Univerità degli studi di Padova
Dipartimento di Scienze Statistiche
Corso di Laurea Magistrale in

Scienze Statistiche

Tesi

Analisi Bayesiana dei Valori Estremi con Distribuzioni a Priori Informative

Relatore: Prof. Antonio Canale
Dipartimento di Scienze Statistiche
Correlatori: Prof.ssa Ilaria Prosdocimi
Prof.ssa Isadora Antoniano Villalobos

Laureando: Simone Meneghello
Matricola N. 2055679

Anno Accademico 2022/23

Indice

Introduzione

L'analisi statistica dei valori estremi ha assunto un'importanza crescente in diversi campi di applicazione, dai cambiamenti climatici all'ingegneria strutturale, dall'analisi del rischio finanziario alle scienze ambientali. Essa si occupa di comprendere e modellare eventi rari che, nonostante la loro bassa probabilità di occorrenza, possono avere impatti significativi e spesso devastanti. Tra le varie categorie di eventi estremi sono inclusi numerosi fenomeni ambientali, quali frane, alluvioni e tempeste. L'efficacia delle misure preventive contro le loro conseguenze potenzialmente catastrofiche dipende in modo cruciale da studi dettagliati e mirati alla specifica natura dei dati in questione.
La distribuzione generalizzata dei valori estremi (GEV) è uno degli strumenti chiave in questo contesto, poiché fornisce un quadro flessibile per descrivere le distribuzioni dei valori massimi in campioni di dati. Tuttavia, l'inferenza tradizionale per la GEV basata sulla massima verosimiglianza può avere limitazioni, specialmente quando la dimensione del campione è piccola. L'approccio bayesiano, che combina informazioni a priori con le informazioni fornite dai dati, offre una soluzione promettente sopratutto nel contesto di campioni piccoli. In particolare, l'uso di distribuzioni a priori informative può migliorare la precisione e la robustezza delle stime, specialmente quando le informazioni a priori sono basate su conoscenze scientifiche o su studi precedenti.
Nel seguente caso specifico si decide di portare una conoscenza legata al parametro di forma della distribuzione generalizzata dei valori estremi. Il valore di tale parametro, solitamente definito come ξ a seconda del valore che presenta, definisce una delle tre distribuzioni che vengono raggruppate nella GEV, e che descrive tre comportamenti distinti dei valori estremi: positivo (Fréchet), zero (Gumbel) e negativo (Weibull). La scelta di tale parametro e quindi del comportamento dei valori estremi è determinante nella modellazione statistica nel seguente contesto. Una strategia che comunemente si adotta in questi casi è quella di lavorare direttamente con la generica distribuzione GEV che racchiude tutte e tre i comportamenti, ma questo tipo di approccio non considera mai il tipo di modello Gumbel poiché tale scelta è ridotta a un singolo punto in uno spazio parametrico continuo.
La distribuzione di Gumbel assume un ruolo fondamentale nell'analisi dei valori estremi, essendo la distribuzione asintotica per il massimo di variabili che seguono distribuzioni stocastiche molto comuni. Questa caratteristica ne ha favorito la popolarità e l'adozione in molteplici campi scientifici, come l'idrologia, dove è spesso impiegata direttamente per la

modellazione, sfruttando inoltre delle conoscenze derivate da ricerche pregresse nel settore. Restringersi all'utilizzo del modello Gumbel però può portare a delle stime distorte e a previsioni sbagliate quando il modello generatore dei dati è un altro della famiglia GEV. Ciò potrebbe portare a una pericolosa sottostima dei rischi associati ad eventi catastrofici. Nella seguente tesi si propone un approccio informativo per l'analisi bayesiana dei valori estremi che incorpora specificamente il caso Gumbel nella modellazione. Questo si realizza assegnando una distribuzione a priori al parametro di forma, che pone un'elevata probabilità a zero, ma consente al contempo che le evidenze derivanti dai dati suggeriscano differenti comportamenti asintotici. Per fare questo vengono utilizzate delle a priori di contrazione, di tipo continuo e del tipo Spike and Slab, sfruttando alcune proposte che vengono usualmente utilizzate nell'ambito della regressione lineare, ma adattandole al caso d'interesse. Queste metodologie appaiono particolarmente utili per l'analisi del parametro ξ, in particolare quando l'ipotesi di base presuppone che tale parametro sia nullo, pur concedendo la flessibilità di una sua potenziale variazione. Ci si concentra sull'efficacia di tali metodi, valutando la loro adattabilità e il comportamento all'interno del contesto di interesse, per stabilire se possano costituire delle basi per il modello proposto.
Oltre alla definizione di modelli per singoli campioni, questa tesi introduce un'estensione per dati strutturati gerarchicamente, in cui i campioni sono interdipendenti o condividono caratteristiche comuni. Il primo capitolo descrive brevemente la teoria dei valori estremi, la distribuzione GEV e alcuni metodi di stima di tale modello mentre il secondo capitolo si concentra sulla descrizione di alcuni metodi di selezione delle variabili bayesiana nel caso della regressione lineare, attraverso l'utilizzo di distribuzioni a priori di contrazione. Nel terzo capitolo vengono descritti i metodi proposti, mostrando i modelli dettagliatamente e le distribuzioni a priori che vengono suggerite per risolvere la problematica d'interesse legata al parametro di forma della GEV, mostrando anche i metodi utilizzati per il calcolo delle distribuzioni a posteriori. Il Capitolo 4 presenta uno studio di simulazione condotto per esaminare l'efficacia dei modelli proposti in vari scenari realistici, fornendo una valutazione dettagliata delle prestazioni dei modelli in condizioni differenti e verosimili. Infine, il Capitolo 5 si dedica all'analisi dei dati relativi alle precipitazioni forniti dall'ARPA per le regioni del Veneto e del Trentino-Alto Adige, applicando i modelli sviluppati. Questo consente non solo di caratterizzare il comportamento delle piogge nelle aree in esame, ma anche di dimostrare concretamente l'utilizzo dei modelli proposti nel contesto reale.
Le metodologie proposte, le simulazioni e le analisi svolte sono state svolte utilizzando il software statistico R, portando alla creazione di una serie di funzioni utilizzabili e modificabili per future analisi o ulteriori studi. La finalità di questo studio è che le tecniche introdotte possano servire come strumenti utili nell'analisi bayesiana dei valori estremi, nel caso in cui sia opportuno il loro impiego, raccomandando cautela nell'utilizzo data l'applicazione ad un settore che richiede grande attenzione.

Capitolo 1

Teoria e Distribuzione Generalizzata dei Valori Estremi

La teoria dei valori estremi è diventata negli ultimi anni una disciplina molto importante per le scienze applicate. Tale teoria si basa sullo studio di eventi estremi, cioè quei rari ed insoliti avvenimenti che si discostano dalla norma. La caratteristica che contraddistingue un'analisi dei valori estremi è l'obiettivo, che risulta essere quello di quantificare il comportamento stocastico di un processo a livelli insolitamente elevati - o bassi-. Si è quindi interessati all'analisi delle code della distribuzione dei dati, anziché sulle regioni centrali. I valori estremi sono di particolare interesse in molteplici settori, poiché possono avere impatti significativi sulla sicurezza, sull'economia e sull'ambiente. Esempi di eventi estremi includono alluvioni tempeste, onde oceaniche, picchi di consumo energetico e gravi crolli di mercato finanziario. Le tecniche che trattano questa tipologia di valori risultano essere quindi ampiamente utilizzate in vari campi, quali l'idrologia, ingegneria, meteorologia, telecomunicazioni, finanza e altre ancora.
L'uso degli strumenti statistici convenzionali per la previsione di eventi rari ha riscontrato una crescente insoddisfazione in ambito scientifico, portando quindi a prediligere un tipo di analisi specifica per tale problematica. La capacità di quantificare in modo accurato l'incidenza di eventi eccezionali è cruciale per valutare e mitigare i rischi associati a tali eventi. Le metodologie per la modellazione dei valori estremi consistono nell'adottare un modello asintotico per descrivere la variazione stocastica nei livelli estremi del processo.
I valori estremi per definizione sono poco frequenti e richiedono stime per livelli del processo che superano notevolmente quelli già osservati. La quantità di osservazioni a disposizione inoltre è spesso poco numerosa e l'obiettivo è di sviluppare modelli statistici che possano catturare in modo accurato e affidabile la probabilità di accadimento di eventi estremi, anche quando i dati osservati sono limitati. Questa necessità comporta un inevitabile utilizzo dell'estrapolazione per nuovi valori e la teoria dei valori estremi offre una categoria di modelli progettati per agevolare tale estrapolazione. Questi presentano sicuramente delle limitazioni ai quali è importante fare attenzione ma ad oggi non sono state proposte alter-

native più credibili. Un'ampia descrizione della teoria dei valori estremi viene fornita in *"An introduction to Statistical Modeling of Extreme Values"* Coles (2001).

1.1 Distribuzione Generalizzata dei Valori Estremi

In questo capitolo verrà trattata la classica teoria degli estremi, concentrandosi sul caso del massimo a blocchi nel contesto di osservazioni indipendenti. Il modello presentato è la base della teoria dei valori estremi e si concentra sullo studio di

$$M_n = \max\{X_1, \ldots, X_n\},$$

dove $X_1, \ldots, X_n$ è una sequenza di variabili casuali indipendenti con una distribuzione comune F. Nelle applicazioni X_i rappresentano valori di un processo misurato nel tempo, quindi M_n rappresenta il massimo del processo per n unità di tempo di osservazione. L'interesse è quindi quello di studiare la distribuzione di M_n, la quale in teoria può essere derivata per tutti i valori di n:

$$F(x) = P(X_1 \leq z) \times \cdots \times P(X_1 \leq z) = F(z)^n.$$

L'approccio che si utilizza nella pratica, quando la distribuzione F è ignota, è quello di lavorare con una famiglia di modelli che approssimino F^n, la quale può essere stimata sulla base dei soli valori estremi osservati. Questo approccio è simile all'utilizzo del teorema del limite centrale, approssimando la distribuzione della media campionaria attraverso la distribuzione normale. Considerazioni asinotitche suggeriscono che se esiste una sequenza di costanti $a_n > 0$ e b_n tale che

$$P\left(\frac{M_n - b_n}{a_n} \leq z\right) \to G(z) \quad \text{con } n \to \infty, \tag{1.1}$$

dove G è una funzione di probabilità, allora G è un membro della famiglia dei valori estremi generalizzati (GEV), la cui funzione di ripartizione è:

$$G(z) = \exp\left\{-\left[1 + \xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}\right\}, \tag{1.2}$$

definita in $\{z : 1 + \xi(z-\mu)/\sigma\}$, dove $\mu \in \mathbb{R}, \sigma > 0$ e $\xi \in \mathbb{R}$. In questa famiglia distributiva i parametri di posizione e di scala sono rispettivamente μ e σ, mentre ξ è il parametro di forma. Questa famiglia è formata da tre classi di distribuzioni che si contraddistinguono tra loro per il valore assunto dal parametro ξ : Gumbel, Fréchet e Weibull.

Con $\xi < 0$ la classe di distribuzione risultante sarà la **Weibull** mentre con $\xi > 0$ si ottiene la **Fréchet**. In entrambi i casi la funzione di ripartizione è definita come in equazione (1.2). Infine il sottoinsieme della famiglia GEV con $\xi = 0$ viene interpetato come il limite di (1.2) con $\xi \to 0$, portando quindi alla distribuzione **Gumbel** definita come:

$$G(z) = \exp\left\{-\exp\left[-\left(\frac{z-\mu}{\sigma}\right)\right]\right\}. \tag{1.3}$$

Quest'ultima classe distributiva presenta quindi solo i parametri di posizione e scala a differenza delle altre. Come si può notare in Figura(1.1) ciascuna delle classi presentate ha forme e comportamenti differenti, che corrispondono ai comportamenti differenti delle code della distribuzione F degli X_i.
Innanzitutto si può notare che il supporto è molto diverso a seconda del tipo di distribuzione. Infatti la distribuzione di Gumbel ha un supporto sull'intero dominio dei numeri reali $\mathbb{R}$, mentre le distribuzioni di Fréchet e Weibull presentano supporti limitati, la prima inferiormente e la seconda superiormente.
È interessante il comportamento della distribuzione G a z_+, l'estremo superiore. Per la Weibull la distribuzione a z_+ è finita, mentre per Fréchet e Gumbell la distribuzione a $z_+ = \infty$. Tuttavia la densità di G decade esponenzialmente per la distribuzione Gumbell mentre in modo polinomiale per la Frèchet, il che corrisponde a tassi di decadimento relativamente diversi nella coda di F. La distribuzione Weibull assegna una probabilità elevata ad eventi estremi più bassi, mentre la Fréchet con la coda destra molto pesante, suggerisce una maggiore probabilità di eventi estremi più alti. È chiaro quindi che le tre famiglie rappresentano il comportamento di valori estremi molto differenti.

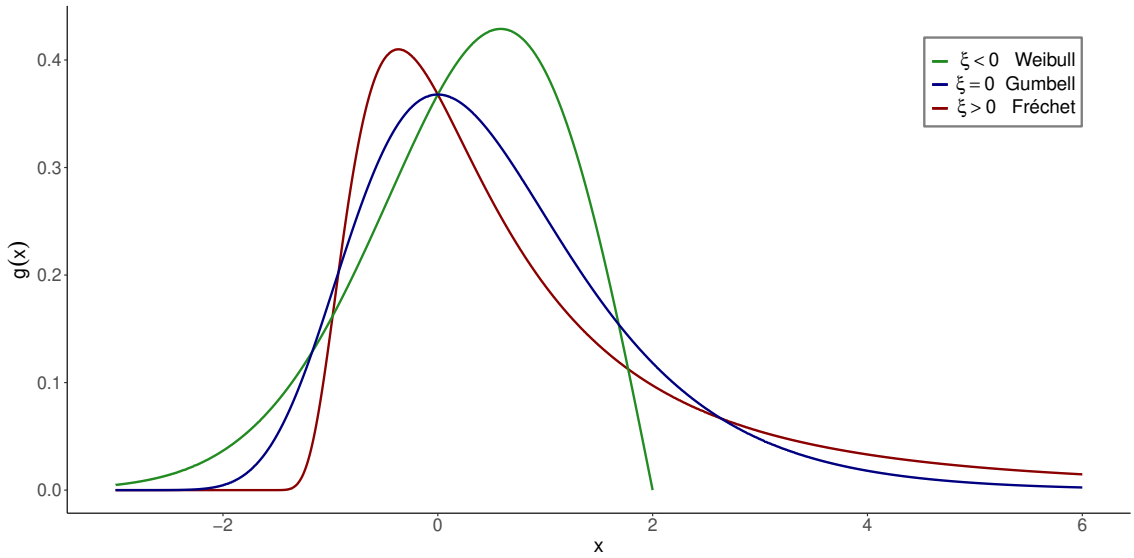

Figura 1.1: Funzione di densità delle tre classi della famiglia GEV: I) Weibull (con $\xi = -0.5$); II) Gumbell (con $\xi = 0$); III) Fréchet (con $\xi = 0.5$).

La distribuzione Gumbell viene spesso usata, anzichè utilizzare l'intera famiglia GEV, come modello per studiare il massimo nell'arco di un anno. In alcuni casi c'è evidenza empirica per supportare questa scelta, in altri casi viene argomentata dal fatto che, dato che molte distribuzioni per X_i porterebbero alla distribuzione limite Gumbell per M_n (come la distribuzione normale, lognormale e gamma ad esempio), questa risulta essere più appropriata rispetto alla generica GEV. Questa strategia risulta rischiosa anche quando il test d'ipotesi e le diagnostiche del modello supportano la riduzione del modello.

È evidente che la stima di ξ risulta molto importante nello studio dei valori estremi; una stima o un'assunzione a priori sul parametro potrebbe portare a risultati fuorvianti ed imprecisi. Ci si concentra ora sullo studio del comportamento di M_n, che non è ancora del tutto definito.
La formula (1.1) pone infatti il problema di conoscere le costanti di normalizzazioni a_n e b_n. Questo viene risolto dalla proprietà della GEV che risulta invariante a trasformazioni di forma e scala:

$$\begin{aligned} P(M_n \leq z) &\approx G\left(\frac{z - b_n}{a_n}\right) \\ &= G^*(z), \end{aligned}$$

dove G^* è un altro membro della famiglia GEV. Quindi se l'approssimazione della distribuzione di $(M_n - b_n)/b_n$ è un membro della famiglia GEV per n grande, la distribuzione di M_n può essere approssimata anch'essa da un differente membro della stessa famiglia. Siccome i parametri della distribuzione devono essere stimati in ogni caso, è irrilevante in pratica che i parametri di G siano differenti da quelli di G^*.
L'approccio che viene seguito per modellare gli estremi di una serie di osservazioni indipendenti $X_1, X_2, ...$ è quindi il seguente: I dati vengono quindi suddivisi in sequenze di osservazioni di lunghezza n, per qualche valore grande di n, creando una sequenza di massimi a blocchi $Z_1, ..., Z_m$. Nell'implementazione del modello è importante la scelta della dimensione dei blocchi n, la quale può risultare critica. La scelta dipende dal classico compromesso tra varianza e distorsione: blocchi troppo piccoli, comportano valori piccoli di n e il modello asintotico sarà scadente, con conseguente *bias* nelle stime e nell'estrapolazione; blocchi troppo grandi generano pochi massimi, portando ad una varianza delle stime elevata. Generalmente considerazioni pragmatiche portano all'adozione di blocchi della lunghezza di un anno ottenendo così dei massimi a blocchi che risultano essere massimi annuali.

1.1.1 Livello di Ritorno

Quando ci si concentra sugli aspetti specifici di un'analisi dei valori estremi, è pertinente valutare i modelli in termini delle loro implicazioni per i futuri valori estremi del processo. A tale scopo è utile calcolare una stima dei quantili degli estremi della distribuzione dei massimi annuale, la quale si ottiene invertendo l'equazione (1.2) e (1.3) :

$$z_p = \begin{cases} \mu - \dfrac{\sigma}{\xi}\left(1 - [-\log(1-p)]^{-\xi}\right), & \text{per } \xi \neq 0; \\ \mu - \sigma \log\left(-\log(1-p)\right), & \text{per } \xi = 0; \end{cases} \tag{1.4}$$

dove $G(z_p) = 1 - p$. Il valore z_p è definito come **livello di ritorno** associato al **periodo di ritorno** $1/p$. Questo significa che il livello z_p rappresenta un valore che ci si aspetta venga superato o raggiunto da un evento estremo, in media, una volta ogni $1/p$ anni, se ad

esempio la lunghezza dei blocchi è di un anno. In altre parole, il livello di ritorno indica il valore soglia al di sopra del quale ci si aspetta che un evento estremo si verifichi in media con una frequenza p nel corso del tempo, o degli anni, come nel caso d'interesse.
Un grafico spesso utilizzato è il *return level plot*, che rappresenta i valori del livello di ritorno sull'asse delle ordinate, contro i valori del livello di ritorno in scala logaritmica ($\log 1/p$) sull'asse delle ascisse. Con il seguente grafico è possibile quindi visualizzare come le code estreme della distribuzione di probabilità si comportano per vari periodi di ritorno. Questo consente di individuare facilmente i valori estremi e di capire quali eventi sono più rari o più frequenti.
Il grafico in Figura 5.3 mostra il *return level plot* per le tre distribuzioni della famiglia GEV. Da questo grafico sono molto evidenti le differenze, già citate precedentemente, nelle caratteristiche delle code delle varie distribuzioni. La scelta di rappresentare il periodo di ritorno in scala logaritmica permette di aumentare l'interesse sui valori che hanno probabilità e quindi evidenziare l'effetto dell'estrapolazione. Data questa caratteristica e la semplicità di interpretazione il *return level plot* risulta particolarmente conveniente sia per la presentazione del modello che per la sua validazione.

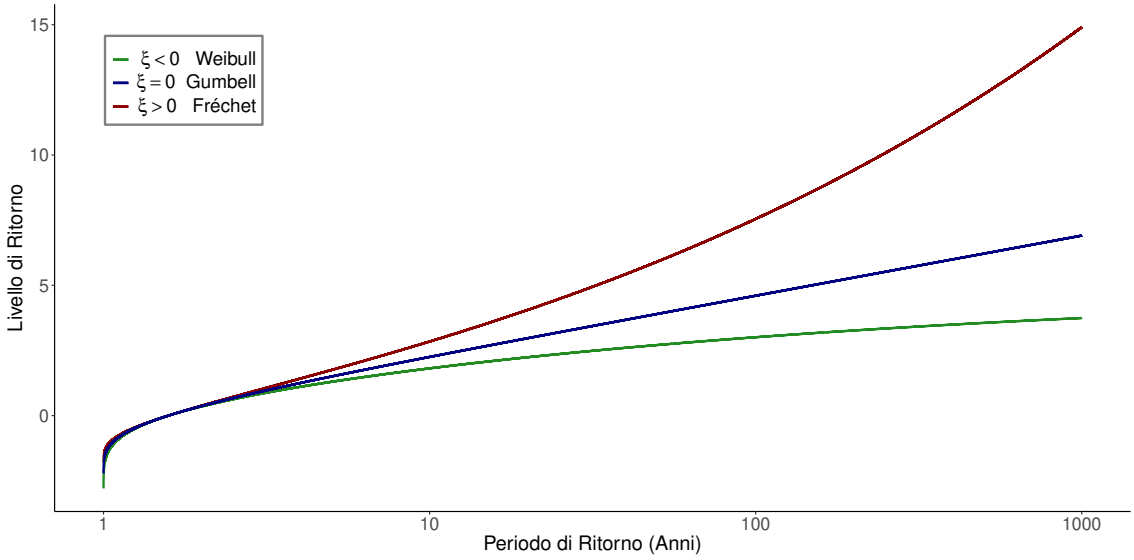

Figura 1.2: livello di ritorno per le varie distribuzioni GEV : I) Weibull (con $\xi = -0.2$); II) Gumbell (con $\xi = 0$); III) Fréchet (con $\xi = 0.2$).

1.2 Inferenza per la Distribuzione GEV

Come visto in precedenza, la distribuzione generalizzata dei valori estremi (GEV) fornisce un modello per la distribuzione dei massimi a blocchi. La sua applicazione consiste nella divisione dei dati in blocchi di lunghezza uguale per adattare la distribuzione GEV alla sequenza di massimi a blocchi; tuttavia, a volte si utilizza direttamente il valore massimo senza procedere alla divisione in blocchi, soprattutto quando si considerano osservazioni su

base annuale. Si assume quindi $Z_1, ..., Z_m$ una sequenza di massimi annuali indipendenti. L'assunzione di indipendenza è valida se le osservazioni rilevate nell'anno X_i sono indipendenti, ma risulta comunque ragionevole anche se gli X_i costituiscono una serie dipendente di valori (Coles, 2001).
Varie tecniche sono state proposte per la stima dei parametri nei modelli per valori estremi. Ciascuna tecnica presenta dei vantaggi e degli svantaggi ma l'utilità generale e l'adattabilità nella costruzione di modelli complessi delle tecniche basate sulla verosimiglianza rendono particolarmente preferibile questo approccio. Per il campione di massimi a blocchi $Z_1, \ldots, Z_m$ la verosimiglianza prende la forma:

$$L(\mu, \sigma, \xi; Z_1, \ldots, Z_m) = \prod_{i=1}^{m} g(z_i|\mu, \sigma, \xi), \tag{1.5}$$

dove g è la densità della distribuzione GEV o Gumbel, a seconda di cosa si decide di modellare. Ci sono due modi contrastanti per ottenere le stime dei parametri dalla funzione di verosimiglianza. Il primo è l'utilizzo della stima di massima verosimiglianza, mentre il secondo prevede l'utilizzo dell'inferenza bayesiana.

1.2.1 Stima di Massima Verosimiglianza

Il metodo della massima verosimiglianza vede la massimizzazione della log-verosimiglianza, per convenienza numerica, del campione di massimi a blocchi indipendenti $Z_1, \ldots, Z_m$ che è definita come:

$$\begin{aligned} l(\mu, \sigma, \xi; Z_1, \ldots, Z_m) = -m \log \sigma - \left(1 + \frac{1}{\xi}\right) \sum_{i=1}^{m} \log \left[1 + \xi \left(\frac{z_i - \mu}{\sigma}\right)\right] \\ - \sum_{i=1}^{m} \left[1 + \xi \left(\frac{z_i - \mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}, \end{aligned} \tag{1.6}$$

condizionato a $1 + \xi(z_i - \mu)/\sigma > 0$, per $i = 1, ..., m$.
Alla combinazione di parametri per la quale la condizione appena riportata viene violata, corrisponde una configurazione in cui almeno uno dei dati osservati supera un estremo della distribuzione, la verosimiglianza risulta quindi zero e la log-verosimiglianza $-\infty$.
Il caso $\xi = 0$ richiede un trattamento separato usando il limite Gumbel della distribuzione GEV, portando alla verosimiglianza:

$$l(\mu, \sigma; Z_1, \ldots, Z_m) = -m \log \sigma - \sum_{i=1}^{m} \left(\frac{z_i - \mu}{\sigma}\right) - \sum_{i=1}^{m} \exp\left[-\left(\frac{z_i - \mu}{\sigma}\right)\right], \tag{1.7}$$

La massimizzazione della log-verosimiglianza definita in equazione (1.6) e (1.7) rispetto ai parametri (μ, σ, ξ) non ha una soluzione analitica ma, dato qualsiasi dataset, la massimizzazione è ottenuta senza particolari complicazioni usando algoritmi standard di ottimizzazione numerica. È necessario comunque assicurarsi che l'algoritmo non si muova verso combinazioni di parametri che violano le condizioni di esistenza della distribuzione

GEV. Infine alcune difficoltà numeriche che potrebbero sorgere dalla valutazione di (1.6) in prossimità di $\xi = 0$ vengono risolti utilizzando direttamente (1.7) per valori di ξ che sono particolarmente vicini a zero.
Una potenziale difficoltà nell'uso della massima verosimiglianza per la GEV riguarda le condizioni di regolarità richieste perchè le usuali proprietà asintotiche associate allo stimatore di massima verosimiglianza siano valide. Queste condizioni non sono soddisfatte dal modello GEV per via degli estremi della distribuzione GEV in quanto questi sono funzione dei parametri. La violazione delle condizioni di regolarità ha come conseguenza la non applicabilità dei risultati asintotici di verosimiglianza standard. Tale problematica è stata affrontata in (Smith, 1985), la cui ricerca ha portato i seguenti risultati:

- quando $\xi > -0.5$, gli stimatori di massima verosimiglianza sono regolari e presentano le usuali prorpietà asintotiche;
- quando $-1 < \xi < -0.5$, gli stimatori sono generalmente ottenibili, ma non presentano le proprietà asintotiche standard;
- quando $\xi < -1$, gli stimatori di massima verosimiglianza sono difficilmente ottenibili.

Il caso $\xi \leq -0.5$ corrisponde a distribuzioni con una coda superiore molto limitata. Questa situazione si incontra raramente applicando la modellazione ai valori estremi, quindi le limitazioni teoriche solitamente non sono un ostacolo nella pratica. La distribuzione approssimata di $(\hat{\mu}, \hat{\sigma}, \hat{\xi})$ è una normale multivariata di media (μ, σ, ξ) e matrice di varianza e covarianza uguale all'inversa della matrice di informazione osservata calcata nella stima di massima verosimiglianza.
Per quanto riguarda i livelli di ritorno citati nella sezione 1.1.1, la loro stima può essere ottenuta utilizzando direttamente i valori stimati dei parametri tramite la massima verosimiglianza. Per la stima degli intervalli di confidenza e della varianza, si ricorre all'utilizzo del metodo delta in combinazione con i principi della teoria classica della verosimiglianza. In ogni caso è necessario essere cauti nell'interpretazione dell'inferenza dei livelli di ritorno.

1.2.2 Inferenza Bayesiana

Un altro metodo per fare inferenza sulla verosimiglianza è l'adozione del paradigma bayesiano. Questo richiede un distribuzione a priori sui parametri (μ, σ, ξ) che rappresenti la conoscenza sui valori dei parametri, prima che i dati siano disponibili. In molti contesti può risultare appropriato utilizzare questo strumento per applicare una conoscenza vera sul processo sotto studio. Molto spesso non essendo a conoscenza di particolari caratteristiche dei parametri, si utilizzano delle distribuzioni a priori con grande varianza, per riflettere la mancanza di conoscenza. Definendo quindi la distribuzione a priori dei parametri con $\Pr(\cdot)$ e la verosimiglianza come in formula (1.5), il teorema di Bayes afferma che

$$p(\mu, \sigma, \xi | Z_1, \ldots, Z_m) \propto L(\mu, \sigma, \xi; Z_1, \ldots, Z_m) \times p(\mu, \sigma, \xi), \tag{1.8}$$

che porta alla distribuzione a posteriori dei parametri, la quale è una modifica della priori, data dall'informazione contenuta nei dati espressi sotto forma di funzione di verosimiglianza. Il risultato di un'analisi bayesiana è un intera distribuzione sull'insieme di parametri, che rappresenta un considerevole vantaggio rispetto al metodo classico: invece di avere una stima puntuale, si ottiene una distribuzione probabilistica completa dei valori dei parametri. Nel caso dovesse essere richiesta una singola stima puntuale solitamente si fornisce una semplice statistica riassuntiva della distribuzione a posteriori, quale la media o la moda. Solitamente in mancanza di particolari conoscenze riguardo i parametri la distribuzione a priori viene scelta impostando $\phi = \log \sigma$, come:

$$p(\mu, \sigma, \xi) = p_\mu(\mu)p_\phi(\phi)p_\xi(\xi), \tag{1.9}$$

dove $p_\mu(\cdot)$, $p_\phi(\cdot)$ e $p_\xi(\cdot)$ sono funzioni di densità della normale con media zero e varianze rispettivamente v_μ, v_ϕ e v_ξ. Il motivo per cui si sceglie di lavorare con ϕ è quello di avere una parametrizzazione facile che permetta di rispettare la positività di σ. La densità a priori (1.9) quindi corrisponde ad una specificazione di indipendenza a priori tra i parametri μ, ϕ e ξ, che può essere resa quasi piatta, e quindi non informativa, scegliendo delle varianze sufficientemente grandi. Una scelta ragionevole è $v_\mu = v_\phi = 10^4$ e $v_\xi = 100$. Questo quindi completa la specificazione del modello.

L'implementazione diretta del teorema di Bayes in generale risulta complicato, e ancor di più per il caso in questione, per via del calcolo della costante di normalizzazione della distribuzione a posteriori. Questa, data (1.8) si ricava calcolando l'integrale:

$$\iiint_\Omega L(\mu, \sigma, \xi; Z_1, \ldots, Z_m) \cdot p(\mu, \sigma, \xi)\, d\mu\, d\sigma\, d\xi,$$

dove Ω rappresenta l'insieme dell'intero spazio parametrico dei parametri (μ, σ, ξ).

Nel caso di modelli complessi come il seguente dove si è interessati a vettori di parametri, questa quantità non è calcolabile analiticamente e può essere molto complessa da calcolare attraverso i metodi numerici classici. In questo contesto, l'impiego di metodi di simulazione come il *Markov Chain Monte Carlo* offre una soluzione efficace per l'approssimazione dell'integrale in questione. Approfondimenti su questa metodologia sono disponibili in Hastings (1970) e Gamerman & Lopes (2006), mentre applicazioni specifiche ai valori estremi sono trattate in Coles (2001). La costruzione delle catene viene usualmente svolta attraverso l'algoritmo di Metropolis-Hastings.

Questo metodo genera un insieme di valori simulati che rappresentano la distribuzione a posteriori dei parametri, consentendo così di ottenere una stima accurata della stessa. Ad esempio, calcolando la media o i quantili del campione ottenuto si può ottenere una stima della media o dei quantili della distribuzione a posteriori. Un vantaggio molto importante che si ottiene utilizzando l'inferenza bayesiana riguarda la previsione. La distribuzione predittiva di un massimo annuale futuro y è:

$$F(y|z) = \int F(y|\mu, \sigma, \xi) f(\mu, \sigma, \xi|Z_1, \ldots, Z_m)\, d\mu\, d\sigma\, d\xi, \tag{1.10}$$

e incorpora l'incertezza sia nel futuro valore di y che nei valori dei parametri stessi. Si ottiene quindi l'analogo del livello di ritorno per il p-esimo anno risolvendo:

$$F(y|z) = 1 - 1/p.$$

La quantità (1.10) potrebbe sembrare di difficile trattazione, tuttavia, essa diventa facilmente approssimabile quando la distribuzione a posteriori viene stimata tramite simulazioni MCMC. Dato il campione $(\mu_1, \sigma_1, \xi_1), \ldots, (\mu_s, \sigma_s, \xi_s)$ di parametri a posteriori ottenuto da tale procedura, è possibile ottenere una stima di (1.10) con:

$$F(y|z) \approx \frac{1}{s} \sum_{i=1}^{s} F(y|\mu_i, \sigma_i, \xi_i).$$

L'analisi bayesiana presenta quindi numerosi vantaggi rispetto a quella basata sulla massima verosimiglianza nel caso dei valori estremi. Essa permette di incorporare ai dati conoscenze pregresse sul processo d'interesse e ottenere una misura di incertezza dovuta alla stima del modello. L'ultimo vantaggio è dovuto al fatto che l'analisi bayesiana non dipende dalle assunzioni di regolarità richieste dalla teoria asintotica della massima verosimiglianza. In particolare, quando $\xi < -0.5$ e la teoria classica della massima varosimiglianza non funziona, l'inferenza bayesiana presenta un valida alternativa.

Capitolo 2

Distribuzioni a Priori di Contrazione

La selezione delle variabili è stata un argomento molto importante nella letteratura statistica degli ultimi decenni, con numerosi articoli pubblicati tra teoria e pratica. Quando si modella una variabile risposta rispetto a delle variabili esplicative, trovare un sottoinsieme di variabili che spieghino meglio una variabile di interesse è un importante aspetto dell'analisi dei dati. Questo consente di ottenere un'interpretazione più chiara, evitare il sovradattamento del modello e la multicollinearità e, infine, di delineare meglio il meccanismo generativo dei dati. La selezione delle variabili è particolarmente importante quando il numero di potenziali predittori è più grande della numerosità campionaria. Nel contesto della regressione lineare, approcci moderni alla selezione delle variabili includono metodi basati sui criteri di informazione, come AIC/BIC, metodi di verosimiglianza penalizzata con riduzione dei coefficienti a zero delle covariate non rilevanti, e approcci bayesiani che usano *a priori* di contrazione per indurre sparsità come misture di due distribuzioni e *a priori* di contrazione continue unimodali.
I metodi bayesiani per la selezione delle variabili hanno diverse caratteristiche interessanti. Questi permettono una modellazione ricca tramite strategie di ricerca stocastica MCMC e forniscono un quadro per integrare diverse fonti di informazione in modo ottimale per ottenere previsioni più accurate; sono estendibili a risposte multivariate e sia in contesti lineari che non lineari; possono gestire lo scenario in cui p è grande e n piccolo; permettono l'uso di distribuzioni a priori per incorporare informazioni accessorie e passate. Il ruolo delle *a priori* di contrazione è quello di concentrare una consistente massa di probabilità a priori vicino allo zero per i parametri di interesse, permettendo che sia solo l'evidenza empirica a determinare quali parametri si distaccano significativamente da tale valore. Tra le molteplici varianti di *a priori* di contrazione, due approcci spiccano per la loro rilevanza e versatilità: la Spike and Slab e l'*a priori* continua.
L'*a priori* Spike and Slab consente di gestire l'incertezza associata ai parametri di un modello combinando due componenti essenziali: Lo "*Spike*" concentrato attorno a zero e la "*Slab*" più diffusa che permette ai parametri di assumere valori lontani da zero se le evidenze dei dati lo supportano. L'*a priori* continua rappresenta un'alternativa che permette

una graduale riduzione dell'importanza dei parametri man mano che si avvicinano a zero. Questo tipo di *a priori* offre una transizione continua da una penalità moderata a una penalità più forte, consentendo ai parametri di assumere una gamma più ampia di valori rispetto alla "Spike and Slab". Di seguito verranno approfondite entrambe le tipologie di *a priori* di contrazione analizzandone meccanismi, caratteristiche e applicazioni nel caso della regressione lineare. Per approfondimenti si veda *Handbook of Bayesian Variable Selection* (Tadesse & Vannucci, 2021) e *Handbook of Statistics Volume 43* (Narisetty, 2020).

2.1 Distribuzioni a priori Spike and Slab

Nel contesto del modello di regressione lineare multipla, una risposta continua y_i, viene modellata attraverso una combinazione lineare di p covariate, $\mathbf{x}_i = (x_1, \ldots, x_p) \in \mathbb{R}^p$:

$$y_i = \alpha + \mathbf{x}_i^T \beta + \epsilon_i, \quad i = 1, \ldots, n, \tag{2.1}$$

con $\epsilon_i \sim N(0, \sigma^2), \beta = (\beta_1, \ldots, \beta_p)^T$ il vettore dei coefficienti di regressione e α l'intercetta. Il problema della selezione delle variabili nasce nel momento in cui è risaputo che non tutte le p covariate sono importanti nello spiegare cambiamenti della risposta e l'identificazione dei predittori importanti risulta essere uno degli obiettivi dell'analisi. Chiaramente, impostare il valore di alcuni coefficienti in (2.1) a zero corrisponde ad escludere dal modello il corrispondente sottoinsieme di predittori. Nel paradigma Bayesiano questo può essere ottenuto imponendo delle misture di priori che inducono sparsità, conosciute come *a priori* Spike and Slab, sui coefficienti β_j. Questa formulazione introduce un vettore latente $\gamma = (\gamma_1, \ldots, \gamma_p)$ di variabili indicatrici latenti:

$$\gamma_j = \begin{cases} 1 & \text{se la variabile } j \text{ è inclusa nel modello,} \\ 0 & \text{altrimenti.} \end{cases}$$

Nella letteratura statistica sono state sviluppate due specificazioni per il seguente caso: La Spike and Slab discreta e continua.

2.1.1 Spike and Slab discreta

La Spike and Slab discreta prevede una mistura di distribuzione come priori sui β_j con una massa puntiforme in zero del tipo:

$$\beta_j | \sigma^2, \gamma_j \sim (1 - \gamma_j)\delta_0(\beta_j) + \gamma_j p_1(\beta_j), \tag{2.2}$$

per $j = 1, \ldots, p$, dove $\delta_0(\cdot)$ è la funzione di Dirac e $p_1(\cdot)$ è una distribuzione a priori diffusa a scelta. In questo caso quindi quando $\gamma_j = 0$, la variabile j-esima viene esclusa, dato che l'*a priori* sul coefficiente corrispondente β_j è una distribuzione a massa puntiforme in zero, mentre quando $\gamma_j = 1$ il predittore viene incluso nel modello attraverso un'*a priori* diffusa. La costruzione dell'*a priori* in (2.2) richiede una specificazione di una distribuzione

a priori di γ. La scelta più semplice e più comune in letteratura è il prodotto di distribuzioni Bernoulli indipendenti con parametro comune ω :

$$p(\gamma|\omega) = \prod_{j=1}^{p} \omega^{\gamma_j}(w-1)^{1-\gamma_j}, \tag{2.3}$$

di conseguenza, il numero attesi di variabili incluse nel modello è pari a $p\omega$. La scelta di ω può essere fatta arbitrariamente per considerazioni a priori sui dati, altrimenti l'incertezza di tale parametro può essere modellata imponendo a sua volta un'*a priori* Beta, $\omega \sim Beta(a, b)$ con a, b scelte in modo appropriato. Un'*a priori* poco informativa può essere ottenuta impostando $a = b = 1$, portando ad un valore atteso a priori pari a $m = a/(a + b) = 0.5$. Scegliendo quindi una distribuzione a priori per α e σ^2 e $p_1(\cdot)$ la specificazione del modello risulta completa. Per semplicità, si assume comunemente l'indipendenza a priori tra i $\beta_j|\gamma$. La scelta della *Slab* $p_1(\cdot)$ è molto varia ma risulta essere molto importante, in quanto determina l'andamento a priori dei parametri che effettivamente entreranno nel modello.

Sono molte le proposte in letteratura, riguardo la scelta della "*Slab*, delle quali si può vedere un'illustrazione in Figura (2.1). Alcune proposte per la distribuzione $p_1(\cdot)$ di β_j quando $\gamma_j = 1$ è la distribuzione gaussiana centrata in zero con varianza $h_j\sigma^2$, la Laplace centrata in zero con parametro di scala $h_j\sigma^2$ e distribuzione Uniforme.

In ognuno dei casi, la distribuzione a posteriori viene calcolata tramite ricerca stocastica MCMC, che permette di esplorare la distribuzione a posteriori ed identificare i modelli con alta probabilità anche nel caso di un elevata numerosità dei predittori. Per effettuare la selezione delle variabili è comunque necessario scegliere una soglia sulla proporzione del parametro γ a posteriori, lavorando sulla probabilità che sia 1 e 0.

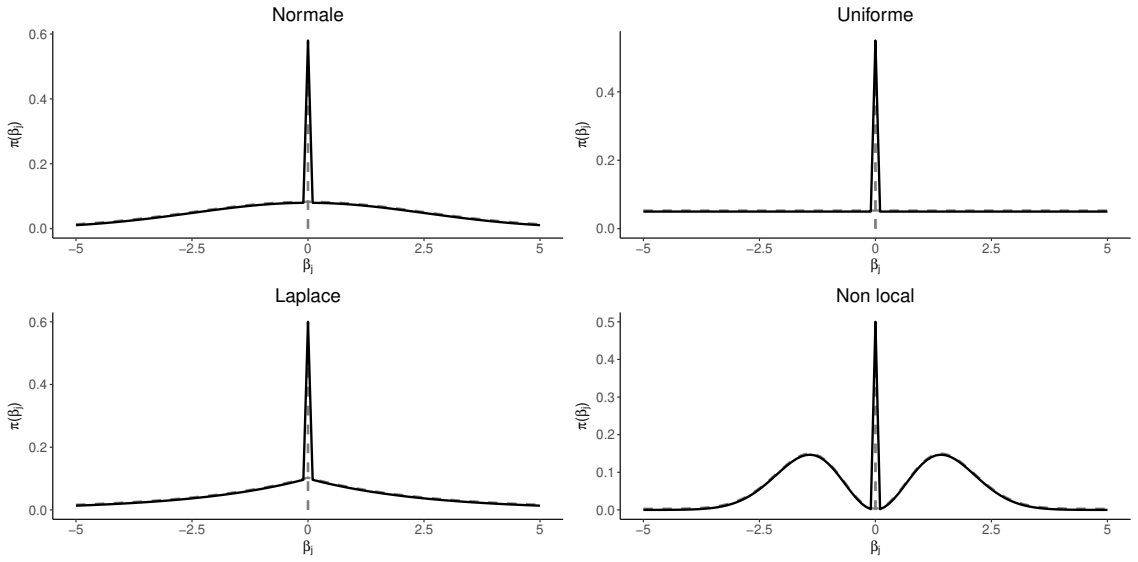

Figura 2.1: Distribuzioni a priori Spike and Slab: I) Gaussiana, II) Uniforme, III) Laplace, IV) Non local.

Il caso Spike and Slab gaussiano comunque risulta essere il più utilizzato per via della semplicità e perché ha la caratteristica di essere coniugato rendendo l'MCMC più semplice. Una proposta interessante per la selezione delle variabili bayesiana è quella di Johnson & Rossell (2012), che suggeriscono l'utilizzo di distribuzioni a priori non locali come *Slab*. La motivazione principale dell'utilizzo di una distribuzione a priori non locale è che la *Slab prior* non dovrebbe avere una massa positiva fissa attorno a zero dal momento che questa dovrebbe rappresentare il segnale con un'intensità diversa da zero. Una *a priori* non locale d'altra parte ha una funzione di densità che tende a zero man mano che il valore del parametro si avvicina a zero.
Due esempi di distribuzioni di questa tipologia sono la *product moment prior* (pMOM) e la *product inverse moment* (piMOM). La pMOM, rappresentata anche in Figura (2.1), è definita come:

$$p(\boldsymbol{\beta}, \tau, \sigma^2) \propto \exp\left\{-\frac{1}{2\tau\sigma^2}\boldsymbol{\beta}^T\boldsymbol{\beta}\right\} \prod_{j=1}^{p} \beta_j^{2r}, \tag{2.4}$$

mentre la piMOM invece è definita come:

$$p(\boldsymbol{\beta}, \tau, \sigma^2) \propto \prod_{j=1}^{p} \beta_j^{-(r+1)} \exp\left\{-\frac{\tau\sigma^2}{\beta_j^2}\right\}, \tag{2.5}$$

con $\tau > 0$ e $r = 1, 2, ...$ come iperparametri, dove r è l'ordine della densità mentre τ rappresenta il parametro di scala che determina la dispersione della densità *a priori* di β. Con queste *a priori* viene mostrato che la distribuzione a posteriori si concentra nel modello reale con probabilità che tende a 1 per $p \leq n$. La distribuzione a posteriori in questo caso non viene calcolata tramite algoritmi standard come Gibbs sampler ma utilizzando un'approssimazione Laplace. Tuttavia, in contesti di alta dimensionalità, il costo computazionale può diventare proibitivo, pertanto sono state sviluppate alternative metodologiche per affrontare questa problematica. Ulteriori approfondimenti e caratteristiche del caso presentato sono presenti in Johnson & Rossell (2012).
La Spike and Slab discreta rappresenta quindi un'ottima soluzione per problemi di selezione delle variabili in contesti ad alta dimensionalità, riscuotendo ampio successo e generando un'elevata quantità di estensioni con diverse tipologie di *Slabs*.

2.1.2 Spike and Slab continua

La costruzione continua della Spike and Slab differisce da quella discreta in equazione (2.2), in quanto assume una mistura di due componenti continue, una concentrata attorno allo zero e l'altra molto più diffusa:

$$\beta_j|\gamma_j, \sigma^2 \sim (1-\gamma_j)p_0(\beta_j) + \gamma_j p_1(\beta_j), \tag{2.6}$$

per $j = 1, \ldots, p$, con $p_0(\cdot)$ funzione di densità concentrata attorno a zero e $p_1(\cdot)$ è la distribuzione a priori diffusa a scelta. La distribuzione a priori $p_0(\cdot)$ pone la maggior parte

della massa di probabilità attorno a zero, definendo il caso in cui sotto l'ipotesi nulla il coefficiente di regressione è trascurabile ma può assumere valori diversi da zero. L'*a priori* $p_1(\cdot)$ invece essendo generalmente una distribuzione piatta e assegnando più massa di probabilità a valori del parametro grandi, corrisponde al segnale. Nel caso della Spike and Slab discreta i metodi per il calcolo della distribuzione a posteriori tendono ad essere computazionalmente onerosi a causa del cambio di dimensione.Ottenendo quindi un modello la cui interpretazione risulta simile, si riesce a diminuire tale carico computazionale.
Solitamente la scelta di $p_0(\cdot)$ e $p_1(\cdot)$ ricade su due distribuzioni gaussiane, la prima concentrata attorno allo zero e la seconda più diffusa. Per la specificazione completa del modello quindi è necessario definire una distribuzione per σ e γ. Come nel caso della Spike and Slab discreta anche in questo contesto solitamente viene scelta una Gamma inversa per σ e una distribuzione Bernoulli per γ come nell'Equazione (2.1.1). Il parametro della Bernoulli ω può essere fissato arbitrariamente oppure trattato come una variabile casuale, la quale segue tipicamente una distribuzione, come la Beta.
Una scelta molto popolare per la Spike and Slab continua è l'*a priori* Spike and Slab LASSO proposta in Ročková & George (2018). Questa pone una mistura di due priori Laplace sui parametri di regressione come segue:

$$\beta_j|\gamma_j \sim (1-\gamma_j)LP(\lambda_0) + \gamma_j LP(\lambda_1), \tag{2.7}$$

dove $LP(\lambda)$ è la distribuzione Laplace con funzione di densità $\psi(\beta|\lambda) = \frac{\lambda}{2}exp(-\lambda|\beta|)$ e $\lambda_0 \gg \lambda_1$. Come si nota in Figura 2.2 la principale differenza tra la versione gaussiana e quella Laplace è data dalle code di quest'ultima che risultano più pesanti.

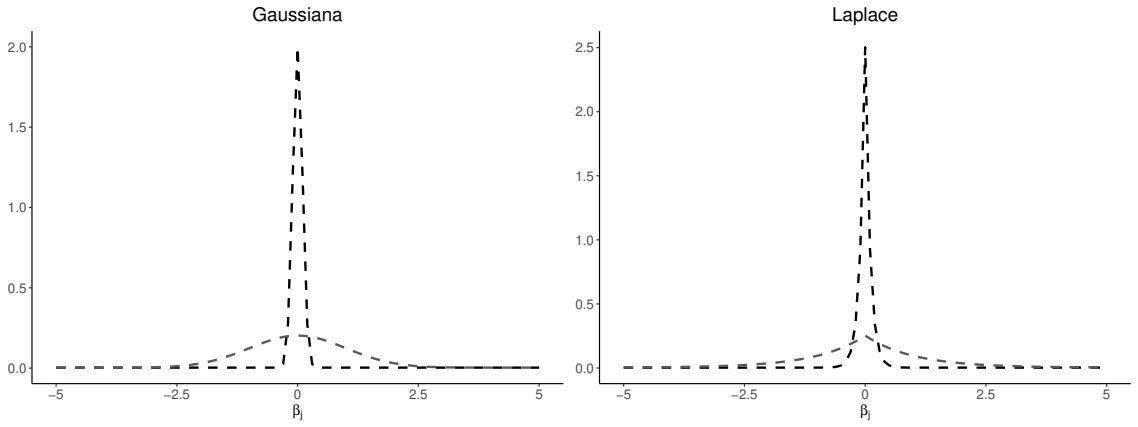

Figura 2.2: Esempi di *a priori* Spike and Slab: Gasussiana e LASSO.

Si nota che scegliendo $\lambda_1 = \lambda_0$ si ottiene la penalizzazione LASSO L_1 mentre con $\lambda_0 \to \infty$ si ottiene un'*a priori* Spike and Slab discreta come caso limite. Una caratteristica dell'*a priori* SSL è la capacità di indurre una transizione continua e non concava tra la verosimiglianza penalizzata e la costruzione della Spike and Slab discreta (con punto di massa a 0). Dato che è una mistura di due distribuzioni Laplace l'*a priori* Spike

ans Slab LASSO può essere vista come un perfezionamento della penalità L_1 del LASSO sui coefficienti. La moda a posteriori così ottenuta con la distribuzione a priori (2.1.2) risulta esattamente sparsa e può essere utilizzata simultaneamente sia per la selezione delle variabili che per la stima dei parametri. Questa proprietà di applicazione di una soglia automatica offre un vantaggio rispetto alle formulazioni della Spike and Slab (2.2). Nel caso discreto infatti non si ottiene la stima sparsa esatta dei coefficienti e tipicamente è richiesto l'utilizzo di una soglia a posteriori per la selezione delle variabili. È risaputo che il LASSO originale soffre di un *bias* di stima, dove coefficienti grandi vengono contratti eccessivamente.
La Spike and Slab LASSO offre quindi il vantaggio di applicare due tipi di contrazione: grande se $|\beta_j|$ risulta piccolo, molto piccolo se $|\beta_j|$ risulta grande. Tale distribuzione a priori quindi propone una modellazione che sta nel mezzo tra il caso Spike and Slab discreto e il caso di utilizzo di *a priori* di contrazione continue, offrendo molta flessibilità nella gestione della sparsità.

2.2 A priori di contrazione continue

Si considera nuovamente il modello di regressione lineare multipla come in Equazione (2.1), nel caso in cui si voglia fare selezione delle variabili, sapendo che non tutte risultano essere importanti nello spiegare la variabile risposta. Le *a priori* di contrazione continue propongono una gestione della sparsità attraverso una contrazione globale dei coefficienti. Queste hanno comunque visto varie modifiche tra le quali le *a priori* di contrazione globale e locale, nella quale sono presenti sia un tipo di contrazione globale, sia locale per i singoli parametri.
Le *a priori* Spike and Slab in caso di sparsità hanno un grande attrattiva interpretativa quindi è importante definire le motivazioni che portano a scegliere un'*a priori* continua. Innanzitutto l'utilizzo della Spike and Slab può far incorrere in problemi computazionali, con complessità proibitive. Una seconda motivazione per l'utilizzo di *a priori* continue è che a volte possono abbinarsi a determinate tipologie di sparsità, dove i parametri sono molto vicini allo zero ma non sono esattamente uguali a zero. Inoltre queste metodologie possono anche portare a risultati molto simili a quelli ottenuti con *a priori* Spike and Slab. I due approcci sono entrambi molto validi in caso di problemi di sparsità, portando quindi a considerare le *a priori* continue come un valido strumento per la stima nei modelli lineari. Di seguito verranno proposti un tipo di *a priori* continua ed una locale e globale.

2.2.1 Laplace prior

La prima tipologia di *a priori* di contrazione continua che viene presentata è la distribuzione Laplace. Nell'ambito della regressione lineare un modello bayesiano che utilizza la seguente prior prende il nome di LASSO bayesiano (Park & Casella, 2008). Questo pone delle priori

Laplace indipendenti sulle componenti di β:

$$p(\beta|\sigma,\lambda) \propto \prod_{j=1}^{p} \frac{\lambda}{2\sigma} \exp\left\{-\frac{\lambda|\beta_j|}{\sigma}\right\},$$

con λ iperparametro, che definisce la quantità di compressione verso lo zero indotto dall'*a priori*. Il nome del modello deriva dal fatto che, come suggerito da Tibshirani (1996), le stime del modello LASSO possono essere interpretate come stima della moda a posteriori quando i parametri di regressione hanno delle Laplace indipendenti ed identicamente distribuite come priori, proprio come il caso in questione.
Per ottenere i risultati e le stime d'interesse solitamente viene implementato un pratico Gibbs sampler utilizzando una rappresentazione gerarchica del modello. La scelta del parametro λ, che nel caso del LASSO ordinario può venire scelto attraverso la convalida incrociata generalizzata e procedimenti di stima e verifica, risulta essere molto rilevante. Il LASSO Bayesiano offre alcune alternative uniche: *empirical Bayes*, attraverso massima verosimiglianza marginale oppure l'uso di una *a priori* diffusa sull'iperparametro. La seconda alternativa vede proposta una classe dell'*a priori* gamma su λ^2 (e non direttamente su λ) nella forma:

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

questa scelta favorisce la coniugatezza e facilita l'applicazione dell'algoritmo di Gibbs sampling per l'estrazione delle variabili. Un'altra possibilità potrebbe essere l'*a priori* impropria invariante in scala $1/\lambda^2$ per λ^2, anche se non viene scelta solitamente in quanto conduce ad una posteriori impropria.

2.2.2 Horseshoe prior

Un'*a priori* diventata sempre più popolare negli ultimi anni per la sua efficacia nella gestione della sparsità è la Horseshoe prior presentata da Carvalho et al. (2009, 2010). L'*a priori* Horseshoe assume che ogni β_j sia condizionatamente indipendente con densità $p_{HS}(\beta_j|\tau)$ dove p_{HS} può essere rappresentato come una mistura di normali:

$$\begin{aligned} (\beta_j|\lambda_j,\tau) &\sim N(0,\lambda_j^2\tau^2), \\ \lambda_j &\sim C^+(0,1), \end{aligned} \tag{2.8}$$

dove $C^+(0,1)$ è la distribuzione half-Cauchy per la deviazione standard λ_j. Nella specificazione del modello sono presenti due parametri: λ_j è il parametro di contrazione locale e τ è il parametro di contrazione globale. Questi parametri permetto di controllare sia la variabilità complessiva dei parametri sia di ciascuno individualmente.
L'*a priori* Horseshoe ha due caratteristiche interessanti che la rendono particolarmente utile come *a priori* di contrazione. Le sue code piatte simili a quelle della distribuzione Cauchy, consentono ai segnali forti di rimanere grandi (quindi non contratti a zero) a

posteriori. Tuttavia il picco alto, che va ad infinito all'origine, fornisce una contrazione significativa per gli elementi di β che sono a zero. Questi elementi rendono la Horseshoe una scelta attraente per gestire vettori sparsi. La densità di β_j in (2.8) è perfettamente definita senza fare riferimento ai λ_j, che possono essere marginalizzati. La stima di β avviene stimando la media a posteriori del modello specificato. Il nome Horseshoe ("Ferro di cavallo") deriva da una particolare caratteristica della distribuzione a priori. Fissando infatti $\sigma^2 = \tau^2 = 1$, si osserva infatti che,

$$E(\beta_j|y) = \int_0^1 (1-k_j) y_j p(k_j|y) dk_j = \{1 - E(k_j|y)\}\, y_j,$$

dove $k_j = 1/(1+\lambda_j^2)$ e $E(k_j|y)$ viene interpretato come l'ammontare di contrazione verso zero, a posteriori. La scelta di definire una Half-Cauchy come priori di λ_j, implica una distribuzione a forma di ferro di cavallo, $Be(1/2, 1/2)$ per il coefficiente di contrazione k_j. La parte destra del ferro di cavallo, $k \approx 0$, non produce alcun tipo di contrazione, descrivendo il segnale. La parte sinistra invece, $k \approx 1$, comporta quasi un totale contrazione a zero evidenziando la mancanza di segnale. L'*a priori* Horseshoe quindi prevede due scenari a priori: la presenza di un forte segnale o la totale mancanza di segnale.
Il coefficiente $k_j \in [0, 1]$ cambia distribuzione a priori a seconda del tipo di distribuzione viene assegnata a λ. Il comportamento di k_j permette di comprendere come ciascun modello cerchi di distinguere tra segnale e rumore. Il grafico in Figura 2.3 permette di osservare l'andamento nel caso dell'*a priori* Horseshoe e la sua peculiare forma a ferro di cavallo.

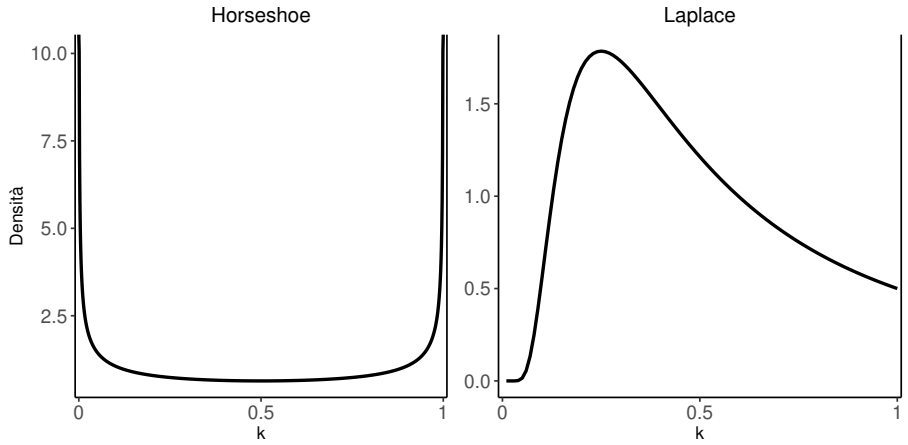

Figura 2.3: Distribuzione a priori implicita $p(k)$ a meno di costanti moltiplicative indotta dall'*a priori* : I) Horseshoe; II) Laplace.

Si osserva quindi il caso in cui $\lambda_j^2 \sim Exp(1/2)$, che comporta delle priori Laplace per ciascun β_j. Questo risulta essere simile al caso precedentemente specificato ma nel quale ad ogni β_j è associato un λ_j diverso. In questo caso la distribuzione di k_j, come si può notare in figura 2.3, risulta molto differente; infatti l' *a priori* Laplace tende ad una costante vicino a $k = 1$, e scompare completamente vicino a $k = 0$.

Come ultima caratteristica, la horseshoe prior, ha il vantaggio di essere libera dalla scelta di iperparametri in quanto risulta essere completamente specificata. Essa risulta quindi un'*a priori* robusta e molto adattabile, portando ottime performance in un ampia carietà di situazioni.

Capitolo 3

A priori di contrazione per la distribuzione GEV

Nei capitoli precedenti è stata descritta la distribuzione Generalizza dei Valori Estremi e il suo utilizzo nel caso di un'analisi sui massimi annuali. Inoltre è stata fatta una panoramica sulle varie tecniche utilizzate nell'analisi bayesiana per selezionare le variabili nel caso della regressione lineare. L'utilizzo di distribuzioni a priori di contrazione essenzialmente serve a verificare l'ipotesi che i parametri di interesse siano o meno significativamente diversi da zero.Il parametro d'interesse sul quale, nella seguente tesi, si vuole lavorare con distribuzioni di contrazione nel caso della distribuzione GEV è il parametro di forma ξ.
Come osservato in precedenza infatti il valore ξ è molto importante in quanto determina profondamente l'andamento della distribuzione, sopratutto nelle code di queste ultime. Nel caso in cui ξ sia uguale a zero si ottiene il caso limite della distribuzione GEV che corrisponde alla distribuzione Gumbel mentre in caso di valori positivi o negativi si ottiene rispettivamente la distribuzione Frèchet e Weibull. La distribuzione Gumbel in particolare viene spesso utilizzata, anche più della famiglia completa GEV, per i massimi annuali, sopratutto in idrologia e nello studio delle precipitazioni. In alcuni casi tale decisione è giustificata dall'evidenza empirica, mentre in altri casi dal fatto che, molte distribuzioni note per una generica variabile X_i portano ad una distribuzione limite Gumbel per $\max\{X_1, \ldots, X_n\}$. Questa strategia può risultare molto rischiosa in molti casi.
Nella pratica analizzando un campione di massimi annuali, per fare inferenza sui modelli dei valori estremi, sono state sviluppate due strategie volte a tenere conto del comportamento dei valori in considerazione.
L'approccio più comune prevede di adattare direttamente la distribuzione GEV ai dati come descritto in Coles (2001),Davison & Smith (1990), Stephenson & Tawn (2004). L'altro approccio prevede di selezionare il tipo di valore estremo, scegliendo tra i tre modelli ($\xi < 0, \xi = 0, \xi > 0$) attraverso un processo di verifica d'ipotesi, ed infine di ottenere il miglior adattamento del modello selezionato (si veda Hosking (1984),Gomes (1987)). Il primo metodo è il più utilizzato ma dato che il tipo Gumbel viene ridotto ad un singolo

valore sullo spazio parametrico, il modello Gumbel non viene mai selezionato. Al contrario, il secondo metodo soffre di due svantaggi: il test d'ipotesi a priori assume che il modello vero sia il Gumbel ($\xi = 0$); ignora l'incertezza intrinseca nella scelta del tipo di modello sull'inferenza successiva per ξ e la distribuzione GEV $G(z)$ definita in Equazione (1.2).
Una difesa dell'approccio che adatta direttamente la distribuzione GEV ai dati, ignorando il tipo Gumbel, è che quest'ultimo è un modello limite per il quale la penultima approssimazione è la distribuzione GEV con un valori di ξ molto piccolo ma diverso da zero. Questo argomento è valido per l'adattamento della GEV ai dati ma non per l'estrapolazione (solitamente molto importante in un'analisi dei valori estremi). Più si procede nell'estrapolazione infatti migliore è l'approssimazione data dal limite di Gumbel (Smith, 1987). Oltretutto in vari casi studio, trattando determinati dati come quelli legati ad eventi idrologici, non è raro che gli esperti assumano che con buona probabilità il modello alla base sia il Gumbel. Nella pratica è necessario quindi un modello che riconosca la possibilità del tipo Gumbel all'interno della più ampia famiglia GEV.
Per fare questo si sceglie di utilizzare un approccio bayesiano, perché oltre ai vantaggi descritti nel Capitolo 1, permette in modo molto semplice di incorporare una conoscenza sulla struttura dei dati alla modellazione. Nel seguente caso si vuole incorporare al modello un conoscenza riguardo la tipologia di distribuzione d'interesse, ed in particolare che il modello possa probabilmente essere Gumbel. Nell'adattamento ai dati del modello non ci si restringe però al caso $\xi = 0$ ma si vuole lasciare ai dati la possibilità di rifiutare il caso della distribuzione Gumbel e seguire le altre due tipologie. Viene proposto quindi l'utilizzo di alcune delle distribuzioni a priori di contrazione discusse nel capitolo precedente.

3.1 Specificazione del modello

Dato il campione di massimi annuali $\mathbf{Z} = Z_1, \ldots, Z_m$, si definisce la verosimiglianza utilizzando la distribuzione generalizzata dei valori estremi:

$$L(\mu, \sigma, \xi; Z) = \prod_{i=1}^{m} g(z_i|\mu, \sigma, \xi), \tag{3.1}$$

con $g(\cdot)$ densità della GEV definita come in (1.5) e con $g(z_i|\mu, \sigma, \xi = 0)$ distribuzione Gumbel. L'analisi bayesiana prevede lo studio della distribuzione a posteriori, definita come:

$$p(\mu, \sigma, \xi|Z) \propto L(\mu, \sigma, \xi; Z) \times p(\mu, \sigma, \xi), \tag{3.2}$$

dove è necessario definire la distribuzione a priori $p(\mu, \sigma, \xi)$, nella quale è possibile incorporare ai dati conoscenze pregresse sui dati. Si sceglie di proporre una distribuzione a priori che assume l'indipendenza dei tre parametri a priori.

Le distribuzioni marginali dei parametri sono:

$$\mu \sim N(0, v_\mu^2), \tag{3.3}$$
$$\phi \sim N(0, v_\phi^2), \tag{3.4}$$

con $\phi = \log(\sigma)$ e $v_\mu = v_\phi = 10^2$. Tale proposta segue quella fatta in (Coles, 2001), nel caso non ci fossero conoscenze particolari a priori riguardo i parametri. Nel modello proposto infatti le distribuzioni a priori vengono tenute non informative, consentendo alle stime di seguire liberamente i dati. L'*a priori* informativa desiderata riguarda il parametro ξ, che sarà selezionata da un gruppo di *a priori* di contrazione, sia discrete sia continue.

3.2 A priori di contrazione per ξ

Vengono di seguite riproposte 5 tipologie di distribuzioni a priori di contrazione adattandole al contesto GEV, in modo specifico per il parametro ξ, tra le quali ci sono distribuzioni a priori continue, Spike and Slab discrete e continue. Risulta necessario però effettuare delle modifiche alle *a priori* nel Capitolo 2 per adattarle al seguente caso, nel quale non si vuole valutare l'uguaglianza o meno a zero degli elementi di un vettore di parametri bensì di un singolo parametro scalare ξ.
Di segui vengono mostrate le distribuzioni a priori modificate per l'utilizzo nell'analisi bayesiana della distribuzione generalizzata dei valori estremi:

- La *a priori* **Laplace** in questo caso assume la seguente forma:

$$p(\xi|\lambda) \propto \frac{\lambda}{2} \exp\{-\lambda|\xi|\}, \tag{3.5}$$

  dove λ è un iperparametro che definisce la quantità di contrazione indotta dalla a priori che viene lasciato variare a seconda dei dati, imponendo una ulteriore distribuzione a priori Gamma(α,β) per λ^2;

- La distribuzione a priori **Spike and Slab discreta** introduce la variabile latente binaria γ che definisce il caso Gumbel ($\xi = 0$) se presenta valori pari a 0, e la generica GEV se assume valori pari a 1. La distribuzione a priori per il parametro ξ è quindi una mistura di distribuzioni con una massa puntiforme in zero:

$$\xi|v_\xi^2, \gamma \sim (1-\gamma)\delta_0(\xi) + \gamma p_1(\xi), \tag{3.6}$$
$$\gamma|\omega \sim \text{Bernoulli}(\omega), \tag{3.7}$$

  con $\delta_0(\cdot)$ funzione di Dirac a $\xi = 0$ e $p_1(\cdot)$ funzione a priori diffusa a scelta per $\xi|\gamma = 1$. La definizione di ω può dipendere da scelte che vengono fatto arbitrariamente date alcune conoscenze pregresse sui dati. Se non si vuole scegliere una probabilità a priori si può decidere di modellare l'incertezza di tale parametro assegnandogli una distribuzione a priori: $\omega \sim Beta(a, b)$. Per la specificazione completa del modello

è quindi necessario definire la *Slab* $p_1(\cdot)$, ed in questo caso vengono scelte due delle distribuzioni descritte nel Capitolo 2.
Innanzitutto si propone la distribuzione normale $\xi|\gamma = 1 \sim N(0, v_\xi)$, considerando il suo successo nella selezione delle variabili in ambito bayesiano. Infine si propone la distribuzione non locale pMOM definita come $p_1(\xi|\tau, v_\xi^2) \propto \exp\left\{-\xi^2/(2\tau v_\xi^2)\right\} \xi^{2r}$ la quale non presenta una massa centrata in zero bensì una moda positiva ed una negativa. In entrambi i casi v_ξ viene posto pari a 2 per ottenere una variabilità che permetta di considerare tutti i valori plausibili nello spazio parametrico di ξ.

- La distribuzione a priori **Spike and Slab LASSO**, ovvero la versione continua della Spike and Slab definisce una variabile latente γ che permette di definire la *a priori* come mistura su ξ di due componenti continue, una concentrata attorno allo zero e l'altra molto più diffusa. In questo caso le distribuzioni sono Laplace:

$$\xi|\gamma \sim (1-\gamma)LP_\xi(\lambda_0) + \gamma LP_\xi(\lambda_1), \tag{3.8}$$

$$\gamma|\omega \sim \text{Bernoulli}(\omega), \tag{3.9}$$

con $\lambda_0 \gg \lambda_1$, iperparametri che definiscono la quantità di contrazione indotta dalle due componenti mistura (in questo caso si propone $\lambda_0 = 20$ e $\lambda_1 = 1/2$). Come nel caso discreto, si sceglie una distribuzione a priori Bernoulli per la variabile γ di probabilità ω, la quale a sua volta si assume abbia una distribuzione non informativa $Beta(a, b)$, con $a = b = 1$.

- Infine la *a priori* **Horseshoe** nel caso di utilizzo nell'ambito della distribuzione GEV è definita:

$$\begin{aligned} (\xi|\lambda, \tau) &\sim N(0, \lambda^2\tau^2), \\ \lambda &\sim C^+(0, 1), \end{aligned} \tag{3.10}$$

dove $C^+(0,1)$ è la distribuzione Half-Cauchy per la deviazione standard λ della gaussiana. In questo caso non si parla di parametro di contrazione globale e locale, in quanto l'unico parametro sul quale la a priori agisce è ξ. La scelta di τ rimane comunque importante, e viene fatta seguendo il criterio ottimale nel caso della regressione lineare, come descritto in Van der Pas et al. (2017) che pone $\tau = p_n/n\sqrt{\log(n/p_n)}$, con p_n, numero di parametri che sono diversi da zero, che in questo caso viene posto a 1. Questa scelta, pur non essendo verificata come ottimale anche in caso di valori estremi, viene fatta in quanto permette di ottenere una varianza della a priori che dipenda dalla numerosità del campione, permettendo una maggior contrazione con campioni molto numerosi.
La distribuzione così definita ha le caratteristiche della sua versione descritta in precedenza ma non permette di fare le considerazioni sull'ammontare di contrazione verso zero del parametro d'interesse, come fatto in Sezione 2.2.2 nel caso del modello

di regressione lineare. I risultati sul valore atteso di $\xi|z$ non sono ottenibili analiticamente come in quel caso, quindi non è possibile verificare direttamente l'effetto della contrazione $k = 1/(1+\lambda^2)$ su di esso. La distribuzione a forma di ferro di cavallo di $Beta(1/2, 1/2)$ di k (Figura 2.3) rimane comunque un comportamento interessante e utile per il caso d'interesse.

In Figura 3.1 vengono quindi illustrate tutte le distribuzioni appena descritte, messe a confronto con la classica *a priori* non informativa normale. Risulta quindi possibile anche graficamente osservare le differenze tra le varie proposte delle distribuzioni a priori per il parametro ξ, così da avere una chiara visione dei modelli descritti precedentemente.

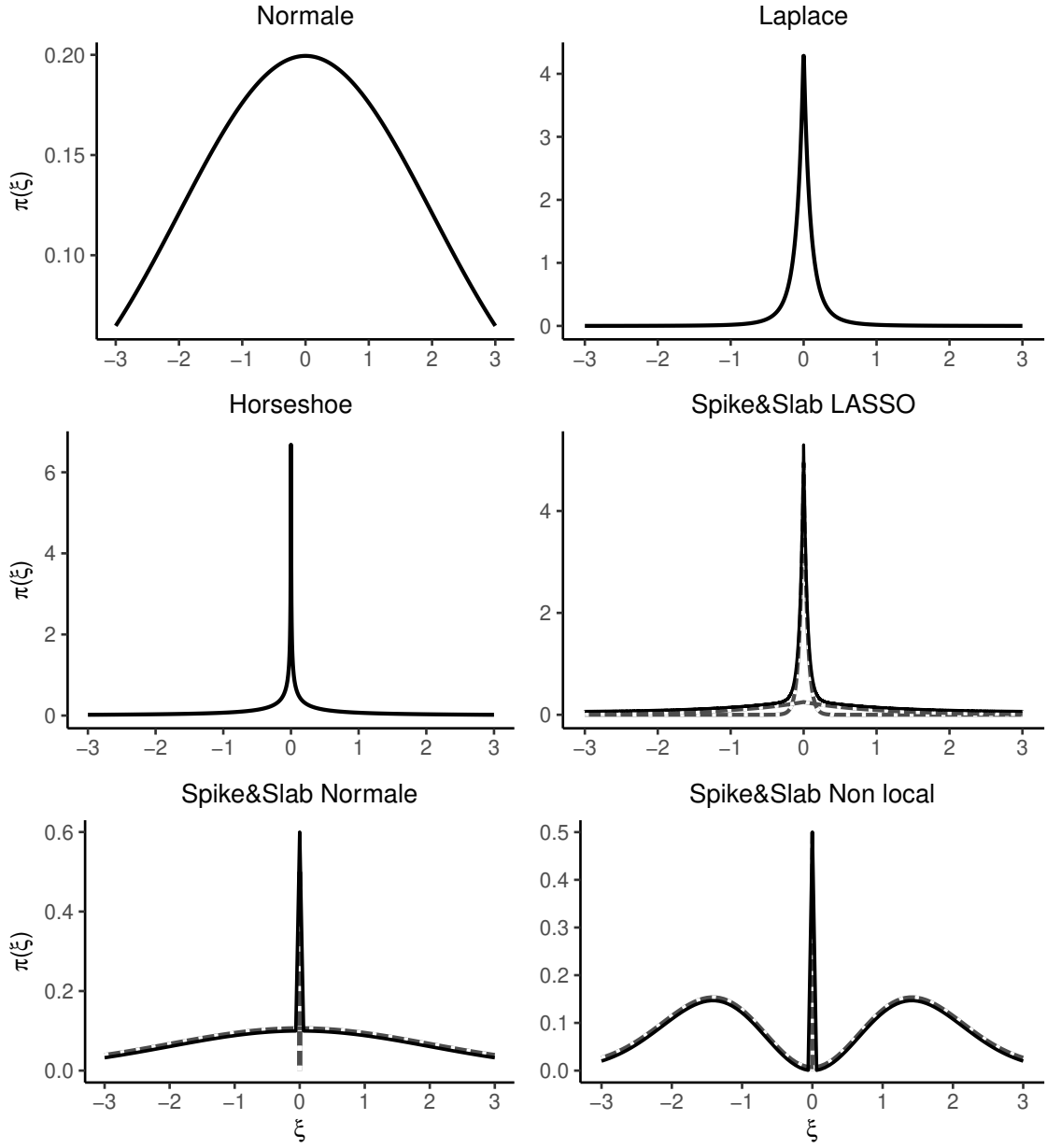

Figura 3.1: Grafici delle distribuzioni a priori di contrazione proposte per il parametro ξ rispetto alla a priori non informativa Normale.

3.3 Stima a posteriori

La densità a posteriori di (μ, σ, ξ) come è ben noto, non può essere calcolata direttamente per via del calcolo dell'integrale che permette di ottenere la costante di normalizzazione. Questo infatti in modelli complessi come il seguente dove si è interessati a vettori di parametri, risulta complesso da calcolare tramite integrazione numerica. Si utilizza il metodo di simulazione *Markov Chain Monte Carlo* che permette di approssimare la distribuzione a posteriori dei parametri attraverso una catena markoviana che abbia $p(\mu, \sigma, \xi|Z)$ come distribuzione limite e ne permette la facile simulazione.
Per generare quindi una sequenza di valori $(\mu_1, \sigma_1, \xi_1), (\mu_2, \sigma_2, \xi_2), \ldots$ indipendenti ed identicamente distribuiti si utilizza l'algoritmo Metropolis-Hastings. Di seguito vengono descritte le due metodologie utilizzate per la generazione delle catene che considerano le differenti strutture dei modelli.

3.3.1 MCMC con a priori continue e Spike and Slab LASSO

Per i modelli con a priori **Laplace**, **Horseshoe** e **Spike and Slab LASSO** si usa il classico algoritmo Metropolis-Hastings con passeggiata casuale su ciascun parametro. Per riassumere i tre casi si definisce $\boldsymbol{\theta} = (\mu, \sigma, \xi, \phi)$ un vettore $1 \times d$ dei parametri e ϕ l'iperparametro che varia per ciascun modello:

$$\phi = \begin{cases} \lambda & \text{per la Laplace e la Horseshoe;} \\ \omega & \text{per la Spike and Slab LASSO.} \end{cases} \tag{3.11}$$

L'algoritmo prevede di ottenere i risultati attraverso un numero di iterazioni elevato arbitrario M. Si inizia con un vettore iniziale della catena $\boldsymbol{\theta}_0$ e ad ogni iterazione, per $i = 1, \ldots, M$ si pone $\boldsymbol{\theta}_i = \boldsymbol{\theta}_{i-1}$ e si eseguono i seguenti passaggi per ogni elemento del vettore dei parametri $\theta_{i,j}$ con $j = 1, \ldots, d$:

- Si simula $\theta_j^*|\boldsymbol{\theta}_i \sim \mathcal{N}(\theta_{i,j}, \epsilon_j^2)$, dove ϵ_j è una costante specifica;
- Si definisce $\boldsymbol{\theta}^* = \boldsymbol{\theta}_i$ con $\theta_{i,j}$ sostituito da θ_j^*;
- Si calcola $\alpha_j = \min\left(1, \frac{p(\boldsymbol{\theta}^*|z)}{p(\boldsymbol{\theta}_i|z)}\right)$;
- Con probabilità α_j si pone $\theta_{i,j} = \theta_j^*$, altrimenti si lascia $\theta_{i,j} = \theta_{i-1,j}$.

Si salva quindi il vettore $\boldsymbol{\theta}_i$, si incrementa i di 1 e si ricomincia con un'altra iterazione. Le costanti ϵ_j indicano la variabilità delle proposte della passeggiata casuale per ciascuna variabile e vanno poste con valori arbitrari tali per cui l'algoritmo possa esplorare interamente lo spazio campionario, ottenendo un tasso di accettazione che varia tra il 25% ed il 50%. Tali valori vanno modificati per ogni campione e modello differente, ogni volta che si stima il modello.

Il seguente algoritmo permette di ottenere quindi un campione rappresentativo della distribuzione a posteriori d'interesse, con la quale è possibile lavorare per ottenere i risultati e le stime richieste. È importante ricordare che prima di essere utilizzato il campione necessita di una pulizia, per via della sua costruzione, tramite una catena markoviana, verificandone la convergenza e ripulendolo dai primi valori generati (*burnin*), così che i valori siano indipendenti ed identicamente distribuiti.

3.3.2 MCMC con a priori Spike and Slab discreta

Per quanto riguarda l'algoritmo per le due **Spike and Slab discrete** proposte è necessario definire alcuni passaggi aggiuntivi. È necessario quindi dividere i parametri d'interesse in due gruppi, $\boldsymbol{\psi} = (\mu, \sigma, \omega)$, γ e ξ con i parametri definiti in (3.7) e (3.6). Si inizia con un vettore iniziale $\boldsymbol{\psi}_0$ e dei valori iniziali ξ_0, $\gamma_0 = 1$. Ad ogni iterazione per $i = 1, \ldots, M$ si pone $\boldsymbol{\psi}_i = \boldsymbol{\psi}_{i-1}$, $\xi_i = \xi_{i-1}$ e $\gamma_i = \gamma_{i-1}$, e si eseguono i seguenti passaggi per ogni elemento del vettore dei parametri $\psi_{i,j}$ con $j = 1, \ldots, d$:

- Si simula $(\psi_j^* | \boldsymbol{\psi}_i, \xi_i, \gamma_i) \sim \mathcal{N}(\psi_{i,j}, \epsilon_j)$, dove ϵ_j è una costante specifica, come definito in precedenza;
- Si definisce $\boldsymbol{\psi}^* = \boldsymbol{\psi}_i$ con $\psi_{i,j}$ sostituito da ψ_j^*;
- Si calcola $\alpha_j = \min\left(1, \frac{p(\boldsymbol{\psi}^*, \xi_i, \gamma_i | z)}{p(\boldsymbol{\psi}_i, \xi_i, \gamma_i | z)}\right)$;
- Con probabilità α_j si pone $\psi_{i,j} = \psi_j^*$, altrimenti si lascia $\psi_{i,j} = \psi_{i-1,j}$.

Se per l'aggiornamento dei parametri μ, σ, ω il procedimento è il medesimo del precedente, così non è per ξ e ω. Prima di procedere con l'aggiornamento di ξ e γ, è necessario evidenziare la dipendenza di ξ dal parametro γ. Quando $\gamma = 0$ infatti il valore di ξ è anch'esso zero, mentre se $\gamma = 1$ allora ξ può variare. Si definisce quindi un metodo per generare γ e di conseguenza un metodo per proporre ξ.
Essendo una variabile discreta γ non può essere proposta attraverso passeggiata casuale gaussiana ma è necessario definirne la *full-conditional*:

$$p(\gamma | \mu, \sigma, \xi, \omega; Z) \propto L(\mu, \sigma, \xi; Z) \left[(1-\gamma)\delta_0(\xi) + \gamma p_1(\xi)\right] \omega^{\gamma} (1-\omega)^{(1-\gamma)}. \tag{3.12}$$

Questa è una distribuzione di Bernoulli i cui parametri sono calcolabili nel seguente modo:

$$\begin{aligned} p(\gamma = 1 | \mu, \sigma, \xi, \omega; Z) &\propto L(\mu, \sigma, \xi; Z)\, \omega\, p_1(\xi); \\ p(\gamma = 0 | \mu, \sigma, \xi, \omega; Z) &\propto L(\mu, \sigma, \xi; Z)\, (1-\omega) \delta_0(\xi). \end{aligned}$$

Il problema che si riscontra però condizionandosi anche a ξ è che si ottengono delle catene che non convergono. Nel caso in cui il valore di ξ precedente sia diverso da zero, la probabilità che $\gamma | \mu, \sigma, \xi, \omega, Z$ sia zero è nulla, per via della presenza della Dirac. Per risolvere tale problematica risulta quindi conveniente condizionarsi solo ai dati e agli altri

parametri d'interesse per la generazione di γ, generando dunque da $p(\gamma|\mu,\sigma,\omega)$ che è anch'essa una Bernoulli:

$$p(\gamma|\mu,\sigma,\omega,Z) = \int_{-\infty}^{+\infty} p(\gamma|\mu,\sigma,\xi,\omega;Z)\,d\xi.$$

Semplificando quindi per ciascun caso si ottiene:

$$p(\gamma=1|\mu,\sigma,\omega,Z) \propto \int L(\mu,\sigma,\xi;Z)\omega p_1(\xi)\,d\xi = P_1, \tag{3.13}$$

$$p(\gamma=0|\mu,\sigma,\omega,Z) \propto L(\mu,\sigma,\xi=0;Z)(1-\omega)\delta_0(\xi=0) = P_0. \tag{3.14}$$

Per ottenere la distribuzione esatta della a posteriori è necessario normalizzare le due quantità attraverso (P_1+P_0). Si ottiene quindi che $\xi|\mu,\sigma,\gamma,Z \sim Bern(P_1/(P_1+P_0))$ Dopo aver definito una proposta per γ si definisce la *full-conditional* per ξ:

$$p(\xi|\gamma=1,\mu,\sigma,Z) = \mathbb{1}\{\xi=0\},$$
$$p(\xi|\gamma=0,\mu,\sigma,Z) \propto L(\mu,\sigma,\xi;Z)\,\omega\,p_1(\xi).$$

Come già sottolineato, quando $\gamma=0$, il parametro ξ è vincolato ad essere uno, altrimenti a posteriori a una distribuzione continua. Date queste osservazioni quindi è possibile definire gli ultimi passaggi della generazione delle catene.
Dopo aver generato $\boldsymbol{\psi}_i=(\mu_i,\sigma_i,\omega_i)$ si prosegue generando $(\gamma^*|\boldsymbol{\psi}_i,\xi_i,Z) \sim Bern(P_{1,i}/(P_{1,i}+P_{0,i}))$ con $P_{1,i}$ e $P_{0,i}$ calcolati come in (3.13) tramite integrazione numerica e come (3.14).

- Se $\gamma^*=0$ allora $\xi^*=0$ e l'iterazione finisce salvando i parametri $(\boldsymbol{\psi}_i,\ \xi_i=0,\gamma_i=0)$
- Se $\gamma^*=1$ allora si propone $\xi^* \sim N(\xi_i,\epsilon_\xi)$ con ϵ_ξ scelto. Si calcola quindi:

  $$\alpha_\xi = \min\left(1, \frac{p(\boldsymbol{\psi}_i,\xi^*,\gamma^*|z)}{p(\boldsymbol{\psi}_i,\xi_i,\gamma_i|z)}\right).$$

  Con probabilità α_ξ si pongono $\xi_i=1$ e $\gamma_i=\gamma^*$, altrimenti si mantengono ξ_i e γ_i invariati rispetto all'iterazione precedente.
  I parametri risultanti quindi sono $(\boldsymbol{\psi}_i,\ \xi_i=\xi^*,\gamma_i=1)$

Terminati i precedenti passaggi si pone $i=i+1$ e si riparte con una nuova iterazione dell'algoritmo. Il calcolo numerico dell'integrale ad ogni iterazione rende questo algoritmo decisamente più oneroso a livello computazionale ma non essendo disponibile una versione analitica di tale quantità risulta l'unica strada percorribile utilizzando MCMC. Come per il caso precedente, dopo una pulizia iniziale del campione ottenuto si può procedere con il suo utilizzo per scopi inferenziali.

3.4 Identificazione del modello Gumbel

Dopo aver ottenuto un campione MCMC della distribuzione a posteriori dei modelli proposti è necessario definire comunque una metodologia per identificare la distribuzione che

definisce il comportamento dei dati. Solitamente analizzando i dati attraverso la distribuzione Gumbel il parametro di forma viene fissato a 0 ma in questo caso nonostante l'utilizzo di metodologie di contrazione, solitamente anche se la distribuzione generatrice dei dati è Gumbel, il parametro ξ risulta comunque una quantità stocastica che può assumere valori differenti da zero. Risulta necessario quindi definire una metodologia che permetta di identificare se la distribuzione a posteriori per ξ sia abbastanza vicina allo zero tale da poter affermare che la distribuzione ottenuta è Gumbel.
Nel caso delle a priori Spike and Slab è possibile estrarre tale informazione andando a vedere la distribuzione a posteriori del parametro $\gamma|Z$, calcolando la proporzione nel campione MCMC di valori di tale parametro pari a 0.
Quando si utilizza una *a priori* di contrazione continua, è necessario invece stabilire una soglia prossima allo zero e determinare la probabilità che il valore assoluto di $\xi|Z$ sia inferiore a questa soglia. Nel contesto di un campione MCMC, tale probabilità si calcola come la frazione dei valori del campione che sono inferiori in valore assoluto alla soglia prefissata. Definita quindi la probabilità che il parametro di forma sia pari a 0, si definisce una soglia che permetta di passare dalla probabilità alla discriminazione del tipo di modello: Gumbel o GEV generico.

3.5 Modello per un insieme di campioni dipendenti

I metodi proposti sono utilizzabili nel momento in cui i dati a disposizione riguardano rilevazioni che verosimilmente hanno caratteristiche simili tra loro, dovute all'appartenenza ad un singolo soggetto, oppure ad una singola stazione di rilevazione, nel caso di eventi atmosferici. In ambito idrologico, come in altri studi scientifici ambientali, lo studio di eventi estremi è fondamentale per prevenire catastrofi naturali e spesso gli esperti del settore assumono che tali valori si distribuiscano come una Gumbel. I modelli proposti quindi possono risultare particolarmente utili in ambiti come il seguente, in quanto, le *a priori* di contrazione permettono di modellare i dati mantenendo l'informazione a priori data dagli studi del settore ma permettendo ai modelli di ottenere risultati differenti qualora i dati richiamino alla più generica famiglia GEV. In questi ambiti i dati a disposizione spesso riguardano più stazioni di rilevazioni, con locazioni e altitudini differenti.
Un possibile approccio potrebbe essere quello di modellare ciascuna stazione di rilevazione in maniera indipendente, ottenendo quindi i risultati per ciascun campione riguardante i singoli luoghi. Può risultare utile però analizzare i dati aggregati per tutte le stazioni, ottenendo comunque delle stime dei parametri per ciascuna stazione, ma sfruttando la capacità di imporre una contrazione locale e globale di alcune a priori di contrazioni proposte. Un approccio di questo tipo permette di complicare il modello, nel caso in cui gli esperti riescano a fornire qualche informazione sulla struttura dei dati in un territorio vasto, nel quale ci si aspetta che il comportamento di questi sia descritto con distribuzione Gumbel nella maggior parte delle locazioni, permettendo ad alcune stazioni di definire comportamenti

differenti. Viene quindi proposta un'estensione alla modellazione bayesiana dei valori estremi con a priori di contrazione, che tenga conto di una struttura gerarchica dei dati. Dato il campione di massimi annuali per J stazioni $\mathbf{Z} = \mathbf{Z}_1, \ldots, \mathbf{Z}_J$ con $\mathbf{Z}_j = Z_{1,j}, \ldots, Z_{m_j,j}$ e m_j la numerosità campionaria di ciascuna stazione, si definisce la verosimiglianza per $(\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\xi})$, per il caso specifico:

$$L(\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\xi}; \mathbf{Z}) = \prod_{j=1}^{J} \prod_{i=1}^{m_j} g(z_{i,j} | \mu_j, \sigma_j, \xi_j). \tag{3.15}$$

Si vuole quindi ottenere la distribuzione a posteriori dei parametri, definendo le quantità necessarie, come segue:

$$\begin{aligned} p(\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\xi} | \mathbf{Z}) &\propto L(\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\xi}; \mathbf{Z}) \times p(\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\xi}), \\ p(\boldsymbol{\mu}, \boldsymbol{\phi}, \boldsymbol{\xi}) &= \prod_{j}^{J} p(\mu_j) p(\phi_j) p(\xi_j), \\ \mu_j &\sim N(0, v_\mu^2), \\ \phi_j &\sim N(0, v_\phi^2), \end{aligned}$$

con $\phi_j = \log \sigma_j$ e $v_\mu = v_\phi$ che vengono posti pari a 100 così da ottenere delle a priori diffuse non informative per tali parametri. Queste risultano uguali per tutti i gruppi di osservazioni e indipendenti tra ciascuno di questi lasciando completamente libere le stime dei parametri per la media e la varianza della distribuzione generalizzata dei valori estremi. La distribuzione a priori degli ξ_j invece non sarà uguale per tutti i J gruppi, ma dovrà avere un parametro di contrazione globale uguale per tutti i gruppi e uno che si modifica a seconda di ciascuno di questi. Una a priori di contrazione che permette tale tipo di struttura è la Horseshoe, che nel seguente caso per ciascun ξ_j si definisce come segue:

$$\begin{aligned} (\xi_j | \lambda_j, \tau) &\sim N(0, \lambda_j^2 \tau^2), \\ \lambda_j &\sim C^+(0, 1), \end{aligned} \tag{3.16}$$

dove $C^+(0,1)$ è la distribuzione Half-Cauchy per la deviazione standard λ, che definisce il parametro di contrazione locale, e τ il parametro che definisce la contrazione globale. Quest'ultimo viene posto pari a $p_n/n\sqrt{\log(n/p_n)}$, valore ottimale nel caso della Horseshoe per la regressione lineare come descritto in Van der Pas et al. (2017), con p_n il numero di parametri che ci si attende siano diversi da zero ed n il numero di osservazioni totali osservate. Tale scelta, nonostante non sia verificato sia ottimale nel caso dei valori estremi, permette di avere un parametro di contrazione globale che dipende dalla numerosità campionaria e dal numero di gruppi che ci si attende abbiano valori che si distribuiscono come una Gumbel. Senza particolari conoscenze riguardo i dati risulta comunque una scelta pragmatica quella di porre $p_n = J$ così da non ottenere una contrazione globale eccessivamente elevata, lasciando a λ_j il compito di contrarre il valore di ξ_j dei singoli gruppi.

Un modello di questo tipo quindi permette di concentrarsi sul comportamento di ciascuna stazione di rilevazione ma tenendo in considerazione anche la presenza di altri gruppi di osservazioni.
La distribuzione a posteriori per i parametri, come per i precedenti casi, si ottiene utilizzando il metodo di simulazione *Markov Chain Monte Carlo* attraverso l'algoritmo Metropolis-Hastings con passeggiata casuale su ciascun parametro come in sezione (3.3.1). Tale metodo nonostante sia molto oneroso in termini di costo computazionale e di tempi di calcolo quanto più la numerosità dei gruppi aumenta, risulta essere comunque il metodo che permette di ottenere delle catene con caratteristiche soddisfacenti, permettendo un'esplorazione completa dello spazio campionario.

Capitolo 4

Simulazioni

Viene effettuato uno studio di simulazione per comprendere meglio il funzionamento dei modelli proposti. Nel caso in questione è interessante valutare l'andamento di questi, dipendentemente dal vero valore di ξ_0 e dalla numerosità del campione. Nell'analisi dei valori estremi infatti, l'interesse è focalizzato sullo studio dei massimi annuali, quindi per ottenere campioni numerosi, sono necessari tempi di rilevazione molto lunghi. Vengono quindi valutate tre numerosità che solitamente è possibile avere analizzando dati reali, e che rappresentano casi di rilevazioni del fenomeno d'interesse brevi o lunghi: 10, 25 e 50 osservazioni. Per quanto riguarda invece il vero valore del parametro di interesse si valutano sia casi in cui ξ sia pari o meno a zero. I valori valutati sono $-0.5, -0.3, -0.1, 0, 0.1, 0.3, 0.5$ così da osservare l'andamento dei modelli nel caso in cui il vero modello sia Fréchet, Weibull e Gumbel, senza valutare però valori troppo grandi in valore assoluto di ξ poiché ci si concentra sui casi più complessi di identificazione del modello e non quando i massimi hanno un andamento più chiaro. Come ultimo aspetto quindi rimane da definire i valori che rappresentano la media e la varianza dei dati, scegliendo una singola coppia di valori invariata nelle varie simulazioni con $\mu_0 = 0$ e $\sigma_0 = 1$.
Per ciascuna combinazione dei 7 valori di ξ e delle 3 dimensioni campionarie vengono quindi generati 100 campioni di massimi annuali dalla distribuzione generalizzata dei valori estremi. I dati generati vengono utilizzati quindi per stimare i parametri alla base del processo generatore attraverso un'analisi Bayesiana, utilizzando i modelli proposti, con le *a priori* di contrazioni continue e Spike and Slab discrete.
Come confronto rispetto ai metodi classici viene utilizzato anche il modello bayesiano che prevede l'utilizzo di una *a priori* diffusa normale per il parametro ξ, la cui varianza non viene scelta pari a 100 come in Coles (2001), per avere un confronto onesto con i modelli proposti. Il parametro di forma della GEV infatti si assume non segua valori particolarmente elevati in valore assoluto, in quanto questo comporta svariate problematiche nella definizione della verosimiglianza. Come standard di riferimento per la valutazione delle metodologie proposte quindi si sceglie il modello bayesiano che prevede la definizione di una *a priori* per ξ centrata in zero, con varianza pari a 2. Si ottiene quindi utilizzando cia-

scuno dei modelli, su un campione di valori estremi, un campione MCMC rappresentativo della distribuzione a posteriori del modello, sul quale verranno poi calcolate determinate varie metriche per valutarne la prestazione.

4.1 Metriche di confronto

I risultati teorici considerati in questo caso sono di natura bayesiana frequentista Ghosal & Van der Vaart (2017), nel senso che si assume che esista un qualche valore reale dei parametri μ_0, σ_0, ξ_0 alla base del processo di generazione dei dati, e le procedure bayesiane vengono valutate per la misura e la precisione con cui sono in grado di trovare il vero vettore dei parametri. Le misure considerate, che definiscono le prestazioni dei modelli, nonostante possano sembrare bayesiane, portano tutte alla valutazione secondo criteri frequentisti tradizionali.
I criteri principali sui quali vengono valutati le *a priori* di contrazione e quella diffusa sono dati dalla capacità della distribuzione a posteriori ottenuta dai modelli di (i) stimare correttamente il parametro ξ sottostante e di (ii) quantificare l'incertezza rimanente sull'esatto valore del parametro. Inoltre per il caso specifico vengono utilizzate delle statistiche propriamente utilizzate nell'analisi dei valori estremi come (iii) il livello di ritorno corrispondente a determinati anni per valutare la prestazione generale del modello sulla totalità dei parametri (μ, σ, ξ) e soprattutto sull'estrapolazione. Infine viene valutata la (iv) capacità dei modelli di discriminare se il vero modello sottostante la generazione dei dati sia effettivamente Gumbel oppure sia uno tra Frèchet e Weibull.
Per ognuna delle 15 diverse combinazioni di numerosità e ξ, vengono creati gruppi di 100 campioni. Questi campioni sono successivamente utilizzati per generare rispettivamente gruppi di 100 distribuzioni a posteriori stimati tramite l'algoritmo MCMC per ciascun modello. Successivamente, per ciascuno di questi gruppi di 100 distribuzioni a posteriori, vengono calcolate le metriche specifiche per valutare l'efficacia dei modelli proposti. In questo modo, per ciascun modello si otterranno 15 valori per ciascuna metrica. Di seguito, verranno descritte in dettaglio le metriche selezionate per valutare la validità dei modelli proposti.
Per valutare (i) la capacità della distribuzione di stimare correttamente il parametro reale di ξ è necessario scegliere uno stimatore di tale parametro. Si sceglie quindi di prendere la media a posteriori del parametro definita come $\hat{\xi}$ e, definendo ξ_0 il vero valore del parametro si calcola l'errore quadratico medio ottenuto dalle 100 repliche:

$$\sum_{i=1}^{100} (\hat{\xi}_i - \xi_0)^2. \tag{4.1}$$

Questo permette di valutare in media quanto il valore stimato si discosta dal parametro sottostante la generazione del campione dei dati.
Per (ii) quantificare l'incertezza del parametro vengono studiati gli intervalli di credibilità,

definendo per ciascuna distribuzione a posteriori il sottoinsieme dello spazio parametrico $\hat{C}(1-\alpha)$ tale per cui la massa di probabilità corrispondente sia $1-\alpha$. In particolare per ξ si sceglie l'intervallo di massima credibilità a posteriori (HPD) per gestire eventuali distribuzioni a posteriori sbilanciate. Per valutare quindi la bontà del modello si vuole ottenere la probabilità che tale intervallo contenga il vero valore del parametro ξ_0 calcolando:

$$\frac{1}{100}\sum_{i=1}^{100}\mathbb{I}\{\xi_0 \in \hat{C}_i(1-\alpha)\}, \quad \text{con } \alpha = 0.05. \tag{4.2}$$

Per (iii) valutare la bontà delle stime complessive sulla totalità dei parametri si utilizza il livello di ritorno, per comprendere quanto le stime dei quantili della distribuzione GEV d'interesse siano precise. Data la distribuzione a posteriori per i parametri μ, σ, ξ, per ogni periodo di ritorno p si ottiene una distribuzione del livello di ritorno z_p, con z_p definito nell'Equazione (1.4). Per ottenere una stima puntuale del livello di ritorno quindi è necessario definire uno stimatore quale la media o la mediana della distribuzione di z_p, mentre per una stima intervallare si utilizza l'intervallo di credibilità. Se si fa variare il valore del periodo di ritorno è possibile quindi ottenere il grafico del livello di ritorno, con una valutazione anche dell'incertezza associata come in Figura 4.1.

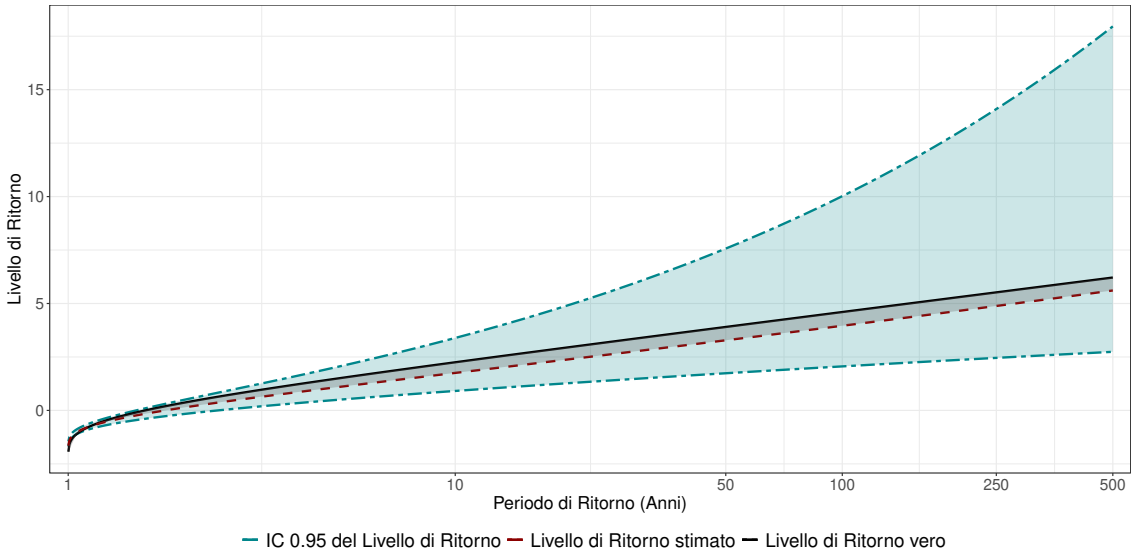

Figura 4.1: Esempio di grafico del Livello di Ritorno stimato, reale(μ_0, σ_0, ξ_0), e il suo Intervallo di Credibilità 0.95. Le due aree colorate definiscono le metriche per la bontà del modello stimato.

Il livello di ritorno stimato è quindi interpretabile come una funzione $\hat{z}_q(p)$, con $q = 0.025, 0.5, 0.975$ per identificare se si sta stimando l'intervallo superiore, inferiore o la mediana delle distribuzioni ad ogni punto con p che varia. Un indice che permette quindi di valutare quanto l'estrapolazione di valori sia buona, identificando $z_0(p)$, la funzione del livello di ritorno ottenuta con i valori reali (μ_0, σ_0, ξ_0) è data dall'area che divide la curva stimata e quella reale :

$$\int_2^{500} |\hat{z}_{0.5}(p) - z_0(p)|\, dp. \tag{4.3}$$

La scelta della mediana come stima del livello di ritorno è dovuta alla sua robustezza in caso di distribuzioni molto sbilanciate, scenario molto frequente per i livelli di ritorno associati a periodi grandi. Oltre a confrontare la stima del livello di ritorno con quella ottenuta con i veri parametri sottostanti al modello generatore dei dati è necessario anche valutare il grado di incertezza associato alle stime, che viene calcolato attraverso l'area compresa tra la curve che definiscono gli intervalli superiori e inferiori di credibilità:

$$\int_2^{500} |\hat{z}_{0.975}(p) - \hat{z}_{0.025}(p)|\, dp. \tag{4.4}$$

Infine come ultime metriche vengono valutate la capacità del modello di identificare la natura della distribuzione GEV, discriminando tra il modello Gumbel e la famiglia più ampia della distribuzione generalizzata dei valori estremi. Innanzitutto si valuta se l'intervallo di confidenza definito precedentemente $\hat{C}(1-\alpha)$ contiene o meno lo zero:

$$\frac{1}{100}\sum_{i=1}^{100} \mathbb{I}\{0 \in \hat{C}_i(1-\alpha)\}, \text{ con } \alpha = 0.05.$$

Questa metrica comunque definisce solo parzialmente la capacità del modello di discriminare i valori di ξ uguali o diversi da zero. Un altro metodo di confronto che viene scelto prevede ulteriori considerazioni da fare sulle distribuzioni a posteriori dei dati. Innanzitutto fissato un modello, vengono creati 3 gruppi di distribuzioni a posteriori, che descrivono i risultati ottenuti dalle repliche con differenti numerosità. All'interno di ciascun gruppo i risultati sono riferiti a campioni generati da distribuzioni GEV con valori uguali o diversi da zero ma uguali numerosità. Per ciascun campione MCMC rappresentativo della distribuzione a posteriori, si adotta una procedura di troncamento che prevede l'assegnazione di un valore soglia, che in questo caso viene scelto pari a 0.05, al di sotto del quale le osservazioni del parametro ξ in valore assoluto vengono direttamente poste a zero.
Questa operazione permette di generare delle probabilità che il valore del parametro venga stimato uguale o diverso da zero, a seconda che rientri o meno nella soglia prestabilita. L'obiettivo principale di questo processo è valutare la bontà della stima dicotomica del tipo di distribuzione generatrice dei dati (Gumbel =1 oppure generica GEV=0) basata su ciascun campione MCMC come singola osservazione. Conoscendo il vero valore del parametro, è possibile calcolare vari indici diagnostici per i modelli utilizzati, tra cui sensibilità e specificità, e rappresentarli graficamente attraverso la curva ROC. Questo studio di simulazione permette quindi di comprendere meglio il comportamento delle varie proposte anche e sopratutto rispetto al metodo classico non informativo.

4.2 Risultati

I risultati delle simulazioni rispetto alle metriche definite precedentemente, vengono mostrati sotto forma di tabella e illustrati attraverso dei grafici per una migliore interpretabilità. Innanzitutto si osserva il comportamento dei modelli in Tabella 4.1 e in Figura 4.2, dove sono rispettivamente riportati i valori dell'errore quadratico come descritto nell'Equazione (4.1), e i boxplot degli scarti in valore assoluto per ciascuna delle 100 repliche corrispondenti ad ogni combinazione di numerosità campionaria e valore reale di ξ per ciascun modello. Oltre ad uno scostamento della stima dal vero valore ξ che diminuisce globalmente per tutti i modelli, rispetto all'aumento della numerosità campionaria è possibile notare dei comportamenti ben visibili per ciascuno di questi.
Innanzitutto si osserva un comportamento globale decisamente migliore nel caso dell'utilizzo di *a priori* di contrazione continue e della Spike and Slab continua rispetto alle *a priori* Spike and Slab discrete. In particolare come atteso i modelli che utilizzano la Laplace, la Horseshoe e la Spike and Slab Lasso permettono di ottenere delle stime molto vicine al vero parametro quando la distribuzione del campione è Gumbel, ottenendo un netto miglioramento rispetto all'utilizzo di una *a priori* non informativa. Tale comportamento è generalizzabile per tutti i veri valori del parametro ξ considerati, quando la numerosità campionaria risulta essere scarsa, ovvero con 10 osservazioni, dove le distribuzioni del valore assoluto degli scarti presentano quasi in tutti i casi una bassa variabilità, e una concentrazione vicina allo zero, quando si utilizzano *a priori* di contrazione continue o la Spike and Slab continua.
Con numerosità più elevate le prestazioni del modello bayesiano con *a priori* normale sembrano migliorare soprattutto all'aumento dello scostamento del vero valore di ξ da zero. I modelli con *a priori* Laplace, Horseshoe e Spike and Slab comunque nel caso di valori pari a 0.1 e -0.1 di ξ permettono di ottenere delle stime del parametro di forma migliori rispetto a quello normale, perdendo la loro efficacia relativa a quest'ultimo solo quando la distribuzione generatrice dei dati presenta ξ particolarmente diversi da zero. L'errore di stima rimane comunque basso e leggermente più alto rispetto al classico modello bayesiano solo in alcuni casi, definendo delle prestazioni buone anche quando il campione si distribuisce come una generica GEV.
Il modello bayesiano che utilizza la Spike and Slab normale sembra avere un comportamento comparabile agli altri proposti tranne per i casi in cui viene utilizzato con campioni provenienti da una Weibull, nei quali sembrerebbe stimare il valore del parametro di forma sempre a zero. Il modello Spike and Slab pMOM mostra un comportamento analogo in tali situazioni, tuttavia, tende a produrre errori più variabili e generalmente maggiori rispetto agli altri modelli, risultando comparabile solamente in specifiche circostanze. I modelli con *a priori* di contrazione continua e Spike and Slab continua sembrano quindi essere i migliori in termini di errori di stima di ξ. Si va quindi ad osservare il comportamento dell'intervallo di credibilità per il parametro di forma per ciascun modello.

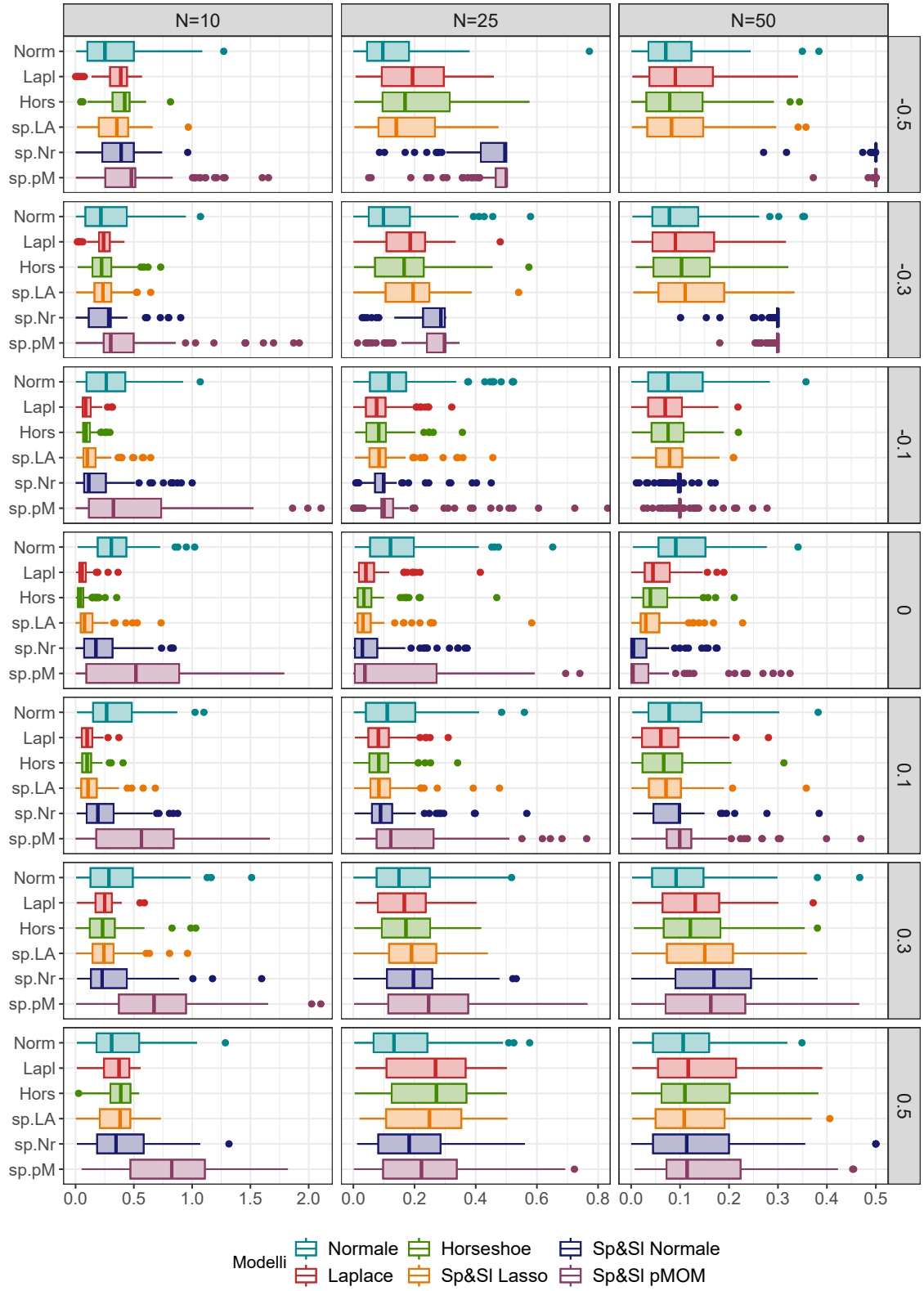

Figura 4.2: Boxplot di $|\hat{\xi}_i - \xi_0|$. I valori delle righe corrispondono ai veri valori del parametro ξ_0 mentre le colonne definiscono i valori della numerosità campionaria.

In Tabella 4.2 si osservano le proporzioni di intervalli che contengono il vero valore del parametro per ciascuna combinazione di numerosità e valore reale di ξ per ciascun modello. Si nota che il modello con *a priori* normale diffusa globalmente sembra registrare le migliori prestazioni, anche se i modelli proposti con *a priori* di contrazione sembrano generare degli intervalli che contengono il vero valore del parametro con alta probabilità con valori anche diversi da zero, con un peggioramento con veri valori pari a 0.5. I modelli con *a priori* Spike and Slab discrete come visto in precedenza sembrano avere una pessima copertura nel caso di modelli Weibull e comunque sembrano essere i peggiori rispetto agli altri modelli proposti in questi termini.
Oltre a verificare la copertura, si va a vedere la proporzione degli intervalli di credibilità che contengono il valore 0 e si notano in Tabella 4.5 dei risultati abbastanza coerenti rispetto al modello sottostante ai dati. Ci si aspetta infatti che la proporzione di intervalli che contengono lo zero diminuisca quanto più è grande lo scostamento del vero valore del parametro da zero, aiutata anche da un aumento di numerosità del campione, che dovrebbe portare una riduzione della grandezza dell'intervallo. Questo accade in modo simmetrico sia per valori positivi che negativi di ξ per la maggior parte dei modelli.
Le Spike and Slab discrete evidenziano ancora una volta le scarse prestazioni con valori negativi del parametro di forma della GEV. Nonostante il modello non informativo presenti le proporzioni più basse di intervalli che contengono lo zero quando la distribuzione non è la Gumbel, anche gli altri modelli, con *a priori* di contrazione presentano dei buoni risultati. Seppur la numerosità di intervalli che contiene zero, in questi casi risulti più elevata rispetto al modello standard, non si ottengono mai delle differente particolarmente elevate. Tali risultati sembrano rassicuranti riguardo la quantità di contrazione a zero indotta dai modelli con *a priori* informative, in quanto non sembra sovrastare il segnale dei dati.
Dopo aver studiato la capacità dei modelli di stimare il vero valore del parametro ξ si passa quindi a valutare la bontà delle stime globali dei parametri della distribuzione. Per fare questo il livello di ritorno può essere un ottimo indice che tenga conto della stima di tutti i parametri in gioco nella distribuzione generalizzata dei valori estremi. Per ciascuna replicazione, per calcolare la bontà di adattamento globale di ciascun modello viene calcolata l'area tra la curva del livello di ritorno stimata e quella reale del modello sottostante i dati e l'area tra le curve che definiscono l'intervallo di credibilità del livello di ritorno , come definito nelle Equazioni (4.1) e (4.1).
Vengono quindi mostrate le distribuzioni di tali quantità, in scala logaritmica per una miglior visibilità, attraverso l'utilizzo di boxplot in Figura 4.3 e 4.4. In Tabella 4.3 e 4.4 viene rappresentato rispettivamente il rapporto delle mediane di area e distanza per ciascun gruppo di repliche rispetto a tale quantità per il modello con *a priori* non informativa normale. Così facendo risulta quindi visibile se i modelli funzionino meglio o peggio, se presentano valori rispettivamente minori o maggiori di 1, rispetto allo standard di riferimento non informativo.
Si comincia quindi con le considerazioni legate alla distanza della curva del livello di ritorno

reale rispetto a quella stimata. Questa innanzitutto si nota che globalmente aumenta per tutti i modelli all'aumentare del valore reale di ξ. Si nota inoltre che, anche a livello di bontà globale oltre che per il singolo parametro ξ, il modello Spike and Slab non locale non sembra presentare delle buone prestazioni. Si osserva poi che le curve di ritorno stimate da tutti i modelli con *a priori* di contrazione hanno una distanza maggiore dalla reale rispetto al modello con *a priori* non informativa nel caso in cui il parametro reale ξ abbia valori minori di -0.1. Sembrerebbe quindi che tali modelli fatichino nella stima di campioni provenienti dalla Weibull, con le Spike and Slab discrete che però sembrano presentare le prestazioni peggiori.
In tutti gli altri casi per qualsiasi numerosità del campione i modelli informativi specificati, a parte la Spike and Slab pMOM presentano distanze mediamente minori, in corrispondenza di ξ vicini e uguali a 0, e comparabili al modello non informativo con valori grandi di ξ, presentando però spesso una variabilità più bassa di tali distanze. Sembrerebbe quindi che i modelli proposti permettano di ottenere un miglioramento globale nelle stime nel caso in cui le osservazioni siano Gumbel, ma comunque senza distorcere le stime quando il valore del parametro si discosta particolarmente da zero.
Si osserva quindi l'area definita dagli intervalli di credibilità del livello di ritorno. Questa globalmente aumenta all'aumentare del valore di ξ, per la natura della distribuzione GEV che presenta delle code a destra sempre più pesanti proporzionalmente all'aumento del parametro di forma. La Spike and Slab pMOM sembra ancora una volta presentare i risultati peggiori rispetto agli altri modelli, mentre la Spike and Slab Normale sopratutto in corrispondenza di valori piccoli o uguali a zero di ξ mostra degli intervalli non particolarmente grandi, con una varianza però abbastanza elevata. I modelli con *a priori* di contrazione Laplace, Horseshoe, Spike and Slab lasso sembrano comunque diminuire l'incertezza riguardo alla stima del livello di ritorno rispetto al modello non informativo, facendo registrare dei netti miglioramenti per alcuni casi. Questi miglioramenti comunque sono da prendere con cautela perché degli intervalli di credibilità per il livello di ritorno più piccoli possono anche essere una conseguenza della contrazione verso zero del parametro di forma.
Tali risultati comunque combinati a quelli legati alla distanza della stima del livello di ritorno rispetto al vero valore e all'errore di stima del parametro di forma, permettono di essere comunque soddisfatti delle prestazioni di alcuni modelli proposti. Il modello Laplace Horseshoe e Spike and Slab lasso tendenzialmente sembrano portare dei miglioramenti sostanziali rispetto al modello bayesiano non informativo ,nel caso in cui i dati provengano da una distribuzione Gumbel. Inoltre sembra che la contrazione a zero non sia così forte da portare una distorsione evidente nel caso in cui il modello sottostante sia Frechét o Weibull, ottenendo delle stime accurate anche a livello globale.

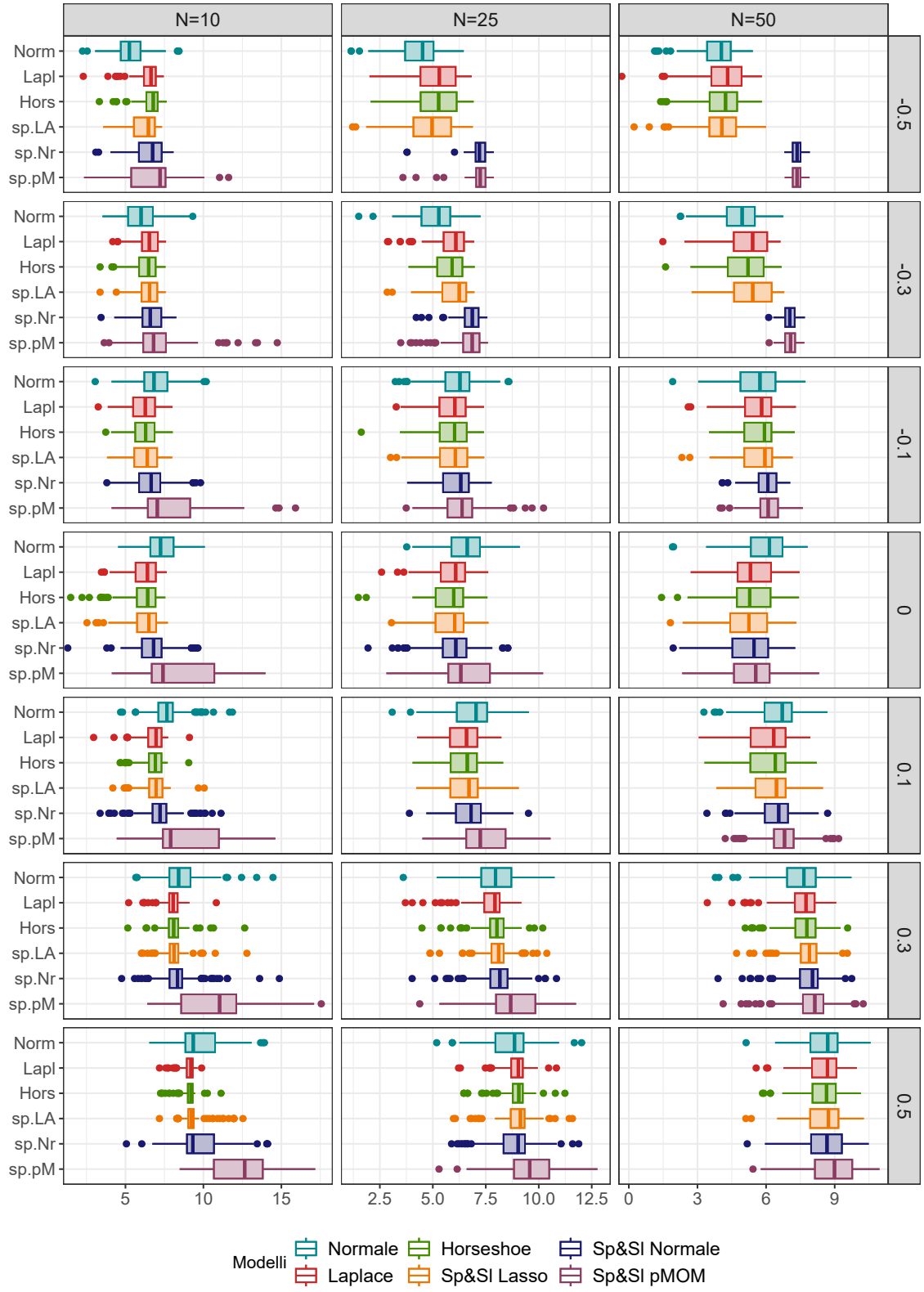

Figura 4.3: Boxplot della distanza della curva del livello di ritorno reale rispetto a quello stimato dal modello, in scala logaritmica per una miglior visualizzazione. I valori delle righe corrispondono ai veri valori del parametro ξ_0 mentre le colonne definiscono i valori della numerosità campionaria.

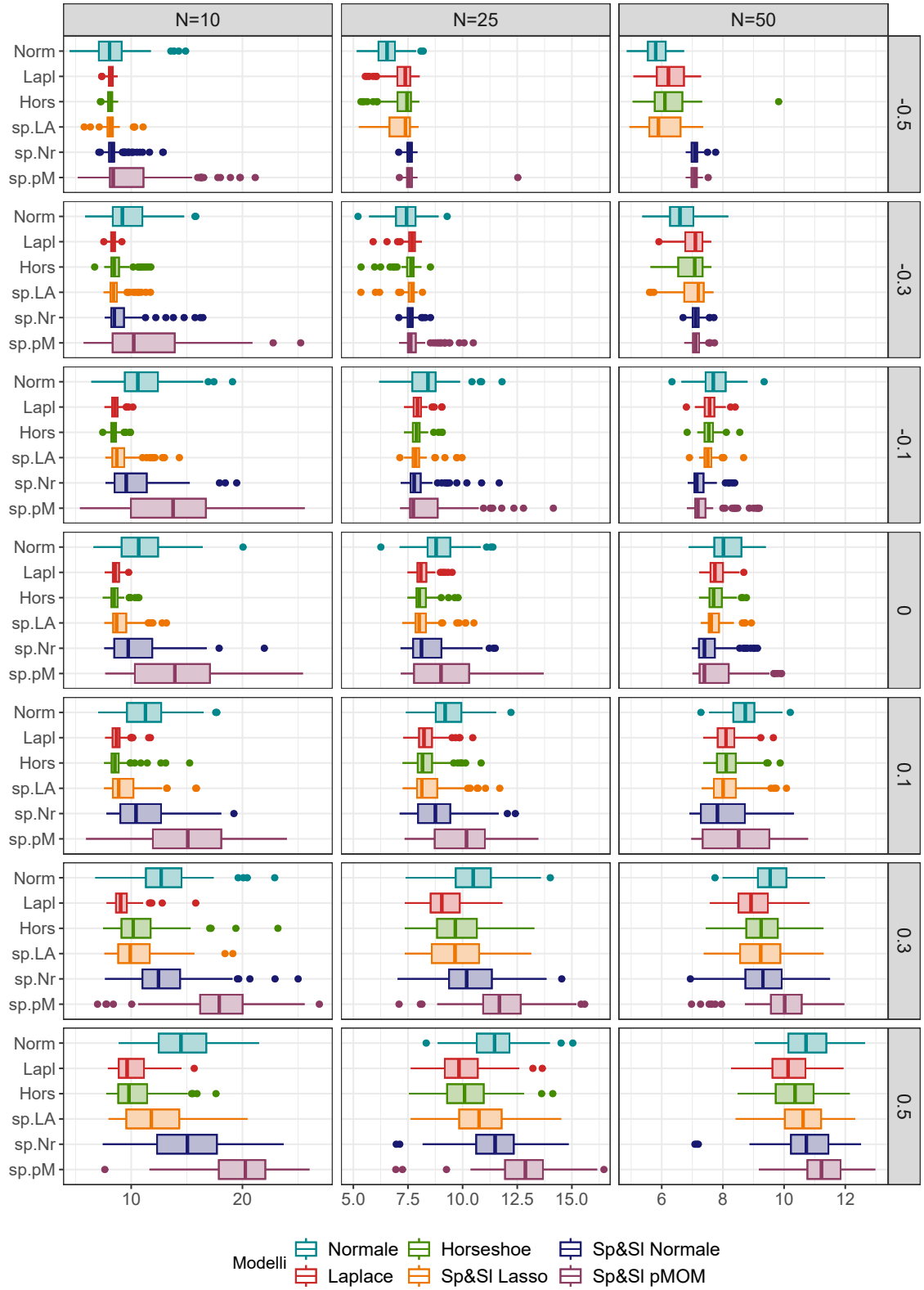

Figura 4.4: Boxplot dell'Area ottenuta da intervalli di credibilità per il livello di ritorno da 2 a 500 anni, in scala logaritmica per una miglior visualizzazione. I valori delle righe corrispondono ai veri valori del parametro ξ_0 mentre le colonne definiscono i valori della numerosità campionaria.

		ξ_0						
N	Modello	-0.5	-0.3	-0.1	0	0.1	0.3	0.5
10	Normale	0.97	0.96	0.97	0.98	0.99	0.98	0.99
	Laplace	0.67	0.85	1.00	1.00	1.00	0.86	0.62
	Horseshoe	0.57	0.96	1.00	1.00	1.00	0.96	0.63
	SpS Lasso	0.87	0.91	1.00	1.00	1.00	0.89	0.85
	SpS Norm	0.79	0.79	0.87	1.00	0.96	0.97	0.98
	SpS pmom	0.74	0.69	0.73	0.89	0.80	0.86	0.86
25	Normale	0.98	0.94	0.94	0.95	0.94	0.97	0.95
	Laplace	0.79	0.82	0.99	1.00	0.97	0.78	0.67
	Horseshoe	0.85	0.90	0.99	1.00	0.96	0.91	0.74
	SpS Lasso	0.89	0.81	0.99	1.00	0.97	0.79	0.83
	SpS Norm	0.34	0.44	0.69	1.00	0.80	0.93	0.93
	SpS pmom	0.22	0.41	0.51	0.98	0.82	0.86	0.93
50	Normale	0.94	0.90	0.93	0.95	0.96	0.94	0.98
	Laplace	0.86	0.81	0.93	1.00	0.94	0.78	0.83
	Horseshoe	0.90	0.82	0.91	1.00	0.90	0.84	0.88
	SpS Lasso	0.89	0.77	0.90	1.00	0.91	0.77	0.92
	SpS Norm	0.02	0.05	0.28	1.00	0.56	0.81	0.92
	SpS pmom	0.01	0.10	0.24	0.99	0.58	0.87	0.93

Tabella 4.2: percentuale di Intervalli di Credibilità HPD con $\alpha = 0.05$ che contengono ξ_0 su 100 repliche.

		ξ_0						
N	Modello	-0.5	-0.3	-0.1	0	0.1	0.3	0.5
10	Normale	0.183	0.151	0.162	0.159	0.167	0.208	0.203
	Laplace	0.143	0.063	0.014	0.008	0.015	0.070	0.136
	Horseshoe	0.160	0.076	0.013	0.007	0.015	0.098	0.145
	SpS Lasso	0.139	0.073	0.036	0.026	0.033	0.090	0.149
	SpS Norm	0.169	0.094	0.095	0.097	0.100	0.167	0.231
	SpS pmom	0.353	0.354	0.462	0.587	0.526	0.684	0.781
25	Normale	0.027	0.028	0.039	0.035	0.034	0.047	0.045
	Laplace	0.056	0.037	0.011	0.007	0.011	0.036	0.083
	Horseshoe	0.059	0.036	0.011	0.006	0.012	0.041	0.084
	SpS Lasso	0.048	0.042	0.016	0.009	0.015	0.046	0.081
	SpS Norm	0.206	0.069	0.016	0.012	0.020	0.050	0.059
	SpS pmom	0.218	0.071	0.048	0.055	0.061	0.105	0.081
50	Normale	0.014	0.017	0.015	0.016	0.016	0.018	0.021
	Laplace	0.019	0.020	0.008	0.005	0.007	0.023	0.028
	Horseshoe	0.017	0.021	0.008	0.004	0.008	0.024	0.027
	SpS Lasso	0.017	0.025	0.008	0.003	0.008	0.028	0.027
	SpS Norm	0.246	0.086	0.009	0.002	0.011	0.036	0.031
	SpS pmom	0.249	0.087	0.012	0.008	0.018	0.035	0.036

Tabella 4.1: EQM calcolato come in (4.1), su 100 repliche per il parametro ξ rispetto al vero valore ξ_0.

		ξ_0						
N	Modello	-0.5	-0.3	-0.1	0	0.1	0.3	0.5
10	Normale	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	Laplace	1.128	0.432	0.126	0.121	0.072	0.026	0.008
	Horseshoe	1.129	0.476	0.111	0.108	0.063	0.081	0.010
	SpS Lasso	1.104	0.440	0.148	0.135	0.088	0.061	0.073
	SpS Norm	1.261	0.491	0.346	0.396	0.425	0.776	1.833
	SpS pmom	1.399	2.721	22	25	45	182	348
25	Normale	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	Laplace	2.340	1.295	0.618	0.487	0.382	0.235	0.200
	Horseshoe	2.533	1.236	0.595	0.457	0.355	0.440	0.263
	SpS Lasso	2.359	1.285	0.565	0.452	0.343	0.431	0.498
	SpS Norm	2.861	1.183	0.533	0.507	0.647	0.753	1.032
	SpS pmom	2.774	1.204	0.517	1.243	2.692	3.254	4.029
50	Normale	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	Laplace	1.515	1.661	0.877	0.755	0.545	0.536	0.553
	Horseshoe	1.359	1.601	0.851	0.736	0.537	0.731	0.684
	SpS Lasso	1.096	1.794	0.813	0.678	0.500	0.718	0.897
	SpS Norm	3.595	1.660	0.573	0.542	0.411	0.801	1.001
	SpS pmom	3.593	1.665	0.593	0.545	0.802	1.656	1.694

Tabella 4.3: Rapporto dell'area tra le curve degli intervalli HPD del livello di ritorno stimato per il modello con *a priori* non informativa, sulla stessa quantità calcolata per i modelli con *a priori* informative.

		ξ_0						
N	Modello	-0.5	-0.3	-0.1	0	0.1	0.3	0.5
10	Normale	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	Laplace	4.065	1.725	0.579	0.440	0.499	0.700	0.874
	Horseshoe	4.669	1.610	0.597	0.449	0.485	0.736	0.885
	SpS Lasso	3.463	1.726	0.657	0.487	0.513	0.740	0.911
	SpS Norm	4.446	1.809	0.838	0.659	0.653	0.911	1.028
	SpS pmom	7.249	2.251	1.268	1.207	1.282	13	28
25	Normale	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	Laplace	2.216	2.250	0.800	0.572	0.641	0.956	1.196
	Horseshoe	2.126	1.876	0.800	0.522	0.671	1.068	1.243
	SpS Lasso	1.569	2.629	0.835	0.548	0.697	1.177	1.289
	SpS Norm	14	4.920	1.054	0.575	0.786	1.235	1.149
	SpS pmom	14	4.962	1.123	0.722	1.191	2.146	2.034
50	Normale	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	Laplace	1.310	1.600	1.073	0.442	0.696	1.095	0.989
	Horseshoe	1.204	1.324	1.189	0.424	0.733	1.160	0.972
	SpS Lasso	1.022	1.589	1.244	0.411	0.794	1.268	1.022
	SpS Norm	27	8.164	1.386	0.511	0.856	1.468	0.983
	SpS pmom	27	8.351	1.400	0.549	1.101	1.620	1.382

Tabella 4.4: Rapporto della distanza tra la curva del livello di ritorno vera e stimata per il modello con *a priori* non informativa, sulla stessa quantità calcolata per i modelli con *a priori* informative.

		ξ_0						
N	Modello	-0.5	-0.3	-0.1	0	0.1	0.3	0.5
10	Normale	0.78	0.93	0.99	0.98	0.98	0.91	0.73
	Laplace	1.00	1.00	1.00	1.00	1.00	0.98	0.96
	Horseshoe	1.00	0.99	1.00	1.00	1.00	0.98	0.96
	SpS Lasso	0.98	1.00	1.00	1.00	1.00	0.98	0.92
	SpS Norm	1.00	1.00	1.00	1.00	1.00	0.97	0.83
	SpS pmom	0.97	0.96	0.94	0.89	0.87	0.71	0.46
25	Normale	0.20	0.60	0.93	0.95	0.89	0.55	0.41
	Laplace	0.66	0.94	1.00	1.00	0.98	0.83	0.59
	Horseshoe	0.72	0.93	1.00	1.00	0.98	0.81	0.58
	SpS Lasso	0.65	0.94	0.99	1.00	0.98	0.83	0.59
	SpS Norm	1.00	1.00	1.00	1.00	0.98	0.78	0.58
	SpS pmom	1.00	1.00	0.99	0.98	0.95	0.60	0.35
50	Normale	0.00	0.22	0.89	0.95	0.86	0.32	0.02
	Laplace	0.08	0.50	0.97	1.00	0.96	0.61	0.11
	Horseshoe	0.11	0.54	0.99	1.00	0.97	0.67	0.14
	SpS Lasso	0.11	0.58	0.98	1.00	0.98	0.71	0.14
	SpS Norm	1.00	1.00	1.00	1.00	0.98	0.71	0.18
	SpS pmom	1.00	1.00	1.00	0.99	0.94	0.58	0.11

Tabella 4.5: Tabella della percentuale di intervalli HPD che contengono lo zero, calcolati sulle 100 repliche per ciascuna combinazione di numerosità e valore di ξ.

Oltre a verificare la capacità di stimare i parametri e di estrapolazione di valori futuri, qualità comunque fondamentali nel caso in cui di analisi di valori estremi, si sceglie di studiare la capacità dei modelli proposti di portare effettivamente un miglioramento interpretativo rispetto ad un'analisi bayesiana non informativa. Si vuole comprendere la capacità dei modelli di discriminare dati provenienti da una distribuzione Gumbel oppure da una più generica distribuzione generalizzata dei valori estremi.
Per fare questo è necessario trovare un metodo che permetta di passare da una distribuzione a posteriori continua per i valori di ξ ad una dicotomica che identifichi distribuzioni Gumbel con valori pari ad 1 e distribuzioni GEV generiche con valori pari a 0. Si sceglie quindi la soglia 0.05 per la quale le osservazioni dei campioni MCMC delle distribuzioni a posteriori in valore assoluto risultano minori, vengono poste direttamente pari a zero. Si ottiene quindi una proporzione di valori uguali a zero per ciascun modello in ciascuno scenario definito per ogni replicazione.
Interpretando questa proporzione come la probabilità che il modello sottostante sia Gumbel e conoscendo effettivamente il valore del parametro ξ è possibile valutare i modelli sulla base della loro capacità classificare la famiglia del modello GEV. Per ciascuna numerosità

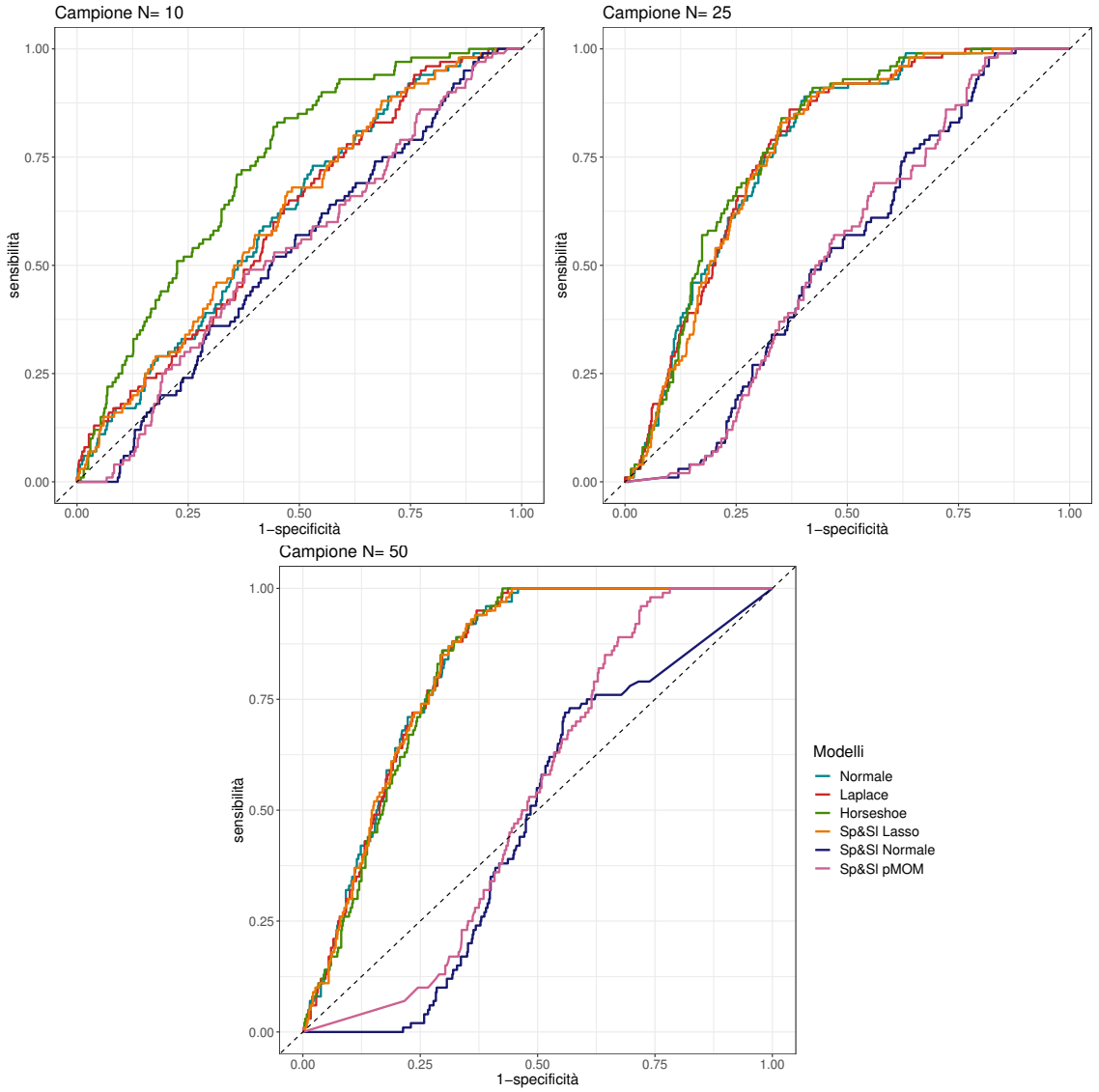

Figura 4.5: Curve ROC per la classificazione del tipo di distribuzione (Gumbel o GEV generica) da parte dei modelli proposti.

campionaria e ciascun modello quindi si confrontano le 100×7 probabilità stimate con il vettore con valore 1 se il campione corrispondente è stato generato da una Gumbel e 0 se dalla più generica GEV. Per passare dalla probabilità di essere Gumbel o meno alla classificazione effettiva è necessario definire un'ulteriore soglia sulla proporzione di ξ uguali a zero per la quale le probabilità maggiori di tale soglia definiscono classificazione pari a 1 e zero altrimenti.

Viene innanzitutto valutata la bontà globale della classificazione al variare della soglia attraverso una curva ROC per ciascuna delle numerosità considerate in fase di simulazione, illustrata in Figura 4.5. Si osserva innanzitutto che l'andamento globale dei modelli con *a priori* Spike and Slab discrete è decisamente deludente rispetto agli altri, con degli andamenti della curva che stanno anche sotto la diagonale, con delle prestazioni quasi

comparabili agli altri modelli solo quando la numerosità è bassa. I modelli con *a priori* Laplace e Spike and Slab LASSO invece presentano un andamento della curva ROC nel complesso molto simile al modello con *a priori* non informativa, per qualsiasi numerosità del campione considerata. Infine si osserva come il modello con *a priori* Horseshoe presenti un andamento particolarmente migliore nel caso in cui siano disponibili solo 10 osservazioni, sottolineandone la buona capacità di classificazione, mentre con numerosità maggiori presenta risultati comparabili ai modelli concorrenti. La curva ROC in ogni caso non risulta essere uno strumento che permette di definire effettivamente se la classificazione di un modello possa essere migliore di altre, in quanto nella realtà la soglia, che in questo caso viene fatta variare, deve essere scelta in modo arbitrario.
Una scelta ragionevole può essere di porre tale soglia pari a 0.5, così da ottenere un interpretazione per cui, se più della metà dei valori del parametro di scala risulta essere uguale a 0, la distribuzione di tale parametro è degenere a 0 e il modello stimato è Gumbel. Fissando quindi la soglia a 0.5 è possibile ottenere per ciascun modello, data una numerosità campionaria, la tabella di errata classificazione che risulta particolarmente utile per analizzare la capacità del modello di discriminare tra le due classi e per calcolare diverse metriche di valutazione delle prestazioni del modello.
In Tabella 4.6 si mostrano diverse metriche che permettono di valutare e comprendere il modo in cui i vari modelli classificano il tipo di distribuzione dei dati osservati. È necessario ricordare che nelle simulazioni considerate il numero di campioni generati da una Gumbel sono in minoranza rispetto a quelli generati dalla Frechèt e dalla Weibull quindi valutare i modelli rispetto all'accuratezza può portare a valutazioni sbagliate riguardo alle prestazioni dei modelli. Risulta fondamentale quindi andare ad osservare nello specifico la percentuale di casi in cui i modelli classificano i dati come Gumbel in modo corretto o come generica distribuzione GEV. Queste quantità corrispondono nel caso considerato a sensibilità e specificità ed identificano la percentuale di corretti positivi e corretti negativi. Due metriche che possono risultare molto utili in casi sbilanciati come questo sono l'F1 score, che considera sia la proporzione di veri positivi sia il richiamo la proporzione di reali positivi identificati come tali, e il Kappa di Cohen che tiene conto del fatto che la concordanza tra stimato e reale possa essere casuale (Cohen, 1960).
Osservando la Tabella 4.6 si nota innanzitutto che, come era prevedibile, il modello con *a priori* non informativa classifica tutti i campioni simulati come provenienti dalla distribuzione generalizzata dei valori estremi, ottenendo un ottima accuratezza, dovuta però al fatto che appunto nel seguente caso la maggior parte dei campioni non proviene dalla Gumbel. Si nota infatti che la sensibilità per tale modello è pari a 0 , come anche l'F1 score e il Kappa di Cohen, a sottolineare come tale modello non permetta di discriminare tra le due famiglie di distribuzione, qualsiasi sia la numerosità del campione. Il modello con *a priori* Laplace presenta lo stesso problema con basse numerosità del campione, ma la contrazione a zero sembra aumentare con campioni più grandi, portando comunque a classificare alcuni campioni come Gumbel, comunque con scarsi risultati.

N=10	Normale	Laplace	Horseshoe	SpS Lasso	SpS Norm	SpS pMoM
Errore Class	0.143	0.143	0.364	0.281	0.343	0.301
Accuratezza	0.857	0.857	0.636	0.719	0.657	0.699
Sensibilità	0.000	0.000	0.720	0.300	0.340	0.280
Specificità	1.000	1.000	0.622	0.788	0.710	0.768
F1 Score	0.000	0.000	0.361	0.233	0.221	0.210
K Cohen	0.000	0.000	0.187	0.071	0.034	0.038

N=25	Normale	Laplace	Horseshoe	SpS Lasso	SpS Norm	SpS pMoM
Errore Class	0.143	0.164	0.280	0.323	0.521	0.497
Accuratezza	0.857	0.836	0.720	0.677	0.479	0.503
Sensibilità	0.000	0.030	0.700	0.760	0.570	0.590
Specificità	1.000	0.970	0.723	0.663	0.463	0.488
F1 Score	0.000	0.050	0.417	0.402	0.238	0.253
K Cohen	0.000	0.000	0.270	0.243	0.015	0.037

N=50	Normale	Laplace	Horseshoe	SpS Lasso	SpS Norm	SpS pMoM
Errore Class	0.143	0.190	0.243	0.269	0.584	0.576
Accuratezza	0.856	0.810	0.757	0.731	0.416	0.424
Sensibilità	0.000	0.350	0.630	0.770	0.800	0.850
Specificità	1.000	0.887	0.778	0.725	0.352	0.353
F1 Score	0.000	0.345	0.426	0.450	0.281	0.297
K Cohen	0.000	0.234	0.292	0.311	0.060	0.080

Tabella 4.6: Metriche per la valutazione della classificazione dei modelli proposti, con soglia di probabilità 0.5

I due modelli con *a priori* Spike and Slab discreta presentano delle prestazioni abbastanza simili per tutte le dimensioni del campione, con un miglioramento discreto con numerosità maggiori, ottenendo comunque scarsi risultati. Un comportamento interessante che seguono entrambi i modelli è quello di discriminare meglio i modelli GEV con campioni di 10 osservazioni, mentre all'aumentare di questi, la classificazione per entrambe le famiglie si eguaglia fino a raggiungere una miglior classificazione delle distribuzioni Gumbel con campioni di 50 osservazioni.
I modelli con *a priori* Horseshoe e Spike and Slab Lasso sembrano essere i migliori nel discriminare le distribuzioni d'interesse, entrambi mantengono con tutte le numerosità campionarie un sensibilità e una specificità che si aggira attorno allo 0.65/0.75 con un graduale miglioramento all'aumentare del numero di osservazioni. Anche in termini di F1 score e Kappa di Cohen, mediocri in valore assoluto, ma rispetto agli altri modelli , molto migliori. Nonostante questo comunque il modello Horseshoe sembra essere il più stabile

e funzionare meglio con basse numerosità, caso comunque molto frequente nello studio di valori estremi.
Le prestazioni del modello bayesiano con un *a priori* di contrazione Horseshoe risultano essere promettenti, conferendo maggior valore all'approccio complesso precedentemente proposto, soprattutto nell'analisi di insiemi di campioni dipendenti. L'efficacia del modello è stata verificata attraverso una simulazione studiata per singoli campioni e si è estesa l'indagine all'andamento del modello complesso.
Per l'esperimento sono stati generati 200 campioni, ciascuno composto da 25 osservazioni di valori massimi con $\mu = 0$ e $\sigma = 1$, variando i valori di ξ. Nello specifico, per 100 campioni i valori di ξ sono stati estratti da una distribuzione Gamma(0.5,1.5), mentre per i restanti 100 campioni ξ è stato impostato a zero. Di conseguenza, la distribuzione generatrice dei dati è alternativamente Frèchet o Gumbel, escludendo il caso Weibull, meno comune in studi idrologici, ambito nel quale i modelli con *a priori* di contrazione trovano notevole applicabilità.
L'adattamento del modello mediante tecniche MCMC ha permesso di ottenere campioni dei parametri a posteriori. Impostando una soglia di 0.05, i valori degli ξ_j a posteriori minori di tale soglia in valore assoluto sono stati ridotti a zero. Ciò ha condotto alla determinazione di probabilità per ciascun ξ_j di essere nullo, e i campioni con più del 50% di zeri sono stati identificati come appartenenti alla distribuzione di Gumbel. Confrontando questi risultati con i veri valori dei ξ_j, è stato possibile calcolare metriche diagnostiche per valutare l'accuratezza del modello nella distinzione tra le distribuzioni Gumbel e GEV generica. La Tabella 4.7 riporta tali metriche.

Errata Classificazione	**Accuratezza**	**Sensibilità**	**Specificità**
0.25	0.75	0.73	0.77

Tabella 4.7: Metriche di classificazione del modello per campioni dipendenti.

Le prestazioni del modello complesso si allineano a quelle osservate per i singoli campioni, evidenziando un miglioramento nelle metriche di sensibilità e specificità. L'impiego dell'*a priori* di contrazione Horseshoe con una contrazione globale del parametro ξ uguale per tutti i campioni sembra favorire un'assegnazione più accurata dei modelli alla distribuzione corrispondente.
L'analisi si estende anche allo studio dell'errore quadratico medio (EQM) nella stima dei parametri per ciascun campione, rispetto ai veri valori dei parametri. La Figura 4.6 illustra le prestazioni del modello in termini di EQM. I risultati evidenziano errori generalmente bassi, tendenti allo zero per i veri valori del parametro nulli, e leggermente più elevati per valori non nulli. Tuttavia, anche in presenza di deviazioni dal modello di Gumbel, la stima del parametro di forma non si discosta eccessivamente dal valore atteso.
Inoltre, è stata valutata la percentuale degli intervalli di confidenza HPD al 95% che includono il vero valore del parametro: 0.93 nel caso di parametri nulli e 0.83 per parametri

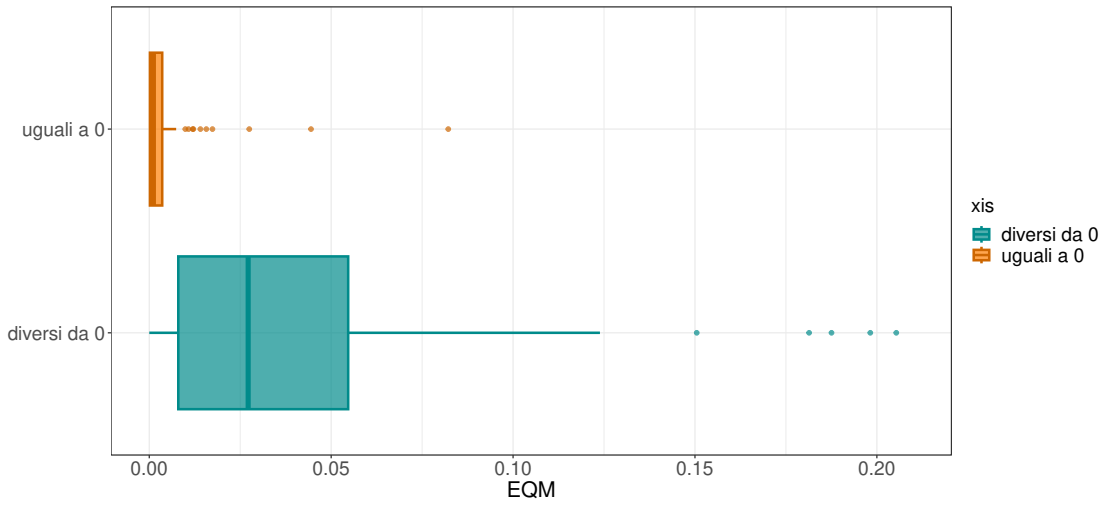

Figura 4.6: Errore quadratico medio per il parametro di forma di ogni campione simulato.

non nulli, con una proporzione complessiva dell'0.88.

Queste osservazioni confermano l'efficacia del modello bayesiano con *a priori* Horseshoe con contrazione globale comune, particolarmente vantaggioso in scenari in cui si considerano più campioni dipendenti provenienti dallo stesso territorio. Tale approccio è quindi particolarmente utile quando la struttura dei dati lo consente, suggerendo un potenziale miglioramento nella contrazione dei parametri di forma, come emerso dall'analisi complessiva dei campioni. Benché lo studio di simulazione condotto possa apparire preliminare, esso fornisce indicazioni preliminari positive sulla validità del modello in scenari realistici, come quelli che saranno esplorati in successive analisi applicate.

Capitolo 5

Analisi dei dati ARPA sulle piogge

Una scienza nella quale lo studio dei valori estremi risulta particolarmente rilevante è l'idrologia, la quale analizza il comportamento di corsi d'acqua, le precipitazioni e i tutti i fenomeni legati all'acqua, nonché la previsione di inondazioni e progettazione di infrastrutture idriche. Gli studi idrologici sono fondamentali per comprendere e gestire le risorse idriche e prevenire disastri legati all'acqua, attraverso anche lo studio dei valori estremi legati a fenomeni idrici.
In questo ambito ed in particolare nello studio delle precipitazioni i modelli proposti nella seguente tesi posso risultare particolarmente utili, in quanto la modellazione dei valori estremi legati alle piogge spesso vengono relegati all'utilizzo del solo modello Gumbel. Questa scelta è data anche dal fatto che spesso gli idrologi, attraverso vari studi assumono spesso che tale modello sia corretto, con eventuali ripercussioni catastrofiche, nel caso in cui i dati si distribuiscano come una Fréchet. Un' assunzione così forte infatti può portare a stime del livello di ritorno sbagliate e magari sottostimare il pericolo di possibili piogge abbondanti Coles et al. (2003). La scelta invece di modellare i dati attraverso la distribuzione generalizzata dei valori estremi, come già spiegato, non permette una buona estrapolazione in caso i dati siano distribuiti come una Gumbel.
Si propone quindi l'approccio Bayesiano con utilizzo di *a priori* di contrazione per il parametro di forma per i massimi annuali delle piogge, così da incorporare l'informazione data dagli idrologi ma permettendo comunque al modello di modellare delle distribuzioni differenti dalla Gumbel in caso i dati portino a tali conclusioni.
I modelli vengono applicati ad uno studio sui dati delle piogge ottenuti dall'Agenzia Regionale per la Prevenzione e Protezione Ambientale. In particolare sono disponibili le rilevazioni delle quantità di acqua delle piogge, rilevate dai pluviometri situati in una vasta area del nord Italia, ad ogni ora. L'area di studio si trova in particolare nell'area nord orientale dell'Italia, si estende per circa $32000 km^2$ e comprende la regione Veneto e le provincie di Bolzano e Trento. L'area è caratterizzata da un'elevata eterogeneità climatica e morfologica del territorio, con determinate zone montuose, collinose e pianeggianti.
Lo studio in particolare riguarda osservazioni continue sulla pioggia in millimetri, aggre-

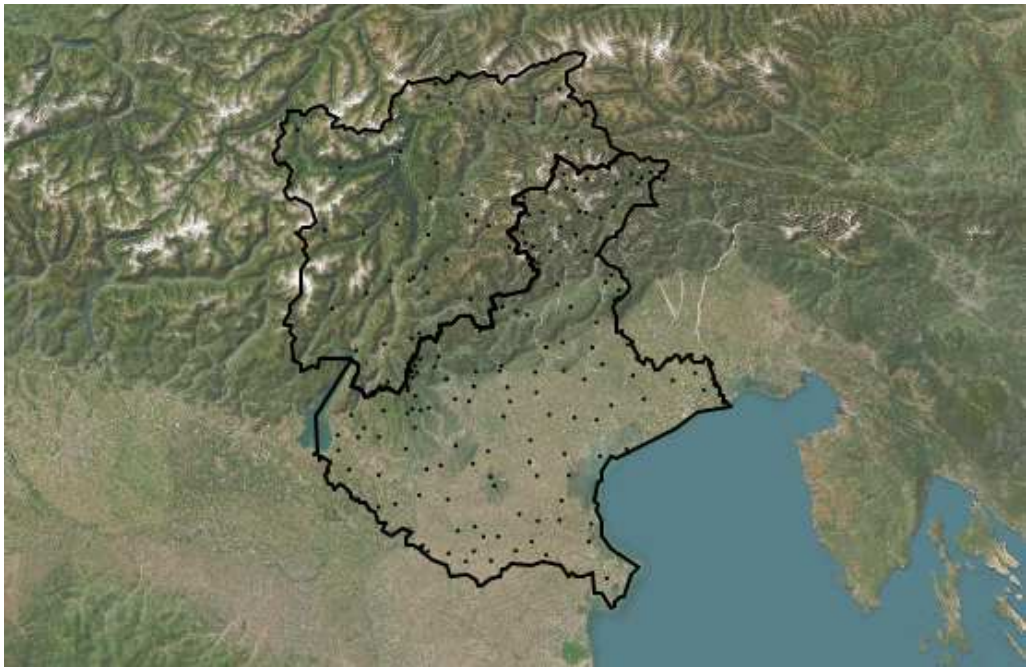

Figura 5.1: Mappa dell'area d'interesse dei dati, con le posizioni delle stazioni di rilevazione.

gate per ogni ora, raccolte presso 174 pluviometri riscaldati situati in tutto il territorio di studio. Sono inoltre presenti le coordinate di ciascuna stazione di rilevazione e l'altitudine. Le serie orarie non coprono tutte lo stesso intervallo temporale ma differiscono per ogni stazione con un periodo di rilevazione totale che varia da un minimo di 14 anni e un massimo di 37 anni.

Innanzitutto per effettuare un analisi dei valori estremi è necessario ottenere le osservazioni necessarie e si sceglie, seguendo quanto viene spesso fatto in analisi di questo tipo, di ottenere i campioni di massimi annuali per ciascuna stazione. Si ottiene così un insieme di osservazioni che identificano i massimi annuali per ciascuna stazione di rilevazione, con numerosità differenti a seconda dei vari periodi di rilevazione. Con tali osservazioni viene effettuata un'analisi bayesiana sulle piogge nelle regioni d'interesse attraverso 3 modelli. Si definisce la verosimiglianza per l'insieme di osservazioni delle diverse stazioni e la distribuzione a priori per la completa specificazione del modello:

$$L(\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\xi}; \mathbf{Z}) = \prod_{j=1}^{174} L(\mu_j, \sigma_j, \xi_j; \mathbf{Z}_j),$$

$$\pi(\mu_j, \phi_j, \xi_j) = \pi(\mu_j)\pi(\phi_j)\pi(\xi_j),$$

con $\phi_j = \log(\sigma_j)$ e una specificazione diversa delle distribuzioni a priori per ciascuno dei tre modelli. La distribuzione a priori nei primi due modelli per i parametri di posizione μ_j e il logaritmo dei parametri di scala ϕ_j è una normale centrata in zero con varianza diffusa pari a 10^4. È nella specificazione della distribuzione a priori per il parametro di forma che i due modelli differiscono.

Nel primo modello infatti i parametri di forma seguono *a priori* di contrazione Horseshoe nella quale l'iperparametro differisce per ciascuna stazione di rilevazione:

$$\begin{aligned}(\xi_j|\lambda_j,\tau_j) &\sim N(0,\lambda_j^2\tau_j^2),\\ \lambda_j &\sim C^+(0,1),\\ \tau_j &= 1/n_j\sqrt{\log(n_j)},\end{aligned} \tag{5.1}$$

con n_j la numerosità del campione $j-$esimo. Ogni gruppo di osservazioni quindi presenta una varianza che è combinazione di una quantità stocastica ed una deterministica differente per ogni stazione. Tale quantità aumenta al diminuire della numerosità del campione delle singole stazioni, imponendo una quantità di contrazione di base minore se la quantità di dati a disposizione per il singolo gruppo è bassa. Questo tipo di modellazione quindi considera singolarmente le stazioni senza tenere in considerazione in nessun modo la presenza delle altre, come descritto in sezione 3.2.
Il secondo modello invece corrisponde al caso gerarchico descritto in sezione 3.4 che vede una specificazione del tipo :

$$\begin{aligned}(\xi_j|\lambda_j,\tau) &\sim N(0,\lambda_j^2\tau^2),\\ \lambda_j &\sim C^+(0,1),\\ \tau &= 174/N\sqrt{\log(N/174)},\end{aligned} \tag{5.2}$$

con N la numerosità totale del campione per tutte le stazioni osservate. Questo tipo di modellazione permettere di fare un'assunzione che tenga conto della presenza di tutte le stazioni di rilevazione, ponendo a priori una quantità di contrazione globale uguale per tutti i gruppi di osservazioni e una parte che cambia tra gruppo e gruppo definita da λ_j.
La scelta della Horseshoe in entrambi i casi è dovuta al fatto che permette questo tipo di modifica del modello e soprattutto al fatto che è risultato essere il più efficace tra quelli proposti dallo studio di simulazione.
L'ultimo modello utilizzato infine prevede una modifica al precedente che permetta di incorporare alle stime dell'informazione relativa alle stazioni di rilevazione dei pluviometri. Per ognuna di queste infatti sono disponibili la locazione geografica in termini di latitudine e longitudine e l'altitudine misurati in metri sopra il livello del mare. Tali informazioni vengono incorporate per modellare il parametro di locazione e scala di ciascuna stazione, dato che comunque è noto che la quantità di acqua che cade durante le piogge è determinata anche dalle caratteristiche del luogo considerato. Da notare che le stime dei parametri della distribuzione generalizzata dei valori estremi risultano correlate tra loro e un'errata stima della media o della varianza potrebbe portare delle gravi conseguenze nella stima del parametro di forma.
Per non perdere quindi troppa informazione e permettere alle singole stazioni di presentare stime del parametro di posizione e scala differenti dalla relazione lineare con le covariate aggiuntive si aggiunge alle variabili anche l'identificativo delle stazioni. Data la verosimiglianza definita precedentemente, si definisce la matrice $\mathbf{X}$ di dimensione $174 \times (3+p)$ con

$p = 173$ di variabili riferite alle stazioni, dove le prime 3 colonne sono riferite in sequenza all'altitudine, alla longitudine e all'altitudine mentre le altre sono indicatrici dell'appartenenza alle stazioni. Si nota che il numero di indicatrici è minore di 1 rispetto al numero di gruppi in quanto l'informazione del primo gruppo è contenuta nell'intercetta per ottenere un modello identificabile. Il modello modificato assume che i parametri di scala e posizione siano il risultato di una combinazione lineare delle caratteristiche della stazione:

$$\mu_j = \beta_0 + \beta_1 x_{j,1} + \beta_2 x_{j,2} + \beta_3 x_{j,3} + \ldots + \beta_p x_{j,p},$$
$$\sigma_j = \exp\left\{\alpha_0 + \alpha_1 x_{j,1} + \alpha_2 x_{j,2} + \alpha_3 x_{j,3} + \ldots + \alpha_p x_{j,p}\right\},$$

con β_0 e α_0 rispettivamente le intercette per la relazione di μ e σ con le covariate. L'utilizzo dell'esponenziale nel caso del parametro σ permette di rispettare il vincolo di positività di quest'ultimo. Risulta quindi necessario definire le distribuzioni a priori per i parametri d'interesse, scegliendo delle *a priori* gaussiane non informative centrate in zero con varianza 10^4 per i β_j e α_j con $j = 0, 1, 2, 3$. Per i parametri ξ_j invece si mantiene invariata la struttura della distribuzione a priori Horseshoe definita in (5.2).
Ognuno dei modelli descritti propone una struttura sempre più complessa rispetto al precedente provando ad integrare quanta più informazione dai dati osservati. Oltre alla stima dei modelli, viene anche verificata la loro validità attraverso l'utilizzo del livello di ritorno. Non conoscendo però il modello generatore dei dati in questo caso risulta conveniente il confronto delle stime di tali quantità rispetto al livello di ritorno empirico. Questo si ottiene attraverso il calcolo della funzione di ripartizione empirica dei valori osservati, ed essendo in un caso in cui il numero di campioni a disposizione è elevato risulta possibile effettuare una valutazione della bontà totale del modello oppure della validità per ciascun campione. In Figura 5.2 viene illustrato un esempio della curva del livello di ritorno stimata, insieme al rispettivo intervallo di credibilità e il livello di ritorno empirico.

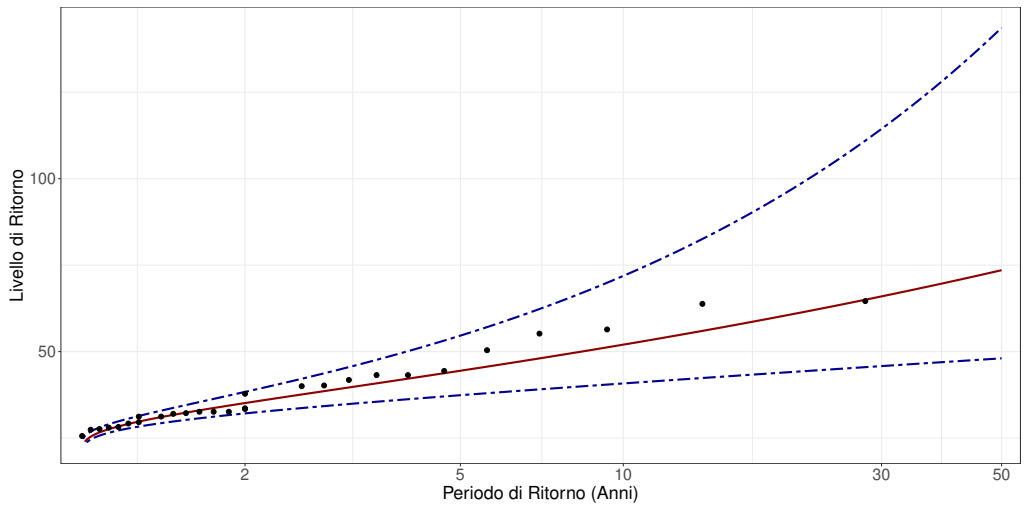

Figura 5.2: Esempio di livelli di ritorno empirici (pallini neri) e stimati (linea rossa) con rispettivi intervalli di credibilità (linee tratteggiate blue) con $\alpha = 0.05$.

Vengono calcolate quindi due metriche con tali quantità, che ci permettono di comprendere l'adattamento dei modelli ai dati disponibili. Innanzitutto viene calcolato lo scarto quadratico medio dei livelli di ritorno stimati rispetto a quelli empirici per ciascuna stazione, sulla falsa riga di quanto calcolato durante le simulazioni precedentemente fatte. Si sceglie quindi di calcolare l'errore di stima pesandolo anche rispetto all'importanza del livello di ritorno d'interesse.
Tale quantità infatti nel caso di valori estremi risulta essere interessante solo in relazione a periodi di ritorno elevati, piuttosto che per periodi brevi. Tale caratteristica vuole essere considerata in un indice di bontà del modello in questo caso e questo può essere fatto attraverso l'utilizzo di una modifica della convergenza di Kullback-Leibler. Per fare questo si lavora direttamente con le probabilità associate ai valori di ritorno, che corrispondono all'inverso del periodo di ritorno in anni. Si definisce quindi $P(z)$ l'inverso del periodo di ritorno corrispondente al livello di ritorno empirico z, definita come la probabilità che un evento estrema assuma valori maggiori di z_i. Definendo quindi $Q(z)$ come la stima dell'inverso del periodo di ritorno per il valore z, si propone per ciascun campione osservato:

$$KL_j = \sum_{i=1}^{n_j} \frac{1}{P(z_{i;j})} \log\left(\frac{P(z_{i;j})}{Q(z_{i;j})}\right) \qquad \text{per, } j = 1, \ldots, 174.$$

Così facendo tale metrica permette di verificare l'informazione persa nella stima del livello di ritorno empirico attraverso il modello, penalizzando però maggiormente gli errori corrispondenti a livelli di ritorno legati a periodi di ritorno elevati. I modelli presentano dei risultati molto simili tra loro, anche se il modello congiunto sembrerebbe mostrare degli errori nella stima del livello di ritorno più variabili rispetto agli altri due modelli proposti e mediamente più alti.

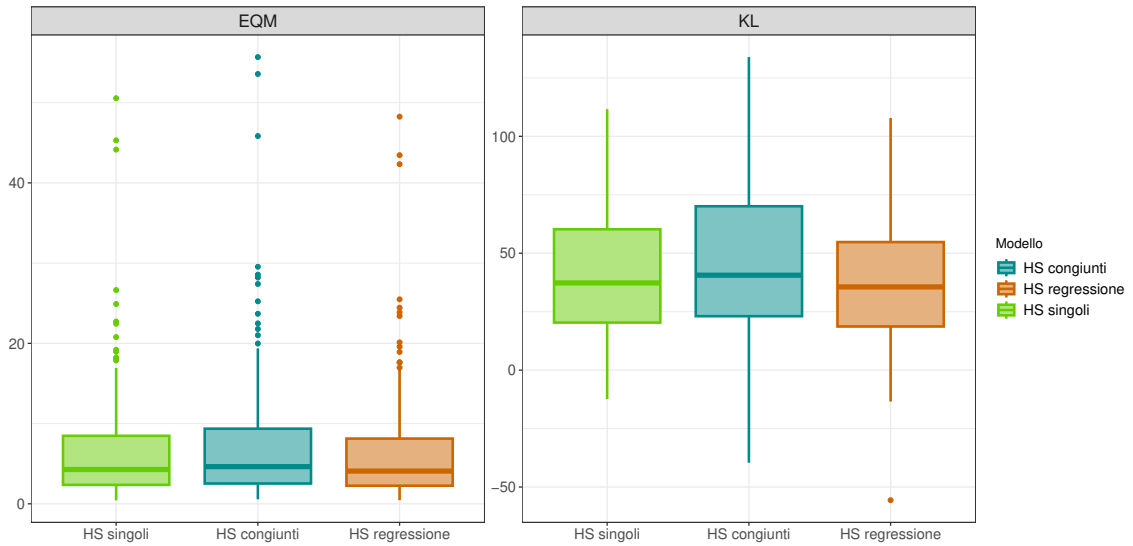

Figura 5.3: Boxplot dell'errore quadratico medio e della distanza di Kullback-Leibler modificata, dei livelli di ritorno stimati rispetto a quelli empirici per le 174 stazioni.

Il modello con *a priori* di contrazione congiunto con l'aggiunta delle covariate sembra essere il migliore, portando ad ottenere delle stime migliori dei livelli di ritorno sia in termini di errore quadratico che di distanza di Kullback-Leibler con penalizzazione maggiore per valori più estremi.
Dopo aver analizzato la bontà delle stime si osservano quindi i risultati ottenuti dai modelli. Innanzitutto si osservano le stime dei modelli, ed in particolare la categorizzazione delle stazioni rispetto alle distribuzioni della famiglia GEV come in Tabella 5.1.

	Fréchet	Weibull	Gumbel
Horseshoe singoli	63	16	95
Horseshoe congiunti	45	9	120
Horseshoe regressione	55	12	107

Tabella 5.1: Tabella delle frequenze di stazioni i cui campioni vengono categorizzati rispettivamente come distribuzioni Fréchet, Weibull o Gumbel.

Si nota che modellando ciascuna stazione di rilevazione singolarmente, imponendo una contrazione legata solo alla numerosità del singolo campione sembrerebbe portare una contrazione a zero minore rispetto agli altri casi. Il modello congiunto sembra invece portare delle stime dei parametri di forma più vicini allo zero mentre aggiungendo le informazioni legate ad altitudine e longitudine si osserva un effetto che sta a metà tra i due modelli. Nella maggior parte dei territori comunque il comportamento dei valori estremi sembra seguire la distribuzione Gumbel piuttosto che una generica GEV, con una buona frequenza di Fréchet e una bassa numerosità di Weibull.
I risultati del modello vengono illustrati in Figura 5.4 , nelle quali vengono rappresentate le stime del modello per ciascuna delle 147 stazioni di rilevazione. Le gradazioni di blue e azzurro dei punti definisce il valore del parametro di posizione stimato, con un gradiente vicino al blu scuro per piogge mediamente più intense e azzurro per le meno intense; la grandezza del diametro rappresenta la stima del parametro di scala con dimensioni minori per piogge meno variabili e maggiori per le più variabili. Infine il tipo di distribuzione viene rappresentato dai punti colorati al centro dei punti grigi, con un colore verde si indica una distribuzione di tipo Weibull, il bianco indica la Gumbel e i punti rossi indicano la Fréchet. Per quest'ultima distribuzione inoltre viene rappresentato anche il valore del parametro di forma , con una trasparenza maggiore per valori più vicini allo zero che però diminuisce man mano che il valore del parametro aumenta.
Si nota che nelle zone della pianura padana le piogge sembrano essere mediamente più intense rispetto ad altre zone, con dei picchi di intensità nella zona settentrionale del Veneto. Queste zone presentano inoltre una grande variabilità con la peculiarità però di seguire nella maggioranza una distribuzione Gumbel, il che comporta una probabilità di osservare dei valori estremi particolarmente inusuali, non particolarmente elevata. Delle zone che invece sembrano particolarmente a rischio in quanto presentano delle piogge con

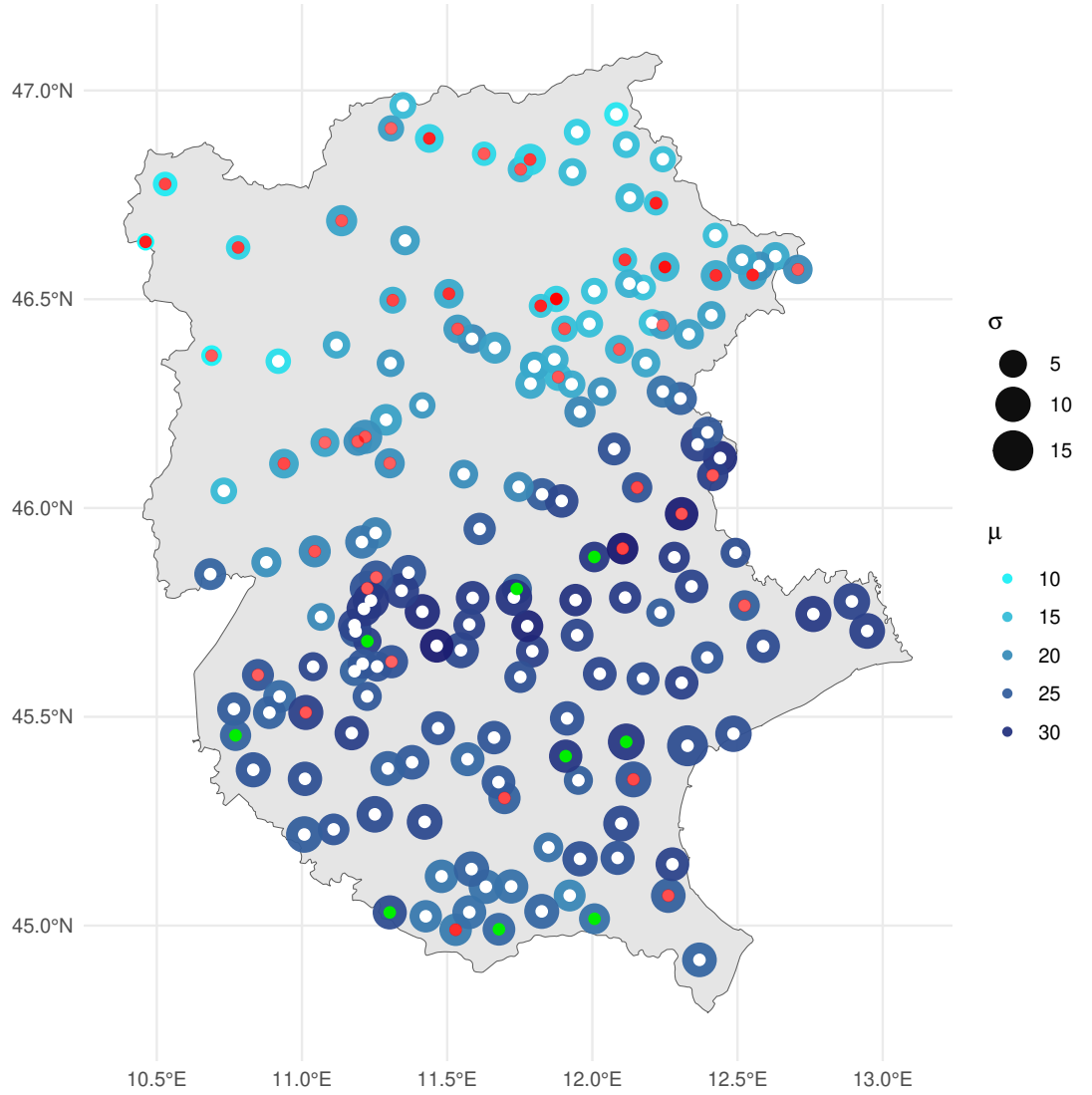

Figura 5.4: Stime modello con *a priori* Horseshoe e parametro di contrazione globale comune tra stazioni. La gradazione di blu e azzurro rappresenta le stime del parametro μ, la grandezza dei punti la stima di σ e il punto più piccolo al centro dei punti indica il tipo di distribuzione: Weibull (Verde), Gumbel (Bianco) e Fréchet (Rosso).

una media intensità alta e soprattutto che seguono una distribuzione di Fréchet, sono Verona insieme a determinate aree in prossimità del lago di Garda e a sud della catena del monte Baldo e le zone in prossimità di Belluno. In queste zone inoltre sembrerebbe dalle stime ottenute che oltre a piogge mediamente intense si verifichino comunemente delle piogge inconsuete e di una straordinaria intensità.
Nelle zone del Trentino-Alto Adige si osserva invece un andamento dei massimi annuali delle piogge molto differente. Si osserva infatti che mediamente in tali zone il massimo registrato delle precipitazioni mediamente non presenta valori particolarmente elevati e inoltre si osserva una varianza minore rispetto alla regione Veneto. Nonostante questo comportamento, si registra un gran numero di stazioni le cui osservazioni vengono categorizzate

nel modello Fréchet, con quindi un'alta probabilità del verificarsi di piogge particolarmente estreme. La conformazione del territorio, che risulta principalmente montuoso sembra quindi portare ad osservare dei comportamenti molto instabili delle precipitazioni, le quali pur non essendo troppo intense possono registrare degli intensi aumenti repentini.
Sembrerebbe quindi anche da queste analisi utile, stimare i parametri di scala e posizione in funzione della posizione geografica in cui ci si trova e l'altitudine della stazione di rilevazione, così da provare a migliorare le stime e ottenere delle informazioni riguardanti le caratteristiche delle precipitazioni massime annuali rispetto alla conformazione del territorio. Per fare questo il terzo modello proposto precedentemente risulta particolarmente utile. Tale complicazione del modello risulta inoltre soddisfacente in termini di bontà di adattamento dei dati, in quanto sembra portare un miglioramento nelle stime oltre che interpretativo.
In Tabella 5.2 si mostrano le stime ottenute dai modelli con i relativi intervalli di credibilità con probabilità 0.95 così da comprendere se l'effetto è risulta davvero significativo.

	μ		$\log \sigma$	
	Coef.	I.C.	Coef.	I.C.
Longitudine	0.217	(-0.407; 0.539)	0.021	(-0.071; 0.154)
Latitudine	0.027	(0.007; 0.061)	0.034	(0.018; 0.067)
Altitudine	-3.533	(-6.109; -0.981)	-0.622	(-0.850; -0.405)

Tabella 5.2: Stime del modello con *a priori* Horseshoe e covariate per il parametro di scala e posizione delle distribuzioni GEV

I risultati mostrano che la longitudine non sembra influire sul livello medio di piogge massime annuali e nemmeno sulla variabilità di queste mentre l'altitudine sembra avere un effetto positivo su entrambi. L'altitudine invece ha un effetto negativo su entrambi i parametri, come si poteva evincere dalle precedenti analisi.
Un effetto positivo della latitudine sembrerebbe inaspettato dai risultati ottenuti dal modello precedente e da quanto osservato in Figura 5.4, ma tale assunzione fuorviante è dovuta al fatto che nelle regioni più a Nord sono anche in questo caso quelle con altitudini maggiori. L'aumento della media di acqua nelle precipitazioni che sembrava essere legato all'aumento della latitudine invece ha un forte legame con l'altitudine della zona, e tale relazione è stato possibile osservarla solo complicando il modello ulteriormente ed utilizzando anche le covariate disponibili.
In Figura 5.5 viene mostrata la mappa del veneto e del Trentino-Alto Adige, colorata secondo la stima del modello bayesiano che utilizza le variabili delle stazioni. In questo caso si mostrano solo le stime ottenute attraverso la combinazione lineare delle stime rispetto alle variabili latitudine, longitudine e altitudine della mappa. Dalla figura si osserva quanto sottolineato precedentemente, con dei valori medi dei massimi annuali che diminuisce a seconda dell'altitudine registrata nel territorio, con la regione Veneto che risulta essere un

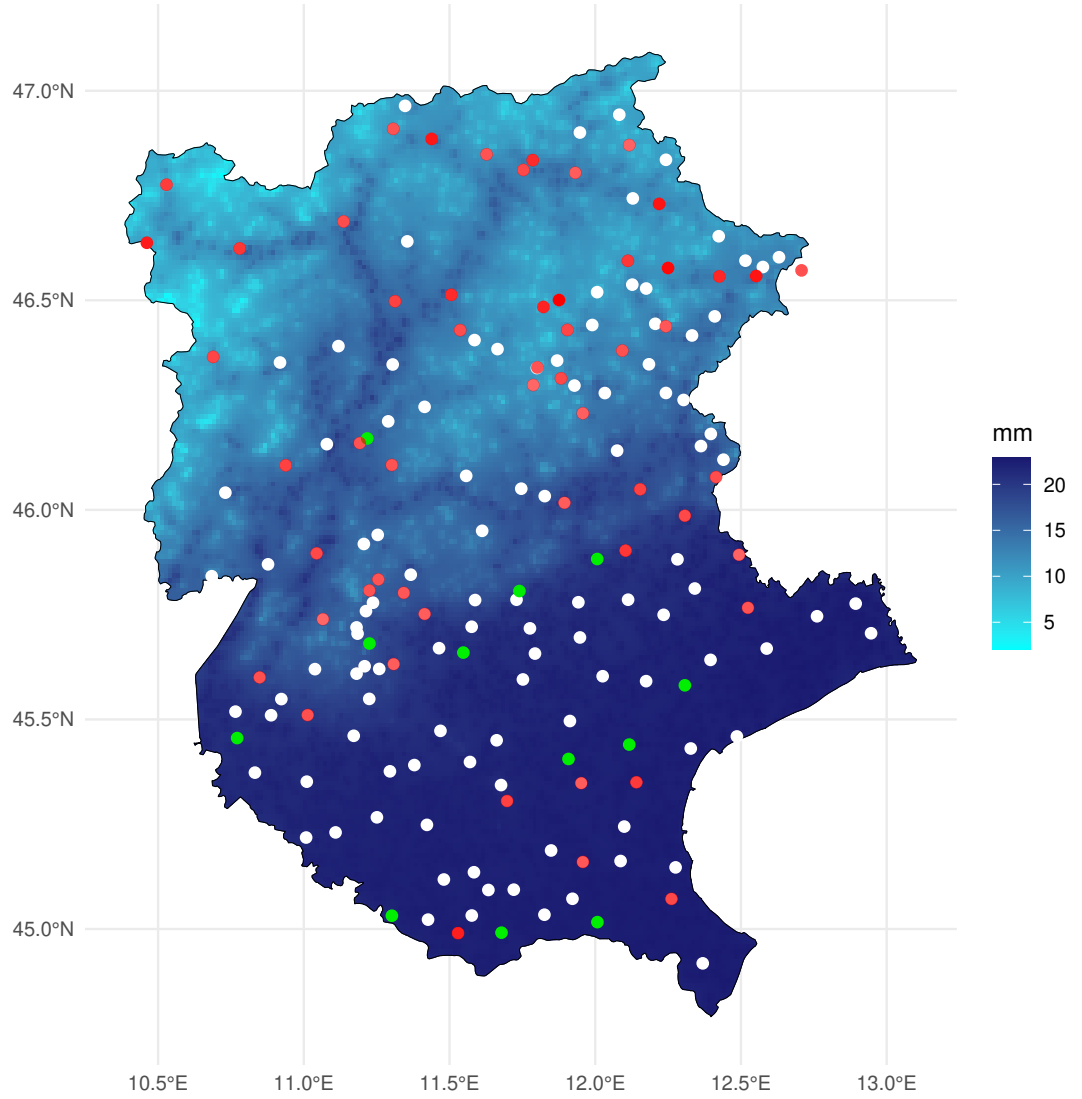

Figura 5.5: Stima della media ottenuta dal Modello con *a priori* Horseshoe e parametro di contrazione globale comune alle stazioni con utilizzo delle covariate per la stima di μ e σ. La gradazione di blu e azzurro rappresenta la media di precipitazioni stimata per ogni locazione geografica rispetto a latitudine, longitudine e altitudine: $\hat{\beta}_0 + \hat{\beta}_1 x_{j,1} + \hat{\beta}_2 x_{j,2} + \hat{\beta}_3 x_{j,3}$.
I Punti definiscono definiscono la distribuzione stimata per le stazioni di rilevazione: Weibull (Verde), Gumbel (Bianco) e Fréchet (Rosso).

luogo in cui le piogge mediamente risultano essere molto intense però con poca probabilità di osservare estremi particolarmente anomali. Il Trentino-Alto Adige in contrapposizione presenta mediamente dei massimi annuali delle precipitazioni meno elevati, con una buona probabilità di eventi particolarmente elevati rispetto la media; comportamento dovuto alle alte quote del territorio. Oltre ad una descrizione delle stime nel territorio infine vengono mostrati i livelli di ritorno relativi a 50 e 1000 anni in Figura 5.6. Nel breve periodo, si nota che il livello di ritorno delle precipitazioni nel Veneto è superiore rispetto al Trentino-Alto Adige. Questo è dovuto al fatto che, nell'interpolazione a breve

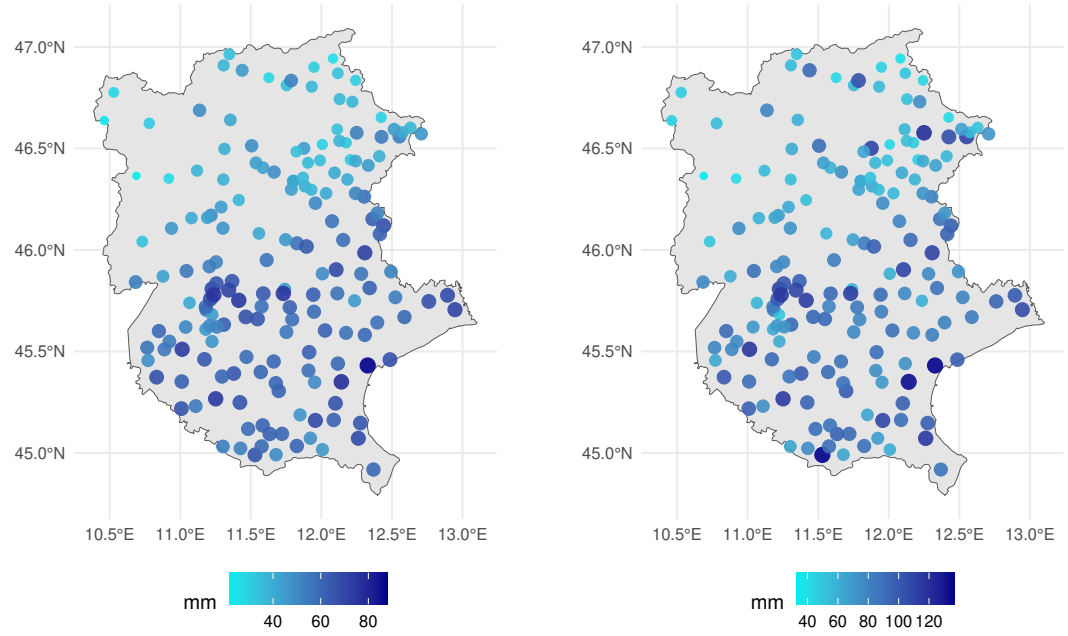

Figura 5.6: Stime del livello di ritorno per ogni stazione di rilevazione relative al periodo di ritorno di 50 anni (Figura a destra) e 1000 anni (Figurata a sinistra).

distanza temporale, la media e la varianza della distribuzione GEV risultano prevalenti. Tuttavia, valutando il livello di ritorno per periodi più lunghi, il cui valore quindi viene superato con probabilità molto bassa, l'influenza di code più pesanti, come nel caso della Fréchet, diventa determinante. Quando si considerano periodi molto lunghi, è evidente che la discrepanza nei livelli di ritorno tra le stazioni del Trentino-Alto Adige e del Veneto si riduce. Inoltre in alcune aree settentrionali, dove erano state registrate stime elevate del parametro di forma, le stime superano persino quelle delle zone più meridionali, le cui distribuzioni stimate sono risultate per la maggior parte Gumbel.

Al termine dell'analisi condotta sui dati ARPA relativi alle precipitazioni, si può affermare con fiducia che i modelli proposti offrono una modellazione adeguata e accurata di tali dati ed una buona interpretazione dei risultati. L'utilizzo di *a priori* di contrazione per il parametro di forma della distribuzione GEV ha permesso di identificare le zone in cui i massimi annuali delle piogge hanno un comportamento che segue effettivamente la distribuzione Gumbel, discriminandoli da territori con valori del parametro di forma positivi o negativi. L'utilizzo inoltre del modello congiunto permette di tenere in considerazione la presenza di tutte le stazioni del territorio d'interesse e l'utilizzo delle covariate di identificare un legame tra caratteristiche morfologiche e la media e la variabilità dei massimi annuali delle precipitazioni. Questo fornisce una base per ulteriori considerazioni e applicazioni nell'ambito idrologico.

Conclusioni

Nella presente tesi, è stata studiata l'importanza dell'analisi statistica dei valori estremi, specialmente nell'ambito degli studi idrologici. Si è fornita una panoramica delle diverse metodologie esistenti e si è proposto un nuovo metodo di analisi bayesiana che sfrutta le distribuzioni a priori informative. Si è posto particolare interesse sulla distribuzione generalizzata dei valori estremi (GEV), con un'attenzione specifica al parametro di forma, ξ, e alle implicazioni delle assunzioni su di esso. Questo parametro è cruciale per descrivere il comportamento dei valori estremi e, in base al suo valore, definisce tre distinti tipi di comportamenti asintotici. Si è concentrata l'attenzione sul caso in cui ξ è zero, che corrispondente alla distribuzione Gumbel, un caso limite all'interno della GEV. Nonostante in alcuni ambiti di ricerca la distribuzione Gumbel non sia considerata, preferendo l'uso diretto della GEV come modello unico, riducendo tale distribuzione ad un singolo punto in uno spazio parametrico continuo, in ambito idrologico la Gumbel è spesso preferita per le sue proprietà vantaggiose.
È stato proposto un approccio che bilancia questi due metodi, utilizzando il paradigma bayesiano. In particolare, sono state adottate distribuzioni a priori di contrazione che inseriscono informazioni a priori sul possibile valore di ξ, assegnando una notevole massa probabilistica allo zero ma permettendo ai dati di influenzare tale assunzione. Tra le distribuzioni a priori di contrazione esaminate, quali Horseshoe, Laplace e le varianti Spike and Slab (sia continue che discrete), si è condotto uno studio di simulazione per valutare le loro performance in diversi scenari, sia per variazioni nella numerosità campionaria che nel vero valore di ξ. L'obiettivo era che i modelli sviluppati potessero chiaramente discriminare il caso Gumbel quando questo riflette la realtà dei dati, senza comprimere eccessivamente il parametro di forma della GEV quando è diverso da zero.
I metodi introdotti sono stati confrontati con il tradizionale approccio bayesiano non informativo, il quale adotta una distribuzione a priori non informativa che lascia il parametro ξ libero di variare senza restrizioni imposte a priori. I risultati hanno mostrato che quando il valore reale del parametro ξ è zero, generalmente i metodi proposti superano in prestazione il modello bayesiano non informativo. Invece, quando ξ differisce da zero, le prestazioni delle metodologie introdotte si mostrano sostanzialmente equiparabili a quelle del modello con *a priori* diffusa. Le versioni discrete del modello Spike and Slab hanno evidenziato certe limitazioni, in particolar modo quando il vero modello generativo dei dati non corrisponde

a quello di Gumbel, mostrando prestazioni non particolarmente soddisfacenti nel caso del modello Weibull. Per quanto riguarda il modello con la distribuzione a priori di Laplace, nonostante offra stime ragionevoli, sembra non facilitare un'efficace discriminazione tra i diversi tipi di modelli, specialmente quando la dimensione del campione è limitata. Si è notato infine che gli approcci che utilizzano le distribuzioni a priori Horseshoe e Spike and Slab continua tendono a fornire stime dei parametri accurate, riducendone la variabilità e migliorando la discriminazione tra il modello Gumbel e la GEV generica. In particolare, la Horseshoe si è distinta per le sue performance promettenti in campioni di piccole dimensioni, una situazione comune negli studi di valori estremi, proponendosi come la proposta più efficace.

La necessità di trattare dati correlati, come quelli tipici degli studi idrologici che provengono da diverse località all'interno di uno stesso territorio, ha stimolato la formulazione di un modello capace di riconoscere una struttura gerarchica nei dati. Questo approccio prevede l'analisi individuale dei campioni per le stime dei parametri di scala e posizione, mantenendo al contempo per la stima degli ξ un parametro di contrazione globale condiviso per tutti i gruppi di osservazioni. Il modello utilizza la distribuzione a priori Horseshoe, che offre parametri di contrazione sia globali sia locali, rendendolo particolarmente adatto per la modellazione richiesta, come confermato dai risultati positivi ottenuti nello studio di simulazione. Questa metodologia consente di ottenere una compressione dei dati più uniforme e coerente.

Per mostrare un'applicazione pratica dei modelli proposti, si è condotta un'analisi sui dati reali delle precipitazioni forniti dall'ARPA, relativi a misurazioni orarie raccolte in diverse stazioni meteorologiche del Veneto e del Trentino-Alto Adige. L'analisi si è concentrata sui massimi annuali di ciascuna stazione, creando così un insieme di campioni rappresentativi dei picchi di precipitazioni. Tre diversi approcci di modellazione sono stati applicati a questi dati. Inizialmente, ogni stazione è stata analizzata separatamente, seguita dall'impiego del modello congiunto che integra una componente di contrazione globale. Infine, questo ultimo modello è stato adattato per includere informazioni geografiche specifiche, come coordinate e altitudine. Le stime hanno rivelato differenze significative nel comportamento delle precipitazioni attribuibili sia all'altitudine che alla posizione geografica, evidenziando distinzioni nette tra aree montuose e zone pianeggianti o collinari.

In sintesi, questa tesi ha presentato una prospettiva alternativa sulla modellazione bayesiana dei valori estremi, proponendo un nuovo strumento applicabile nel campo idrologico, che possa offrire vantaggi interpretativi e predittivi sui valori estremi futuri. Gli strumenti sviluppati risultano essere soddisfacenti, in particolare quando si utilizza l'*a priori* Horseshoe, e come si è osservato possono essere adattati per beneficiare di progressi futuri o modifiche basate su ricerche successive. Nonostante i risultati incoraggianti, si raccomanda prudenza nell'uso dei modelli, data l'importanza cruciale dell'accuratezza nelle analisi in questo ambito, dove errori di stima potrebbero portare a sottostimare il rischio di catastrofi naturali future.

Bibliografia

Carvalho, C. M., Polson, N. G. & Scott, J. G. (2009). Handling Sparsity via the Horseshoe. In *Proceedings of the Twelth International Conference on Artificial Intelligence and Statistics*, D. van Dyk & M. Welling, eds., vol. 5 of *Proceedings of Machine Learning Research*. Hilton Clearwater Beach Resort, Clearwater Beach, Florida USA: PMLR.

Carvalho, C. M., Polson, N. G. & Scott, J. G. (2010). The Horseshoe Estimator for Sparse Signals. *Biometrika* **97**, 465–480.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement* **20**, 37–46.

Coles, S. (2001). *An Introduction to Statistical Modeling of Extreme Values*. Springer Series in Statistics. London: Springer-Verlag.

Coles, S., Pericchi, L. R. & Sisson, S. (2003). A fully probabilistic approach to extreme rainfall modeling. *Journal of Hydrology* **273**, 35–50.

Davison, A. C. & Smith, R. L. (1990). Models for Exceedances Over High Thresholds. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **52**, 393–425.

Gamerman, D. & Lopes, H. F. (2006). *Markov chain Monte Carlo: stochastic simulation for Bayesian inference*. CRC press.

Ghosal, S. & Van der Vaart, A. (2017). *Fundamentals of Nonparametric Bayesian Inference*, vol. 44. Cambridge University Press.

Gomes, M. I. (1987). Extreme value theory—statistical choice. In *Colloq. Math. Soc. János Bolyai*, vol. 45.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**, 97–109.

Hosking, J. R. (1984). Testing whether the shape parameter is zero in the generalized extreme-value distribution. *Biometrika* **71**, 367–374.

Johnson, V. E. & Rossell, D. (2012). Bayesian Model Selection in High-dimensional Settings. *Journal of the American Statistical Association* **107**, 649–660.

Narisetty, N. N. (2020). Chapter 4 - Bayesian Model Selection for High-Dimensional Data. In *Principles and Methods for Data Science*, A. S. Srinivasa Rao & C. Rao, eds., vol. 43 of *Handbook of Statistics*. Elsevier, pp. 207–248.

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

Ročková, V. & George, E. I. (2018). The Spike-and-Slab Lasso. *Journal of the American Statistical Association* **113**, 431–444.

Smith, R. L. (1985). Maximum likelihood estimation in a class of nonregular cases. *Biometrika* **72**, 67–90.

Smith, R. L. (1987). Approximations in extreme value theory. *Preprint, Univ. North-Carolina* .

Stephenson, A. & Tawn, J. (2004). Bayesian Inference for Extremes: Accounting for the Three Extremal types. *Extremes* **7**, 291–307.

Tadesse, M. G. & Vannucci, M. (2021). Handbook of Bayesian Variable Selection .

Van der Pas, S., Szabó, B. & Van der Vaart, A. (2017). Adaptive posterior contraction rates for the horseshoe .