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

Analisi delle curve di mortalità prematura e senescente: classificazione dei paesi attraverso Model Based Clustering

Relatore:
Stefano Mazzuco

Laureando:
Riccardo Barato
Matricola N. 2055165

Anno accademico 2023/2024

Indice

Introduzione e obbiettivi della tesi

Monitorare i tassi di mortalità nei paesi, è un'attività di fondamentale importanza per una serie di ragioni cruciali che vanno ben oltre il semplice interesse statistico.
I dati raccolti, forniscono un quadro prezioso sulla salute e il benessere di una popolazione e costituiscono uno strumento imprescindibile per la pianificazione ospedaliera, la valutazione dell'efficacia delle politiche pubbliche e la risposta alle emergenze sanitarie. Il monitoraggio dei tassi di mortalità è inoltre, essenziale per identificare cambiamenti demografici, tendenze e fattori di rischio che possono influenzare la salute delle comunità. In ambito demografico, per monitorare e valutare le performance di un paese e della sua popolazione, si fa largo uso di metriche quantitative. Tra queste, il livello di mortalità prematura rappresenta un importante indicatore, poiché permette di catturare una specifica dimensione del carico di decessi all'interno di un paese, che possono essere etichettati come non necessari o, in qualche maniera, evitabili. Tuttavia, nonostante le sfaccettature chiave che questo concetto riveste, ottenere una sua misurazione chiara e precisa risulta tutt'altro che semplice. La letteratura sottolinea come l'attuazione pratica di questo concetto possa esercitare un'influenza significativa sulle decisioni e sulle scelte del paese sotto osservazione. Questa difficoltà è dovuta alla natura latente del concetto di mortalità prematura. Non risulta infatti, una caratteristica osservabile come evento singolo, ma è un concetto che emerge dall'analisi di dati relativi alla mortalità in una popolazione in senso generale. Di conseguenza, caratteristiche come l'aspettativa media di vita di una popolazione, indicatori economici come reddito, livello di istruzione e occupazione e indici di qualità della vita percepita, concorrono tutti a definire un valore operativo e affidabile per la misura di interesse.
Ad oggi, la letteratura demografica distingue la mortalità naturale da quella prematura attraverso approcci non sempre privi di svantaggi o limitazioni, affidandosi a soglie numeriche o a metodologie che non offrono sempre un'assoluta distinzione tra le due componenti all'interno di uno stesso paese. L'obbiettivo della tesi, è quello di superare queste limitazioni proponendo un metodo con il quale identificare gruppi di paesi simili, sulla base della naturale distribuzione delle morti naturali e solo in seguito, definire la componente di mortalità prematura specifica per ogni paese.

Questo metodo avanzato offre una visione più dettagliata e adattabile rispetto ai tradizionali approcci statistici, consentendo una più profonda comprensione delle componenti di mortalità, basata su informazioni a priori e sull'identificazione di gruppi di paesi simili nella naturale curva dei decessi per anzianità.
La tesi è articolata come segue.
Il primo Capitolo è dedicato alla presentazione del concetto di mortalità prematura e dell'importanza che questa misura riveste nei contesti sociali e politici all'interno dell'organizzazione di un paese. Viene presentata sotto due specifici punti di vista, sottolineando per ognuno i limiti e i relativi punti di forza. Il Capitolo 2, vede delinearsi il metodo con il quale è stato affrontato il problema, oltre che una breve presentazione dei dati utilizzati e delle operazioni preliminari adottate per la preparazione alle fasi di analisi e simulazione successiva. Si prosegue con il Capitolo 3, nel quale viene presentato il modello scelto, specificandone la forma e le relative distribuzioni a priori adottate per la fase di simulazione. Questa sezione è stata volutamente scomposta in due parti: la prima a riferimento della componente prematura del modello e la seconda legata ai decessi per longevità. Per quest'ultima, è stato dedicato un approfondimento alla SUN, una particolare forma funzionale ottenuta nel calcolo di una delle full conditional del modello. Sempre in questo capitolo, viene data un'idea dell'algoritmo di Gibbs Sampling utilizzato per la stima, soffermandosi anche sulla scelta dei valori iniziali e degli iperparametri delle distribuzioni utilizzate. Il quarto Capitolo, presenta il clustering e l'algoritmo di expectation-maximization, con il quale si è affrontato il raggruppamento dei paesi. Il Capitolo 5 è dedicato ai risultati ottenuti, sia in termini di stime, sia per ciò che riguarda la qualità delle simulazioni. Si conclude con delle osservazioni di carattere generale sull'intero lavoro, nel Capitolo 6.

Capitolo 1

Mortalità prematura

La mortalità prematura vanta gran interesse nelle analisi comparative a livello nazionale ed internazionale. Non a caso l'*ONU* (Organizzazione delle Nazioni Unite), nell'obbiettivo 3.4 dell'agenda dello sviluppo sostenibile [2], pone la riduzione di un terzo entro il 2030 del numero di morti premature tramite il trattamento e la promozione del benessere e della salute mentale. Il monitoraggio di questa componente della mortalità è particolarmente significativo, poiché coinvolge una fascia di individui che si trovano in una fase della vita in cui ci si aspetterebbe generalmente uno stato di buona salute e un'aspettativa di vita più lunga. In questa prospettiva, l'analisi della mortalità prematura fornisce importanti riferimenti sulla salute di una popolazione, sull'efficacia delle politiche sanitarie e sulla necessità di interventi mirati a prevenire le cause di decesso che colpiscono in età giovanile.
Malgrado il fascino dal punto di vista concettuale risulta difficile avere un'idea operativa per tale indicatore. La letteratura offre interpretazioni disparate, a seconda dei contesti e dagli enti promotori della ricerca. Volendo citarne alcuni, l'*OECD* (Organizzazione per la cooperazione e lo sviluppo economico), identifica la mortalità prematura come anni potenziali di vita persi (*PYLL*) prima dei 70 anni [3]. Il *GBD* (Global Burden of Disease) identifica il valore, moltiplicando il numero di decessi per un'aspettativa di vita standard globale, all'età in cui avviene la morte [4]. Una procedura comune adottata da altri enti come l'*OMS* (Organizzazione Mondiale della Sanità) o l'*Eurostat*, è quella di stabilire delle soglie numeriche specifiche al di sotto o entro le quali, una morte viene etichettata come prematura. Nei casi citati si ha rispettivamente la finestra 30-70 anni [5], e la soglia massima dei 65 anni [6], nel definire una morte precoce.
L'evolversi della ricerca in ambito accademico, ha portato l'adozione di approcci sostanzialmente differenti, i quali derivano direttamente dalla natura stessa dell'aspetto sotto osservazione. In origine Lexis [7], seguito poi da Kannisto [8] [9] e Cheung et al. [10], abbandonano l'idea di una soglia limite che risultava essere troppo stringente in alcuni

contesti, e propongono una metodologia che evidenzia due curve di distribuzione parzialmente sovrapposte, una in rappresentanza della componente prematura e la seconda che evidenzia le morti per longevità.
Emergono quindi due filosofie di approccio distinte; la prima di carattere *assoluto*, mentre la seconda di carattere più pragmatico e *relativo*.

1.1 Mortalità prematura come approccio assoluto

Il monitoraggio della mortalità prematura tramite un approccio di tipo assoluto, risulta quello di più semplice comprensione. Viene adottato dalle principali organizzazioni e istituzioni di livello internazionale e prevede venga fissato un valore soglia di età, entro il quale si considera un decesso come prematuro. Solitamente il valore varia tra i 65 e i 70 anni, scelto in maniera indipendente dall'ente di analisi. Questo porta a non avere un unico limite standardizzato e di conseguenza le unità statistiche incluse nell'insieme prematuro, possono avere un percorso clinico molto differente.
Un esempio è offerto dall'*Eurostat* che considera tra i decessi prematuri due categorie, identificandole come morti *amenable* e morti *preventable* [11]. Entrambe le tipologie indicano una componente di decessi evitabili, avvenuti quindi in maniera prematura rispetto all'età dell'individuo. Nello specifico però, la prima tipologia riguarda quei decessi evitabili dal punto di vista clinico, magari attraverso la somministrazione di un trattamento maggiormente adeguato alla situazione clinica del paziente, mentre la seconda tipologia riguarda l'aspetto della prevenzione sociale, legata ai comportamenti del singolo individuo come dipendenze e screening medici.
Ciò che emerge quindi, è la soggettività con la quale si distinguono le categorie dei decessi, e sebbene la procedura decisionale di una soglia possa apparire semplice e immediata, questo approccio non tiene conto in alcun modo delle caratteristiche intrinseche della popolazione a cui si riferisce. Di conseguenza, l'identificazione di una soglia cronologica, rimane una questione controversa.

1.2 Mortalità prematura come approccio relativo

L'approccio di tipo relativo risulta sostanzialmente differente dal precedente. Ha visto delinearsi nel tempo numerose versioni, partendo da una prima idea di Lexis [7] (Figura 1.1, immagine in alto), ampliata successivamente da una proposta più complessa, da parte di Pearson [12] (Figura 1.1, immagine in basso).
Storicamente i modelli per l'analisi tendono a lavorare con i tassi di mortalità ($m(x)$). Più recentemente, alcuni lavori tra cui quelli appena citati considerano invece come quantità

da modellare, il numero di decessi delle tavole di mortalità ($d(x)$).
Lexis introduce l'idea che la componente prematura possa essere ricavata come eccedente tra la vera distribuzione dei decessi di una popolazione e la singola componente delle morti per longevità. Per quest'ultima (Figura 1.1, area grigia) assume in oltre, un andamento simmetrico ottenuto a partire dalla coda destra dell'intera distribuzione dei decessi e prevedendo, per le età inferiori, il profilo presunto della curva (linea tratteggiata).
L'area in eccesso (Figura 1.1, area rossa) tra la curva effettiva e l'ipotetica rappresentazione della componente naturale, costituisce la quota di morti premature nella popolazione di riferimento. È in questi termini che l'approccio si definisce *relativo*, proprio perchè è ottenuto come differenza tra le due curve menzionate.

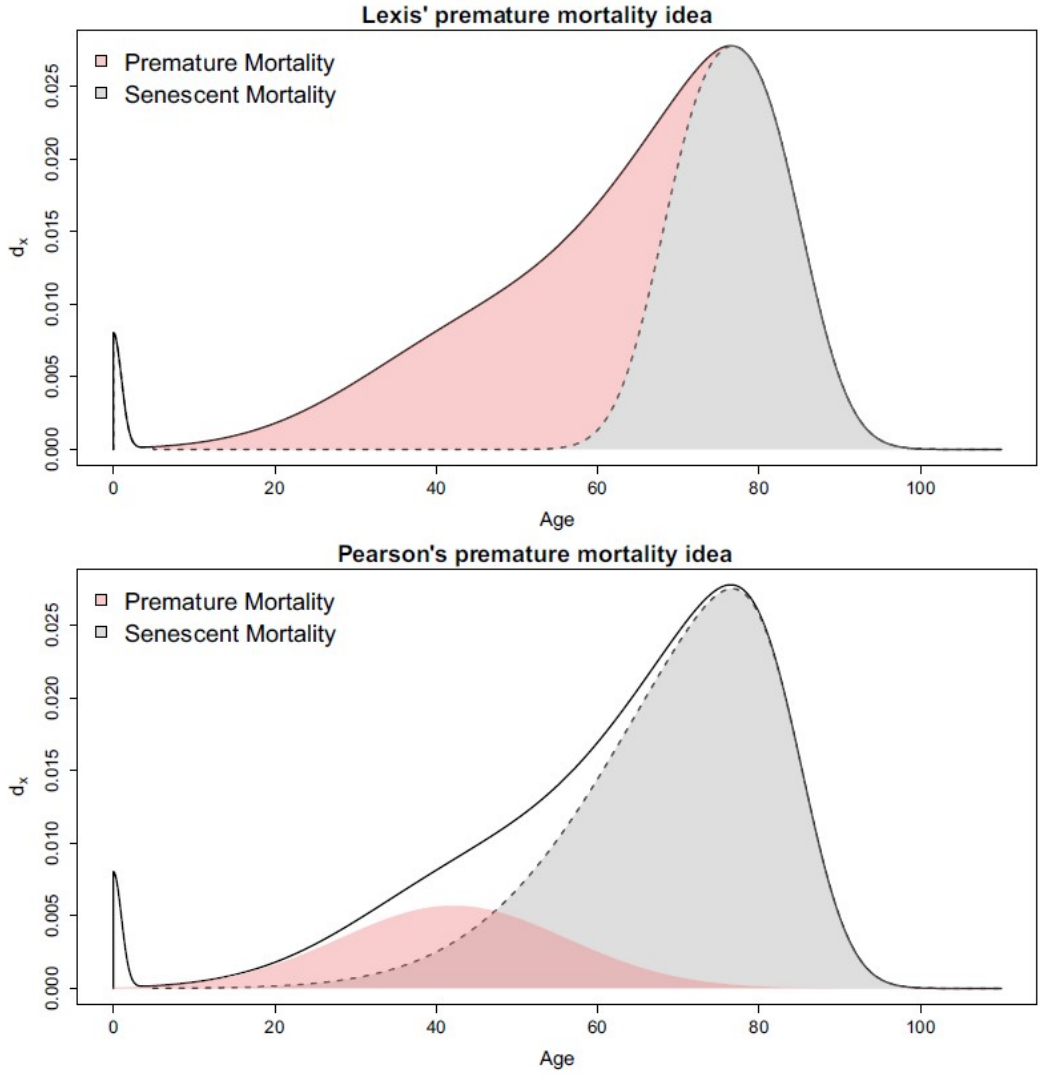

Figura 1.1: Da Mazzuco et al. [1]; rappresentazioni tramite approccio relativo per la misurazione della mortalità prematura (area rossa) secondo Lexis e Pearson.

La proposta di Lexis verrà poi corretta da Pearson, secondo il quale l'idea di una curva simmetrica per descrivere la mortalità prematura risulta irrealistica e troppo limitante. Avanza quindi una nuova interpretazione, proponendo una distribuzione asimmetrica che possa meglio rappresentare l'andamento dei decessi naturali (Figura 1.1, immagine in basso).
L'identificazione di questi andamenti, la si dovrà al contributo di Zanotto et al. [13], il quale formulerà un primo modello statistico a partire dalle distribuzioni delle età alla morte. Tale scelta deriva dalle ricerche di Basellini e Camarda [14], Mazzuco et al. [15] e Pascariu et al. [16] che considerando questa curva come conveniente da adattare.
Si tratta del modello mistura a tre componenti:

$$d(x) = \eta f_I(x) + (1-\eta)\alpha f_m(x, \xi_m, \omega_m, \lambda_m) + (1-\eta)(1-\alpha) f_M(x, \xi_M, \omega_M, \lambda_M) \quad (1.1)$$

avente:

- $f_I(x) = \frac{\sqrt{2}}{\pi} exp(-x^2)$ **mortalità infantile**;
- $f_m(x, \xi_m, \omega_m, \lambda_m) = \frac{2}{\omega_m}\phi(\frac{x-\xi_m}{\omega_m})\Phi(\lambda_m \frac{x-\xi_m}{\omega_m})$ **mortalità prematura**;
- $f_M(x, \xi_M, \omega_M, \lambda_M) = \frac{2}{\omega_M}\phi(\frac{x-\xi_M}{\omega_M})\Phi(\lambda_M \frac{x-\xi_M}{\omega_M})$ **mortalità senescente**.

La distribuzione dei decessi per età $d(x)$, viene adattata combinando le tre componenti f_I, f_m e f_M. Queste rappresentano rispettivamente le distribuzioni della morte all'età x, per la quota infantile, per le morti premature e per i decessi naturali, ai quali si farà riferimento in seguito con il termine di morti senescenti. In particolare per f_m e f_M viene utilizzata una distribuzione normale asimmetrica, avente ξ, ω e λ come parametri della distribuzione, mentre ϕ e Φ a indicare rispettivamente funzione di distribuzione e funzione di ripartizione di una normale. Il caso asimmetrico rappresenta un'estensione della distribuzione normale semplice, alla quale viene aggiunto un parametro di asimmetria a indicare verso quale lato si sviluppa maggiormente una coda della distribuzione. Questa tipologia di curve risulta più adatta in contesti come quello in analisi.
Nei termini del lavoro proposto, la stima del parametro α rappresenta quella di maggiore interesse, poiché indica la quota di mortalità prematura nell'insieme di riferimento.
Già nello scorso decennio, Siler [17], Heligman e Pollard [18] hanno proposto una scomposizione dei decessi basata sui tassi di mortalità $m(x)$ relativi alle componenti citate, integrando a questi, anche fattori aggiuntivi, al fine di descrivere ulteriori fasi e aspetti della vita, rendendo così la modellazione più accurata. La letteratura più recente ha visto nel lavoro di Basellini e Camarda [19] un modello mistura basato sui valori dei decessi presenti nelle tavole di mortalità $d(x)$.
Quest'ultimo, come i precedenti non dispone però di una stima per la quota di mortalità

prematura che viene invece offerta dall'approccio del modello 1.1.
L'adozione di un approccio relativo consente quindi una modellazione più flessibile, evitando di vincolarsi a una soglia numerica che potrebbe risultare eccessivamente restrittiva per determinati paesi o gruppi di individui.

1.3 Limiti etici ed operativi

Entrambi gli approcci proposti, presentano alcune criticità operative che, dato il campo di studio, si riflettono anche sul punto di vista etico.
L'approccio assoluto, con la scelta di una soglia numerica, risulta inequivocabile nella distinzione delle diverse mortalità, senza però tener conto in alcun modo delle caratteristiche specifiche della popolazione di interesse (Mazzuco et al. [1]). Se si volesse essere più puntuali scegliendo un valore specifico per paese, questi non sarebbero più confrontabili avendo parametri di riferimento differenti. Al contrario, un unico valore per tutti, porterebbe ad uniformare troppo paesi con caratteristiche di longevità molto differenti, con conseguenti difficoltà nelle decisioni sociali.
Oltre la singola stima delle componenti, l'aspetto di maggior rilevanza è che la vita di un singolo individuo, e quindi la sua morte, avrebbe un peso specifico differente a seconda dello stato di appartenenza, risultando meritevole di essere evitata in un paese piuttosto che in un altro (Mazzuco et al. [1]). Il concetto di morte meritevole in alcuni paesi rispetto ad altri solleva interrogativi profondi sulle disparità socio-economiche e sulla qualità dell'assistenza sanitaria. L'introduzione di soglie fisse potrebbe comportare il rischio di discriminare comunità che già affrontano sfide significative, contribuendo a perpetuare disuguaglianze globali. La variabilità delle condizioni di vita, delle risorse e delle infrastrutture sanitarie tra le nazioni può influenzare significativamente il tasso di mortalità prematura e senescente. Paesi con accesso limitato a cure mediche di alta qualità, istruzione e condizioni socio-economiche sfavorevoli possono trovarsi in una posizione svantaggiata, con una maggiore incidenza di morti premature rispetto a paesi più sviluppati. Un approccio di carattere endogeno, basato quindi sulle curve stesse dei decessi, porta a distinzioni delle mortalità direttamente collegate alle metodologie di analisi. Il rischio, è quello di risultare errate qualora si inglobassero informazioni provenienti da paesi con caratteristiche demografiche molto differenti.
Data la delicatezza e l'importanza dell'ambito nel quale si sta indagando, la definizione di criteri standardizzati per la determinazione di tali indicatori rappresenta un compito complesso. La comprensione delle curve di mortalità riveste un'importanza cruciale negli interessi di uno stato, non solo per decisioni governative efficienti, ma soprattutto al fine

di progettare politiche sanitarie mirate ed efficaci per l'intera popolazione, cercando di promuovere l'equità e il rispetto per la dignità umana in ogni paese.

Capitolo 2

Metodo e dati

In questa capitolo, partendo dal lavoro di Mazzuco, Suhrcke e Zanotto [1], verrà presentato il metodo adottato per affrontare il problema della mortalità prematura, combinando gli aspetti vantaggiosi di entrambi gli approcci enunciati in precedenza e limitando il più possibile quelli negativi.

2.1 Metodo

L'idea adottata è quella di delineare un modello che permetta di unire gruppi di paesi simili sulla base della distribuzione della mortalità senescente, lasciando invece la componente prematura libera di variare in maniera indipendente da paese a paese. Sebbene con un obbiettivo differente, la scelta di un fattore comune per più unità risulta ragionevole anche in base al modello previsivo di Li-Lee [20]. Per questo modello gli autori propongono un tasso di mortalità specifico per paese, unito ad un fattore comune per il gruppo di loro appartenenza. In questi termini, il fattore comune legato alla distribuzione dei decessi in età longeve, permette ai paesi membri dello stesso gruppo di essere confrontabili tra loro, senza però uniformarsi totalmente grazie alla componente prematura variabile. Anche dal punto di vista logico, si può pensare che paesi simili, per caratteristiche sociali, sviluppo in ambito sanitario e di longevità della propria popolazione, possano condividere in maniera rapida i progressi medici raggiunti.
È sulla base di questi ragionamenti che viene proposto il modello gerarchico a mistura, riportato in seguito:

$$d_j(x) = \alpha_j f_j^m(x, \mu_j^m, \sigma_j^m, \gamma_j^m) + (1 - \alpha_j) f^M(x, \mu^M, \sigma^M, \gamma^M) \tag{2.1}$$

dove con $d_j(x)$ si indica la distribuzione dei decessi per età x nella tavola della vita del paese j.

Questo modello presenta una formulazione gerarchica che rispetto al 1.1, ignora la componente di mortalità infantile dei primi anni di vita ed è costituito da una mistura di due componenti f^m e f^M, rispettivamente la componente di mortalità prematura avente parametri distinti per ogni paese, e la mortalità senescente che invece ha un valore comune per tutti i paesi etichettati come simili tra loro. Il peso delle singole componenti è rappresentato dal parametro α_j (e $1 - \alpha_j$), avente anch'esso una struttura gerarchica.
L'aspetto innovativo di questa tesi, rispetto al lavoro di Mazzuco et al. [1], risiede nell'identificazione dei gruppi di paesi simili tra loro attraverso un clustering di tipo model based. Precedentemente in [1], la scelta per la composizione dei cluster avviene identificando tre gruppi di paesi con mortalità senescente, *bassa*, *media* e *alta*, calcolata unicamente sulla base della speranza di vita alla nascita. Un buon adattamento del modello ai dati suggerisce quindi una buona composizione del cluster, al contrario se un paese avesse un profilo di mortalità molto diverso dagli altri del gruppo, il modello non si adatterebbe adeguatamente ai dati.
Per questo lavoro, durante la procedura di stima, esplicitata nel dettaglio nel Capitolo 3, viene inserito un processo di clustering, basato via via sulle stime parziali dei parametri della componente senescente. Ad ogni iterazione, l'algoritmo produrrà una stima delle etichette per ognuno dei paesi considerati, fino al raggiungimento della convergenza. Il clustering è basato sulla verosimiglianza della componente senescente perché sarà questa, come già spiegato, a determinare la similarità tra gli stati.
Il numero dei cluster evidenziati in Mazzuco et al. [1], verranno poi confrontati con quelli ricavati dalla nuova procedura, sia dal punto di vista della composizione sia in termini di valori stimati. Il numero ottimale dei gruppi, sarà scelto massimizzando la log verosimiglianza del modello.

2.2 Descrizione del dataset

I dati utilizzati in questo lavoro sono forniti dal database di mortalità dei paesi dell'America Latina (LAMBdA - Latin American Mortality database)[21]. Si tratta di un progetto sostenuto dal *National Institute on Aging*, che ha operato negli anni per la raccolta e la creazione di un database, con oltre 15 decadi di censimenti, registrazioni e conteggi totali dei decessi. Questi ultimi, rappresentano le tavole di mortalità, o tabelle di mortalità.
In generale si tratta di strumenti utilizzati in demografia per analizzare e prevedere i tassi di mortalità all'interno di una popolazione o di un gruppo di persone in uno specifico periodo. Le tavole forniscono informazioni importanti sull'andamento della sopravvivenza e dei decessi a diverse età, risultando utili per le pianificazioni finanziarie o la previdenza sociale. Solitamente sono basate su dati storici, per fasce di età di ampiezza pari a 5 anni.

Queste tavole offrono non solo la situazione demografica nel periodo di interesse, ma anche una visione più approfondita dei fattori esterni che posso aver modellato la struttura sociale nel tempo. Rispetto ai decessi osservati in una popolazione reale, i valori presenti all'interno delle tavole, fanno riferimento ai decessi in una coorte fittizia. Non risentono quindi della diversa struttura per età delle varie popolazioni e nemmeno della diversa numerosità, trovando quindi maggiore applicabilità nei contesti operativi.
Per questo lavoro, si sono considerate le tavole di mortalità relative a 17 paesi dell'America Latina, per individui in una finestra di età da 0 a 85+ anni. I paesi a disposizione per questo progetto sono: Argentina, Brasile, Cile, Colombia, Costa Rica, Cuba, Repubblica Dominicana, Ecuador, El Salvador, Guatemala, Messico, Nicaragua, Panama, Paraguay, Perù, Uruguay e Venezuela. Un estratto dei dati grezzi a disposizione viene riportato nella tabella 2.1.

	ctry	ctry.code	year	age	mx	qx	lx	Lx	Tx	ex
229	Argentina	2020	2005	0	0.02	0.02	100000	98436	7045162	70.45
230	Argentina	2020	2005	1	0.00	0.00	98273	392392	6946725	70.69
231	Argentina	2020	2005	5	0.00	0.00	97981	489530	6554333	66.89
232	Argentina	2020	2005	10	0.00	0.00	97844	488814	6064804	61.98
233	Argentina	2020	2005	15	0.00	0.01	97675	487147	5575990	57.09
234	Argentina	2020	2005	20	0.00	0.01	97164	484199	5088843	52.37
235	Argentina	2020	2005	25	0.00	0.01	96489	480638	4604644	47.72
236	Argentina	2020	2005	30	0.00	0.01	95736	476623	4124006	43.08
237	Argentina	2020	2005	35	0.00	0.01	94879	471806	3647383	38.44
238	Argentina	2020	2005	40	0.00	0.02	93800	465360	3175577	33.85
239	Argentina	2020	2005	45	0.01	0.03	92283	455854	2710217	29.37
240	Argentina	2020	2005	50	0.01	0.04	89966	441257	2254363	25.06
241	Argentina	2020	2005	55	0.01	0.06	86394	418911	1813106	20.99
242	Argentina	2020	2005	60	0.02	0.10	80953	386040	1394195	17.22
243	Argentina	2020	2005	65	0.03	0.15	73151	340243	1008155	13.78
244	Argentina	2020	2005	70	0.05	0.22	62522	279455	667912	10.68
245	Argentina	2020	2005	75	0.08	0.34	48708	201968	388457	7.98
246	Argentina	2020	2005	80	0.14	0.52	32079	118737	186488	5.81
247	Argentina	2020	2005	85	0.23	1.00	15415	67752	67752	4.40

Tabella 2.1: Estratto della tavola della mortalità dell'Argentina per l'anno 2005, dal database LAMBdA [21].

Prima di procedere nel descrivere le componenti del modello e le procedure di stima, i dati raccolti sono stati elaborati. Si sono considerati i dati più recenti a disposizione, relativi all'anno 2005 e sulla base delle operazioni svolte in Mazzuco et al.[1], dalla forma compatta osservabile nella tabella, si è passati ad una forma disaggregata. I valori dei decessi sono stati resi individuali, ipotizzando una distribuzione uniforme per ogni classe di età, fatta eccezione per la prima e l'ultima classe, per le quali si è scelta invece una

distribuzione esponenziale. Inoltre, dato che il modello proposto 2.1, non prevede nessuna componente infantile, i dati relativi ad individui con un'età inferiore ai 5 anni non sono stati considerati. In conclusione, con circa 550 osservazioni per ognuno dei diciassette paesi (9217 unità statistiche totali), si è potuto passare alla fase di stima del modello. La quantità di osservazioni utilizzate deriva solamente da ragioni di tipo computazionale. Essendo i dati riferiti a coorti fittizzie le cui numerosità sono arbitrarie, si è cercato di mantenere un compromesso tra informazione utilizzata e tempi di esecuzione ragionevoli.

2.3 Strumenti e software

Per l'intera realizzazione del progetto, si è fatto uso del software di analisi statistica `R`, (`R` Development Core Team, 2008) in versione 4.3.2 (2023-10-31 ucrt). Il codice utilizzato è disponibile nell'Appendice, in coda al lavoro.

Capitolo 3

Modello

3.1 Implementazione

Nella sezione 2.1, è stato presentato un modello mistura gerarchico a due componenti. Queste due componenti, f^m e f^M, rappresentano rispettivamente la componente di mortalità prematura e senescente della distribuzione di mortalità del paese j-esimo.
Le distribuzioni scelte sono una normale; $f_j^m \sim \mathcal{N}(\mu_j, \sigma_j^2)$ e una normale asimmetrica; $f^M \sim \mathcal{SN}(\xi, \omega^2, \gamma)$. Quest'ultima in particolare, è in accordo con la proposta di Pearson [12], il quale suggerisce una distribuzione non simmetrica per la componente senescente.
I parametri μ_j e ξ rappresentano i termini di posizione delle due distribuzioni, σ_j^2 e ω^2, i termini di scala, che per analogia vengono indicati entrambi con il termine ^{2}, e infine γ, parametro di forma della distribuzione normale asimmetrica. In oltre verrà indicato, sempre per questioni di notazione, $\theta = (\alpha_j, \mu_j, \sigma_j^2, \xi, \omega^2, \gamma)$, l'insieme di tutti i parametri del modello comprensivo anche del termine di mistura α.
In maniera esplicita quindi, la distibuzione delle morti per il j-esimo paese all'età x è:

$$d_j(x) = \alpha_j \, \mathrm{N}(\mu_j, \sigma_j^2) + (1 - \alpha_j) \, \mathrm{SN}(\xi, \omega^2, \gamma) \tag{3.1}$$

Il modello viene adattato tramite un approccio bayesiano, pertanto è necessario affiancare alla verosimiglianza considerata, delle distribuzioni a priori per ogni singolo parametro in θ. A differenza del lavoro originale [1], in cui viene utilizzato per la stima dei parametri un approccio Hamiltonian Monte Carlo, qui si procede in maniera differente. La scelta delle distribuzioni a priori viene fatta nell'ottica di semplificare il calcolo delle full conditional attraverso un approccio tipo Gibbs Sampling.
Il Gibbs Sampling è un potente strumento nel campo dell'inferenza statistica bayesiana, utilizzato nella stima di distribuzioni di probabilità complesse. Questo metodo consente di esplorare lo spazio dei parametri in modo efficiente, convergendo gradualmente verso

la distribuzione di equilibrio desiderata. Il campionamento di Gibbs è infatti un caso specifico dell'algoritmo Metropolis-Hastings in cui le proposte vengono sempre accettate e l'aggiornamento di un parametro o su una serie di parametri di interesse, avviene sulla base dei dati osservati. Avendo nel modello un numero di parametri non indifferente, la possibilità di sfruttare forme coniugate non eccessivamente complesse, permette di convergere verso una distribuzione target in maniera efficace.
Mantenendo una notazione ordinata, verranno presentate le distribuzioni a priori e le conseguenti full conditional per i singoli parametri, partendo dal termine di mistura α.

3.2 Parametro di mistura

Ogni singola x_i contribuisce con il proprio valore all'insieme delle morti premature o senescenti, non entrambe. Per poter generare valori per α è necessario l'utilizzo di variabili latenti z_i, che indicano se l'osservazione x_i corrispondente, contribuisce alla componente di decessi prematuri o a quelli senescenti. Formalmente significa che vi è una struttura del tipo:

$$z_i|\pi \sim \mathrm{B}(\pi) \tag{3.2}$$

e quindi:

$$x_i|z_i = 1, X \sim \mathrm{N}(\mu_j, \sigma_j^2) \quad \text{e} \quad x_i|z_i = 0, X \sim \mathrm{SN}(\xi, \omega^2, \gamma) \tag{3.3}$$

In questa formulazione, il ruolo precedentemente svolto da α viene ora assunto dal parametro π. Tuttavia, a fini pratici e per mantenere la coerenza nella notazione, si continuerà a utilizzare π per riferirsi al termine di mistura, sia nel modello, sia in θ.
È necessario in oltre, essendo un'analisi bayesiana, specificare la distribuzione a priori per questo nuovo termine del modello. Nello specifico si ha:

- $\pi \sim$ Be(1,1)

Questa distribuzione beta non risulta particolarmnete informativa ma offre un vantaggio nel calcolo della full conditional del parametro. Ora combinando la priori per π con la verosimiglianza corrispondente alla struttura latente appena descritta, si ottiene:

$$f(\theta|X,Z) \propto \prod_{i=1}^{n} \pi^{z_i}(1-\pi)^{1-z_i} \left[\frac{1}{\sqrt{2\pi\sigma_j^2}} \exp\left(-\frac{(x-\mu_j)^2}{2\sigma_j^2}\right)\right]^{z_i} \left[\frac{2}{\omega}\phi\left(\frac{x_i-\xi}{\omega}\right)\Phi\left(\gamma\frac{x_i-\xi}{\omega}\right)\right]^{1-z_i} \tag{3.4}$$

con $X = (x_1, x_2, \ldots, x_n)$ e $Z = (z_1, z_2, \ldots, z_n)$.
Dal punto di vista formale, prima di ricavare le distribuzioni marginali di ogni singolo parametro andrebbe scritta l'intera distribuzione congiunta.

Poiché i parametri sono indipendenti tra loro, si è deciso di facilitare la scrittura e la comprensione delle singole componenti mantenendo momentaneamente solo la prima priori. Di conseguenza, si è potuto derivare la full conditional del parametro π in modo più agevole, marginalizzando la 3.4 rispetto a quest'ultimo.
Si ricava:

$$f(\pi|Z) \propto \pi^{\sum_{i=1}^{n} z_i}(1-\pi)^{n-\sum_{i=1}^{n} z_i} \tag{3.5}$$

che corrisponde alla distribuzione beta:

$$\text{Be}(1+\sum_{i=1}^{n} z_i \ , \ 1+n-\sum_{i=1}^{n} z_i) \tag{3.6}$$

La distribuzione ottenuta per il parametro della componente di mistura, risulta facilmente implementabile all'interno di un algoritmo di Gibbs Sampling, permettendo la stima della quota di mortalità prematura per ogni paese j.
Ad ogni iterazione t dell'algoritmo, la probabilità di un'osservazione del paese j-esimo di contribuire alla componente prematura ($z_i = 1$) risulta proporzionale alla probabilità al passo precedente ($\pi_j^{(t-1)}$) moltiplicata per un valore $f(x_i|\theta^{(t-1)}, z_i)$. Questo viene calcolato come il rapporto della densità prematura sul totale della densità 3.1, utilizzando per i parametri le stima al passo precedente ($\theta^{(t-1)}$). Questo accorgimento, prima di aggiornare il parametro di mistura e procedere con le nuove stime dei parametri, consente di utilizzare il contributo dei dati e le informazioni che questi portano.
Prima di proseguire, occorre evidenziare come la densità considerata, coinvolga sempre un unico paese j-esimo per volta. Questo per sottolineare come la quantità di osservazioni utilizzate, all'interno delle sommatorie per l'aggiornamento dei parametri siano relative alle sole unità del paese in analisi j.
Per le rimanenti distribuzioni a priori e full conditional si è preferito affrontare la loro presentazione in due sezioni dedicate, cercando di mantenere distinte le componenti relative alla mortalità prematura da quelle relative alla mortalità senescente.

3.3 Componente prematura

Per i parametri di posizione e scala della componente prematura del modello, la scelta delle distribuzioni a priori è stata dettata da valutazioni di tipo operativo. Si è sfruttata la forma della densità normale, scegliendo delle distribuzioni ad essa coniugate.
Si ha:

- $\mu \sim \text{N}(\mu_0, \sigma_0^2)$ per il parametro di posizione;
- $\sigma^2 \sim \text{I-Ga}(a, b)$ per il parametro di scala.

Data la natura delle due distribuzioni a priori scelte, le full conditional corrispondenti mantengono la stessa forma funzionale, però con i parametri aggiornati.
Marginalizzando per μ e σ^2, il prodotto tra la verosimiglianza e le distribuzioni appena presentate si ottiene:

$$\mathrm{N}\left((\frac{\sum_{i=1}^{n} x_i z_i}{\sigma^2} + \frac{\mu_0}{\sigma_0^2})/(\frac{\sum_{i=1}^{n} z_i}{\sigma^2} + \frac{1}{\sigma_0^2})\ ,\ (\frac{\sum_{i=1}^{n} z_i}{\sigma^2} + \frac{1}{\sigma_0^2})^{-1}\right) \tag{3.7}$$

per il parametro di posizione e

$$\text{I-Ga}\left(a + \frac{\sum_{i=1}^{n} z_i}{2}\ ,\ b + \frac{\sum_{i=1}^{n}(x_i - \mu)^2 z_i}{2}\right) \tag{3.8}$$

per il parametro di scala.
Le distribuzioni 3.7 e 3.8, come per le rispettive priori, fanno riferimento ad una distribuzione normale e ad una distribuzione gamma inversa. Anche qui, come per il termine di proporzione della mistura, si fa riferimento al solo paese j-esimo in analisi.

3.4 Componente senescente

La componente di mortalità senescente risulta più delicata nella scelta delle distribuzioni a priori. Questo perché sulla convergenza dell'algoritmo e quindi sulla stima corretta dei parametri ξ, ω^2 e γ, si basa una buona identificazione dei cluster tra paesi.
Come riferimenti principali a supporto della scelta delle priori sono stati utilizzati il lavoro di Canale, Pagui e Scarpa del 2016 [22], nel quale viene illustrato un approccio bayesiano per l'inferenza su una distribuzione normale asimmetrica, e il lavoro di Mazzuco e Keilman [23], in particolare il capitolo 5; *Projecting Proportionate Age–Specific Fertility Rates via Bayesian Skewed Processes* curato da Aliverti, Durante e Scarpa.

In maniera analoga con quanto fatto per la componente prematura, data l'indipendenza tra i parametri in gioco, si è preferito procedere considerando solamente la componente dei decessi per cause naturali del modello 3.1. Ciò significa avere una variabile casuale univariata X con distribuzione normale asimmetrica, indicata come $X \sim \mathrm{SN}(\xi, \omega^2, \gamma)$. La funzione di densità di probabilità di questa distribuzione è:

$$f(x, \xi, \omega^2, \gamma) = \frac{2}{\omega}\phi\left(\frac{x-\xi}{\omega}\right)\Phi\left(\gamma\frac{x-\xi}{\omega}\right) \tag{3.9}$$

con ϕ e Φ funzione di densità e di ripartizione di una normale e ξ, ω^2, γ parametri rispettivamente di posizione, scala e forma.
L'adozione di una distribuzione simmetrica, in contesti di ambito medico, economico o come quello in esame, risulta spesso irrealistica.
Negli ultimi decenni infatti, la letteratura ha perseguito con maggior interesse la scelta di distribuzioni parametriche più flessibili in termini di asimmetria e curtosi, in modo da poter rappresentare più fedelmente le realtà in esame. Nello specifico, la distribuzione normale asimmetrica è stata ampliamente generalizzata da molti autori. Tra questi si ricordano i contributi di Azzalini e Capitanio [25] [26], Azzalini e Dalla Valle [27], Branco e Dey [28] e Genton e Loperfido [29]. Tra tutti, di particolare rilevanza è il lavoro di Arellano-Valle e Azzalini [30], nel quale vengono unificate in un unico articolo alcune proposte esistenti, fornendo anche importanti considerazioni per la famiglia di distribuzioni Unified Skew Normal, utile in seguito.

3.4.1 Distribuzione SUN (Unified Skew Normal)

Per la definizione delle distribuzioni a priori e delle relative full conditional, si procede, in linea con Canale, Pagui e Scarpa [22], considerando inizialmente ξ e ω^2 come valori noti pari a 0 e 1 e concentrando l'attenzione sul parametro di forma γ.
Prima di procedere, per uniformità con quanto riportato nell'articolo di riferimento, per il termine γ si farà riferimento con α. Questa rappresenta solo una scelta di notazione, per facilitare la comprensione dei passaggi.
Fatta questa premessa, la verosimiglianza relativa alla distribuzione 3.9, per un campione iid $X = (x_1, x_2, \ldots, x_n)$ risulta:

$$\mathcal{L}(\alpha) = \prod_{i=1}^{n} 2\phi(x_i)\Phi(\alpha x_i) \tag{3.10}$$

Come priori per il termine α, in accordo anche con quanto deciso in Mazzuco et al. [1], si è optato per la distribuzione normale asimmetrica seguente:

$$\mathrm{SN}(\alpha_0, \psi_0, \lambda_0) = \frac{2}{\psi_0}\phi\left(\frac{\alpha - \alpha_0}{\psi_0}\right)\Phi\left(\lambda_0\frac{\alpha - \alpha_0}{\psi_0}\right) \tag{3.11}$$

dove $\alpha_0, \psi_0, \lambda_0$ rappresentano gli iperparametri di posizione, scala e forma, mentre ϕ e Φ la funzione di densità e di ripartizione di una normale. La scelta dei valori per questi iperparametri verrà approfondita in seguito nella sezione 3.5.
Ora, moltiplicando tra loro la componente di verosimiglianza 3.10 con la priori 3.11 e

marginalizzando per il parametro di forma α, si ottiene la seguente full conditional:

$$\begin{aligned} f(\alpha|x) &\propto \phi\left(\frac{\alpha-\alpha_0}{\psi_0}\right)\Phi\left(\lambda_0\frac{\alpha-\alpha_0}{\psi_0}\right)\prod_{i=1}^{n}\Phi\left(\alpha x_i\right) \\ &\propto \phi\left(\frac{\alpha-\alpha_0}{\psi_0}\right)\Phi_{n+1}\left(\begin{bmatrix} x\alpha_0 \\ 0 \end{bmatrix}+\begin{bmatrix} x \\ \lambda_0/\psi_0 \end{bmatrix}(\alpha-\alpha_0)\,;I_{n+1}\right). \end{aligned} \tag{3.12}$$

La distribuzione appena ottenuta appartiene alla classe *Unified Skew Normal*, a cui si farà riferimento in seguito utilizzando l'acronimo SUN. Questa classe di distribuzioni estende la famiglia delle distribuzioni normali, per includere la possibilità di avere asimmetria (skewness) nei dati. In più, con il termine *unificato*, si fa riferimento al fatto che questa classe di distribuzioni racchiuda in sè varie tipologie di asimmetrie, tipiche di altre forme funzionali [30]. Dopo una procedura di normalizzazione e riscrivendola in maniera compatta la full conditional ottenuta, risulta essere una:

$$\mathrm{SUN}_{1,n+1}\left(\alpha_0,\gamma,\psi_0,1,\Delta,\Gamma\right) \tag{3.13}$$

dove $\Delta=[\delta_i]_{i=1,\dots,n+1}$ è un vettore di dimensione $n+1$ i cui elementi $\delta_i=\psi_0 z_i\left(\psi_0^2 z_i^2+1\right)^{-1/2}$ con $z=\left(\psi_0 x^{\mathrm{T}},\lambda_0\right)^{\mathrm{T}}$, $\gamma=\left(\Delta_{1:n}\alpha_0\psi_0^{-1},0\right)$ è un vettore di dimensione $n+1$ contenente i primi n valori di Δ e l'$n+1$-esimo elemento pari a zero , infine $\Gamma=I-D(\Delta)^2+\Delta\Delta^{\mathrm{T}}$, dove $D(V)$ rappresenta una matrice diagonale, avente gli elementi coincidenti con quelli del vettore V.
Per chiarire la relazione tra la distribuzione 3.12 e la sua formulazione compatta in 3.13 viene riportata la generica densità per $Z\sim\mathrm{SUN}_{d,m}(\xi,\gamma,\omega,\Omega,\Delta,\Gamma)$.
La relativa funzione di probabilità è:

$$f(Z;\xi,\gamma,\omega,\Omega,\Delta,\Gamma)=\phi_d(z-\xi;\omega\Omega\omega)\frac{\Phi_m\left(\gamma+\Delta\Omega^{-1}\omega^{-1}(z-\xi);\Gamma-\Delta\Omega^{-1}\Delta^{\mathrm{T}}\right)}{\Phi_m(\gamma;\Gamma)} \tag{3.14}$$

dove $\Phi_d(;\Sigma)$ è la funzione di ripartizione d-variata di una normale con matrice di varianza e covarianza Σ, Ω, Γ, e $\Omega^*=\left((\Gamma,\Delta)^{\mathrm{T}},\left(\Delta^{\mathrm{T}},\Omega\right)^{\mathrm{T}}\right)$ sono matrici di correlazione, e ω rappresenta una matrice diagonale di dimensioni $d\times d$.
Ora, in ordine con quanto riportato per 3.13, si considerino $\xi=\alpha_0$, $\omega=\psi_0$, e le generiche dimensioni d e m pari a 1 e $n+1$, in modo da avere la densità Z coincidente alla densità per il parametro di forma α della componente di mortaltà senescente del modello.
Per ogni ulteriore passaggio formale, si rimanda all'Appendice del lavoro di Canale, Pagui e Scarpa [22], nella quale sono esplicitate tutte le relazioni per ricavare gli elementi descritti in 3.13.

Prima di procedere con l'inferenza sul vettore completo dei parametri, occorre sottolineare come poter generare in maniera efficiente dalla distribuzione appena presentata.
I parametri ξ e ω^2 vengono mantenuti momentaneamente noti, con valori pari a 0 e 1.
La letteratura ha prodotto abbondanti risultati teorici sulle rappresentazioni stocastiche per le famiglie di distribuzioni asimmetriche. Nello specifico, per la classe di distribuzioni SUN, con l'obbiettivo di perseguire un risultato teorico efficiente in termini di simulazione, viene fatto riferimento alla Sezione 2.1 di Arellano-Valle e Azzalini (2006) [30] e al riepilogo in Canale, Pagui e Scarpa [22]. Viene ricavato il seguente Lemma:

Lemma 3.4: Sia $V_0 \sim \text{LTN}_q(-\gamma; 0, \Gamma)$ e $V_1 \sim N(0, \Omega)$ con V_0 indipendente da V_1 e dove la notazione $\text{LTN}_d(\tau; \mu, \Sigma)$ indica una distribuzione normale d-variata con media μ e matrice di varianza e covarianza Σ troncata sotto τ. Ora se

$$Y = \xi + \omega \left(\Delta^{\mathrm{T}} \Gamma^{-1} V_0 + \sqrt{1 - \Delta^{\mathrm{T}} \Gamma^{-1} \Delta} V_1 \right),$$

allora $Y \sim \text{SUN}_{1,q}(\xi, \gamma, \omega, 1, \Delta, \Gamma)$.

Si fa presente che i termini γ, Δ e Γ fanno riferimento a quelli descritti per la full conditional di α (3.13), mentre ξ e ω presentati nel Lemma, si rifanno ai termini scalari α_0 e ψ_0 sempre della relazione 3.13. Di conseguenza la dimensione q è pari a $n + 1$.
Inoltre, è utile ricordare che tutti i vettori a cui si è fatto riferimento in precedenza, sono vettori colonna. Questo è importante, soprattutto per le operazioni matriciali richieste dal Lemma 3.4, al fine di poter ottenere per Y, e quindi come valore della full conditional per α, un termine scalare.
Per simulare dalla distribuzione gaussiana troncata multivariata ($n+1$ dimensionale), occorre scegliere un algoritmo di campionamento efficiente. Negli ultimi anni, la ricerca si è mossa in questa direzione producendo algoritmi di campionamento a fette [31] o basati su Monte Carlo Hamiltoniano [32]. Nel caso in questione, dato che la dimensione della normale troncata cresce con n, è ragionevole aspettarsi un valore piuttosto grande. Da Botev 2017 [33] si sa che campionare in maniera indipendente da una normale troncata non risulta banale quando la dimensione è maggiore di qualche centinaio. Si è quindi scelto di sfruttare un approccio Hamiltoniano, in particolare quello presentato in [34].
L'articolo [34] con il relativo pacchetto, propone due algoritmi di generazione che rappresentano lo stato dell'arte per generare da una normale multivariata troncata quando la dimensione dei parametri supera qualche centinaio.
Gli algoritmi proposti sfruttano entrambi soluzioni analitiche della dinamica hamiltoniana e sono un Monte Carlo Hamiltoniano Armonico (*Harmonic-HMC*) e un Monte Carlo

Hamiltoniano a Zigzag (*Zigzag-HMC*).
Tra i metodi citati, è stato scelto lo Zigzag-HMC sulla base dei tempi di esecuzione dichiarati nell'articolo e soprattutto sulla base delle dimensioni della distribuzione che iterativamente è stato necessario generare.
Un secondo punto che ha richiesto particolare attenzione, è stato l'inversione della matrice Γ, di dimensioni $n+1 \times n+1$. In generale, l'inversione di una matrice richiede un numero di operazioni il cui ordine di grandezza è pari a $O\left(n^3\right)$. Questo fa si che risulti proibitivo procedere con l'inversione in maniera classica per il caso in analisi.
Per ovviare al problema, data la particolare forma con cui è stata rappresentata la matrice Γ, si è potuto ricorrere alla formula di Sherman–Morrison. Si tratta di un risultato fondamentale dell'algebra lineare che fornisce una modalità efficiente per calcolare l'inverso di una matrice. Per il caso in questione l'applicazione della formula risulta:

$$\begin{aligned}\Gamma^{-1} &= \left(I - D(\Delta)^2 + \Delta\Delta^{\mathrm{T}}\right)^{-1} \\ &= D\left(\Delta'\right) - \frac{1}{1+\sum_{i=1}^{n}\delta_i^2\left(1-\delta_i^2\right)^{-1}} D\left(\Delta'\right)\Delta\Delta^{\mathrm{T}} D\left(\Delta'\right) \\ &= D\left(\Delta'\right) - \frac{1}{1+\sum_{i=1}^{n}\delta_i^2\left(1-\delta_i^2\right)^{-1}}\tilde{\Delta}\end{aligned} \tag{3.15}$$

dove Δ' è il vettore di dimensione $n+1$ i cui elementi sono $\left(1-\delta_i^2\right)^{-1}$ e $\tilde{\Delta}$ è una matrice $n+1 \times n+1$ con elementi $\tilde{\delta}_{ij} = \delta_i\delta_j\left(1-\delta_i^2\right)^{-1}\left(1-\delta_j^2\right)^{-1}$, per $i,j = 1,\ldots,n+1$.
Viene sottolineato in Canale, Pagui e Scarpa [22], che questa espressione non è valida in generale per il modello SUN ma è una conseguenza della specificazione precedente.

A questo punto si hanno tutti gli elementi per la stesura dell'algoritmo di campionamento. Prima di procedere però, occorre puntualizzare alcune questioni. Come già anticipato, fino ad ora è stata trattata solamente la full conditional per il parametro di forma della componente senescente del modello. Nella sezione successiva verranno esplicitate anche le priori e le conseguenti full conditional, per il parametro di posizione ξ e per il parametro di scala ω^2. Le osservazioni che contribuiscono alla componente senescente del modello 3.1, sono quelle la cui corrispondente variabile z_i assume un valore pari a 0 (specificazione 3.3). Essendo, la componente senescente comune per tutti i paesi dello stesso cluster, le dimensioni delle matrici e dei vettori presentati in questa sezione seguiranno questo ordine di grandezza, e non più la dimensione specifica di un singolo paese, come nel caso della mortalità prematura.

3.5 Stima e iperparametri

Con riferimento ai termini di posizione e scala della componente di mortalità senescente, considerati fino a questo momento come valori noti, vengono ora presentate le priori e le relative full conditional. In accordo con Canale, Pagui e Scarpa [22], le priori scelte sono:

- $\xi \sim \mathrm{N}(\xi_0, k\omega^2)$ per il parametro di posizione;
- $\omega^2 \sim \text{I-Ga}(A, B)$ per il parametro di scala.

Anche in questo caso, la scrittura I-Ga(A, B) fa riferimento ad una distribuzione gamma inversa di parametri A e B.
Un caso particolare del Lemma 3.4 suggerisce l'introduzione di $\eta_1, \ldots, \eta_n$ variabili latenti. Condizionatamente a queste variabili latenti η_i, si può considerare la generica i-esima osservazione come proveniente da una distribuzione normale, con media $\xi + \omega\delta\,|\eta_i|$ e varianza $(1-\delta^2)\,\omega^2$.
L'apporto dato da questa interpretazione, permette una forma coniugata per le distribuzioni dei parametri di posizione e scala delle morti senescenti, facilitando l'implementazione dell'algoritmo di campionamento Gibbs Sampling. Limitandosi alla componente senescente i passi sui quali quest'ultimo si basa sono:

- viene generata ogni η_i dalla relativa distribuzione full conditional:

$$\eta_i \sim TN_0\left(\delta\,(x_i - \xi)\,, \omega^2\left(1-\delta^2\right)\right),$$

  dove δ è $\alpha/\sqrt{\alpha^2+1}$ e $TN_\tau\,(\mu, \sigma^2)$ è una normale troncata sotto τ con media μ varianza σ^2.

- per il parametro di posizione ξ, viene generato un valore dalla rispettiva full conditional, che risulta:

$$\mathrm{N}\left(\hat{\mu}, \hat{\kappa}\omega^2\right), \tag{3.16}$$

  mentre per ω^2 da:

$$\text{I-Ga}(A + (n+1)/2, B + \hat{b}) \tag{3.17}$$

  dove

$$\hat{\mu} = \frac{\kappa \sum_{i=1}^{n} (x_i - \delta\eta_i) + (1-\delta^2)\,\xi_0}{n\kappa + (1-\delta^2)},$$

$$\hat{\kappa} = \frac{\kappa\,(1-\delta^2)}{n\kappa + (1-\delta^2)}$$

$$\hat{b} = \frac{1}{2\,(1-\delta^2)} \left\{ \delta^2 \sum_{i=1}^{n} \eta_i^2 - 2\delta \sum_{i=1}^{n} \eta_i\,(x_i - \xi) + \sum_{i=1}^{n} (x_i - \xi)^2 + \frac{1-\delta^2}{\kappa}\,(\xi - \xi_0)^2 \right\}.$$

- si genera infine un valore per il parametro di asimmetria α proveniente da:

$$\alpha \sim \pi\left(\alpha \mid x^*\right),$$

con $x_i^* = (x_i - \xi)/\omega$ per $i = 1, \ldots, n$, e con $\pi(\alpha \mid x)$, la full conditional corrispondente descritta in 3.13.

Si pone ora l'attenzione nella scelta degli iperparametri e sui valori iniziali adottati per le distribuzioni coinvolte nel modello. L'idea è quella di offrire una visione chiara e dettagliata dei termini coinvolti sia per la componente prematura, sia per quella senescente. Riguardo gli iperparametri si è cercato di proporre quelli più coerenti e significativi, rifacendosi dove possibile al lavoro di Mazzuco et al. [1]. Questo passaggio non è sempre stato possibile a causa della differente parametrizzazione utilizzata nell'ambito della mortalità prematura. Nel lavoro di riferimento, viene adottata una parametrizzazione centrata secondo quanto indicato da Azzalini e Capitanio [24], mentre per il lavoro proposto è stata preferita una parametrizzazione diretta.
Per i termini π_i relativi alle proporzioni di mistura del modello, si è scelto per ognuno dei diciassette paesi un valore generato uniformemente tra 0.1 e 0.3.
Per la componente del modello distribuita come una normale, si sono scelti invece, valori comuni per tutti i paesi. Nel dettaglio il parametro di posizione μ_j è stato posto pari a 50 e quello di scala σ_j^2 pari a 144. Per le rispettive priori in 3.3 si è optato per distribuzioni non troppo specifiche. La curva per μ_j è stata centrata con μ_0 pari sempre a 50 e con iperparametro di scala, σ_0^2, pari a 49. Gli iperparametri a e b di σ_j^2 sono stati scelti pari a 144 e 15696. Per questi, come anche per la distribuzione gamma inversa della componente senescente, si è cercato di mantenere la stessa informatività adottata in Mazzuco et al. [1]. Si sono infatti ricavati valore atteso e varianza dalla distribuzione del lavoro citato, per poi eguagliarli a quelli della distribuzione scelta e ricavare gli iperparametri necessari.
Passando ai valori relativi alla distribuzione normale asimmetrica, si è cercato di aiutare l'avvio della procedura di clustering differenziando i valori di ξ. A seconda della quantità di cluster testati, sono stati scelti per ξ valori compresi tra 85 e 95, mentre valori fissi pari a 625 per ω^2 e a -4.5 per il parametro di asimmetria Figura 3.1.
Per gli iperparametri si ha invece; ξ_0 equivalente ai valori scelti per ξ, A e B pari a 470 e 292500, ricavati come descritto in precedenza, α_0, ψ_0 e λ_0 dell'equazione 3.11, rispettivamente -4.5, 0.5 e 1. Infine, per k di 3.16 un valore di 0.25.

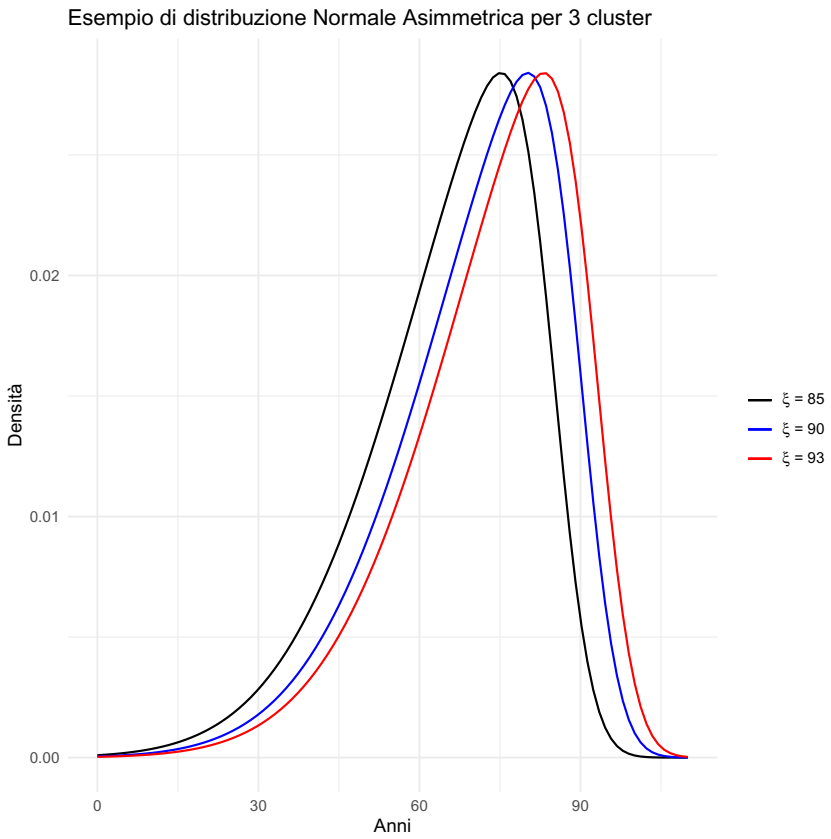

Figura 3.1: Esempio di distribuzione normale asimmetrica per la componente senescente del modello con 3 cluster. Parametro di posizione ξ pari a 85, 90 e 93.

Nel complesso, data la dimensione campionaria a disposizione si è lasciato fosse la verosimiglianza a guidare maggiormente il processo di stima, senza implementare né particolari controlli durante le diverse fasi di stima, né scelte distributive troppo mirate. Se lo studio riguardasse popolazioni di entità inferiore converrebbe implementare maggiori controlli.

Capitolo 4

Clustering

In questo capitolo si andrà a presentare l'utilizzo della componente senescente del modello, per definire un clustering di tipo model based, al fine di identificare un possibile numero di gruppi latenti tra i paesi del Centro America considerati.

4.1 Model Based Clustering

Con model based clustering si fa riferimento ad una tecnica di apprendimento non supervisionato che mira a identificare attraverso la modellazione statistica delle possibili strutture latenti nei dati. Il metodo ha come assunzione principale quella di considerare i dati come provenienti da una distribuzione probabilistica sconosciuta. Attraverso una procedura iterativa mira quindi a partizionare le unità statistiche nei rispettivi gruppi, sfruttando proprio questa struttura latente. L'assegnazione dei punti, al rispettivo gruppo di appartenenza, avviene considerando l'etichetta del cluster come un parametro da stimare attraverso un'opportuna procedura inferenziale [35].
La distribuzione ignota costituisce una mistura finita di distribuzioni, cioè una combinazione di componenti parametriche, ognuna necessaria a rappresentare uno specifico sottogruppo all'interno della popolazione. Le singole distribuzioni, possono appartenere a classi differenti o variare semplicemente nei valori assunti dai parametri. Quest'ultima casistica è quella proposta nella tesi. L'obbiettivo, è stimare i parametri assunti dalla mistura e poi utilizzare questi per calcolare le probabilità di appartenenza al cluster (Everitt et al. [36]).
Il modello utilizzato in questo lavoro si basa sulle distribuzioni normali asimmetriche, proprio per identificare quei gruppi di paesi con una più simile mortalità naturale. Di conseguenza, le unità statistiche da dividere in gruppi sono costituite dai diciassette paesi dell'America Centrale considerati.

In generale un modello a mistura finita con G componenti è tale se considerato un vettore casuale Y, per ogni sua realizzazione y è possibile scrivere la relativa densità come:

$$f(y|\theta) = \sum_{g=1}^{G} \tau_g f_g(y|\theta_g) \tag{4.1}$$

dove $\tau_g > 0$, tale che $\sum_{g=1}^{G} \tau_g = 1$, è la proporzione di miscelazione per la g-esima componente, $f_g(y|\theta_g)$ è la densità del g-esimo componente, e $\theta = (\tau_1, \ldots, \tau_G, \theta_1, \ldots, \theta_G)$ è il vettore dei parametri. In accordo con le motivazioni teoriche precedentemente espresse, le densità $f_1(y|\theta_1), f_2(y|\theta_2), \ldots, f_G(y|\theta_G)$ fanno riferimento alla componente senescente del modello, descritta da una distribuzione normale asimmetrica. Ne consegue che volendo scrivere la verosimiglianza del modello si ottiene:

$$L(\theta) = \prod_{i=1}^{n} \sum_{g=1}^{G} \tau_g f(y_i|\xi_g, \omega_g^2, \gamma_g) \tag{4.2}$$

dove con ξ_g, ω_g^2 e γ_g si fa riferimento ai parametri di posizione, scala e forma della normale asimmetrica, mentre $\tau_1, \ldots, \tau_G$ rappresentano le probabilità di appartenenza al cluster g. Questo concetto di appartenenza di ogni singola unità ad un gruppo può essere rappresentato con z_i pari 1 se l'osservazione, appartiene al gruppo e 0 altrimenti.
z_i è realizzazione di una variabile casuale z_i con distribuzione multinomiale a G categorie e relative probabilità $\tau_1, \ldots, \tau_G$. La probabilità di appartenenza $z_{ig} = 1$ date le osservazioni y_i è pari a:

$$P\left[z_{ig} = 1 \mid y_i\right] = \frac{\tau_g f(y_i|\xi_g, \omega_g, \gamma_g)}{\sum_{h=1}^{G} \tau_h f(y_i|\xi_g, \omega_g, \gamma_g)} \tag{4.3}$$

Quest'ultima formulazione risulta fondamentale all'interno della procedura. Permette infatti ad ogni passo dell'algoritmo di utilizzare le stime ottenute in maniera iterativa per i parametri del modello nel calcolo del vettore di probabilità di appartenenza ai gruppi. Sarà poi sulla base di questi valori, che le osservazioni verranno associate al relativo cluster di appartenenza. Negli anni, la letteratura ha suggerito molti approcci per la stima delle distribuzione delle miscele sia di tipo frequentista che di carattere bayesiano. Quello che però ha dominato per i risultati, è un approccio basato sulla massimizzazione della verosimiglianza (Titterington et al. [37]). Nello specifico viene massimizzata la verosimiglianza di 4.1 attraverso la Massimizzazione delle Aspettative (Expectation-Maximization o algoritmo EM) [38]. Si tratta di un algoritmo ampiamente utilizzato nei problemi di model based clustering poiché offre una soluzione flessibile alle differenti forme introdotte per la rappresentazione delle componenti di mistura, adattandosi bene ai requisiti e alle sfide di queste tipologie di problemi.

4.2 Algoritmo EM

L'algoritmo EM è un algoritmo utilizzato per stimare i parametri di modelli statistici quando si hanno dati incompleti o mancanti, nel caso in analisi le etichette di appartenenza ai gruppi.
In generale è possibile considerare per i dati osservati $y_1, \ldots, y_n$ delle z_{ig} a indicare l'appartenenza al componente g-esimo. Ciò significa che:

$$z_{ig} = \begin{cases} 1 \text{ se } y_i \text{ appartiene al gruppo } g \\ 0 \text{ altrimenti} \end{cases}$$

Il vettore $z_i = (z_{i1}, \ldots, z_{iG})$ rappresenta il vettore delle etichette di appartenenza ai G gruppi da stimare.
Ora considerando osservazioni ed etichette come indipendenti, è possibile scrivere la verosimiglianza per l'algoritmo EM. Per il lavoro in questione in realtà si è fatto uso della log verosimiglianza. Questa scelta ha permesso di lavorare con valori più facilmente trattabili rendendo le operazioni più agevoli. Si ottiene quindi:

$$l_c(\theta) = \sum_{i=1}^{n} \sum_{g=1}^{G} z_{ig} \left[\log \tau_g + \log f(y_i | \xi_g, \omega_g^2, \gamma_g) \right] \tag{4.4}$$

Il funzionamento dell'algoritmo può essere riassunto in due passaggi iterativi: il primo è un "*E-step*", o passo di aspettativa, nel quale vengono calcolate le aspettative (o probabilità) delle variabili latenti date le osservazioni e i parametri correnti del modello. Il secondo è un "*M-step*", o passo di massimizzazione, nel quale vengono aggiornati i parametri del modello in modo da massimizzare la verosimiglianza dei dati. In questa fase vengono utilizzate le stime ottenute nella prima parte, cioè durante la fase di aspettativa. Questo significa sostituire le z_{ig} in 4.4 con i loro valori attesi:

$$\hat{z}_{ig}^{(q)} = \frac{\hat{\tau}_g^{(q-1)} f\left(y_i \mid \hat{\xi}_g^{(q-1)}, \hat{\omega^2}_g^{(q-1)}, \hat{\gamma}_g^{(q-1)}\right)}{\sum_{g=1}^{G} \hat{\tau}_g^{(q-1)} f\left(y_i \mid \hat{\xi}_g^{(q-1)}, \hat{\omega^2}_g^{(q-1)}, \hat{\gamma}_g^{(q-1)}\right)} \tag{4.5}$$

per $i = 1, \ldots, n$ e $g = 1, \ldots, G$, dove $\hat{\tau}_g^{(q)}$ è il valore di τ dopo la q-esima iterazione dell'algoritmo EM. Si noti che, nel passaggio E, il condizionamento avviene sulle stime correnti dei parametri, quindi l'uso di "$\hat{}$" per i parametri in 4.5.

La parentesi teorica appena presentata introduce in maniera generale quelli che sono i passaggi principali di un algoritmo EM. Con riferimento a quanto svolto e ricordando che l'obbiettivo della tesi era di implementare una procedura di clustering all'interno dell'algoritimo di stima dei parametri, occorre puntualizzare alcuni aspetti.
L'aggiornamento delle stime ha portato ad utilizzare i valori calcolati ad ogni passo del Gibbs Sampler per la fase "E" dell'algoritmo proposto in questa sezione. In questo modo, i valori via via stimati per i parametri del modello, vengono utilizzati all'iterazione successiva per calcolare le probabilità di appartenenza ad ogni gruppo (3.2).
Inoltre, l'assegnazione iterativa dei paesi ad un particolare cluster, comporta ripercussioni anche sui valori stimati. Questo perché le osservazioni utilizzate nella stima dei parametri di posizione, scala e forma di ogni gruppo, fanno riferimento a tutti i paesi appartenenti al gruppo in analisi e dichiarate a concorrere per la componente senescente. Inserire un paese distante per caratteristiche, ad un particolare gruppo, potrebbe influire in maniera significativa sulle successive stime.
Un ulteriore appunto riguarda il criterio di arresto che solitamente viene implementato quando si utilizza un algoritmo di questo tipo. La scelta di non inserire un criterio di arresto, è dovuta alla volontà di ottenere comunque un numero minimo di replicazioni per la stima dei singoli parametri. Sebbene non sia la scelta più ottimizzata, si è preferito lasciare libero l'algoritmo di continuare anche in condizioni in cui il guadagno in termini di verosimiglianza risultasse minimo.

Capitolo 5

Risultati

Si procede ora con la presentazione dei risultati ottenuti. Questi verranno commentati sia in termini di valori stimati, sia per ciò che riguarda la convergenza e la qualità delle simulazioni. L'algoritmo è stato testato con un numero di replicazioni pari a 10.000 e con un termine di burnin pari a 3.000. Questa scelta è stata fatta per ricercare un buon grado di convergenza lasciando le catene libere di esplorare il più possibile lo spazio parametrico.

5.1 Analisi e confronto

Malgrado gli sforzi e le accortezze nella fase di progettazione del modello, i risultati del lavoro non sono stati del tutto soddisfacenti.
L'intera analisi è stata effettuata più volte, utilizzando un numero di possibili cluster da individuare, compresi tra i 2 e 5. Specialmente con un numero di cluster superiore, ad esempio nei casi di 4 o 5, le osservazioni convergono velocemente all'interno di due soli gruppi per poi collassare in un unico insieme comprensivo di tutti i paesi considerati.
La convergenza di tutte le unità da clusterizzare all'interno di un unico gruppo, non rappresenta di per sé un problema. Questo aspetto sottolinea solamente come i profili dei singoli paesi non presentino una componente di variabilità così accentuata da appartenere a gruppi differenti, almeno in termini di mortalità senescente.
Test fatti con punti iniziali differenti in termini di valori di ξ, parametro di posizione della componente senescente, hanno confermato questo risultato anche quando il numero di gruppi da ricercare era ridotto solamente a 2 o 3.
Una possibile spiegazione della convergenza ad un unico gruppo tra gli stati, può essere attribuita alla forma iniziale della componente normale asimmetrica. In Figura 3.1, è riportato un esempio con tre cluster. Quello che si nota è che le tre distribuzioni condividono la maggior parte della propria massa di probabilità, specialmente confrontando la curva blu centrale, con le altre.

Anche prendendo in considerazione i soli profili più esterni (curva rossa e curva nera) e scegliendo punti iniziali più separati, la procedura non ha evidenziato particolari distinzioni portando a non preferire per nessun paese un gruppo piuttosto che un altro.
Un altro aspetto di rilievo riguarda le variabili latenti z_i responsabili dell'assegnazione delle singole osservazioni di ogni paese alla componente senescente o a quella prematura. Se le osservazioni dichiarate a far parte della componente di mortalità naturale non sono ben individuate tra tutte quelle disponibili per il singolo paese, la possibilità di identificare una distribuzione favorevole alla divisione in gruppi viene a mancare già dopo poche replicazioni dell'algoritmo. Questo, implica che un gruppo possa non venir più considerato proprio perché le osservazioni su cui si è calcolata la curva senescente siano in realtà legate a mortalità più basse.
Per valutare il lavoro dal punto di vista delle stime, si è optato per analizzare i dati solamente ricercando 2 o, più spesso, 3 gruppi latenti tra i paesi, così da poterli identificare come fatto in Mazzuco et. al. [1] in paesi con mortalità senescente bassa media o elevata.
In primo luogo viene analizzata la convergenza delle catene.
Di seguito, in Figura 5.1, si riporta a scopo illustrativo solamente l'esempio dell'Ecuador. La diagnostica di convergenza dei parametri relativi agli altri sedici paesi in analisi viene allegata in Appendice.

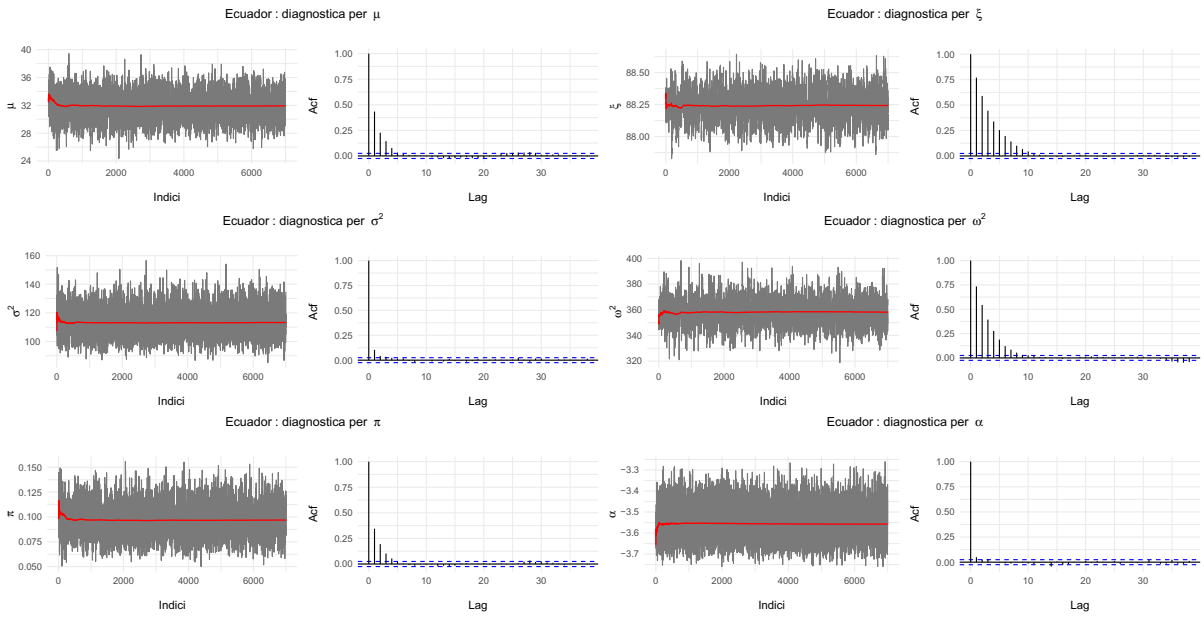

Figura 5.1: Diagnostica di convergenza per i parametri relativi all'Ecuador.

Le catene esplorano quasi tutte in maniera efficiente lo spazio parametrico, convergendo in breve ad un valore stabile della stima. Per il caso proposto solamente quelle per π e per ξ rimangono in un intorno più ristretto. Questo comportamento non risulta però, condiviso su larga scala dagli altri paesi che dispongono comunque di una buona compo-

nente di mixing anche per questi parametri, come pure per la componente di correlazione il cui comportamento risulta soddisfacente. Si può notare come i grafici di correlazione mostrino un andamento che tende ad esaurirsi in maniera piuttosto veloce per tutti i parametri. La componente di diagnostica è stata valutata anche in termini di effective size per valutare l'indipendenza dei campioni generati e in termini distributivi. La forma a campana unimodale ottenuta per la distribuzione dei valori simulati indica che le catene stanno esplorando le regioni corrette. Per tutti gli atri paesi considerati nell'analisi, non si sono evidenziate particolari differenze.
Si analizzano ora i valori stimati per i parametri del modello. Come anticipato il termine π_j, relativo alla proporzione di mistura per ogni paese risulta quello di maggiore interesse poiché permette di individuare la quota di decessi prematuri all'interno del paese j-esimo. Di seguito, si riporta una tabella riassuntiva delle stime per paese su scala percentuale, proponendo nella seconda riga, un confronto con i termini rilevati in Mazzuco et al. [1]. Tra parentesi viene indicato un intervallo di confidenza hpd. I valori hpd indicati risentono in maniera diretta della numerosità campionaria utilizzata. Intervalli più stringenti, potrebbero essere ottenuti considerando la vera numerosità dei decessi nell'anno 2005 e non i valori tabulati nelle tavole.

	Argentina	**Brasile**	**Cile**	**Colombia**	**Costa Rica**	**Cuba**
1	5.16 (2.7-8.36)	9.68 (6.5-13.38)	4.08 (2.07-6.57)	8.56 (5.88-11.73)	4.50 (2.51-6.93)	3.03 (1.37-5.22)
2	26.00	23.00	11.00	16.00	21.00	27.00
	Repubblica Dominicana	**Ecuador**	**El Salvador**	**Guatemala**	**Messico**	**Nicaragua**
1	8.68 (5.63-12.19)	9.69 (6.73-12.97)	22.30 (18.11-26.96)	15.31 (11.67-19.43)	6.7 (4.17-9.73)	12.74 (9.01-16.89)
2	20.00	26.00	34.00	24.00	33.00	26.00
	Panama	**Paraguay**	**Perù**	**Uruguay**	**Venezuela**	
1	5.54 (3.39-8.13)	7.26 (4.53-10.47)	6.92 (4.35-10.04)	5.00 (2.61-8.03)	11.72 (8.55-15.32)	
2	23.00	21.00	16.00	21.00	23.00	

Tabella 5.1: Valori stimati su scala percentuale per il parametro π_j per ogni paese (riga 1). Valori per il medesimo parametro ottenuti in Mazzuco et al. [1] (riga 2). Tra parentesi, l'intervallo hpd.

Ciò che emerge ad un primo sguardo è che i valori stimati siano tutti inferiori rispetto a quelli nell'articolo di riferimento. Occorre sottolineare però, che le stime utilizzate come riferimento sono fortemente soggette al gruppo di appartenenza del rispettivo paese. Un paese come il Cile, i cui valori sono stati calcolati sia per il gruppo a media che ad alta mortalità, vede una forte variazione nel parametro stimato, rendendo difficile un confronto diretto con i valori ottenuti per questo lavoro.
Per i parametri μ_J e σ_j^2 le stime sono riassunte nella tabella 5.2. I valori ottenuti non evidenziano differenze sostanziali nella componente prematura del modello per i diciassette paesi considerati. In generale le stime di μ_j variano tra età comprese di 31 e 35 anni, mentre il termine di scala σ_j^2 tra 110 e 115.

	Argentina	**Brasile**	**Cile**	**Colombia**	**Costa Rica**	**Cuba**
μ_j	35.30	35.55	34.44	32.45	31.62	32.76
σ_j^2	114.31 (10.69)	115.31 (10.74)	110.91 (10.53)	112.14 (10.59)	111.89 (10.58)	111.58 (10.56)
	Repubblica Dominicana	**Ecuador**	**El Salvador**	**Guatemala**	**Messico**	**Nicaragua**
μ_j	34.62	31.93	34.68	33.95	33.75	36.57
σ_j^2	114.15 (10.68)	113.22 (10.64)	117.24 (10.83)	115.04 (10.73)	111.87 (10.58)	117 (10.82)
	Panama	**Paraguay**	**Perù**	**Uruguay**	**Venezuela**	
μ_j	31.20	34.54	33.03	35.43	32.67	
σ_j^2	110.86 (10.53)	113.57 (10.66)	113.52 (10.65)	112.94 (10.63)	113.32 (10.65)	

Tabella 5.2: Valori stimati per i parametri μ_j e σ_j^2 per ogni paese. Tra parentesi, è indicato il valore della deviazione standard.

Rispetto i valori iniziali, la componente di verosimiglianza ha influito notevolmente, guidando le stime verso valori inferiori per entrambi i termini. Volendo contestualizzare il più possibile le stime ottenute, il calcolo di un intervallo di confidenza al 95% per le età dei decessi prematuri, produce per tutti i paesi intervalli tra gli 11 e i 58 anni di età. Queste metriche risultano abbastanza in accordo anche con i valori di mortalità prematura considerati tramite l'identificazione di una soglia numerica, in un approccio di tipo assoluto (Sezione 1.1).
Riguardo la componente di mortalità senescente i parametri ricavati sono solamente tre, in riferimento all'unico gruppo comprendente tutti i paesi. Nello specifico si ha per ξ, parametro di posizione, un valore pari a 88.24 (hpd: 87.83 - 88.44), per ω^2 un valore di 358.11 (hpd: 318.79 - 375.23) (standard deviation: 18.92) e infine per α, un valore di asimmetria pari a -3.56 (hpd: -3.76 - -3.40). La stima dei valori della componente senescente, risentono maggiormente della divisione in cluster data la costruzione e l'intento dell'algoritmo. Nonostante questo, nelle differenti prove, anche mantenendo un numero di cluster costanti, il gruppo individuato alla fine del processo non era sempre il medesimo. Questo aspetto sottolinea ulteriormente come i differenti punti di partenza per ξ, come quelli citati per i 3 cluster in figura 3.1, non siano in realtà distintivi dei gruppi scelti, proprio perché la maggior parte della probabilità è condivisa.
In Figura A5 è rappresentato l'adattamento del modello proposto (curva arancione), rispetto al numero di decessi per ognuno dei diciassette paesi considerati. Vengono riportate anche la singola componente prematura (curva verde) e la sola componente senescente (curva blu).
L'adattamento del modello è abbastanza buono per la maggior parte dei paesi riportati. Particolare rilevanza si ha per Brasile, Repubblica Dominicana, Ecuador e Nicaragua per i quali la componente di mortalità prematura risulta ben colta. Dove il modello sembra fare più difficoltà è invece per età più longeve. Nello specifico, il modello identifica correttamente la forma, risultando però meno preciso nell'identificazione dei picchi di età dei decessi per vecchiaia. Specialmente per Costa Rica, Ecuador Messico e Panama il

modello tende a sottostimare la frequenza dei decessi per età attorno ai 90 anni. Tuttavia, si precisa che non essendo riusciti a definire cluster distinti, si costringono i paesi a più alta mortalità a condividere la componente senescente con i paesi a più bassa mortalità con un conseguente risultato negativo in termini di adattamento come si evidenzia per Costa Rica o Panama.

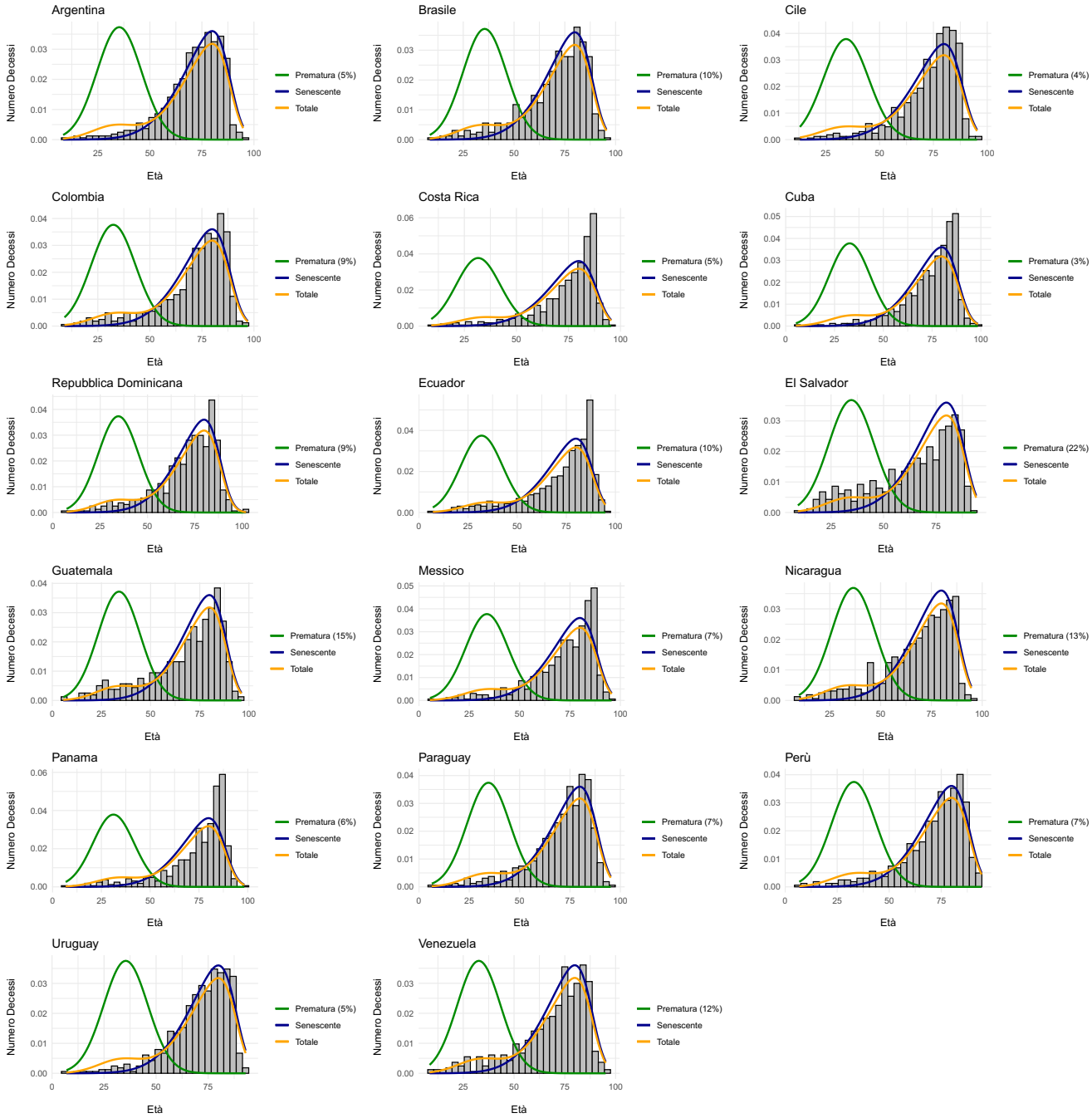

Figura 5.2: Rappresentazione delle componenti di mortalità per il modello proposto per i 17 paesi del Centro America considerati.

Come ultima analisi si è voluto testare nuovamente la capacità del modello di identificare cluster latenti in un gruppo di paesi considerati, sulla base della similarità della loro componente di mortalità senescente. Per fare questo, si è considerato un nuovo set di dati relativo a quindici paesi in tutto il mondo. I valori fanno riferimento all'anno 2018 e sono disponibili in *Human Fertility Databas* (https://www.mortality.org). I nuovi paesi considerati sono: Australia, Bielorussia, Bulgaria, Canada, Cile, Danimarca, Finlandia, Francia, Giappone, Irlanda, Italia, Lituania, Spagna, Svezia e Stati Uniti. Sebbene l'idea fosse considerare paesi che per caratteristiche di longevità fossero notoriamente molto più eterogenei, i risultati non sono stati migliori. Considerando sempre un numero di cluster pari a 2 o a 3, è stato mantenuto per più iterazioni un numero disgiunto di insiemi, rispetto al precedente set di paesi, per poi convergere tutti nuovamente all'interno di un unico gruppo. Per questo motivo, i risultati in termini di stima dei parametri e convergenza dell'algoritmo non vengono discussi. L'adattamento del modello a questo nuovo set è consultabile in Appendice.

Capitolo 6

Conclusioni

Il monitoraggio dei tassi di mortalità, in particolare della componente prematura è un aspetto di fondamentale importanza nelle politiche organizzative di un paese. La possibilità di individuare in maniera precisa una quota di morti evitabili, va da sé essere un aspetto imprescindibile nella pianificazione sanitaria di uno stato.
In questa tesi si è cercato di implementare un modello che fornisse non solo le stime dei singoli parametri, ma anche una possibile divisione dei paesi analizzati in gruppi latenti, sulla base della distribuzione dei decessi naturali. L'idea di identificare gruppi simili tra loro, permetterebbe agli stati con un maggior numero di decessi prematuri di uniformarsi a quelli con quote inferiori facenti parte dello stesso cluster modificando le proprie politiche interne, e sensibilizzando alla prevenzione nelle età maggiormente a rischio.
Se da una parte, il lavoro proposto vede un buon adattamento del modello alle distribuzioni dei decessi per le differenti età, l'identificazione di gruppi latenti sottostanti ai paesi considerati, non risultà del tutto apprezzabile. Come già sottolineato nei capitoli precedenti, sebbene possa effettivamente non esserci una così netta divisione in gruppi delle mortalità senescenti, i risultati riscontrati con il secondo set di paesi risultano incerti. Prima considerazione del perchè i risultati siano incoraggianti solo in parte, può essere attribuita alla scelta delle distribzuioni a priori ed in particolare, nell'assegnazione dei valori iniziali. Volendo rimanere non informativi e lasciare siano i dati a definire eventuali gruppi, risulta difficile per l'algoritmo identificare insiemi sulla base di distribuzioni che condividono la maggior parte della propria probabilità. In maniera opposta, essere troppo puntuali nella scelta dei valori, avvicinerebbe l'analisi ad un approccio di tipo assoluto, dove le distribuzioni a priori risulterebbero paragonabili alle soglie numeriche.
Altro aspetto di interesse per un possibile miglioramento della procedura riguarda la distinzione delle osservazioni premature e senescenti tramite le variabili latenti z_i. Questo passaggio risulta cruciale nel determinare le due componenti di mistura del modello. Nonostante la probabilità di generare le z_i sia aggiornata ad ogni passo tramite la verosi-

miglianza, età molto alte potrebbero comunque contribuire alla componente prematura, impoverendo di conseguenza le informazioni necessarie ad una corretta identificazione delle curve necessarie al clustering.
In conclusione, l'analisi condotta conferma l'importanza e le implicazioni che il concetto di mortalità riveste all'interno di un paese. I limiti, ma soprattutto i punti di forza evidenziati nel lavoro, possano essere un nuovo punto di partenza per ricerche future.

Bibliografia

[1] S. Mazzuco, M. Suhrcke, e L. Zanotto, «How to measure premature mortality? A proposal combining "relative" and "absolute" approaches», Popul Health Metrics, vol. 19, fasc. 1, p. 41, dic. 2021, doi: 10.1186/s12963-021-00267-y.

[2] United Nations. Transforming our world: the 2030 agenda for sustainable development. United Nations: Technical report; 2015.

[3] OECD. Health at a Glance 2009: OECD Indicators. OECD Publishing. 2009.

[4] CJL Murray, M. Ezzati, AD Flaxman, S. Lim, R. Lozano, C. Michaud, M. Naghavi, JA Salomon, K. Shibuya, T. Vos, D. Wikler, AD Lopez. Gbd 2010: design, definitions, and metrics. The Lancet. 2012;380(9859):2063–6. https:// doi. org/ 10. 1016/ S0140-6736(12) 61899-6.

[5] WHO: Targets and Indicators for Health 2020. WHO Regional Office for Europe; 2016.

[6] Eurostat: Health Statistics—Atlas on Mortality in the European Union.European Communities, 2009.

[7] J. Veron, J.-M. Rohrbasser, e J. Mandelbaum, «Wilhelm Lexis: The Normal Length of Life as an Expression of the "Nature of Things"», Population (English Edition, 2002-), vol. 58, fasc. 3, p. 303, mag. 2003, doi: 10.2307/3246676.

[8] V. Kannisto, «Measuring the compression of mortality», DemRes, vol. 3, p. 6, set. 2000, doi: 10.4054/DemRes.2000.3.6.

[9] V. Kannisto, «Mode and Dispersion of the Length of Life», Population: An English Selection, vol. 13, fasc. 1, pp. 159–171, 2001.

[10] S. L. K. Cheung, J.-M. Robine, E. J.-C. Tu, e G. Caselli, «Three dimensions of the survival curve: horizontalization, verticalization, and longevity extension», Demography, vol. 42, fasc. 2, pp. 243–258, mag. 2005, doi: 10.1353/dem.2005.0012..

[11] Eurostat: Amenable and preventable deaths statistics. Technical report, Eurostat Statistics Explained. 2016.

[12] K. Pearson. The chances of death and other studies in evolution, vol. I. London: Edward Arnold; 1897.

[13] L. Zanotto, V. Canudas-Romo, S. Mazzuco. A mixture-function mortality model: illustration of the evolution of premature mortality. Eur J Popul. 2020 (to appear).

[14] U. Basellini & C. G. Camarda (2019) Modelling and forecasting adult age-at-death distributions, Population Studies, 73:1, 119-138, DOI: 10.1080/00324728.2018.1545918

[15] S. Mazzuco, B. Scarpa & L. Zanotto (2018) A mortality model based on a mixture distribution function, Population Studies, 72:2, 191-200, DOI: 10.1080/00324728.2018.1439519

[16] D. M. Pascariu, A. Lenart & V. Canudas-Romo (2019) The maximum entropy mortality model: forecasting mortality using statistical moments, Scandinavian Actuarial Journal, 2019:8, 661-685, DOI: 10.1080/03461238.2019.1596974

[17] W. Siler A competing-risk model for animal mortality. Ecology. 1979;60(4):750–7. https:// doi. org/ 10. 2307/ 19366 12.

[18] L. Heligman, JH Pollard. The age pattern of mortality. J Inst Actuar. 1980;107(1):49–80.

[19] CG Camarda, U. Basellini Smoothing, decomposing and forecasting mortality rates. Eur J Popul. 2021. https:// doi. org/ 10. 1007/ s10680- 021- 09582-4.

[20] Li, Nan, and R.Lee. "Coherent Mortality Forecasts for a Group of Populations: An Extension of the Lee-Carter Method." Demography 42, no. 3 (2005): 575–94. http://www.jstor.org/stable/4147363.

[21] A. Palloni, G. Pinto, H. Beltrán-Sánchez. Latin American mortality database (LAMBdA). Madison: University of Wisconsin; 2014.

[22] A. Canale, E. C. K. Pagui & B. Scarpa (2016) Bayesian modeling of university first-year students' grades after placement test, Journal of Applied Statistics, 43:16, 3015-3029, DOI: 10.1080/02664763.2016.1157144

[23] S. Mazzuco e N. Keilman, A c. di, Developments in Demographic Forecasting, vol. 49. in The Springer Series on Demographic Methods and Population Analysis, vol. 49. Cham: Springer International Publishing, 2020. doi: 10.1007/978-3-030-42472-5.

[24] A. Azzalini e A. Capitanio, "The Skew-Normal and Related Families".

[25] A. Azzalini and A. Capitanio, Statistical applications of the multivariate skew normal distribution, J. R. Stat. Soc. Ser. B 61 (1999), pp. 579–602.

[26] A. Azzalini and A. Capitanio, Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t-distribution, J. R. Stat. Soc. Ser. B 65 (2003), pp. 367–389.

[27] A. Azzalini and A. Dalla Valle, The multivariate skew-normal distribution, Biometrika 83 (1996), pp. 715–726.

[28] M. Branco and D.Dey, A general class ofmultivariate skew-elliptical distributions, J.Multivariate Anal. 79 (2001), pp. 93–113.

[29] M.G. Genton and N.M.R. Loperfido, Generalized skew-elliptical distributions and their quadratic forms, Ann. Inst. Statist. Math. 57 (2005), pp. 389–401.

[30] R. B. Arellano-Valle e A. Azzalini, «On the Unification of Families of Skew-normal Distributions», Scandinavian J Statistics, vol. 33, fasc. 3, pp. 561–574, set. 2006, doi: 10.1111/j.1467-9469.2006.00503.x.

[31] M.W. Liechty and J. Lu,Multivariate normal slice sampling, J. Comput.Graph. Statist. 19 (2010), pp. 281–294.

[32] A. Pakman and L. Paninski, Exact Hamiltonian Monte Carlo for truncated multivariate Gaussians, J. Comput. Graph. Statist. 23 (2014), pp. 518–542

[33] Z. I. Botev, «The Normal Law Under Linear Restrictions: Simulation and Estimation via Minimax Tilting», Journal of the Royal Statistical Society Series B: Statistical Methodology, vol. 79, fasc. 1, pp. 125–148, feb. 2016, doi: 10.1111/rssb.12162.

[34] Z. Zhang, A. Chin, A. Nishimura, e M. A. Suchard, «hdtg: An R package for high-dimensional truncated normal simulation». arXiv, 22 settembre 2022. Consultato: 23 febbraio 2024. [Online]. Disponibile su http://arxiv.org/abs/2210.01097.

[35] C. Bouveyron, G. Celeux, T. B. Murphy, and A. E. Raftery. Model-Based Clustering and Classification for Data Science: With Applications in R. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, England, 2019a. doi: 10.1017/9781108644181.

[36] S. B. Everitt, S. Landau, M. Leese, and D. Stahl. Finite Mixture Densities as Models for Cluster Analysis, chapter 6, pages 143–186. John Wiley & Sons, Ltd, 2011. doi: 10.1002/ 9780470977811.ch6.

[37] D. M. Titterington, S. Afm, A. F. Smith, U. Makov, et al. Statistical analysis of finite mixture distributions, volume 198. John Wiley & Sons Incorporated, New York, 1985.

[38] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society: Series B (Methodological), 39(1):1–22, 1977. doi: 10.1111/j.2517-6161.1977.tb01600.x.

Appendice

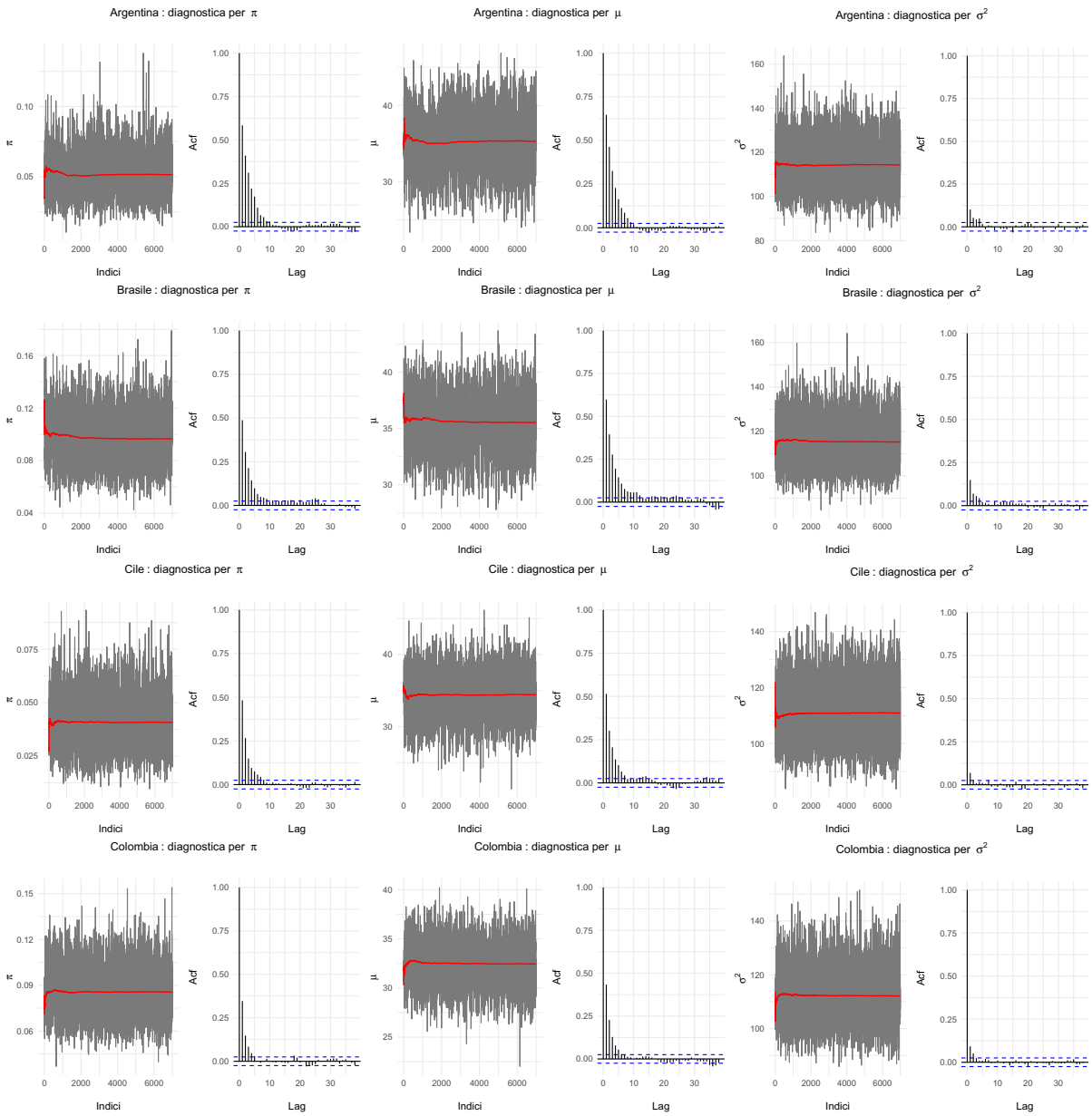

Figura A1: Diagnostica di convergenza per i parametri relativi alla componente premtura per Argentina, Brasile, Cile e Colombia.

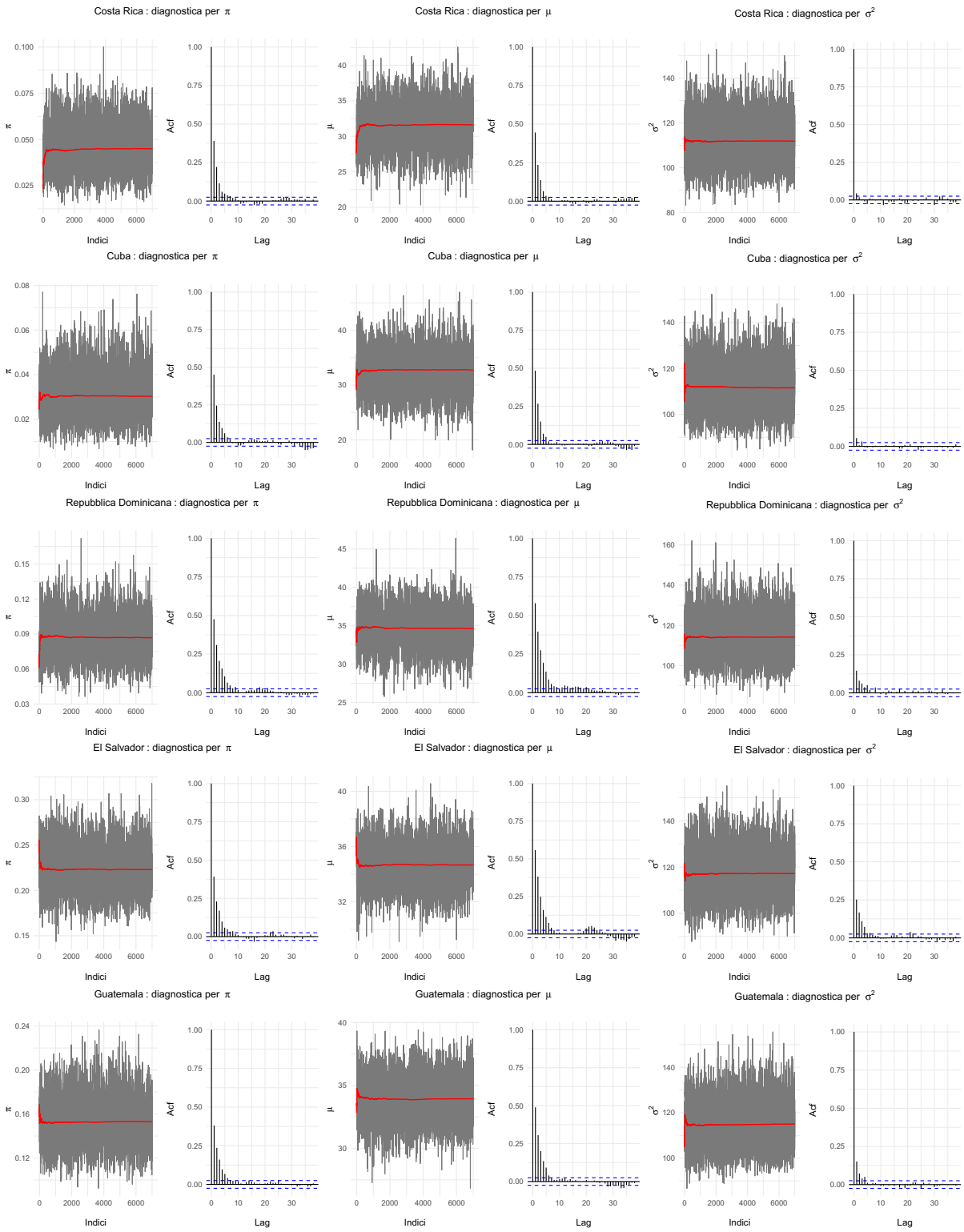

Figura A2: Diagnostica di convergenza per i parametri relativi alla componente premtura per Costa Rica, Cuba, Repubblica Dominicana, El Salvador e Guatemala.

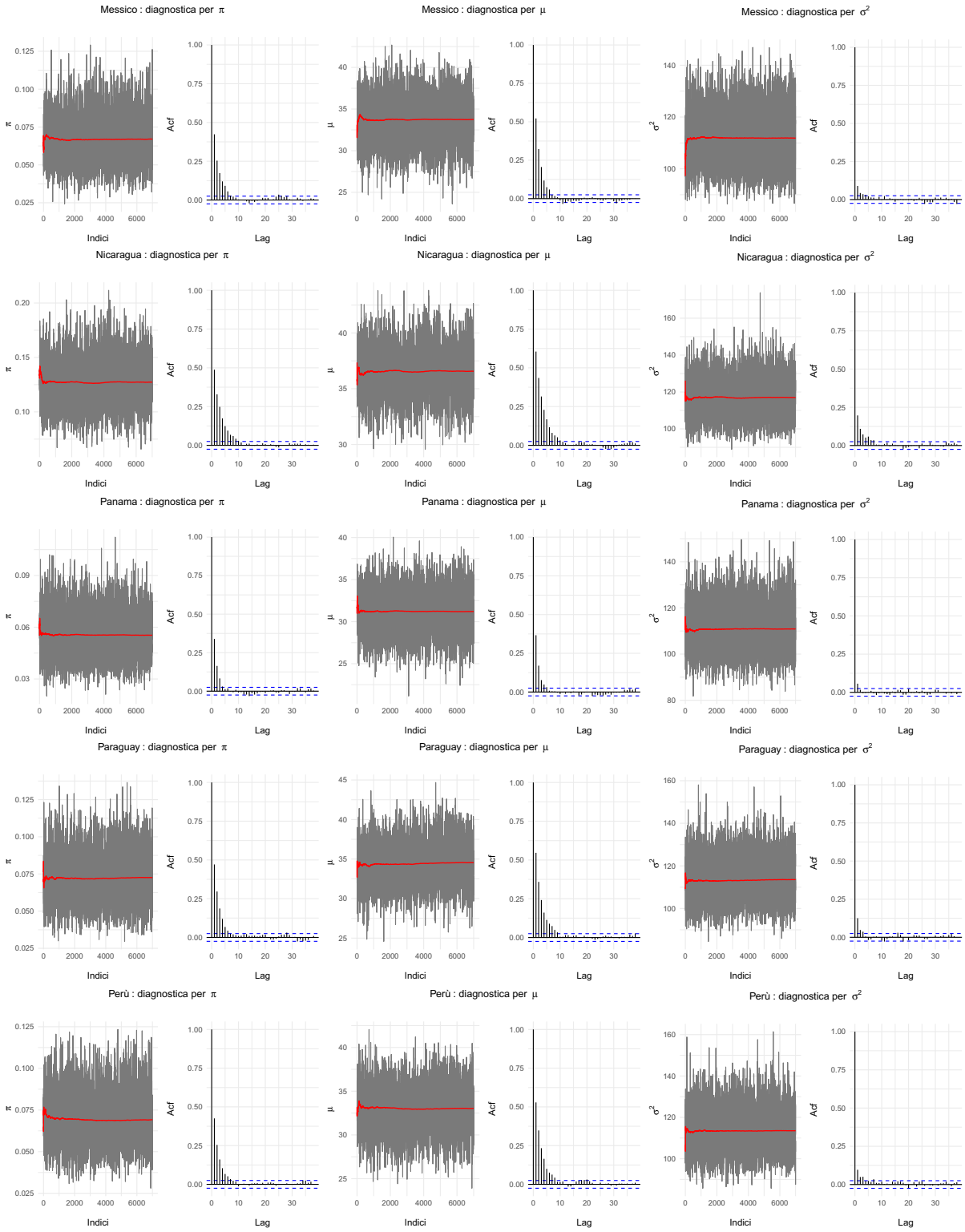

Figura A3: Diagnostica di convergenza per i parametri relativi alla componente premtura per Messico, Nicaragua, Panama, Paraguay e Perù.

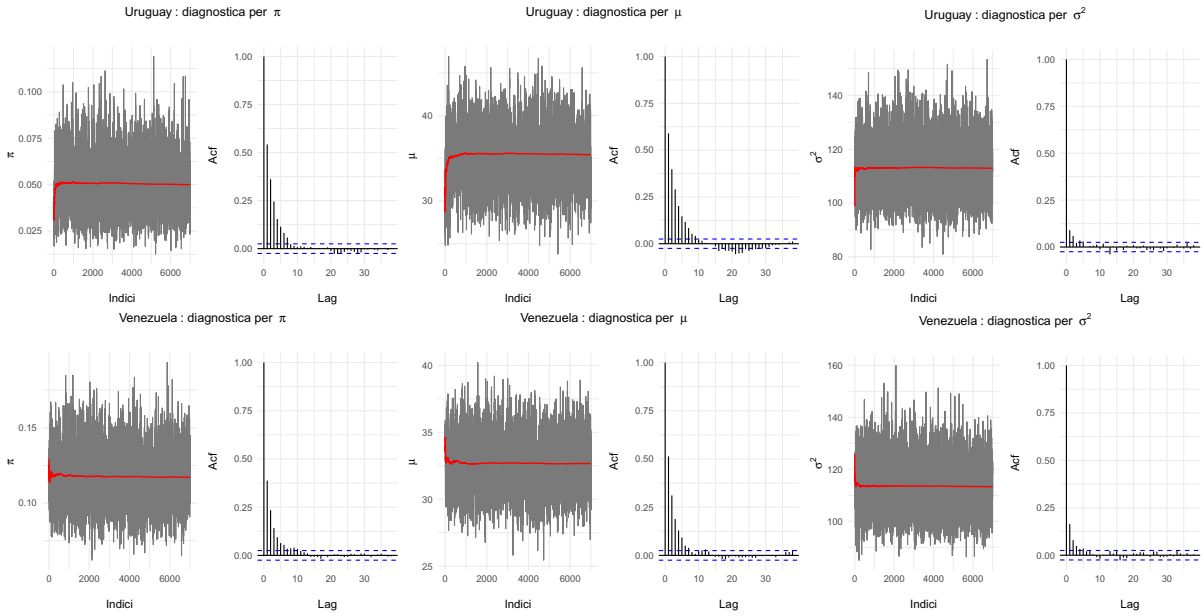

Figura A4: Diagnostica di convergenza per i parametri relativi alla componente premtura per Uruguay e Venezuela.

	Australia	**Bielorussia**	**Bulgaria**	**Canada**	**Cile**
1	1.82 (0.03-3.2)	12.18 (4.97-16.31)	8.07 (2.36-11.91)	2.34 (0.27-3.81)	2.93(0.62-4.56)
	Danimarca	**Finlandia**	**Francia**	**Giappone**	**Irlanda**
1	1.76 (0.02-3.21)	1.95 (0.18-3.3)	2.28 (0.09-3.86)	1.44 (0.01-2.65)	1.80 (0.17-3.11)
	Italia	**Lituania**	**Spagna**	**Svezia**	**Stati Uniti**
1	1.31 (0-2.42)	9.19 (3.57-12.53)	0.95 (0-2.11)	1.47(0-2.67)	5.12 (1.98-7.28)

Tabella A1: Valori stimati su scala percentuale per il parametro π_j per ogni paese del secondo set considerato. Tra parentesi, l'intervallo hpd.

	Australia	**Bielorussia**	**Bulgaria**	**Canada**	**Cile**
μ_j	38.81	48.70	46.36	36.74	38.61
σ_j^2	110.54 (10.51)	116.5 (10.79)	117.65 (10.85)	110.25 (10.5)	110.94 (10.53)
	Danimarca	**Finlandia**	**Francia**	**Giappone**	**Irlanda**
μ_j	39.70	37.78	39.17	39.21	37.55
σ_j^2	111.25 (10.55)	110.77 (10.52)	111.05 (10.54)	110.72 (10.52)	110.61 (10.52)
	Italia	**Lituania**	**Spagna**	**Svezia**	**Stati Uniti**
μ_j	37.78	44.12	43.41	38.21	38.83
σ_j^2	110.6 (10.52)	115.76 (10.76)	110.53 (10.51)	110.6 (10.52)	111.12 (10.54)

Tabella A2: Valori stimati per i parametri μ_j e σ_j^2 per ogni paese del secondo set considerato. Tra parentesi, è indicato il valore della deviazione standard.

Per la componente senescente si ha: ξ pari a 91.53 (hpd: 91.16 - 91.71), ω^2 pari a 293.48 (hpd: 265.02 - 307.32) e α pari a -3.91 (hpd: -4.20 - -3.72).

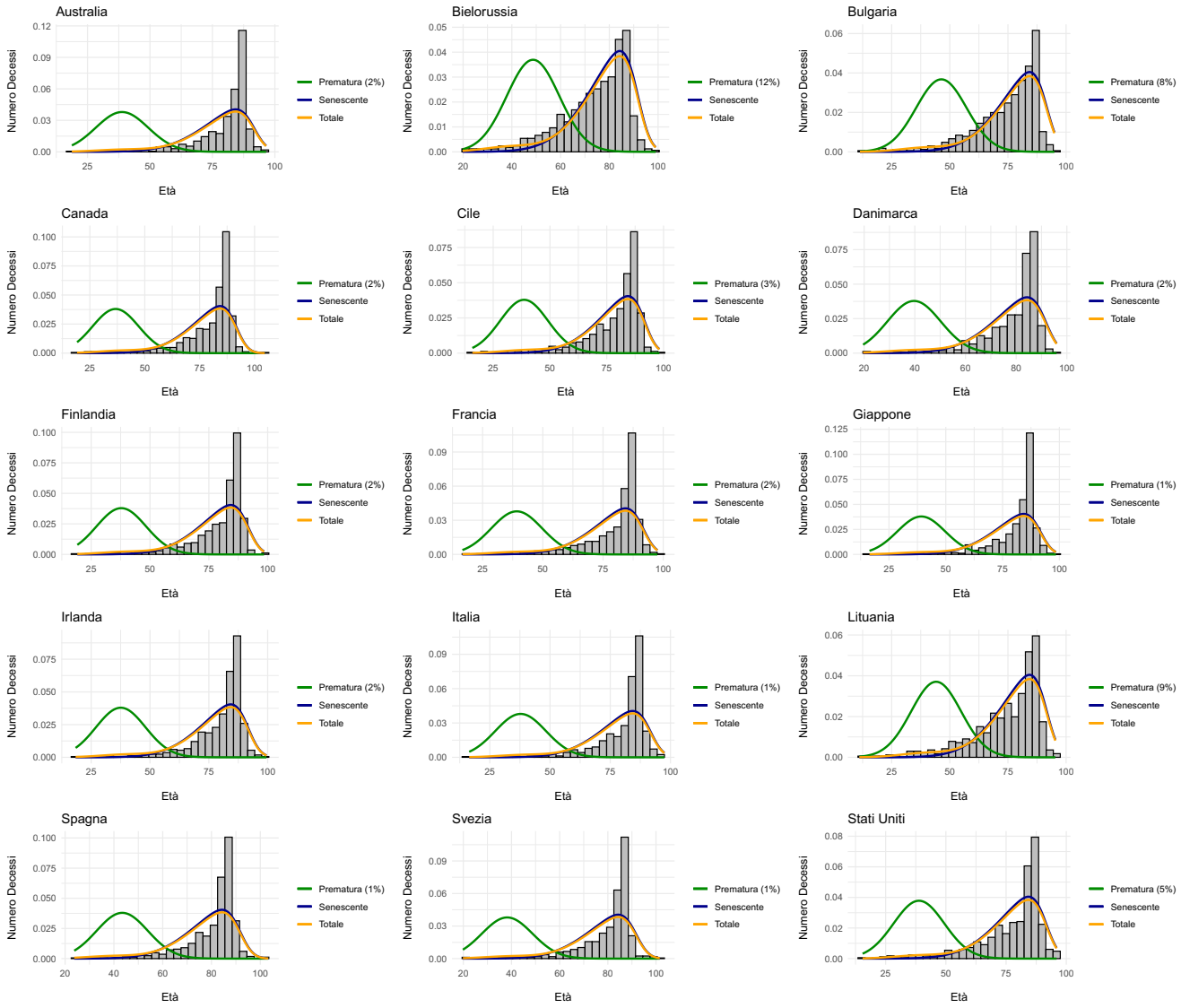

Figura A5: Rappresentazione delle componenti di mortalità per il modello proposto per il secondo set di paesi.

In conclusione, vengono riportate le funzioni e i frammenti di codice principali, utilizzati nella realizzazione della tesi.

```
library(sn)
library(truncnorm)
library(hdtg)

dskn_custom = function(y, xi, omega, sk){
 return(sum(dsn(y, xi = xi, omega = omega, alpha = sk, log = T)))
}

NORM_prematura = function(y, sigma, mu0, sigma0, a, b){
 media = rnorm(1, mean = (((mu0/sigma0) + (sum(y)/sigma)) /
 ((1/sigma0) + (length(y)/sigma))),
 sd  = sqrt(1/((1/sigma0) + (length(y)/sigma))))
 varianza = 1/rgamma(1, (a + length(y)/2),
 (b + (sum((y-media)^2)/2)))
 return((list(media = media, varianza = varianza)))
}

RSUN = function(N0, alpha0, gamma, phi0, delta.cap){
 GAMMA = matrix(diag(1, N0 + 1) - diag(delta.cap^2) +
 (delta.cap%*%t(delta.cap)), (N0 + 1))
 zeros = numeric(N0+1)
 V0 = zigzagHMC(n = 1, mean = zeros, prec = GAMMA,
 lowerBounds = -gamma,
 upperBounds = rep(Inf, length(zeros)))
 V1 = rnorm(1)
 A  = diag(1/(1 - delta.cap^2))
 B  = 1/(1 + sum((delta.cap^2) / (1 - delta.cap^2)))
 C1 = outer(delta.cap, delta.cap)
 C2 = outer((1 / (1 - delta.cap^2)), (1 / (1 - delta.cap^2)))
 C  = matrix(C1*C2, length(delta.cap))
 GAMMA_1 = A - B * C
 alpha = alpha0 + phi0*(t(delta.cap) %*% GAMMA_1 %*% t(V0) +
 V1 %*% sqrt(1-t(delta.cap) %*% GAMMA_1 %*% delta.cap))
 return(alpha)
}
```

```
gibbs = function(R, y, country, J, G,
pi0, mu, sigma, mu0, sigma0, a, b,
xi, omega, alpha, xi0, k, A, B,
alpha0, phi0, lambda0){

 N = length(y)
 #Cluster
 CLS = matrix(NA, R, J) #matrice salvataggio cluster
 pi_g = rep(1 / G, G) #proporzione dei cluster
 cls = sample(1:G, J, prob = pi_g, replace = T)

 #Proporzione mistura
 pi = matrix(NA, R+1, J)
 pi[1,] = pi0

 #Mortalita prematura
 mu_m = matrix(NA, R+1, J) #MEDIA prematura
 mu_m[1,] = mu
 sigma_m = matrix(NA, R+1, J) #VARIANZA prematura
 sigma_m[1,] = sigma

 #Mortalita senescente
 xi_M = matrix(NA, R+1, G) #POSIZIONE senescente
 xi_M[1,] = xi
 omega_M = matrix(NA, R+1, G) #SCALA senescente
 omega_M[1,] = rep(omega, G)
 alpha_M = matrix(NA, R+1, G) #FORMA senescente
 alpha_M[1,] = rep(alpha, G)

 #GIBBS
 for (r in 1:R){

  zi = numeric(N)

  #MORTALITA PREMATURA
  for(j in 1:J){
```

```
 # #variabili latenti prematura/senescente
 PZ1 = pi[r,j] * apply(matrix(y[country == j], nrow = 1), 2,
 dnorm, mean = mu_m[r,j], sd = sqrt(sigma_m[r,j])) /
 ((pi[r,j] * apply(matrix(y[country == j], nrow = 1), 2, dnorm,
 mean = mu_m[r,j], sd = sqrt(sigma_m[r,j]))) +
 ((1 - pi[r,j]) * apply(matrix(y[country == j], nrow = 1), 2, dsn,
 xi = xi_M[r,cls[j]], omega = sqrt(omega_M[r,cls[j]]),
 alpha = alpha_M[r,cls[j]])))

 zi[country == j] = sapply(PZ1, function(p) rbinom(1, 1, p))
 stime = NORM_prematura(y[country == j][zi[country == j] == 1],
 sigma_m[r, j], mu0, sigma0 , a, b)

 mu_m[(r+1), j] = stime$media
 sigma_m[(r+1), j] = stime$varianza

 #proporzione mistura prematura/senescente
 pi[(r+1), j] = rbeta(1, (1 + sum(zi[country == j])),
 (1 + length(y[country == j]) - sum(zi[country == j])))

 #CLUSTERING
 probabilities = numeric(length(pi_g))
 for (g in sort(unique(cls))) {
  probabilities[g] = pi_g[g] *
  dskn_custom(y[country == j][zi[country == j] == 0],
  xi_M[r, g], sqrt(omega_M[r, g]), alpha_M[r, g])
 }
 probabilities = probabilities / sum(probabilities)
 CLS[r,j] = sample(1:G, 1, prob = probabilities)
}

cls = CLS[r,] #cluster
for (g in sort(unique(cls))){
 #dati "parziali" di questo cluster con zi=0 (senescente)
 y_g0 = y[country %in%
 c(which(cls == g))][zi[country %in% c(which(cls == g))]==0]
```

```
  N0 = length(y_g0)

  #MORTALITA SENESCENTE
  eta = numeric(N0) #variabili latenti
  delta = alpha_M[r, g]/sqrt(1 + alpha_M[r, g]^2)

  for(i in 1:N0){
   eta[i] = rtruncnorm(1, a = 0,
   mean = delta*(y_g0[i] - xi_M[r, g]),
   sd   = sqrt(omega_M[r, g]*(1 - delta^2)))
  }

  mu.hat = (k * sum(y_g0 - (delta * eta)) +
  ((1 - delta^2) * xi0)) / ((N0 * k) + (1 - (delta^2)))
  k.hat  = (k * (1 - delta^2)) / ((N0 * k) + (1 - delta^2))
  b.hat  = (1 / (2 * (1 - (delta^2)))) * (((delta^2) * sum(eta^2)) -
  (2 * delta * sum(eta * (y_g0 - xi_M[r, g]))) +
  sum((y_g0 - xi_M[r, g])^2) +
  (((1 - (delta^2))/k) * ((xi_M[r, g] - xi0)^2)))

  xi_M[(r+1), g] = rnorm(1, mu.hat, sd = sqrt(k.hat * omega_M[r, g]))
  omega_M[(r+1), g] = 1/rgamma(1, A + ((N0 + 1)/2), B + b.hat)

  y.star = (y_g0 - xi_M[r, g])/sqrt(omega_M[r, g])
  z = c(phi0*y.star, lambda0)
  delta.cap = (phi0 * z) / sqrt(((phi0 * z)^2) + 1)
  gamma = c(delta.cap[1:N0] * (alpha0 / phi0), 0)

  alpha_M[(r+1), g] = RSUN(N0, alpha0, gamma, phi0, delta.cap)

  if (r%%100 == 0 | r == 1){
   cat("Iterazione:", r, "- cluster:", g, "di", length(which(pi_g != 0)), "\n")
  }
 }#iterazione del cluster g
 pi_g = as.numeric(table(factor(cls, levels = factor(1:G))) / length(cls))
}#iterazione replicazione r
```

```
 return(list(pi = pi[2:(R+1),],      #proporzione mistrura
           mu_m = mu_m[2:(R+1),],    #media PREMATURA
         sigma_m = sigma_m[2:(R+1),], #varianza PREMATURA
           xi_M = xi_M[2:(R+1),],    #posizione SENESCENTE
         omega_M = omega_M[2:(R+1),], #scala SENESCENTE
         alpha_M = alpha_M[2:(R+1),], #forma SENESCENTE
         cluster = CLS))             #cluster dei paesi
}
```