2)^3_+ + \cdots + \alpha_{k+1} (x - \xi_{k+1})^3_+ \tag{6}$$

si calcolano α_k e α_{k+1} come:

$$\alpha_k = \frac{\alpha_2(\xi_1 - \xi_k) + \alpha_3(\xi_2 - \xi_k) + \cdots + \alpha_{k-1}(\xi_{k-2} - \xi_{k-1})}{(\xi_k - \xi_{k-1})}.$$

$$\alpha_{k+1} = \frac{\alpha_2(\xi_1 - \xi_{k-1}) + \alpha_3(\xi_2 - \xi_{k-1}) + \cdots + \alpha_{k-1}(\xi_{k-2} - \xi_{k-1})}{(\xi_{k-1} - \xi_k)}.$$

Si è fornita la duplice natura tramite cui caratterizzare una NCS perché, nel contesto dell'analisi di sopravvivenza, si è soliti prediligere la rappresentazione espressa nella relazione (4), per distinguere le due rappresentazioni la relazione (4) viene appositamente chiamata *Restricted Cubic Splines*, per non creare ambiguità con la relazione espressa nella (6).

In virtù della relazione (5), se si considera una RCS con 3 nodi, allora si ha

che $x_1 = x$ e $j = 1, \ldots, k-2$, con $k = 3$, ovvero $j = 1$. Sicché:

$$x_2 = (x - \xi_1)_+^3 - \frac{(x - \xi_1)_+^3(\xi_3 - \xi_1)}{(\xi_3 - \xi_2)} + \frac{(x - \xi_3)_+^3(\xi_2 - \xi_j)}{(\xi_3 - \xi_2)},$$

dunque, per la relazione (4) si ha che:

$$f(x) = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2$$

ovvero

$$f(x) = \alpha_0 + \alpha_1 x + \alpha_2 \left[(x - \xi_1)_+^3 - \frac{(x - \xi_1)_+^3(\xi_3 - \xi_1)}{(\xi_3 - \xi_2)} + \frac{(x - \xi_3)_+^3(\xi_2 - \xi_j)}{(\xi_3 - \xi_2)}\right]$$

e se l'intercetta è nulla si ha:

$$f(x) = \alpha_1 x + \alpha_2 \left[(x - \xi_1)_+^3 - \frac{(x - \xi_1)_+^3(\xi_3 - \xi_1)}{(\xi_3 - \xi_2)} + \frac{(x - \xi_3)_+^3(\xi_2 - \xi_j)}{(\xi_3 - \xi_2)}\right]. \quad (7)$$

La relazione (7) chiarisce a cosa corrispondono i coefficienti che coinvolgono le RCS nel Sarculator. Sotto questa rappresentazione infatti, α_1 è il coefficiente associato alla componente lineare, mentre α_2 incorpora l'effetto non lineare della RCS. Questo è il modo in cui sono stati ricavati i coefficienti dell'età e del diametro della lesione riportati in Tabella 1.

In contesti di analisi di sopravvivenza viene utilizzata ampiamente la rappresentazione (7) piuttosto che la (6), in quanto si preferisce incorporare l'effetto non lineare in un'unica componente. Per ulteriori dettagli circa le due rappresentazioni fornite si rimanda al paragrafo 2.4.5 di Harrell et al. (2001).

3 Modellizzazione bayesiana

Rispetto all'ambito classico, ove ϑ è visto come parametro oggetto di inferenza, in ambito bayesiano ϑ è considerata una variabile aleatoria (v.a.). Pertanto, la statistica bayesiana porta a un cambio di paradigma: non solo le realizzazioni $\boldsymbol{x}$ sono una v.a. ma lo è anche ϑ, oggetto di inferenza e, pertanto, al contrario della statistica classica, anche a ϑ è necessario associare una legge di distribuzione.

La funzione di verosimiglianza $\mathscr{L}(\boldsymbol{x}; \vartheta)$, in ambito bayesiano viene scritta come una legge di distribuzione condizionata: $\mathscr{L}(\boldsymbol{x} \mid \vartheta)$.

Denominata $\pi(\vartheta)$ la legge di distribuzione per ϑ è dunque necessario ricondursi alla legge congiunta $\pi(\boldsymbol{x}, \vartheta)$, di $\boldsymbol{x}$ e ϑ. Per facilitare la trattazione si considererà ϑ come unico parametro e $\boldsymbol{x}$ come vettore nei seguenti paragrafi, ricordando che, qualora $\boldsymbol{\vartheta}$ fosse un vettore di parametri e $\mathbf{X}$ una matrice, varrebbero esattamente le medesime considerazioni. Infatti a partire dal capitolo 4 si farà necessariamente e inevitabilmente riferimento a $\boldsymbol{\vartheta}$ come vettore di parametri e a $\mathbf{X}$ come una

matrice di dati osservati (standardizzati) e non già un solo vettore. Tuttavia per rendere più agevole l'introduzione all'ambito bayesiano si è preferito alleggerire la notazione in questa prima parte.

3.1 Distribuzione a priori e a posteriori

In virtù del teorema di Bayes, definita $\pi(\vartheta)$ la legge di distribuzione **a priori** per ϑ, e definita la legge di distribuzione congiunta $\pi(\boldsymbol{x}, \vartheta)$, la distribuzione a posteriori di ϑ risulta:

$$\pi(\vartheta \mid \boldsymbol{x}) = \frac{\mathscr{L}(\boldsymbol{x} \mid \vartheta)\pi(\vartheta)}{\int_\Theta \mathscr{L}(\boldsymbol{x} \mid \vartheta)\pi(\vartheta)d\vartheta},$$

posto $m(\boldsymbol{x}) = \int_\Theta \mathscr{L}(\boldsymbol{x} \mid \vartheta)\pi(\vartheta)d\vartheta$ si ha:

$$\pi(\vartheta \mid \boldsymbol{x}) = \frac{\pi(\boldsymbol{x}, \vartheta)}{m(\boldsymbol{x})}$$

con $m(\boldsymbol{x})$ detta distribuzione predittiva iniziale, dal momento che, integrando rispetto al supporto della v.a. ϑ si elimina la dipendenza di $m(\boldsymbol{x})$ da ϑ medesimo. Ne discende che, posto $\frac{1}{m(\boldsymbol{x})} = c$, con $c > 0$ c, è lecita la riscrittura:

$$\pi(\vartheta \mid \boldsymbol{x}) = c \times \mathscr{L}(\boldsymbol{x} \mid \vartheta)\pi(\vartheta)$$

o, equivalentemente:

$$\pi(\vartheta \mid \boldsymbol{x}) \propto \mathscr{L}(\boldsymbol{x} \mid \vartheta)\pi(\vartheta),$$

con $\propto$ simbolo indicante la proporzionalità tra la distribuzione a posteriori per ϑ e il prodotto tra la verosimiglianza condizionata e la distribuzione a priori per ϑ. Si osservi che $\mathscr{L}(\boldsymbol{x} \mid \vartheta)\pi(\vartheta)$ è anche detto nucleo della distribuzione a posteriori. La nozione di proporzionalità risulterà utile in quanto, nell'ambito della statistica bayesiana, molto raramente è possibile calcolare analiticamente

$$m(\boldsymbol{x}) = \int_{\Theta} \mathscr{L}(\boldsymbol{x} \mid \vartheta)\pi(\vartheta)\mathrm{d}\vartheta.$$

Si pone pertanto il problema di come campionare da una distribuzione a posteriori che, tuttavia, non si riesce a determinare analiticamente a meno di una costante moltiplicativa. In tal senso la nozione di nucleo della distribuzione a posteriori ricoprirà un ruolo di primaria importanza per il campionamento dalla a posteriori in quanto, per campionare da essa, non è strettamente necessario il calcolo di $m(\boldsymbol{x})$. Infatti, è sufficiente riuscire a derivare analiticamente la forma che assume il nucleo della a posteriori per essere in grado di campionare dall'intera distribuzione attraverso opportuni metodi computazionali.

3.2 Metodi computazionali

3.2.1 Metodi di approssimazione di un funzionale della a posteriori

Supponendo dunque che non si riesca a calcolare $m(\boldsymbol{x})$ e, conseguentemente, che non si riesca a derivare l'espressione analitica della distribuzione a posteriori, a meno della costante di normalizzazione, verrà brevemente fornita una disamina sui metodi più utilizzati per generare comunque dalla distribuzione a posteriori

o approssimarne un suo funzionale $g(\vartheta)$. Verranno invece approfonditi più dettagliatamente i metodi MCMC (*Monte Carlo Markov Chains*), in particolare il metodo di Metropolis-Hastings che si è rivelato il più adatto ai fini di generare un campione dalla distribuzione a posteriori nel contesto oggetto d'analisi.

I principali metodi per approssimare un funzionale di ϑ quando non è nota la costante di normalizzazione $m(\boldsymbol{x})$ sono:

1. Metodo di Laplace
2. *Monte Carlo Importance Sampling*

Per il metodo di Laplace si veda Tierney and Kadane (1986) per una disamina approfondita. Un altro metodo di approssimazione che, a differenza dei due metodi precedenti, non è volto ad approssimare un funzionale della distribuzione a posteriori ma la distribuzione medesima è il Metodo di approssimazione Normale.

Si fornirà una breve disamina sul *Monte Carlo Importance Sampling* in quanto propedeutico alla comprensione del *Bridge Sampling* che verrà trattato nel paragrafo 4.6.1, e sul metodo dell'approssimazione normale utile nella fase di specificazione della matrice di varianza-covarianza della *proposal distribution*, si veda il paragrafo 4.4.2.

Monte Carlo Importance Sampling

Sia $\vartheta_1, \vartheta_2, \ldots, \vartheta_n$ un campione casuale, (iid), dalla distribuzione a posteriori $\pi(\vartheta \mid \boldsymbol{x})$. Di nuovo, al fine di effettuare l'inferenza su un funzionale di ϑ, $g(\vartheta)$, si può valutare

$$\mathbb{E}^{\pi(\cdot|\boldsymbol{x})}\left[g(\vartheta)\right]$$

e pensare di approssimarlo tramite metodo Monte Carlo con

$$\overline{g}_n = \frac{1}{n}\sum_{i=1}^{n} g(\vartheta_i).$$

Tuttavia, per applicare il metodo Monte Carlo servirebbe campionare dalla distribuzione a posteriori che però non è nota a causa della costante di normalizzazione. L'idea del *Monte Carlo Importance Sampling* è dunque quella di scegliere una funzione "sufficientemente vicina", $h(\vartheta)$, alla distribuzione a posteriori avente il medesimo supporto. Scelta $h(\vartheta)$, detta funzione di importanza (da cui il nome *Importance Sampling*), si estraggono $\vartheta'_1, \vartheta'_2, \ldots, \vartheta'_n$ realizzazioni iid da $h(\vartheta)$. In altri termini si estrae un campione Monte Carlo da $h(\vartheta)$. Dunque si approssima $\mathbb{E}^{\pi(\cdot|\boldsymbol{x})}\left[g(\pi\left(\vartheta\right)\right]$ con

$$\tilde{g}_n = \frac{1}{n}\sum_{i=1}^{n} w_i g\left(\vartheta'_i\right) \qquad \text{dove} \qquad w_i = \frac{\pi\left(\vartheta'_i \mid \boldsymbol{x}\right)}{h\left(\vartheta'_i\right)}.$$

Se $h(\vartheta) = \pi\left(\vartheta \mid \boldsymbol{x}\right)$ allora $w_i = 1$ e quindi verrebbe applicato il metodo Monte Carlo standard. Invece, tipicamente, essendo noto solo il nucleo della distribuzione a posteriori esiste almeno un ϑ'_i tale che $w_i \neq 1$.

Si osservi che w_i contiene ancora l'intera distribuzione a posteriori, ciononostante è possibile dimostrare che l'approssimazione *Monte Carlo Importance Sampling* di $g\left(\vartheta\right)$ è data da

$$\tilde{g}_n = \frac{(1/n)\sum_{i=1}^{n} g(\vartheta_i^*)\mathscr{L}(\boldsymbol{x} \mid \vartheta_i^*)}{(1/n)\sum_{i=1}^{n} \mathscr{L}(\boldsymbol{x} \mid \vartheta_i^*)},$$

in cui è sufficiente conoscere la funzione di verosimiglianza.

Dimostrazione

Considerando la definizione di distribuzione a posteriori, è lecita la seguente

riscrittura:

$$\begin{aligned}\mathbb{E}^{\pi(\cdot|\boldsymbol{x})}\left[g\left(\vartheta\right)\right] &= \int_{\Theta} g\left(\vartheta\right) \frac{\mathscr{L}(\boldsymbol{x} \mid \vartheta)\pi\left(\vartheta\right)}{m(\boldsymbol{x})} \mathrm{d}\vartheta \\ &= \frac{1}{m(\boldsymbol{x})} \int_{\Theta} g\left(\vartheta\right) \mathscr{L}(\boldsymbol{x} \mid \vartheta)\pi\left(\vartheta\right) \mathrm{d}\vartheta \\ &= \frac{\int_{\Theta} g\left(\vartheta\right) \mathscr{L}(\boldsymbol{x} \mid \vartheta)\pi\left(\vartheta\right) \mathrm{d}\vartheta}{\int_{\Theta} \mathscr{L}(\boldsymbol{x} \mid \vartheta)\pi\left(\vartheta\right) \mathrm{d}\vartheta},\end{aligned}$$

da cui è possibile porre $g_1\left(\vartheta\right) = g\left(\vartheta\right)\mathscr{L}(\boldsymbol{x} \mid \vartheta)$, $g_2\left(\vartheta\right) = \mathscr{L}(\boldsymbol{x} \mid \vartheta)$, e porre la funzione di importanza, $h(\vartheta)$, esattamente pari alla distribuzione a priori $\pi\left(\vartheta\right)$, in modo tale da poter estrarre $\vartheta'_1, \vartheta'_2, \ldots, \vartheta'_n$ da h. Si ottiene pertanto:

$$\begin{aligned}\mathbb{E}^{\pi(\cdot|\boldsymbol{x})}\left[g\left(\vartheta\right)\right] &= \frac{\int_{\Theta} g\left(\vartheta\right) \mathscr{L}(\boldsymbol{x} \mid \vartheta)\pi\left(\vartheta\right) \mathrm{d}\vartheta}{\int_{\Theta} \mathscr{L}(\boldsymbol{x} \mid \vartheta)\pi\left(\vartheta\right) \mathrm{d}\vartheta} \\ &\approx \frac{\mathbb{E}^{\pi(\cdot)}\left[g_1\left(\vartheta\right)\right]}{\mathbb{E}^{\pi(\cdot)}\left[g_2\left(\vartheta\right)\right]} = \frac{\mathbb{E}^{\pi(\cdot)}\left[g\left(\vartheta\right)\mathscr{L}(\boldsymbol{x} \mid \vartheta)\right]}{\mathbb{E}^{\pi(\cdot)}\left[\mathscr{L}(\boldsymbol{x} \mid \vartheta)\right]} \\ &\approx \frac{(1/n)\sum_{i=1}^{n} g(\vartheta_i^*)\mathscr{L}(\boldsymbol{x} \mid \vartheta_i^*)}{(1/n)\sum_{i=1}^{n} \mathscr{L}(\boldsymbol{x} \mid \vartheta_i^*)}. \qquad \square\end{aligned}$$

Si osservi che una logica molto simile verrà applicata al paragrafo 4.6.1 per la costruzione del *bridge sampling*.

3.2.2 Metodo dell'approssimazione normale

Prima di approfondire la trattazione sui metodi computazionali che permettano di campionare dalla distribuzione a posteriori in modo decisamente più raffinato, si concluderà questa breve disamina sui metodi di approssimazione con il metodo dell'approssimazione normale che, rispetto al metodo di Laplace e al *Monte Carlo Importance Sampling*, non approssima un funzionale della distribuzione a posteriori

ma l'intera distribuzione.

Sotto le note condizioni di regolarità, e per $n \to \infty$, è lecita l'approssimazione riportata da Van der Vaart (2000):

$$\boldsymbol{\vartheta} \approx \mathcal{N}\left(\tilde{\boldsymbol{\vartheta}}, \tilde{\boldsymbol{\Sigma}}\right),$$

ovvero, la distribuzione a posteriori può essere approssimata come una normale multivariata, avente come media la moda a posteriori $\tilde{\boldsymbol{\vartheta}}$, e come matrice di varianza-covarianza $\tilde{\boldsymbol{\Sigma}}$. Il problema si riduce a capire come ricavare, dunque, i parametri per la normale. Si osservi, a riguardo, che la moda a posteriori $\tilde{\boldsymbol{\vartheta}}$ coincide con il punto di massimo della funzione di verosimiglianza, la quale è nota se si sceglie come distribuzione a priori una distribuzione non informativa di Laplace, ovvero una qualsiasi funzione del tipo $\pi(\boldsymbol{\vartheta}) = c$, con $c \in \mathbb{R}$. Infatti il problema con la a priori non informativa di Laplace risiede solitamente nel fatto che, essendo una distribuzione impropria (non integra a uno sul suo supporto), essa può essere impiegata in fase di elicitazione se e solo se la distribuzione a posteriori risulta propria. Dal momento che, in questo caso, la distribuzione a posteriori è una normale, si ha la garanzia che integrerà a uno, pertanto la distribuzione non informativa di Laplace, benché impropria, è applicabile.

Infine, la matrice $\tilde{\boldsymbol{\Sigma}}$ è l'inversa della matrice di elementi $-\frac{\partial^2}{\partial\vartheta_i \partial\vartheta_j} \log \pi(\boldsymbol{\vartheta} \mid \boldsymbol{x})\Big|_{\boldsymbol{\vartheta}=\tilde{\boldsymbol{\vartheta}}}$, che coincide con l'informazione osservata.

Il metodo dell'approssimazione fornisce l'occasione per mostrare come l'informazione osservata di Fisher, che in ambito frequentista rappresenta l'hessiana valutata nel punto di massima verosimiglianza, possa essere vista. in ambito bayesiano, come una misura di accuratezza della distribuzione a posteriori

per ogni campione dato, seguendo l'interpretazione riportata da Gelman et al. (1995). Infatti, si consideri la distribuzione log a posteriori $\log \pi(\boldsymbol{\vartheta} \mid \mathbf{X})$ e si applichino gli sviluppi di Taylor arrestati al second'ordine, centrati sulla moda a posteriori $\tilde{\boldsymbol{\vartheta}}$; si ha:

$$\log \pi(\boldsymbol{\vartheta} \mid \boldsymbol{x}) \approx \log \pi(\tilde{\boldsymbol{\vartheta}} \mid \boldsymbol{x}) + \frac{1}{2} \left(\boldsymbol{\vartheta} - \tilde{\boldsymbol{\vartheta}}\right)' \boldsymbol{\mathcal{H}}\Big|_{\boldsymbol{\vartheta}=\tilde{\boldsymbol{\vartheta}}} \left(\boldsymbol{\vartheta} - \tilde{\boldsymbol{\vartheta}}\right)$$

ed essendo $\log \pi(\boldsymbol{\vartheta} \mid \mathbf{X})$ funzione di $\boldsymbol{\vartheta}$, il primo termine è costante mentre il secondo è proporzionale al logaritmo di una funzione di densità normale, da cui l'approssimazione:

$$\boldsymbol{\vartheta} \approx \mathcal{N}\left(\tilde{\boldsymbol{\vartheta}}, \tilde{\boldsymbol{\Sigma}}\right)$$

con $\tilde{\boldsymbol{\Sigma}} = -\boldsymbol{\mathcal{H}}\Big|^{-1}_{\boldsymbol{\vartheta}=\tilde{\boldsymbol{\vartheta}}} = \boldsymbol{\mathcal{I}}^{-1}\left(\tilde{\boldsymbol{\vartheta}}\right)$ con $\boldsymbol{\mathcal{I}}^{-1}\left(\tilde{\boldsymbol{\vartheta}}\right)$ informazione osservata. Questa considerazione si rivelerà utile in fase di specificazione della matrice di varianza-covarianza della *proposal distribution*, si veda il paragrafo 4.4.2.

3.3 Metodi MCMC: *Monte Carlo Markov Chains*

3.3.1 Richiami sulle catene di Markov e loro proprietà

Dal momento che i metodi MCMC si basano sulle catene di Markov, si richiameranno i concetti utili per la comprensione dei meccanismi che stanno alla base di tali metodi computazionali. Giova, a tal fine, partire dalla seguente definizione.

Definizione 3.3.1 (Processo Stocastico) *Sia T un generico insieme e sia $\mathcal{S}$ l'insieme rappresentante lo spazio degli stati; sia $(\Omega, \mathscr{F}, \mathbb{P})$ uno spazio probabilistico,*

un processo stocastico $X = (X_t : t \in T)$ *è una collezione di variabili aleatorie definite sullo spazio probabilistico e indicizzate da* t *e tali che* $X_t : \Omega \to \mathcal{S}, \quad \forall t \in T$.

Si ricordi che lo spazio degli stati non è altro che l'insieme dei possibili valori assunti dal fenomeno. La definizione di processo stocastico è estremamente generale, non a caso si possono trovare definizioni diverse formalmente ma del tutto equivalenti, si veda a riguardo Durrett and Durrett (1999). Si osservi inoltre che non sono posti vincoli su che tipo d'insieme debba essere T e nemmeno su che tipo di insieme debba essere lo spazio degli stati $\mathcal{S}$.

In questa sezione si analizzeranno in particolare i processi Markoviani a tempo discreto $X = (X_n : n \geq 0)$ (T è un'insieme finito o, al più, infinito ma numerabile). Si ricordi che, in generale, un processo stocastico $X = (X_t : t \geq 0)$ è detto processo Markoviano se soddisfa la **Proprietà di Markov**.
Sia $(\Omega, \mathcal{F}, \mathbb{P})$ uno spazio di probabilità, con $(\mathcal{F}_s \in I)$ per un insieme di indici I totalmente ordinato; e sia $(S, \mathcal{S})$ uno spazio misurabile. Un processo stocastico adattato, $(X : \Omega \to S, \quad t \in I)$, gode della proprietà di Markov se per ogni $A \in \mathcal{S}$ e per ogni $s, t \in I$, con $s < t$,

$$\mathbb{P}\left(X_t \in A \mid \mathcal{F}_s\right) = \mathbb{P}\left(X_t \in A \mid X_s\right).$$

Se $\mathcal{S}$ è un insieme discreto e $I = \mathbb{N}$ allora un processo gode della proprietà di Markov se

$$\mathbb{P}\left(X_{t+h} = j \mid X_s = i, s \leq t\right) = \mathbb{P}\left(X_{t+h} = j \mid X_t = i\right),$$

con $i, j \in \mathcal{S}$.

Si osservi che la proprietà di Markov garantisce che la probabilità di transizione a tempo $t+h$ dallo stato i allo stato j dato che al tempo $s \leq t$ il processo si trovava in i coincide con la probabilità di transizione dallo stato i allo stato j a tempo $t+h$ dato che il processo si trovavo in i all'istante t. I processi Markoviani sono caratterizzati cioè dall'assenza di memoria, non importano gli stati occupati negli istanti precedenti al tempo presente t, la probabilità di transitare dallo stato i allo stato j dipende solo dal fatto che il processo all'istante t (presente) si trovi nello stato i.

Si osservi che, per quanto attiene ai metodi computazionali necessari da un punto di vista bayesiano, la trattazione avverrà considerando lo spazio degli stati $\mathcal{S}$ come un insieme discreto. In questo caso, i processi Markoviani a tempo discreto e continuo prendono il nome di catene di Markov.

La trattazione, dunque, prenderà in considerazione le catene di Markov a tempo discreto, richiamandone brevemente caratteristiche e proprietà fondamentali.

3.3.1.1 Catene di Markov a tempo discreto

Definizione 3.3.2 (Catena di Markov) *Sia* $X = (X_n : n \geq 0)$ *un processo Markoviano a tempo discreto. Sia* $\mathcal{S}$ *un insieme discreto, allora il processo* $X = (X_n : n \geq 0)$ *è detto catena di Markov a tempo discreto e gode della proprietà di Markov:*

$$\mathbb{P}\left(X_{n+1} = j \mid X_n = i_n, X_{n-1} = i_{n-1}, \ldots, X_0 = i_0\right) = \mathbb{P}(X_{n+1} \mid X_n = i)$$

Inoltre se la catena di Markov è tale che $\mathbb{P}(X_{n+1} = j \mid X_n = i) = p(i,j)$*, cioè non dipende da* n *allora la catena è* ***omogenea****.*

Dunque una catena di Markov omogenea non solo è caratterizzata dall'assenza di memoria ma anche dal fatto che la probabilità di transizione tra gli stati non dipende dall'istante n al quale la catena si trova in uno stato ma solo dallo stato in cui si trova.

D'ora in avanti verranno considerate solo catene di Markov omogenee. Per caratterizzarle sono sufficienti i seguenti due elementi:

1. $\mu_i = \mathbb{P}(X_0 = i)$ con $i \in \mathcal{S}$, $\quad \mu_i \geq 0 \quad \sum_{i \in \mathcal{S}} \mu_i = 1$. Cioè le probabilità con cui la catena parte da un generico stato. Si osservi che se è noto lo stato iniziale della catena allora tale variabile è degenere (sottocaso).
2. Se $\mathcal{S}$ è finito allora il secondo elemento è costituito da una matrice stocastica $\mathbf{P}$ di dimensione $n \times n$ detta matrice di transizione.
   Chiamato $p(i,j) = \mathbb{P}(X_{n+1} = j \mid X_n = i)$ il generico elemento della matrice, che non dipende dall'istante n, e indicante la probabilità di passare da uno stato all'altro, la matrice di transizione ha forma:

$$\mathbf{P} = \begin{pmatrix} p(1,1) & p(1,2) & \dots & p(1,n) \\ p(2,1) & p(2,2) & \dots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ p(n,1) & p(2,n) & \dots & p(n,n) \end{pmatrix}$$

Dunque, gli elementi caratterizzanti univocamente una catena di Markov a tempo discreto omogenea sono $(\boldsymbol{\mu}, \mathbf{P})$. Da questi elementi è possibile definire la catena di Markov a tempo discreto in uno spazio degli stati finito. Nel caso più generale si definisce la matrice di transizione a $m + n$ passi: $\mathbf{P}^{n+m}(i,j)$.

In particolare, il generico elemento della matrice può essere definito tramite le **equazioni di Chapman-Kolmogorov**, si veda Ross (2014) per dettagli, come:

$$p^{n+m}(i,j) = \sum_{h \in \mathcal{S}} p^{(m)}(i,h) p^{(n)}(h,j).$$

Quindi, la probabilità di transizione dallo stato i allo stato j in $m+n$ passi può essere vista come la somma del prodotto di tutte le possibili probabilità di transizione (anche dette traiettorie) dallo stato i allo stato $h \quad \forall h \in \mathcal{S}$ in m passi e tutte le possibili probabilità di transizione dallo stato h allo stato j in n passi.

3.3.1.2 Classificazione degli stati e proprietà

La classificazione degli stati permette di comprendere come si comporta la catena nel lungo periodo, permette quindi di enucleare il comportamento limite della catena, che è quello che trova una diretta connessione con i metodi MCMC.

Definizione 3.3.3 (Accessibilità di uno stato) *Lo stato j è detto accessibile da i, e si indicherà l'accessibilità con $i \to j$, se la probabilità di transizione dallo stato i allo stato j in n passi è positiva:* $\exists n \geq 0 : p^{(n)}(i,j) > 0$.

Definizione 3.3.4 (Stati comunicanti) *Due stati i, j sono detti comunicanti, $i \longleftrightarrow j$, se i comunica con j e j comunica con i, cioè se $i \to j$ e $j \to i$.*

La relazione di comunicabilità è a tutti gli effetti una relazione di equivalenza, infatti soddisfa le seguenti proprietà:

1. $i \longleftrightarrow i$; proprietà riflessiva.
2. Se $i \longleftrightarrow j$ allora $j \longleftrightarrow i$; simmetria.
3. se $i \longleftrightarrow j$ e $j \longleftrightarrow k$ allora $i \longleftrightarrow k$; proprietà transitiva.

Dal momento che la relazione è una relazione di equivalenza essa produce delle partizioni tra loro disgiunte, al cui interno sono contenuti tutti gli stati che comunicano tra loro. Il problema che si pone è capire quando gli stati comunicano tra loro da un punto di vista operativo. A tal fine si considerino le seguenti definizioni e i seguenti risultati.

Definizione 3.3.5 (Insieme Chiuso) *Un insieme $\mathcal{C} \subseteq \mathcal{S}$ è detto chiuso se* $p^{(n)}(i,j) = 0 \quad \forall i \in \mathcal{C},\ \forall n \ e\ j \notin \mathcal{C}$.

In un insieme chiuso, quindi, la probabilità di transizione da un qualsiasi stato appartenente a esso a uno che non vi appartenga è nulla per qualunque numero di passi.

Definizione 3.3.6 (Insieme Irriducibile) *Un insieme $\mathcal{A} \subseteq \mathcal{S}$ è detto irriducibile se tutti gli stati di $\mathcal{A}$ sono comunicanti.*

Si osservi che le partizioni costruite sulla base della relazione di equivalenza sono, per definizione, irriducibili.

Per introdurre il concetto di stato transitorio e stato ricorrente si considerino $X_0 = i$, e la funzione che probabilizza l'evento di ritornare ad i per la prima volta dopo n passi, essendo partiti da i:

$$f_i(n) = \mathbb{P}\left(X_n = i, X_{n-1} \neq i, \ldots, X_1 \neq i \mid X_0 = i\right).$$

Si consideri quindi la seguente funzione:

$$f_i = \sum_{n=1}^{\infty} f_i(n),$$

che rappresenta, pertanto, la probabilità di ritornare ad i per ciascun valore di n. In altri termini, è come se si stesse considerando la probabilità di tornare allo

stato i, essendo partiti dallo stato i, prima o poi. Si forniscono ora le seguenti definizioni:

Definizione 3.3.7 (Stato ricorrente) *Uno stato è detto ricorrente se* $f_i = 1$.

Uno stato ricorrente è quindi uno stato tale per cui se la catena ha già transitato per quello stato ritornerà nel medesimo stato infinite volte.

Definizione 3.3.8 (Stato transitorio) *Uno stato è detto transitorio se* $f_i < 1$.

Intuitivamente se f_i rappresenta la probabilità di ritornare prima o poi in i, il suo complemento a uno, $1 - f_i$, rappresenta la probabilità di non ritornare più allo stato i. Essendo la probabilità non nulla per uno stato transitorio, prima o poi lo stato verrà abbandonato.

Definizione 3.3.9 (Periodo di uno stato) *Si consideri l'insieme che raccoglie tutti i numeri di passi tali per cui la catena possa tornare allo stato i per la prima volta, essendo partita dallo stato i:*

$$\mathcal{I}_i = \left\{ n \geq 1 : p^{(n)}(i,i) > 0 \right\}.$$

Si definisce periodo di uno stato il massimo comun divisore di $\mathcal{I}_i$*. Gli stati aventi massimo comun divisore maggiore di* 1 *sono detti* ***stati periodici****. Stati aventi massimo comun divisore pari a* 1 *sono detti stati* ***aperiodici****.*

Definiti gli elementi teorici utili ai fini della classificazione degli stati, si forniscono indicazioni operative per la classificazione degli stati:

1. Sia $\mathcal{C} \subseteq \mathcal{S}$ un sottoinsieme finito degli stati, allora deve contenere almeno uno stato ricorrente.
2. Se $i \longleftrightarrow j$ e i è uno stato ricorrente, allora anche j è uno stato ricorrente.
3. Se $i \longleftrightarrow j$ e i è uno stato transitorio, allora anche j è uno stato transitorio.

4. Se ogni sottoinsieme $\mathcal{C} \subseteq \mathcal{S}$ è irriducibile, allora ogni insieme contiene solo stati dello stesso tipo.
5. Se $\mathcal{C} \subseteq \mathcal{S}$ è chiuso e irriducibile, allora contiene solo stati ricorrenti (per la 1 e per la 3).
6. Se $\mathcal{C} \subseteq \mathcal{S}$ è aperto e irriducibile, allora è composto da soli stati transitori.
7. Se $i \longleftrightarrow j$, allora i e j hanno lo stesso periodo.

Sulla base delle definizioni e dei risultati appena forniti è possibile analizzare il comportamento della distribuzione e l'equilibrio della catena nel lungo periodo. A tal fine giova introdurre la seguente

Definizione 3.3.10 (Distribuzione stazionaria o invariante) *Si consideri una catena di Markov a tempo discreto, con $\mathcal{S}$ composto da k stati, si denoti con π una probabilità, con $\mathbf{P}$ la matrice di transizione, e si consideri il vettore k-dimensionale*

$$\boldsymbol{\pi} = \begin{pmatrix} \pi(1) \\ \vdots \\ \vdots \\ \pi(k) \end{pmatrix} \quad con \quad \pi(i) > 0 \quad e \quad \sum_{i=1}^{k} \pi(i) = 1.$$

Se $\boldsymbol{\pi}$ è tale che $\boldsymbol{\pi}\mathbf{P} = \boldsymbol{\pi}'$, allora la distribuzione di $\boldsymbol{\pi}$ si definisce stazionaria (o invariante).

Si osservi che se $X_0 \sim \boldsymbol{\pi}$ allora $X_n \sim \boldsymbol{\pi} \quad \forall n \geq 0$. Ovvero, se si estrae il valore iniziale dalla distribuzione di $\boldsymbol{\pi}$ allora anche tutti gli altri valori provengono dalla distribuzione invariante. La catena di Markov risulta, conseguentemente, a componenti identicamente distribuite.

Se si pone quindi $X_0 = \boldsymbol{\pi}$ allora la distribuzione limite sarà certamente

quella di $\boldsymbol{\pi}$. Si osservi che, anche se si parte da una qualsiasi altra distribuzione stazionaria, la distribuzione limite è sempre quella di $\boldsymbol{\pi}$ in virtù della proprietà di Markov, per la quale il valore iniziale della catena non è rilevante.

Si enuncia, infine, il seguente

Teorema 3.1 *Se* $X = (X_n, n \geq 0)$ *è una catena di Markov a tempo discreto, aperiodica e irriducibile allora esiste ed è unica la distribuzione invariante di* $\boldsymbol{\pi}$ *tale che* $\boldsymbol{\pi}\mathbf{P} = \boldsymbol{\pi}'$. *Inoltre, le probabilità di transizione a n passi convergono a* $\boldsymbol{\pi}\mathbf{P} = \boldsymbol{\pi}'$. *Cioè*

$$p^{(n)}(i,j) \underset{n\to\infty}{\longrightarrow} \pi(j) \quad j = 1, \ldots, k$$

Osservazione: il teorema descrive, sotto le assunzioni esplicitate, che asintoticamente il comportamento della probabilità di transizione non dipende più nemmeno dallo stato in cui si trova la catena ma solo dallo stato di arrivo della catena. L'osservazione è valida anche per le marginali, che già di per sé non dipendono da i.

Inoltre, si osservi che se non tutti gli stati sono comunicanti, cioè la catena è riducibile, allora:

- se la catena origina da insiemi transitori, prima o poi li abbandona.
- se la catena origina da insiemi chiusi, dunque irriducibili, allora rimarrà in tali insiemi, e quindi si considerano solo gli stati degli insiemi chiusi per calcolare la distribuzione limite, giacché gli altri insiemi sono irrilevanti poiché irraggiungibili.

Da tali osservazioni si può concludere che il teorema può essere applicato a qualunque catena Markoviana omogenea a tempo discreto.

Un ultimo risultato, utile soprattutto per quanto attiene ai metodi computa-

zionali bayesiani di cui si discuterà nei paragrafi 3.3.1.4, 3.3.1.5, e 3.3.1.7, è dato dalla

Estensione della legge dei grandi numeri per catene Markoviane: sia $X = (X_n, n \geq 0)$ una catena di Markov omogenea, aperiodica, irriducibile, con distribuzione stazionaria $\boldsymbol{\pi}$, allora:

$$\frac{1}{n}\sum_{i=1}^{n} f(X_i) \xrightarrow{p} \mathbb{E}^{\boldsymbol{\pi}}\left[f(X)\right] = \sum_{j=1}^{k} f(j)\pi(j).$$

Si tratta di un'estensione dal momento che si richiede l'omogeneità della catena e non già iid.

Si osservi che sono state trattate le proprietà delle catene di Markov a tempo discreto, ma queste sono facilmente estendibili al caso delle catene di Markov a tempo continuo, per dettagli si veda Ross (2014). Terminati i richiami sugli elementi utili per comprendere i metodi computazionali bayesiani si procede fornendo una giustificazione sulla validità di tali metodi.

3.3.1.3 Giustificazione della validità dei MCMC

I metodi che verranno esposti nei paragrafi 3.3.1.4, 3.3.1.5, 3.3.1.7, come già chiarito, si basano sulle catene di Markov e sull'applicazione del metodo Monte Carlo. Tali metodi sono costruiti su un impianto teorico che ne garantisce la validità. In particolare i metodi MCMC trovano la loro giustificazione nel teorema ergodico che è, in sostanza, l'analogo della legge dei grandi numeri per le catene di Markov. Risulta pertanto opportuno evidenziare i risultati che giustificano il ricorso ai metodi MCMC. In particolare, sia: $(\boldsymbol{X}^n)_{n\geq 1}$ una catena di Markov omogenea, irriducibile, con nucleo di transizione p e legge di distribuzione

stazionaria π, allora π è l'unica distribuzione stazionaria. Inoltre, l'intuizione del teorema ergodico, per i cui dettagli si veda Robert et al. (1999), è la seguente: sia $(\boldsymbol{X}^n)_{n\geq 1}$ una catena di Markov omogenea, irriducibile con nucleo di transizione p e legge di distribuzione stazionaria π, sia g una funzione a valori reali tali che $\mathbb{E}\left(\mid g(x) \mid\right) < \infty$ e sia

$$\overline{g_n} = \frac{1}{n}\sum_{i=1}^{n} g\left(x^i\right),$$

allora

$$\mathbb{P}\left\{\overline{g_n} \underset{n\to\infty}{\longrightarrow} \mathbb{E}^{\pi}\left[g(x)\right]\right\} = 1.$$

Se la catena di Markov è anche aperiodica, allora la distanza tra il kernel di transizione in n passi e la distribuzione iniziale tende a zero. Formalmente, sia $\mathcal{S}$ lo spazio degli stati e sia $\mathcal{A} \subseteq S$ allora:

$$\|p^n(x,\cdot) - \pi(\cdot)\| \underset{n\to\infty}{\longrightarrow} 0 \quad \forall x \in \mathcal{S},$$

con

$$\|p^n(x,\cdot) - \pi(\cdot)| = 2 \sup_{\mathcal{A}\in\mathcal{S}} |p^n(x,\mathcal{A}) - \pi(\mathcal{A})|.$$

Si osservi che quanto appena esposto garantisce l'unicità della distribuzione stazionaria, inoltre il teorema ergodico gioca un ruolo fondamentale in quanto è noto che da una catena di Markov omogenea e irriducibile si avranno campioni identicamente distribuiti, ma per poter applicare il metodo Monte Carlo le osservazioni devono essere anche indipendenti; il teorema di ergodicità, di fatto, garantisce la possibilità di applicare il metodo Monte Carlo dichiarando la convergenza in probabilità. Infine, il fatto che la distanza tra il kernel di transizione in n passi e la distribuzione iniziale tenda a zero sottolinea come, per $n \to \infty$, le

estrazioni dal kernel di transizione possono essere considerate come provenienti dalla distribuzione iniziale stazionaria π. Questi tre risultati forniscono una valida ed esaustiva giustificazione all'utilizzo dei metodi MCMC.

In particolare, considerando congiuntamente i tre risultati si deduce che per poter campionare dalla distribuzione a posteriori, conoscendone solo il nucleo, sarà necessario avviare una catena, definire un periodo di *burn-in* pari a m sufficientemente elevato; non considerare le estrazioni fino a m, in quanto la garanzia di provenienza dalla distribuzione a posteriori la si ha per $n \to \infty$; considerare le estrazioni che avvengono dopo l'm-sima fino alla n-sima come estratte dalla distribuzione π, e trattarle con metodo Monte Carlo. Si osservi che per poter applicare il metodo Monte Carlo non deve essere solo elevato il periodo di *burn-in* ma anche il numero di estrazioni aggiuntive n, altrimenti il metodo fornisce stime non robuste. Per schematizzare:

$$\underbrace{\boldsymbol{\vartheta}^0, \boldsymbol{\vartheta}^1, \boldsymbol{\vartheta}^2, \ldots, \boldsymbol{\vartheta}^m}_{\text{periodo di burn-in}}, \quad \underbrace{\boldsymbol{\vartheta}^{m+1}, \boldsymbol{\vartheta}^{m+2}, \ldots, \boldsymbol{\vartheta}^{m+n}}_{\text{Campione considerabile estratto da } \pi}$$

3.3.1.4 Gibbs Sampling

Sia dato un vettore aleatorio $\boldsymbol{\vartheta} = (\vartheta_1, \ldots, \vartheta_d)$, si supponga di conoscere solo il nucleo della distribuzione a posteriori, se si riescono a ricavare tutte le *full conditional distributions*, ovvero:

$$\pi\left(\vartheta_i \mid \vartheta_1, \vartheta_2, \ldots, \vartheta_{i-1}, \vartheta_{i+1}, \ldots, \vartheta_d, \boldsymbol{x}\right),$$

allora è possibile campionare dalla distribuzione a posteriori tramite la procedura del *Gibbs Sampling*, che consta dei seguenti passi:

1. Si inizializza la catena arbitrariamente:

$$\boldsymbol{\vartheta}_0 = \left(\vartheta_1^0, \ldots, \vartheta_d^0\right).$$

2. Si ricava il punto successivo della traiettoria:

$$\boldsymbol{\vartheta}_1 = \left(\vartheta_1^1, \ldots, \vartheta_d^1\right),$$

dove

$$\begin{aligned}
&\vartheta_1^1 \quad \text{estratto dalla } \textit{full conditional} \quad \pi\left(\vartheta_1 \mid \vartheta_2^0, \vartheta_3^0 \ldots, \vartheta_d^0, \boldsymbol{x}\right), \\
&\vartheta_2^1 \quad \text{estratto dalla } \textit{full conditional} \quad \pi\left(\vartheta_2 \mid \vartheta_1^1, \vartheta_3^0 \ldots, \vartheta_d^0, \boldsymbol{x}\right), \\
&\quad \vdots \\
&\vartheta_d^1 \quad \text{estratto dalla } \textit{full conditional} \quad \pi\left(\vartheta_d \mid \vartheta_1^1, \vartheta_2^1 \ldots, \vartheta_{d-1}^1, \boldsymbol{x}\right).
\end{aligned}$$

3. Si reitera il ciclo di generazione descritto al punto 2 fino a convergenza.

Si può dimostrare che per n ed m sufficientemente elevati il *Gibbs Sampling* fornisce una catena di Markov omogenea. Nel caso oggetto d'analisi non ci si addentrerà più nel dettaglio dell'algoritmo, in quanto le *full conditionals* sono non determinabili analiticamente e si rimanda, per ulteriori dettagli, a Robert et al. (2007). Si analizzerà più dettagliatamente, invece, un altro metodo MCMC che verrà utilizzato ai fini delle analisi: l'algoritmo Metropolis-Hastings.

3.3.1.5 Metropolis-Hastings

Il metodo Metropolis-Hastings, per i cui dettagli si raccomanda la lettura dei lavori di Metropolis et al. (1953) e Hastings (1970), si basa sulla logica di un tipico algoritmo del tipo accettazione-rifiuto. In particolare, sia $q(\boldsymbol{\vartheta}, \boldsymbol{\vartheta}^*)$ la funzione di densità di probabilità di transizione da $\boldsymbol{\vartheta}$ a $\boldsymbol{\vartheta}^*$, dove $\boldsymbol{\vartheta}^*$ è un nuovo candidato, generato dalla distribuzione $q(\boldsymbol{\vartheta}, \boldsymbol{\vartheta}^*)$ che prende il nome di *proposal distribution*. La *proposal distribution* è una funzione che considera lo stato in cui si trova la catena $\boldsymbol{\vartheta}$ e propone un candidato per l'iterazione successiva: $\boldsymbol{\vartheta}^*$. Il candidato proposto ha probabilità di accettazione, $\alpha\left(\boldsymbol{\vartheta}, \boldsymbol{\vartheta}^*\right)$, definita da:

$$\alpha\left(\boldsymbol{\vartheta}, \boldsymbol{\vartheta}^*\right) = \min\left\{\frac{\pi(\boldsymbol{\vartheta}^* \mid \mathbf{X}) q(\boldsymbol{\vartheta}^*, \boldsymbol{\vartheta})}{\pi(\boldsymbol{\vartheta} \mid \mathbf{X}) q(\boldsymbol{\vartheta}, \boldsymbol{\vartheta}^*)}, 1\right\}.$$

Si osservi che nella probabilità di accettazione la distribuzione a posteriori compare nel rapporto $\pi(\boldsymbol{\vartheta}^* \mid \mathbf{X})/\pi(\boldsymbol{\vartheta} \mid \mathbf{X})$, dunque è sufficiente conoscere il nucleo della distribuzione a posteriori in quanto le costanti di normalizzazione $m(\mathbf{X})$, si semplificano. Si osservi inoltre che, se viene scelta una *proposal distribution* simmetrica, allora: $q(\boldsymbol{\vartheta}^*, \boldsymbol{\vartheta}) = q(\boldsymbol{\vartheta}, \boldsymbol{\vartheta}^*)$, donde:

$$\alpha\left(\boldsymbol{\vartheta}, \boldsymbol{\vartheta}^*\right) = \min\left\{\frac{\pi(\boldsymbol{\vartheta}^* \mid \mathbf{X}) q(\boldsymbol{\vartheta}^*, \boldsymbol{\vartheta})}{\pi(\boldsymbol{\vartheta} \mid \mathbf{X}) q(\boldsymbol{\vartheta}, \boldsymbol{\vartheta}^*)}, 1\right\} = \min\left\{\frac{\pi(\boldsymbol{\vartheta}^* \mid \mathbf{X})}{\pi(\boldsymbol{\vartheta} \mid \mathbf{X})}, 1\right\}. \tag{8}$$

La probabilità di accettazione si riduce al minimo tra il rapporto tra i nuclei e uno. Il minimo è necessario in quanto non si ha alcuna garanzia che il rapporto sia inferiore a uno, infatti se così non fosse allora $\alpha\left(\boldsymbol{\vartheta}, \boldsymbol{\vartheta}^*\right)$ non potrebbe più esser definita probabilità. Pertanto, se $\frac{\pi(\boldsymbol{\vartheta}^* \mid \mathbf{X}) q(\boldsymbol{\vartheta}^*, \boldsymbol{\vartheta})}{\pi(\boldsymbol{\vartheta} \mid \mathbf{X}) q(\boldsymbol{\vartheta}, \boldsymbol{\vartheta}^*)} > 1$ allora si pone $\alpha\left(\boldsymbol{\vartheta}, \boldsymbol{\vartheta}^*\right) = 1$.

In generale la procedura Metropolis-Hastings consta dei seguenti step:

1. Si inizializza la catena in corrispondenza di un punto $\boldsymbol{\vartheta}^0$.
2. Si genera $\boldsymbol{\vartheta}^*$ dalla *proposal distribution* $q\left(\boldsymbol{\vartheta}^0, \cdot\right)$.
3. Si genera $u \sim U(0,1)$.
4. Se $u < \alpha\left(\boldsymbol{\vartheta}, \boldsymbol{\vartheta}^*\right)$ allora si pone $\boldsymbol{\vartheta}^1 = \boldsymbol{\vartheta}^*$ (si accetta il candidato) altrimenti si pone $\boldsymbol{\vartheta}^1 = \boldsymbol{\vartheta}^0$ (si rifiuta il candidato).
5. Si reitera da 2 a 4 fino a che non si ritiene raggiunta la convergenza.

Si osservi che, non conoscendo la vera distribuzione a posteriori, quando si parla di convergenza, non si parla di convergenza in senso stretto. Non conoscendo la distribuzione a posteriori ciò che si può monitorare è la stabilizzazione della catena. Si parlerà quindi di un buon *mixing* della catena quando si otterrà la stabilizzazione dell'oscillazione intorno a un certo valore. Per monitorare il *mixing* vengono utilizzati i *trace plots* che mostrano le oscillazioni della catena per ciascuna iterazione. In Figura 6 viene riportato l'esempio di un buon *mixing* e di un *mixing* di una catena che mostra una forte componente di autocorrelazione.

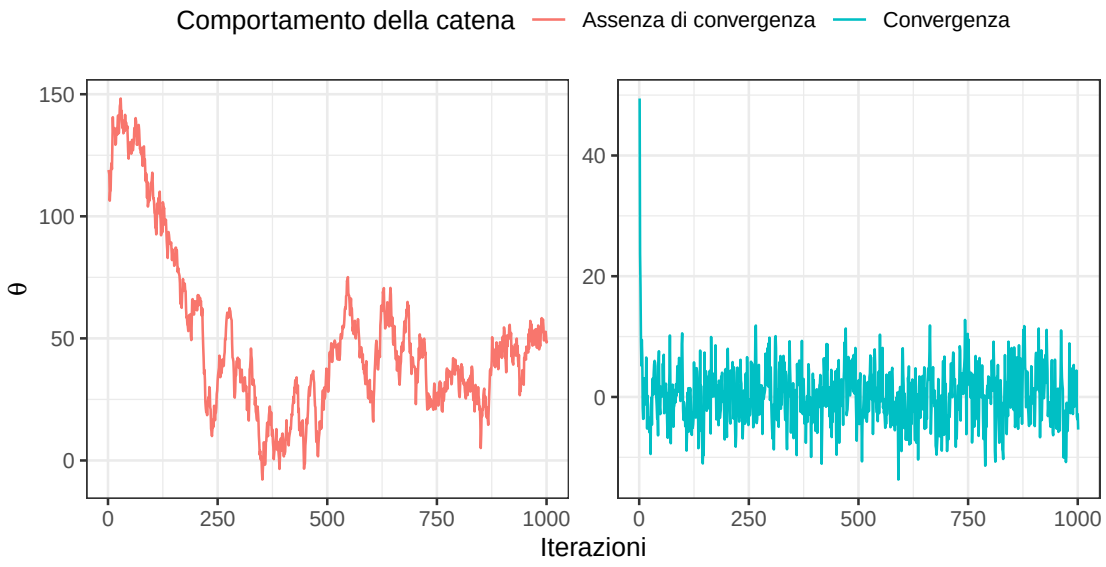


Figura 6: Trace Plot per il monitoraggio del mixing della catena: esempio di un buon mixing sulla destra e di un mixing che non è arrivato a convergenza dopo 1000 iterazioni sulla sinistra

Essendo un metodo basato sulle catene di Markov, anche il Metropolis-Hastings muove dall'idea che la catena sia caratterizzata da una distribuzione stazionaria. Sorge spontaneo chiedersi se il Metropolis-Hastings fornisca delle garanzie teoriche riguardo la stazionarietà. Ciò premesso risulta utile, al fine di comprendere se vi siano garanzie di stazionarietà, introdurre il concetto di reversibilità. Una catena di Markov è detta reversibile se la probabilità di transizione è invariante rispetto all'ordinamento temporale. In altri termini, la reversibilità implica che la probabilità di transizione della catena da uno stato a un altro non dipende dall'istante ma solo dallo stato in cui si trova la catena. Si può dimostrare che la reversibilità implica la stazionarietà, sicché se la catena è reversibile allora è anche stazionaria. Per enucleare in modo più rigoroso il legame tra stazionarietà e reversibilità si ricordi che il kernel di transizione del *Metropolis-Hastings*, come evidenziato da Carlin and Chib (1995), risulta:

$$p\left(\boldsymbol{\vartheta}, \boldsymbol{\vartheta}^*\right) = q\left(\boldsymbol{\vartheta}, \boldsymbol{\vartheta}^*\right) \alpha\left(\boldsymbol{\vartheta}, \boldsymbol{\vartheta}^*\right) + \delta_{\boldsymbol{\vartheta}}(\boldsymbol{\vartheta}^*) \int_{\mathbb{R}^p} q\left(\boldsymbol{z} \mid \boldsymbol{\vartheta}\right) \left[1 - \alpha\left(\boldsymbol{z} \mid \boldsymbol{\vartheta}\right)\right] \mathrm{d}\boldsymbol{z},$$

dove $\delta_{\boldsymbol{\vartheta}}(\boldsymbol{\vartheta}^*)$ è un punto con massa di probabilità in $\boldsymbol{\vartheta}$ e

$$\int_{\mathbb{R}^p} q\left(\boldsymbol{z} \mid \boldsymbol{\vartheta}\right) \left[1 - \alpha\left(\boldsymbol{z} \mid \boldsymbol{\vartheta}\right)\right] \mathrm{d}\boldsymbol{z}$$

rappresenta il rischio di rimanere in $\boldsymbol{\vartheta}$ e non transitare a $\boldsymbol{\vartheta}^*$.

Dato il kernel di transizione, è possibile dimostrare che se il kernel di transizione è reversibile allora la distribuzione limite della catena è stazionaria. Il problema si riduce quindi a dimostrare che il kernel di transizione è reversibile. A tal fine come dimostrato nel lavoro di Chib and Greenberg (1995), il kernel risulta

reversibile rispetto alla distribuzione a posteriori se:

$$\pi(\boldsymbol{\vartheta} \mid \mathbf{X}) q(\boldsymbol{\vartheta}, \boldsymbol{\vartheta}^*) = \pi(\boldsymbol{\vartheta}^* \mid \mathbf{X}) q(\boldsymbol{\vartheta}^*, \boldsymbol{\vartheta}).$$

Tale condizione può essere verificata agevolmente. Infatti:

1. se, $\pi(\boldsymbol{\vartheta} \mid \mathbf{X}) q(\boldsymbol{\vartheta}, \boldsymbol{\vartheta}^*) < \pi(\boldsymbol{\vartheta}^* \mid \mathbf{X}) q(\boldsymbol{\vartheta}^*, \boldsymbol{\vartheta})$ allora

$$\alpha(\boldsymbol{\vartheta}, \boldsymbol{\vartheta}^*) = 1 \quad \text{e} \quad \alpha(\boldsymbol{\vartheta}^*, \boldsymbol{\vartheta}) = \frac{\pi(\boldsymbol{\vartheta} \mid \mathbf{X}) q(\boldsymbol{\vartheta}, \boldsymbol{\vartheta}^*)}{\pi(\boldsymbol{\vartheta}^* \mid \mathbf{X}) q(\boldsymbol{\vartheta}^*, \boldsymbol{\vartheta})},$$

   sicché:

$$\begin{aligned}
\pi(\boldsymbol{\vartheta} \mid \mathbf{X}) q(\boldsymbol{\vartheta}, \boldsymbol{\vartheta}^*) \alpha(\boldsymbol{\vartheta}, \boldsymbol{\vartheta}^*) &= \pi(\boldsymbol{\vartheta} \mid \mathbf{X}) q(\boldsymbol{\vartheta}, \boldsymbol{\vartheta}^*), \\
\pi(\boldsymbol{\vartheta}^* \mid \mathbf{X}) q(\boldsymbol{\vartheta}^*, \boldsymbol{\vartheta}) \alpha(\boldsymbol{\vartheta}^*, \boldsymbol{\vartheta}) &= \pi(\boldsymbol{\vartheta}^* \mid \mathbf{X}) q(\boldsymbol{\vartheta}^*, \boldsymbol{\vartheta}) \frac{\pi(\boldsymbol{\vartheta} \mid \mathbf{X}) q(\boldsymbol{\vartheta}, \boldsymbol{\vartheta}^*)}{\pi(\boldsymbol{\vartheta}^* \mid \mathbf{X}) q(\boldsymbol{\vartheta}^*, \boldsymbol{\vartheta})} \\
&= \pi(\boldsymbol{\vartheta} \mid \mathbf{X}) q(\boldsymbol{\vartheta}, \boldsymbol{\vartheta}^*).
\end{aligned}$$

2. Se invece $\pi(\boldsymbol{\vartheta} \mid \mathbf{X}) q(\boldsymbol{\vartheta}, \boldsymbol{\vartheta}^*) > \pi(\boldsymbol{\vartheta}^* \mid \mathbf{X}) q(\boldsymbol{\vartheta}^*, \boldsymbol{\vartheta})$ allora

$$\alpha(\boldsymbol{\vartheta}^*, \boldsymbol{\vartheta}) = 1 \quad \text{e} \quad \alpha(\boldsymbol{\vartheta}, \boldsymbol{\vartheta}^*) = \frac{\pi(\boldsymbol{\vartheta}^* \mid \mathbf{X}) q(\boldsymbol{\vartheta}^*, \boldsymbol{\vartheta})}{\pi(\boldsymbol{\vartheta} \mid \mathbf{X}) q(\boldsymbol{\vartheta}, \boldsymbol{\vartheta}^*)},$$

   sicché

$$\begin{aligned}
\pi(\boldsymbol{\vartheta} \mid \mathbf{X}) q(\boldsymbol{\vartheta}, \boldsymbol{\vartheta}^*) \alpha(\boldsymbol{\vartheta}, \boldsymbol{\vartheta}^*) &= \pi(\boldsymbol{\vartheta} \mid \mathbf{X}) q(\boldsymbol{\vartheta}, \boldsymbol{\vartheta}^*) \frac{\pi(\boldsymbol{\vartheta}^* \mid \mathbf{X}) q(\boldsymbol{\vartheta}^*, \boldsymbol{\vartheta})}{\pi(\boldsymbol{\vartheta} \mid \mathbf{X}) q(\boldsymbol{\vartheta}, \boldsymbol{\vartheta}^*)} \\
&= \pi(\boldsymbol{\vartheta}^* \mid \mathbf{X}) q(\boldsymbol{\vartheta}^*, \boldsymbol{\vartheta}), \\
\pi(\boldsymbol{\vartheta}^* \mid \mathbf{X}) q(\boldsymbol{\vartheta}^*, \boldsymbol{\vartheta}) \alpha(\boldsymbol{\vartheta}^*, \boldsymbol{\vartheta}) &= \pi(\boldsymbol{\vartheta}^* \mid \mathbf{X}) q(\boldsymbol{\vartheta}^*, \boldsymbol{\vartheta}).
\end{aligned}$$

Segue che

$$\pi(\boldsymbol{\vartheta} \mid \mathbf{X})q(\boldsymbol{\vartheta}, \boldsymbol{\vartheta}^*) = \pi(\boldsymbol{\vartheta}^* \mid \mathbf{X})q(\boldsymbol{\vartheta}^*, \boldsymbol{\vartheta}).$$

La condizione per la reversibilità del kernel rispetto alla distribuzione a posteriori è soddisfatta, pertanto la catena converge a una distribuzione stazionaria. Si conclude che il Metropolis-Hastings assicura la convergenza della catena ad una distribuzione stazionaria.

Il problema che rimane da affrontare, e a cui non si può fornire una risposta univoca, riguarda la scelta della *proposal distribution*. Solitamente viene scelta sulla base del problema oggetto d'analisi, tuttavia ciò non implica che la scelta di una distribuzione come *proposal distribution*, o l'inizializzazione dei suoi parametri, sia di scarso rilievo, al contrario. Infatti, con una buona scelta della distribuzione e una buona inizializzazione dei parametri della distribuzione, altro punto rilevante e di non facile approccio, si può giungere a convergenza molto più velocemente rispetto a un'altra distribuzione con un'altra inizializzazione. Soprattutto quando il vettore delle v.a. da campionare consta di numerose componenti, una buona specificazione della *proposal distribution* può portare notevoli vantaggi in termini di efficienza, costo computazionale e velocità di raggiungimento della convergenza. Si osservi altresì che comunque inizializzata la distribuzione, il teorema ergodico garantisce, al divergere delle iterazioni, la convergenza. Il vero fulcro del problema è se l'onere computazionale per raggiungerla sia sostenibile.

Sebbene non sia possibile fornire indicazioni specifiche circa il miglior modo per scegliere la *proposal distribution* e inizializzarla in generale, è comunque possibile enucleare quali siano i due tipi di approcci prevalenti per l'utilizzo del Metropolis-Hastings:

1. La *proposal distribution* è visibile come una *Random Walk*: $\boldsymbol{\vartheta}^* = \boldsymbol{\vartheta} + \boldsymbol{\varepsilon}$ con $\boldsymbol{\varepsilon}$ parametro di disturbo che segue una distribuzione da specificare. In questo caso il candidato viene generato a partire dal valore assunto nello stato precedente dalla catena, perturbato da $\boldsymbol{\varepsilon}$.
2. La *proposal distribution* assume forma $q(\boldsymbol{\vartheta}, \boldsymbol{\vartheta}^*) = h(\boldsymbol{\vartheta}^*)$; la funzione $h(\boldsymbol{\vartheta}^*)$ non dipende dallo stato occupato dalla catena al passo precedente (per questo è anche denominato *Independent Metropolis-Hastings*). Si osservi che si ottiene comunque un campione Monte Carlo perché, se è vero che quando si genera il candidato lo si genera indipendentemente da $\boldsymbol{\vartheta}$, è altresì vero che è su di esso che viene basata la decisione di accettare o rifiutare il candidato.

Si osservi inoltre che un tasso di accettazione troppo basso è indice di inefficienza in quanto il *mixing* della catena risulta troppo lento, e ci vorrebbero innumerevoli iterazioni per arrivare a convergenza. Al contempo, un tasso di accettazione troppo alto è da evitare giacché significherebbe campionare solo dai valori ad alta densità della distribuzione a posteriori, e quindi le code sarebbero sottorappresentate.

Come primo approccio di campionamento si considererà proprio un *Random Walk* MH multivariato. In particolare, si considererà come *proposal distribution* una distribuzione Normale $p + 2$ variata dopo aver specificato in modo oculato il vettore delle medie, $\tilde{\boldsymbol{\vartheta}} = \boldsymbol{\mu}$, e la matrice di varianza-covarianza, $\boldsymbol{\Sigma}$, per la cui disamina si rimanda ai paragrafi 4.4.1 e 4.4.2. La *proposal* per generare il candidato, condizionatamente al valore di λ (parametro di precisione) risulta:

$$\boldsymbol{\vartheta}^* \mid \boldsymbol{\vartheta} \sim \mathcal{N}_{p+2}\left(\tilde{\boldsymbol{\vartheta}}, \frac{2.38^2}{p}\tilde{\boldsymbol{\Sigma}}\right).$$

Il fattore $\frac{2.38^2}{p+2}$ rende ottimo il tasso di accettazione nel caso ad elevata dimensionalità, per dettagli si rimanda al paragrafo 3.3.1.7 e al lavoro di Roberts and Rosenthal (2007). Inoltre, poiché la Normale è simmetrica, la probabilità di accettazione si riduce a

$$\alpha\left(\boldsymbol{\vartheta}, \boldsymbol{\vartheta}^*\right) = \min\left\{\frac{\pi(\boldsymbol{\vartheta}^* \mid \mathbf{X})}{\pi(\boldsymbol{\vartheta} \mid \mathbf{X})}, 1\right\}$$

L'algoritmo RW MH consta allora dei seguenti passi:

1. Si inizializza la catena in corrispondenza di un punto $\boldsymbol{\vartheta}^0$.
2. Si genera $\boldsymbol{\vartheta}^*$ dalla *proposal distribution* $q\left(\boldsymbol{\vartheta}^0, \cdot\right)$.
3. Si genera $u \sim U(0,1)$.
4. Se $u < \alpha\left(\boldsymbol{\vartheta}, \boldsymbol{\vartheta}^*\right)$ allora si pone $\boldsymbol{\vartheta}^1 = \boldsymbol{\vartheta}^*$ (si accetta il candidato), altrimenti si pone $\boldsymbol{\vartheta}^1 = \boldsymbol{\vartheta}^0$ (si rifiuta il candidato).
5. Si reitera da 2 a 4 fino a che non si ritiene raggiunta la convergenza.

3.3.1.6 *Thinning Period*

Sovente capita che anche a fronte di un elevato numero di iterazioni la catena possa non raggiungere la convergenza in un tempo ragionevole, a causa di una forte componente di autocorrelazione tra i candidati. In tal caso è possibile aumentare il numero di iterazioni, fino a che la catena non converga. Tuttavia tale approccio rischia di condurre a un onere computazionale considerevole e spesso non sostenibile.

Un'altra possibilità per ovviare a tale problematicità è cercare di decorrelare i valori che vengono accettati. L'idea è che, fisso restando il numero di iterazioni, se si considerano i valori da campionare ogni k valori generati, allora è più probabile che questi siano meno correlati rispetto a due candidati successivi. Si scartano quindi i valori di $\boldsymbol{\vartheta}$ generati tra un periodo e il successivo. Il periodo di scarto tra un $\boldsymbol{\vartheta}$ e il successivo è detto *thinning period*. In particolare, l'algoritmo diventa così composto dai seguenti passi:

1. Si inizializza la catena in corrispondenza di un punto $\boldsymbol{\vartheta}^0$.
2. Si genera $\boldsymbol{\vartheta}^*$ dalla *proposal distribution* $q\left(\boldsymbol{\vartheta}^0, \cdot\right)$.
3. Si genera $u \sim U(0,1)$.
4. Se $u < \alpha\left(\boldsymbol{\vartheta}, \boldsymbol{\vartheta}^*\right)$ allora si pone $\boldsymbol{\vartheta}^1 = \boldsymbol{\vartheta}^*$ (si accetta il candidato), altrimenti si pone $\boldsymbol{\vartheta}^1 = \boldsymbol{\vartheta}^0$ (si rifiuta il candidato).
5. Dopo il periodo di *burn-in* si salva $\boldsymbol{\vartheta}$ solo ogni k iterazioni, dove k è il *thinning period*.
6. Si reitera da 2 a 5 fino a che non si ritiene raggiunta la convergenza.

3.3.1.7 *Metropolis Adjusted Langevin Algorithm* (MALA)

Talvolta, soprattutto in contesti ad elevata dimensionalità, per quanto si cerchi di ottimizzare l'inizializzazione della *proposal distribution*, l'alto numero di variabili induce la *proposal* a generare candidati estremamente simili. Ciò è dovuto al fatto che all'aumentare del numero di variabili la matrice di varianza-covarianza della *proposal distribution* tende a presentare valori di covarianza elevata e valori di varianza bassi. L'effetto prodotto è un pessimo *mixing* della catena, che mostra un andamento caratterizzato da forte autocorrelazione. Di conseguenza servirebbe

un numero di iterazioni estremamente elevato, il che renderebbe il campionamento computazionalmente oneroso, per far convergere la catena e ottenere un buon *mixing*, si veda Beskos and Stuart (2009). In altri termini la *proposal* tende a generare candidati simili e dunque a produrre catene fortemente autocorrelate, poiché non riesce ad esplorare efficacemente altri punti dello spazio parametrico, in quanto non tiene conto della struttura della distribuzione a posteriori.

Per far sì che la catena converga in un numero ragionevole di iterazioni ed ottenere un buon *mixing* sono state sviluppate, recentemente, diverse tecniche che possono essere trovate in letteratura. Interessanti e precursori di molte altre varianti, a titolo di esempio, sono i lavori di Haario et al. (2001) e di Roberts and Rosenthal (2001), che pur basandosi sul Metropolis-Hastings e quindi sul principio dell'accettazione-rifiuto, fossero in grado di proporre candidati appartenenti allo spazio parametrico effettuando un'esplorazione più estensiva dello spazio stesso, incorporando l'informazione sulla forma della distribuzione a posteriori. Tali tecniche inducono la *proposal distribution* a generare candidati molto meno simili tra loro, esplorando in modo più efficiente lo spazio parametrico. Ciò si traduce in una decorrelazione della catena e porta al raggiungimento della convergenza a parità di numero di iterazioni. Quindi, invece che incrementare considerevolmente il numero di iterazioni e, conseguentemente, l'onere computazionale, il quale può risultare insostenibile, tali metodi agiscono sulla specificazione della *proposal distribution* inducendola a proporre candidati meno correlati tra loro.

In particolare, i metodi più recenti sfruttano la salita del gradiente, il quale incorpora l'informazione della struttura della distribuzione a posteriori per esplorare più efficacemente lo spazio parametrico. Si osservi che i metodi basati sulla salita stocastica del gradiente sono largamente utilizzati anche nell'ambito del

Machine Learning al fine di allenare efficacemente un modello particolarmente complesso (si pensi ad algoritmi di *Boosting*, quali, a titolo di esempio, XGBoost, LightGBM, CatBoost, per dettagli relativi a tali metodi si raccomanda un'attenta lettura dei lavori di Ferreira and Figueiredo (2012) e Mayr et al. (2014))

Fra i diversi algoritmi disponibili hanno trovato ampio utilizzo il Metropolis-Hastings basato sul metodo Hemiltoniano che, come suggerito dal nome, sfrutta proprio la dinamica hemiltoniana per esplorare efficacemente lo spazio parametrico, si veda Neal et al. (2011), e il *Metropolis Adjusted Langevin Algorithm* (MALA). Il funzionamento del MALA verrà brevemente enucleato in quanto algoritmo scelto per il campionamento nel caso oggetto d'analisi, per ulteriori dettagli si rimanda a Girolami and Calderhead (2011).

Si procederà, in particolare, illustrando il funzionamento del MALA e del MALA pre-condizionato, in quanto è stato l'algoritmo selezionato per migliorare il *mixing* della catena.

Il **MALA** sfrutta, in particolare, il gradiente della distribuzione log a posteriori,

$$\nabla_{\boldsymbol{\vartheta}} \log \pi(\boldsymbol{\vartheta} \mid \mathbf{X}) = \nabla_{\boldsymbol{\vartheta}} \log \pi(\mathbf{X} \mid \boldsymbol{\vartheta}) + \nabla_{\boldsymbol{\vartheta}} \log \pi(\boldsymbol{\vartheta})$$

che non richiede la conoscenza della costante di normalizzazione e che fornisce la direzione e il tasso di incremento della funzione di interesse.

Dunque, sia $\epsilon > 0$, "sufficientemente piccolo", sia $\boldsymbol{z}^{(j-1)} \sim \mathcal{N}_p(\mathbf{0}, \mathbf{I}_p)$, e detto $\boldsymbol{\vartheta}^{(j-1)}$ il vettore di parametri della catena al passo $j-1$, il valore della catena al passo successivo, j, sarà aggiornato come:

$$\boldsymbol{\vartheta}^{(j)} = \boldsymbol{\vartheta}^{(j-1)} + \frac{\epsilon^2}{2}\nabla_{\boldsymbol{\vartheta}} \log \pi(\boldsymbol{\vartheta}^{(j-1)} \mid \mathbf{X}) + \epsilon \boldsymbol{z}^{(j-1)},$$

che porta a un incremento di $\pi(\boldsymbol{\vartheta} \mid \mathbf{X})$ (da cui il nome salita del gradiente). Incorporando il gradiente all'interno della procedura MCMC si induce la *proposal distribution* a proporre candidati sempre più prossimi a valori per i quali la distribuzione a posteriori mostra densità più elevata.

L'intuizione appena fornita risulta logica, tuttavia la sua validità si poggia su un solido apparato teorico basato sulla diffusione di Langevin. Infatti, sia $\pi(\boldsymbol{\vartheta} \mid \mathbf{X})$ la funzione di densità della distribuzione invariante da cui si desidera campionare (si veda il teorema 3.1 per le assunzioni e le implicazioni).

L'equazione di Langevin fonda le sue radici nella fisica e, in prima istanza, descrive il moto browniano di una particella immersa in un fluido. Sia $\left(\vartheta^{(t)}\right)_{t\geq 0}$ un processo stocastico a tempo continuo; si utilizza tale notazione per facilitare il passaggio successivo al MALA, ($\boldsymbol{\vartheta}^{(t)}$ identifica la posizione della particella). Sia $P(\boldsymbol{\vartheta})$ l'energia potenziale della particella e sia $B^{(t)}$ un moto browniano p dimensionale (per approfondimenti sulla dinamica del moto browniano si veda Wang and Uhlenbeck (1945), i quali forniscono ampie fondamenta teoriche), l'equazione di Langevin può essere scritta come equazione differenziale stocastica, si veda Langevin (1908), nel modo seguente:

$$\mathrm{d}\boldsymbol{\vartheta}^{(t)} = \underbrace{-\nabla P(\boldsymbol{\vartheta})\mathrm{d}t}_{\text{drift}} + \underbrace{\sqrt{2}\mathrm{d}B^{(t)}}_{\text{diffusione}}, \tag{9}$$

dove il gradiente negativo dell'energia rappresenta, per la definizione fisica di lavoro, il vettore di tutte le forze di deriva agenti sulla particella.

Al fine di comprendere come evolve la funzione di densità di probabilità

di $\boldsymbol{\vartheta}^{(t)}$ nel tempo è sufficiente considerare l'equazione di Fokker-Plank associata all'equazione stocastica differenziale di Langevin, si veda (Ichimaru, 2018, p. 231). Si tratta di un'equazione alle derivate parziali che descrive proprio l'evoluzione della funzione di densità di probabilità nel tempo, sotto l'effetto di forze di deriva e forze di rumore:

$$\frac{\partial f(\boldsymbol{\vartheta}, t)}{\partial t} = \frac{\partial}{\partial \boldsymbol{\vartheta}} \left[\frac{\partial}{\partial \boldsymbol{\vartheta}} P(\boldsymbol{\vartheta}) f(\boldsymbol{\vartheta}, t) \right] + \frac{\partial f(\boldsymbol{\vartheta}, t)}{\partial \boldsymbol{\vartheta}^2}.$$

Quando $\frac{\partial f(\boldsymbol{\vartheta},t)}{\partial t}$ raggiunge il punto in cui la distribuzione a essa associata è invariante non potrà più deviare da tale punto (proprio perché la distribuzione è invariante). Sicché, quando $\frac{\partial f(\boldsymbol{\vartheta},t)}{\partial t}$ raggiunge il punto di invarianza allora $\frac{\partial f(\boldsymbol{\vartheta},t)}{\partial t} = 0$, poiché non subisce più variazioni nel tempo. Quindi, dal momento che $\pi(\boldsymbol{\vartheta} \mid \mathbf{X})$ è stata posta come funzione di densità della distribuzione invariante, deve risultare che:

$$\frac{\partial f(\boldsymbol{\vartheta}, t)}{\partial t} = \frac{\partial}{\partial \boldsymbol{\vartheta}} \left[\frac{\partial}{\partial \boldsymbol{\vartheta}} P(\boldsymbol{\vartheta}) \pi(\boldsymbol{\vartheta} \mid \mathbf{X}) + \frac{\partial \pi(\boldsymbol{\vartheta} \mid \mathbf{X})}{\partial \boldsymbol{\vartheta}} \right] = 0,$$

che è soddisfatta se e solo se:

$$\frac{\partial}{\partial \boldsymbol{\vartheta}} P(\boldsymbol{\vartheta}) \pi(\boldsymbol{\vartheta} \mid \mathbf{X}) + \frac{\partial \pi(\boldsymbol{\vartheta} \mid \mathbf{X})}{\partial \boldsymbol{\vartheta}} = 0,$$

che è possibile dimostrare, per dettagli si veda (Landau and Lifshitz, 2013, par. 28, Part. I), avere come soluzione:

$$\pi(\boldsymbol{\vartheta} \mid \mathbf{X}) \propto \exp\left[-P(\boldsymbol{\vartheta})\right],$$

che rappresenta il nucleo di una distribuzione di Boltzman. Il che implica che è

possibile campionare dai cosiddetti *Energy-Based Models* della forma:

$$\pi(\boldsymbol{\vartheta} \mid \mathbf{X}) = \frac{e^{-\mathcal{E}(\boldsymbol{\vartheta})}}{Z},$$

dove $\mathcal{E}(\boldsymbol{\vartheta}) = P(\boldsymbol{\vartheta})$, ovvero si considera l'energia totale del sistema, e non già solo quella potenziale, che assegna a tutti i suoi possibili stati, $\boldsymbol{\vartheta}^{(1)}, \ldots, \boldsymbol{\vartheta}^{(n)}$, i rispettivi valori di energia. Z è la costante di normalizzazione che può essere ignorata dal momento che l'equazione differenziale stocastica di Langevin (9) richiede la conoscenza del solo gradiente di $\pi(\boldsymbol{\vartheta} \mid \mathbf{X})$.

Ricordando che $\pi(\boldsymbol{\vartheta} \mid \mathbf{X})$ può essere riscritta come $\exp[\log \pi(\boldsymbol{\vartheta} \mid \mathbf{X})]$, è lecito porre $P(\boldsymbol{\vartheta}) = \mathcal{E}(\boldsymbol{\vartheta}) = -\log \pi(\boldsymbol{\vartheta} \mid \mathbf{X})$.

L'equazione differenziale stocastica di Langevin (9) assume allora la seguente forma:

$$\mathrm{d}\boldsymbol{\vartheta}^{(t)} = \frac{1}{2} \nabla \log \pi(\boldsymbol{\vartheta} \mid \mathbf{X}) \mathrm{d}t + \mathrm{d}B^{(t)},$$

e dunque la distribuzione stazionaria dell'equazione di diffusione di Langevin è proprio la distribuzione a posteriori $\pi(\boldsymbol{\vartheta} \mid \mathbf{X})$.

Risulta necessario, a fini implementativi, disporre di una versione discreta della diffusione di Langevin, che è possibile ricavare tramite il metodo di Eulero-Maruyama, si veda Kloeden et al. (1992). La versione a tempo discreto che ne discende ha la forma dell'equazione (10). Si osservi che l'approssimazione discreta non assicura più la convergenza a $\pi(\boldsymbol{\vartheta} \mid \mathbf{X})$, dal momento che esiste un *trade-off* tra l'accuratezza dell'approssimazione, che cresce tanto più $\epsilon \to 0$, e l'efficienza, che cresce al crescere di ϵ. Tale criticità viene risolta trattando la distribuzione

$$\boldsymbol{\vartheta}^{(j)} = \boldsymbol{\vartheta}^{(j-1)} + \frac{\epsilon^2}{2}\nabla_{\boldsymbol{\vartheta}} \log \pi(\boldsymbol{\vartheta}^{(j-1)} \mid \mathbf{X}) + \epsilon \boldsymbol{z}^{(j-1)} \tag{10}$$

come *proposal density* di un algoritmo Metropolis-Hastings.

il MALA può quindi essere visto come uno specifico algoritmo MH con *proposal distribution* della forma:

$$\boldsymbol{\vartheta}^* \mid \boldsymbol{\vartheta} \sim \mathcal{N}_p \left(\tilde{\boldsymbol{\vartheta}} + \frac{\epsilon^2}{2p^{\frac{1}{3}}} \nabla_{\boldsymbol{\vartheta}} \log \pi(\boldsymbol{\vartheta} \mid \mathbf{X}) \ , \ \frac{\epsilon^2}{p^{\frac{1}{3}}} \mathbf{I}_p \right),$$

con $\epsilon > 0$ parametro di *tuning*, che va accuratamente selezionato per ottenere la convergenza in modo efficiente. Si noti che rispetto al RW MH, in questo caso la *proposal distribution* non è simmetrica, pertanto la probabilità di accettazione assume la canonica forma:

$$\alpha\left(\boldsymbol{\vartheta}, \boldsymbol{\vartheta}^*\right) = \min \left\{ \frac{\pi(\boldsymbol{\vartheta}^* \mid \mathbf{X}) q(\boldsymbol{\vartheta}^*, \boldsymbol{\vartheta})}{\pi(\boldsymbol{\vartheta} \mid \mathbf{X}) q(\boldsymbol{\vartheta}, \boldsymbol{\vartheta}^*)}, 1 \right\}.$$

Si pone dunque come problema quello di capire quale sia il valore ottimo di ϵ, che è possibile risolvere guardando al comportamento asintotico del MALA. Il teorema di Roberts and Rosenthal (2007) garantisce che: definito un processo stocastico a tempo continuo $(Z_t)_{t \geq 0}$ e parametrizzato in modo da effettuare transizioni ogni $p^{-1/3}$ unità, esso converge in distribuzione a W, dove W è un processo diffuso che soddisfa l'equazione differenziale stocastica:

$$\mathrm{d}W^{(t)} = \sqrt{h(\epsilon)}\mathrm{d}B^{(t)} + \frac{h(\epsilon)\nabla \log f(W^{(t)})}{2}\mathrm{d}t,$$

dove $h(\epsilon)$ è la velocità di diffusione ed assume la seguente forma:

$$h(\epsilon) = \epsilon^2 2\Phi(-\mathcal{J}\epsilon^3)$$

per qualche costante $\mathcal{J}$ che dipende solo da f.

La funzione $h(\epsilon)$ è strettamente legata alla varianza asintotica e proprio per questo si cerca il valore di ϵ che massimizza $h(\epsilon)$. Inoltre, se nel caso descritto al paragrafo 3.3.1.5 l'algoritmo presentava una complessità nell'ordine di un $\mathcal{O}(p)$ il MALA presenta un grado di complessità nell'ordine di un $\mathcal{O}(p^{1/3})$, il che lo rende teoricamente più accurato e dunque più efficiente, soprattutto in problemi ad elevata dimensionalità.

Al fine di selezionare l'ϵ più appropriato giova notare che il tasso di accettazione del MH $A_p(\epsilon)$ è dato da:

$$A_p(\epsilon) = \lim_{R\to\infty} \frac{\sum_{r=1}^{R} \mathbb{I}_{\alpha_r > u_r}(\boldsymbol{\vartheta}^*_{\boldsymbol{r}})}{R},$$

con

$$\mathbb{I}_{\alpha_r > u_r}(\boldsymbol{\vartheta_r}) = \begin{cases} 1 & \text{se } \alpha_r\left(\boldsymbol{\vartheta}, \boldsymbol{\vartheta}^*\right) > u_r \\ \\ 0 & \text{se } \alpha_r\left(\boldsymbol{\vartheta}, \boldsymbol{\vartheta}^*\right) < u_r \end{cases},$$

con $u_r \sim U(0,1)$. Ma allora:

$$\lim_{p\to\infty} A_p(\epsilon) = A(\epsilon) = 2\Phi(-\mathcal{J}\epsilon^3)$$

e quindi il tasso di accettazione è legato al parametro di *tuning*. Donde, $h(\epsilon)$ può essere calcolata senza conoscere la costante $\mathcal{J}$. Infatti l'ϵ_{opt} può essere calcolato o tramite strategia di *trial and errors*, oppure tramite metodi adattivi in modo tale

che $A(\epsilon_{\text{opt}}) \approx 0.574$ che è stato dimostrato essere il valore ottimo per il MALA, per dettagli si faccia sempre riferimento a Roberts and Rosenthal (2007).

In realtà l'algoritmo sin qui presentato assume che la distribuzione a posteriori sia a componenti iid. Ciò non è detto che si verifichi, pertanto, per quanto si possa raffinare il valore di ϵ il *mixing* potrebbe sempre presentare assenza di convergenza in un numero ragionevole di iterazioni. Per risolvere il problema è sufficiente apportare una piccola modifica all'algoritmo: infatti, se si considera nota la matrice di varianza-covarianza a posteriori è possibile riparametrizzare $\boldsymbol{\vartheta}$ in modo tale che gli elementi di $\boldsymbol{\Sigma}$ siano ortogonali. Se si applica il MALA sulla riparametrizzazione ortogonale, il problema della mancata verifica dell'assunzione di iid non può più sussistere, vista l'ortogonalità. In questo caso applicare la rotazione tramite la decomposizione della matrice $\boldsymbol{\Sigma}$ è una semplice operazione algebrica che però rende il MALA efficace, oltre che efficiente.

Visto quanto appena esposto è possibile apportare una piccola modifica alla *proposal distribution* del MALA per poter utilizzare il MALA, nella sua formulazione originaria, sulla parametrizzazione ortogonale. Si denominerà, pertanto, MALA pre-condizionato l'algoritmo avente come *proposal distribution*

$$\boldsymbol{\vartheta}^* \mid \boldsymbol{\vartheta} \sim \mathcal{N}_p \left(\tilde{\boldsymbol{\vartheta}} + \frac{\epsilon^2}{2p^{\frac{1}{3}}} \boldsymbol{\Sigma} \nabla_{\boldsymbol{\vartheta}} \log \pi(\boldsymbol{\vartheta} \mid \mathbf{X}) \ , \ \frac{\epsilon^2}{p^{\frac{1}{3}}} \boldsymbol{\Sigma} \right).$$

La matrice di varianza-covarianza a posteriori è non nota ma può essere approssimata con diverse tecniche.

3.4 Approccio classico e bayesiano ai dati ad elevata dimensionalità

In ambito frequentista le tecniche per gestire dati ad elevata dimensionalità sono innumerevoli. In questo paragrafo verranno enucleati i metodi di natura modellistica che trovano un parallelismo con i metodi bayesiani per gestire dati ad elevata dimensionalità. Non si farà riferimento, pertanto, ad algoritmi propri del *Machine Learning* o del *Deep Learning*, che ben gestiscono i dati ad elevata dimensionalità ma a un costo computazionale non indifferente e, soprattutto, a costo della perdita di interpretabilità. Per costruzione tali algoritmi costituiscono delle *black-box* che si rivelano efficaci a scopo previsivo ma, a causa della loro complessità, risulta quasi sempre impossibile fornire delle interpretazioni. Inoltre, in quanto algoritmi di apprendimento, soffrono di mancanza di garanzie inferenziali. A tal proposito si veda Efron and Hastie (2021) che riportano un'interessante discussione sull'importanza delle variabili relativamente alle Foreste Casuali, ma estendibile a tutti gli algoritmi che procedono attribuendo un livello di importanza alle variabili stesse. Mancando di garanzia inferenziali, gli algoritmi di *Machine Learning* dovrebbero essere evitati se l'obiettivo è la selezione di variabili, e gli autori enucleano in modo eccellente la motivazione, con un esempio tanto semplice quanto efficace da un punto di vista della comprensione.

Si discuteranno ora due metodi molto utilizzati in ambito frequentista che si basano sull'idea di un'ottimizzazione vincolata:

1. Regressione Ridge.
2. Regressione LASSO (*Least Absolute Shrinkage and Selection Operator*).

Queste tecniche di regressione si sfruttano quando il numero di variabili è di molto superiore al numero di osservazioni. Gli usuali modelli di regressione vengono costruiti sotto l'ipotesi che $p << n$, pena la sovrastima. In altri termini, più il numero di variabili si avvicina al numero di osservazioni disponibili più il rischio di interpolare i dati, piuttosto che estrapolare informazione da essi, aumenta.

Se $p >> n$ qualsiasi modello di regressione non riuscirebbe ad essere stimato, dal momento che si hanno a disposizione più variabili che osservazioni. In questi casi, per rendere possibile la costruzione di un modello di regressioni si *"comprimono"* i parametri, ovvero si penalizzano i parametri in modo tale da riuscire comunque a massimizzare la funzione di verosimiglianza. Da un punto di vista interpretativo tale *shrinkage* rende le stime dei coefficienti distorte, per questo di solito non è possibile accompagnare le stime dei coefficienti a dei test che ne determinino la significatività. Infatti la distorsione fa cadere tutte le ipotesi che vigono nel campo degli stimatori non distorti. Al contempo, tuttavia, la distorsione dei coefficienti rende possibile la costruzione di un modello che altrimenti sarebbe impossibile da specificare e che, accettando un *bias* più elevato a livello di stima dei coefficienti, produce stime meno variabili. Il *bias* è imputabile al fatto che non si è più nella classe degli stimatori non distorti. Si può anche dimostrare, si veda per dettagli il Teorema di Theobald (1974), che esiste almeno un valore del parametro di *shrinkage*, λ, tale per cui l'errore quadratico medio della regressione Ridge è inferiore a quello della regressione standard.

Da un punto di vista operativo la penalizzazione può essere necessaria anche se il numero di variabili non supera di molto il numero di osservazioni. Un criterio generale per stabilire se occorre introdurre un fattore di penalizzazione è

considerare il reciproco del *condition number*: $K = \frac{\delta_{max}}{\delta_{min}}$, ovvero il rapporto tra il più grande e il più piccolo valore singolare della matrice $\mathbf{X}'\mathbf{X}$, per i dettagli riguardanti la decomposizione in valori singolari si rimanda al paragrafo 7.1. Nel caso in cui $1/K$ risulti più basso della precisione della "macchina" è necessario introdurre un fattore di penalizzazione per poter procedere alla regressione. Si noti che se $p >> n$ non vi è necessità di calcolare il *condition number*, in quanto necessariamente si dovrà procedere con tecniche di penalizzazione. Si osservi che, mentre il termine LASSO è definito come un acronimo, il termine Ridge indica proprio il fatto che, in contesti ad alta dimensionalità, per rendere la matrice del disegno invertibile, si aggiunge una "cresta" sulla diagonale principale della matrice, per tutti i dettagli sulla regressione Ridge si veda McDonald (2009).

Le due regressioni poc'anzi citate differiscono per il modo in cui penalizzano i coefficienti. Al fine di comprendere le diverse penalizzazioni si introduce il concetto di norma. In generale la norma L_q di un vettore $\boldsymbol{x}$ è denotata come:

$$L_q = \|\boldsymbol{x}\|_q = \left(\sum_{j=1}^{n} \mid x_j \mid^q \right)^{\frac{1}{q}}.$$

La differenza tra la regressione Ridge e la regressione LASSO risiede proprio nella penalizzazione adottata: la Ridge adotta come penalizzazione la norma $\mathrm{L}_2 = \|\boldsymbol{x}\|_2^2 = \sum_{j=1}^{n} x_j^2$, la regressione LASSO, per i cui dettagli si rimanda al lavoro di Tibshirani (1996), adotta come penalizzazione la norma $\mathrm{L}_1 = \|\boldsymbol{x}\|_1 = \sum_{j=1}^{n} \mid x_j \mid$. La differenza tra le due penalizzazioni risiede nel modo in cui comprimono i coefficienti: la norma L_1 può permettere di azzerare determinati coefficienti al fine di massimizzare la verosimiglianza ed effettua uno *shrinkage* sui coefficienti molto elevato. In virtù di questa proprietà discende l'aggettivo di *Selection Operator*,

perché potendo azzerare i coefficienti, di fatto, sta eseguendo una selezione di variabili. La regressione Ridge agisce in modo meno marcato: la norma L_2 non consente di azzerare i coefficienti, se non nel caso limite in cui il parametro di *shrinkage* tenda a infinito, ma può condurli molto vicini allo zero, di fatto annullandone quasi totalmente l'effetto da un punto di vista delle previsioni del modello.

Quale tipo di penalizzazione scegliere dipende dal problema oggetto d'analisi ma anche dai dati stessi: si ricordi che l'operatore LASSO non può selezionare più variabili che osservazioni, la penalizzazione Ridge non è affetta da questo problema.

Inoltre, se è pur vero che l'operatore LASSO può, potenzialmente, effettuare una selezione delle variabili, è altresì vero che non si può utilizzare la regressione LASSO per selezionare le variabili sugli stessi dati su cui si andrà ad effettuare l'inferenza, in quanto, da un punto di vista frequentista, il LASSO non fornisce alcuna garanzia inferenziale circa la selezione delle variabili effettuata. Per avere garanzie inferenziali sulla selezione di variabili effettuata dal LASSO bisognerebbe applicare metodi quali, a titolo di esempio: la *stability selection*, si veda Meinshausen and Bühlmann (2010), o la tecnica del *data splitting*, per dettagli si veda Thall et al. (1997). Tali metodiche forniscono garanzie inferenziali applicate con il LASSO, ma il LASSO da solo non ne fornisce alcuna.

Chiariti i concetti che stanno alla base delle due principali regressioni penalizzate conosciute ed usate, si declineranno ora, formalmente, dal punto di vista frequentista e dal punto di vista bayesiano.

Dal punto di vista frequentista, definita $\mathscr{L}(\mathbf{X};\boldsymbol{\vartheta})$ la funzione di verosimiglianza di un generico modello, indicando con $l(\mathbf{X};\boldsymbol{\vartheta})$ la corrispondente funzione di log

verosimiglianza e denotato con λ il parametro di *shrinkage*, ovverosia il peso da dare alla penalizzazione (tanto più alto è λ tanto maggiore sarà la penalizzazione), la regressione Ridge e la regressione LASSO possono essere viste, semplicemente, come dei problemi di ottimizzazione. Più precisamente possono essere visti come problemi di ottimizzazione convessa (strettamente convessa la Ridge, in quanto $q \geq 1$ e convessa, non strettamente, la regressione LASSO, giacché la sparsità richiede $q \leq 1$ e la convessità $q \geq 1$ e la LASSO è caratterizzata da $q = 1$). In particolare:

$$\text{Regressione Ridge:} \qquad \hat{\boldsymbol{\vartheta}}_{Ridge} = \arg\max_{\boldsymbol{\vartheta}} \quad l(\mathbf{X}; \boldsymbol{\vartheta}) - \lambda \sum_{j=1}^{p} \vartheta_j^2$$

$$\text{Regressione LASSO:} \qquad \hat{\boldsymbol{\vartheta}}_{LASSO} = \arg\max_{\boldsymbol{\vartheta}} \quad l(\mathbf{X}; \boldsymbol{\vartheta}) - \lambda \sum_{j=1}^{p} \mid \vartheta_j \mid$$

In entrambi i casi, tipicamente, il λ ottimo viene scelto tramite *cross-validation* dopo aver specificato un grigliato di valori di λ. Si osservi inoltre che, generalmente, $\mathbf{X}$ rappresenta la matrice dei dati standardizzati. La standardizzazione è di considerevole importanza dal momento che, in assenza di essa, la soluzione sarebbe influenzata dall'unità di misura delle diverse variabili.

Ferma restando l'ipotesi di standardizzazione della matrice dei dati, e ferme restando le considerazioni fatte poc'anzi, ciò che cambia in ambito bayesiano è il modo con cui si approcciano i problemi tali per cui $p >> n$. In particolare, considerata la funzione di verosimiglianza di un generico modello, $\mathscr{L}(\mathbf{X} \mid \boldsymbol{\vartheta})$, se si vuole introdurre una penalizzazione sui coefficienti, essendo quest'ultimi in ambito bayesiano delle variabili aleatorie, si eliciterà una distribuzione a priori che effettui

lo *shrinkage* sui coefficienti.

Non è difficile dimostrare, per ulteriori dettagli si veda Hsiang (1975), che se si vuole ottenere una penalizzazione di tipo Ridge allora bisogna specificare una distribuzione a priori del tipo:

$$\boldsymbol{\vartheta} \sim \mathcal{N}_p\left(\mathbf{0}, \frac{\sigma^2}{2\lambda}\mathbf{I}_p\right) \sim \mathcal{N}_p\left(\mathbf{0}, \frac{1}{2\lambda}\mathbf{I}_p\right) \iff \pi(\boldsymbol{\vartheta}) = \sqrt{\frac{\lambda^p}{\pi^p}}\ \exp\left(-\lambda \sum_{j=1}^{p} \beta_j^2\right) \tag{11}$$

(si ricordi che per dati standardizzati $\sigma^2 = 1$), da cui discende che la distribuzione a posteriori risulta:

$$\pi(\boldsymbol{\vartheta} \mid \mathbf{X}) \propto \mathscr{L}(\mathbf{X} \mid \boldsymbol{\vartheta})\pi(\boldsymbol{\vartheta}) \propto \mathscr{L}(\mathbf{X} \mid \boldsymbol{\vartheta}) \exp\left(-\lambda \sum_{j=1}^{p} \beta_j^2\right)$$

e quindi la distribuzione log a posteriori è:

$$\log \pi(\boldsymbol{\vartheta} \mid \mathbf{X}) = l(\mathbf{X} \mid \boldsymbol{\vartheta}) - \lambda \sum_{j=1}^{p} \vartheta_j^2 + c,$$

con c costante additiva che racchiude i termini non caratterizzanti il nucleo della distribuzione log a posteriori. Si osservi che i due problemi risultano equivalenti, semplicemente in ambito bayesiano si considera il vettore di parametri aleatorio e si elicita una distribuzione a priori per esso in modo da caratterizzarne il comportamento. Il vantaggio del metodo bayesiano risiede nella sua semplicità di interpretazione e nella sua flessibilità. La distribuzione a priori specificata in (11) risulta sensata: viene specificata come distribuzione sui parametri oggetto di inferenza una normale centrata in zero; si parte dunque da una condizione in cui tutti i parametri risultano nulli. Se si elicita un λ elevato si garantisce

alla distribuzione una varianza molto ridotta, il che porterà a dei coefficienti che rimarranno molto vicini a zero. Contrariamente, se si specifica un valore di λ basso, si lascerà che la distribuzione a priori abbia una varianza elevata e quindi, con molta facilità, i coefficienti si distanzieranno dal punto di partenza (0). Si evince come sia molto più semplice e lineare il ragionamento in ambito bayesiano rispetto al ragionamento in termini di ottimizzazione convessa che avviene in ambito frequentista.

In ambito bayesiano non è insolito elicitare λ con una *hyperprior*, che solitamente e convenzionalmente viene posta come una $\mathcal{C}^+(0,1)$ (cioè una *half-Cauchy*). Non è tuttavia insolito procedere creando una sequenza di λ e, successivamente al campionamento di $\boldsymbol{\vartheta}$, procedere con strategie che verranno illustrate nel capitolo 4.

Similmente alla regressione Ridge, per la penalizzazione LASSO in ambito bayesiano si procede elicitando la distribuzione a priori per $\boldsymbol{\vartheta}$ con una distribuzione *double exponential* centrata in 0 e con parametro di scala τ, del tipo:

$$\pi(\boldsymbol{\vartheta}) = \exp\left(-\frac{1}{2\tau}\sum_{j=1}^{p} \mid \vartheta_j \mid\right),$$

da cui, ponendo $\lambda = \frac{\sigma^2}{\tau}$ e considerando la distribuzione log a posteriori si ottiene:

$$\log \pi(\boldsymbol{\vartheta} \mid \mathbf{X}) = l(\mathbf{X} \mid \boldsymbol{\vartheta}) - \lambda \sum_{j=1}^{p} \mid \vartheta_j \mid + c.$$

Si osservi che λ continua ad avere un'interpretazione analoga al caso della penalizzazione Ridge. Per ulteriori dettagli sulla penalizzazione LASSO da un punto di vista bayesiano si veda Park and Casella (2008).

La logica che si cela dietro l'approccio bayesiano come si è potuto evincere nel corso di questo paragrafo è pulita ed elegante oltre che di immediata interpretazione.

4 Metodi

4.1 Scelta del modello di sopravvivenza

Nell'ambito dell'analisi di sopravvivenza sono diversi i modelli che è possibile considerare per modellare dati del tipo tempo all'evento. Essi differiscono, prevalentemente, per come viene impostata la modellizzazione della funzione di rischio. Uno stimatore che ha trovato, e trova tutt'ora, ampio utilizzo in questo contesto è il modello di Cox (1972) altresì conosciuto con il nome di modello a rischi proporzionali, si veda per una trattazione esaustiva Breslow (1975). La popolarità dello stimatore di Cox risiede nella sua natura semiparametrica: infatti, l'unica assunzione che prevede consiste nella proporzionalità tra il rischio e l'effetto delle variabili nel tempo, oltre la linearità del previsore, l'indipendenza tra le coppie (t_i, δ_i). In particolare, non viene imposta alcuna distribuzione sulla funzione di rischio, da qui la natura semiparametrica dello stimatore. Al fine di comprendere

le ragioni che rendono tale stimatore uno tra i più usati nel contesto dell'analisi di sopravvivenza, si consideri la forma che esso assume: chiamato $h_0(t)$ il rischio *baseline*, ed indicando con β il coefficiente di regressione e con x la variabile esplicativa (si prende in considerazione al momento il caso univariato per semplificare l'idea interpretativa), lo stimatore è definito come:

$$h(t) = h_0(t) \exp(x\beta)$$

e per l'assunzione di proporzionalità risulta:

$$\frac{h_{X=x+1}(t)}{h_{X=x}(t)} = \frac{h_0(t) \exp[(x+1)\beta]}{h_0(t) \exp(x\beta)} = \exp(\beta), \tag{12}$$

da cui si evince chiaramente come lo stimatore richieda solo la presenza di proporzionalità tra l'effetto delle covariate e il rischio nel tempo. Inoltre, dalla relazione (12) si ha un'immediata interpretazione dei coefficienti stimati. Infatti, esponenziando i coefficienti, si ottengono gli *Hazard Ratio*, utili a stabilire se un fattore risulti protettivo o di rischio. Gli *Hazard Ratio* quantificano inoltre l'impatto di una variazione della variabile di interesse sul rischio e, conseguentemente, sulla probabilità di sopravvivenza nel tempo. Infine, per comprendere il motivo della popolarità di cui tutt'oggi gode tale stimatore è sufficiente osservare il modo in cui vengono stimati i coefficienti. Essendo un modello semiparametrico la stima dei coefficienti avviene massimizzando non già la funzione di verosimiglianza ma la funzione di verosimiglianza parziale. Formalmente, siano $t_{(j)}$, con $j = 1, ..., T$ i tempi d'evento ordinati e sia R_j il numero di pazienti a rischio a tempo j. La verosimiglianza parziale può essere interpretata come la probabilità che un

individuo con profilo di covariate $\boldsymbol{x}_j$ esperisca l'evento a tempo t_j dato che il numero di pazienti a rischio a tempo t_j è pari a R_j e che si è già verificato un evento a tempo t_j:

$$\mathscr{PL}(\boldsymbol{\beta} \mid \mathbf{X}) = \prod_{j=1}^{J} \frac{\exp(\boldsymbol{x}_j'\boldsymbol{\beta})}{\sum_{i \in R_j} \exp(\boldsymbol{x}_i'\boldsymbol{\beta})},$$

da cui si può osservare come la funzione di verosimiglianza parziale dipenda solo dai coefficienti e non già dal rischio.

Non richiedendo una stima diretta della funzione di rischio e, pertanto, la specificazione di una distribuzione su di esso, lo stimatore di Cox risulta uno strumento flessibile, con poche assunzioni da verificare, e largamente impiegato nell'analisi di sopravvivenza. Tuttavia, se ciò lo rende uno stimatore di facile impiego in ambito frequentista, lo rende al contempo, per come è costruito, inutilizzabile in ambito bayesiano, a meno di specificare una distribuzione per la funzione di rischio medesima. Si osservi tuttavia, che specificare una distribuzione sulla funzione di rischio renderebbe poco sensato l'utilizzo del modello di Cox dal momento che il suo punto di forza risiede proprio nell'assenza di necessità di specificare una distribuzione sul rischio. Utilizzarlo in ambito bayesiano farebbe perdere il vantaggio che conserva in ambito frequentista. Inoltre, nel contesto dell'analisi di sopravvivenza, vi sono numerosi modelli di natura prettamente parametrica che meglio si prestano all'utilizzo in un contesto bayesiano. Il modello basato sulla distribuzione di Weibull risulta un'alternativa parametrica flessibile e robusta allo stimatore semiparametrico di Cox.

4.1.1 Modello di Weibull

Il modello di sopravvivenza di Weibull deve il nome all'omonima distribuzione. In particolare, si denoti con T il tempo all'evento, se $T \sim \text{Weibull}(\alpha, \epsilon)$ allora la funzione di densità, per $t \geq 0$, risulta:

$$f\left(t, \alpha, \epsilon\right) = \frac{\alpha}{\epsilon}\left(\frac{t}{\epsilon}\right)^{\alpha-1} \exp\left[-\left(\frac{t}{\epsilon}\right)^{\alpha}\right], \tag{13}$$

con α parametro di forma ed ϵ parametro di scala.

Sebbene la (13) sia la parametrizzazione più conosciuta per la distribuzione di Weibull, a fini modellistici si è preferito adottare un'altra parametrizzazione. In particolare, ponendo $\gamma = \epsilon^{-\alpha}$ si ha, per $t \geq 0$:

$$f\left(t, \alpha, \gamma\right) = \alpha\gamma t^{\alpha-1} \exp\left(-t^{\alpha}\gamma\right).$$

La funzione di sopravvivenza risulta quindi:

$$\begin{aligned} S(t, \alpha, \gamma) = \mathbb{P}\left(T > t\right) &= \int_{t}^{+\infty} \alpha\gamma z^{\alpha-1} \exp\left(-z^{\alpha}\gamma\right) \mathrm{d}z \\ &= -\exp\left(-\gamma z^{\alpha}\right)\Big|_{t}^{+\infty} \\ &= -\lim_{z\to+\infty} -\exp\left(-\gamma z^{\alpha}\right) + \exp\left(-\gamma t^{\alpha}\right) \\ &= \exp\left(-\gamma t^{\alpha}\right), \end{aligned}$$

e per la (3) si ha:

$$h(t, \alpha, \gamma) = \frac{f(t, \alpha, \gamma)}{S(t, \alpha, \gamma)} = \alpha\gamma t^{\alpha-1}.$$

Si consideri c l'evento indicante la censura, e sia

$$\delta_i = \begin{cases} 1 & \text{se} \quad t_i \leq c \\ \\ 0 & \text{se} \quad t_i > c \end{cases} .$$

δ_i è un indicatore che è pari a uno se al tempo t_i il paziente non ha esperito la censura, zero altrimenti. Il modello di base e il modello indotto possono quindi essere formalizzati come segue:

Modello base : $\left\{\mathcal{X}, f(t, \alpha, \gamma) = \alpha\gamma t^{\alpha-1} \exp\left(-\gamma t^{\alpha}\right)\right\}$ con $\gamma = e^{\boldsymbol{x}'\beta}$

Modello indotto : $\left\{\mathcal{X}^{(n)}, f(\boldsymbol{t}, \boldsymbol{\delta}, \alpha, \boldsymbol{\gamma}) = \prod_{i=1}^{n} \left(\alpha\gamma_i t_i^{\alpha-1}\right)^{\delta_i} \exp\left(-\gamma_i t_i^{\alpha}\right)\right\}$ con $\gamma_i = e^{\boldsymbol{x}_i'\boldsymbol{\beta}}$.

Ponendo $\gamma_i = e^{\boldsymbol{x}_i'\boldsymbol{\beta}}$ si giunge alla specificazione della verosimiglianza come:

$$\mathscr{L}(\mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta}; \boldsymbol{\beta}, \alpha) = \prod_{i=1}^{n} \left(\alpha\gamma_i t_i^{\alpha-1}\right)^{\delta_i} \exp\left(-\gamma_i t_i^{\alpha}\right),$$

dove $\mathbf{X}$ rappresenta la matrice avente come colonne le variabili esplicative, ovvero la matrice dei dati.

La parametrizzazione scelta non è l'unica possibile ma è particolarmente adatta a fini modellistici. Infatti, riparametrizzare $\gamma_i = e^{\boldsymbol{x}_i'\boldsymbol{\beta}}$ permette, una volta ottenuto un campione per $\boldsymbol{\beta}$, di interpretare l'esponenziale del coefficiente in termini di *Hazard Ratio* (HR). Si consideri, a riguardo, la j-sima variabile esplicativa. Fisse restando le altre esplicative, si ha, per ogni i:

$$HR(\beta_j) = \frac{\alpha e^{(x_{ij}+1)\beta_j} t_i^{\alpha-1}}{\alpha e^{x_{ij}\beta_j} t_i^{\alpha-1}}$$

$$= \frac{e^{(x_{ij}+1)\beta_j}}{e^{x_{ij}\beta_j}}$$

$$= e^{x_{ij}\beta_j+\beta_j-x_{ij}\beta_j} = e^{\beta_j},$$

sicché, l'interpretazione risulta quella usuale: per un incremento unitario della variabile x_j, il rischio aumenta di e^{β_j}. Infine, per fornire l'interpretazione dei coefficienti è sufficiente guardare all'$HR(\beta_j)$ infatti:

1. se $HR(\beta_j) > 1$ allora il regressore $\boldsymbol{x}_j$ rappresenta un fattore di rischio per l'evento.
2. se $HR(\beta_j) < 1$ allora il regressore $\boldsymbol{x}_j$ rappresenta un fattore protettivo dall'evento.
3. se $HR(\beta_j) = 1$ allora il regressore $\boldsymbol{x}_j$ è neutrale rispetto all'evento.

4.2 Scelta delle distribuzioni a priori e loro elicitazione

Per l'elicitazione delle distribuzioni a priori per i parametri α, β, sono state scelte due distribuzioni indipendenti, sicché

$$\pi(\boldsymbol{\beta}, \alpha \mid \mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta}) \propto \mathscr{L}(\mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta} \mid \boldsymbol{\beta}, \alpha)\pi(\alpha)\pi(\boldsymbol{\beta}).$$

Si osservi che in un contesto di dati ad elevata dimensionalità (in cui $p >> n$) è necessario elicitare distribuzioni appropriate su $\boldsymbol{\beta}$ al fine di poter rendere possibile

il processo di modellizzazione, in particolare è necessario imporre distribuzioni che comprimano i coefficienti. Inoltre, l'elicitazione delle distribuzioni è stata effettuata in modo da rendere più agevole la specificazione della *proposal distribution*. In particolare: per l'elicitazione di α non sussistono particolari vincoli e, pertanto, dal momento che si procederà implementando un RW Metropolis-Hastings gaussiano multivariato e il MALA pre-condizionato, si è propeso, onde evitare complicazioni che subentrerebbero con altre distribuzioni per le quali sarebbe necessario valutare il Jacobiano, elicitare una distribuzione log-normale:

$$\alpha \sim \text{lognorm}\left(\mu, \sigma^2\right),$$

in modo tale che $\log(\alpha) \sim N(\mu, \sigma^2)$.

Per quanto riguarda l'elicitazione della distribuzione a priori di $\boldsymbol{\beta}$ risulta necessario, come anticipato, ricorrere a una distribuzione a priori che comprima i parametri per rendere possibile la regressione. Si è propeso, dunque, per:

$$\boldsymbol{\beta}_a \sim \mathcal{N}_2\left(\mathbf{0}, \sigma^2_{\boldsymbol{\beta}}\mathbf{I}_2\right) \qquad \boldsymbol{\beta}_b \sim \mathcal{N}_{p-1}\left(\mathbf{0}, \frac{1}{2\lambda}\mathbf{I}_{p-1}\right)$$

ovvero per una normale bivariata per $\boldsymbol{\beta}_a = (\beta_0, \beta_1)$, dove β_0, β_1 rappresentano rispettivamente i coefficienti associati all'intercetta e al predittore lineare estratto dal Sarculator. La varianza $\sigma^2_{\boldsymbol{\beta}}$ è un valore elevato arbitrario, in quanto varianza associata all'intercetta e varianza associata al predittore lineare: tali coefficienti infatti non vengono penalizzati in fase di regressione e si elicita una varianza sufficientemente elevata in modo che la distribuzione ad essi associata possa risultare una a priori *vaga*, anche detta, quasi non informativa. Associando una varianza elevata alla distribuzione a priori si permette al modello di stimare i

coefficienti ignorando quasi completamente il contributo della a priori (da qui il termine non informativa). D'altra parte non avrebbe senso penalizzare anche tali coefficienti dal momento che il Sarculator è risultato un ottimo indice prognostico anche in fase di validazione esterna come riportato da Callegaro et al. (2019).

Per quanto riguarda i $p-1$ coefficienti restanti, $\boldsymbol{\beta}_b$, la a priori gaussiana multivariata elicitata, con vettore delle medie pari a zero e matrice di varianza-covarianza pari a $\frac{1}{2\lambda}\mathbf{I}_{p-1}$, corrisponde a imporre una penalizzazione tramite norma L_2 su tutti i coefficienti. Tale penalizzazione dipende da λ, che è detto parametro di precisione o di *shrinkage*. Lo *shrinkage* è il fattore che comprime i coefficienti verso lo zero: all'aumentare di λ aumenta la contrazione verso lo zero dei coefficienti. Riscrivendo in forma compatta, per il vettore $p+1$-dimensionale, $\boldsymbol{\beta}$, è stata elicitata la seguente distribuzione a priori:

$$\boldsymbol{\beta} \sim \mathcal{N}_{p+1}\left[\mathbf{0}, \begin{pmatrix} \sigma^2_{\beta}\mathbf{I}_2 & \mathbf{0} \\ \mathbf{0} & \frac{1}{2\lambda}\mathbf{I}_{p-1} \end{pmatrix}\right].$$

Si osservi che la distribuzione a priori elicitata per i $p-1$ coefficienti corrisponde a una penalizzazione Ridge, si veda Van Erp et al. (2019). Infatti, in ambito frequentista, il problema verrebbe affrontato come una minimizzazione vincolata. Più nel dettaglio, il problema, in ambito frequentista, come enucleato nel paragrafo 3.4 assumerebbe la seguente forma:

$$\hat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta},\alpha} \quad l(\boldsymbol{\beta}, \alpha \mid \mathbf{X}) - \lambda \sum_{j=2}^{p} \beta_i^2,$$

che è, di fatto, la medesima forma che si ottiene elicitando la distribuzione a priori Normale $p-1$ variata con media $\mathbf{0}$ e matrice di varianza-covarianza $\frac{1}{2\lambda}\mathbf{I}_{p-1}$,

per i $p-1$ coefficienti. Il vantaggio in ambito bayesiano è che è possibile specificare direttamente, tramite la matrice di varianza-covarianza, quali coefficienti penalizzare e quali non penalizzare. Si osservi che, poiché è stata elicitata una distribuzione a priori corrispondente a una penalità di tipo Ridge, è necessario standardizzare la matrice dei dati $\mathbf{X}$, in modo che il *range* di variazione delle esplicative non influenzi lo *shrinkage*. Pertanto, d'ora in avanti, la matrice dei dati, $\mathbf{X}$, è da intendersi come matrice dei dati standardizzata.

Dal momento che i dati oggetto d'analisi sono ad elevata dimensionalità potrebbe sorgere spontaneo domandarsi perché tra tutte le distribuzioni a priori esistenti che effettuano *shrinkage* sia stata scelta proprio la Ridge.

Tale quesito merita un supplemento di riflessione, conoscenza del dominio applicativo entro il quale si stanno analizzando i dati, conoscenza del tipo di dato stesso e la conoscenza di come lavorano le distribuzioni a priori bayesiane che effettuano *shrinkage*.

Una prima importante considerazione risiede nelle caratteristiche intrinseche di cui sono dotate le distribuzioni a priori che permettono, in modo differente, di imporre una penalizzazione sui coefficienti. Si consideri, a titolo di esempio, la distribuzione di Laplace. Essa è l'analogo bayesiano della regressione LASSO ma, se nel caso frequentista il LASSO ha in effetti la proprietà di poter azzerare esattamente i coefficienti sulla base del parametro di *shrinkage*, ciò non vale nell'ambito bayesiano se si considera la media a posteriori dei coefficienti. In ambito bayesiano, e in un contesto ad elevata dimensionalità caratterizzato da $p >> n$, elicitare una distribuzione a priori di Laplace o una Normale che equivalga a una penalizzazione Ridge non induce risultati significativamente differenti. La distribuzione di Laplace si caratterizza per una maggior massa di probabilità

in zero e per code più pesanti. Ma per quanta massa venga posta in zero, in ambito bayesiano, nemmeno una a priori di Laplace riesce ad azzerare la media a posteriori dei coefficienti. Al fine di ottenere effettivamente l'azzeramento dei coefficienti, considerando una distribuzione a priori di Laplace bisognerebbe considerare la moda a posteriori, *Maximum a posteriori*. Si osservi tuttavia che sebbene la moda a posteriori possa effettivamente risultare pari a zero, ciò non vale per gli intervalli di credibilità intorno alla moda medesima. In altri termini, in ambito bayesiano, considerare la moda a posteriori al fine di effettuare selezione di variabili implicherebbe ignorare l'incertezza intorno al valor modale. Ciò si tradurrebbe in un *bias* di selezione.

Questo spiega perché, nel caso in oggetto, si è proceduto elicitando una distribuzione a priori Normale che effettuasse uno *shrinkage* sui coefficienti, consapevoli del fatto che non li avrebbe azzerati, ma d'altra parte non lo avrebbe fatto nemmeno la a priori di Laplace. Si consideri inoltre che la penalizzazione derivante dalla distribuzione di Laplace induce altresì la perdita della proprietà dell'oracolo, si veda Park and Casella (2008).

Riassumendo: la distribuzione di Laplace aumenta la penalizzazione e la massa di probabilità intorno a zero ma nessun coefficiente avrà media a posteriori esattamente uguale a zero. Sarebbe quindi necessario, come nel caso della regressione Ridge, utilizzare comunque un criterio arbitrario per stabilire se una variabile è da selezionare.

Recentemente, diverse distribuzioni a priori che penalizzassero i coefficienti sono state studiate. La differenza tra tali distribuzioni risiede nel grado e nel tipo di penalizzazione e non già nella capacità di azzerare o meno i coefficienti in modo esatto. Un esempio di distribuzione a priori particolarmente interessante

al fine di comprendere come differenti distribuzioni a priori possano penalizzare i coefficienti, è la *Horsehoe Prior* introdotta da Carvalho et al. (2009) o la *Regularized Horseshoe*, una variazione della *Horseshoe* proposta da Piironen and Vehtari (2017). La distribuzione *horsehoe* non è scrivibile analiticamente in forma chiusa, tuttavia la sua specificazione gerarchica ha forma:

$$\beta_j \mid \lambda_j, \tau \sim \mathcal{N}(0, \lambda_j^2 \tau^2), \tag{14}$$

con:

$$\begin{aligned} \beta_j \mid \lambda_j, \tau &\sim \mathcal{N}(0, \lambda_j^2 \tau^2) \\ \lambda_j &\sim \mathcal{C}^+(0, 1) \\ \tau &\sim \mathcal{C}^+(0, 1). \end{aligned}$$

Tale distribuzione è di particolare interesse nel contesto di selezione di variabili perché: è una distribuzione continua, attribuisce a ciascun coefficiente una penalizzazione differente, ed è una distribuzione a priori che conduce a una specificazione del modello gerarchica, elicitando una *hyperprior* sulla varianza dei coefficienti e una ulteriore *hyperprior* sul parametro di scala della *hyperprior*. Il vantaggio di tale distribuzione deriva dal fatto che non è necessario elicitare nessun iperparametro in quanto, per costruzione, su essi viene imposta una *hyperprior*. Inoltre è costruita in modo tale da poter porre, da un punto di vista asintotico, una massa di probabilità infinita in zero; è di fatto la distribuzione a priori continua che più si avvicina ad azzerare effettivamente i coefficienti. Presenta anche code molto più pesanti di una distribuzione a priori di Laplace, di una t-Student o di una Normale. La natura delle code porta a incrementare ulteriormente la penalizzazione dei coefficienti. Tuttavia, l'aspetto più interessante è dovuto alla

varianza della distribuzione dei coefficienti: con riferimento alla equazione (14), la varianza dei coefficienti risulta regolata da due parametri, uno specifico per ogni coefficiente λ_j e uno globale τ, comune a tutti i coefficienti. La presenza di un parametro di *shrinkage* globale e di uno locale è ciò che permette alla distribuzione a priori di porre una penalizzazione quasi nulla su certi coefficienti. Se un coefficiente è molto associato alla variabile risposta il parametro di *shrinkage* locale può controbilanciare la penalizzazione globale in modo da annullare l'effetto di penalizzazione su tale coefficiente (la *regularized horshoe* invece incentiva lo *shrinkage* anche per le variabili che sono fortemente rilevanti).

Ne discende che, in contesti di selezione di variabili la distribuzione *horseshoe* sia particolarmente utile, perché sebbene anch'essa necessiti di criteri arbitrari per stabilire se includere o meno una variabile, può fornire indicazioni più chiare circa quali potrebbero essere le variabili da selezionare. Il condizionale è d'obbligo dal momento che, in contesti in cui la numerosità campionaria è limitata, come nel caso in oggetto, lo *shrinkage* globale può dominare completamente l'effetto di quello locale, per dettagli si veda Bhadra et al. (2019). Se il numero di variabili è molto elevato rispetto al numero di osservazioni lo *shrinkage* locale non riuscirà a sottrarre all'effetto di quello globale le variabili, anche se quest'ultime sono effettivamente associate alla risposta, ciò potrà avvenire solo per quelle variabili la cui associazione con la risposta è estremamente forte.

Si osservi, tuttavia, che ciò non è desiderabile in un contesto in cui come ipotesi iniziale non si hanno pochi predittori molto informativi ma, al contrario, molti predittori poco informativi, si veda Tibshirani and Wasserman (2016). Questo è il caso delle variabili radiomiche: molte variabili sono tra loro correlate e molte volte catturano effetti e caratteristiche molto simili tra loro, le variabili di

texture tenderanno a essere molto correlate tra loro, così come quelle morfologiche.

In termini più espliciti, la radiomica non rappresenta un contesto in cui è ragionevole ipotizzare sparsità ma, al contrario, abbondanza e ridondanza sia di segnale che di rumore.

Un'altra distribuzione a priori che poteva essere considerata era la *hyperlasso*, la quale risolve le principali criticità della distribuzione di Laplace, soddisfa la proprietà dell'oracolo ma, nuovamente, ipotizza sparsità, per dettagli si consiglia la lettura di Griffin and Brown (2011).

Alla luce di tali considerazioni non si ravvisano motivi validi per elicitare una distribuzione a priori diversa dalla Ridge. Il campione oggetto d'analisi si caratterizza per una bassa numerosità campionaria e un numero molto elevato di variabili, sulle quali non è ragionevole imporre l'ipotesi di sparsità. Con queste premesse sembra più ragionevole elicitare una distribuzione a priori che penalizzi tutti i coefficienti e che non ipotizzi sparsità. In Figura 7 vengono riportati i grafici delle diverse distribuzioni a priori che effettuano lo *shrinkage* per visualizzare le principali differenze.

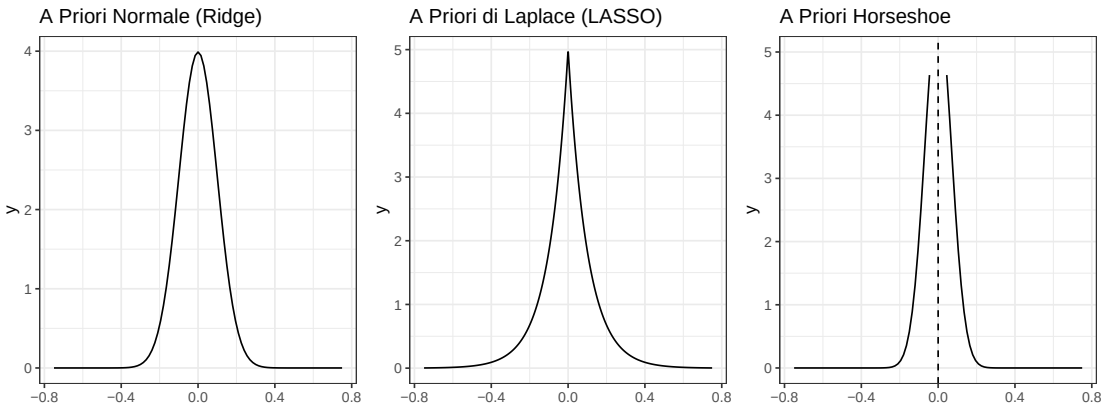


Figura 7: Forma delle distribuzioni a priori che effettuano shrinkage, il parametro di shrinkage è posto pari a 0.1 in tutti e tre i casi

La domanda, che discende da quanto appena esplicitato, è come effettuare selezione di variabili attraverso delle distribuzioni a priori che non inducono sparsità? La domanda non ha una risposta esatta: da un punto di vista teorico qualsiasi criterio che si scelga per effettuare la selezione introduce un *bias*. Infatti, teoricamente, sulla base di distribuzioni a priori continue non è possibile effettuare una selezione di variabili senza introdurre un *bias* di selezione. D'altra parte, se vengono utilizzati gli stessi dati per stimare un modello e selezionare variabili, l'errore entra nel processo per la semplice ragione di aver utilizzato i dati per più di uno scopo: il modello viene costruito sui dati e la rilevanza o meno delle variabili viene dedotta dal modello costruito. Da un punto di vista bayesiano la fonte dell'errore di selezione è facile da intuire: selezionando alcune variabili attraverso un criterio arbitrario, per quanto ragionevole esso possa essere, si trascura l'incertezza relativa alle altre variabili non rientranti nella selezione.

La selezione di variabili introduce sempre un *bias* di selezione. Fa eccezione la distribuzione a priori *spike and slab*, la quale riesce effettivamente a porre esattamente a zero i coefficienti. Tuttavia è una distribuzione derivante da una mistura di una v.a. continua e una discreta: La *spike and slab* utilizza una bernoulliana per probabilizzare la rilevanza o meno di una variabile, se la variabile non è rilevante il coefficiente associato viene posto a zero, se è rilevante la distribuzione elicitata è una normale vaga, in modo che il coefficiente possa essere determinato sulla base dei dati. Sebbene tale distribuzione sembri funzionare molto bene da un punto di vista intuitivo, da un punto di vista pratico i contesti in cui possono sorgere numerosi problemi sono proprio quelli in cui la numerosità campionaria è molto bassa e il numero di variabili molto elevato; proprio il caso oggetto d'analisi. Ciò è intuitivamente comprensibile: essendoci molto rumore,

e segnale largamente correlato, non sempre risulta possibile procedere con la regressione vista la a priori bernoulliana che caratterizza la *spike and slab* e, anche quando ciò avviene, non sempre la distribuzione è in grado di caratterizzare correttamente il segnale. Per ulteriori dettagli sulla a priori *spike and slab* si raccomandano i lavori di George and McCulloch (1997) e Ishwaran and Rao (2005). Onde evitare l'introduzione di qualsiasi tipo di *bias*, sembra pertanto più opportuno considerare tutti i coefficienti penalizzati campionati dalla distribuzione a posteriori e considerare, a posteriori, un modello che tenga conto dell'incertezza di tutti i modelli, sulla base dei criteri che verranno enucleati nei paragrafi 4.6, 4.6.1, 4.6.3 e 4.7.

Enucleate le ragioni che hanno condotto a elicitare proprio le distribuzioni a priori specificate poc'anzi, l'ultimo nodo da sciogliere è come elicitare gli iperparametri di tali distribuzioni. In altri termini verrà esposto il razionale dietro alla elicitazione di μ, σ^2, λ. Per quanto concerne il parametro di precisione, λ, non è insolito attribuirgli una *hyperprior Half Cauchy*, si faccia nuovamente riferimento a Van Erp et al. (2019). In questo caso, tuttavia, si è preferito optare per una sequenza di λ su cui valutare il modello, in modo che potesse poi essere possibile specificare un modello dato dalla combinazione dei modelli con i coefficienti corrispondenti a un diverso valore del parametro di precisione.

I parametri μ, σ^2 sono stati elicitati partendo da considerazioni teoriche sulla distribuzione di Weibull. Infatti, se $\alpha = 1$ allora la funzione di rischio risulterebbe costante, in quanto la distribuzione derivante sarebbe una esponenziale, si veda Hallinan Jr (1993). Si è dunque proceduto elicitando gli iperparametri della distribuzione a priori di α in modo tale che la media della distribuzione a priori di α corrispondesse proprio a uno. Questo consente di partire da una condizione

di neutralità rispetto alla specificazione della funzione di rischio. Impostando quindi una varianza elevata σ^2_α, in modo tale da rendere la log-normale quasi non informativa, si è lasciato che il parametro si aggiornasse basandosi maggiormente sulla base della verosimiglianza. Affinché $\mathbb{E}(\alpha) = 1$ si devono specificare gli iperparametri della log-normale come segue: $\mu = -\log(\sigma^2_\alpha+1)/2$ e $\sigma^2 = \log(\sigma^2_\alpha+1)$. Si ricordi infatti che, se: $\alpha \sim \text{lognorm}(\mu, \sigma^2)$ allora:

$$\mathbb{E}(\alpha) = \exp\left(\mu + \frac{\sigma^2}{2}\right) \quad \text{e} \quad \text{Var}(\alpha) = \exp\left(2\mu + \sigma^2\right)\left(\exp(\sigma^2) - 1\right).$$

Per dettagli relativi alla distribuzione log-normale si rimanda a Heyde (1963). Si ha quindi:

$$\begin{cases} \mathbb{E}(\alpha) = \exp\left(\mu + \frac{\sigma^2}{2}\right) = 1 \\ \\ \text{Var}(\alpha) = \exp(2\mu + \sigma^2)(\exp(\sigma^2) - 1) = \sigma^2_\alpha \end{cases} \Leftrightarrow$$

$$\begin{cases} \mu = -\frac{\sigma^2}{2} \\ \\ \exp(\sigma^2) = \sigma^2_\alpha + 1 \end{cases} \Leftrightarrow \begin{cases} \mu = -\frac{\log(\sigma^2_\alpha+1)}{2} \\ \\ \sigma^2 = \log(\sigma^2_\alpha + 1) \end{cases}.$$

In ultimo, dal momento che la funzione di densità della log-normale è

$$f(\alpha) = \frac{1}{\alpha\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(\log(\alpha) - \mu)^2}{2\sigma^2}\right\},$$

e che il logaritmo di una log-normale è una Normale di medesimi parametri, è possibile riparametrizzare ponendo $\xi = \log(\alpha)$ e considerare la distribuzione a priori per ξ, con $\xi \sim \mathcal{N}\left(-\frac{\log(\sigma^2_\alpha+1)}{2}, \log(\sigma^2_\alpha + 1)\right)$.

4.3 Distribuzione a posteriori

Definita la funzione di verosimiglianza con la parametrizzazione opportuna e ricordando che $\gamma_i = e^{\boldsymbol{x}_i'\boldsymbol{\beta}}$, e che si è posto $\log(\alpha) = \xi$, è possibile ricavare il nucleo della distribuzione a posteriori:

$$
\begin{aligned}
\pi(\boldsymbol{\beta}, \xi \mid \mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta}) &\propto \mathscr{L}(\mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta} \mid \boldsymbol{\beta}, \xi)\pi(\xi)\pi(\boldsymbol{\beta}) \\
&\propto \prod_{i=1}^{n} \left(e^{\xi}\gamma_i t_i^{e^{\xi}-1}\right)^{\delta_i} \exp\left(-\gamma_i t_i^{e^{\xi}}\right) \frac{\exp\left(-\frac{\left[\xi+\log(\sigma_\alpha^2+1)/2\right]^2}{2\log(\sigma_\alpha^2+1)}\right)}{\sqrt{2\pi\log(\sigma_{\alpha+1}^2)}} \times \\
&\qquad \times \frac{\lambda^{(p-1)/2}}{\pi^{(p-1)/2}} \exp\left(-\lambda\sum_{j=2}^{p}\beta_j^2\right) \frac{\exp\left[-(\beta_0^2+\beta_1^2)/(2\sigma_{\boldsymbol{\beta}}^2)\right]}{2\pi\sigma_{\boldsymbol{\beta}}^2} \\
&\propto \prod_{i=1}^{n} \left(e^{\xi}e^{\boldsymbol{x}_i'\boldsymbol{\beta}}t_i^{e^{\xi}-1}\right)^{\delta_i} \exp\left(-e^{\boldsymbol{x}_i'\boldsymbol{\beta}}t_i^{e^{\xi}}\right) \exp\left(-\frac{\left[\xi+\log(\sigma_\alpha^2+1)/2\right]^2}{2\log(\sigma_\alpha^2+1)}\right) \times \\
&\qquad \times \frac{\lambda^{(p-1)/2}}{2\pi\sigma_{\boldsymbol{\beta}}^2} \exp\left(-\lambda\sum_{j=2}^{p}\beta_j^2\right) \exp\left(-\frac{\beta_0^2+\beta_1^2}{2\sigma_{\boldsymbol{\beta}}^2}\right) \\
&\propto \prod_{i=1}^{n} \left(e^{\xi+\boldsymbol{x}_i'\boldsymbol{\beta}}t_i^{e^{\xi}-1}\right)^{\delta_i} \exp\left(-e^{\boldsymbol{x}_i'\boldsymbol{\beta}}t_i^{e^{\xi}}\right) \exp\left(-\frac{\xi^2}{2\log(\sigma_\alpha^2+1)} - \frac{\xi}{2}\right) \times \\
&\qquad \times \lambda^{(p-1)/2} \exp\left(-\lambda\sum_{j=2}^{p}\beta_j^2 - \frac{\beta_0^2+\beta_1^2}{2\sigma_{\boldsymbol{\beta}}^2}\right) \\
&\propto \exp\left(\sum_{i=1}^{n}\delta_i\xi + \delta_i\boldsymbol{x}_i'\boldsymbol{\beta} - e^{\boldsymbol{x}_i'\boldsymbol{\beta}}t_i^{e^{\xi}}\right) \prod_{i=1}^{n} t_i^{\delta_i(e^{\xi}-1)} \times \\
&\qquad \times \exp\left(-\frac{\xi^2}{2\log(\sigma_\alpha^2+1)} - \frac{\xi}{2} - \lambda\sum_{j=2}^{p}\beta_j^2 - \frac{\beta_0^2+\beta_1^2}{2\sigma_{\boldsymbol{\beta}}^2}\right).
\end{aligned}
$$

Le costanti moltiplicative sono state eliminate in quanto non caratterizzanti il nucleo della distribuzione a posteriori.

4.4 Definizione della *proposal distribution*

Dal momento che non è possibile ricavare per via analitica le distribuzioni *full conditionals* e quindi, al fine di campionare dalla distribuzione a posteriori il *Gibbs Sampling* è precluso, la scelta del metodo computazionale di campionamento ricade, in via naturale, sul Metropolis-Hastings. In particolare si è propeso, in prima istanza, per un *Random Walk* MH scegliendo come *proposal distribution* una gaussiana $p+2$ variata, di media $\boldsymbol{\mu}$ e matrice di varianza-covarianza $\boldsymbol{\Sigma}$. Chiamata $q\left(\boldsymbol{\vartheta}, \boldsymbol{\vartheta}^*\right)$ la funzione di densità della *proposal distribution*, il vettore dei parametri candidati ad ogni iterazione, condizionatamente al vettore dei parametri del passo precedente, ha legge di distribuzione:

$$\boldsymbol{\vartheta}^* \mid \boldsymbol{\vartheta} \sim \mathcal{N}_{p+2}\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right),$$

dove $p+2$ è dato dalla somma del numero di coefficienti $\boldsymbol{\beta}$ (considerando anche l'intercetta) e del parametro di forma della Weibull. Si osservi che la distribuzione normale multivariata è una distribuzione simmetrica, il che implica: $q\left(\boldsymbol{\vartheta} \mid \boldsymbol{\vartheta}^*\right) = q\left(\boldsymbol{\vartheta}^* \mid \boldsymbol{\vartheta}\right)$, donde, la probabilità di accettazione definita nella (8) può essere semplificata come:

$$\alpha\left(\boldsymbol{\vartheta}, \boldsymbol{\vartheta}^*\right) = \min\left\{\frac{\pi(\boldsymbol{\vartheta}^* \mid \mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta}) q\left(\boldsymbol{\vartheta}^*, \boldsymbol{\vartheta}\right)}{\pi(\boldsymbol{\vartheta} \mid \mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta}) q\left(\boldsymbol{\vartheta}, \boldsymbol{\vartheta}^*\right)}, 1\right\} \Longrightarrow \alpha\left(\boldsymbol{\vartheta}, \boldsymbol{\vartheta}^*\right) = \min\left\{\frac{\pi(\boldsymbol{\vartheta}^* \mid \mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta})}{\pi(\boldsymbol{\vartheta} \mid \mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta})}, 1\right\}.$$

Ciò permette di calcolare più rapidamente la probabilità di accettazione una volta generato il candidato.

Come noto, si veda a titolo di esempio il lavoro di Sherlock et al. (2010), la

convergenza dell'algoritmo è fortemente influenzata dall'inizializzazione di $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ che, pertanto, deve essere sufficientemente oculata per non incorrere in un numero di iterazioni eccessivo per raggiungere la convergenza. Un'opzione ragionevole è quella di inizializzare $\boldsymbol{\mu}$ come $\tilde{\boldsymbol{\vartheta}} = \arg\max_{\boldsymbol{\vartheta}} \log \pi(\boldsymbol{\vartheta} \mid \mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta})$ per ogni λ della sequenza. I valori di $\tilde{\boldsymbol{\vartheta}}_\lambda$, tali per cui si ottiene l'ottimo della distribuzione log a posteriori per i diversi valori di λ, si possono quindi utilizzare come media di ciascuna *proposal distribution*.

In altri termini, si considera una sequenza di valori di λ, si procede, per ogni λ, ottimizzando $\arg\max_{\boldsymbol{\vartheta}} \log \pi(\boldsymbol{\vartheta} \mid \mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta})$ ottenendo $\tilde{\boldsymbol{\vartheta}}$, si utilizza il vettore ottimo trovato per calcolare la matrice dell'informazione osservata, e si inizializzano quindi la media e la matrice di varianza-covarianza della *proposal*.

La scelta di considerare la distribuzione log a posteriori rispetto alla distribuzione a posteriori è stata presa in ragione del fatto che:

1. l'ottimizzazione della distribuzione log a posteriori risulta più stabile e meno onerosa computazionalmente;
2. dalla distribuzione log a posteriori è possibile ricavare una stima della matrice di varianza-covarianza sensata per la *proposal distribution*, proprio a partire dal punto di massimo utilizzato per inizializzarne la media.

Per quanto concerne il secondo punto, nel paragrafo 3.2.2 si è già enucleato il ruolo ricoperto dall'informazione osservata per la stima della matrice di varianza-covarianza di una distribuzione normale:

$$-\boldsymbol{\mathcal{H}}_{\log \pi(\cdot \mid \mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta})}\Big|_{\boldsymbol{\vartheta} = \tilde{\boldsymbol{\vartheta}}_\lambda} = \boldsymbol{\mathcal{I}}\left(\tilde{\boldsymbol{\vartheta}}_\lambda\right) \Rightarrow \boldsymbol{\Sigma}_\lambda \approx \tilde{\boldsymbol{\Sigma}}_\lambda = \boldsymbol{\mathcal{I}}^{-1}\left(\tilde{\boldsymbol{\vartheta}}_\lambda\right). \tag{15}$$

Verrà utilizzato proprio tale risultato per inizializzare la matrice di varianza-

covarianza delle *proposal distributions*. Ne discende che l'inizializzazione delle *proposal distributions* è data da

$$\boldsymbol{\vartheta}^*{}_\lambda \mid \boldsymbol{\vartheta}_\lambda \sim \mathcal{N}_{p+2}\left(\tilde{\boldsymbol{\vartheta}}_\lambda, \boldsymbol{\mathcal{I}}^{-1}\left(\tilde{\boldsymbol{\vartheta}}_\lambda\right)\right).$$

È possibile dimostrare che per valori elevati di p, come nel caso oggetto d'analisi, è opportuno adottare una versione più efficiente della matrice di varianza-covarianza: in particolare la *proposal distribution* per un elevato numero di variabili assume forma:

$$\boldsymbol{\vartheta}^*{}_\lambda \mid \boldsymbol{\vartheta}_\lambda \sim \mathcal{N}_{p+2}\left(\tilde{\boldsymbol{\vartheta}}_\lambda, \frac{2.38^2}{p+2}\boldsymbol{\mathcal{I}}^{-1}\left(\tilde{\boldsymbol{\vartheta}}_\lambda\right)\right).$$

Per ulteriori dettagli si rimanda all'articolo di Roberts and Rosenthal (2001). Si preferisce rimandare per approfondimenti alla fonte citata, piuttosto che derivare il fattore $\frac{2.38^2}{p+2}$, dal momento che la sua derivazione segue le linee di quella, più complicata, enucleata per il MALA al paragrafo 3.3.1.7.

Intuitivamente, il fattore $\frac{2.38^2}{p+2}$ serve a non rendere nullo, o quasi nullo, il tasso di accettazione, che sarebbe tale al crescere del numero di variabili. Tale forma della matrice di varianza-covarianza serve a rendere più efficiente il campionamento dalla distribuzione a posteriori e, teoricamente, a non dover estendere di troppo il numero di iterazioni necessarie per raggiungere la convergenza.

Rimane quindi da calcolare il punto di massimo della distribuzione log a posteriori per ciascun valore di λ specificato. Si osservi che si è scelto di considerare 100 distinti valori di λ nell'intervallo $[e^5, e^{10.5}]$.

L'intervallo in oggetto è stato scelto verificando, dopo la fase di campionamento, quale fosse il valor minimo di λ tale per cui il modello risultasse identificabile. Tale valore è stato fissato come estremo inferiore dell'intervallo. Il valor massimo

è stato scelto in modo tale che la penalizzazione si traducesse nella quasi totale capacità di azzeramento dei coefficienti soggetti alla penalizzazione.

Il processo per la definizione della sequenza ha pertanto richiesto, da un punto di vista metodologico, un elevato grado di attenzione e non è stato lineare, anzi. Le informazioni ottenute in fase di campionamento sono state utilizzate per raffinare gli estremi dell'intervallo, fino a giungere a un intervallo quanto più ottimale possibile che fornisse, in particolare, solo modelli identificabili.

4.4.1 Distribuzione log a posteriori e ottimizzazione

D'ora in avanti il pedice λ verrà omesso da $\boldsymbol{\vartheta}_\lambda$ per non appesantire la notazione. Si ricordi tuttavia che il processo che verrà descritto per la derivazione del punto di ottimo, per la derivazione dell'informazione osservata, per la decomposizione di quest'ultima e per l'inizializzazione delle cento distinte *proposal distributions* è da intendersi identico, ed è stato ripetuto, per ogni valore di λ.

La distribuzione log a posteriori, $\log \pi(\boldsymbol{\beta}, \xi \mid \mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta})$, è facilmente ricavabile a partire dalla distribuzione a posteriori. Inoltre presenta il vantaggio di rendere la distribuzione più gestibile sia dal punto di vista analitico che da quello computazionale. Si ha:

$$\log \pi(\boldsymbol{\beta}, \xi \mid \mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta}) = \sum_{i=1}^{n} \left(\delta_i \xi + \delta_i \boldsymbol{x}_i' \boldsymbol{\beta} - e^{\boldsymbol{x}_i' \boldsymbol{\beta}} t_i^{e^\xi} + \delta_i (e^\xi - 1) \log t_i \right) + \\ - \frac{\xi^2}{2 \log(\sigma_\alpha^2 + 1)} - \frac{\xi}{2} - \lambda \sum_{j=2}^{p} \beta_j^2 - \frac{\beta_0^2 + \beta_1^2}{2\sigma_{\boldsymbol{\beta}}^2} + c,$$

con c costante additiva. Ne discende che il punto di ottimo, $\tilde{\boldsymbol{\vartheta}} = \left(\tilde{\boldsymbol{\beta}}, \tilde{\xi} \right)$ è il punto tale per cui il gradiente si annulla.

Formalmente: $\tilde{\boldsymbol{\vartheta}} \mid \nabla_{\boldsymbol{\vartheta}} \log \pi(\tilde{\boldsymbol{\beta}}, \tilde{\xi} \mid \mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta}) = 0$. Risulta dunque necessario risolvere un sistema di $p+2$ equazioni:

$$
\begin{cases}
\frac{\partial \log \pi(\cdot|\mathbf{X},\boldsymbol{t},\boldsymbol{\delta})}{\partial \xi} = 0 \\
\frac{\partial \log \pi(\cdot|\mathbf{X},\boldsymbol{t},\boldsymbol{\delta})}{\partial \beta_0} = 0 \\
\frac{\partial \log \pi(\cdot|\mathbf{X},\boldsymbol{t},\boldsymbol{\delta})}{\partial \beta_1} = 0 \\
\frac{\partial \log \pi(\cdot|\mathbf{X},\boldsymbol{t},\boldsymbol{\delta})}{\partial \beta_2} = 0 \\
\quad \vdots \\
\frac{\partial \log \pi(\cdot|\mathbf{X},\boldsymbol{t},\boldsymbol{\delta})}{\partial \beta_j} = 0 \\
\quad \vdots \\
\frac{\partial \log \pi(\cdot|\mathbf{X},\boldsymbol{t},\boldsymbol{\delta})}{\partial \beta_n} = 0
\end{cases}
\Longrightarrow
\begin{cases}
\sum_{i=1}^{n} \left[\delta_i + \log t_i (\, e^{\xi}\delta_i - e^{\boldsymbol{x}_i'\boldsymbol{\beta}} t_i^{e^{\xi}} e^{\xi} \,)\right] - \frac{\xi}{\log(\sigma_\alpha^2+1)} - \frac{1}{2} = 0 \\
\sum_{i=1}^{n} \left(\delta_i x_{i0} - x_{i0} e^{\boldsymbol{x}_i'\boldsymbol{\beta}} t_i^{e^{\xi}}\right) - \frac{\beta_0}{\sigma_{\boldsymbol{\beta}}^2} = 0 \\
\sum_{i=1}^{n} \left(\delta_i x_{i1} - x_{i1} e^{\boldsymbol{x}_i'\boldsymbol{\beta}} t_i^{e^{\xi}}\right) - \frac{\beta_1}{\sigma_{\boldsymbol{\beta}}^2} = 0 \\
\sum_{i=1}^{n} \left(\delta_i x_{i2} - x_{i2} e^{\boldsymbol{x}_i'\boldsymbol{\beta}} t_i^{e^{\xi}}\right) - 2\lambda\beta_2 = 0 \\
\quad \vdots \\
\sum_{i=1}^{n} \left(\delta_i x_{ij} - x_{ij} e^{\boldsymbol{x}_i'\boldsymbol{\beta}} t_i^{e^{\xi}}\right) - 2\lambda\beta_j = 0 \\
\quad \vdots \\
\sum_{i=1}^{n} \left(\delta_i x_{ip} - x_{ip} e^{\boldsymbol{x}_i'\boldsymbol{\beta}} t_i^{e^{\xi}}\right) - 2\lambda\beta_p = 0
\end{cases}
$$

Il sistema non è risolvibile analiticamente ed è necessario ricorrere all'approssimazione numerica, per la quale è stato scelto come algoritmo di ricerca del minimo della $-\log \pi(\boldsymbol{\beta}, \xi \mid \mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta})$, procedura equivalente a massimizzare la $\log \pi(\boldsymbol{\beta}, \xi \mid \mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta})$, un algoritmo dei minimi quadrati non lineari adattivo, per i cui dettagli si veda Dennis Jr et al. (1981). La procedura è stata reiterata sulla sequenza dei 100 distinti valori di λ e sono stati selezionati i vettori $\tilde{\boldsymbol{\vartheta}}$ tali da minimizzare $-\log \pi(\boldsymbol{\beta}, \xi \mid \mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta})$.

4.4.2 Stima della matrice di varianza-covarianza

Trovato il $\tilde{\boldsymbol{\vartheta}}$ ottimo tramite l'annullamento del gradiente è dunque possibile ricavare l'informazione osservata e inizializzare la matrice di varianza-covarianza in virtù della relazione (15). A tal fine si considerino gli elementi che compongono la matrice hessiana $\boldsymbol{\mathcal{H}}$ cambiati di segno. Si osservi che da un punto di vista implementativo si è scelto, per ragioni di efficienza, di procedere per via analitica, ricavando tutte le derivate parziali utili per l'hessiana e valutandole in corrispondenza del punto di massimo della log a posteriori. Si osservi che rispetto alle funzioni di indubbia utilità che stimano, con metodi di approssimazione numerica, i valori dell'hessiana in un punto specifico, in questo caso si è propeso non già per un'approssimazione ma per l'esattezza. La scelta si è rivelata anche molto più efficiente da un punto di vista computazionale: una volta codificate le derivate seconde parziali, procedere per via esatta piuttosto che per ottimizzazione numerica ha ridotto il tempo medio di stima della hessiana da circa 45 minuti a circa 21 secondi. Per rendere più compatta la notazione si definisca una funzione di $\tilde{\beta}_j$ del tipo:

$$f(\tilde{\beta}_j) = \begin{cases} \lambda & \text{se } j \geq 2 \\ \\ \frac{1}{2\sigma^2_\beta} & \text{se } j < 2 \end{cases}, \quad \text{con} \quad j = 0, \dots p.$$

In tal modo si penalizzano o meno i coefficienti a seconda di quale coefficiente si sta considerando. Posto $j, k = 0, \dots p \ \wedge \ j \neq k$ si ha, quindi, che gli elementi che compongono la matrice hessiana cambiata di segno e valutata nel punto di ottimo risultano:

$$-\frac{\partial^2}{\partial\xi^2}\log\pi(\boldsymbol{\beta},\xi\mid\mathbf{X},\boldsymbol{t},\boldsymbol{\delta})\Big|_{\boldsymbol{\vartheta}=\tilde{\boldsymbol{\vartheta}}}=\sum_{i=1}^{n}\left[e^{\boldsymbol{x}_i'\tilde{\boldsymbol{\beta}}}\log t_i\left(e^{\tilde{\xi}}t_i^{e^{\tilde{\xi}}}+e^{2\tilde{\xi}}t_i^{e^{\tilde{\xi}}}\log t_i\right)-e^{\tilde{\xi}}\delta_i\log t_i\right]+\frac{1}{\log(\sigma_\alpha^2+1)};$$

$$-\frac{\partial^2}{\partial\beta_j^2}\log\pi(\boldsymbol{\beta},\xi\mid\mathbf{X},\boldsymbol{t},\boldsymbol{\delta})\Big|_{\boldsymbol{\vartheta}=\tilde{\boldsymbol{\vartheta}}}=\sum_{i=1}^{n}\left(x_{ij}^2e^{\boldsymbol{x}_i'\tilde{\boldsymbol{\beta}}}t_i^{e^{\tilde{\xi}}}\right)+2f(\tilde{\beta}_j);$$

$$-\frac{\partial^2}{\partial\beta_j\partial\xi}\log\pi(\boldsymbol{\beta},\xi\mid\mathbf{X},\boldsymbol{t},\boldsymbol{\delta})\Big|_{\boldsymbol{\vartheta}=\tilde{\boldsymbol{\vartheta}}}=-\frac{\partial^2}{\partial\xi\partial\beta_j}\log\pi(\boldsymbol{\beta},\xi\mid\mathbf{X},\boldsymbol{t},\boldsymbol{\delta})\Big|_{\boldsymbol{\vartheta}=\tilde{\boldsymbol{\vartheta}}}=\sum_{i=1}^{n}x_{ij}e^{\boldsymbol{x}_i'\tilde{\boldsymbol{\beta}}+\xi}t_i^{e^{\tilde{\xi}}}\log t_i;$$

$$-\frac{\partial^2}{\partial\beta_j\partial\beta_k}\log\pi(\boldsymbol{\beta},\xi\mid\mathbf{X},\boldsymbol{t},\boldsymbol{\delta})\Big|_{\boldsymbol{\vartheta}=\tilde{\boldsymbol{\vartheta}}}=-\frac{\partial^2}{\partial\beta_k\partial\beta_j}\Big|_{\boldsymbol{\vartheta}=\tilde{\boldsymbol{\vartheta}}}\log\pi(\boldsymbol{\beta},\xi\mid\mathbf{X},\boldsymbol{t},\boldsymbol{\delta})=\sum_{i=1}^{n}x_{ij}x_{ik}e^{\boldsymbol{x}_i'\tilde{\boldsymbol{\beta}}}t_i^{e^{\tilde{\xi}}},$$

con $\tilde{\boldsymbol{\vartheta}}=\left(\tilde{\boldsymbol{\beta}},\tilde{\xi}\right)$. Posto $\ell=\log\pi(\boldsymbol{\beta},\xi\mid\mathbf{X},\boldsymbol{t},\boldsymbol{\delta})$ si può facilmente ricavare $-\boldsymbol{\mathcal{H}}\Big|_{\boldsymbol{\vartheta}=\tilde{\boldsymbol{\vartheta}}}$ come:

$$\begin{pmatrix}
-\frac{\partial^2\ell}{\partial\xi^2}\Big|_{\tilde{\boldsymbol{\vartheta}}} & -\frac{\partial^2\ell}{\partial\beta_0\partial\xi}\Big|_{\tilde{\boldsymbol{\vartheta}}} & \cdots & \cdots & \cdots & \cdots & -\frac{\partial^2\ell}{\partial\beta_p\partial\xi}\Big|_{\tilde{\boldsymbol{\vartheta}}}\\
-\frac{\partial^2\ell}{\partial\xi\partial\beta_0}\Big|_{\tilde{\boldsymbol{\vartheta}}} & \frac{1}{\sigma_\beta^2}+\sum_{i=1}^{n}x_{i0}^2e^{\boldsymbol{x}_i'\tilde{\boldsymbol{\beta}}}t_i^{e^{\tilde{\xi}}} & -\frac{\partial^2\ell}{\partial\beta_0\partial\beta_1}\Big|_{\tilde{\boldsymbol{\vartheta}}} & \cdots & \cdots & \cdots & -\frac{\partial^2\ell}{\partial\beta_0\partial\beta_p}\Big|_{\tilde{\boldsymbol{\vartheta}}}\\
\vdots & -\frac{\partial^2\ell}{\partial\beta_1\partial\beta_0}\Big|_{\tilde{\boldsymbol{\vartheta}}} & \frac{1}{\sigma_\beta^2}+\sum_{i=1}^{n}x_{i1}^2e^{\boldsymbol{x}_i'\tilde{\boldsymbol{\beta}}}t_i^{e^{\tilde{\xi}}} & -\frac{\partial^2\ell}{\partial\beta_1\partial\beta_2}\Big|_{\tilde{\boldsymbol{\vartheta}}} & \cdots & \cdots & \vdots\\
\vdots & \vdots & -\frac{\partial^2\ell}{\partial\beta_2\partial\beta_1}\Big|_{\tilde{\boldsymbol{\vartheta}}} & 2\lambda+\sum_{i=1}^{n}x_{i2}^2e^{\boldsymbol{x}_i'\tilde{\boldsymbol{\beta}}}t_i^{e^{\tilde{\xi}}} & \ddots & \cdots & \vdots\\
\vdots & \vdots & \vdots & -\frac{\partial^2\ell}{\partial\beta_3\partial\beta_2}\Big|_{\tilde{\boldsymbol{\vartheta}}} & \ddots & \ddots & \vdots\\
\vdots & \vdots & \vdots & \vdots & \ddots & \ddots & -\frac{\partial^2\ell}{\partial\beta_{p-1}\partial\beta_p}\Big|_{\tilde{\boldsymbol{\vartheta}}}\\
-\frac{\partial^2\ell}{\partial\xi\partial\beta_p}\Big|_{\tilde{\boldsymbol{\vartheta}}} & -\frac{\partial^2\ell}{\partial\beta_p\partial\beta_0}\Big|_{\tilde{\boldsymbol{\vartheta}}} & \cdots\cdots & \cdots & \cdots & -\frac{\partial^2\ell}{\partial\beta_p\partial\beta_{p-1}}\Big|_{\tilde{\boldsymbol{\vartheta}}} & 2\lambda+\sum_{i=1}^{n}x_{i0}^2e^{\boldsymbol{x}_i'\tilde{\boldsymbol{\beta}}}t_i^{e^{\tilde{\xi}}}
\end{pmatrix}$$

in cui è stata riportata esplicitamente la diagonale principale per evidenziare come l'intercetta e il predittore lineare non siano stati soggetti a penalizzazione.

Infine, in virtù della relazione (15), è sufficiente considerare l'inversa dell'hessiana cambiata di segno e valutata in corrispondenza del massimo per ottenere

un'approssimazione della matrice di varianza-covarianza della *proposal distribution* per un valore di λ. Sono quindi stati inizializzati i parametri della *proposal distribution*.

Si osservi tuttavia che risulterebbe computazionalmente oneroso e inefficiente generare realizzazioni dalle *proposal distributions* così specificate. Infatti si noti che per generare ciascun candidato da $\mathcal{N}_{p+2}\left(\tilde{\boldsymbol{\vartheta}}, \frac{2.38^2}{p+2}\boldsymbol{\mathcal{I}}^{-1}\left(\tilde{\boldsymbol{\vartheta}}\right)\right)$ ad ogni iterazione bisognerebbe generarlo come segue:

$$\boldsymbol{\vartheta}^* = \tilde{\boldsymbol{\vartheta}} + \mathbf{A}\boldsymbol{Z},$$

con $\boldsymbol{Z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{p+2})$ normale multivariata standard, e $\mathbf{A}$ matrice di dimensione $(p+2) \times (p+2)$ tale che $\mathbf{A}\mathbf{A}' = \frac{2.38^2}{p+2}\boldsymbol{\mathcal{I}}^{-1}\left(\tilde{\boldsymbol{\vartheta}}\right)$, che esiste certamente in quanto la matrice di varianza-covarianza è simmetrica, quadrata e semidefinita positiva, si veda Wilkinson et al. (2013). Ad ogni iterazione bisognerebbe quindi calcolare la matrice $\mathbf{A}$. Tuttavia, ciò renderebbe l'algoritmo inefficiente dal momento che la matrice $\mathbf{A}$ può essere calcolata al di fuori del ciclo di iterazioni. In questo caso per identificare tale matrice si è scelto di utilizzare la Decomposizione ai Valori Singolari (*Singular Value Decomposition*, SVD). Si poteva anche optare per la Decomposizione Spettrale, che rappresenta un caso particolare della SVD, tuttavia risulta computazionalmente più efficiente procedere con la decomposizione ai valori singolari.

La procedura viene ripetuta per ogni valore di λ giungendo quindi a 100 matrici $\mathbf{A}$ distinte.

4.4.3 Decomposizione ai valori singolari della matrice di varianza-covarianza della *proposal distribution*

Dal momento che la matrice di varianza-covarianza $\boldsymbol{\mathcal{I}}^{-1}\left(\tilde{\boldsymbol{\vartheta}}\right)$ è quadrata, simmetrica e semidefinita positiva, lo è anche la matrice $\tilde{\boldsymbol{\Sigma}} = \frac{2.38^2}{p+2}\boldsymbol{\mathcal{I}}^{-1}\left(\tilde{\boldsymbol{\vartheta}}\right)$. Ricorrendo le ipotesi per l'applicazione del teorema di decomposizione spettrale, per i cui dettagli si rimanda al paragrafo 7.1 in Appendice, è lecita la riscrittura:

$$\begin{aligned}
\tilde{\boldsymbol{\Sigma}} &= \frac{2.38^2}{p+2}\boldsymbol{\mathcal{I}}^{-1}\left(\tilde{\boldsymbol{\vartheta}}\right) \\
&= \frac{2.38^2}{p+2}\mathbf{V}\mathbf{D}^{-1}\mathbf{V}' \\
&= \frac{2.38^2}{p+2}\mathbf{V}\mathbf{D}^{-1/2}\mathbf{V}'\mathbf{V}\mathbf{D}^{-1/2}\mathbf{V}' \\
&= \frac{2.38^2}{p+2}\mathbf{V}\mathbf{D}^{-1/2}\mathbf{D}^{-1/2}\mathbf{V}' = \mathbf{A}\mathbf{A}' \quad \text{con} \quad \mathbf{A} = \frac{2.38}{\sqrt{p+2}}\mathbf{V}\mathbf{D}^{-1/2}.
\end{aligned}$$

Quindi, per la generazione del candidato, sarà sufficiente ricavare prima la matrice $\mathbf{A}$ e durante le iterazioni della procedura MCMC utilizzare la matrice calcolata prima dell'inizio delle iterazioni. Si osservi che su decine di migliaia di iterazioni questo passaggio può condurre a un risparmio di tempo di considerevole rilevanza.

Si è pertanto giunti al termine della procedura necessaria al fine di inizializzare ciascuna delle cento *proposal distribution*.

4.4.4 Pseudo-algoritmo Inizializzazione *proposal*

Algoritmo 1: Log a posteriori e ricerca dell'ottimo per la sequenza di $\boldsymbol{\lambda}$

Function `loglikelihood(`$\boldsymbol{\vartheta}$, $\mathbf{X}$, $\boldsymbol{t}$, $\boldsymbol{\delta}$`)`**:**

  return $\sum_{i=1}^{n} \vartheta_0 \boldsymbol{\delta} + \mathbf{X}' \boldsymbol{\vartheta}_{-0} \boldsymbol{\delta} + \boldsymbol{\delta} e^{\vartheta_0 - 1} \log(\boldsymbol{t}) - e^{\mathbf{X}' \boldsymbol{\vartheta}_{-0}} \boldsymbol{t} e^{\vartheta_0}$

End Function

Function `logprior(`$\boldsymbol{\vartheta}$,σ^2_β, σ^2_β, l`)`**:**

  return $-\frac{\vartheta_0^2}{2\log(\sigma^2_\alpha + 1)} + \frac{\vartheta_0}{2} - l \sum_{j=3}^{p} \vartheta_j^2 - \frac{\vartheta_1^2 + \vartheta_2^2}{2\sigma^2_\beta}$

End Function

Function `logposterior(`$\boldsymbol{\vartheta}$, $\mathbf{X}$, $\boldsymbol{t}$, $\boldsymbol{\delta}$, σ^2_α, σ^2_β, l`)`**:**

  return `loglikelihood(`$\boldsymbol{\vartheta}, \mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta}$`)` +`logprior(`$\boldsymbol{\vartheta}, \sigma^2_\alpha, \sigma^2_\beta, l$`)`

End Function

Function `logpostmin(`$\boldsymbol{\vartheta}$, $\mathbf{X}$, $\boldsymbol{t}$, $\boldsymbol{\delta}$, σ^2_α, σ^2_β, l`)`**:**

  return − `logposterior(`$\boldsymbol{\vartheta}, \mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta}, \sigma^2_\alpha, \sigma^2_\beta, l$`)`

End Function

$a = 10.5$

$b = 5$

$\boldsymbol{\lambda} \leftarrow$ sequenza di lunghezza 100 da e^a a e^b

betamat $\leftarrow$ matrice 100 righe e numero di colonne $= p + 2$ con p numero di esplicative

for $i \leftarrow 0$ **to** 100 **do**

  $\tilde{\boldsymbol{\vartheta}}_{\lambda_i} = \arg\min_{\boldsymbol{\vartheta}}$ `logpostmin(`$\boldsymbol{\vartheta}, \mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta}, \sigma^2_\alpha = 16, \sigma^2_\beta = 600, l = \lambda_i$`)`

  betamat[i,] $\leftarrow \tilde{\boldsymbol{\vartheta}}_{\lambda_i}$

end

Algoritmo 2: Inizializzazione delle *proposal distributions*

Derivazione analitica della hessiana e valutazione nell'ottimo $-\boldsymbol{\mathcal{H}}\Big|_{\boldsymbol{\vartheta}=\tilde{\boldsymbol{\vartheta}}} = \boldsymbol{\mathcal{I}}\left(\tilde{\boldsymbol{\vartheta}}\right)$

Function `Amatrix(`$\boldsymbol{\vartheta}$`,`σ_α^2`, `σ_β^2`, `l`, `$\mathbf{X}$`, `$\boldsymbol{t}$`, `$\boldsymbol{\delta}$`):`

$H \leftarrow$ matrice vuota $(p+2) \times (p+2)$

$H[1,1] \leftarrow \sum_{i=1}^{n} \left[e^{\boldsymbol{x}_i' \tilde{\boldsymbol{\vartheta}}_{-0}} \log t_i \left(e^{\tilde{\vartheta_0}} t_i^{e^{\tilde{\vartheta_0}}} + e^{2\tilde{\vartheta_0}} t_i^{e^{\tilde{\vartheta_0}}} \log t_i \right) - e^{\tilde{\vartheta_0}} \delta_i \log t_i \right] + \frac{1}{\log(\sigma_\alpha^2+1)}$

for $j \leftarrow 2$ **to** $p+2$ **do**

  $H[2:(p+1),1] \leftarrow \sum_{i=1}^{n} x_{ij} e^{\boldsymbol{x}_i' \tilde{\boldsymbol{\vartheta}}_{-0}} t_i^{e^{\tilde{\vartheta}_0}} \log t_i$ vettorizzato per efficienza

end

for $j \leftarrow 2$ **to** $p+2$ **do**

  $diag(H)[j] \leftarrow \sum_{i=1}^{n} \left(x_{ij}^2 e^{\boldsymbol{x}_i' \tilde{\boldsymbol{\vartheta}}_{-0}} t_i^{e^{\tilde{\vartheta}_0}} \right) + 2f(\tilde{\vartheta}_{-0,j})$

end

for $j \leftarrow 2$ **to** $p+2$ **do**

  for $k \leftarrow 1$ **to** $j-1$ **do**

    $H[j,k] \leftarrow \sum_{i=1}^{n} x_{ij} x_{ik} e^{\boldsymbol{x}_i' \tilde{\boldsymbol{\vartheta}}_{-0}} t_i^{e^{\tilde{\vartheta}_0}}$

  end

end

$H_{\boldsymbol{\vartheta}}$ Si ottiene completando, per simmetria, la matrice triangolare superiore

Decomposizione ai valori singolari

$\mathbf{V} \leftarrow \text{SVD}(H_{\boldsymbol{\vartheta}})[u]$ matrice degli autovettori normalizzati

$\mathbf{d} \leftarrow \text{SVD}(H_{\boldsymbol{\vartheta}})[\delta]$ vettore dei valori singolari

$\mathbf{D} = diag\left(1/\sqrt{\mathbf{d}}\right)$

Matrice di Varianza-covarianza

$\tilde{\boldsymbol{\Sigma}} \leftarrow \frac{2.38^2}{p+2} \mathbf{V}\mathbf{D}^2\mathbf{V}'$

Matrice per generare dalla *proposal distribution* in modo efficiente

$\mathbf{A} \leftarrow \frac{2.38}{\sqrt{p+2}} \mathbf{V}\mathbf{D}$

return A

End Function

4.4.5 Pseudo-algoritmo RW Metropolis - Hastings

Algoritmo 3: Random Walk Metropolis - Hastings per ciascun λ

$a = 10.5$
$b = 5$
$\boldsymbol{\lambda} \leftarrow$ sequenza di lunghezza 100 da e^a a e^b
sequenza $\leftarrow$ sequenza da $(burnin + thinning)$ a $(R + burnin)$, con passo pari al *thinning*.

Function `RMH`(*burnin, R,* τ, $\boldsymbol{\vartheta}$,σ^2_α, σ^2_β, $\boldsymbol{\lambda}$, $\mathbf{X}$, $\boldsymbol{t}$, $\boldsymbol{\delta}$)**:**
    out $\leftarrow$ array composto da 100 matrici con numero di righe R/τ e numero di colonne $= p + 2$ con p numero di esplicative e τ *thinning period*

    for $i \leftarrow 1$ **to** *length*$(\boldsymbol{\lambda})$ **do**

        inizializzazione
        $\boldsymbol{\vartheta}$ = betamat$[i,]$
        $\mathbf{A} \leftarrow$ `Amatrix`$(\boldsymbol{\vartheta}, \mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta}, \sigma^2_\alpha = 16, \sigma^2_\beta = 600, l = \lambda_i)$
        logp $\leftarrow$ `logposterior`$(\boldsymbol{\vartheta}, \mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta}, \sigma^2_\alpha = 16, \sigma^2_\beta = 600, l = \lambda_i)$
        index $\leftarrow 0$

        for $r \leftarrow 1$ **to** $burn + R$ **do**

            $\boldsymbol{\vartheta}^* = \boldsymbol{\vartheta} + \mathbf{A}\boldsymbol{Z}$ generazione candidato
            logpnew $\leftarrow$ `logposterior`$(\boldsymbol{\vartheta}^*, \mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta}, \sigma^2_\alpha = 16, \sigma^2_\beta = 600, l = \lambda_i)$
            $\alpha = \min\{\exp(logpnew - logp), 1\}$
            $u \leftarrow u \sim U(0, 1)$

            if $u < \alpha$ **then**
                $\boldsymbol{\vartheta} = \boldsymbol{\vartheta}^*$ accetta il candidato
                logp $\leftarrow$ logpnew
            end
            if $r > burnin$ & $r \in sequenza$ **then**
                $index \leftarrow index + 1$
                out$[index, , i] \leftarrow \boldsymbol{\vartheta}$
            end
        end

    end

    return out
End Function

4.5 MALA pre-condizionato

Come già esplicitato nel paragrafo 3.3.1.7, nei casi ad elevata dimensionalità, per quanto si possa cercare di inizializzare in modo sensato la *proposal distribution* nella sua versione normale, talvolta ciò non è sufficiente a garantire la convergenza in un numero di iterazioni ragionevole da un punto di vista di onere computazionale. Per questo motivo, e tenuto conto dei risultati ottenuti con la specificazione delle *proposal distributions* specificate nei paragrafi precedenti (si veda, a riguardo, la Figura 9 al paragrafo 5.1), si è proceduto specificando anche *proposal distributions* basate su metodi più elaborati ed efficienti, che permettessero di raggiungere la convergenza della catena a parità di numero di iterazioni. A tal fine è stata considerata una versione più elaborata del Metropolis-Hastings (il MALA), capace di sfruttare la salita del gradiente tramite la discretizzazione delle dinamiche di diffusione di Langevin, per far sì che le *proposal distributions* si concentrassero sulle zone a più alta densità della distribuzione a posteriori (per dettagli circa il funzionamento del MALA e la specificazione che verrà ora presentata si rimanda al paragrafo 3.3.1.7).

L'apparato teorico sin qui descritto, la derivazione analitica delle matrici di varianza-covarianza valutate nei relativi punti di ottimo e i medesimi punti di ottimo ottenuti tramite ottimizzazione numerica, facilitano la specificazione delle diverse *proposal distributions*. Infatti, per quanto visto nel paragrafo 3.3.1.7, fissato un valore di λ per semplicità espositiva, (fermo restando che ciò che verrà esposto vale per ogni valore della sequenza dei cento valori di λ), la *proposal distribution* risulta:

$$\boldsymbol{\vartheta}^* \mid \boldsymbol{\vartheta} \sim \mathcal{N}_{p+2}\left(\tilde{\boldsymbol{\vartheta}} + \frac{\epsilon^2}{2(p+2)^{\frac{1}{3}}}\boldsymbol{\Sigma}\nabla_{\boldsymbol{\vartheta}} \log \pi(\boldsymbol{\vartheta} \mid \mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta}) \ , \ \frac{\epsilon^2}{(p+2)^{\frac{1}{3}}}\boldsymbol{\Sigma}\right),$$

dove il gradiente valutato nell'ottimo ha forma:

$$\nabla_{\boldsymbol{\vartheta}} \log \pi(\boldsymbol{\vartheta} \mid \mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta})\Big|_{\boldsymbol{\vartheta}=\tilde{\boldsymbol{\vartheta}}} = \begin{pmatrix} \sum_{i=1}^n \Big[\delta_i + e^{\tilde{\xi}} \log t_i \Big(\delta_i - e^{\boldsymbol{x}_i'\tilde{\boldsymbol{\beta}} t_i^{e^{\tilde{\xi}}}} \Big)\Big] - \frac{\tilde{\xi}}{\log(\sigma_\alpha^2+1)} - \frac{1}{2} \\ \sum_{i=1}^n \Big(\delta_i x_{i0} - x_{i0} e^{\boldsymbol{x}_i'\tilde{\boldsymbol{\beta}}} t_i^{e^{\tilde{\xi}}}\Big) - \frac{\tilde{\beta}_0}{\sigma_{\boldsymbol{\beta}}^2} \\ \sum_{i=1}^n \Big(\delta_i x_{i1} - x_{i1} e^{\boldsymbol{x}_i'\tilde{\boldsymbol{\beta}}} t_i^{e^{\tilde{\xi}}}\Big) - \frac{\tilde{\beta}_1}{\sigma_{\boldsymbol{\beta}}^2} \\ \sum_{i=1}^n \Big(\delta_i x_{i2} - x_{i2} e^{\boldsymbol{x}_i'\tilde{\boldsymbol{\beta}}} t_i^{e^{\tilde{\xi}}}\Big) - 2\lambda\tilde{\beta}_2 \\ \vdots \\ \sum_{i=1}^n \Big(\delta_i x_{ij} - x_{ij} e^{\boldsymbol{x}_i'\tilde{\boldsymbol{\beta}}} t_i^{e^{\tilde{\xi}}}\Big) - 2\lambda\tilde{\beta}_j \\ \vdots \\ \sum_{i=1}^n \Big(\delta_i x_{ip} - x_{ip} e^{\boldsymbol{x}_i'\tilde{\boldsymbol{\beta}}} t_i^{e^{\tilde{\xi}}}\Big) - 2\lambda\tilde{\beta}_p \end{pmatrix},$$

con $\boldsymbol{\Sigma} \approx \tilde{\boldsymbol{\Sigma}} = \boldsymbol{\mathcal{I}}^{-1}(\tilde{\boldsymbol{\vartheta}})$ e ϵ parametro soggetto a *tuning*. Si ricordi che specificare scorrettamente ϵ può tradursi nell'assenza di convergenza, pertanto ϵ dovrebbe essere specificato in modo che il tasso di accettazione, che dipende dalla velocità di diffusione, sia circa pari a 57.4%, si veda a riguardo il paragrafo 3.3.1.7. La *proposal distribution* risulta:

$$\boldsymbol{\vartheta}^* \mid \boldsymbol{\vartheta} \sim \mathcal{N}_{p+2}\left(\tilde{\boldsymbol{\vartheta}} + \frac{\epsilon^2}{2(p+2)^{\frac{1}{3}}}\boldsymbol{\mathcal{I}}^{-1}(\tilde{\boldsymbol{\vartheta}})\nabla_{\boldsymbol{\vartheta}} \log \pi(\boldsymbol{\vartheta} \mid \mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta})\Big|_{\boldsymbol{\vartheta}=\tilde{\boldsymbol{\vartheta}}} \ , \ \frac{\epsilon^2}{(p+2)^{\frac{1}{3}}}\boldsymbol{\mathcal{I}}^{-1}(\tilde{\boldsymbol{\vartheta}})\right),$$

in cui tutte le quantità coinvolte sono già state definite nei paragrafi del Capitolo 4 e dunque, la *proposal distribution* risulta già inizializzata. Fa eccezione ϵ, per la

cui specificazione si procederà secondo un paradigma di aggiustamento, solo nella fase di *burn-in*, decrescente al crescere del numero di iterazioni.

Anche in questo caso la matrice di varianza-covarianza è data da $\boldsymbol{\Sigma} \approx \tilde{\boldsymbol{\Sigma}} = \boldsymbol{\mathcal{I}}^{-1}(\tilde{\boldsymbol{\vartheta}})$ e, sempre in virtù del teorema di decomposizione spettrale, è scrivibile come:

$$\boldsymbol{\Sigma} \approx \tilde{\boldsymbol{\Sigma}} = \boldsymbol{\mathcal{I}}^{-1}(\tilde{\boldsymbol{\vartheta}}) = \mathbf{V}\mathbf{D}^{-1/2}\mathbf{D}^{-1/2}\mathbf{V}',$$

sicché per generare da una normale $(p+2)$-variata, con media

$$\tilde{\boldsymbol{\vartheta}} + \frac{\epsilon^2}{(p+2)^{\frac{1}{3}}}\boldsymbol{\mathcal{I}}^{-1}(\tilde{\boldsymbol{\vartheta}})\nabla_{\boldsymbol{\vartheta}} \log \pi(\boldsymbol{\vartheta} \mid \mathbf{X})\Big|_{\boldsymbol{\vartheta}=\tilde{\boldsymbol{\vartheta}}},$$

è possibile generare dalla *proposal distribution* senza dover invertire, per ogni iterazione, l'intera matrice di varianza-covarianza, attraverso la matrice:

$$\mathbf{A} = \mathbf{V}\mathbf{D}^{-1/2}.$$

Donde, il generico candidato si genera facilmente dall'equazione:

$$\boldsymbol{\vartheta}^{*(j)} = \boldsymbol{\vartheta}^{(j-1)} + \frac{\epsilon^2}{(p+2)^{\frac{1}{3}}}\nabla_{\boldsymbol{\vartheta}} \log \pi(\boldsymbol{\vartheta} \mid \mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta})\Big|_{\boldsymbol{\vartheta}=\boldsymbol{\vartheta}^{(j-1)}} + \frac{\epsilon}{(p+2)^{\frac{1}{6}}}\mathbf{A}\mathbf{Z}. \qquad (16)$$

Si osservi che, in ciascuna iterazione è oggetto di aggiornamento anche il gradiente. Infine, al contrario della *proposal distribution* specificata nel caso del RW MH, nel MALA pre-condizionato la *proposal distribution* è manifestamente non simmetrica. Pertanto la probabilità di accettazione ha forma:

$$\alpha\left(\boldsymbol{\vartheta}, \boldsymbol{\vartheta}^*\right) = \min\left\{\frac{\pi(\boldsymbol{\vartheta}^* \mid \mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta})q\left(\boldsymbol{\vartheta}^*, \boldsymbol{\vartheta}\right)}{\pi(\boldsymbol{\vartheta} \mid \mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta})q\left(\boldsymbol{\vartheta}, \boldsymbol{\vartheta}^*\right)}, 1\right\}.$$

Si ricordi che ϵ è un parametro ignoto. Dunque, per rendere operativa l'equazione (16) è necessario specificare un valore di ϵ. In contesti a bassa dimensionalità ϵ può essere stabilito tramite un processo di *trial and errors*. Infatti, sapendo che il tasso di accettazione asintotico ottimale per il MALA è di circa il 57.4%, è possibile procedere con diverse prove fino a giungere a un valore di ϵ che faccia registrare un tasso di accettazione sufficientemente prossimo a quello ottimale. In questo caso, il problema oggetto d'analisi richiederebbe di effettuare, per ogni valore di λ specificato, innumerevoli prove per raggiungere dei risultati soddisfacenti, risulta dunque chiaro che non è possibile procedere tramite un processo di *trial and errors* in contesti ad alta dimensionalità quale quello oggetto d'analisi.

Una possibilità alternativa è quella di implementare un algoritmo rientrante nella classe degli algoritmi adattivi che, stabilito un numero di iterazioni, aggiorna ogni k iterazioni il valore di ϵ in modo che il tasso di accettazione alle iterazioni successive si "adatti" o, in altri termini, converga al valore ottimale. Sono state proposte diverse versioni dell'*Adaptive Metropolis*, estendibili al MALA, che sfruttano l'aggiornamento ogni k passi del valore di ϵ in funzione del tasso di accettazione registrato nelle k iterazioni precedenti per migliorare quello delle iterazioni successive, per dettagli si veda Roberts and Rosenthal (2009). Queste strategie nascono per fornire una stima della matrice di varianza-covarianza a ogni iterazione nel caso di un RW MH. Nulla vieta, tuttavia, di integrarle all'interno del MALA, in modo da ottenere un *Adaptive*-MALA pre-condizionato. Denominiamo tale variante del MALA come A-MALA *pre-conditioned*. Si potrebbe obiettare che, in questo modo, si altera il processo generativo della catena e dunque non è assicurata la convergenza. I medesimi autori, a riguardo, dimostrano che, se ϵ viene aggiornato con valori decrescenti all'aumentare del numero di iterazioni, la

convergenza garantita dal teorema ergodico viene preservata. Un adattamento decrescente di ϵ è noto col nome di *diminishing adaptation*. Se al crescere del numero di iterazioni l'aggiornamento di ϵ diventa marginale allora la convergenza è garantita.

Si osservi tuttavia che gli autori, sotto l'ipotesi di *diminishing adaptation*, dimostrano che viene preservata la proprietà ergodica, e che quindi la catena converge, ma non assicurano che converga alla distribuzione stazionaria di interesse. In altri termini, il *mixing* potrebbe suggerire l'avvenuta convergenza, tuttavia tale convergenza potrebbe essere avvenuta a una distribuzione diversa rispetto alla distribuzione a posteriori.

In virtù di tale considerazione, si è ritenuto opportuno adottare una strategia di *tuning* adattiva del valore di ϵ solo per quei valori di $\boldsymbol{\vartheta}$ che vengono scartati nella fase di *burn-in* in modo tale che, terminata tale fase, termini con essa anche il *tuning* di ϵ, così da non applicare l'algoritmo adattivo anche alle iterazioni di campionamento effettivo e da non alterare la convergenza alla distribuzione stazionaria.

Inoltre, tenuto conto del fatto che è ragionevole aspettarsi che al variare di λ, in particolare per valori più piccoli di λ, il numero di iterazioni necessarie affinché ϵ si stabilizzi aumenti, si è proceduto rendendo adattivo anche il numero di iterazioni del *burn-in* stesso. In particolare, stabilito un valore di *tolerance* $\iota = 0.054$, si è introdotta una *stopping rule* che fermasse le iterazioni di *burn-in* solo se venivano soddisfatte, congiuntamente, le seguenti condizioni:

1. il numero di iterazioni di *burn-in* risulta non inferiore a 40000;
2. ϵ rimane costante, dunque non viene aggiornato, per 500 iterazioni consecutive;

3. se le condizioni 1 e 2 non vengono soddisfatte si arresta il periodo di *burn-in* a 150000 iterazioni.

Si osservi che 50 è stato fissato come numero di iterazioni dopo le quali verificare se aggiornare il parametro ϵ. Inoltre, il criterio tramite il quale verificare se, ogni 50 iterazioni, fosse opportuno aggiornare ϵ, è stato implementato nel modo seguente: sia r l'r-esima iterazione di *burn-in* e sia $\epsilon^{(j-1)}$ il valore di ϵ derivante dal precedente passo di aggiornamento, allora:

$$\begin{cases} \epsilon^{(j)} = \epsilon^{(j-1)} + \min\left(0.01, \frac{1}{\sqrt{r_j}}\right) & \text{se} \quad \epsilon^{(j)} \notin [0.574, 0.547 + \iota] \\ \epsilon^{(j)} = \epsilon^{(j-1)} - \min\left(0.01, \frac{1}{\sqrt{r_j}}\right) & \text{se} \quad \epsilon^{(j)} \notin [0.547 - \iota, 0.547) \\ \epsilon^{(j)} = \epsilon^{(j-1)} & \text{se} \quad \epsilon^{(j)} \in [0.547 - \iota, 0.547 + \iota] \end{cases} .$$

La *stopping rule* richiede, per arrestare la fase di *burn-in*, che il numero di iterazioni risulti maggiore di 40000 e inferiore a 150000 e che per 500 iterazioni il valore di ϵ rimanga costante o, equivalentemente, che 10 valori consecutivi di ϵ siano uguali. Si osservi inoltre che l'aggiornamento è decrescente con il numero di iterazioni, quindi nella fase di *burn-in* è pienamente rispettata l'ipotesi di *diminishing adaptation*.

Si osservi infine che adottare questa strategia per l'ottimizzazione di ϵ riduce di molto l'onere computazionale, dal momento che per la maggior parte dei valori di λ sono sufficienti 60000 iterazioni di *burn-in*.

Tale procedura viene ripetuta per ogni valore di λ specificato, solo nella fase di *burn-in*. Pertanto, l'ottimizzazione di ϵ sarà sicuramente sotto-performante rispetto agli usuali algoritmi adattivi che modificano ϵ anche nella fase di campionamento vera e propria, tuttavia, sebbene si abbia maggior variabilità intorno al tasso di accettazione, ciò viene compensato dalla convergenza alla vera distribuzione

stazionaria in fase di campionamento. Infatti dopo il periodo di *burn-in* non viene più aggiornato ϵ.

Dalle osservazioni sin qui fatte si può pertanto concludere che, la procedura di ottimizzazione di ϵ per ogni λ preserva la proprietà di ergodicità e dunque la convergenza nella fase di *burn-in* e, al tempo stesso, non altera la struttura della catena nella fase di campionamento, durante la quale ϵ risulta fissato. Si sacrifica parzialmente l'accuratezza a livello di ottimizzazione di ϵ per il tasso di convergenza asintotico ottimale, ma si preserva l'integrità strutturale e la convergenza alla distribuzione stazionaria della catena di Markov in fase di campionamento. Infine, il punto ϵ^0 di inizializzazione è stato posto pari a 1. Il numero di iterazioni di campionamento è stato fissato pari a 75000 con un *thinning period* di 15, in modo da decorrelare parzialmente i candidati e ottenere un campione di ampiezza 5000 per ciascun vettore di parametri, e per ciascun λ.

Definiti i dettagli teorici dell'algoritmo che verrà implementato si procede fornendo lo pseudocodice nel caso dell' A-MALA pre-condizionato con ottimizzazione del parametro di *step-size*: ϵ.

Per quanto riguarda l'algoritmo che determina l'ottimo a priori per ogni valore di λ si faccia pure riferimento all'Algoritmo 1, in quanto non sono presenti variazioni.

4.5.1 Pseudo Algoritmo: A-MALA pre-condizionato

Algoritmo 4: Matrici di Varianza-Covarianza e Gradiente

Derivazione analitica della hessiana e valutazione nell'ottimo $-\boldsymbol{\mathcal{H}}\Big|_{\boldsymbol{\vartheta}=\tilde{\boldsymbol{\vartheta}}} = \boldsymbol{\mathcal{I}}(\tilde{\boldsymbol{\vartheta}})$

Function `Amatrix`($\boldsymbol{\vartheta}$,σ_α^2, σ_β^2, l, $\mathbf{X}$, $\boldsymbol{t}$, $\boldsymbol{\delta}$)**:**

$H \leftarrow$ matrice vuota $(p+2)\times(p+2)$

$H[1,1] \leftarrow \sum_{i=1}^{n}\left[e^{\boldsymbol{x}_i'\tilde{\boldsymbol{\vartheta}}_{-0}}\log t_i\left(e^{\tilde{\vartheta_0}}t_i^{e^{\tilde{\vartheta_0}}}+e^{2\tilde{\vartheta_0}}t_i^{e^{\tilde{\vartheta_0}}}\log t_i\right)-e^{\tilde{\vartheta_0}}\delta_i\log t_i\right]+\frac{1}{\log(\sigma_\alpha^2+1)}$

for $j \leftarrow 2$ **to** $p+2$ **do**

  $H[2:(p+1),1] \leftarrow \sum_{i=1}^{n} x_{ij}e^{\boldsymbol{x}_i'\tilde{\boldsymbol{\vartheta}}_{-\mathbf{0}}+\tilde{\vartheta_0}}t_i^{e^{\tilde{\vartheta}_0}}\log t_i$ vettorizzato per efficienza

end

for $j \leftarrow 2$ **to** $p+2$ **do**

  $diag(H)[j] \leftarrow \sum_{i=1}^{n}\left(x_{ij}^2e^{\boldsymbol{x}_i'\tilde{\boldsymbol{\vartheta}}_{-0}}t_i^{e^{\tilde{\vartheta}_0}}\right)+2f(\tilde{\vartheta}_{-0,j})$

end

for $j \leftarrow 2$ **to** $p+2$ **do**

  for $k \leftarrow 1$ **to** $j-1$ **do**

    $H[j,k] \leftarrow \sum_{i=1}^{n} x_{ij}x_{ik}e^{\boldsymbol{x}_i'\tilde{\boldsymbol{\vartheta}}_{-0}}t_i^{e^{\tilde{\vartheta}_0}}$

  end

end

$H_{\boldsymbol{\vartheta}}$ Si ottiene completando, per simmetria, la matrice triangolare superiore

$\mathbf{V} \leftarrow \mathrm{SVD}(H_{\boldsymbol{\vartheta}})[u]$ $\mathbf{d} \leftarrow \mathrm{SVD}(H_{\boldsymbol{\vartheta}})[\delta]$; $\mathbf{D} = diag\left(1/\sqrt{\mathbf{d}}\right)$ **SVD**

$\tilde{\boldsymbol{\Sigma}} \leftarrow \mathbf{VD}^2\mathbf{V}'$; $\mathbf{A} \leftarrow \mathbf{VD}$

return A

End Function

Derivazione analitica del gradiente e valutazione nell'ottimo $\nabla_{\boldsymbol{\vartheta}}\log\pi(\boldsymbol{\vartheta}\mid\mathbf{X},\boldsymbol{t},\boldsymbol{\delta})\Big|_{\boldsymbol{\vartheta}=\tilde{\boldsymbol{\vartheta}}}$

Function `lgrad`($\boldsymbol{\vartheta}$,σ_α^2, σ_β^2, l, $\mathbf{X}$, $\boldsymbol{t}$, $\boldsymbol{\delta}$)**:**

$\mathrm{gr}[1] \leftarrow \sum_{i=1}^{n}[\delta_i+\log t_i(\,e^{\tilde{\vartheta_0}}\delta_i-e^{\boldsymbol{x}_i'\tilde{\boldsymbol{\vartheta}}_{-0}}t_i^{e^{\tilde{\vartheta_0}}}e^{\tilde{\vartheta_0}}\,)]-\frac{\tilde{\vartheta_0}}{\log(\sigma_\alpha^2+1)}-\frac{1}{2}$

$\mathrm{gr}[2] \leftarrow \sum_{i=1}^{n}\left(\delta_i x_{i0}-x_{i0}e^{\boldsymbol{x}_i'\tilde{\boldsymbol{\vartheta}}_{-0}}t_i^{e^{\tilde{\vartheta_0}}}\right)-\frac{\tilde{\vartheta_1}}{\sigma_{\boldsymbol{\beta}}^2}$

$\mathrm{gr}[3] \leftarrow \sum_{i=1}^{n}\left(\delta_i x_{i1}-x_{i1}e^{\boldsymbol{x}_i'\tilde{\boldsymbol{\vartheta}}_{-0}}t_i^{e^{\tilde{\vartheta_0}}}\right)-\frac{\tilde{\vartheta_2}}{\sigma_{\boldsymbol{\beta}}^2}$

for $j \leftarrow 4$ **to** p **do**

  $\mathrm{gr}[j] \leftarrow \sum_{i=1}^{n}\left(\delta_i x_{ij}-x_{ij}e^{\boldsymbol{x}_i'\tilde{\boldsymbol{\vartheta}}_{-0}}t_i^{e^{\tilde{\vartheta_0}}}\right)-2\lambda\tilde{\vartheta}_{-0,j}$

end

return gr

End Function

Algoritmo 5: Pseudo-algoritmo A-MALA pre-condizionato con *tuning*

Function `MALA`(*burnin*, R, τ, $\boldsymbol{\vartheta}$,σ^2_α, σ^2_β,$\boldsymbol{\lambda}$, $\mathbf{X}$, $\boldsymbol{t}$, $\boldsymbol{\delta}$, *bat*, *target* $= 0.574$, *tolerance* $= 0.054$)**:**

- out $\leftarrow$ array: 100 matrici: righe= R/τ, colonne $= p+2$. p numero variabili; τ *thinning period*
- **for** $i \leftarrow 1$ **to** *length*($\boldsymbol{\lambda}$) **do**
  - $\boldsymbol{\vartheta}$ = betamat$[i,]$
  - $\mathbf{A} \leftarrow$ `Amatrix`$(\boldsymbol{\vartheta}, \mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta}, \sigma^2_\alpha, \sigma^2_\beta, l = \lambda_i)$
  - $\mathbf{S} \leftarrow \mathbf{A}\mathbf{A}'$; $\mathbf{S} \leftarrow \mathbf{S}^{-1}$
  - logp $\leftarrow$ `logposterior`$(\boldsymbol{\vartheta}, \mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta}, \sigma^2_\alpha, \sigma^2_\beta, l = \lambda_i)$
  - lgradiente $\leftarrow \mathbf{S}$ `lgrad`$(\boldsymbol{\vartheta}, \mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta}, \sigma^2_\alpha, \sigma^2_\beta, l = \lambda_i)$
  - accepted $\leftarrow 0$; *batch* $\leftarrow 1$;
  - index $\leftarrow 0$; $\epsilon \leftarrow 1$; $j \leftarrow 0$; $r \leftarrow 0$
  - **while** $r \leq burnin + R$ **do**
    - $r \leftarrow r + 1$
    - **if** *batch* $=$ *bat* & $r < burnin + 1$ **then**
      - $j \leftarrow j + 1$
      - **if** *accepted* /*bat* $>$ *target* $+$ *tolerance* **then**
        - $\epsilon \leftarrow \epsilon + \min\left(0.01, \sqrt{\frac{1}{r}}\right)$
      - **end**
      - **if** *accepted* /*bat* $<$ *target* $-$ *tolerance* **then**
        - $\epsilon \leftarrow \epsilon - \min\left(0.01, \sqrt{\frac{1}{r}}\right)$
      - **end**
      - `adaptivemonitoring` $[j, i] \leftarrow \epsilon$; *batch* $\leftarrow 0$; accepted $\leftarrow 0$
      - **if** $r \geq 40000$ & `all`$((\epsilon =$ `adaptivemonitoring` $[(j - r + 1) : j, i]) =$ *TRUE* **then**
        - $r \leftarrow burnin$
      - **end**
    - **end**
    - *batch* $\leftarrow$ *batch* $+ 1$
    - $\boldsymbol{\vartheta}^* \leftarrow \boldsymbol{\vartheta} + \frac{\epsilon^2}{2(p+2)^{1/3}}$ lgradiente $\sqrt{\frac{\epsilon^2}{2(p+2)^{1/3}}}\mathbf{AZ}$
    - $\text{logp}_{new} \leftarrow$ `logposterior`$(\boldsymbol{\vartheta}^*, \mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta}, \sigma^2_\alpha = 16, \sigma^2_\beta = 600, l = \lambda_i)$
    - $\text{lgradiente}_{new} \leftarrow \mathbf{S}$ `lgrad`$(\boldsymbol{\vartheta}^*, \mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta}, \sigma^2_\alpha = 16, \sigma^2_\beta = 600, l = \lambda_i)$
    - $\text{diff}_{old} \leftarrow \boldsymbol{\vartheta} - \boldsymbol{\vartheta^*} - \frac{\epsilon^2}{2(p+2)^{1/3}}$ lgradiente_{new}
    - $\text{diff}_{new} \leftarrow \boldsymbol{\vartheta^*} - \boldsymbol{\vartheta} - \frac{\epsilon^2}{2(p+2)^{1/3}}$ lgradiente
    - $\text{q}_{old} \leftarrow \text{diff}_{old}\ \mathbf{S}^{-1}\ \text{diff}_{old}\ \frac{(p+2)^{1/3}}{\epsilon^2}$
    - $\text{q}_{new} \leftarrow \text{diff}_{new}\mathbf{S}^{-1}\ \text{diff}_{new}\frac{(p+2)^{1/3}}{\epsilon^2}$
    - $\alpha \leftarrow \min\ (1, \exp(\text{logp}_{new} - \text{logp} + \text{q}_{old} - \text{q}_{new}\ ))$
    - $u \leftarrow u \sim U(0, 1)$
    - **if** $u < \alpha$ **then**
      - logp $\leftarrow \text{logp}_{new}$
      - lgradiente $\leftarrow \text{lgradiente}_{new}$
      - $\boldsymbol{\vartheta} = \boldsymbol{\vartheta}^*$ si accetta il valore
      - accepted $\leftarrow$ accepted+1 si aggiorna il contatore dei valori accettati
    - **end**
    - **if** $r >$ *burnin* & $r \in$ *sequenza* **then**
      - *index* $\leftarrow$ *index* $+ 1$
      - out$[index, , i] \leftarrow \boldsymbol{\vartheta}$
    - **end**
  - **end**
- **end**
- **return** out

End Function

4.6 Confronto tra modelli

Una volta completato il campionamento dei coefficienti per i cento valori di λ specificati si ottengono cento distinti modelli. Si pone pertanto il problema di capire quale, fra i cento modelli, rappresenti il miglior modello a posteriori, oppure come combinare insieme le informazioni raccolte dai diversi modelli. A tal fine è possibile procedere per via esatta o per via approssimata. Dopo aver fornito una disamina relativa al metodo esatto ci si concentrerà maggiormente sul metodo approssimato, procedendo sotto l'assunzione che tale metodo, se ben implementato, possa portare a risultati molto simili a quello esatto. Il motivo per cui si procede per via approssimata risiede nell'onere computazionale imputabile al campionamento: se per la fase di stima dei parametri risulta sufficiente e anzi, abbondante, un campione di 5000 realizzazione per parametro, in fase di stima della verosimiglianza marginale, necessaria per procedere per via esatta, il campione richiesto è molto più ampio. Si consideri che suddividendo il campionamento per blocchi di λ, non è stato sufficiente nemmeno un campione di 50000 realizzazioni per parametro, campionato su 500000 iterazioni con un *thinning period* pari a 10. Si consideri inoltre che realizzare tale campionamento per ogni valore di λ ha comportato un onere computazionale considerevole.

Per questo motivo, invece che procedere per forza bruta, si è preferito adottare una procedura che, a costo di considerazioni teoriche aggiuntive utili a comprenderne la validità, potesse consentire di procedere più agevolmente con la parte inferenziale.

Ciò premesso, verrà quindi identificato il miglior modello, verrà costruito un ulteriore modello attraverso una mistura dei cento modelli con pesi pari alla

probabilità a posteriori associata a ciascun modello, che verrà stimata per via approssimata per l'impossibilità di procedere per via esatta, ma anche per motivi che verranno enucleati nel paragrafo 4.6.2. In altri termini, il metodo approssimato si presenta in prima istanza come unica via percorribile, tuttavia presenta vantaggi non indifferenti dal punto di vista delle proprietà che possono esserne derivate.

Si osservi che la strategia di costruzione del modello mistura fonda le basi sull'idea che, sebbene per λ non sia stata specificata nessuna distribuzione a priori, l'intero modello possa essere visto come un modello gerarchico, in cui il modello stesso, $\mathcal{M}_\lambda$, diviene una variabile aleatoria. Da un punto di vista bayesiano tale approccio è ragionevole: l'incertezza associata alla scelta del miglior modello induce a ritenere il modello, correttamente, fonte di aleatorietà e per questo viene considerato anch'esso una variabile casuale a cui viene assegnata una distribuzione a priori discreta uniforme del tipo $\mathcal{M}_\lambda \sim U(0,1)$. Imporre tale distribuzione a priori, di fatto, equivale a non privilegiare nessun modello a priori, cosa ragionevole dal momento che, non avendo alcuna informazione a priori circa la bontà dei modelli essi risultano equiprobabili. La strategia usata per costruire tale modello è anche nota con il nome di *Bayesian Model Averaging*, per dettagli si veda Raftery et al. (1997) e Wasserman (2000) e il paragrafo 4.7 in cui viene fornita un'attenta disamina del caso oggetto d'analisi.

Anzitutto si fornisce la seguente:

Definizione 4.6.1 (Fattore di Bayes (FB)) *Siano $\mathcal{M}_i, \mathcal{M}_j$ due modelli, e siano $\mathbf{X}$ la matrice dei dati osservati, $\boldsymbol{t}$ il vettore dei tempi osservati e $\boldsymbol{\delta}$ il vettore degli indici di evento e censura. Si definisce Fattore di Bayes tra il $\mathcal{M}_i$ e $\mathcal{M}_j$ il*

rapporto tra le verosimiglianze marginali dei modelli:

$$FB_{ij} = \frac{\mathscr{L}(\mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta} \mid \mathcal{M}_i)}{\mathscr{L}(\mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta} \mid \mathcal{M}_j)}$$

Si analizzerà in prima istanza il metodo utile ad estrarre le probabilità che ciascun modello sia quello ottimo a posteriori per via esatta, basato sul *Bridge Sampling*. Si considererà poi un'altra possibile strategia, che approssima la probabilità a posteriori del modello con il WAIC (*Watanable-Akaike Informative Criteria*). Si noti che, al contrario del campionamento dalla distribuzione a posteriori, per il quale non risulta strettamente necessario conoscere la costante di normalizzazione $m(\mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta})$, essa risulta invece cruciale quando si vuole effettuare un confronto tra modelli attraverso il Fattore di Bayes.

Una via analoga per procedere al confronto tra modelli è quella di considerare la probabilità a posteriori di ciascun modello, si ha:

$$\underbrace{\pi(\mathcal{M}_\lambda \mid \mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta})}_{\text{prob. a posteriori del modello}} = \underbrace{\frac{\mathscr{L}(\mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta} \mid \mathcal{M}_\lambda)}{\sum_\lambda \mathscr{L}(\mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta} \mid \mathcal{M}_\lambda)\pi(\mathcal{M}_\lambda)}}_{\text{fattore di aggiornamento}} \times \underbrace{\pi(\mathcal{M}_\lambda)}_{\text{prob. a priori del modello}} . \tag{17}$$

Nuovamente però, per giungere a $m(\mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta})$, bisognerebbe risolvere il seguente integrale:

$$\int_\Theta \pi(\boldsymbol{\vartheta} \mid \mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta}) \mathrm{d}\boldsymbol{\vartheta} = \int_{\Theta \subseteq \mathbb{R}^{p+2}} \mathscr{L}(\mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta} \mid \boldsymbol{\vartheta}) \pi(\boldsymbol{\vartheta}) \mathrm{d}\boldsymbol{\vartheta},$$

che è un integrale in $p + 2 = 2147$ variabili, per il quale persiste la difficoltà di calcolo e di approssimazione anche solo per via numerica, esattamente come nel caso dell'approssimazione della distribuzione a posteriori. Da qui l'esigenza e la rilevanza dei metodi che verranno enucleati nei paragrafi 4.6.1 e 4.6.3. Si osservi

che l'integrale in oggetto rappresenta proprio l'integrale della verosimiglianza marginale.

4.6.1 *Bridge Sampling*

Il *Bridge Sampling* è un metodo di stima delle costanti di normalizzazioni che può essere visto come una generalizzazione dei metodi di stima più semplici usualmente utilizzati, come, a titolo di esempio, lo stimatore Monte Carlo, o il MCIS. La principale differenza tra tali metodi risiede nell'oggetto di stima e nel modo in cui si giunge ad essa: se per le tecniche di stima più semplici viene effettuato il campionamento a partire da una distribuzione di interesse, il *Bridge Sampling* necessita di stimare due distinte costanti di normalizzazione per stimare il Fattore di Bayes, dal momento che esso è composto dal rapporto tra due densità le cui costanti di normalizzazione sono ignote, essendo ignote quelle delle distribuzioni a posteriori. Pertanto, il *Bridge sampling* giunge alla stima delle due costanti attraverso il campionamento da due distinte distribuzioni. Il problema, in questo caso, risiede nell'accuratezza della stima: tanto più le due distribuzioni di interesse sono simili tanto più accurata risulta la stima. Pertanto, il fulcro per raggiungere un buon grado di accuratezza è campionare una costante di normalizzazione alla volta, usando come seconda distribuzione una *proposal distribution* che sia simile alla distribuzione di interesse. Rimane da capire quali distribuzioni siano le più appropriate al fine di applicare il *Bridge Sampling*.

Siano, quindi, $h(\boldsymbol{\vartheta})$ una funzione detta *bridge function* e $g(\boldsymbol{\vartheta})$ la densità della *proposal distribution* scelta in modo opportuno come seconda distribuzione, allora:

$$1 = \frac{\int_\Theta \mathscr{L}(\mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta} \mid \boldsymbol{\vartheta}) m(\mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta}) h(\boldsymbol{\vartheta}) g(\boldsymbol{\vartheta}) \mathrm{d}\boldsymbol{\vartheta}}{\int_\Theta \mathscr{L}(\mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta} \mid \boldsymbol{\vartheta}) \pi(\boldsymbol{\vartheta}) h(\boldsymbol{\vartheta}) g(\boldsymbol{\vartheta}) \mathrm{d}\boldsymbol{\vartheta}}, \qquad \text{da cui}$$

$$m(\mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta}) = \frac{\int_\Theta \mathscr{L}(\mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta} \mid \boldsymbol{\vartheta}) \pi(\boldsymbol{\vartheta}) h(\boldsymbol{\vartheta}) g(\boldsymbol{\vartheta}) \mathrm{d}\boldsymbol{\vartheta}}{\int_\Theta h(\boldsymbol{\vartheta}) g(\boldsymbol{\vartheta}) \pi(\boldsymbol{\vartheta} \mid \mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta}) \mathrm{d}\boldsymbol{\vartheta}}$$

$$= \frac{\mathbb{E}_{g(\boldsymbol{\vartheta})} \left[\mathscr{L}(\mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta} \mid \boldsymbol{\vartheta}) \pi(\boldsymbol{\vartheta}) h(\boldsymbol{\vartheta}) \right]}{\mathbb{E}_{\pi(\boldsymbol{\vartheta} \mid \mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta})} \left[g(\boldsymbol{\vartheta}) h(\boldsymbol{\vartheta}) \right]}$$

donde

$$m(\mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta}) = \frac{\mathbb{E}_{g(\boldsymbol{\vartheta})} \left[\mathscr{L}(\mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta} \mid \boldsymbol{\vartheta}) \pi(\boldsymbol{\vartheta}) h(\boldsymbol{\vartheta}) \right]}{\mathbb{E}_{\pi(\boldsymbol{\vartheta} \mid \mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta})} \left[g(\boldsymbol{\vartheta}) h(\boldsymbol{\vartheta}) \right]} \approx \frac{\frac{1}{n_2} \sum_{i=1}^{n_2} \mathscr{L}(\mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta} \mid \tilde{\boldsymbol{\vartheta}}_i) \pi(\tilde{\boldsymbol{\vartheta}}_i) h(\tilde{\boldsymbol{\vartheta}}_i)}{\frac{1}{n_1} \sum_{j=1}^{n_1} g(\boldsymbol{\vartheta}_i^*) h(\boldsymbol{\vartheta}_i^*)}. \tag{18}$$

Per stimare $m(\mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta})$, quindi, è necessario estrarre n_1 campioni dalla distribuzione a posteriori, $\boldsymbol{\vartheta}_1^*, \boldsymbol{\vartheta}_2^*, \ldots, \boldsymbol{\vartheta}_{n_1}^*$, e n_2 campioni dalla *proposal distribution*, $\tilde{\boldsymbol{\vartheta}}_1, \tilde{\boldsymbol{\vartheta}}_2, \ldots, \tilde{\boldsymbol{\vartheta}}_{n_2}$. Dunque occorre specificare in modo oculato la *bridge function* e la *proposal distribution*. A tal fine, si considera come *bridge function* la funzione proposta da Meng and Wong (1996, p.831-860):

$$h(\boldsymbol{\vartheta}) = \frac{C}{s_1 \mathscr{L}(\mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta} \mid \boldsymbol{\vartheta}) \pi(\boldsymbol{\vartheta}) + s_2 m(\mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta}) g(\boldsymbol{\vartheta})}, \tag{19}$$

dove $s_1 = \frac{n_1}{n_1+n_2}$, $s_2 = \frac{n_2}{n_1+n_2}$, con C una generica costante, il cui calcolo non è richiesto in quanto la *bridge function* compare sia a numeratore che a denominatore del rapporto della (18). La funzione (19) viene scelta in quanto minimizza l'errore quadratico medio relativo:

$$\mathrm{RE}^2 = \frac{\mathbb{E}\left[\left(\hat{m}(\mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta}) - m(\mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta}) \right)^2 \right]}{m(\mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta})^2}.$$

Per ulteriori dettagli dimostrativi si veda Meng and Wong (1996, p.837).

Si osservi che l'equazione (19) mostra che la *bridge function* dipende strettamente dalla quantità che si vuole stimare: $m(\mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta})$. Tuttavia, si può procedere per via iterativa, fino a convergenza (operativamente fino a che non viene raggiunto il limite di tolleranza impostato) in modo tale che al passo $k+1$ la stima $\hat{m}(\mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta})$ risulti data da:

$$\hat{m}(\mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta})^{(k+1)} = \frac{\frac{1}{n_2} \sum_{i=1}^{n_2} \frac{\mathscr{L}(\mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta} | \tilde{\boldsymbol{\vartheta}}_i) \pi(\tilde{\boldsymbol{\vartheta}}_i)}{s_1 \mathscr{L}(\mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta} | \tilde{\boldsymbol{\vartheta}}_i) \pi(\tilde{\boldsymbol{\vartheta}}_i) + s_2 \hat{m}(\mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta})^{(k)} g(\tilde{\boldsymbol{\vartheta}}_i)}}{\frac{1}{n_1} \sum_{j=1}^{n_1} \frac{g(\vartheta_j^*)}{s_1 \mathscr{L}(\mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta} | \boldsymbol{\vartheta}_j^*) \pi(\boldsymbol{\vartheta}_j^*) + s_2 \hat{m}^{(k)}(\mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta}) g(\boldsymbol{\vartheta}_j^*)}} \tag{20}$$

Si è pertanto giunti allo stimatore bridge della costante di normalizzazione. Si osservi che da un punto di vista implementativo risulta più conveniente utilizzare la (20), trasformando le quantità in logaritmi, per evitare problemi di stabilità numerica (per approfondire si veda Gronau et al., 2017, Appendix B).

Mancherebbe da specificare una *proposal distribution* opportuna, che sia strettamente associata alla distribuzione di densità a posteriori. In questo caso le scelte possono essere diverse ma risulta chiaro che la scelta più ragionevole risulti essere quella di prendere come *proposal distribution* una distribuzione normale multivariata avente come media, e come matrice di varianza-covarianza, le rispettive quantità calcolate sulla base del campione estratto a posteriori, ovvero su $\boldsymbol{\vartheta}_1^*, \boldsymbol{\vartheta}_2^*, \ldots, \boldsymbol{\vartheta}_{n_1}^*$. La funzione $g(\boldsymbol{\vartheta})$ è la densità di una $\mathcal{N}\left(\bar{\boldsymbol{\vartheta}}^*, \boldsymbol{\Sigma}_{\bar{\boldsymbol{\vartheta}}^*}\right)$. In questo caso risulta evidente la stretta relazione sussistente tra le due distribuzioni.

La normale multivariata non è l'unica scelta possibile, per dettagli circa la specificazione di *proposal* alternative si veda Meng and Schilling (2002).

4.6.2 BIC e metodo approssimato

Un altro approccio per stimare la probabilità a posteriori di un modello, $\pi(\mathcal{M}_\lambda \mid \mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta})$, discende, in via naturale, dalla procedura di derivazione del *Bayesian Information Criterion* (BIC). Notoriamente il BIC è stato derivato nell'ambito dell'approccio bayesiano da Schwarz (1978) ed ha forma:

$$\text{BIC} = -2 \log \mathscr{L} (\mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta} \mid \boldsymbol{\vartheta}, \mathcal{M}_\lambda) + (p + 2) \log n, \tag{21}$$

in cui si considera il caso oggetto d'analisi in cui $p + 2$ rappresenta il numero di parametri stimati (in generale la formula è del tutto equivalente, con p generico numero di parametri).

Il BIC è in stretta relazione asintotica con la verosimiglianza marginale. Si consideri infatti $\mathscr{L}(\mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta} \mid \mathcal{M}_\lambda)$, si ha:

$$\mathscr{L}(\mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta} \mid \mathcal{M}_\lambda) = \int_\Theta \mathscr{L} (\mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta} \mid \boldsymbol{\vartheta}, \mathcal{M}_\lambda) \, \pi (\boldsymbol{\vartheta} \mid \mathcal{M}_\lambda) \, \mathrm{d}\boldsymbol{\vartheta}, \tag{22}$$

con $\pi (\boldsymbol{\vartheta} \mid \mathcal{M}_\lambda)$ distribuzione a priori di $\boldsymbol{\vartheta}$ condizionatamente al modello $\mathcal{M}_\lambda$. A partire dalla (22) è lecita la riscrittura:

$$\mathscr{L}(\mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta} \mid \mathcal{M}_\lambda) = \int_\Theta \exp \left[\log \mathscr{L} (\mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta} \mid \boldsymbol{\vartheta}, \mathcal{M}_\lambda) \right] \pi (\boldsymbol{\vartheta} \mid \mathcal{M}_\lambda) \, \mathrm{d}\boldsymbol{\vartheta}.$$

Applicando gli sviluppi di Taylor arrestati al second'ordine per $\log \mathscr{L} (\mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta} \mid \boldsymbol{\vartheta}, \mathcal{M}_\lambda)$ e arrestati al prim'ordine per $\pi (\boldsymbol{\vartheta} \mid \mathcal{M}_\lambda)$, intorno al punto di massimo $\dot{\boldsymbol{\vartheta}}$ (massimo a posteriori, coincidente con la moda a posteriori o la stima di massima

verosimiglianza), e sostituendo gli sviluppi nell'equazione (22), si ha che:

$$\begin{aligned}\mathscr{L}(\cdot \mid \mathcal{M}_\lambda) &= \int_\Theta \exp\left[\log \mathscr{L}(\mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta} \mid \boldsymbol{\vartheta}, \mathcal{M}_\lambda)\right] \pi(\boldsymbol{\vartheta} \mid \mathcal{M}_\lambda)\, \mathrm{d}\boldsymbol{\vartheta} \\ &= \int_\Theta \exp\left[\log \mathscr{L}\left(\mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta} \mid \dot{\boldsymbol{\vartheta}}, \mathcal{M}_\lambda\right) - \frac{n}{2}\left(\boldsymbol{\vartheta} - \dot{\boldsymbol{\vartheta}}\right)' \mathcal{I}(\dot{\boldsymbol{\vartheta}}) \left(\boldsymbol{\vartheta} - \dot{\boldsymbol{\vartheta}}\right) + \cdots\right] \times \\ &\qquad \left[\pi\left(\dot{\boldsymbol{\vartheta}} \mid \mathcal{M}_\lambda\right) + \left(\boldsymbol{\vartheta} - \dot{\boldsymbol{\vartheta}}\right) \boldsymbol{\nabla} \pi(\boldsymbol{\vartheta} \mid \mathcal{M}_\lambda)(\dot{\boldsymbol{\vartheta}}) + \cdots\right] \mathrm{d}\boldsymbol{\vartheta}\end{aligned}$$

con $\mathcal{I}\left(\dot{\boldsymbol{\vartheta}}\right)$ informazione osservata. Sicché, in virtù del metodo di Laplace, per il quale si rimanda al paragrafo 3.2.1, l'equazione (22) può essere approssimata come:

$$\mathscr{L}(\mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta} \mid \mathcal{M}_\lambda) \approx e^{\log \mathscr{L}(\mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta} \mid \boldsymbol{\vartheta}, \mathcal{M}_\lambda)} \pi\left(\dot{\boldsymbol{\vartheta}} \mid \mathcal{M}_\lambda\right) \int_\Theta e^{-\frac{n}{2}\left(\boldsymbol{\vartheta} - \dot{\boldsymbol{\vartheta}}\right)' \mathcal{I}(\dot{\boldsymbol{\vartheta}}) \left(\boldsymbol{\vartheta} - \dot{\boldsymbol{\vartheta}}\right)} \mathrm{d}\boldsymbol{\vartheta}. \quad (23)$$

Osservando che

$$\int_\Theta \exp\left[-\frac{n}{2}\left(\boldsymbol{\vartheta} - \dot{\boldsymbol{\vartheta}}\right)' \mathcal{I}(\dot{\boldsymbol{\vartheta}}) \left(\boldsymbol{\vartheta} - \dot{\boldsymbol{\vartheta}}\right)\right] \mathrm{d}\boldsymbol{\vartheta} = (2\pi)^{(p+2)/2} n^{-p-2} \left|\mathcal{I}\left(\dot{\boldsymbol{\vartheta}}\right)\right|^{-1/2}$$

e sostituendo in (23) si ha:

$$\mathscr{L}(\mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta} \mid \mathcal{M}_\lambda) \approx \exp\left[\log \mathscr{L}(\mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta} \mid \boldsymbol{\vartheta}, \mathcal{M}_\lambda)\right] \pi\left(\dot{\boldsymbol{\vartheta}} \mid \mathcal{M}_\lambda\right) (2\pi)^{\frac{p+2}{2}} n^{-\frac{p+2}{2}} \left|\mathcal{I}\left(\dot{\boldsymbol{\vartheta}}\right)\right|^{-\frac{1}{2}}$$

da cui:

$$\begin{aligned}-2\log \mathscr{L}(\mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta} \mid \mathcal{M}_\lambda) \approx &-2\log \mathscr{L}(\mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta} \mid \boldsymbol{\vartheta}, \mathcal{M}_\lambda) + (p+2)\log n + \\ &+ \log\left|\mathcal{I}(\dot{\boldsymbol{\vartheta}})\right| - (p+2)\log(2\pi).\end{aligned} \quad (24)$$

Considerando che, per $n \to \infty$, $\log\left|\mathcal{I}(\dot{\boldsymbol{\vartheta}})\right| = O(1)$ e $-(p+2)\log(2\pi) = O(1)$, la

(24) diventa:

$$-2\log\mathscr{L}(\mathbf{X},\boldsymbol{t},\boldsymbol{\delta}\mid\mathcal{M}_\lambda)\approx-2\log\mathscr{L}\left(\mathbf{X},\boldsymbol{t},\boldsymbol{\delta}\mid\boldsymbol{\vartheta},\mathcal{M}_\lambda\right)+(p+2)\log n, \tag{25}$$

da cui, per la (21) e la (25) si ha

$$-2\log\mathscr{L}(\mathbf{X},\boldsymbol{t},\boldsymbol{\delta}\mid\mathcal{M}_\lambda)\approx\mathrm{BIC}_\lambda.$$

Sicché, la verosimiglianza marginale, condizionatamente al modello $\mathcal{M}_\lambda$, è approssimabile come:

$$\mathscr{L}(\mathbf{X},\boldsymbol{t},\boldsymbol{\delta}\mid\mathcal{M}_\lambda)\approx\exp\left(-\frac{1}{2}\mathrm{BIC}_\lambda\right).$$

La procedura di derivazione è stata fatta considerando $p+2$ parametri, per rendere diretta la connessione con il caso oggetto d'analisi. Per ulteriori dettagli dimostrativi e interpretativi si consiglia il paragrafo 9.1 del testo di Konishi and Kitagawa (2008).

Da quanto appena enucleato discende che il BIC è in stretta relazione con la verosimiglianza marginale. In particolare, approssima asintoticamente il valore della verosimiglianza marginale. Supponendo che a priori non si abbiano informazioni per privilegiare nessuno dei modelli stimati risulta sensato attribuire alla probabilità a priori di ciascun modello, $\pi(\mathcal{M}_\lambda)$, una distribuzione a priori uniforme di parametri 0 e 1: $\mathcal{M}_\lambda\sim U(0,1)$. Ne discende che, a priori, i modelli hanno tutti la medesima probabilità di essere il modello ottimo.

Quindi, in virtù della relazione (17) è possibile ottenere, per ogni modello, la

probabilità a posteriori approssimata come:

$$\pi(\mathcal{M}_\lambda \mid \mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta}) = \frac{\mathscr{L}(\mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta} \mid \mathcal{M}_\lambda)}{\sum_\lambda \mathscr{L}(\mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta} \mid \mathcal{M}_\lambda)} \approx \frac{\exp\left(-\frac{1}{2}\mathrm{BIC}_\lambda\right)}{\sum_{\lambda=1}^{\Lambda} \exp\left(-\frac{1}{2}\mathrm{BIC}_\lambda\right)} \tag{26}$$

con $\Lambda = 100$ numero di modelli stimati.

Procedere per via asintotica presenta vantaggi notevoli dal punto di vista computazionale: non è necessario campionare innumerevoli valori dalla distribuzione a posteriori per stimare la verosimiglianza marginale, la quale richiede una numerosità campionaria maggiore rispetto a quella richiesta per la caratterizzazione della distribuzione a posteriori; dal BIC è possibile risalire alla probabilità a posteriori approssimata per ciascun modello. Lo svantaggio risiede nell'approssimazione stessa: se con il *Bridge Sampling* è possibile stabilire con esattezza la probabilità a posteriori del modello, con il BIC la si può solo approssimare. Tuttavia, e ciò verrà maggiormente chiarito nel paragrafo 4.6.3, stabilita la relazione che sussiste tra la verosimiglianza marginale stimata tramite *Bridge Sampling*, la verosimiglianza marginale asintotica, e il BIC, procedere per via approssimata permette di sfruttare la medesima relazione per ottenere sempre delle probabilità ma, al posto del BIC, di utilizzare altri indici che siano maggiormente informativi del potere predittivo del modello. In altri termini, per via approssimata è possibile distorcere la stima della verosimiglianza marginale per ottenere delle probabilità a posteriori che siano influenzate più dalla generalizzabilità del modello piuttosto che da quanto l'approssimazione è vicina al vero valore della verosimiglianza marginale.

Enucleata l'origine dell'approssimazione in esame, si osservi che in contesti ad elevata dimensionalità caratterizzati da una bassa numerosità campionaria,

proprio come nel caso oggetto d'analisi, il BIC non si rileva un indice di bontà del modello robusto. In particolare, per $p >> n$ il BIC tende ad attribuire probabilità più alte a modelli che sembrano mostrare sovrastima. Ne consegue che il BIC porta a una selezione che sovrastima la verosimiglianza marginale e dunque anche la probabilità a posteriori che il modello selezionato sia ottimo. Si consiglia, sui problemi inerenti all'utilizzo del BIC in contesti ad elevata dimensionalità, il lavoro di Broman and Speed (2002), in cui viene effettuata una disamina approfondita dei problemi del BIC in contesti ad elevata dimensionalità.

In virtù di tale criticità, sono state proposte diverse estensioni del BIC al caso ad alta dimensionalità. L'*Extended Bayesian Information Criteria*, EBIC, di Chen and Chen (2008), ne è un esempio. L'EBIC, tuttavia, presenta problemi analizzati da Owrang and Jansson (2018), i quali propongono un altro indice per la risoluzione dei problemi di cui sembra affetto l'EBIC. Tuttavia la soluzione proposta risulta applicabile solo nel caso dei modelli lineari e presenta a sua volta delle criticità. La versione del BIC per modelli ad alta dimensionalità ad oggi meno soggetta a problematicità è la versione robusta dell'EBIC proposta da Gohain and Jansson (2023). Nuovamente, tuttavia, l'indice è applicabile ai soli modelli lineari e sebbene risolva gran parte dei problemi che affliggono invece gli altri indici, e sia molto meno soggetto alla sovrastima, l'ipotesi sulla quale si regge è che nei dati sia presente sparsità, ipotesi che non sempre risulta corretta a priori e che quindi dovrebbe indurre a non utilizzare tale indice a fini di selezione.

A causa delle limitazioni che presentano il BIC e le sue estensioni in contesti ad elevata dimensionalità, si è preferito utilizzare altri indici per valutare la bontà del modello. In particolare, si è ritenuto più appropriato considerare due indici che guardassero maggiormente alla generalizzabilità del modello da un punto di vista

predittivo. In particolare, è stato considerato un indice basato sulla devianza, il *Deviance Information Criterion* (DIC), e il *Watanable-Akaike Information Criterion* (WAIC), anche detto *Widely Applicable Information Criterion*.

4.6.3 DIC e WAIC come alternative al BIC

Un approccio sensato per la stima delle probabilità a posteriori dei modelli consiste nel calcolare le probabilità sulla base dell'accuratezza predittiva associata a ciascun modello. Il razionale consiste nel considerare degli indici di accuratezza del modello che siano quanto più possibile non distorti rispetto alla quantificazione dell'incertezza previsiva e che, di conseguenza, forniscano informazioni utili per la selezione sulla base dell'accuratezza predittiva, stimata non già sul campione sul quale il modello è stato costruito ma sulle osservazioni cosiddette *out-of-sample*, ovvero nuove osservazioni.

Quanto appena esposto assomiglia molto al classico approccio di convalida incrociata ampiamente usato in statistica e nel *Machine Learning*. Si osservi, a riguardo, che un approccio possibile è proprio quello della convalida incrociata per la stima dell'accuratezza previsiva del modello. Tuttavia è altresì evidente come nel caso oggetto d'analisi tale procedura, per i cui dettagli in ambito bayesiano si rimanda a Vehtari and Lampinen (2002), risulti troppo onerosa da un punto di vista computazionale. L'ideale sarebbe quindi trovare delle misure di accuratezza previsiva in grado di tenere conto di osservazioni *out-of-sample* senza tuttavia dover necessariamente passare per la convalida incrociata.

In questo contesto risulta sensato considerare l'*Expected log predictive density*. Infatti sussiste una stretta relazione tra l'informazione di Kullback-Leibler (D_{KL})

e l'*Expected log predictive density*. In particolare, il modello con la più bassa divergenza di Kullback-Leibler corrisponde al modello con l'*Expected log predictive density* maggiore. Di conseguenza il modello tale per cui viene massimizzata la *Expected log predictive density* è il modello con la maggior probabilità a posteriori. Per ulteriori dettagli circa la divergenza di Kullback-Leibler si rimanda a Csiszár (1975).

Al fine di calcolare l'*Expected log predictive density* per ciascun modello, si denoti con $\tilde{\boldsymbol{x}}_i$ una nuova osservazione, sia f una funzione rappresentante il meccanismo generatore dei dati. L'*Expected log predictive density* per la nuova osservazione ha forma:

$$\text{elpd} = \mathbb{E}_f \log \pi(\boldsymbol{\vartheta} \mid \tilde{\boldsymbol{x}}_i) = \int_{\mathcal{X}} \log \pi(\boldsymbol{\vartheta} \mid \tilde{\boldsymbol{x}}_i) f(\tilde{\boldsymbol{x}}_i) \mathrm{d}\boldsymbol{x}.$$

Sia la *Expected log pointwise predictive density* per n nuove osservazioni definita come:

$$\mathcal{L} = \sum_{i=1}^{n} \mathbb{E}_f \log \pi(\boldsymbol{\vartheta} \mid \tilde{\boldsymbol{x}}_i) = \sum_{i=1}^{n} \int_{\mathcal{X}} \log \pi(\boldsymbol{\vartheta} \mid \tilde{\boldsymbol{x}}_i) f(\tilde{\boldsymbol{x}}_i) \mathrm{d}\boldsymbol{x}.$$

Da un punto di vista applicativo, il vero meccanismo generatore dei dati, f, non è noto, come non risulta noto il vettore di parametri $\boldsymbol{\vartheta}$. In altri termini, $\mathcal{L}$ rappresenta la quantificazione esatta ma non calcolabile. Ciò che è possibile fornire è una stima di $\mathcal{L}$, che sia quanto più possibile non distorta. Si denoti con $\dot{\mathcal{L}}$ la *log pointwise predictive density* ovvero la stima, tramite la distribuzione a posteriori, dell'accuratezza del modello. Si ha:

$$\dot{\mathcal{L}} = \sum_{i=1}^{n} \log \int_{\Theta} \mathscr{L}(\boldsymbol{x}_i, t_i, \delta_i \mid \boldsymbol{\vartheta}) \pi(\boldsymbol{\vartheta} \mid \mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta}) \mathrm{d}\boldsymbol{\vartheta}.$$

Da un punto di vista computazionale è necessaria una versione discreta di $\dot{\mathcal{L}}$. A tal fine si assuma che S, ovvero il numero di osservazioni campionate dalla a posteriori, sia sufficientemente elevato, allora è lecita la seguente discretizzazione:

$$\dot{\mathcal{L}} = \sum_{i=1}^{n} \log \left(\frac{1}{S} \sum_{s=1}^{S} \mathscr{L}(\boldsymbol{x}_i, t_i, \delta_i \mid \boldsymbol{\vartheta}^s) \right) \tag{27}$$

Essendo $\dot{\mathcal{L}}$ calcolata sui dati su cui il modello è costruito, essa è necessariamente una quantità che sovrastima la vera $\mathcal{L}$. Risulta pertanto necessario trovare delle tecniche che, a partire dalla stima di $\dot{\mathcal{L}}$, apportino un aggiustamento tenendo conto del sovra-adattamento e che riescano a produrre una stima di $\mathcal{L}$ quanto più possibile vicina al suo vero valore. Partendo dal presupposto che il *gold standard* sarebbe rappresentato dalla convalida incrociata, è comunque possibile adottare degli indici che aggiustino $\dot{\mathcal{L}}$ per l'*overfitting* e che, contemporaneamente, si comportino in modo molto simile alla convalida incrociata, fornendo quindi risultati del tutto analoghi. In quest'ottica vengono enucleati i due seguenti criteri informativi: il DIC e il WAIC, i quali, a partire dalla stima distorta $\dot{\mathcal{L}}$, introducono, in modo diverso, un fattore di aggiustamento per correggere la distorsione.

4.6.3.1 DIC

Il DIC stima $\mathcal{L}$ facendo *plug-in* della media a posteriori di $\tilde{\boldsymbol{\vartheta}} = \mathbb{E}^{\pi(\boldsymbol{\vartheta}|\cdot)}(\boldsymbol{\vartheta} \mid \mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta})$ e introducendo un fattore di correzione, p_{DIC}, interpretabile come il numero effettivo di parametri. Si ha pertanto che la nuova misura di accuratezza predittiva è data da:

$$\log \mathscr{L}\left(\mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta} \mid \tilde{\boldsymbol{\vartheta}}\right) - p_{\text{DIC}},$$

con

$$p_{\text{DIC}} = 2\left(\log \mathscr{L}\left(\mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta} \mid \tilde{\boldsymbol{\vartheta}}\right) - \frac{1}{S}\sum_{s=1}^{S} \log \mathscr{L}\left(\mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta} \mid \boldsymbol{\vartheta}^s\right)\right),$$

dove $\boldsymbol{\vartheta}^s$ rappresenta il vettore campionario s-simo ottenuto dal campionamento MCMC. Vi è anche un altro modo per calcolare il p_{DIC}, per il quale si rimanda a Van Der Linde (2005), in quanto la versione implementata in questo lavoro è quella appena riportata. Il DIC in senso stretto è definito in termini della devianza piuttosto che in termini della *log predictive density*, infatti ha forma:

$$\text{DIC} = -2\log \mathscr{L}\left(\mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta} \mid \tilde{\boldsymbol{\vartheta}}\right) + 2p_{\text{DIC}}.$$

Il DIC è largamente utilizzato in virtù della sua semplicità di impiego. Infatti è sufficiente valutare la log verosimiglianza in corrispondenza del massimo a posteriori, e in corrispondenza di ciascun valore estratto nella fase di campionamento. Per una disamina più approfondita riguardo le caratteristiche del DIC si raccomanda il lavoro di Spiegelhalter et al. (2002) e quello di Ando (2007).

Si osservi che nel DIC si considera nel fattore di aggiustamento lo scostamento, in media, dalla log verosimiglianza valutata nel punto di massimo a posteriori per ciascun campione MCMC. In altri termini, ci si condiziona a un unico valore della distribuzione a posteriori, il massimo, non tenendo in considerazione delle differenze esistenti tra altri punti della distribuzione. Per questo motivo, sebbene nella pratica diano informazioni spesso simili, al DIC viene in genere preferito il WAIC che, a differenza del DIC, non si condiziona a un unico valore della distribuzione a posteriori.

4.6.3.2 WAIC

Il WAIC, introdotto da Watanabe (2013), parte dalla stima della *log pointwise predictive density* riportata in (27) e, per fornire una stima non distorta di $\mathcal{L}$, aggiunge un fattore di penalizzazione, p_{WAIC}, sicché, indicata con $\hat{\mathcal{L}}$ la stima non distorta di $\mathcal{L}$, si ha che:

$$\hat{\mathcal{L}} = \dot{\mathcal{L}} - p_{\text{WAIC}},$$

con

$$p_{\text{WAIC}} = 2\sum_{i=1}^{n}\left(\log \mathbb{E}^{\pi(\boldsymbol{\vartheta}|\cdot)}\mathscr{L}(\boldsymbol{x}_i, t_i, \delta_i \mid \boldsymbol{\vartheta}) - \mathbb{E}^{\pi(\boldsymbol{\vartheta}|\cdot)}\log \mathscr{L}(\boldsymbol{x}_i, t_i, \delta_i \mid \boldsymbol{\vartheta})\right),$$

da cui discende la discretizzazione a scopo computazionale:

$$p_{\text{WAIC}} = 2\sum_{i=1}^{n}\left(\log\left(\frac{1}{S}\sum_{s=1}^{S}\mathscr{L}(\boldsymbol{x}_i, t_i, \delta_i \mid \boldsymbol{\vartheta}^s)\right) - \frac{1}{S}\sum_{i=1}^{S}\log \mathscr{L}(\boldsymbol{x}_i, t_i, \delta_i \mid \boldsymbol{\vartheta}^s)\right).$$

Similmente al DIC, si è soliti esprimere il WAIC in termini di devianza, quindi:

$$\text{WAIC} = -2\dot{\mathcal{L}} + 2p_{\text{WAIC}}.$$

Si osservi che, come nel caso del DIC, anche per il WAIC il p_{WAIC} può essere calcolato attraverso un'altra formulazione, in particolare con la varianza campionaria della a posteriori:

$$p_{\text{WAIC2}} = \sum_{i=1}^{n}\frac{1}{S-1}\sum_{s=1}^{S}\left(\log \mathscr{L}(\boldsymbol{x}_i, t_i, \delta_i \mid \boldsymbol{\vartheta}^s) - \frac{1}{S}\sum_{i=1}^{S}\log \mathscr{L}(\boldsymbol{x}_i, t_i, \delta_i \mid \boldsymbol{\vartheta}^s)\right)^2$$

Si è proceduto implementando entrambe le versioni del p_{WAIC}, tuttavia si è ritenuto opportuno utilizzare i risultati della prima opzione in quanto più stabili e interpretabili in termini di numero di parametri stimati.

Rispetto al DIC, il WAIC effettua la media a partire dalla distribuzione a posteriori piuttosto che condizionarsi a un solo valore, come l'ottimo a posteriori. Questa caratteristica è di cruciale importanza soprattutto in ottica previsiva, giacché il WAIC effettivamente valuta la bontà delle previsioni e aggiusta $\dot{\mathcal{L}}$ sulla basa di quest'ultime e non già sulla base delle performance della *log density predictive*, per la quale viene fatto *plug-in* dell'ottimo a posteriori. Infatti, da un punto di vista bayesiano, il WAIC permette di evitare di usare la distribuzione predittiva a posteriori. Per quanto concerne il DIC questo passaggio successivo è d'obbligo, a riguardo si consiglia il lavoro di Vehtari et al. (2017).

Infine, è possibile dimostrare che il WAIC è l'equivalente asintotico della *Leave-One-Out Cross-validation* in ambito bayesiano, ovvero la forma di validazione interna più fine possibile, per i dettagli dimostrativi si rimanda al lavoro di Watanabe and Opper (2010).

Date le premesse, si è deciso comunque di calcolare sia il DIC che il WAIC anche per scopo comparativo, in particolare per valutare se fossero concordi nella selezione del modello. Si è infine deciso di utilizzare il WAIC per calcolare le probabilità a posteriori dei modelli secondo la relazione (26):

$$\pi(\mathcal{M}_\lambda \mid \mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta}) = \frac{\mathscr{L}(\mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta} \mid \mathcal{M}_\lambda)}{\sum_\lambda \mathscr{L}(\mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta} \mid \mathcal{M}_\lambda)} \approx \frac{\exp\left(-\frac{1}{2}\text{WAIC}_\lambda\right)}{\sum_{\lambda=1}^{\Lambda} \exp\left(-\frac{1}{2}\text{WAIC}_\lambda\right)}, \tag{28}$$

con $\Lambda = 100$ numero di modelli stimati. In altri termini, si è considerata la relazione valida a livello asintotico tra la probabilità a posteriori di un modello

e il BIC corrispondente al modello stesso ed è stata estesa la relazione derivata nel paragrafo 4.6.2 al WAIC. Si è scelto il WAIC in quanto dovrebbe essere una misura più accurata rispetto al BIC relativamente al potere di generalizzabilità del modello. Si osservi che un approccio simile, ma basato sull'AIC, è stato già suggerito da Burnham and Anderson (2004).

Enucleate le caratteristiche del WAIC è possibile cogliere il vantaggio dell'approccio asintotico: esso permette di stimare le probabilità a posteriori di ciascun modello, esattamente come il *Bridge Sampling*. Tuttavia, l'approssimazione riguarda indirettamente le probabilità a posteriori dal momento che il BIC approssima in prima istanza la verosimiglianza marginale. Da questo punto di vista, nulla vieta di distorcere la stima della verosimiglianza marginale per ottenere delle probabilità a posteriori che non dipendano da quanto l'approssimazione della verosimiglianza marginale risulti affidabile ma, bensì, che siano basate su una stima quanto meno distorta possibile dell'accuratezza predittiva di ciascun modello, considerando al posto del BIC, per esempio, proprio il WAIC, che asintoticamente rispecchia la misura di accuratezza della LOO-CV in ambito bayesiano.

L'approccio risulta pertanto ragionevole: non si è interessati tanto alla stima della verosimiglianza marginale, quanto più ad ancorare le probabilità a posteriori a una misura di accuratezza che tenga conto di osservazioni *out-of-sample*.

Concludendo, se procedere per via approssimata in prima istanza può apparire una scelta forzata per ragioni computazionali, contemporaneamente si rivela un approccio sensato e per certi versi capace di aggiungere qualcosa dal punto di vista informativo rispetto alla procedura esatta.

4.7 Miglior modello e *Bayesian Model Averaging*

Si pone infine il problema di come utilizzare le probabilità a posteriori approssimate.

In prima istanza è possibile valutare il modello associato alla probabilità a posteriori più elevata, che rappresenta una possibilità per la selezione del modello. Tuttavia è importante notare che la selezione di un singolo modello, quando si è in presenza di diversi candidati tra cui effettuare la selezione, introduce inevitabilmente un *bias* di selezione. Selezionare un modello sulla base delle probabilità a posteriori può essere sensato se, fra numerosi candidati, uno di essi si associa a una probabilità a posteriori molto più elevata rispetto alle probabilità associate agli altri modelli. Al contrario, se si è in presenza di numerosi candidati, in cui non vi è presenza di un modello con una probabilità a posteriori molto più elevata delle altre, ma le probabilità a posteriori risultano molto simili per un sottoinsieme consistente di modelli, allora procedere selezionando un solo modello accentuerebbe in modo sostanziale il *bias* di selezione.

Si consideri inoltre che in un contesto in cui i candidati hanno probabilità a posteriori molto simili tra loro selezionare un singolo modello, e stimarne i parametri, può portare a una significativa perdita di informazioni, e di conseguenza a una stima viziata dei parametri oggetto di inferenza. Nel contesto oggetto di analisi si hanno 100 modelli candidati con probabilità a posteriori molto simili, segue che ciascuna probabilità a posteriori sarà bassa: se nessun modello domina nettamente sugli altri è ragionevole aspettarsi che la singola probabilità a posteriori

non superi il 5%. Si osservi che il valore preso è puramente arbitrario e serve solo per comprendere la natura del problema oggetto d'analisi, qualsiasi altra probabilità sufficientemente bassa può essere adottata e il ragionamento rimarrebbe valido. Fatta la dovuta precisazione, si ipotizzi di selezionare proprio il modello con la probabilità a posteriori più elevata che, a titolo di esempio, è stata posta pari al 5%. Da un punto di vista inferenziale la procedura si configurerebbe come segue: sia $\hat{\mathcal{M}}_{\lambda_k}$ il modello con la massima probabilità a posteriori, allora il vettore di parametri ottimo su cui basare l'inferenza sarebbe $\hat{\boldsymbol{\vartheta}}_{\lambda_k}$, che rappresenta la media dei vettori di parametri $\boldsymbol{\vartheta}_{\lambda_k}$ campionati con MCMC. Ne discende che le stime e le previsioni prodotte utilizzando il modello $\hat{\mathcal{M}}_{\lambda_k}$ non tengono conto dell'incertezza relativa alla selezione. In altri termini, il modello $\hat{\mathcal{M}}_{\lambda_k}$ a cui si associa una probabilità a posteriori pari al 5% viene utilizzato per la stima e per le previsioni come se fosse l'unico modello possibile e, dunque, come se a esso fosse associata una probabilità di essere il modello esatto del 100%. La stima e le previsioni risultano dunque condizionate ad un unico modello.

In contesti in cui le probabilità a posteriori sono molto vicine tra loro, per ottenere delle stime dei parametri robuste e delle previsioni per nuove osservazioni consistenti, è necessario tenere conto dell'incertezza espressa dalle probabilità a posteriori di ciascun modello e non condizionarsi ad un unico modello.

L'approccio del *Bayesian Model Averaging* (BMA) tiene conto non solo dell'incertezza associata alla stima dei parametri ma anche di quella sussistente tra i diversi modelli. Invece che considerare la distribuzione dei parametri di un singolo modello per la stima o la previsione, il BMA considera le distribuzioni dei parametri e delle previsioni di tutti i modelli, e le combina attraverso una media ponderata, in cui i pesi sono forniti proprio dalle probabilità a posteriori associate

a ciascun modello. Per riferimenti riguardanti il BMA si consigliano i primi lavori di Jeffreys (1998), da p.295, e Jevons (1877), da p. 290; per una trattazione più esaustiva e associata ai metodi computazionali più recenti si vedano i lavori di Hoeting et al. (1999), Raftery et al. (1997) e Fragoso et al. (2018).

Formalmente, siano $\mathcal{M}_1, \ldots, \mathcal{M}_\lambda, \ldots, \mathcal{M}_\Lambda$ i modelli candidati, si indichino con $\pi(\mathcal{M}_1 \mid \mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta}), \ldots, \pi(\mathcal{M}_\lambda \mid \mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta}), \ldots, \pi(\mathcal{M}_\Lambda \mid \mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta})$ le probabilità a posteriori associate a ciascun modello. Posto $\Lambda = 100$ si ha che la probabilità a posteriori di $\boldsymbol{\vartheta}$, condizionatamente ai dati, ai 100 modelli candidati, e alle probabilità a posteriori ad essi associate è data da:

$$\pi_{\text{BMA}}(\boldsymbol{\vartheta} \mid \mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta}) = \sum_{\lambda=1}^{\Lambda} \pi(\boldsymbol{\vartheta}_\lambda \mid \mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta}) \pi(\mathcal{M}_\lambda \mid \mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta}), \tag{29}$$

da cui segue che la media a posteriori di $\boldsymbol{\vartheta}$ ha forma:

$$\tilde{\boldsymbol{\vartheta}}_{\text{BMA}} = \frac{1}{S} \sum_{s=1}^{S} \boldsymbol{\vartheta}^s = \frac{1}{S} \sum_{s=1}^{S} \sum_{\lambda=1}^{\Lambda} \boldsymbol{\vartheta}_\lambda^s \pi(\mathcal{M}_\lambda \mid \mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta}),$$

con S numero di campioni dalla a posteriori.

In generale, indicando con Δ la quantità di interesse inferenziale, sia esso il vettore dei parametri ottimo, $\tilde{\boldsymbol{\vartheta}}_{\text{BMA}}$, sia esso un indicatore di bontà previsiva o le previsioni stesse, essa si ottiene considerando la media rispetto alla distribuzione $\pi_{\text{BMA}}(\boldsymbol{\vartheta} \mid \mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta})$, sicché la media a posteriori di Δ è data da:

$$\tilde{\Delta}_{\text{BMA}} = \frac{1}{S} \sum_{s=1}^{S} \sum_{\lambda=1}^{\Lambda} \Delta_\lambda^s \pi(\mathcal{M}_\lambda \mid \mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta}).$$

Si osservi che la relazione (29) è una mistura delle 100 distribuzioni a posteriori ottenute con campionamento MCMC, dove i pesi della mistura sono forniti proprio

dalle probabilità a posteriori associate a ciascun modello. Da non confondere quindi la mistura di distribuzioni a posteriori con una convoluzione di distribuzioni. Al fine di generare i campioni della distribuzione in oggetto sarà sufficiente estrarre 5000 vettori di parametri dalle 100 distribuzioni a posteriori associate ai modelli, attribuendo come probabilità di estrazione dalla j-sima distribuzione la probabilità a posteriori corrispondente al j-simo modello.

Per quanto riguarda le probabilità a posteriori, $\pi(\mathcal{M}_\lambda \mid \mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta})$, si sfrutterà, come ampiamente discusso, la relazione (28), con un accorgimento: dal punto di vista computazionale se si introducesse il WAIC con il suo valore originario si andrebbe incontro alla valutazione di una quantità che la macchina riconosce come nulla al numeratore (anche se nulla non è). Per questo motivo giova ricorrere a una normalizzazione dei valori. In particolare risulta di estrema utilità definire ciascun WAIC in funzione del minimo WAIC, in modo tale che:

$$\overline{\mathrm{WAIC}}_\lambda = \mathrm{WAIC}_\lambda - \min\left(\mathbf{WAIC}\right).$$

Infine, dal momento che a priori non si ha a disposizione alcuna informazione in grado di fornire indicazioni circa quali modelli pesare maggiormente, si ritiene opportuno attribuire a ciascun modello una a priori $U(0,1)$, sicché ciascun modello, a priori, risulti equiprobabile. Dunque per la relazione (28) si ha che:

$$\pi(\mathcal{M}_\lambda \mid \mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta}) = \frac{\exp\left(-\frac{1}{2}\overline{\mathrm{WAIC}}_\lambda\right)}{\sum_{\lambda=1}^{\Lambda} \exp\left(-\frac{1}{2}\overline{\mathrm{WAIC}}_\lambda\right)}, \tag{30}$$

che computazionalmente risulta calcolabile.

4.8 Sarculator vs *Bayesian Model Averaging*

In ultima istanza si condurrà un test di ipotesi per verificare se il modello costruito attraverso il BMA, includente anche le variabili radiomiche, sia preferibile al modello contenente solo il Sarculator come indice prognostico. Il test è utile dal momento che permette di comprendere se le variabili radiomiche aggiungono potere prognostico al Sarculator stesso.

Per procedere con il test è risultato conveniente fare riferimento all'apparato teorico sin qui costruito. Infatti, con riferimento all'equazione (17), e per quanto enucleato nel paragrafo 4.7, sia $\mathcal{M}_0$ il modello contenente solo il Sarculator e l'intercetta, ovvero il modello tale per cui il parametro di precisione per i coefficienti di regressione associati alle variabili radiomiche è pari a $\lambda = \infty$. Attribuendo a $\mathcal{M}_0$ una probabilità a priori di essere il miglior modello pari al 50%, e ripartendo il restante 50% di probabilità, uniformemente, su ciascuno dei 100 modelli, sicché la probabilità a priori di ciascun modello risulti pari allo 0.5%, è possibile calcolare le rispettive probabilità a posteriori.

Sia $\mathcal{M}_\lambda$ il generico modello ottenuto in corrispondenza di λ, allora il test di ipotesi diviene il seguente:

$$\mathrm{H}_0 : \mathcal{M}_0 \quad \text{è sufficiente} \quad vs \quad \mathrm{H}_1 : \exists\lambda \mid \mathcal{M}_\lambda \quad \text{è migliore di} \quad \mathcal{M}_0.$$

Si osservi che il seguente sistema di ipotesi è equivalente a verificare se, posta la probabilità a priori di $\mathcal{M}_0$ pari al 50%, $\pi(\mathcal{M}_0) = 0.5$, e la probabilità a priori di ciascun $\mathcal{M}_\lambda$ pari a $\frac{1}{2}\frac{1}{\Lambda}$, con $\Lambda = 100$, ovvero $\pi(\mathcal{M}_\lambda) = 0.005 \quad \forall\lambda$, allora la

probabilità a posteriori di $\mathcal{M}_0$ è maggiore o, al più, uguale a 1/2. Formalmente:

$$\mathrm{H}_0 : \pi\left(\mathcal{M}_0 \mid \mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta}\right) \geq \frac{1}{2} \;\Big|\; \pi\left(\mathcal{M}_0\right) = \frac{1}{2} \quad vs \quad \mathrm{H}_1 : \pi\left(\mathcal{M}_0 \mid \mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta}\right) < \frac{1}{2} \;\Big|\; \pi\left(\mathcal{M}_0\right) = \frac{1}{2}$$

Al fine di calcolare la probabilità a posteriori $\pi\left(\mathcal{M}_0 \mid \mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta}\right)$ si sfrutterà la relazione (30), in modo tale che:

$$\pi\left(\mathcal{M}_0 \mid \mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta}\right) = \frac{\mathscr{L}\left(\mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta} \mid \mathcal{M}_0\right) \pi\left(\mathcal{M}_0\right)}{\mathscr{L}\left(\mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta} \mid \mathcal{M}_0\right) \pi\left(\mathcal{M}_0\right) + \sum_{\lambda}^{\Lambda} \mathscr{L}\left(\mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta} \mid \mathcal{M}_\lambda\right) \pi\left(\mathcal{M}_\lambda\right)},$$

dove le verosimiglianze marginali non sono note e quindi, ricorrendo la (30), si ottiene la seguente approssimazione:

$$\pi\left(\mathcal{M}_0 \mid \mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta}\right) \approx \frac{\exp\left(-\frac{1}{2}\overline{\mathrm{WAIC}}_0\right) \times \frac{1}{2}}{\exp\left(-\frac{1}{2}\overline{\mathrm{WAIC}}_0\right) \times \frac{1}{2} + \sum_{\lambda}^{\Lambda} \exp\left(-\frac{1}{2}\overline{\mathrm{WAIC}}_\lambda\right) \times \frac{1}{2\Lambda}}. \tag{31}$$

5 Risultati

In questo capitolo verrà dapprima indagato il comportamento del campionamento effettuato tramite metodi MCMC per verificare il raggiungimento della convergenza. Si paragonerà il *mixing* del RW MH con quello dell'A-MALA pre-condizionato.

Si effettuerà anche la diagnostica circa la componente adattiva per verificare che l'ipotesi di *diminishing adaptation* sia rispettata. Si procederà quindi valutando i risultati del miglior modello e quelli derivanti dal *Bayesian Model Averaging*. Si verificherà poi se le variabili radiomiche aggiunte al Sarculator siano preferibili al solo modello con il Sarculator, per rispondere alla domanda originaria: la radiomica aumenta il potere prognostico del Sarculator?

Si costruiranno infine le curve di sopravvivenza stimate per due pazienti, per fornire un esempio di come utilizzare il BMA per ottenere delle previsioni, e la curva di sopravvivenza globale con relativo intervallo di credibilità al 95%.

5.1 Diagnostica A-MALA pre-condizionato

Al fine di inizializzare i vettori delle medie delle *proposal distributions* è stata specificata una sequenza di 100 valori di λ. Per ciascuno di questi si è proceduto all'approssimazione numerica del vettore ottimo iniziale. In Figura 8 si riporta l'andamento, al variare di λ espresso in scala logaritmica, dei valori ottimi di ciascun parametro.

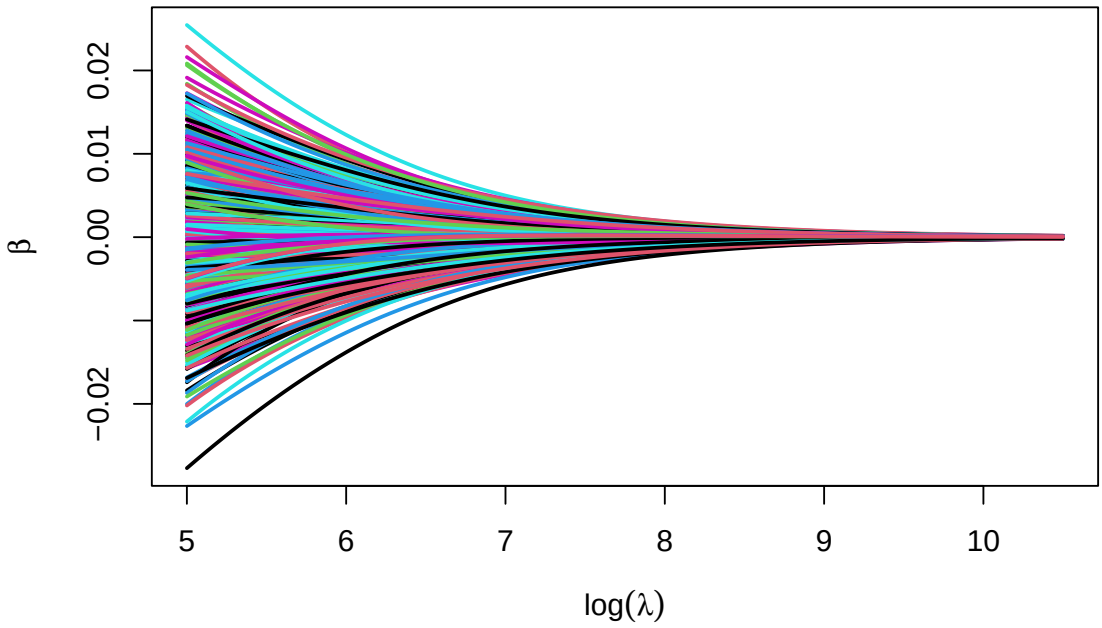


Figura 8: Inizializzazione del vettore dei parametri, durante l'analisi pre-ottimale, il grafico mostra l'andamento dell'inizializzazione dei vettori dei coefficienti radiomici al variare del parametro di precisione espresso in scala logaritmica

Si osservi che gli estremi dell'intervallo $[e^5, e^{10.5}]$ sono stati selezionati attraverso un processo di *trial and errors*. In particolare, l'intervallo iniziale era sufficientemente ampio $[e^2, e^{10.5}]$, in modo da esplorare differenti valori di penalizzazione, da una penalizzazione molto marcata, $e^{10.5}$, a una penalizzazione molto moderata e^2. In seguito si è osservato il comportamento del campionamento effettuato tramite metodo MCMC e si è aumentato gradualmente l'estremo inferiore

dell'intervallo, in modo tale che i campioni mostrassero tutti un buon *mixing* e non presentassero problemi di identificabilità.

Si è osservato infatti che penalizzazioni troppo basse portavano a un contesto di quasi non identificabilità del modello. A partire dai vettori di ottimo iniziale si è anche proceduto calcolando, per via esatta, il valore corrispondente delle matrici di varianza-covarianza delle diverse *proposal distributions* sfruttando la relazione tra la distribuzione log a posteriori e l'informazione osservata di Fisher.

Con riferimento alla Figura 8 si può osservare come le penalizzazioni siano tutte molte elevate e i coefficienti relativi alle variabili radiomiche siano molto prossimi a zero. In particolare, la massima penalizzazione porta a un vettore di inizializzazione quasi nullo.

Per il primo algoritmo di campionamento utilizzato, il *Random Walk Metropolis Hastings*, con *proposal distributions* specificate come riportato nel capitolo 4, vengono riportati in Figura 9 i *trace plots* di alcune variabili ($\log \lambda = 10.39$). Il valore di λ per cui si mostrano i grafici risulta utile per confrontare la convergenza dell'algoritmo in oggetto con quella dell'A-MALA pre-condizionato. Si evince come la convergenza sia lontana dall'essere raggiunta. Ne consegue che si sarebbe dovuto aumentare in modo computazionalmente non sostenibile il numero delle iterazioni per giungere a convergenza. Tale andamento caratterizza anche tutti gli altri valori di λ.

Risulta pertanto chiara l'esigenza di adottare un metodo di campionamento più efficiente, che si traduca in un buon *mixing*, fisso restando il numero di iterazioni.

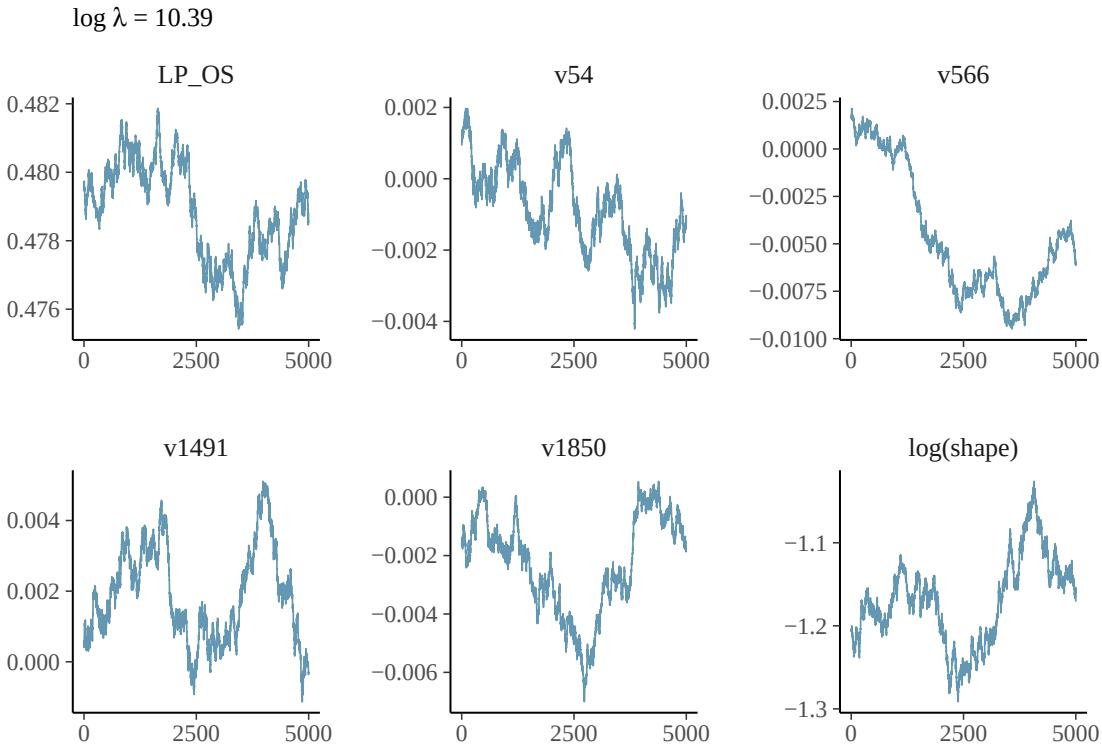
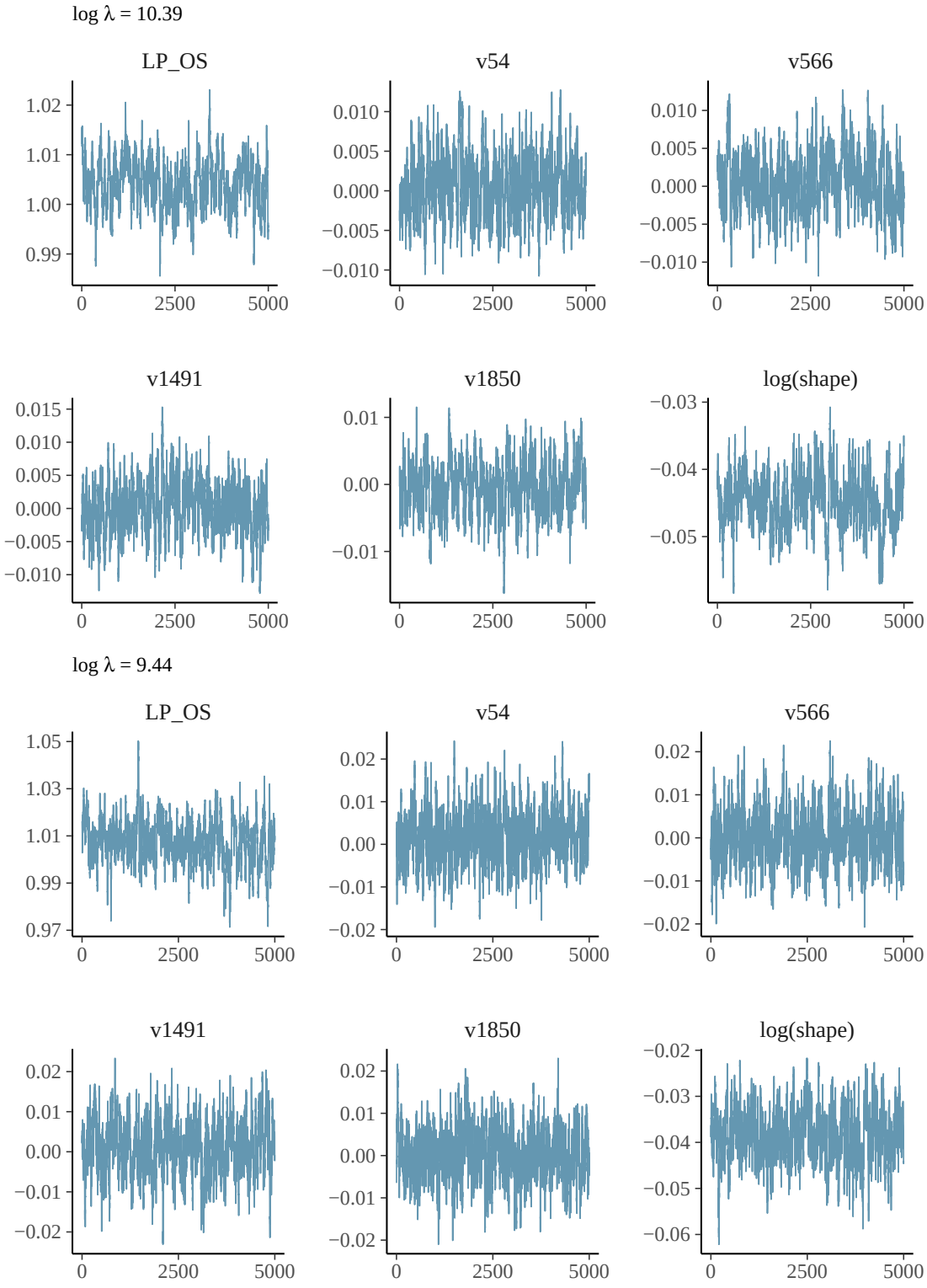


Figura 9: Trace plots per il monitoraggio della convergenza, per alcune variabili, relativi al RW-MH

Si è quindi implementata una versione ad hoc del MALA pre-condizionato i cui risultati, in termini di convergenza, vengono riportati in Figura 10 per un sottocampione di quattro valori di λ. Per fornire un'idea del comportamento complessivo del campionamento, per ciascun valore di λ, sono stati riportati i *trace plots* per quattro valori che fossero rappresentativi del grado di penalizzazione nell'intervallo considerato. In particolare, il primo valore del parametro di precisione corrisponde a una forte penalizzazione $\lambda = e^{10.39}$, il secondo e il terzo valore a penalizzazioni intermedie $\lambda = e^{9.44}$, e $\lambda = e^{7.94}$, l'ultimo valore rappresenta infine una penalizzazione molto blanda $e^{5.56}$. Si può notare come l'A-MALA implementato porti a convergere le catene per ciascuno dei 4 valori di λ. In generale la convergenza risulta buona per ciascuno dei 100 valori di λ.


log λ = 10.39
LP_OS
v54
v566
v1491
v1850
log(shape)
log λ = 9.44
LP_OS
v54
v566
v1491
v1850
log(shape)

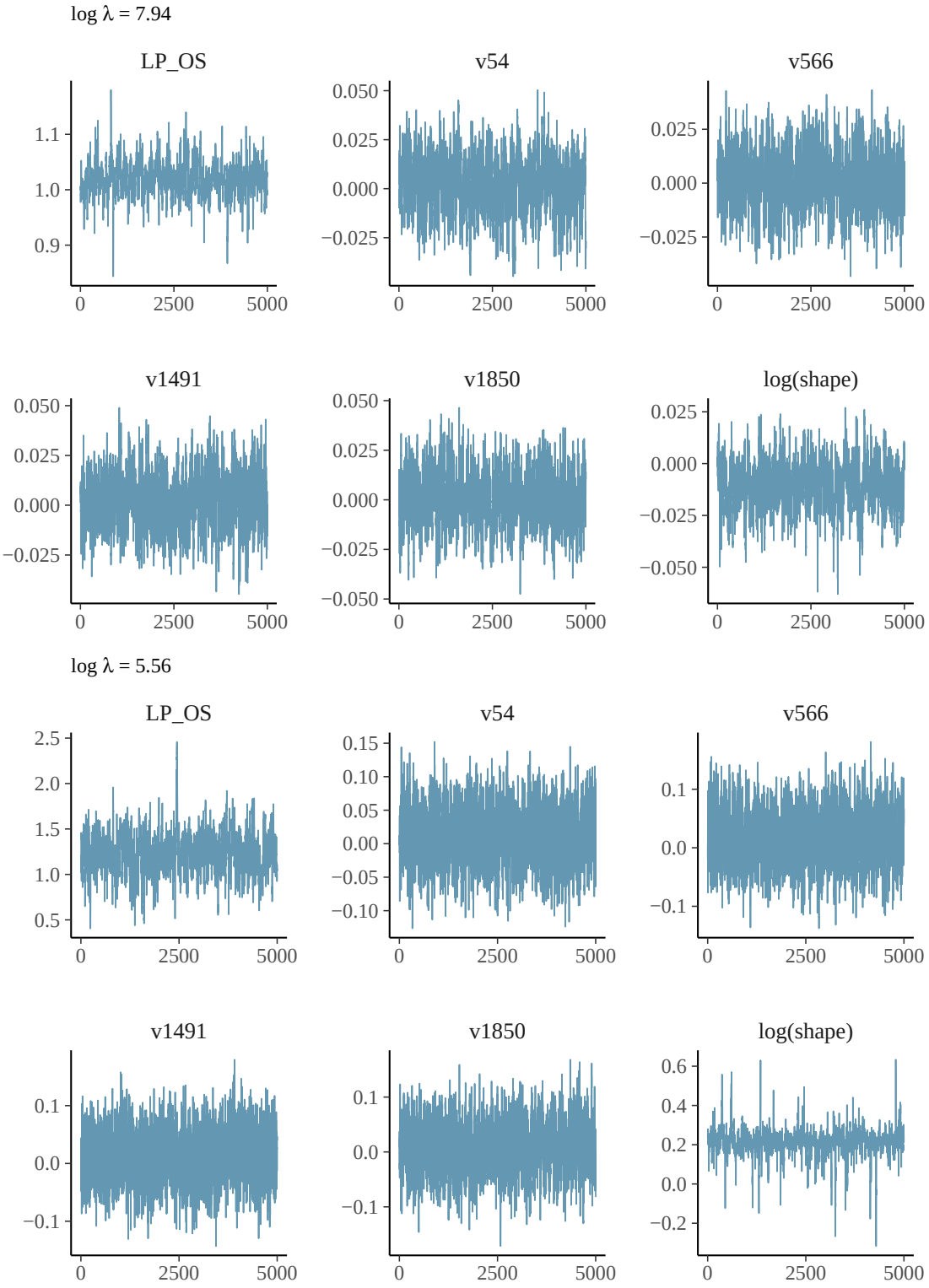


Figura 10: Trace plots per il monitoraggio della convergenza per 4 distinti valori del parametro di precisione, per alcune variabili, relativi all' A-MALA pre-condizionato

Per quanto riguarda il tasso di accettazione ottimale, è stato enucleato nel paragrafo 4.5 che, asintoticamente, ovvero per un numero sufficiente di iterazioni, esso risulta pari al 57.4%. I risultati relativi ai 100 valori di λ mostrano un tasso di accettazione mediano, registrato dopo la fase di *burn-in*, di circa 56.8%. Il primo quartile dei valori del tasso di accettazione risulta pari a circa il 56% e il terzo quartile pari a circa il 57.8%. Tali risultati dimostrano come la strategia adattiva scelta per l'ottimizzazione del parametro di *step-size* sia stata efficace per ogni valore di λ, portando ad un tasso di accettazione effettivo molto vicino a quello ottimale teorico.

In Figura 11 si riporta anche il grafico dell'*Effective Sample Size* al variare di λ espresso in scala logaritmica.

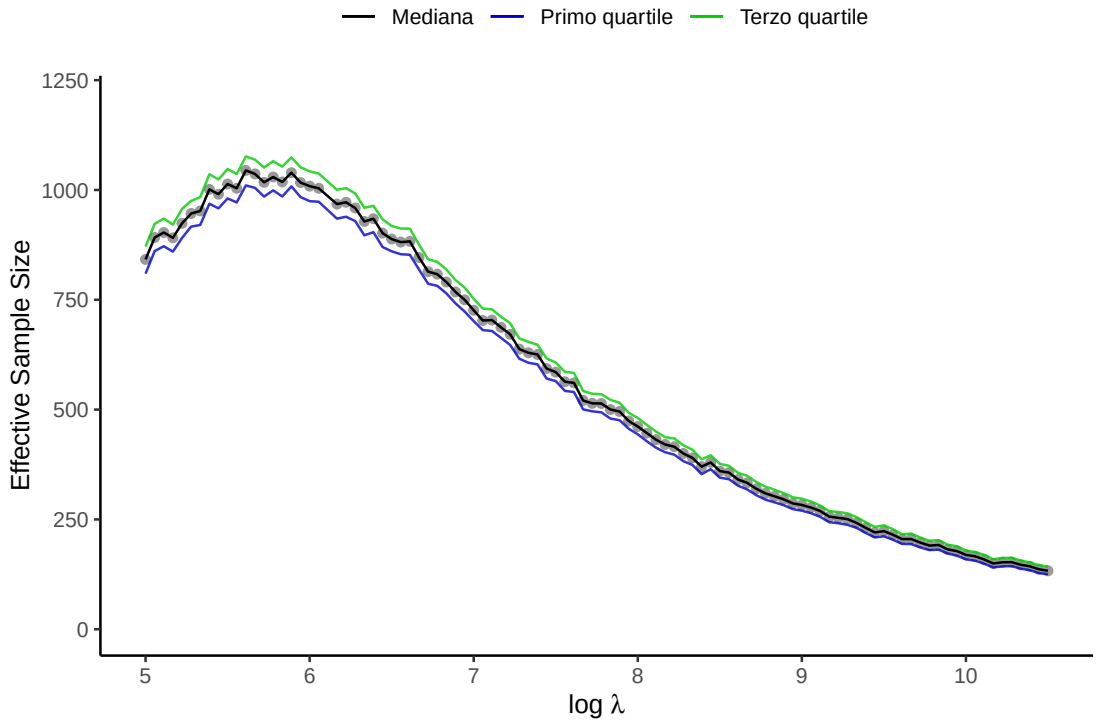


Figura 11: Effective sample size al variare del parametro di precisione, espresso in scala logaritmica

Alte penalizzazioni comportano una ESS ridotta. Come atteso, in altri

termini, quando la penalizzazione diventa molto elevata i coefficienti tendono ad avere un intervallo di variazione intorno allo zero sempre più ridotto, ne consegue che le transizioni ammissibili tra gli stati della catena si riducono notevolmente, e i candidati mostrano conseguentemente un aumento dell'autocorrelazione e una diminuzione dell'ESS.

5.1.1 Diagnostica relativa alla componente adattiva

In Figura 12 si riportano le variazioni dello *step-size* in funzione delle iterazioni di *burn-in* per i quattro valori di λ specificati nel paragrafo 5.1. Sono stati scelti i medesimi valori di λ in quanto rappresentativi di tutta la sequenza oltre che per coerenza espositiva.

I risultati mostrano che l'assunzione di *diminishing adaptation* risulta rispettata: all'aumentare del numero di iterazioni gli aggiustamenti apportati al parametro di *step-size* tendono ad essere sempre più marginali. A titolo di esempio, la differenza tra gli ultimi due valori dello *step-size* per $\log \lambda = 10.39$ risulta pari a 0.0046 a fronte di una differenza in corrispondenza delle prime iterazioni di 0.01. Valori simili si ottengono con riferimento ai valori $\log \lambda = 9.44$ e $\log \lambda = 7.94$. Naturalmente, proseguendo per più iterazioni, il campionamento per $\log \lambda = 5.56$ mostra una differenza fra gli ultimi due valori di *step-size* ancora più marginale, pari a circa 0.0026.

Un'altra considerazione rilevante riguarda l'iterazione in cui cessa il *burn-in*, avendo reso anche il numero di iterazioni di *burn-in* variabile in funzione della stabilizzazione del parametro di *step-size*, per dettagli si rimanda al paragrafo 4.5. Si può osservare come al diminuire della penalizzazione cresca il numero di

iterazioni di *burn-in* necessario per giungere alla stabilizzazione del parametro di *step-size*. Ciò indica che per raggiungere la convergenza l'algoritmo necessita di più iterazioni. Tale risultato è in linea con le aspettative teoriche: più si procede nella direzione di imporre sempre meno penalizzazione sui coefficienti più ci si avvicina a condizioni di quasi non identificabilità. In particolare la matrice di varianza-covarianza della *proposal*, pur non essendo singolare, è prossima alla singolarità. Ciò induce più variabilità sui candidati proposti, sicché se questo da un lato si traduce in un aumento dell'ESS, come si evince dalla Figura 11, dall'altro induce l'algoritmo a proporre candidati non sempre ottimali e dunque risulta necessario incrementare il numero di iterazioni al fine di stabilizzare il parametro di *step-size* e ottenere una probabilità di accettazione prossima a quella asintoticamente ottimale.

I risultati relativi all'iterazione in corrispondenza della quale viene applicata la *stopping rule* dimostrano in modo più eloquente quanto appena enucleato. Infatti, in media, sono sufficienti 55106 iterazioni di *burn-in* per stabilizzare il parametro di *step-size*. Si consideri inoltre che il terzo quartile della distribuzione delle iterazioni di arresto è pari a 56400. Ciò indica che per il 75% dei valori di λ sono sufficienti meno di 70000 iterazioni di *burn-in* per giungere a convergenza. Solo per 18 valori sono necessarie più di 70000 iterazioni per convergere in fase di *burn-in*. Di tali valori il 72% si trova tra i 25 valori di λ più bassi. Vi sono infine 4 valori di λ tali da raggiungere il limite massimo imposto per le iterazioni di *burn-in* e sono tutti e 4 tra gli 8 valori di λ più bassi specificati. Questi risultati sembrano suggerire che rendere adattivo anche il numero di iterazioni di *burn-in* risulti utile per ridurre l'onere computazionale senza tuttavia rinunciare alla convergenza.

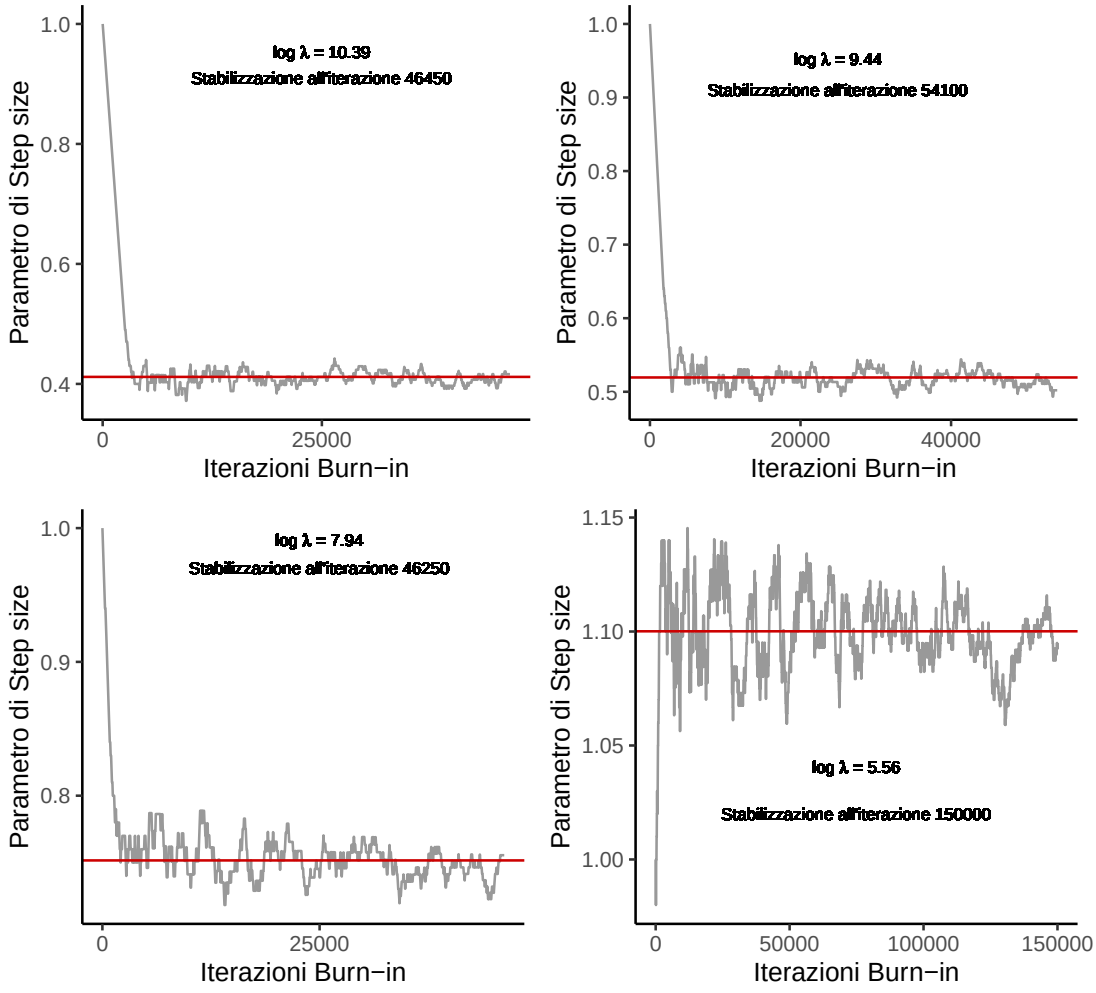


Figura 12: Andamento del parametro di step-size all'aumentare delle iterazioni, per differenti valori del parametro di precisione

5.2 Miglior modello e *Model Averaging*

Prima di analizzare i risultati relativi al miglior modello e al modello ottenuto tramite BMA, si fornisce un esempio delle performance di accuratezza prognostica di quattro modelli rappresentativi, corrispondenti ai quattro distinti valori di λ considerati in precedenza, in modo da evidenziare le differenze distribuzionali degli indici al variare della penalizzazione. In Figura 13 vengono riportati gli

indici di accuratezza prognostica per i quattro distinti modelli. Si osservi che sulle ordinate i modelli sono identificati riportandoli in ordine di penalizzazione decrescente.

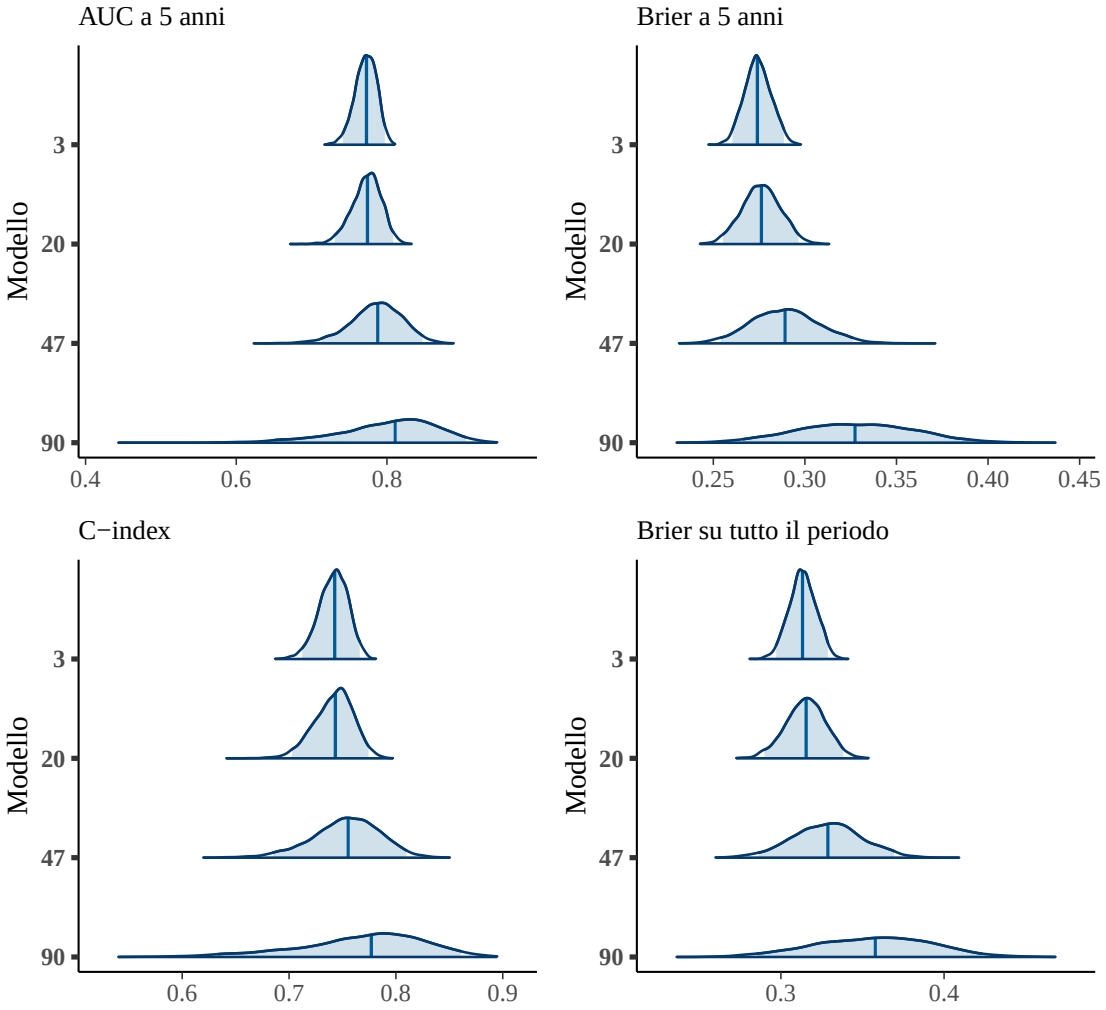


Figura 13: Distribuzione a posteriori degli indicatori di prognosi, con intervalli di credibilità al 95%, per quattro modelli rappresentativi. Sulle ordinate vengono riportati gli indici delle penalizzazioni in ordine decrescente: tanto più basso l'indice della penalizzazione, tanto più alta è la penalizzazione. In particolare: 3 corrisponde al terzo valore del parametro di precisione, 20 al ventesimo, 47 al quarantasettesimo e 90 al novantesimo.

Si noti che, al diminuire della penalizzazione, le distribuzioni degli indici di accuratezza prognostica diventano più ampie, così come i rispettivi intervalli di credibilità. Osservando le distribuzioni riportate in Figura 13 si rileva come, al

diminuire della penalizzazione, le distribuzioni presentino una mediana centrata su valori più elevati ma, al contempo, una variabilità molto più elevata, il che si traduce in una maggior incertezza intorno al valore medio, o mediano, a posteriori. Valori elevati di λ si associano, pertanto, a un minor grado di incertezza intorno al valore medio, o mediano, a posteriori. Contrariamente, valori più bassi di penalizzazione si associano a un grado di incertezza più elevato.

Da quanto emerso è ragionevole aspettarsi che il miglior modello verrà identificato in corrispondenza di valori del parametro di precisione elevati, in quanto, essendo il WAIC una misura che asintoticamente approssima la LOO-CV, esso tenderà a privilegiare modelli che abbiano maggior garanzia di generalizzabilità e, conseguentemente, variabilità più bassa.

5.2.1 Miglior modello

In Figura 14 si riporta l'andamento del DIC e del WAIC in funzione delle corrispondenti misure del numero di parametri effettivi. In prima istanza si osservi che il DIC e il WAIC sono concordi nella scelta del miglior modello. Inoltre si osservi che, come atteso, il miglior modello selezionato a posteriori è un modello che presenta una penalizzazione molto elevata: $\log \lambda^{opt} = 10.39$. Ciò conferma che ai fini della generalizzabilità e dell'applicabilità del modello su nuovi dati i due criteri prediligono un modello che, pur presentando AUC a 5 anni e C-index mediamente inferiori rispetto a modelli con penalizzazioni più blande, si caratterizza per una variabilità e un'incertezza intorno alle metriche di accuratezza prognostica inferiore.

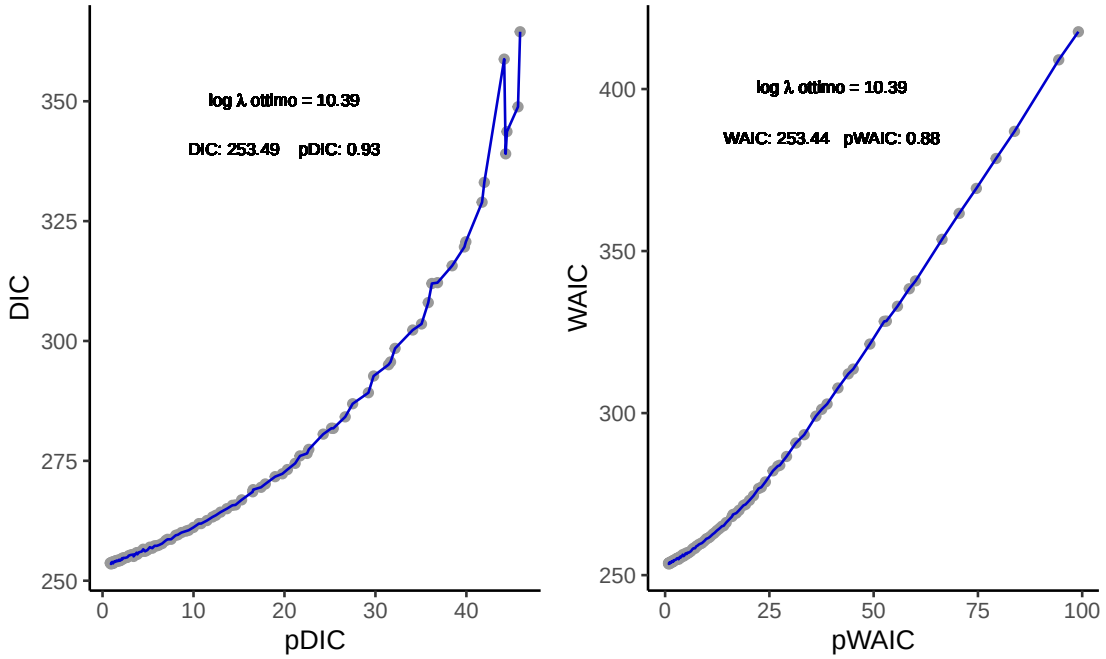


Figura 14: Andamento del DIC e del WAIC in funzione dei gradi di libertà rispettivi, i valori riportati indicano i valori in corrispondenza dei quali viene selezionata la penalizzazione

Un'altra osservazione riguarda il numero di parametri effettivi calcolati per il DIC e per il WAIC. Se per il DIC il massimo del pDIC non supera i 50, per il WAIC si giunge fino a circa 98.9, in corrispondenza della penalizzazione più bassa $\log \lambda = 5$. Ciò potrebbe risultare, dal punto di vista interpretativo, abbastanza anomalo dato che il numero di osservazioni ammonta a 91. In realtà, a differenza del pDIC che apparentemente sembra fornire risultati più in linea con quanto ci si aspetterebbe, il pWAIC risulta più elevato poiché il pDIC non tiene conto pienamente dell'incertezza associata ai parametri stimati. Infatti il pDIC considera la log verosimiglianza nel punto di ottimo a posteriori come valore di riferimento per calcolare il numero di parametri effettivi:

$$p_{\text{DIC}} = 2\left(\log \mathscr{L}\left(\mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta} \mid \tilde{\boldsymbol{\vartheta}}\right) - \frac{1}{S}\sum_{s=1}^{S} \log \mathscr{L}\left(\mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta} \mid \boldsymbol{\vartheta}^s\right)\right).$$

Il pWAIC, invece, considera non già il valore puntuale della log verosimiglianza nel punto di ottimo a posteriori ma il log della media della verosimiglianza valutata in corrispondenza di ciascun vettore di parametri campionato, tenendo dunque in considerazione l'incertezza associata a ciascun vettore di parametri:

$$p_{\text{WAIC}} = 2\sum_{i=1}^{n}\left(\log\left(\frac{1}{S}\sum_{s=1}^{S}\mathscr{L}(\boldsymbol{x}_i \mid \boldsymbol{\vartheta}^s)\right) - \frac{1}{S}\sum_{i=1}^{S}\log\mathscr{L}(\boldsymbol{x}_i \mid \boldsymbol{\vartheta}^s)\right).$$

Risulta pertanto chiaro che più la penalizzazione diminuisce più alta sarà l'incertezza associata alla stima e, conseguentemente, più elevato sarà il pWAIC. Nella condizione limite, in cui la penalizzazione scelta porti a una matrice di varianza-covarianza della *proposal* non singolare ma prossima alla singolarità, il pWAIC incorpora tale informazione come un aumento dell'incertezza. Conseguentemente, per il minimo valore di penalizzazione specificato, $\log\lambda = 5$, il pWAIC incorpora l'informazione di elevata incertezza associata alla stima. In altri termini, un pWAIC superiore al numero delle osservazioni indica che il modello è quasi non identificabile, tanto da attribuire un numero di parametri effettivi più elevato del numero di osservazioni. La presenza di un unico valore di pWAIC che ecceda il numero di osservazioni è dovuto al lavoro di specificazione della sequenza che è stato fatto prima di procedere con la valutazione modellistica. Il processo di specificazione attraverso *trial and errors* ha consentito di individuare il limite inferiore di λ.

Venendo invece all'interpretazione del numero di parametri effettivi associati al miglior modello è importante chiarire a cosa corrisponda il pWAIC. A tal fine

si ricorda la struttura delle distribuzioni a priori assegnate ai parametri:

$$\boldsymbol{\beta} \sim \mathcal{N}_{p+1}\left[\mathbf{0}, \begin{pmatrix} \sigma_\beta^2 \mathbf{I}_2 & \mathbf{0} \\ \mathbf{0} & \frac{1}{2\lambda}\mathbf{I}_{p-1} \end{pmatrix}\right], \quad \xi \sim \mathcal{N}\left(-\frac{\log(\sigma_\alpha^2+1)}{2}, \log(\sigma_\alpha^2+1)\right).$$

Per 2144 parametri viene posto un vincolo sulla matrice di varianza-covarianza della distribuzione a priori. Tale vincolo impone che la varianza sia estremamente ridotta in modo tale che i parametri radiomici siano penalizzati e molto vicini a 0, al fine di rendere possibile la regressione (si consideri, a titolo di esempio, il λ ottimo: $\lambda = e^{10.39}$). Vi sono poi 3 parametri non soggetti a penalizzazione, quindi senza vincoli: l'intercetta, il predittore lineare del Sarculator e il log del parametro di *shape* della Weibull. Si osservi che il numero di parametri effettivo calcolato con il pWAIC risulta pari a 0.88. Tale valore è dovuto alla struttura delle distribuzioni a priori. Infatti, essendo la numerosità campionaria contenuta, l'informazione a priori posta sui 2144 parametri radiomici domina globalmente l'effetto rilevato dal pWAIC.

Al fine di evidenziare quali siano le distribuzioni a posteriori dei parametri in corrispondenza del miglior modello, in Figura 15 si riporta la distribuzione di un sottocampione di 100 parametri. Si osservi che, sebbene dal punto di vista grafico le distribuzioni sembrino molto variabili, osservando con più attenzione l'asse delle ascisse si evince come in realtà esse presentino un basso grado di variabilità. Ciò è coerente con il valore di penalizzazione elevato associato al miglior modello.

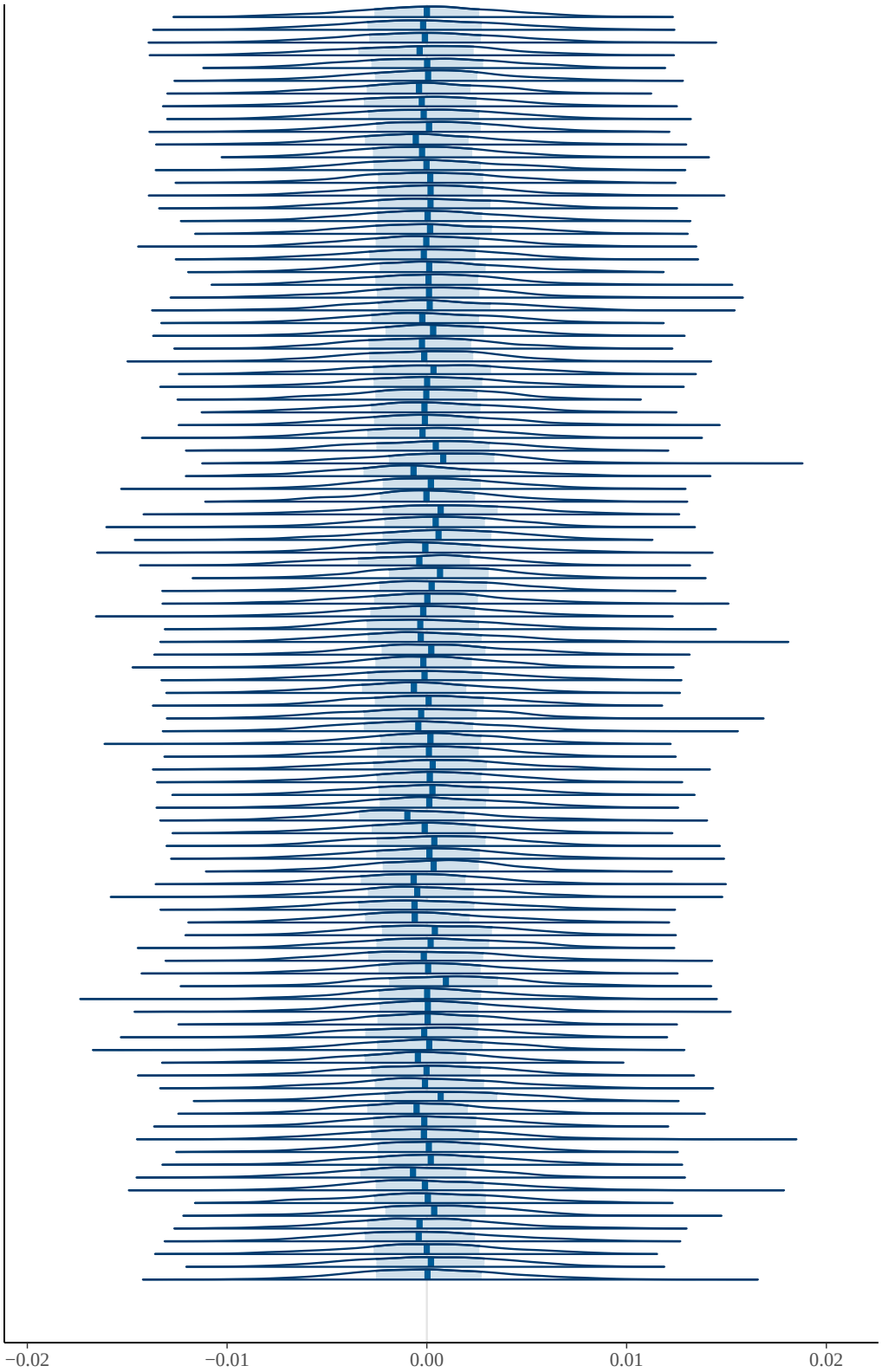


Figura 15: Distribuzione a posteriori di un sottocampione casuale di 100 parametri del modello ottimo, identificato con il WAIC

In Figura 16 si riportano le distribuzioni a posteriori di alcune variabili radiomiche, scelte casualmente, e delle variabili di interesse, ovvero il parametro di *shape* della Weibull e il predittore lineare derivato dal Sarculator.

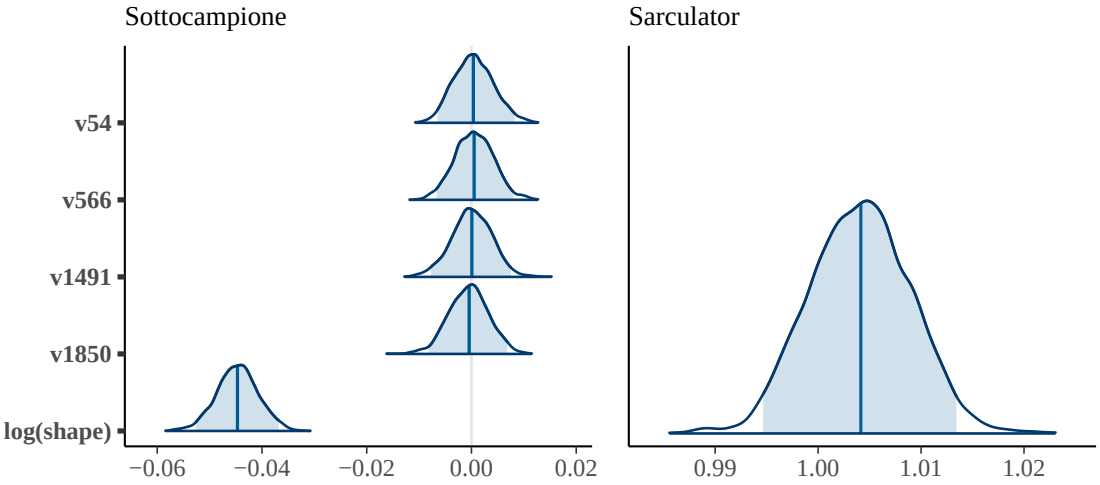


Figura 16: Distribuzione a posteriori di alcuni specifici parametri del modello ottimo, identificato con il WAIC, e relativi intervalli di credibilità al 95%

Da quanto emerge dalla Figura 16, il miglior modello effettua uno *shrinkage* marcato sui parametri radiomici, le cui distribuzioni sono infatti centrate sullo zero, mentre per quanto riguarda il parametro di *shape*, la distribuzione si distanzia dallo zero, così come quella del Sarculator. Si osservi altresì la bassa variabilità distribuzionale di tutti i parametri in oggetto, il che riduce di molto il grado di incertezza intorno al valor medio.

In Figura 17 si riportano invece le distribuzioni degli indici di accuratezza prognostica associate al miglior modello. Si osservi che la variabilità intorno al valore mediano risulta molto bassa, si ha quindi un basso grado di incertezza circa l'accuratezza prognostica a livello di tutti gli indici. In particolare, il modello presenta un AUC mediano a 5 anni di circa 0.773, un Brier Score mediano a

5 anni di circa 0.274, un C-index mediano pari a circa 0.742, e un Brier Score valutato su tutto il periodo di circa 0.313.

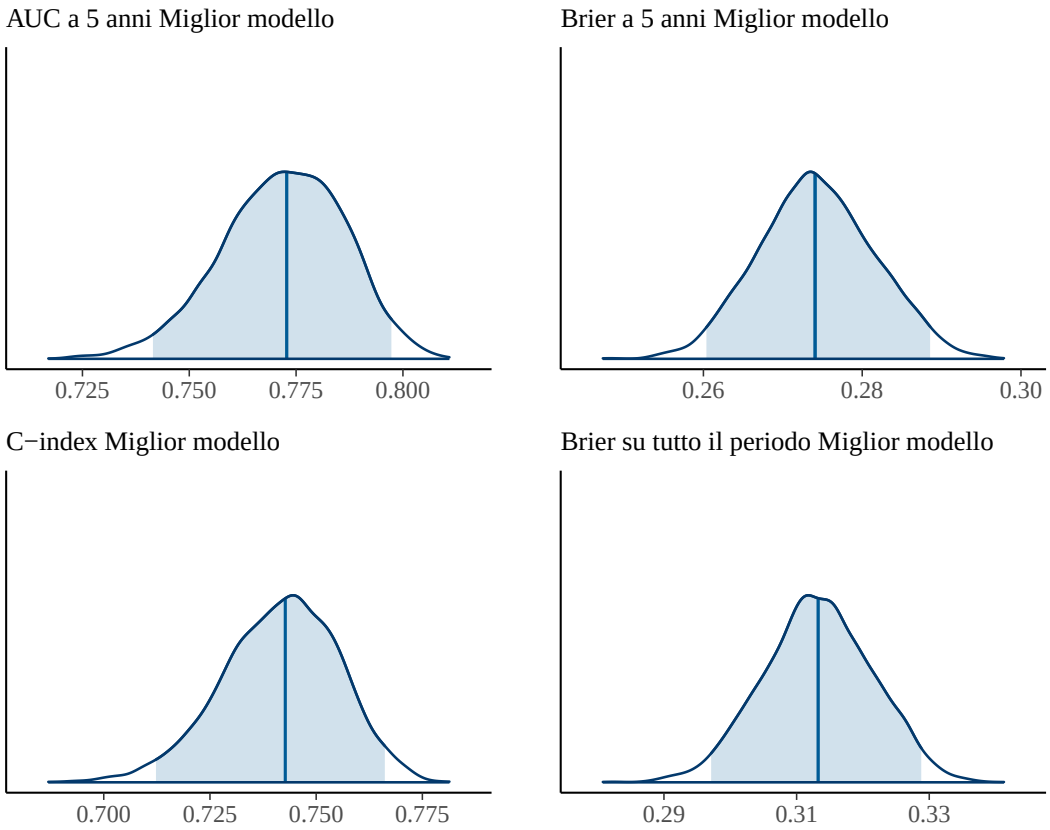


Figura 17: Distribuzione a posteriori degli indici di performance del modello ottimo, identificato con il WAIC, e relativi intervalli di credibilità al 95%

Tutte le misure di accuratezza risultano buone, sia in termini di accuratezza delle previsioni sia in termini di calibrazione del modello.

Nonostante i buoni risultati registrati in termini di *performance* e variabilità dal miglior modello, la selezione di un singolo modello non tiene conto dell'incertezza insita nel processo di selezione stesso. Si è quindi ritenuto opportuno adottare, per la costruzione del modello, un approccio più robusto, che consentisse di incorporare l'incertezza derivante dal processo di selezione. Per perseguire tale scopo si è proceduto con il *Bayesian Model Averaging*.

5.2.2 *Model Averaging*

L'idea di utilizzare il BMA nasce dal fatto che si cerca un modello in grado di tener conto dell'incertezza associata a tutti i modelli candidati per la selezione. Dalla Figura 14 risulta chiaro come non vi sia una distinzione netta tra il miglior modello e altri modelli con valori di penalizzazione elevati. I primi 20 modelli risultano avere valori di WAIC molto simili al modello ottimo, che fa registrare un WAIC di 253.44 (il ventesimo modello fa registrare un WAIC di 257.6). Si intuisce che per valori simili di WAIC scegliere il modello migliore piuttosto che il secondo modello non porta a un cambiamento rilevante. Selezionare un modello in questo contesto potrebbe risultare una soluzione non ottimale. Da qui l'idea di sfruttare la relazione (30) per calcolare la probabilità a posteriori di ciascun modello e ponderare le stime, le previsioni, e gli indici di accuratezza per tali probabilità in modo da ottenere un modello che tenga conto del grado di incertezza associato a ciascun modello.

Si osservi altresì che procedere in questo modo consente di tenere conto anche dei modelli che mostrano indici di accuratezza prognostica più elevata. Facendo riferimento alla Figura 13, risulta chiaro che ponderare, a titolo di esempio, il novantesimo modello con la relativa probabilità a posteriori e incorporarlo all'interno di un modello mediato può portare a un incremento degli indici di accuratezza prognostica.

In Figura 18 vengono riportate le distribuzioni a posteriori di alcuni parametri dopo l'applicazione del BMA.

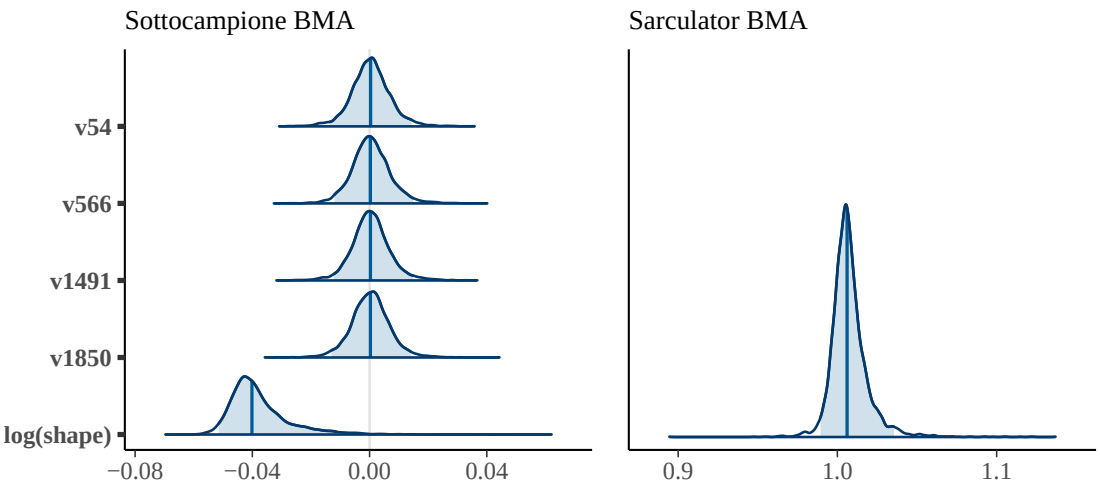


Figura 18: Distribuzione a posteriori di alcuni specifici parametri del BMA, ottenute ponderando ciascun parametro per la probabilità a posteriori di ciascun modello, e relativi intervalli di credibilità al 95%

Per quanto attiene ai parametri di interesse, al Sarculator viene attribuito un valore medio di circa 1.0075 a cui corrisponde un *Hazard Ratio* di circa 2.7389. Il parametro di forma della Weibull derivante dal BMA si ottiene esponenziando il parametro ξ dato che, per fini di campionamento, si è posto $\xi = \log \alpha$. Dunque il parametro di forma stimato risulta pari a circa 0.9628 e, per quanto enucleato nel paragrafo 2.4, ciò implica che il rischio decresce nel tempo. Il risultato è in linea con la letteratura e con il campione oggetto d'analisi dal momento che i pazienti affetti da sarcomi sono soliti sviluppare l'evento entro i primi 5 anni dalla diagnosi. Superata la soglia dei 5 anni gli eventi diminuiscono considerevolmente. In altri termini, il tempo funge da fattore protettivo relativamente all'insorgenza dell'evento.

Sembra controintuitivo ma più il tempo passa meno è probabile che un paziente affetto da sarcomi agli arti muoia, giacché le morti si verificano quasi interamente entro i primi 5 anni.

In Figura 19 si riportano le distribuzioni degli indici di accuratezza prognostica

ottenuti tramite BMA.

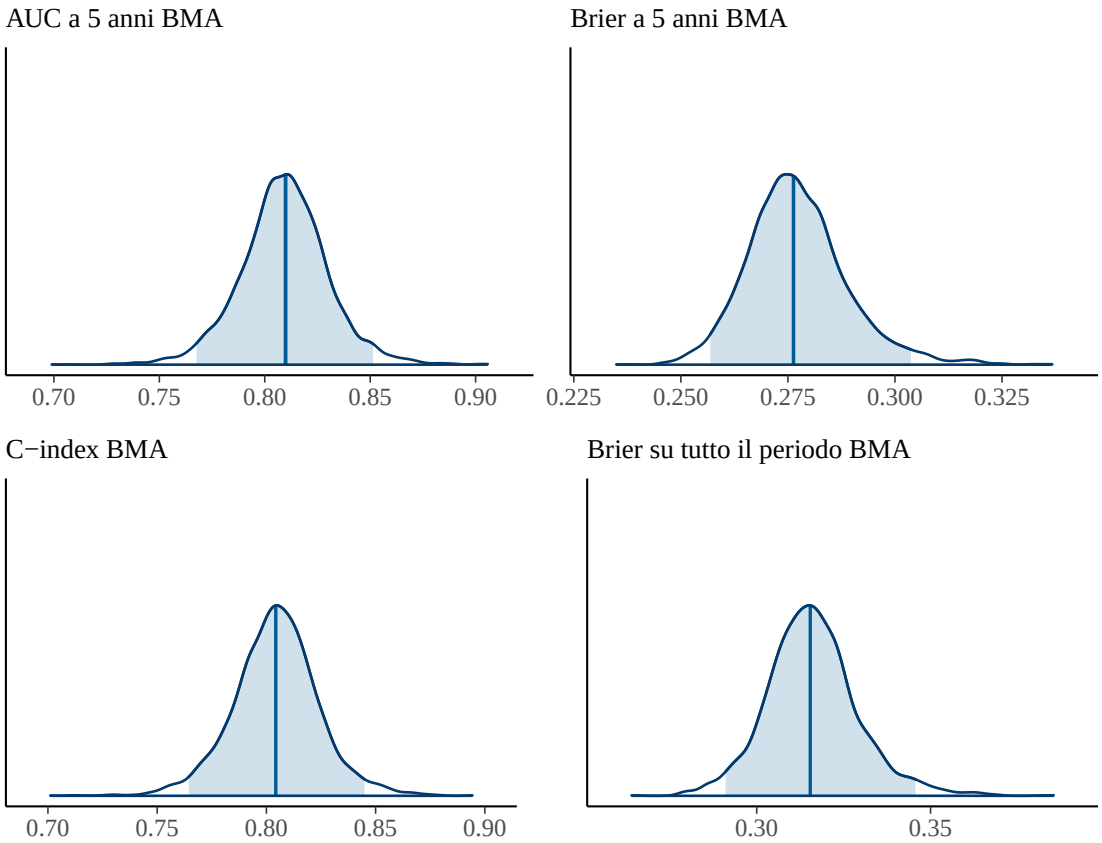


Figura 19: Distribuzione a posteriori degli indici di performance del BMA e relativi intervalli di credibilità al 95%

Si osservi che per quanto riguarda l'AUC a 5 anni e il C-index, Il BMA porta a un incremento dei valori medi e mediani. L'incremento risulta considerevole soprattutto se si guarda al C-index, che guadagna circa 6 punti percentuali rispetto al miglior modello. Per fornire un'idea più chiara delle differenze intercorrenti tra il BMA e il miglior modello in termini di accuratezza prognostica si rimanda al paragrafo 5.3.

5.3 Confronto tra Miglior modello e BMA

Verranno ora confrontati i risultati relativi al miglior modello e al BMA. In Figura 20 vengono riportate le distribuzioni di alcuni coefficienti del miglior modello e del BMA. In generale si può apprezzare come il BMA tenda a produrre, per quanto riguarda i parametri soggetti a penalizzazioni, distribuzioni più simmetriche e leggermente più variabili, in quanto risultato di misture di normali. Per quanto riguarda i coefficienti di interesse le stime sono sostanzialmente equivalenti per quanto riguarda il Sarculator (l'intervallo di credibilità al 95% del miglior modello è pari a $[0.9947, 1.0134]$, rispetto all'intervallo del BMA che è pari a $[0.9894, 1.0355]$).

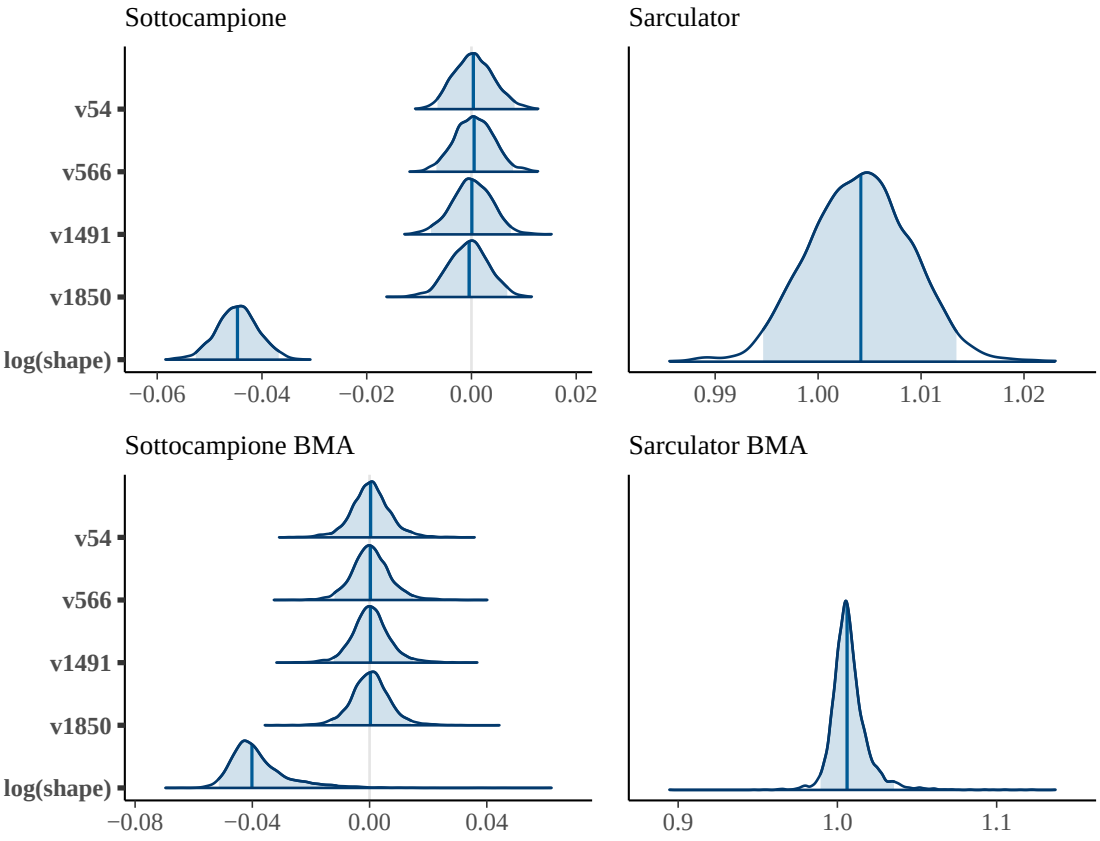


Figura 20: Distribuzioni a posteriori di alcuni dei parametri ottenute con BMA a confronto con le distribuzioni dei medesimi parametri del modello ottimo, identificato con il WAIC, e relativi intervalli di credibilità al 95%

Il parametro di forma della Weibull stimato con BMA risulta avere una distribuzione con una coda più pesante a destra, ma presenta un valor medio prossimo a quello del modello ottimo. In Figura 21 si riportano le distribuzioni a posteriori degli indici di accuratezza prognostica ottenuti tramite BMA a confronto con quelle del miglior modello.

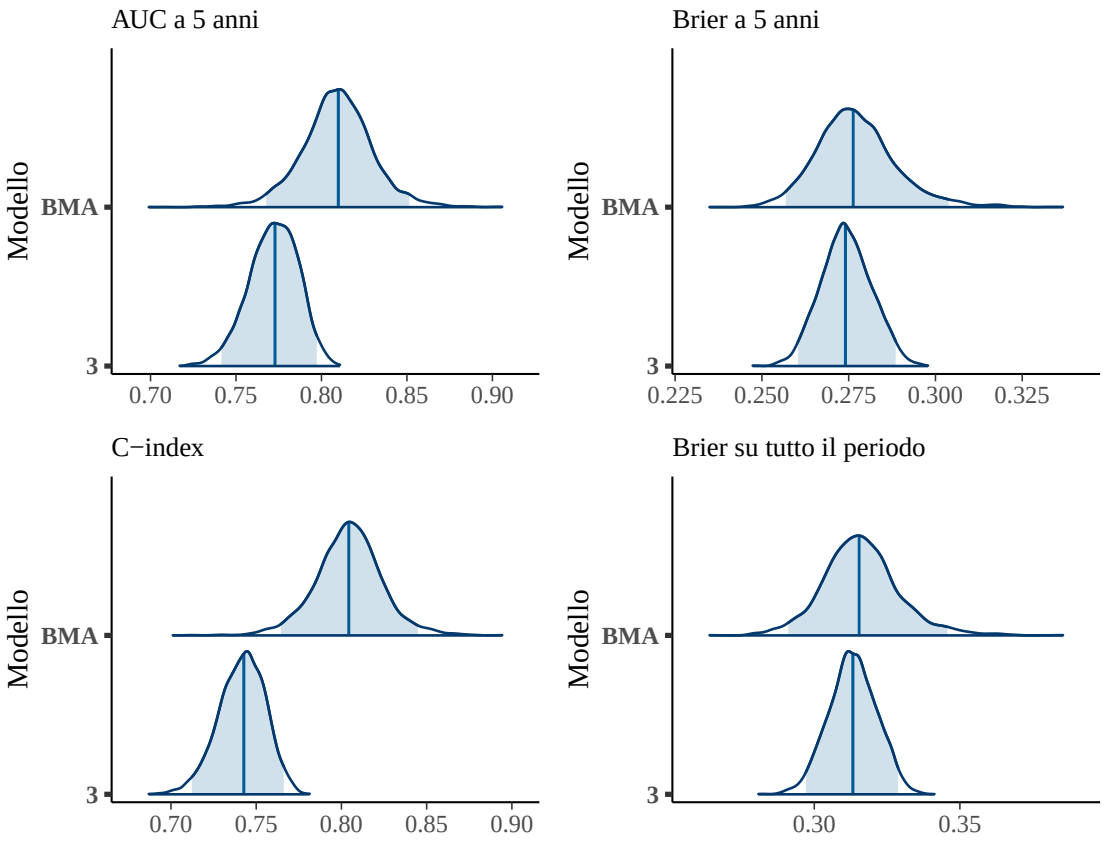


Figura 21: Distribuzione a posteriori degli indici di accuretezza per il BMA e per il modello ottimo a confronto

La differenza in termini di AUC, e soprattutto di C-index, risulta considerevole. Si rileva invece un lieve peggioramento degli indici di calibrazione, si veda la Tabella 5 per ulteriori dettagli in cui vengono riportati i valori medi degli indici di accuratezza e i rispettivi intervalli di credibilità al 95%. Tenendo conto

dell'incertezza associata a tutti i modelli, il BMA presenta distribuzioni con code più lunghe e si riscontra un lieve incremento degli intervalli di credibilità.

Metrica	BMA		Modello ottimo ($\log \lambda = 10.39$)	
AUC a 5 anni	0.809	$(0.768, 0.851)$	0.772	$(0.741, 0.797)$
Brier Score a 5 anni	0.277	$(0.257, 0.304)$	0.274	$(0.260, 0.289)$
Brier Score totale	0.316	$(0.291, 0.346)$	0.313	$(0.297, 0.329)$
C-index	0.804	$(0.764, 0.845)$	0.743	$(0.713, 0.771)$

Per ciascun indice si riporta la media a posteriori e l'intervallo di credibilità al 95%

Tabella 5: Performance del modello ottimo e del modello ottenuto tramite *Bayesian Model Averaging*

I risultati mostrano che il BMA migliora l'AUC a 5 anni e il C-index e che tiene conto anche dell'incertezza di tutti i modelli. Ciò rende il BMA meno soggetto alla sovrastima. Il lieve peggioramento in termini di Brier Score a 5 anni e su tutto il periodo per quanto concerne il valor medio a posteriori risulta trascurabile: con riferimento alle stime riportate in Tabella 5 si può concludere che dal punto di vista della calibrazione ciò che varia sono gli intervalli di credibilità che divengono più ampi per tener conto dell'incertezza associata a ciascun modello.

Pare dunque chiaro che, complessivamente, propendere per il BMA sia stata una scelta che ha portato considerevoli vantaggi dal punto di vista della gestione dell'incertezza, generando un modello meno soggetto alla sovrastima e presentante un aumento dell'accuratezza prognostica.

In Figura 22 si riporta la calibrazione a 5 anni e su tutto il periodo considerato, mediata con BMA. I rischi sono stati suddivisi in quartili. Si osservi che il modello pare ben calibrato, con una lieve tendenza alla sottostima per le previsioni del rischio su tutto il periodo d'osservazione.

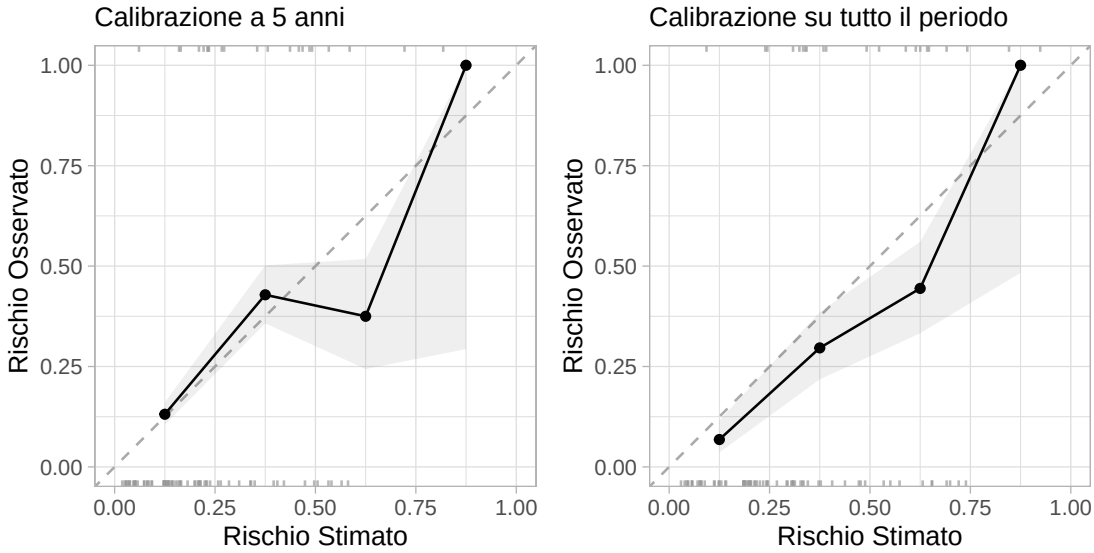


Figura 22: Calibrazione a 5 anni e su tutto il periodo ottenuta con BMA, e intervalli di credibilità al 95%

Dalla Figura 22 è possibile dedurre che cosa abbia indotto l'aumento del C-Index. Essendo esso, infatti, una misura di concordanza globale del modello beneficia della calibrazione su tutto il periodo. Infatti, se si considera tutto il periodo di osservazione, il modello sembra sottostimare leggermente i rischi in modo più sistematico rispetto a quanto avviene a 5 anni; al contempo però, la calibrazione tra il secondo e il terzo quartile di rischio risulta decisamente migliore rispetto alla medesima valutata a 5 anni. Diversamente, l'AUC a 5 anni risente della minor calibrazione del modello a 5 anni tra il secondo e il terzo quartile di rischio e, conseguentemente, fa registrare un aumento più contenuto.

In ultima istanza si procederà con la verifica dell'obiettivo principale del presente lavoro. Si chiarirà perciò se la radiomica aggiunge potere prognostico al Sarculator. Per rispondere a tale quesito si confronterà il modello che include solo il Sarculator con i modelli contenenti sia il Sarculator che la radiomica come descritto al paragrafo 4.8.

5.4 Scelta del modello

In Figura 23 vengono riportati i *trace plots* e le distribuzioni a posteriori dei parametri del modello contenente solo il Sarculator.

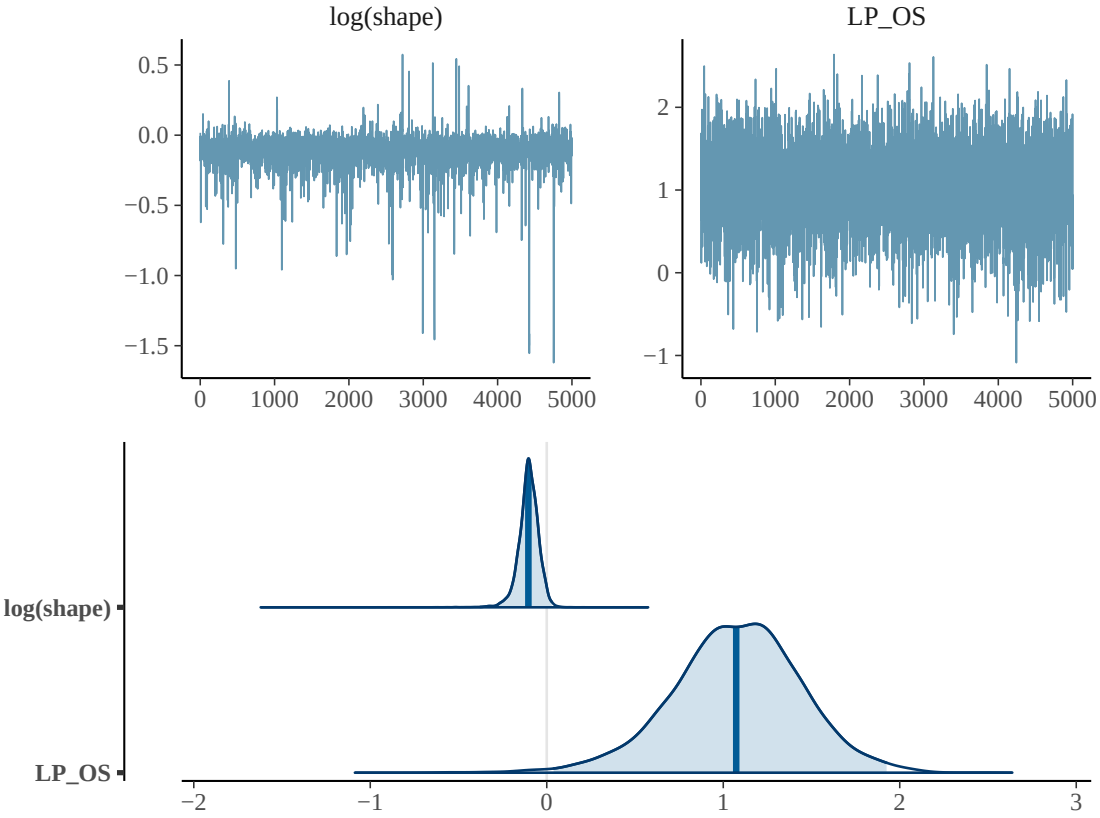


Figura 23: Trace plots e distribuzione a posteriori dei parametri di interesse del modello con solo il Sarculator e relativi intervalli di credibilità al 95%

Per la stima dei parametri del modello è stato applicato il medesimo algoritmo applicato agli altri modelli presentanti parametri penalizzati. In questo caso le uniche distribuzioni a priori rimangono quelle per ξ, β_0, β_1, con β_1, parametro del predittore lineare del Sarculator, che ovviamente sono distribuzioni a priori vaghe, che non inducono alcuna penalizzazione.

I *trace plots* manifestano l'avvenuta convergenza, le distribuzioni mostrano molta incertezza riguardo la stima del coefficiente associato al Sarculator. Per quanto attiene alla stima del log del parametro di forma, essa sembra non distanziarsi in modo sostanziale dallo zero. Si è quindi proceduto calcolando Il WAIC del modello che è risultato pari a 266.79. Si osservi che già per il valore di WAIC fatto registrare, paragonandolo agli altri valori precedentemente ottenuti, si potrebbe concludere che la radiomica aggiunge potere predittivo, poiché sono preferibili molti modelli che includono le variabili radiomiche con penalizzazione dei coefficienti ad esse associati. Comunque, per essere più rigorosi e certi del risultato, si è proceduto come descritto al paragrafo 4.8, attribuendo dunque una probabilità a priori al modello contenente solo il Sarculator pari al 50%, dal momento che esso risulta un indice prognostico già validato, e si è ripartito uniformemente l'altro 50% sui 100 modelli. Ciascuno dei 100 modelli, a priori, presenta quindi una probabilità di essere il modello ottimo pari allo 0.5%. Il contesto così costruito è un contesto che predilige molto il modello con solo il Sarculator, a priori e, pertanto, per poter propendere per l'ipotesi che la radiomica aggiunta al Sarculator risulti effettivamente rilevante a fini prognostici è necessario che vi sia una forte evidenza a sostegno di tale ipotesi.

Stabilite le probabilità a priori, si è proceduto con il calcolo delle probabilità a posteriori per ciascun modello secondo la relazione (31). Si è quindi proceduto conducendo il test di ipotesi, di cui si riporta per semplicità espositiva il sistema corrispondente nelle due formulazioni equivalenti.

$$\mathrm{H}_0 : \mathcal{M}_0 \quad \text{è sufficiente} \quad vs \quad \mathrm{H}_1 : \exists \lambda \mid \mathcal{M}_\lambda \quad \text{è migliore di} \quad \mathcal{M}_0.$$

$$\mathrm{H}_0 : \pi\left(\mathcal{M}_0 \mid \mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta}\right) \geq \frac{1}{2} \;\middle|\; \pi\left(\mathcal{M}_0\right) = \frac{1}{2} \quad vs \quad \mathrm{H}_1 : \pi\left(\mathcal{M}_0 \mid \mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta}\right) < \frac{1}{2} \;\middle|\; \pi\left(\mathcal{M}_0\right) = \frac{1}{2}$$

Si osservi che la prima formulazione risulta utile a fini concettuali. La seconda formulazione è equivalente alla prima, tuttavia consente di formalizzare e rendere operativo il sistema di ipotesi.

Il risultato ottenuto dall'applicazione del test è il seguente:

$$\pi\left(\mathcal{M}_0 \mid \mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta}\right) \approx 0.0049.$$

Dunque, dal momento che $\pi\left(\mathcal{M}_0 \mid \mathbf{X}, \boldsymbol{t}, \boldsymbol{\delta}\right) < \frac{1}{2}$ si rifiuta l'ipotesi nulla per cui il Sarculator da solo sia preferibile a un modello che incorpori il Sarculator stesso e le informazioni derivanti dalla radiomica. Pertanto, è lecito affermare che la radiomica aggiunge potere prognostico al Sarculator.

A completamento del risultato appena ottenuto, che indica la rilevanza delle variabili radiomiche a fini prognostici, si è deciso, fisse restando le probabilità a priori specificate nel sistema di ipotesi, di fornire una rappresentazione grafica che consentisse di paragonare, in termini di probabilità a posteriori, tutti i modelli con quello contenente solo il Sarculator. A tal fine, in Figura 24 vengono riportate le probabilità a posteriori ottenute, ordinate in senso crescente, per ciascun modello. Dal grafico si può notare come il modello che presenta solo il Sarculator abbia una probabilità inferiore a molti modelli che includono anche la radiomica, nonostante esso a priori fosse stato fortemente pesato rispetto agli altri modelli. Si ricordi

infatti che la probabilità a priori attribuita al modello con solo il Sarculator è del 50%, contro lo 0.5% attribuita a ciascuno degli altri modelli. Si può inoltre osservare che quando la penalizzazione diventa consistente la radiomica riesce ad aggiungere informazioni, diversamente, per penalizzazioni blande prevale il rumore, e questo è in linea con quanto ci si aspetta per le variabili di natura radiomica, caratterizzate da molto segnale poco informativo. Ne consegue che, per poter aggiungere potere prognostico, i coefficienti associati a tali variabili devono necessariamente essere penalizzati per sopprimere il più possibile il rumore.

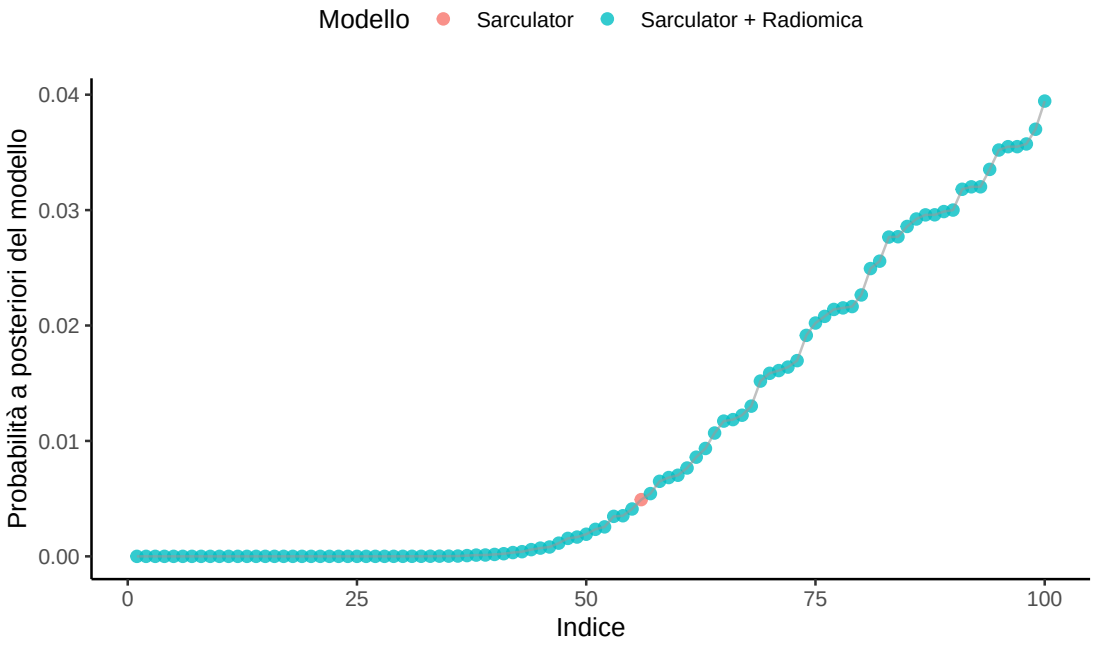


Figura 24: Grafico delle probabilità a posteriori di ciascun modello. Al modello con solo il Sarculator è stata attribuita una probabilità a priori pari al 50%, la restante percentuale è stata ripartita uniformemente tra i modelli penalizzati.

In particolare, il modello con solo il Sarculator, su 101 modelli possibili, risulta al quarantaquattresimo posto. In altri termini, a livello di singolo modello, i modelli che presentano una penalizzazione maggiore o uguale a $\lambda = e^{8.17}$ sono tutti strettamente preferibili al modello con solo il Sarculator. Inoltre, se a priori

la probabilità che il modello con solo il Sarculator fosse ottima ammontava al 50%, a posteriori la probabilità che sia ottimo ammonta allo 0.49%, contro la massima probabilità a posteriori per il miglior modello pari al 4.00%.

In conclusione, il test di ipotesi condotto fornisce un'indicazione chiara a sostegno dell'inclusione dell'informazione derivante dalle variabili radiomiche all'interno di un modello prognostico che contenga anche il Sarculator, pertanto è lecito concludere che la radiomica risulta un elemento che aumenta il potere prognostico del Sarculator.

5.5 Curve di Sopravvivenza

Verificato che la radiomica e il Sarculator sono preferibili al modello con solo il Sarculator, si produrranno ora le curve di sopravvivenza per due pazienti, presi a titolo di esempio, per mostrare come si comporta il BMA da un punto di vista previsivo e verrà fornita anche la curva di sopravvivenza globale stimata tramite BMA. Nell'analisi di sopravvivenza la costruzione di una curva di sopravvivenza per un paziente di cui si è osservato l'esito non risulta semplice come fare previsioni in altri contesti di regressione o classificazione.

Solitamente oltre alle curve stimate per paziente, nell'ambito dell'analisi di sopravvivenza, si è interessati alla curva di sopravvivenza stimata globale che indica, mediamente, la sopravvivenza di una popolazione nel tempo, tenuto conto dei fattori che incidono direttamente sulla sopravvivenza stessa. Si è quindi provveduto ad applicare il BMA sulle previsioni ottenute per ciascun paziente per fornire una stima della curva di sopravvivenza globale riportata in Figura 25.

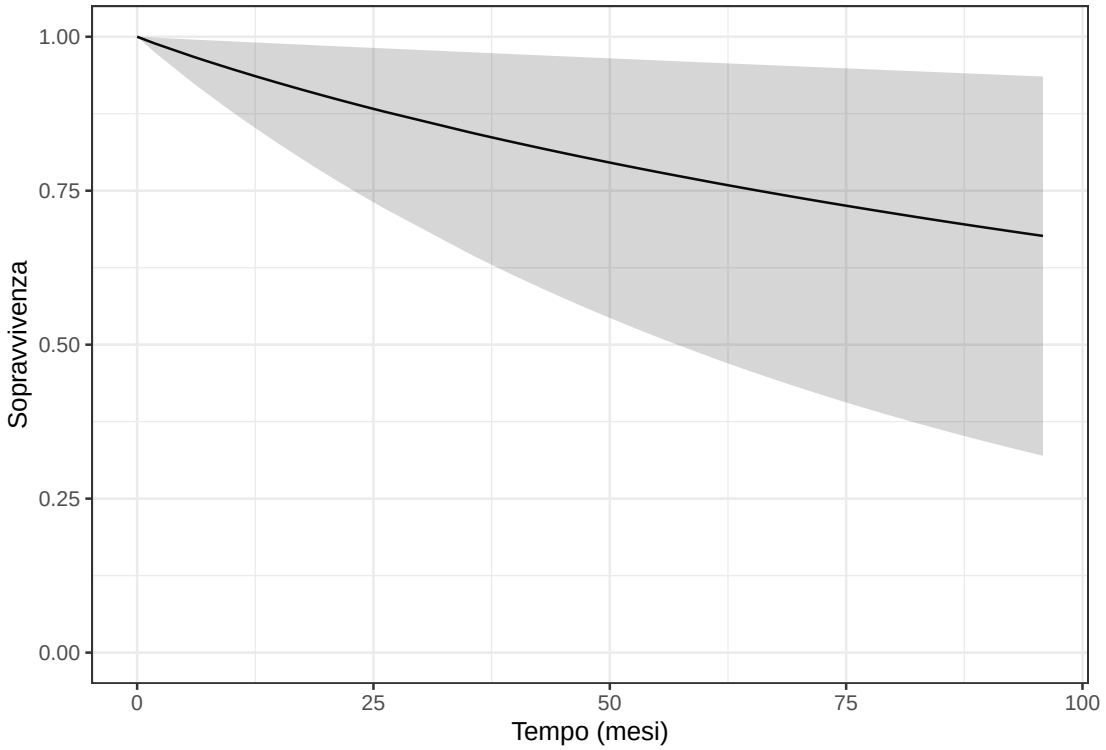


Figura 25: Curva di Sopravvivenza globale stimata dal BMA con credible set al 95%

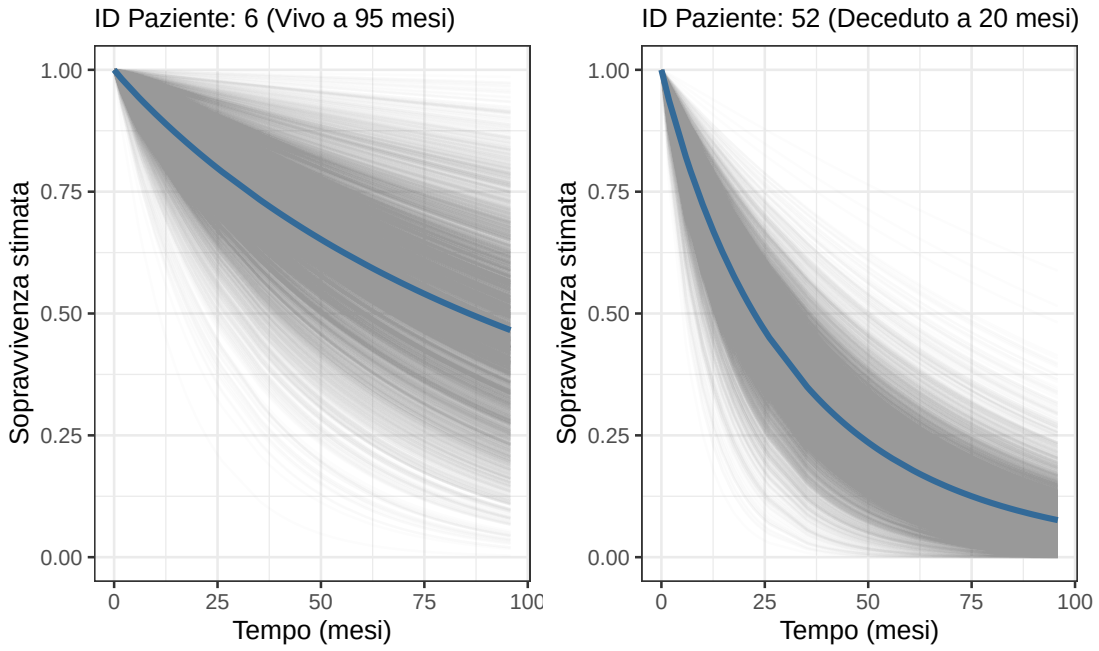


Figura 26: Curva di Sopravvivenza stimata (in blu) dal BMA per due pazienti con differenti caratteristiche. In grigio si riportano le 5000 previsioni ottenute tramite BMA e MCMC per ciascun istante temporale

Si osservi che, diversamente da problemi di regressione o classificazione, nell'ambito dell'analisi di sopravvivenza per ciascun $\boldsymbol{\vartheta}_{\lambda}^{s}$, e per ciascun individuo, è necessario produrre non già una singola previsione ma tante previsioni quanti sono gli istanti di tempo. Di fatto si stima una curva e non già un singolo valore. Ne discende che il BMA è stato applicato dopo aver prodotto la stima della sopravvivenza, per ciascun paziente, per ciascun $\boldsymbol{\vartheta}_{\lambda}^{s}$, in corrispondenza di ciascun istante temporale. In altri termini, per la sopravvivenza globale e il relativo intervallo di credibilità al 95%, il BMA è stato applicato condizionatamente a ciascun istante temporale. Infine, per mostrare come stimare le curve di sopravvivenza per singolo paziente tramite BMA, si riporta in Figura 26 la stima delle curve di sopravvivenza di due pazienti e le relative 5000 realizzazioni. La curva blu rappresenta la media delle 5000 previsioni, in ciascun istante temporale, ottenute tramite BMA ed è la stima che viene fornita per i due pazienti dal modello costruito.

6 Discussioni

Il lavoro di tesi sin qui condotto ha portato ad approfondire differenti aspetti della statistica che raramente trovano una cornice entro la quale si sviluppano interamente. In particolare, i dati oggetto d'analisi si caratterizzavano per l'elevata dimensionalità del dominio delle variabili ma, al contempo, come usuale negli studi clinici, erano caratterizzati da una bassa numerosità campionaria. Solitamente quando si affrontano problemi di dati ad elevata dimensionalità si ipotizza, implicitamente, che anche le osservazioni a disposizione risultino numerose, anche nei casi in cui $p >> n$. Se da un lato questo porta ad un onere computazionale più elevato ai fini delle analisi, dall'altro la dimensione campionaria è tale che anche se il numero di variabili supera di molto la dimensione campionaria, a partire dal campione è possibile comunque procedere seguendo un approccio ormai standardizzato e ampiamente utilizzato: dividere i dati in due sottoinsiemi. In questo caso il paradigma prevede un insieme di dati di *training*, sul quale selezionare le variabili,

eventualmente quantificando l'incertezza tramite *cross-validation*, e un insieme di dati di test, su cui si procede con l'inferenza. Ciò solitamente è possibile perché la dimensione campionaria è tale da sovrarappresentare la popolazione, pertanto il sottocampione utilizzato per la procedura di selezione di variabili contiene al suo interno tutte le informazioni utili a procedere con la selezione. Diversamente, quando il campione è caratterizzato da una bassa numerosità campionaria, la divisione in sottoinsiemi può portare, come nel caso oggetto d'analisi, a campioni non più rappresentativi della popolazione in esame e quindi a una selezione di variabili fortemente viziata e potenzialmente non ottima. In altri termini, quando si hanno a disposizione dimensioni campionarie ridotte e un numero di variabili molto elevato, le informazioni contenute nel campione non permettono di effettuare una selezione tramite il canonico paradigma di *data splitting* che continui a fornire garanzie inferenziali. Questo è un problema rilevante che, tuttavia, viene spesso ignorato, applicando una procedura LASSO che però non fornisce garanzie inferenziali circa la selezione. Non è insolito, inoltre, che il LASSO venga utilizzato sugli stessi dati per selezionare e per fare inferenza, introducendo così un *bias* dovuto alla selezione e uno dovuto all'assenza di garanzie inferenziali.

Nel corso della tesi si è cercato invece di sviluppare un metodo che, tenuto conto del tipo di dato a disposizione, della numerosità campionaria e del numero di variabili, fosse in grado di predire la sopravvivenza di pazienti affetti da sarcomi agli arti senza effettuare selezione di variabili e dunque senza introdurre il *bias* ad essa associato. Si è inoltre costruito un modello ad hoc, attraverso l'elicitazione delle distribuzioni a priori, che non penalizzasse il Sarculator in quanto indice prognostico già sottoposto a numerose validazioni in cui ha mostrato una buona capacità prognostica. Si è dunque tenuto conto dell'informazione a priori su quale

variabile era già noto fosse rilevante e si è applicata la penalizzazione solo alle variabili radiomiche.

Per poter specificare distribuzioni a priori adatte allo scopo e *proposal distributions* inizializzate secondo solide considerazioni teoriche si è dovuto optare per un metodo di campionamento MCMC (il MALA), che è stato anch'esso implementato ad hoc, in modo che l'implementazione fosse ottima per il problema oggetto d'analisi. Se da un lato ciò ha portato a complicazioni teoriche ed applicative, dall'altra ha consentito di calarsi all'interno di ogni singolo passaggio utile nel contesto bayesiano per giungere al risultato finale, senza fare ricorso all'utilizzo di algoritmi già implementati e poco flessibili per il caso oggetto d'analisi. Inoltre, la letteratura su metodi bayesiani applicati a dati di sopravvivenza ad elevata dimensionalità ma con bassa numerosità campionaria è molto limitata. Ciò ha portato ad analizzare nel dettaglio tutte le opzioni teoriche possibili, con uno sforzo non indifferente dal punto di vista della costruzione di un approccio solido e ben strutturato.

I risultati hanno mostrato che la radiomica aggiunge effettivamente potere prognostico al Sarculator, composto solo da variabili cliniche e che, dunque, andrebbe tenuta in considerazione per la costruzione di un indice prognostico composito che si prefigga l'obiettivo di massimizzare l'accuratezza prognostica.

Infine, essendo il campione in oggetto costruito appositamente per rappresentare esattamente la popolazione tipica di pazienti affetti da sarcomi agli arti, tenendo conto delle incidenze dei diversi istotipi tumorali, del *grading*, e della dimensione della lesione, non è stato possibile effettuare una suddivisione dei dati e validare il modello costruito. La validazione esterna rappresenta, dunque, un obiettivo futuro per sviluppare ulteriormente questo lavoro di ricerca. Tuttavia

si osservi che per la natura stessa dell'approccio bayesiano, che tiene conto dell'incertezza in fase di costruzione del modello tramite le distribuzioni a priori e a posteriori, il rischio che il modello incorra nella sovrastima è minimo diversamente dal rischio che si configurerebbe nell'ambito frequentista. Inoltre, a riguardo, applicando il BMA si tiene conto dell'incertezza anche dei modelli meno probabili, il che riduce ulteriormente il rischio che il modello sovrastimi. Infine, la modalità stessa con cui sono state stimate le probabilità a posteriori di ciascun modello, utilizzando il WAIC che quantifica più la generalizzabilità di un modello piuttosto che l'adattamento del modello ai dati usati per la stima, al posto del BIC, fornisce un ulteriore garanzia circa la stabilità dei risultati ottenuti.

Un possibile sviluppo futuro può riguardare la matrice dei dati utilizzati. In contesti come quello presentato sarebbe interessante riuscire ad elaborare delle strategie di *pre-screening* sulla matrice dei dati, senza utilizzare la variabile risposta, per riuscire a diminuire il numero di variabili coinvolte nel processo di modellizzazione. In questo contesto avrebbe senso anche porre una *hyperprior* sul parametro di precisione, o utilizzare distribuzioni a priori come quelle riportate nel paragrafo 4.2, per poi applicare il metodo della proiezione delle previsioni, come descritto da Piironen et al. (2020), al fine di selezionare le variabili rilevanti proiettandole su uno spazio ridotto rispetto allo spazio delle variabili del modello originario, minimizzando pertanto il *bias* di selezione.

Dal punto di vista della radiomica, effettuare un *pre-screening* della matrice dei dati è sicuramente un elemento di primaria importanza che sarà oggetto di ricerche future.

Concludendo, la tesi ha toccato e messo in comunicazione differenti ambiti della statistica attraverso un approccio bayesiano ad hoc: dati ad elevata dimensio-

nalità, dati con bassa numerosità campionaria, analisi di sopravvivenza e variabili di natura omica, ponendo l'accento sulla gestione dell'incertezza e proponendo un modello che fosse quanto più possibile accurato e che poggiasse su solide basi teoriche.

7 Appendice

7.1 Decomposizione Spettrale e Decomposizione ai Valori Singolari

Si enunceranno i due teoremi nell'ambito dell'insieme dei reali $\mathbb{R}$ dal momento che la matrice di varianza-covarianza della *proposal distribution* non presenta numeri complessi. Si osservi tuttavia che i due teoremi e la connessione tra essi è estendibile, con l'imposizione di ulteriori ipotesi, anche nel campo complesso, $\mathbb{C}$, si veda Brake et al. (2019).

Teorema 7.1 (Decomposizione Spettrale) *Sia* $\mathbf{A}$ *una matrice quadrata, simmetrica a valori reali di dimensione* $k \times k$*. Allora* $\mathbf{A}$ *è decomponibile nel prodotto di tre matrici:*

$$\mathbf{A} = \mathbf{VDV}' = \sum_{j=1}^{k} d_j \boldsymbol{v_j} \boldsymbol{v_j}'$$

Con d_j *si denota il* j*-simo autovalore associato ad* $\mathbf{A}$*,* $\mathbf{D}$ *è una matrice diagonale di dimensione* $k \times k$ *avente sulla diagonale principali i* k *autovalori di* $\mathbf{A}$*, ordinati in senso decrescente:*

$$\mathbf{D} = \begin{pmatrix} d_1 & 0 & \cdots & 0 \\ 0 & d_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & d_k \end{pmatrix},$$

con $\boldsymbol{v}_j$ *si denota il* j*-simo autovettore normalizzato (*$\|\boldsymbol{v}_j\| = 1$*) associato all'autovalore* d_j*;* $\mathbf{V} = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_j, \ldots \boldsymbol{v}_k)$*, inoltre* $\mathbf{V}$ *è una matrice ortogonale:* $\mathbf{V}\mathbf{V}' = \mathbf{V}'\mathbf{V} = \mathbf{I}_k$.

Il teorema 7.1 si applica a matrici quadrate e, in effetti, la matrice di varianza-covarianza lo è, tuttavia, da un punto di vista implementativo è più efficiente ricorrere alla generalizzazione del teorema e sfruttare la decomposizione ai valori singolari, che è applicabile a matrici rettangolari (di cui le matrici quadrate sono un caso particolare).

Teorema 7.2 (Decomposizione ai Valori Singolari SVD) *Sia* $\mathbf{A}$ *una matrice rettangolare, di dimensioni* $m \times k$*, a valori reali con rango* $rg(\mathbf{A})$ *tale che* $rg(\mathbf{A}) = r \leq \min(m, k)$ *allora esiste una matrice* $\mathbf{U}$ *di dimensioni* $m \times m$ *e una matrice* $\mathbf{V}$ *di dimensioni* $k \times k$ *tali che:* $\mathbf{U}\mathbf{U}' = \mathbf{I}_m$*,* $\mathbf{V}\mathbf{V}' = \mathbf{I}_k$ *e che:*

$$\mathbf{A} = \mathbf{U}\boldsymbol{\Delta}\mathbf{V}' = \sum_{j=1}^{\min(m,k)} \delta_j \boldsymbol{u_j} \boldsymbol{v_j}'$$

con $\mathbf{U} = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_j, \ldots \boldsymbol{u}_m)$ *e* $\mathbf{V} = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_j, \ldots \boldsymbol{v}_k)$*, dove* $\boldsymbol{u}_j, \boldsymbol{v}_j$ *denotano rispettivamente gli autovettori normalizzati delle due matrici, e* $\boldsymbol{\Delta}$ *matrice di dimensione* $m \times k$ *avente elemento di posizione* (j, j) *pari a* $\delta_j = \sqrt{d_j} \geq 0 \quad \forall j = 1, \ldots, \min(m, k)$*, e gli altri elementi pari a zero.* d_j *rappresenta il* j*-simo autovalore della matrice* $\mathbf{A}'\mathbf{A}$ *e* δ_j *è detto valore singolare.*

Dimostrazione

Se $rg(\mathbf{A}) = r \leq \min(m, k)$ allora la matrice $\mathbf{A}'\mathbf{A}$ è simmetrica e semidefinita positiva. Ne discende che la matrice $\mathbf{A}$ è diagonalizzabile e, per il teorema di

decomposizione spettrale, è lecita la riscrittura:

$$\mathbf{A}'\mathbf{A} = \mathbf{V}\mathbf{D}\mathbf{V}' = \sum_{j=1}^{k} d_j \boldsymbol{v_j}\boldsymbol{v_j}' = \sum_{j=1}^{k} \delta_j^2 \boldsymbol{v_j}\boldsymbol{v_j}'.$$

Segue che $rg(\mathbf{A}) = rg(\mathbf{A}'\mathbf{A})$ poiché $\mathbf{D}$ è diagonale. Detto δ_j il j-simo valore singolare $\delta_j = \sqrt{d_j} \geq 0$ allora per la j-sima coppia autovalore-autovettore vale l'identità: $\mathbf{A}'\mathbf{A}\boldsymbol{v_j} = \delta_j^2 \boldsymbol{v_j}$. Ma allora, poiché la matrice è diagonalizzabile deve essere anche a rango pieno. Essendo la matrice a rango pieno: $\delta_j > 0, \quad \forall j$, e quindi è possibile definire per ogni j un vettore m-dimensionale, $\boldsymbol{u_j}$, come:

$$u_j = \frac{\mathbf{A}\boldsymbol{v_j}}{\delta_j},$$

e tale che, per costruzione, sia l'autovettore a norma unitaria della matrice $\mathbf{A}\mathbf{A}'$. Indicando con $\mathbf{V}$ la matrice avente come colonne gli autovettori normalizzati $\boldsymbol{v_j}$ e $\mathbf{U}$ la matrice avente come colonne gli autovettori normalizzati $\boldsymbol{u_j}$, definita la matrice diagonale $\boldsymbol{\Delta}$ di dimensioni $(k \times m)$ con elemento di posto (j, j) il valore singolare δ_j, si conclude che:

$$\mathbf{U} = \mathbf{A}\mathbf{V}\boldsymbol{\Delta}^{-1} \Leftrightarrow \mathbf{U}\boldsymbol{\Delta} = \mathbf{A}\mathbf{V} \Leftrightarrow \mathbf{A} = \mathbf{U}\boldsymbol{\Delta}\mathbf{V}' \qquad \square$$

Si osservi che se $\mathbf{A}$ è quadrata, simmetrica e semidefinita positiva allora $\mathbf{U} = \mathbf{V}$ e $\boldsymbol{\Delta} = \mathbf{D}$ con $\mathbf{D}$ matrice diagonale e avente sulla diagonale principale gli autovalori di $\mathbf{A}$, in questo caso la decomposizione spettrale risulta un caso particola della SVD.

7.2 Codice utilizzato per le analisi

Si riporta il codice utilizzato per le analisi.

```
##-----------------------------
## Matrice dei dati e librerie
##-----------------------------
library(numDeriv)
library(cvwrapr)
library(ggplot2)
library(dplyr)
library(bayesplot)
library(probably)
library(pROC)
library(patchwork)
library(tidymodels)
#load("path del dataset")
db_variable<-recipe( ~., data=db_variable)|>
  step_select(- status_bin, -time_OS, skip = T)|>
  step_normalize(all_numeric_predictors()) |>
  prep() |>
  bake(db_variable)
X <- db_variable|>dplyr::select(-status_bin,- time_OS)
X <- cbind("intercept"=rep(1, nrow(X)),X )
X <- as.matrix(X)
tempo <- db_variable$time_OS
stato <- db_variable$status_bin

##--------------------------
##     Likelihood
##--------------------------

lik <- function(X, theta, t, status){
  eta <- X %*% theta[-1]
  exp( sum(status*theta[1] + status * eta - exp(eta) * t^(exp(theta[1])) ) ) *
    prod(t^( (exp(theta[1])-1) * status ))
}

##--------------------------
##     Log Posterior:
##--------------------------
# theta[1]=eps, theta[2...]=beta
loglik <- function(theta, X, t, status) {
  eta<-X %*% theta[-1]
  eps<-exp(theta[1])

  theta[1]*sum(status) +
  sum(eta[status==1])  +
  (eps-1) * sum(log(t[status==1])) -
  sum( (exp(eta)) * (t^eps) )
}

logprior <- function(theta, sigma_alpha, sigma_beta, l) {
  # shape
  dnorm(theta[1], mean=-log(sigma_alpha+1)/2, sd=sqrt(log(sigma_alpha+1)), log=T ) +
  # Ridge beta
  sum(dnorm(theta[-c(1,2,3)], mean=0, sd=sqrt(1/(2*l)) , log=T )) +
  # Beta non penalizzati
```

```
  sum(dnorm(theta[c(2,3)], mean=0, sd=sqrt(sigma_beta), log=T) )
}

logpost <- function(theta, X, t, status, lambda) {
  loglik(theta=theta, X=X, t=t, status=status) +
  logprior(theta=theta, sigma_alpha=16, sigma_beta=600, l=lambda)
}

##--------------------------
# Gradiente
##--------------------------
lgradient <- function(theta, t, x, s_alpha, s_beta, lambda, status) {

  eta <- exp(x %*% theta[-1])
  eps <- exp(theta[1])
  teps <- t^eps
  grad <- matrix(rep(0,2147), ncol=1, nrow=2147)

  # Derivata rispetto a xi
  grad[1,1]<- sum(status) + eps * sum(log(t[status==1]))  -
            eps * sum( eta * teps * log(t) ) - theta[1]/(log(s_alpha+1)) -1/2
  # Derivate rispetto a beta
  for(j in 1:ncol(x)){
    grad[j+1,1]<-sum(x[status==1,j]) - sum(x[,j] * eta * teps) -
      ifelse(j>2, 2 * theta[j+1] * lambda, theta[j+1]/s_beta )
  }

  return(grad)
}


##-----------------------------------------------------
##   Inizializzazione delle medie delle  Proposals:
##-----------------------------------------------------

# Vettori iniziali delle proposal
#-----------------------------------------
# Definiamo qui i valori di lambda e massimizziamo rispetto alla griglia
# di valori possibili
a =10.5; b = 5
lam <- exp(seq(from = a, to = b, length = 100))

# Massimo log posterior (per proposal):
#--------------------------------------
logpostmax <- function(theta, X, t, status, lambdaseq){
  - logpost(theta=theta, X = X, t = t, status = status, lambda=lambdaseq)
}

beta_mat <- matrix(0,ncol=ncol(X)+1, nrow=length(lam))
logpostlam <- vector(); conv <- vector()

set.seed(123)
for(i in 1:(length(lam))){
  r<- rep(0,ncol(X)+1)
  stores <- nlminb(r,
              objective=function(theta)
                        logpostmax(theta, X = X, t= tempo, status = stato, lambdaseq = lam[i]),
              gradient = function(theta)
                        - lgradient(theta, t=tempo, x=X, s_alpha=16, s_beta=600,
                                    lambda=lam[i], status=stato),
              control=list(iter.max=4000, eval.max=4000)
              )
```

```
  conv[i] <- stores$convergence
  beta_mat[i,] <- stores$par
}
## Monitoraggio valore di lambda tale per cui i x
## coefficienti si azzerano, e raffinamento della griglia
matplot(log(lam), beta_mat[,-c(1,2,3)], type="l", lty=1,
        xlab=expression(log(symbol(l))),
        ylab=expression(beta), lwd=2)

# Definiamo analiticamente la - matrice Hessiana
# Effettuiamo la scomposizione in valori singolari
# Ricaviamo le matrici A per ciascuna proposal
#-----------------------------------------------------
A_matrix <- function(theta, x, status, t,
                     s_alpha, s_beta, lambda){

  subH <- matrix(NA, nrow = 2146, ncol = 2146)
  eta <- x %*% theta[-1]; eps <- exp(theta[1])
  lt <- log(t); teps <- t^(eps)
  exeta <- exp(eta)

  ## -d^2log()/dxi^2_j
  xixi <- sum( exeta * lt * ( eps * teps + (eps^2) * teps * lt ) ) -
                eps * sum( lt[status==1]) + 1/(log(s_alpha+1))
  ## -d^2log()/dxidbeta_j = -d^2log()/dbeta_jdxi
  dbjxi<-vector()
  for(j in 1:ncol(x)){
    dbjxi[j]<-sum( x[,j] * exp( eta + theta[1] ) * teps * lt )
  }

  # Due cicli per rendere più efficiente il codice:
  # Primo ciclo: inseriamo gli elementi sulla diagonale principale.
  # Secondo ciclo: costriamo la matrice triangolare inferiore.
  # Per simmetria ricostruiamo poi quella superiore.
  # Dimezziamo il tempo.

  # d^2log()/dbeta_j^2
  for (j in 1:ncol(x)){
    diag(subH)[j] <-  sum( (x[,j]^2) * exeta * teps) +
                            ifelse(j>2, 2*lambda, 1/s_beta )
  }
  # dlog()/dbeta_j dbeta_k = # dlog()/dbeta_k dbeta_j
  for(j in 2:ncol(x)){
    for(k in 1:(j-1)){
        subH[j,k] <- sum( x[,j] * x[,k] * exeta * teps )
    }
  }
  # Simmetria
  subH[upper.tri(subH)] <- t(subH)[upper.tri(subH)]
  # Aggreghiamo e giungiamo alla -Hessiana
  H <- unname(cbind(dbjxi, subH))
  H <- unname(rbind(matrix(c(xixi, dbjxi), nrow=1), H))
  # Decomposizione ai valori singolari
  #-------------------------------------
  SVD <- svd(H)
  D1<-diag(1/sqrt(SVD$d))
  V<-SVD$u
  A <- V %*% D1
  return(A)
}
```

```
#-------------------------
# MALA con tuning epsilon
#-------------------------
MALA <- function(R, burn_in, lamd, thinning=10, X=X, t=tempo,
                 status=stato, sigma_alpha=16, sigma_beta=600,
                 target = 0.574, bat=50, tolerance=0.054, stop=15,
                 y=y) {
  require(coda)
  p <- ncol(X)+1
  out <- array(NA, c(R/thinning, p,length(lamd)))
  sequenza <- seq((burn_in+1),(R+burn_in), thinning)
  eps_opt<- rep(0,100); tasso_accept<-vector();
  adaptive_monitoring <- matrix(NA, ncol = length(lamd), nrow=burn_in)
  lower_tol <- target - tolerance ; upper_tol <- target + tolerance
  r_stop<-vector()

  r5anni<- matrix(NA,nrow=5000, ncol=100)
  brier5anni <- matrix(NA,nrow=5000, ncol=100)
  rover<- matrix(NA,nrow=5000, ncol=100)
  brierover <- matrix(NA,nrow=5000, ncol=100)
  Cindex <- matrix(NA, nrow=5000, ncol=100)


  for(i in 1:length(lamd) ){
      theta <- beta_mat[i,]
      A <- A_matrix(theta=theta, x=X, status=status, t=t,
                    s_alpha=sigma_alpha, s_beta=sigma_beta,
                    lambda=lamd[i])
      S <- A %*% t(A) ; S1 <- solve(S)
      # Log-posterior
      logp <- logpost(theta=theta, lambda=lamd[i], X=X, t=t, status = status)
      # Calcolo del  gradiente
      lgrad <- c(S %*% lgradient(theta=theta, lambda=lamd[i], x=X,
                                 t=t, status = status, s_alpha=sigma_alpha,
                                 s_beta = sigma_beta))

       # Inizializzazione per il tuning per ogni nuovo lambda
       accepted <- 0
       accettati_R <- 0  # Per tasso accettazione complessivo per ogni lambda
       epsilon <- 1    # Inizializzazione di eps
       batch <- 1             # Inizializzazione della batch size
       index <- 0
       j <- 0
        # MALA Adattivo
        #------------------
       r=1
        while (r <= (burn_in + R)) {

          # Tuning epsilon
          #------------------
          if (r < burn_in+1 & batch == bat) {
            j <- j+1
              if ((accepted / bat) > upper_tol )  {
                epsilon <- epsilon + ifelse(r>5000, sqrt(1/r), 0.01)
               }else if( (accepted / bat) < lower_tol ) {
                epsilon <- epsilon - ifelse(r>5000, sqrt(1/r), 0.01)
               }

            adaptive_monitoring[j , i] <- epsilon
            batch <- 0; accepted <- 0; print(epsilon)
            if(r > 40000 ){
              if(all(epsilon == adaptive_monitoring[(j-stop+1):j , i])==T){
```

```
            r_stop[i] <- r
            r = burn_in
          }
        }
      }
      #aggiorniamo la bach size
      batch <- batch + 1
      # Proseguiamo con il MALA
      sigma2 <- epsilon^2 / p^(1 / 3) ; sigma <- sqrt(sigma2)

      theta_new <- theta + sigma2 / 2 * lgrad + sigma * c(crossprod(A, rnorm(p)))
      logpnew <- logpost(theta=theta_new, lambda=lamd[i], X=X, t=t, status = status)
      lgrad_new <- c(S %*% lgradient(theta=theta_new, lambda=lamd[i], x=X,
                              t=t, status = status, s_alpha=sigma_alpha,
                              s_beta = sigma_beta))

      diffold <- theta - theta_new - sigma2 / 2 * lgrad_new
      diffnew <- theta_new - theta - sigma2 / 2 * lgrad

      qold <- -diffold %*% S1 %*% diffold / (2 * sigma2)
      qnew <- -diffnew %*% S1 %*% diffnew / (2 * sigma2)

      alpha <- min(1, exp(logpnew - logp + qold - qnew))
      if (runif(1) < alpha) {
        logp <- logpnew
        lgrad <- lgrad_new
        theta <- theta_new # Accetta  il valore
        # aggiorniamo quanti sono gli accettati nel burn in
        accepted <- accepted + 1
        # aggiorniamo quanti sono gli accettati dopo il burn in
        if( r %in% sequenza){
          accettati_R <- accettati_R + 1
        }
      }
      # 5000 valori con thinning pari a 15
      if ( r %in% sequenza ) {
        index <- index + 1
        out[index, ,i] <- theta

        r5anni[index,i]<- roc(status,
                              as.vector(1-exp(-(60^(exp(theta[1]))*exp(theta[2]))/
                              (exp(X[,-1]%*%theta[-c(1,2)])) )), plot=F)$auc
        brier5anni[index,i]<- mean((status-
                              as.vector(1-exp(-(60^(exp(theta[1]))*exp(theta[2]))/
                              (exp(X[,-1]%*%theta[-c(1,2)])) ))^2)
        rover[index,i]<- roc(status,
                              as.vector(1-exp(-(max(tempo)^(exp(theta[1]))*exp(theta[2]))/
                              (exp(X[,-1]%*%theta[-c(1,2)])) )), plot=F)$auc
        brierover[index,i]<- mean((status-
                              as.vector(1-exp(-(max(tempo)^(exp(theta[1]))*exp(theta[2]))/
                              (exp(X[,-1]%*%theta[-c(1,2)])) ))^2)
        Cindex[index,i]<- getCindex(
          as.vector(exp(-(max(tempo)^(exp(theta[1]))*exp(theta[2]))/
                              (exp(X[,-1]%*%theta[-c(1,2)])) )),y)
      }
      r=r+1
    }
    eps_opt[i]<-epsilon
    tasso_accept[i] <- thinning*accettati_R/R

}
return(list(out, eps_opt,tasso_accept, adaptive_monitoring,
```

```
              r_stop, r5anni,brier5anni, rover, brierover, Cindex ))
}
```

```
#-----------------------------
# Modello Base
#-----------------------------
logprior0 <- function(theta, sigma_alpha, sigma_beta) {
  # shape
  dnorm(theta[1], mean=-log(sigma_alpha+1)/2, sd=sqrt(log(sigma_alpha+1)), log=T ) +
  # Beta non penalizzati
  sum(dnorm(theta[c(2,3)], mean=0, sd=sqrt(sigma_beta), log=T) )
}

logpost0 <- function(theta, X, t, status, lambda) {
  loglik(theta=theta, X=X, t=t, status=status) +
  logprior0(theta=theta, sigma_alpha=16, sigma_beta=16)
}

lgradient0 <- function(theta, t, x, s_alpha, s_beta, status) {
  eta <- exp(x %*% theta[-1])
  eps <- exp(theta[1])
  teps <- t^eps
  grad <- vector(length = 3)
  # Derivata rispetto a xi
  grad[1]<- sum(status) + eps * sum(log(t[status==1]))  -
            eps * sum( eta * teps * log(t) ) - theta[1]/(log(s_alpha+1)) -1/2
  # Derivate rispetto a beta
    grad[2]<-sum(x[status==1,1]) - sum(x[,1] * eta * teps) - theta[2]/s_beta
    grad[3]<-sum(x[status==1,2]) - sum(x[,2] * eta * teps) - theta[3]/s_beta
  return(grad)
}

logpostmax0 <- function(theta, X, t, status){
  - logpost0(theta=theta, X = X, t = t, status = status)
}

beta_mat0 <- vector(length = 3)
set.seed(123)
r0<- rep(0,3)
stores0 <- nlminb(r,
            objective=function(theta)
                        logpostmax0(theta, X = X[,c(1,2)], t= tempo, status = stato),
            gradient = function(theta)
                        - lgradient0(theta, t=tempo, x=X[,c(1,2)], s_alpha=16, s_beta=16, status=stato),
            control=list(iter.max=4000, eval.max=4000)
          )
beta_mat0 <- stores0$par

A_matrix0 <- function(theta, x, status, t,
                      s_alpha, s_beta){

  H <- matrix(NA, nrow = 3, ncol = 3)
  eta <- x %*% theta[-1]
  eps <- exp(theta[1])
```

```
  lt <- log(t)
  teps <- t^(eps)
  exeta <- exp(eta)
  ## -d^2log()/dxi^2_j
  xixi <- sum( exeta * lt * ( eps * teps + (eps^2) * teps * lt ) ) -
                  eps * sum( lt[status==1]) + 1/(log(s_alpha+1))

  H[1,1] <- xixi
  H[2,1] <- sum( x[,1] * exp( eta + theta[1] ) * teps * lt )
  H[3,1] <- sum( x[,2] * exp( eta + theta[1] ) * teps * lt )
  H[1,2] <- sum( x[,1] * exp( eta + theta[1] ) * teps * lt )
  H[2,2] <- sum( (x[,1]^2) * exeta * teps) +  1/s_beta
  H[3,2] <- sum( x[,1] * x[,2] * exeta * teps )
  H[1,3] <- sum( x[,2] * exp( eta + theta[1] ) * teps * lt )
  H[2,3] <- sum( x[,2] * x[,1] * exeta * teps )
  H[3,3] <- sum( (x[,2]^2) * exeta * teps) +  1/s_beta

  # Decomposizione ai valori singolari
  #------------------------------------
  SVD <- svd(H)
  D1<-diag(1/sqrt(SVD$d))
  V<-SVD$u
  A <- V %*% D1
  # Verifica che tutto sia in ordine e no ci siano autovalori
  # negativi:
  return(A)
}

#---------------------------------------------------------
# MALA MODELLO BASE, UGUALE A QUELLO DEI 100 MODELLI.
# UNICA DIFFERENZA: non vi è penalizzazione
#---------------------------------------------------------

MALA_M0 <- function(R, burn_in, thinning=15, X=X, t=tempo,
                 status=stato, sigma_alpha=16, sigma_beta=600,
                 target = 0.574, bat=50, y=y) {
  require(coda)
  p <- ncol(X)+1; j=0 ;
  out <- matrix(NA, nrow=R/thinning, ncol=p)
  sequenza <- seq((burn_in+1),(R+burn_in), thinning)
  adaptive_monitoring <- vector()
  eps_opt<- 0; tasso_accept<-0;
  r5anni<- matrix(NA,nrow=5000, ncol=1)
  brier5anni <- matrix(NA,nrow=5000, ncol=1)
  rover<- matrix(NA,nrow=5000, ncol=1)
  brierover <- matrix(NA,nrow=5000, ncol=1)
  Cindex <- matrix(NA, nrow=5000, ncol=1)

  theta <- beta_mat0
  A <- A_matrix0(theta=theta, x=X[,c(1,2)], status=status, t=t,
                s_alpha=sigma_alpha, s_beta=sigma_beta)
  S <- A %*% t(A)
  S1 <- solve(S)
  # Log-posterior
  logp <- logpost0(theta=theta, X=X[,c(1,2)], t=t, status = status)
  # Calcolo del  gradiente
  lgrad <- c(S %*% lgradient0(theta=theta,  x=X[,c(1,2)],
                                t=t, status = status, s_alpha=sigma_alpha,
                                s_beta = sigma_beta))
  # Inizializzazione
  accepted <- 0 ; accettati_R <- 0  ; epsilon <- 1
  batch <- 1; index <- 0
```

```
# MALA Adattivo
#------------------
r=1
 while (r <= (burn_in + R)) {
   # Tuning epsilon
   #------------------
   if (r < burn_in+1 & batch == bat) {
       j <- j+1
       if ((accepted / bat) > target+tolerance )  {
         epsilon <- epsilon + ifelse(r>5000, sqrt(1/r), 0.01)
         } else if( (accepted / bat) < target-tolerance ) {
         epsilon <- epsilon - ifelse(r>5000, sqrt(1/r), 0.01)
         }
       adaptive_monitoring[j] <- epsilon
       batch <- 0; accepted <- 0;
       if(r > 40000 ){
         if(all(epsilon == adaptive_monitoring[(j-stop+1):j])==T){
           r = burn_in
         }
       }
     }
     batch <- batch + 1
     # Proseguiamo con il MALA
     sigma2 <- epsilon^2 / p^(1 / 3)
     sigma <- sqrt(sigma2)
     theta_new <- theta + sigma2 / 2 * lgrad + sigma * c(crossprod(A, rnorm(p)))
     logpnew <- logpost0(theta=theta_new, X=X[,c(1,2)], t=t, status = status)
     lgrad_new <- c(S %*% lgradient0(theta=theta_new, x=X[,c(1,2)],
                              t=t, status = status, s_alpha=sigma_alpha,
                              s_beta = sigma_beta))
     diffold <- theta - theta_new - sigma2 / 2 * lgrad_new
     diffnew <- theta_new - theta - sigma2 / 2 * lgrad
     qold <- -diffold %*% S1 %*% diffold / (2 * sigma2)
     qnew <- -diffnew %*% S1 %*% diffnew / (2 * sigma2)
     alpha <- min(1, exp(logpnew - logp + qold - qnew))
     if (is.infinite(alpha)==F & runif(1) < alpha & is.na(alpha)==F) {
       logp <- logpnew ; lgrad <- lgrad_new
       theta <- theta_new ; accepted <- accepted + 1
       # aggiorniamo quanti sono gli accettati dopo il burn in
       if( r %in% sequenza){
         accettati_R <- accettati_R + 1
       }
     }
     if ( r %in% sequenza ) {
       out[index ,] <- theta

       index <- index + 1
       pred.tmp60 <- as.vector(1-exp(-(60^(exp(theta[1]))*
                              (exp(X%*%theta[-1])) )))
       pred.tmpmax <- as.vector(1-exp(-(max(tempo)^(exp(theta[1]))*
                              (exp(X%*%theta[-1])) )))
       r5anni[index,1]<- roc(status, pred.tmp60, plot=F)$auc
       brier5anni[index,1]<- mean((status- pred.tmp60)^2)
       rover[index,1]<- roc(status, pred.tmpmax, plot=F)$auc
       brierover[index,1]<- mean((status - pred.tmpmax)^2)
       Cindex[index,1]<- getCindex(pred.tmpmax,y)
     }
     r=r+1
 }
    eps_opt<-epsilon ; tasso_accept <- thinning*accettati_R/R
return(list(out, eps_opt,tasso_accept,
            r5anni,brier5anni, rover, brierover, Cindex ))
```

```
}
```

```
# Modello Base
results0 <- MALA_M0(R = 75000, burn_in = 150000, thinning = 15,
                X=X[,c(1,2)], status=stato,
                t=tempo, sigma_alpha=16, sigma_beta=600,
                bat=50, target=0.574, tolerance = 0.054, stop=10, y=y)
# 100 Modelli
results <- MALA(R = 75000, burn_in = 150000, thinning = 15,
                lamd=lam, X=X, status=stato,
                t=tempo, sigma_alpha=16, sigma_beta=600,
                bat=50, target=0.574, tolerance = 0.054, stop=10, y=y)
```

```
colnames(results) <- c("log(shape)","Intercetta" ,"LP_OS", colnames(X[,-c(1,2)]))
colnames(results) <- c("log(shape)","Intercetta" ,"LP_OS", colnames(X[,-c(1,2)]))
colnames(cindex) <- paste(1:100, sep = "_")
colnames(auc5) <- paste(1:100, sep = "_")
colnames(brier5) <- paste(1:100, sep = "_")
colnames(aucover) <- paste(1:100, sep = "_")
colnames(brierover) <- paste(1:100, sep = "_")

DIC_fun <- function(M, X=X, t=tempo, status=stato){
  lp <- vector()
  elp <- vector()
  p_DIC <- vector()
  for(i in 1:dim(M)[3]){
    M_i <- as.matrix(M[,,i])
    theta_i <- as.vector(apply(M_i, 2, mean))
    lp[i] <- loglik(theta=theta_i, X=X, t=t, status=status)
    elp[i] <- mean(apply(M_i, 1, function(x) loglik(X=X, theta=x, t=t, status=status) ) )
    p_DIC[i] <- 2*(lp[i] - elp[i])
  }

  DIC<- -2*lp + 2*p_DIC
  names(DIC)<- paste("lambda", 1:100, sep = "_")
  names(p_DIC)<- paste("lambda", 1:100, sep = "_")
  return(list(DIC, p_DIC))
}

WAIC_fun <- function(M, X=X, t=tempo, status=stato){
  lppd <- vector()
  p_WAIC <- vector()
  for(i in 1:dim(M)[3]){
    M_i<-M[,,i]
    lppd_tmp <-vector()
     mean_ind <-vector()
    for(j in 1:nrow(X)){
```

```
      lppd_tmp[j] <- mean(apply(M_i, 1, function(x)
        lik(X=as.vector(X[j,]), theta=x, t=t[j], status=status[j])) )
      mean_ind[j]<- mean(apply(M_i, 1, function(x)
        loglik(X=as.vector(X[j,]), theta=x, t=t[j], status=status[j]) )) # var singoli s
    }
    lppd[i] <- sum(log(lppd_tmp))
    p_WAIC[i] <- 2 * sum(log(lppd_tmp) - mean_ind)
    print(i)
  }
  WAIC <- -2*lppd + 2*p_WAIC
  names(WAIC)<- paste("lambda", 1:100, sep = "_")
  names(p_WAIC)<- paste("lambda", 1:100, sep = "_")
  return(list(WAIC, p_WAIC))
}

# WAIC funzione per modello base
WAIC_fun0 <- function(M, X=X, t=tempo, status=stato){
  lppd_tmp <-NA ; mean_ind <-NA
  S <- dim(M)[1]
  for(j in 1:nrow(X)){
    lppd_tmp[j] <- mean(apply(M_i, 1, function(x)
      lik(X=as.vector(X[j,]), theta=x, t=t[j], status=status[j])) )
    mean_ind[j]<- mean(apply(M_i, 1, function(x)
      loglik(X=as.vector(X[j,]), theta=x, t=t[j], status=status[j]) )) # var singoli s
  }
  lppd<- sum(log(lppd_tmp))
  p_WAIC <- 2 * sum(log(lppd_tmp) - mean_ind)
  WAIC <- -2*lppd + 2*p_WAIC
  return(list(WAIC, p_WAIC))
}

mcmc_trace(resmh,pars=c("LP_OS","v54", "v566", "v1491", "v1850", "log(shape)"))+
scale_x_continuous(n.breaks = 3, limits = c(0,5000))+
ggtitle(expression(paste("log ", lambda , " =", " 10.39")))+
theme(plot.title = element_text(size = 10.5))


tr1 <- mcmc_trace(results[,,3],
                  pars=c("LP_OS","v54", "v566", "v1491", "v1850", "log(shape)"))+
  scale_x_continuous(n.breaks = 3, limits = c(0,5000))+
  ggtitle(expression(paste("log ", lambda , " =", " 10.39")))+
  theme(plot.title = element_text(size = 10.5))
tr2 <- mcmc_trace(results[,,20],
                  pars=c("LP_OS","v54", "v566", "v1491", "v1850", "log(shape)"))+
  scale_x_continuous(n.breaks = 3, limits = c(0,5000))+
  ggtitle(expression(paste("log ", lambda , " =", " 9.44")))+
  theme(plot.title = element_text(size = 10.5))
tr3 <- mcmc_trace(results[,,47],
                  pars=c("LP_OS","v54", "v566", "v1491", "v1850", "log(shape)"))+
  scale_x_continuous(n.breaks = 3, limits = c(0,5000))+
  ggtitle(expression(paste("log ", lambda , " =", " 7.94")))+
  theme(plot.title = element_text(size = 10.5))
tr4 <- mcmc_trace(results[,,90],
                  pars=c("LP_OS","v54", "v566", "v1491", "v1850", "log(shape)"))+
  scale_x_continuous(n.breaks = 3, limits = c(0,5000))+
  ggtitle(expression(paste("log ", lambda , " =", " 5.56")))+
  theme(plot.title = element_text(size = 10.5))
(tr1 + tr2) /(tr3 + tr4)

stop_burn<- ifelse(is.na(stop_burn)==T, 150000, stop_burn)
colnames(admon)<- c(paste("lambda",1:100, sep = "_"))
```

```
# diminishing adaptation
x3<- c(1,seq(50,stop_burn[3],50))
x20 <- c(1, seq(50,stop_burn[20],50))
x47 <- c(1,seq(50,stop_burn[47],50))
x90 <- c(1,seq(50,stop_burn[90],50))

adm1<-
  rbind.data.frame(rep(1,100), admon)|>
  filter(is.na(lambda_3)==F)|>
ggplot(aes(x=x3, y=lambda_3))+
  geom_line(colour="gray60")+
  geom_hline(yintercept = median(admon[,3], na.rm = T), colour="red3")+
  theme_classic()+
  scale_x_continuous(n.breaks = 4)+
  ylab("Parametro di Step size")+
  xlab("Iterazioni Burn-in")+
  geom_text(aes(x=25000, y=0.95),
            label=expression(paste("log ", lambda , " =", " 10.39")), size=2.5)+
  geom_text(aes(x=25000, y=0.91),
            label=paste("Stabilizzazione all'iterazione", stop_burn[3] ),size=2.5)

adm2 <-
  rbind.data.frame(rep(1,100), admon)|>
  filter(is.na(lambda_20)==F)|>
  ggplot(aes(x=x20, y=lambda_20))+
    geom_line(colour="gray60")+
    geom_hline(yintercept = median(admon[,20], na.rm = T), colour="red3")+
    theme_classic()+
    scale_x_continuous(n.breaks = 4)+
    ylab("Parametro di Step size")+
    xlab("Iterazioni Burn-in")+
    geom_text(aes(x=25000, y=0.95),
              label=expression(paste("log ", lambda , " =", " 9.44")), size=2.5)+
    geom_text(aes(x=25000, y=0.91),
              label=paste("Stabilizzazione all'iterazione", stop_burn[20] ), size=2.5)


adm3 <-
  rbind.data.frame(rep(1,100), admon)|>
    filter(is.na(lambda_47)==F)|>
  ggplot(aes(x=x47, y=lambda_47))+
    geom_line(colour="gray60")+
    geom_hline(yintercept = median(admon[,47], na.rm = T), colour="red3")+
    theme_classic()+
    scale_x_continuous(n.breaks = 4)+
    ylab("Parametro di Step size")+
    xlab("Iterazioni Burn-in")+
    geom_text(aes(x=25000, y=0.99),
              label=expression(paste("log ", lambda , " =", " 7.94")), size=2.5)+
    geom_text(aes(x=25000, y=0.97),
              label=paste("Stabilizzazione all'iterazione", stop_burn[47] ), size=2.5)


adm4 <-
  rbind.data.frame(rep(1,100), admon)|>
    filter(is.na(lambda_90)==F)|>
  ggplot(aes(x=x90, y=lambda_90))+
    geom_line(colour="gray60")+
    geom_hline(yintercept = median(admon[,90], na.rm = T), colour="red3")+
    theme_classic()+
    scale_x_continuous(n.breaks = 5)+
    ylab("Parametro di Step size")+
```

```
    xlab("Iterazioni Burn-in")+
    geom_text(aes(x=75000, y=1.04),
              label=expression(paste("log ", lambda , " =", " 5.56")), size=2.5)+
    geom_text(aes(x=75000, y=1.02),
              label=paste("Stabilizzazione all'iterazione", stop_burn[90] ), size=2.5)

(adm1 + adm2) / (adm3 + adm4)

# Tasso di accettazione per ogni lambda
summary(t_ac_compl)

# Effective Sample Size
effss<-data.frame("Minimo"=NA, "Primo quartile"=NA, "Mediana"=NA,
                  "Media"=NA, "Terzo quartile"=NA, "Massimo"=NA)
for(i in 1:length(lam)){
  effss[i,] <- as.vector(summary(effectiveSize(as.mcmc(results[,,i]))))
}
effss["l"] <- log(lam)
effss_plot <-
  effss|>
  ggplot(aes(x=l, y=Mediana))+
  geom_point(colour="gray60")+
  geom_line(aes(y=Mediana, colour="Mediana"))+
  geom_line(aes(y=Primo.quartile, colour="Primo.quartile"), alpha=0.8)+
  geom_line(aes(y=Terzo.quartile, colour="Terzo.quartile"), alpha=0.8)+
  scale_y_continuous(limits = c(0,1200), n.breaks = 5)+
  scale_colour_manual("",
                      breaks = c("Mediana", "Primo.quartile", "Terzo.quartile"),
                      labels = c("Mediana", "Primo quartile", "Terzo quartile"),
                      values = c("black", "blue3", "green3")) +
  theme_classic()+
  xlab(expression(paste("log ", lambda)))+
  ylab("Effective Sample Size")+
  theme(legend.position = "top")
effss_plot


DIC_all <- DIC_fun(M=results, X=X, t=tempo, status=stato)
DIC_db <-data.frame("DIC"=DIC_all[[1]], "pDIC"=DIC_all[[2]])
WAIC_all <- WAIC_fun(M=results, X=X, t=tempo, status=stato)
WAIC_db <-data.frame("WAIC"=WAIC_all[[1]], "pWAIC"=WAIC_all[[2]])

# Esempio performance 4 modelli tipo
mc1<-mcmc_areas(auc5, pars = c("3", "20", "47", "90"), prob = 0.95)+
  ggtitle("AUC a 5 anni")+
  labs(y="Modello")+
  theme(plot.title = element_text(size = 10.5))
mc2<-mcmc_areas(brier5, pars = c("3", "20", "47", "90"), prob = 0.95)+
  ggtitle("Brier a 5 anni")+
  labs(y="Modello")+
  theme(plot.title = element_text(size = 10.5))
mc3 <- mcmc_areas(brierover, pars = c("3", "20", "47", "90"), prob = 0.95)+
  ggtitle("Brier su tutto il periodo")+
  labs(y="Modello")+
  theme(plot.title = element_text(size = 10.5))
mc4<-mcmc_areas(cindex, pars = c("3", "20", "47", "90"), prob = 0.95)+
  ggtitle("C-index")+
  labs(y="Modello")+
  theme(plot.title = element_text(size = 10.5))

(mc1 + mc2) / (mc4+ mc3)
```

```
# Grafici DIC e WAIC
pldic<-
  DIC_db|>
  ggplot(aes(x=pDIC, y=DIC))+
    geom_point(colour="gray60")+
    geom_line(colour="blue3")+
    theme_classic()+
    geom_text(aes(x=20, y=350),
              label=expression(paste("log ",lambda, " ottimo = ", "10.39")), size=2.5)+
    geom_text(aes(x=20, y=340),
              label=paste("DIC:",round(DIC_db$DIC[which.min(DIC_db$DIC)],2), " ",
                          " pDIC:",round(DIC_db$pDIC[which.min(DIC_db$DIC)]+3, 2)
              ), size=2.5)
plwaic<-
  WAIC_db|>
  ggplot(aes(x=pWAIC, y=WAIC))+
    geom_point(colour="gray60")+
    geom_line(colour="blue3")+
    theme_classic()+
    geom_text(aes(x=40, y=400),
              label=expression(paste("log ",lambda, " ottimo =", " 10.39")), size=2.5)+
    geom_text(aes(x=40, y=385),
              label=paste("WAIC:",round(WAIC_db$WAIC[which.min(WAIC_db$WAIC)],2), " ",
                          "pWAIC:",round(WAIC_db$pWAIC[which.min(WAIC_db$WAIC)]+3, 2)
              ), size=2.5)
pldic + plwaic

# Grafico 100 distribuzioni casuali dal miglior modello
mcmcovertbest<-
  mcmc_areas(results[,,3][,-c(1,2,3, sample(4:2147, size = 2044, replace = F))])+
  theme(axis.text.y = element_blank(),
        axis.ticks.y = element_blank())
mcmcovertbest

# Distribuzioni Miglior modello
mcm1<-mcmc_areas(results[,,3],
              pars = c("v54", "v566", "v1491", "v1850", "log(shape)"), prob = 0.95)+
  ggtitle("Sottocampione")+
  theme(plot.title = element_text(size = 10.5))
mcm2<-mcmc_areas(results[,,3], pars = c("LP_OS"), prob = 0.95)+
  theme(axis.text.y = element_blank(),
        axis.ticks.y = element_blank())+
  ggtitle("Sarculator")+
  theme(plot.title = element_text(size = 10.5))
mcm1+mcm2

# Distribuzioni Performance miglior modello
mc1best<-mcmc_areas(auc5, pars = c("3"), prob = 0.95)+
  ggtitle("AUC a 5 anni Miglior modello")+
  labs(y=" ")+
  theme(plot.title = element_text(size = 10.5),
        axis.text.y = element_blank(),
        axis.ticks.y = element_blank())
mc2best<-mcmc_areas(brier5, pars = c("3"), prob = 0.95)+
  ggtitle("Brier a 5 anni Miglior modello")+
  labs(y=" ")+
  theme(plot.title = element_text(size = 10.5),
        axis.text.y = element_blank(),
        axis.ticks.y = element_blank())
mc3best <- mcmc_areas(brierover, pars = c("3"), prob = 0.95)+
  ggtitle("Brier su tutto il periodo Miglior modello")+
  labs(y=" ")+
```

```
  theme(plot.title = element_text(size = 10.5),
        axis.text.y = element_blank(),
        axis.ticks.y = element_blank())
mc4best<-mcmc_areas(cindex, pars = c("3"), prob = 0.95)+
  ggtitle("C-index Miglior modello")+
  labs(y=" ")+
  theme(plot.title = element_text(size = 10.5),
        axis.text.y = element_blank(),
        axis.ticks.y = element_blank())

(mc1best + mc2best) / (mc4best + mc3best)


quantile(auc5[,3], probs = c(0.025, 0.50, 0.975))
quantile(brier5[,3], probs = c(0.025, 0.50, 0.975))
quantile(brierover[,3], probs = c(0.025, 0.50, 0.975))
quantile(cindex[,3], probs = c(0.025, 0.50, 0.975))
mean(auc5[,3])
mean(brier5[,3])
mean(brierover[,3])
mean(cindex[,3])

#----------------------------
# Bayesian Model Averaging
#----------------------------
WAIC_st2<- WAIC_db$WAIC - min(WAIC_db$WAIC)
posterior_prob2 <- exp(-0.5*WAIC_st2)/sum(exp(-0.5*WAIC_st2))
resultsmix<-results
M2 <- resultsmix[,,1]
cbma2 <- cindex[,1]
boverbma2 <- brierover[,1]
auc5bma2 <- auc5[,1]
b5bma2 <- brier5[,1]
ind<-sample(1:dim(resultsmix)[3], size = 5000, prob = posterior_prob2, replace = T)
for(i in 1:nrow(M2)){
    M2[i,] <- as.vector(resultsmix[i,,ind[i]])
    cbma2[i] <- as.numeric(cindex[i,ind[i]])
    boverbma2[i] <- as.numeric(brierover[i,ind[i]])
    auc5bma2[i] <- as.numeric(auc5[i,ind[i]])
    b5bma2[i] <- as.numeric(brier5[i,ind[i]])
}

# Distribuzione coefficienti BMA
mcmbma12<-mcmc_areas(M2,
                  pars = c("v54", "v566", "v1491", "v1850", "log(shape)"), prob = 0.95)+
  ggtitle("Sottocampione BMA")+
  theme(plot.title = element_text(size = 10.5))
mcmbma21<-mcmc_areas(M2, pars = c("LP_OS"), prob = 0.95)+
  theme(axis.text.y = element_blank(),
        axis.ticks.y = element_blank())+
  ggtitle("Sarculator BMA")+
  theme(plot.title = element_text(size = 10.5))
mcmbma12 + mcmbma21

# Coefficienti stimati
mean(M2[,1])
mean(M2[,3])
quantile(M2[,3], probs = c(0.025, 0.975))
quantile(results[,3,3], probs = c(0.025, 0.975))
exp(mean(M2[,1]))
exp(mean(M2[,3]))
```

```
# Performance BMA
db_performance_bma2 <- cbind.data.frame(auc5bma2, b5bma2, boverbma2, cbma2)
mc1bma2<-mcmc_areas(db_performance_bma2, pars = c("auc5bma2"), prob = 0.95)+
  ggtitle("AUC a 5 anni BMA")+
  theme(plot.title = element_text(size = 10.5),
        axis.text.y = element_blank(),
        axis.ticks.y=element_blank())
mc2bma2<-mcmc_areas(db_performance_bma2, pars = "b5bma2", prob = 0.95)+
  ggtitle("Brier a 5 anni BMA")+
  theme(plot.title = element_text(size = 10.5),
        axis.text.y = element_blank(),
        axis.ticks.y=element_blank())
mc3bma2 <- mcmc_areas(db_performance_bma2, pars="boverbma2", prob = 0.95)+
  ggtitle("Brier su tutto il periodo BMA")+
  labs(y=expr(lambda))+
  theme(plot.title = element_text(size = 10.5),
        axis.text.y = element_blank(),
        axis.ticks.y=element_blank() )
mc4bma2<-mcmc_areas(db_performance_bma2, pars=c("cbma2"), prob = 0.95)+
  ggtitle("C-index BMA")+
  theme(plot.title = element_text(size = 10.5),
        axis.text.y = element_blank(),
        axis.ticks.y=element_blank())

(mc1bma2 + mc2bma2) / (mc4bma2+ mc3bma2)

quantile(auc5bma2, probs = c(0.025, 0.50, 0.975))
quantile(b5bma2, probs = c(0.025, 0.50, 0.975))
quantile(boverbma2, probs = c(0.025, 0.50, 0.975))
quantile(cbma2, probs = c(0.025, 0.50, 0.975))
mean(auc5bma2)
mean(cbma2)
mean(boverbma2)
mean(b5bma2)

# Confronto BMA vs Miglior modello
(mcm1+mcm2)/(mcmbma12 + mcmbma21)

auc5 <- cbind.data.frame(auc5,"BMA"=auc5bma2)
brier5 <- cbind.data.frame(brier5,"BMA"=b5bma2)
brierover <-cbind.data.frame(brierover,"BMA"=boverbma2)
cindex <- cbind.data.frame(cindex,"BMA"=cbma2)

conf1<-mcmc_areas(auc5, pars = c("BMA", "3"), prob = 0.95)+
  ggtitle("AUC a 5 anni")+
  labs(y="Modello")+
  theme(plot.title = element_text(size = 10.5))
conf2<-mcmc_areas(brier5, pars = c("BMA", "3"), prob = 0.95)+
  ggtitle("Brier a 5 anni")+
  labs(y="Modello")+
  theme(plot.title = element_text(size = 10.5))
conf3 <- mcmc_areas(brierover, pars = c("BMA", "3"), prob = 0.95)+
  ggtitle("Brier su tutto il periodo")+
  labs(y="Modello")+
  theme(plot.title = element_text(size = 10.5))
conf4<-mcmc_areas(cindex, pars = c("BMA", "3"), prob = 0.95)+
  ggtitle("C-index")+
  labs(y="Modello")+
  theme(plot.title = element_text(size = 10.5))

(conf1 + conf2) / (conf4 + conf3)
```

```
# Calibrazione BMA
R<-vector()
j=1
for(i in 1:dim(M2)[1]){
  R[j:(j+90)] <- 1- exp(- 60^(exp(M2[i,1])) * exp(X%*%M2[i,-1] ))
  j=j+91
}

R2<-
  cbind.data.frame(
  Previsioni=data.frame(R=R, id=rep(1:91,5000))|>
    group_by(id)|>
    dplyr::summarise( Ri= mean(R))|>
    ungroup()|>
    dplyr::select(Ri)|>
    unlist()|>
    unname(),
  Osservati=db_variable$status_bin
  )
R<-vector()
j=1
for(i in 1:dim(M2)[1]){
  R[j:(j+90)] <- 1- exp(- max(tempo)^(exp(M2[i,1])) * exp(X%*%M2[i,-1] ))
  j=j+91
}

R3<-
  cbind.data.frame(
  Previsioni=data.frame(R=R, id=rep(1:91,5000))|>
    group_by(id)|>
    dplyr::summarise( Ri= mean(R))|>
    ungroup()|>
    dplyr::select(Ri)|>
    unlist()|>
    unname(),
  Osservati=db_variable$status_bin
  )
cli<-data.frame(".pred_surv"=1-R3$Previsioni, ".pred_risk"=R3$Previsioni,
           "Class"=factor(ifelse(R3$Osservati==0, "surv", "risk") ) )|>
  as_tibble()

cliover<-
  cli|>cal_plot_breaks(Class, .pred_risk, num_breaks = 4,
                           event_level="first", conf_level = 0.95)+
  ylab("Rischio Osservato")+
  xlab("Rischio Stimato")+
  ggtitle("Calibrazione su tutto il periodo")+
  theme(plot.title = element_text(size = 10.5))

cli2<-data.frame(".pred_surv"=1-R2$Previsioni, ".pred_risk"=R2$Previsioni,
           "Class"=factor(ifelse(R2$Osservati==0, "surv", "risk") ) )|>
  as_tibble()

cli5m<-
  cli2|>cal_plot_breaks(Class, .pred_risk, num_breaks = 4,
                           event_level="first", conf_level = 0.95)+
  ylab("Rischio Osservato")+
  xlab("Rischio Stimato")+
  ggtitle("Calibrazione a 5 anni")+
  theme(plot.title = element_text(size = 10.5))

cli5m + cliover
```

```
# Confronto modello base con BMA
basedens<- mcmc_areas(results0,pars = c("log(shape)", "LP_OS"),  prob = 0.95,
           area_method="scaled height")

basetrace <-mcmc_trace(results0, pars = c("log(shape)", "LP_OS"))
basetrace / basedens

# Test di ipotesi
WAIC0<- WAIC_fun0(M=results0, X=X, t=tempo, status=stato)
WAIC_st3<- c(unname(WAIC0)-min(WAIC_db$WAIC), WAIC_st2)
p0<- (exp(-0.5*WAIC_st3[1])*0.5)/
                            (exp(-0.5*WAIC_st3[1])*0.5 +
                               sum(exp(-0.5*WAIC_st3[-1])*0.5/length(WAIC_st3[-1]) ) )
p1<- (exp(-0.5*WAIC_st3[-1])*0.5/length(WAIC_st3[-1]))/
                            (exp(-0.5*WAIC_st3[1])*0.5 +
                            sum(exp(-0.5*WAIC_st3[-1])*0.5/length(WAIC_st3[-1]) ) )
posterior_probability_test <- c(p0, p1)

# Grafico probabilità a posteriori
posterior_prob_plot<-
data.frame("Prob"= posterior_probability_test,
           "Modello"= c("Sarculator", rep("Sarculator + Radiomica",
                                          length(WAIC_st3[-1])))
           )|>
  arrange(Prob, desc=T)|>
  mutate(Indice=1:length(posterior_probability_test))|>
  ggplot(aes(x=Indice, y=Prob, col=Modello))+
  geom_point(alpha=0.8, size=2)+
  geom_line(alpha=0.6, colour="gray60")+
  ylab("Probabilità a posteriori del modello")+
  theme_classic()+
  theme(legend.position = "top")

posterior_prob_plot


# Overall Survival
tempo_asc<- sort(tempo,decreasing = F)
survmatrix <- matrix(ncol = 3, nrow = (nrow(X)^2)*nrow(M2) )
survmatrix[,1] <- rep(1:nrow(X), each=nrow(X)*nrow(M2))
survmatrix[,2] <- rep(tempo_asc, nrow(X)*nrow(M2))
colnames(survmatrix)<-c("ID","Tempo","Sopravvivenza")

j=1
for(i in 1:nrow(X)){
   for(k in 1:nrow(M2)){
     malphabma <- exp(M2[k,1])
     tmeanbma <- as.vector(M2[k,-1])
     survmatrix[c(j:(j+90)),3] <- exp(- ( tempo_asc^(malphabma) ) *
                                        (as.vector(exp(X[i, ]%*%tmeanbma) ) ) )
     j=j+91
   }
}

survmatrix<-rbind(survmatrix, matrix( c(1:91, rep(0, 91),
                          rep(1, 91)), nrow = 91, ncol = 3, byrow = F ) )

survplot<-data.frame(id=survmatrix[,1], Tempo=survmatrix[,2],
           sopravvivenza=survmatrix[,3])|>
  group_by(Tempo)|>
```

```
  dplyr::summarise("lb"=quantile(sopravvivenza, 0.025),
            "ub"=quantile(sopravvivenza, 0.975),
            "Sopravvivenza"=mean(sopravvivenza)
            )|>
  ggplot(aes(x=Tempo, y=Sopravvivenza, ymin=lb, ymax=ub))+
  geom_line()+
  theme_bw()+
  ylim(c(0,1))+
  geom_ribbon(alpha=0.2)+
  xlab("Tempo (mesi)")

survplot

# Sopravvivenza per 2 pazienti:
mediapz52<-
  rbind.data.frame(
      data.frame(id=survmatrix[,1], Tempo=survmatrix[,2],
              Sopravvivenza=survmatrix[,3])|>
      filter(id==52 & Tempo!=0)|>
      dplyr::mutate(iter=rep(1:5000, each=91)),
      data.frame(id=rep(52,5000), Tempo=rep(0, 5000),
              Sopravvivenza=rep(1,5000), iter=1:5000)
  )|>
  group_by(Tempo)|>
  dplyr::summarise(Sopravvivenza=mean(Sopravvivenza))|>
  mutate(iter=rep(1, 92))


mediapz6<-
  rbind.data.frame(
      data.frame(id=survmatrix[,1], Tempo=survmatrix[,2],
              Sopravvivenza=survmatrix[,3])|>
      filter(id==6 & Tempo!=0)|>
      dplyr::mutate(iter=rep(1:5000, each=91)),
      data.frame(id=rep(6,5000), Tempo=rep(0, 5000),
              Sopravvivenza=rep(1,5000), iter=1:5000)
  )|>
  group_by(Tempo)|>
  dplyr::summarise(Sopravvivenza=mean(Sopravvivenza))|>
  mutate(iter=rep(1, 92))


survplotex52<-
  rbind.data.frame(
      data.frame(id=survmatrix[,1], Tempo=survmatrix[,2],
              Sopravvivenza=survmatrix[,3])|>
      filter(id==52 & Tempo!=0)|>
      dplyr::mutate(iter=rep(1:5000, each=91)),
      data.frame(id=rep(52,5000), Tempo=rep(0, 5000),
              Sopravvivenza=rep(1,5000), iter=1:5000)
  )|>
  group_by(iter)|>
    ggplot(aes(x=Tempo, y=Sopravvivenza, col=iter))+
    geom_line(aes(group=iter), alpha=0.03, colour="gray60")+
    geom_line(data=mediapz52, linewidth=1.2)+
    theme_bw()+
    theme(legend.position = "none")+
    ylim(c(0,1))+
    ylab("Sopravvivenza stimata")+
    xlab("Tempo (mesi)")+
    ggtitle("ID Paziente: 52 (Deceduto a 20 mesi)")+
    theme(plot.title = element_text(size = 10.5))
```

```
survplotex6<-
  rbind.data.frame(
      data.frame(id=survmatrix[,1], Tempo=survmatrix[,2],
              Sopravvivenza=survmatrix[,3])|>
      filter(id==6 & Tempo!=0)|>
      dplyr::mutate(iter=rep(1:5000, each=91)),
      data.frame(id=rep(6,5000), Tempo=rep(0, 5000),
              Sopravvivenza=rep(1,5000), iter=1:5000)
  )|>
  group_by(iter)|>
    ggplot(aes(x=Tempo, y=Sopravvivenza, col=iter))+
    geom_line(aes(group=iter), alpha=0.03, colour="gray60")+
    geom_line(data=mediapz6, linewidth=1.2)+
    theme_bw()+
    theme(legend.position = "none")+
    ylim(c(0,1))+
    ylab("Sopravvivenza stimata")+
    xlab("Tempo (mesi)")+
    ggtitle("ID Paziente: 6 (Vivo a 95 mesi)")+
    theme(plot.title = element_text(size = 10.5))


survplotex6 + survplotex52
```

8 Ringraziamenti

Un grande ringraziamento e riconoscimento va anzitutto a chi ha seguito ogni singolo passaggio dello sviluppo di questa tesi. Grazie al Prof. Tommaso Rigon. Mi ha seguito dall'impostazione iniziale alla revisione finale, è stato sempre disponibile, fulmineo nel fornire soluzioni ai problemi che riscontravo. Mi ha fatto scoprire la bellezza della statistica bayesiana. Ho sempre pensato che la tesi non dovesse essere un semplice pro forma, ma un'occasione da cogliere per mettermi in discussione ed imparare. Ebbene, non penso potessi incontrare Relatore migliore. La stimo e la ammiro molto. Le critiche fatte durante la fase di correzione sono sempre state puntuali, precise e pungenti, ma anche quelle mi hanno motivato, mi hanno fatto venire voglia di riscrivere, rivedere, rileggere. Dedizione, precisione, talento, competenza, correttezza e umanità non sono virtù che è facile trovare in una persona oggi, e Lei le ha tutte. Le auguro davvero di fare la carriera che merita un docente del suo calibro. E auguro a me stesso di avere occasione di imparare ancora e ancora da Lei. Grazie davvero.

Prof.ssa Rosalba Miceli... C'è bisogno che Le dica quanto La ammiri e quanto Le sia grato? Ogni giorno imparo da Lei qualcosa di nuovo. Da quando La conosco non ho mai smesso di imparare. La ammiro per il gruppo di ricerca che sta creando e coordinando e di cui sono fiero di far parte. Lei è la mia correlatrice e il mio capo, ma sarebbe più corretto dire che Lei è la mia mentore. Sono ormai tre anni abbondanti che lavoro con Lei e mi ha sempre trattato con rispetto, umanità e gentilezza. Mi ha fatto sentire a mio agio sin dal primo giorno e mi ritengo davvero fortunato ad avere un capo come Lei. 10 ore di lavoro in media al giorno, rimanere in Istituto fino a sera quasi fino a far suonare l'allarme ogni volta. Eppure quando siamo lì non sembrano affatto 10 ore. Grazie di tutto! E miraccomando... Deve ridurre le ore di lavoro!

A Lidia. No, non basta un grazie. Siamo cresciuti insieme, abbiamo superato qualsiasi cosa. Mi conosci meglio di chiunque altro. Dire grazie, a te, sarebbe troppo scontato, non basterebbe e qualsiasi altra cosa che potrei dire sarebbe banale perché già detta. Quindi... La tesi, che rappresenta 7 mesi di duro lavoro la dedico a te (quindi leggila). Ti meriti tutta la felicità del mondo! Grazie per essere sempre al mio fianco soprattutto nei momenti no. Grazie! Già lo sai, ma non guasta mai ripeterlo, senza di te non sarei arrivato fino a qui. Quindi prenditi anche tu parte del merito! Sono orgoglioso di te, di dove sei arrivata e sono già orgoglioso di dove arriverai! Grazie Lidia per tutto quello che hai fatto e che continui a fare per me! Sei una persona stupenda.

Ai miei genitori e mio fratello, che mi sopportano nonostante con loro non riesco a far emergere quasi mai il lato bello di me. So di dare molte cose per scontate, non lo sono, tutt'altro. Quindi Grazie di sopportarmi e supportarmi in ogni passo che faccio. So che vorreste essere più partecipi e che io ve lo impedisco ma vi sono davvero grato e vi voglio bene in un modo che non riesco a esprimere, ora godetevi la casa nuova, il giardino, e soprattutto il meritato riposo! Finalmente, siete arrivati a raggiungere tutti i vostri traguardi, è il momento di pensare a sé, pensate sempre agli altri, me compreso, ora dovete pensare a voi! Davide! Che Grande! Un

mutuo a 22 anni, una casa col giardino, e una serenità invidiabile. Sono orgoglioso!

Un ringraziamento al mio Neurologo! Dott. Vittorio Martinelli... Ormai siamo a 3 anni e mezzo che ci conosciamo. Grazie per avermi rimesso in sesto, se cammino e posso svolgere una vita praticamente normale è perché non ha mai rinunciato a cercare la terapia adatta quando altri avevano già mollato. E grazie per continuare a cercare di farmi stare sempre meglio, si è preso cura di me a 360 gradi. Grazie!

Un grande Grazie a un mio collega e amico: Federico. Sono contento di aver trovato una persona come te in questo percorso! Ci siamo aiutati a vicenda, sempre. Come dimenticare la presentazione preparata alle 4 del mattino in piedi in un Autogrill a pochi minuti da Milano? Anche la statistica può essere divertente se sai scegliere le persone di cui circondarti! Mi ritengo fortunato di aver trovato un collega, ma prima ancora un amico, con cui condividere idee, pensieri, ragionamenti. Sì, anche in autogrill di ritorno dalla Summer School che hai organizzato egregiamente! Ma questo è il magnifico mondo della statistica. Quindi grazie! Di tutto, anche per il sostegno che mi hai dato al di là dell'ambito accademico.

Un Grazie sentito anche al Prof. Aldo Solari. Docente competente, professionale, umano, che con la sua sottile e pungente ironia ha tenuto uno dei corsi più interessanti e stimolanti di tutta la Magistrale, culminato con un esame nel quale all'ultima domanda non c'era una risposta corretta. Voleva solo vedere come ragionavamo! Penso che dei 180 minuti a disposizione tutti abbiano speso metà del tempo su quella domanda senza risposta. Dopo 5 anni di studi è stato in grado di farci appassionare nuovamente alla statistica, che nel suo corso appariva ogni lezione come qualcosa di nuovo e non ancora esplorato. È stato di grande ispirazione. Grazie!

E infine... No, non ringrazierò me stesso. Perché quello che sono e il punto a cui sono arrivato lo devo alle persone che ho appena ringraziato. Quindi a me stesso auguro di continuare a perseverare e di non mollare mai. Prima o poi la felicità e la serenità arriverà.

Bibliografia

Amin, M. B., Edge, S. B., Greene, F. L., Byrd, D. R., Brookland, R. K., Washington, M. K., Gershenwald, J. E., Compton, C. C., Hess, K. R., Sullivan, D. C., et al. (2017). *AJCC cancer staging manual*, volume 1024. Springer.

Ando, T. (2007). Bayesian predictive information criterion for the evaluation of hierarchical bayesian and empirical bayes models. *Biometrika*, 94(2):443–458.

Azzalini, A. and Scarpa, B. (2012). *Data analysis and data mining: An introduction*. OUP USA.

Beskos, A. and Stuart, A. (2009). Computational complexity of metropolis-hastings methods in high dimensions.

Bhadra, A., Datta, J., Polson, N. G., and Willard, B. (2019). Lasso meets horseshoe. *Statistical Science*, 34(3):405–427.

Bilimoria, K. Y., Stewart, A. K., Winchester, D. P., and Ko, C. Y. (2008). The national cancer data base: a powerful initiative to improve cancer care in the united states. https://www.facs.org/quality-programs/cancer-programs/national-cancer-database/.

Bologna, M., Tenconi, C., Corino, V. D., Annunziata, G., Orlandi, E., Calareso, G., Pignoli, E., Valdagni, R., Mainardi, L. T., and Rancati, T. (2023). Repeatability and reproducibility of mri-radiomic features: A phantom experiment on a 1.5 t scanner. *Medical Physics*, 50(2):750–762.

Brake, D. A., Hauenstein, J. D., Schreyer, F.-O., Sommese, A. J., and Stillman, M. E. (2019). Singular value decomposition of complexes. *SIAM Journal on Applied Algebra and Geometry*, 3(3):507–522.

Breslow, N. E. (1975). Analysis of survival data under the proportional hazards model. *International Statistical Review/Revue Internationale de Statistique*, pages 45–57.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1):1–3.

Broman, K. W. and Speed, T. P. (2002). A model selection approach for the identification of quantitative trait loci in experimental crosses. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 64(4):641–656.

Burnham, K. P. and Anderson, D. R. (2004). Multimodel inference: understanding aic and bic in model selection. *Sociological methods & research*, 33(2):261–304.

Burningham, Z., Hashibe, M., Spector, L., and Schiffman, J. D. (2012). The epidemiology of sarcoma. *Clinical sarcoma research*, 2(1):1–16.

Callegaro, D., Miceli, R., Bonvalot, S., Ferguson, P., Strauss, D. C., Levy, A., Griffin, A., Hayes, A. J., Stacchiotti, S., Le Pechoux, C., et al. (2016). Development and external validation of two nomograms to predict overall survival and occurrence of distant metastases in adults after surgical resection of localised soft-tissue sarcomas of the extremities: a retrospective analysis. *The Lancet Oncology*, 17(5):671–680.

Callegaro, D., Miceli, R., Bonvalot, S., Ferguson, P. C., Strauss, D. C., van Praag, V. V., Levy, A., Griffin, A. M., Hayes, A. J., Stacchiotti, S., et al. (2019). Development and external validation of a dynamic prognostic nomogram for primary extremity soft tissue sarcoma survivors. *EClinicalMedicine*, 17.

Carlin, B. P. and Chib, S. (1995). Bayesian model choice via markov chain monte carlo methods. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 57(3):473–484.

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2009). Handling sparsity via the horseshoe. *Journal of Machine Learning Research - Proceedings Track*, 5:73–80.

Chen, J. and Chen, Z. (2008). Extended bayesian information criteria for model selection with large model spaces. *Biometrika*, 95(3):759–771.

Chib, S. and Greenberg, E. (1995). Understanding the metropolis-hastings algorithm. *The american statistician*, 49(4):327–335.

Coriasco, M., Rampado, O., and Bradac, G. B. (2014). *Elementi di risonanza magnetica: Dal protone alle sequenze per le principali applicazioni diagnostiche*. Springer Science & Business Media.

Corino, V. D., Montin, E., Messina, A., Casali, P. G., Gronchi, A., Marchianò,

A., and Mainardi, L. T. (2018). Radiomic analysis of soft tissues sarcomas can distinguish intermediate from high-grade lesions. *Journal of Magnetic Resonance Imaging*, 47(3):829–840.

Cormier, J. N. and Pollock, R. E. (2004). Soft tissue sarcomas. *CA: a cancer journal for clinicians*, 54(2):94–109.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202.

Cox, D. R. and Oakes, D. (1984). *Analysis of survival data*, volume 21. CRC press.

Csiszár, I. (1975). I-divergence geometry of probability distributions and minimization problems. *The annals of probability*, pages 146–158.

Dennis Jr, J. E., Gay, D. M., and Walsh, R. E. (1981). An adaptive nonlinear least-squares algorithm. *ACM Transactions on Mathematical Software (TOMS)*, 7(3):348–368.

Dobson, R. and Giovannoni, G. (2019). Multiple sclerosis–a review. *European journal of neurology*, 26(1):27–40.

Durrett, R. and Durrett, R. (1999). *Essentials of stochastic processes*, volume 1. Springer.

Efron, B. and Hastie, T. (2021). *Computer age statistical inference, student edition: algorithms, evidence, and data science*, volume 6. Cambridge University Press.

Fanciullo, C., Gitto, S., Carlicchi, E., Albano, D., Messina, C., and Sconfienza, L. M. (2022). Radiomics of musculoskeletal sarcomas: a narrative review. *Journal of Imaging*, 8(2):45.

Fazel, M., Dufresne, A., Vanacker, H., Waissi, W., Blay, J.-Y., and Brahmi, M. (2023). Immunotherapy for soft tissue sarcomas: Anti-pd1/pdl1 and beyond. *Cancers*, 15(6):1643.

Ferreira, A. J. and Figueiredo, M. A. (2012). Boosting algorithms: A review of

methods, theory, and applications. *Ensemble machine learning: Methods and applications*, pages 35–85.

Fragoso, T. M., Bertoli, W., and Louzada, F. (2018). Bayesian model averaging: A systematic review and conceptual classification. *International Statistical Review*, 86(1):1–28.

Gandini, S., Massi, D., and Mandalà, M. (2016). Pd-l1 expression in cancer patients receiving anti pd-1/pd-l1 antibodies: A systematic review and meta-analysis. *Critical reviews in oncology/hematology*, 100:88–98.

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

George, E. I. and McCulloch, R. E. (1997). Approaches for bayesian variable selection. *Statistica sinica*, pages 339–373.

Girolami, M. and Calderhead, B. (2011). Riemann manifold langevin and hamiltonian monte carlo methods. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 73(2):123–214.

Gohain, P. B. and Jansson, M. (2023). Robust information criterion for model selection in sparse high-dimensional linear regression models. *IEEE Transactions on Signal Processing*.

Griffin, J. E. and Brown, P. J. (2011). Bayesian hyper-lassos with non-convex penalization. *Australian & New Zealand Journal of Statistics*, 53(4):423–442.

Gronau, Q. F., Sarafoglou, A., Matzke, D., Ly, A., Boehm, U., Marsman, M., Leslie, D. S., Forster, J. J., Wagenmakers, E.-J., and Steingroever, H. (2017). A tutorial on bridge sampling. *Journal of mathematical psychology*, 81:80–97.

Haario, H., Saksman, E., and Tamminen, J. (2001). An adaptive metropolis algorithm. *Bernoulli*, pages 223–242.

Hallinan Jr, A. J. (1993). A review of the weibull distribution. *Journal of Quality Technology*, 25(2):85–93.

Harrell, F. E. et al. (2001). *Regression modeling strategies: with applications to*

linear models, logistic regression, and survival analysis, volume 608. Springer.

Harrell Jr, F., Lee, K., and Mark, D. (1996). Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine*, 15(4):361–387.

Hastings, W. K. (1970). Monte carlo sampling methods using markov chains and their applications.

Heyde, C. C. (1963). On a property of the lognormal distribution. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 25(2):392–393.

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999). Bayesian model averaging: a tutorial (with comments by m. clyde, david draper and ei george, and a rejoinder by the authors. *Statistical science*, 14(4):382–417.

Hsiang, T. (1975). A bayesian view on ridge regression. *Journal of the Royal Statistical Society Series D: The Statistician*, 24(4):267–268.

Ichimaru, S. (2018). *Basic principles of plasma physics: a statistical approach*. CRC Press.

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

Jain, S., Xu, R., Prieto, V. G., and Lee, P. (2010). Molecular classification of soft tissue sarcomas and its clinical applications. *International journal of clinical and experimental pathology*, 3(4):416.

Jeffreys, H. (1998). *The theory of probability*. OuP Oxford.

Jevons, W. S. (1877). *The principles of science: A treatise on logic and scientific method*, volume 1. Macmillan and Company.

Johnson, R. E., Washington, M. T., Prakash, S., and Prakash, L. (2000). Fidelity of human dna polymerase η. *Journal of Biological Chemistry*, 275(11):7447–7450.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282):457–481.

Kloeden, P. E., Platen, E., Kloeden, P. E., and Platen, E. (1992). *Stochastic differential equations*. Springer.

Konishi, S. and Kitagawa, G. (2008). *Information criteria and statistical modeling*. Springer Science & Business Media.

Landau, L. D. and Lifshitz, E. M. (2013). *Course of theoretical physics*. Elsevier.

Langevin, P. (1908). Sur la théorie du mouvement brownien. *Compt. Rendus*, 146:530–533.

Lodish, H., Berk, A., Raff, M., Lewis, J., Kaiser, C. A., and Krieger, M. (2008). *Molecular Biology of the Cell*. Garland science.

Mayr, A., Binder, H., Gefeller, O., and Schmid, M. (2014). The evolution of boosting algorithms. *Methods of information in medicine*, 53(06):419–427.

McDonald, G. C. (2009). Ridge regression. *Wiley Interdisciplinary Reviews: Computational Statistics*, 1(1):93–100.

Meinshausen, N. and Bühlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 72(4):417–473.

Meng, X.-L. and Schilling, S. (2002). Warp bridge sampling. *Journal of Computational and Graphical Statistics*, 11(3):552–586.

Meng, X.-L. and Wong, W. H. (1996). Simulating ratios of normalizing constants via a simple identity: a theoretical exploration. *Statistica Sinica*, pages 831–860.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087–1092.

Miceli, R., Callegaro, D., Barretta, F., Gronchi, A., and Vergani, R. (2022). Sarculator 2.1.2. https://apps.apple.com/na/app/sarculator/id1052119173, https://play.google.com/store/apps/details?id=it.digitalforest.sarculator&hl=it&gl=US.

Neal, R. M. et al. (2011). Mcmc using hamiltonian dynamics. *Handbook of markov*

chain monte carlo, 2(11):2.

Owrang, A. and Jansson, M. (2018). A model selection criterion for high-dimensional linear regression. *IEEE Transactions on Signal Processing*, 66(13):3436–3446.

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

Peeken, J. C., Bernhofer, M., Spraker, M. B., Pfeiffer, D., Devecka, M., Thamer, A., Shouman, M. A., Ott, A., Nüsslin, F., Mayr, N. A., et al. (2019a). Ct-based radiomic features predict tumor grading and have prognostic value in patients with soft tissue sarcomas treated with neoadjuvant radiation therapy. *Radiotherapy and Oncology*, 135:187–196.

Peeken, J. C., Spraker, M. B., Knebel, C., Dapper, H., Pfeiffer, D., Devecka, M., Thamer, A., Shouman, M. A., Ott, A., von Eisenhart-Rothe, R., et al. (2019b). Tumor grading of soft tissue sarcomas using mri-based radiomics. *EBioMedicine*, 48:332–340.

Piironen, J., Paasiniemi, M., and Vehtari, A. (2020). Projective inference in high-dimensional problems: Prediction and feature selection.

Piironen, J. and Vehtari, A. (2017). Sparsity information and regularization in the horseshoe and other shrinkage priors.

Raftery, A. E., Madigan, D., and Hoeting, J. A. (1997). Bayesian model averaging for linear regression models. *Journal of the American Statistical Association*, 92(437):179–191.

Robert, C. P., Casella, G., and Casella, G. (1999). *Monte Carlo statistical methods*, volume 2. Springer.

Robert, C. P. et al. (2007). *The Bayesian choice: from decision-theoretic foundations to computational implementation*, volume 2. Springer.

Roberts, G. O. and Rosenthal, J. S. (2001). Optimal scaling for various metropolis-hastings algorithms. *Statistical science*, 16(4):351–367.

Roberts, G. O. and Rosenthal, J. S. (2007). Coupling and ergodicity of adaptive markov chain monte carlo algorithms. *Journal of applied probability*, 44(2):458–475.

Roberts, G. O. and Rosenthal, J. S. (2009). Examples of adaptive mcmc. *Journal of computational and graphical statistics*, 18(2):349–367.

Ross, S. M. (2014). *Introduction to probability models*. Academic press.

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics*, pages 461–464.

Sherlock, C., Fearnhead, P., and Roberts, G. O. (2010). The random walk metropolis: linking theory and practice through a case study.

Sinha, S. and Peach, A. H. S. (2010). Diagnosis and management of soft tissue sarcoma. *Bmj*, 341.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 64(4):583–639.

Spraker, M. B., Wootton, L. S., Hippe, D. S., Ball, K. C., Peeken, J. C., Macomber, M. W., Chapman, T. R., Hoff, M. N., Kim, E. Y., Pollack, S. M., et al. (2019). Mri radiomic features are independently associated with overall survival in soft tissue sarcoma. *Advances in radiation oncology*, 4(2):413–421.

Stiller, C., Trama, A., Serraino, D., Rossi, S., Navarro, C., Chirlaque, M., Casali, P., Group, R. W., et al. (2013). Descriptive epidemiology of sarcomas in europe: report from the rarecare project. *European journal of cancer*, 49(3):684–695.

Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., and Bray, F. (2021). Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, 71(3):209–249.

Thall, P. F., Russell, K. E., and Simon, R. M. (1997). Variable selection in regression via repeated data splitting. *Journal of Computational and Graphical Statistics*, 6(4):416–434.

Theobald, C. M. (1974). Generalizations of mean square error applied to ridge regression. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 36(1):103–106.

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

Tibshirani, R. and Wasserman, L. (2016). A closer look at sparse regression. *Lecture notes*.

Tierney, L. and Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the american statistical association*, 81(393):82–86.

Tjorve, E. (2009). Shapes and functions of species–area curves (ii): a review of new models and parameterizations. *Journal of Biogeography*, 36(8):1435–1445.

Trama, A., Badalamenti, G., Baldi, G. G., Brunello, A., Caira, M., Drove, N., Marrari, A., Palmerini, E., Vincenzi, B., Dei Tos, A. P., et al. (2019). Soft tissue sarcoma in italy: From epidemiological data to clinical networking to improve patient care and outcomes. *Cancer Epidemiology*, 59:258–264.

Van Der Linde, A. (2005). Dic in variable selection. *Statistica Neerlandica*, 59(1):45–56.

Van der Vaart, A. W. (2000). *Asymptotic statistics*, volume 3. Cambridge university press.

Van Erp, S., Oberski, D. L., and Mulder, J. (2019). Shrinkage priors for bayesian penalized regression. *Journal of Mathematical Psychology*, 89:31–50.

Vehtari, A., Gelman, A., and Gabry, J. (2017). Practical bayesian model evaluation using leave-one-out cross-validation and waic. *Statistics and computing*, 27:1413–1432.

Vehtari, A. and Lampinen, J. (2002). Bayesian model assessment and comparison using cross-validation predictive densities. *Neural computation*, 14(10):2439–2468.

Voss, R. K., Callegaro, D., Chiang, Y.-J., Fiore, M., Miceli, R., Keung, E. Z., Feig, B. W., Torres, K. E., Scally, C. P., Hunt, K. K., et al. (2022). Sarculator is a good model to predict survival in resected extremity and trunk sarcomas in us patients. *Annals of surgical oncology*, 29(7):4376–4385.

Wallis, W. A. and Roberts, H. V. (2014). *The nature of statistics*. Courier Corporation.

Wang, M. C. and Uhlenbeck, G. E. (1945). On the theory of the brownian motion ii. *Reviews of modern physics*, 17(2-3):323.

Wasserman, L. (2000). Bayesian model selection and model averaging. *Journal of mathematical psychology*, 44(1):92–107.

Watanabe, S. (2013). A widely applicable bayesian information criterion. *The Journal of Machine Learning Research*, 14(1):867–897.

Watanabe, S. and Opper, M. (2010). Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of machine learning research*, 11(12).

Wedeen, V. J., Hagmann, P., Tseng, W.-Y. I., Reese, T. G., and Weisskoff, R. M. (2005). Mapping complex tissue architecture with diffusion spectrum magnetic resonance imaging. *Magnetic resonance in medicine*, 54(6):1377–1386.

Wilkinson, J. H., Bauer, F. L., and Reinsch, C. (2013). *Linear algebra*, volume 2. Springer.

Zhao, S., Su, Y., Duan, J., Qiu, Q., Ge, X., Wang, A., and Yin, Y. (2019). Radiomics signature extracted from diffusion-weighted magnetic resonance imaging predicts outcomes in osteosarcoma. *Journal of Bone Oncology*, 19:100263.