Università degli Studi di Padova

Dipartimento di Scienze Statistiche

Corso di Laurea Magistrale in

Scienze Statistiche

Relazione Finale

La mortalità prematura in Francia: un modello gerarchico bayesiano per l'analisi delle cause di morte

Relatore: Prof. Stefano Mazzuco

Dipartimento di Scienze Statistiche

Correlatore: Dott.ssa Lucia Zanotto

Dipartimento di Scienze Statistiche

Laureanda: Chiara Pastrello

Matricola N. 1106610

Anno Accademico 2016/2017

"Data! Data! Data!" he cried impatiently.
"I can't make bricks without clay."

– sir Arthur Conan Doyle

Indice

Elenco delle figure

Elenco delle tabelle

Introduzione

La mortalità è uno degli aspetti principali che viene studiato in una popolazione. Si tratta di un tema complesso da trattare poiché la forma della distribuzione dei decessi ha subito continue trasformazioni nel tempo e può variare molto a seconda delle caratteristiche del gruppo di individui oggetto di interesse quali genere, età, condizione socio-economica e Paese di appartenenza.

Questo lavoro ha origine dalla tesi di dottorato di Lucia Zanotto (2016) «A mixture model to distinguish mortality components» nella quale è stato proposto un nuovo modello parametrico per studiare la distribuzione dei decessi per età e le sue componenti. In particolare, ispirandosi alla teoria di Pearson (1897) sulle componenti della mortalità, Zanotto ha analizzato i decessi della tavola di mortalità di diversi Paesi attraverso un modello composto da una mistura di tre distribuzioni: una semi-normale per la mortalità infantile e due normali asimmetriche, una per la mortalità accidentale e prematura ed un'altra per la mortalità adulta. L'attenzione è stata focalizzata soprattutto sullo studio delle caratteristiche e dei cambiamenti nel tempo della mortalità accidentale e prematura, componente che non sempre è stata considerata in letteratura in quanto difficile da riconoscere e distinguere dalle altre. Uno dei risultati significativi emersi da tale lavoro è che in Francia negli ultimi anni è stato osservato un aumento del numero di decessi associati alla mortalità prematura, senza tuttavia riuscire a trovare dei motivi adeguati per giustificare questo fenomeno. L'obiettivo principale di questa tesi è dunque quello di accogliere questa domanda di ricerca e provare a fornire una valida spiegazione attraverso l'analisi delle cause di morte. Seguendo questa strada, si è giunti alla costruzione

di un nuovo modello gerarchico bayesiano, il quale ha permesso di chiarire alcuni aspetti della questione e dare una possibile interpretazione a quanto rilevato in precedenza.

Nel Capitolo 1 si accenna brevemente alle teorie demografiche di Lexis (1879) e Pearson (1897) su come discriminare tra le componenti della mortalità; in seguito viene introdotto il modello mistura di Zanotto (2016), la sua specificazione e la stima dei parametri attraverso il metodo della massima verosimiglianza. Inoltre, si riportano i risultati principali emersi dalla stima del modello per diversi Paesi e viene esposto l'argomento principale di questa tesi, ovvero l'aumento della mortalità prematura in Francia a partire dal 1990.

Il Capitolo 2 è dedicato alla descrizione dei dati utilizzati in questo lavoro per l'analisi delle cause di morte. È stato possibile accedere ad una nuova fonte di informazioni, *The Human Cause-of-Death Database* (HCD), una recente banca dati di alta qualità da cui sono state estratte delle serie storiche di mortalità per causa disponibili per la Francia dal 2000 al 2013, contenenti i conteggi per classi d'età dei decessi della popolazione classificati in base ad una lista di 16 gruppi di cause, distintamente per genere. Poiché i dati che entrano nella stima del modello mistura del Capitolo 1 sono i decessi di una tavola di mortalità, si è spiegato in che modo sono state utilizzate le informazioni originali per costruire delle tavole a decremento multiplo da cui ricavare le serie di dati necessarie per stimare i modelli. La rimanente parte del capitolo è dedicata ad alcune analisi esplorative preliminari di tali dati e al confronto della mortalità per causa tra uomini e donne.

All'inizio del Capitolo 3 viene definita brevemente la distribuzione normale asimmetrica, riportando alcuni risultati noti in letteratura quali proprietà, momenti e una parametrizzazione alternativa che risolve alcuni problemi in campo inferenziale e che sarà adottata in seguito. Sulla base di quanto emerso dalle statistiche descrittive, si è deciso di concentrare le analisi soprattutto sugli uomini, colpiti in misura maggiore dalla mortalità prematura rispetto alle donne. Dopo opportuni adattamenti, il modello mistura introdotto nel Capitolo 1 è

stato stimato attraverso il metodo della massima verosimiglianza marginalmente per ciascuna delle cause di morte analizzate per i decessi maschili negli anni 2000-2013. Purtroppo, i risultati non sono stati soddisfacenti a causa di problematiche di diversa natura che hanno reso le stime di massima verosimiglianza ottenute poco attendibili. Tuttavia, queste analisi sono state sfruttate come punto di partenza per la creazione di un nuovo modello.

Nel Capitolo 4 viene presentato il modello gerarchico bayesiano proposto in questa tesi per l'analisi delle cause di morte, motivando le varie scelte che hanno portato alla sua specificazione e descrivendo il metodo utilizzato per stimarlo con il software Stan, un linguaggio di programmazione recentemente sviluppato per supportare l'inferenza bayesiana attraverso metodi Markov Chain Monte Carlo (MCMC) che si basano su algoritmi Hamiltonian Monte Carlo (HMC).

I risultati della stima del modello gerarchico sono presentati nel Capitolo 5, i quali si sono dimostrati molto soddisfacenti e hanno permesso di dare un'interpretazione ragionevole all'aumento della mortalità prematura in Francia.

Infine, nel Capitolo 6 sono riportate in breve alcune considerazioni emerse provando a replicare tutte le precedenti analisi anche per le donne.

In Appendice è mostrato un esempio di codice scritto in Stan utilizzato per stimare il modello gerarchico.

Capitolo 1

Un modello mistura per distinguere le componenti della mortalità

1.1 Alcune teorie demografiche per descrivere la mortalità

La mortalità è cambiata nel tempo e può essere molto diversa tra nazioni. La forma della distribuzione dei decessi non è semplice da approssimare, pertanto per studiarne l'evoluzione ed i cambiamenti è necessario distinguerne le componenti, che solitamente sono cinque. La mortalità *infantile* riguarda i decessi all'età 0, quella *giovanile* i decessi tra il primo anno di vita e l'inizio dell'adolescenza; con mortalità *accidentale* si intende la piccola "gobba" che qualche volta è visibile nella curva della mortalità attorno ai 20-25 anni soprattutto per gli uomini, mentre la mortalità *adulta* comprende le morti nell'ultima fase della vita. I decessi che avvengono tra queste ultime due componenti costituiscono la mortalità *prematura*. La curva della mortalità prematura può essere difficile da distinguere da quella accidentale ed in particolare da quella adulta; infatti, spesso si sovrappone in parte ad esse, rendendo poco visibile dov'è la fine di

una e l'inizio dell'altra e sembrando quasi un'unica distribuzione. Non esiste un chiaro punto di separazione o range d'età che suggerisca quale sia l'esatta posizione di tale componente. Per tutti questi motivi, la mortalità prematura non ha una precisa definizione e non sempre è stata tenuta in considerazione dai più noti modelli parametrici di mortalità, come il modello di Gompertz (1825), quello di Heligman e Pollard (1980), ecc.

Nella letteratura demografica si possono trovare diverse teorie su come discriminare tra le componenti della mortalità, di cui le due principali appartengono a Lexis e Pearson. Lexis (1879) classifica i decessi in tre gruppi: le morti infantili, che avvengono nella primissima infanzia, le "morti normali", che si concentrano attorno ad un valore modale nell'età adulta e costituiscono una curva simmetrica, e le morti premature, che si sovrappongono parzialmente a quelle adulte, in una regione di transizione tra i primi due gruppi. Nella Figura 1.1 si possono distinguere le corrispondenti tre parti nella distribuzione dei decessi secondo Lexis, contrassegnate con simboli diversi.

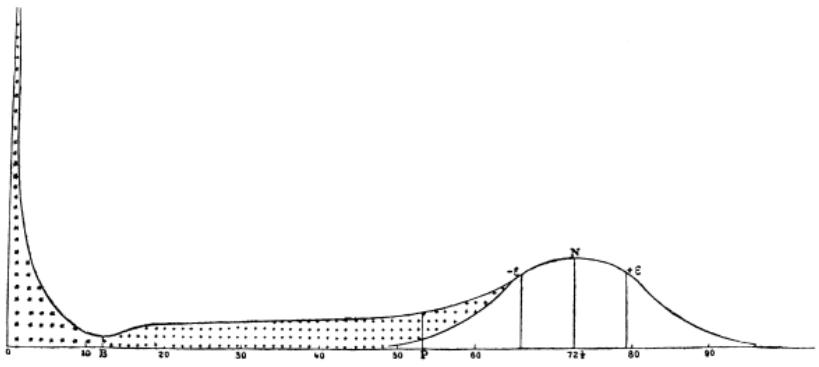

Figura 1.1: Le tre componenti della distribuzione dei decessi per età secondo Lexis (*Fonte*: Lexis, 1879).

Pearson (1897) assegna ad ogni componente della mortalità una distribuzione; a suo parere, la distribuzione dei decessi è composta da una mistura di cinque funzioni con diversi gradi di asimmetria. Egli separa la mortalità infantile da quella giovanile e suddivide la regione di transizione di Lexis in mortalità

accidentale e mortalità prematura, modellate entrambe con una distribuzione normale simmetrica con moda attorno ai 25 e 40 anni, rispettivamente. Infine, Pearson rappresenta la mortalità adulta con una distribuzione asimmetrica, in particolare con asimmetria negativa.

Oltre al numero di componenti considerate, la principale differenza tra le due teorie è che per Lexis la mortalità prematura non ha caratteristiche proprie, è una regione di transizione e può essere identificata solo in relazione alle morti "normali" e infantili, mentre Pearson assegna ad ogni componente la propria distribuzione, pertanto anche la mortalità prematura ha una sua specifica funzione matematica e non è più vista solo come conseguenza delle altre. Un'altra importante differenza riguarda la forma della curva della mortalità adulta: Lexis propone una distribuzione simmetrica, mentre Pearson una asimmetrica, giustificando questa scelta sostenendo che il numero di decessi che si verificano nelle età più avanzate e dunque anche la forma della distribuzione dipendono dall'incidenza delle morti avvenute nelle età precedenti.

Questa tesi ha origine dal lavoro di Zanotto (2016), in cui viene rielaborata la teoria di Pearson sulle componenti della mortalità. Tuttavia, seguendo rigorosamente tale approccio si sarebbe ottenuto un modello mistura con almeno 13 parametri, che avrebbe inevitabilmente comportato problemi di identificabilità. Nel prossimo paragrafo si descrive qual è la soluzione adottata.

1.2 Il modello mistura

Nella tesi di dottorato di Lucia Zanotto 2016 «A mixture model to distinguish mortality components» è stato proposto un nuovo modello parametrico per studiare la distribuzione dei decessi per età e le sue componenti. È stata posta particolare attenzione sull'evoluzione nel tempo della mortalità accidentale e prematura in diversi Paesi, componente che spesso non viene considerata perché è difficile da riconoscere. Il nuovo modello include, infatti, una funzione specifica per modellare la mortalità prematura e distinguerla da quella adulta.

Ispirandosi alla teoria di Pearson ma volendo anche limitare il numero di parametri, Zanotto ha costruito un modello mistura di tre distribuzioni: una semi-normale, per la mortalità infantile, e due normali asimmetriche, una per la mortalità accidentale e prematura ed un'altra per la mortalità adulta.

La distribuzione semi-normale (o HN, da *Half Normal*) ha una forma molto simile a quella della prima parte della distribuzione dei decessi, è definita solo per valori maggiori di 0 ed ha la seguente funzione di densità di probabilità:

$$f_I(x;1) = \frac{\sqrt{2}}{\sqrt{\pi}} \exp\left(-\frac{x^2}{2}\right) \qquad x > 0 \tag{1.1}$$

dove x è l'età alla morte e il pedice I indica che si tratta della mortalità infantile. In origine la (1.1) includeva anche un parametro σ, in seguito fissato pari a 1, valore che permette di descrivere in modo opportuno la prima parte della curva di mortalità e ad evitare l'impiego di un numero eccessivo di parametri.

Per descrivere la mortalità accidentale e prematura (m) e quella adulta (M) sono state adottate due distribuzioni normali asimmetriche (o SN, da *Skew Normal*). Le corrispondenti funzioni di densità di probabilità sono:

$$f_m(x;\theta_m) = \frac{2}{\omega_m}\phi\left(\frac{x-\xi_m}{\omega_m}\right)\Phi\left(\lambda_m\frac{x-\xi_m}{\omega_m}\right) \tag{1.2}$$

$$f_M(x;\theta_M) = \frac{2}{\omega_M}\phi\left(\frac{x-\xi_M}{\omega_M}\right)\Phi\left(\lambda_M\frac{x-\xi_M}{\omega_M}\right) \tag{1.3}$$

dove $\phi(\cdot)$ e $\Phi(\cdot)$ sono le funzioni di densità e di ripartizione della normale standard, il pedice m indica la distribuzione della mortalità accidentale e prematura, e M quella della mortalità adulta. Entrambe le distribuzioni hanno tre parametri: $\theta_m = (\xi_m, \omega_m, \lambda_m)$ e $\theta_M = (\xi_M, \omega_M, \lambda_M)$, dove $\xi_{(\cdot)} \in \mathbb{R}$ è il parametro di posizione, $\omega_{(\cdot)} \in \mathbb{R}^+$ il parametro di scala e $\lambda_{(\cdot)} \in \mathbb{R}$ il parametro di forma.

La famiglia delle normali asimmetriche verrà trattata con maggior dettaglio all'inizio del Capitolo 3 di questa tesi. Tuttavia, questa classe di distribuzioni comprende la normale come caso particolare ($\lambda = 0$), proprietà da cui deriva la scelta di tale funzione per la mortalità adulta, in modo da poter verificare se per rappresentarla è opportuno utilizzare una distribuzione potenzialmente

asimmetrica, in linea con la teoria di Pearson, oppure una simmetrica, in accordo con Lexis.

Per modellare congiuntamente mortalità accidentale e prematura è stata usata un'altra normale asimmetrica, in quanto è sembrata la distribuzione più adatta a cogliere la forma della curva di mortalità nella sua parte centrale.

Il modello proposto da Zanotto (2016) ha origine dalla combinazione delle tre funzioni (1.1), (1.2) e (1.3) con i parametri (o pesi) di mistura $\eta \in [0, 1]$ e $\alpha \in [0, 1]$ ed ha la seguente specificazione:

$$f(x;\theta) = \eta \cdot f_I(x;1) + (1-\eta) \cdot \Big[\alpha f_m(x;\theta_m) + (1-\alpha) f_M(x;\theta_M)\Big] \qquad (1.4)$$

dove θ è il vettore degli 8 parametri e x è l'età alla morte. La distribuzione dei decessi è definita solo per valori positivi di x ed è compresa nell'intervallo $x \in [0, \Omega]$, dove Ω è considerato circa pari a 110 anni. Nella Figura 1.2 si può vedere una rappresentazione semplificata del modello e delle sue componenti.

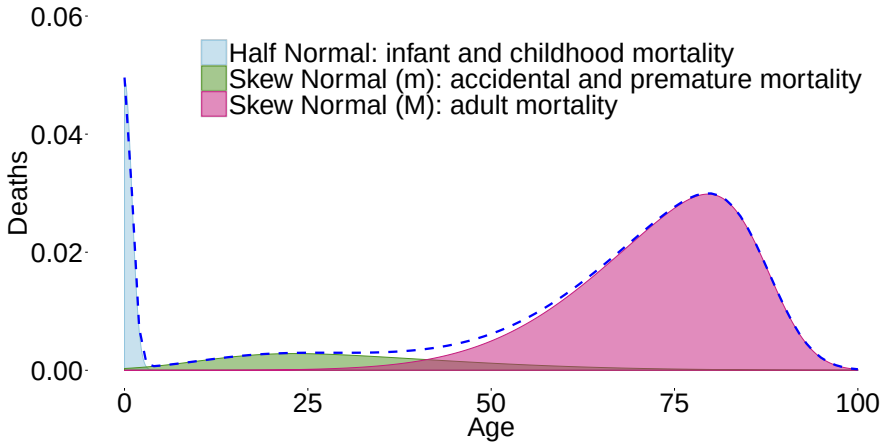

Figura 1.2: Una semplificata distribuzione dei decessi (linea blu tratteggiata) e le tre funzioni del modello mistura (*Fonte*: Zanotto, 2016).

Tutti i parametri del modello (1.4) hanno un'esplicita interpretazione demografica, che facilita l'analisi dei risultati e i confronti; in particolare, il primo parametro di mistura η indica l'intensità della mortalità infantile, mentre il secondo parametro di mistura α descrive l'importanza della mortalità accidentale

e prematura nella distribuzione complessiva, ovvero più elevato è il valore di α, maggiore è il ruolo assunto da questa componente rispetto a quella adulta, e viceversa, come mostrato nella Figura 1.3 per alcuni valori di α.

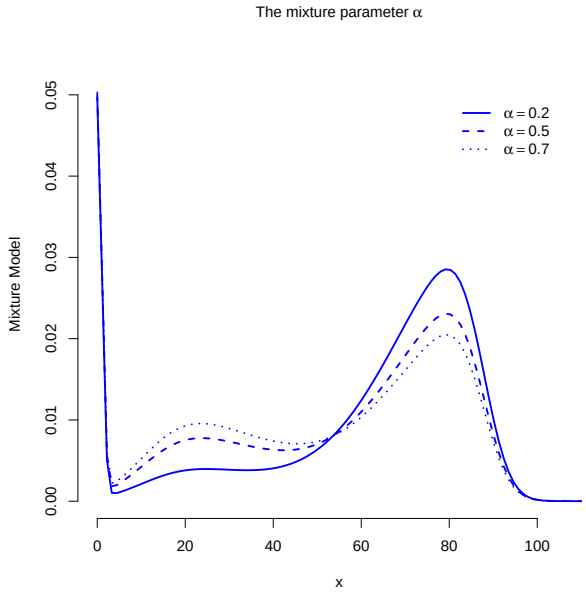

Figura 1.3: Funzione di densità del modello mistura per diversi valori di α (*Fonte*: Zanotto, 2016).

1.3 Il metodo di stima dei parametri

I dati che entrano nella stima del modello (1.4) sono i decessi d_x della tavola di mortalità, della quale si rimandano definizione e funzioni principali al Capitolo 2 di questa tesi. Infatti, i dati a disposizione di Zanotto (2016) sono in forma aggregata: non è nota l'età esatta x del decesso di ciascun individuo, ma si conosce il numero di morti per ogni classe d'età. Nonostante la distribuzione dei decessi per età non abbia una forma semplice da modellare, se opportunamente divisa per la radice l_0 della tavola di mortalità ha il vantaggio di poter essere vista come una funzione di densità di probabilità:

$$\sum_{x=0}^{\Omega} \frac{d_x}{l_0} = 1 \tag{1.5}$$

Sfruttando questo risultato, è stato possibile ricorrere ad un modello mistura per stimarla ed inoltre ottenere la funzione di verosimiglianza in forma esplicita. Poiché gli individui muoiono una sola volta, gli intervalli d'età sono disgiunti e mutuamente esclusivi e ricoprono l'intera durata della vita. Per questi motivi è stata dunque adottata la distribuzione multinomiale; la funzione di verosimiglianza che ne deriva è:

$$L(\theta; d_x) = \prod_{x=0}^{\Omega} p(x;\theta)^{d_x} \tag{1.6}$$

dove d_x indica i decessi della tavola di mortalità all'età x, Ω è l'ultima età di morte considerata e $p(x, \theta)$ è la probabilità di morire nell'intervallo d'età $[x, x + 1)$, che può essere calcolata come l'integrale del modello mistura tra i due estremi, ovvero

$$p(x;\theta) = \int_x^{x+1} f(t;\theta)\, dt \tag{1.7}$$

Pertanto, le stime di massima verosimiglianza dei parametri in θ del modello (1.4) si ottengono massimizzando l'equazione (1.6).

1.4 Risultati principali

Zanotto (2016) ha stimato il modello (1.4) per descrivere la distribuzione dei decessi della popolazione maschile di Svezia, Francia, Germania dell'Est, Repubblica Ceca e di molti altri Paesi per diversi anni a partire dal secolo scorso per i quali erano disponibili i dati. Nel complesso, i risultati riguardanti la mortalità infantile ed adulta hanno confermato quanto già noto da studi precedenti: è emersa una decisiva riduzione della mortalità infantile e una compressione e spostamento verso età più avanzate di quella adulta. In contemporanea a quest'ultimo fatto, anche la distribuzione della mortalità accidentale e prematura si è spostata verso destra, pertanto è stato dimostrato che l'evoluzione della mortalità prematura dipende fortemente dalle trasformazioni di quella adulta, a conferma dello stretto legame esistente tra queste due componenti. Inoltre, per la maggior parte dei Paesi studiati la mortalità accidentale con

piccola "gobba" attorno all'età di 20 anni è praticamente scomparsa, mentre quella prematura in alcuni casi sembra essere addirittura in aumento, a causa di una crescita del numero di decessi registrati al di fuori della prima parte della curva della mortalità adulta. Nella Figura 1.4 si può vedere un esempio di questo fenomeno per la Francia nel 2010.

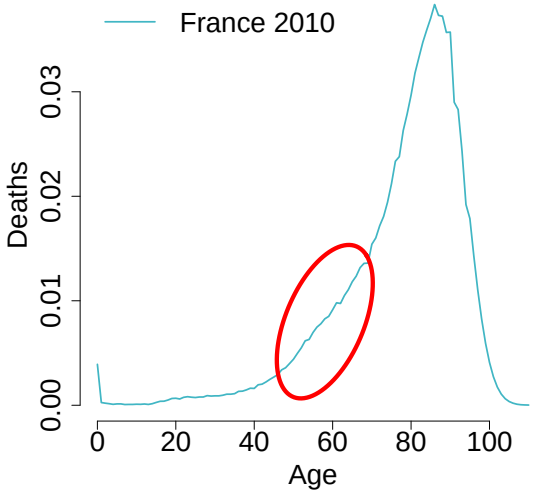

Figura 1.4: Distribuzione dei decessi per età in Francia nel 2010. In rosso sono evidenziati gli eventi in aumento legati alla mortalità prematura (*Fonte*: Zanotto, 2016).

Per riuscire a considerare quest'ultima tipologia di decessi, si è derivato un nuovo concetto di mortalità prematura, non più collocata attorno ai 40 anni come in passato, ma spostata in avanti verso i 50-65 anni.

L'aumento della mortalità prematura è risultato evidente soprattutto in Francia. Con lo scopo di quantificare il contributo della componente f_m (1.2), nella Figura 1.5 è mostrato graficamente l'andamento della percentuale di decessi associati alla mortalità accidentale e prematura in seguito alla stima del modello per i dati disponibili per la Francia dal 1900 al 2013. Esclusi i due picchi visibili in corrispondenza delle guerre mondiali, si osserva un generale abbassamento di questo tipo di mortalità dal 1900 fino al 1990; da quest'ultimo anno in poi, la serie inverte la sua traiettoria e comincia ad accelerare fino a

raggiungere negli anni 2000 valori addirittura più elevati di quelli registrati all'inizio del periodo di studio (Zanotto, 2016).

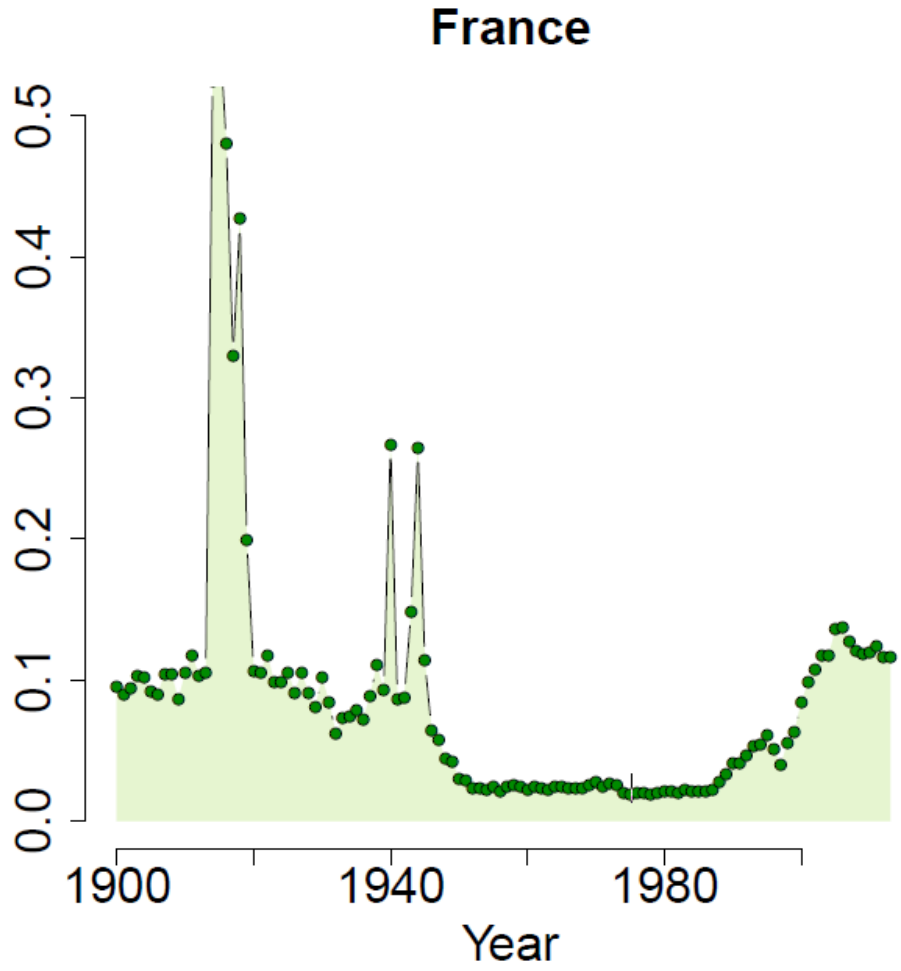

Figura 1.5: Percentuale di decessi legati alla mortalità accidentale e prematura (f_m) (*Fonte*: Zanotto, 2016).

Una domanda sorge allora spontanea: poiché in media la speranza di vita si sta allungando, perché a partire dal 1990 in Francia si osserva un aumento del numero di decessi connessi alla mortalità accidentale e prematura?

In altri studi precedenti, Pearson valutò la possibilità che la mortalità prematura fosse associata a specifiche malattie, ma alla fine non trovò una causa specifica che permettesse di giustificare i decessi che avvengono a circa metà del ciclo di vita. Il modello di Zanotto (2016), nonostante un tentativo

non andato a buon fine alla ricerca di una possibile relazione tra le cause esterne di morte (incidenti stradali, omicidi, suicidi e avvelenamenti) e la funzione che descrive la mortalità prematura, non è riuscito a dare una risposta alla domanda precedente. Questo rimane dunque un problema aperto da quel lavoro e rappresenta il punto di partenza di questo. Il contributo di questa tesi sarà provare ad esplorare varie cause di morte, alla ricerca di una possibile spiegazione dell'aumento della mortalità prematura in Francia negli ultimi anni.

Capitolo 2

Un'altra fonte: le cause di morte

L'evento *morte* non colpisce tutti in ugual misura e può essere anche molto diverso per genere ed età; ad esempio, gli uomini sono più soggetti delle donne a tutte le fasce d'età, la mortalità di solito è più evidente nelle età anziane e nella primissima infanzia, alcune cause di morte hanno maggiore rilevanza rispetto ad altre, e così via.

Nei secoli scorsi, in Italia, la maggior parte dei decessi era causata da malattie di origine infettiva; negli ultimi anni, invece, come riporta l'Istat, le principali cause di morte sono le malattie di natura degenerativa, come le malattie del cuore, le malattie cerebrovascolari e i tumori maligni (Istat, 2014; Istat, 2017).

Il profilo di mortalità per causa di una popolazione varia fortemente in base all'età oltre che al sesso. Si deduce che l'analisi delle cause di morte può essere fondamentale per studiare aspetti specifici e cambiamenti della mortalità, e questa è proprio la strada che abbiamo deciso di seguire per provare a spiegare l'aumento della mortalità prematura in Francia.

2.1 The Human Cause-of-Death Database

Abbiamo a disposizione una nuova fonte di informazione: *The Human Cause-of-Death Database* (HCD).[1] Si tratta di un progetto congiunto dell'Istituto Francese per gli Studi Demografici (INED) di Parigi e dell'Istituto Max Planck per la Ricerca Demografica (MPIDR) di Rostock che fornisce un accesso gratuito e relativamente facile da usare a serie storiche di mortalità specifica per cause di morte, caratterizzate da dati dettagliati e di alta qualità. A differenza di altre banche dati internazionali esistenti sulle cause di morte, nelle quali le serie sono spesso gravemente danneggiate da cambiamenti periodici nella classificazione delle malattie e riflettono discontinuità, lo HCD contiene serie temporali continue con cause di morte classificate in base a una lista costante, permettendo di descrivere l'andamento della mortalità per causa e facilitare le analisi comparative nel tempo e tra Paesi attraverso una metodologia universale e standardizzata.

Attualmente lo HCD contiene serie di dati per 16 Paesi, tra i quali è fortunatamente inclusa anche la Francia. Inoltre, i dati disponibili per la Francia sono relativi agli anni 2000-2013, proprio l'ultima parte del periodo per il quale desideriamo studiare la mortalità prematura.

La classificazione delle cause di morte è in continua evoluzione, in funzione della diagnostica medica. A fini di comparabilità, i dati relativi alla mortalità nello HCD sono classificati in base a tre tipi di liste di cause di morte: breve, intermedia e intera. L'ultima, la più lunga e dettagliata, è specifica per ogni Paese, mentre le prime due sono le stesse per tutti gli Stati e comprendono, rispettivamente, 16 e 104 gruppi di cause. Per le nostre analisi è stato deciso di usare la lista breve, in quanto si ritiene che 16 cause siano sufficienti, mentre 104 risulterebbero forse troppe per i nostri scopi. Nella Tabella 2.1 sono descritte le

[1] *Human Cause-of-Death Database* (2017). French Institute for Demographic Studies (France) and Max Planck Institute for Demographic Research (Germany). URL: `http://www.causeofdeath.org`

16 categorie di cause di morte della lista breve, a cui faremo riferimento da qui in avanti in questa tesi.

Tra tutti i dati disponibili nello HCD per la Francia dal 2000 al 2013, quelli selezionati per le nostre analisi sono i conteggi dei decessi della popolazione classificati nei 16 gruppi di cause di morte e in classi di età, distintamente per uomini e donne. I gruppi di età seguono il formato tipico e più convenzionale delle tavole di mortalità; in particolare, il primo intervallo d'età ha ampiezza un anno (età 0), il secondo 4 anni (1-4), e tutti quelli successivi da 5-9 a 90-94 hanno ampiezza 5 anni, con ultimo intervallo aperto 95+.

Il nostro primo obiettivo è quello di stimare il modello presentato nel Capitolo 1 marginalmente per i dati di ciascuna causa di morte. Poiché tale modello prevede l'utilizzo dei decessi di una tavola di mortalità, è dunque necessario ripartirli tra le diverse cause. Tutto questo è possibile quando esiste per il periodo oggetto di studio una tavola di mortalità generale della popolazione dalla quale ricavare una tavola *a decremento multiplo*.

Le tavole a decremento multiplo sono estensioni di tavole di mortalità standard (a decremento singolo) in cui coesistono decrementi simultanei dovuti a varie cause. Nei prossimi paragrafi verrà descritto il procedimento seguito per costruirle.

2.2 La tavola di mortalità per processi a decremento singolo

La tavola di mortalità è uno degli strumenti più importanti e completi utilizzati in demografia per l'analisi statistica della mortalità.

Questo metodo si presta perfettamente ai confronti tra popolazioni diverse e nel tempo; infatti, in questo modo i risultati non sono influenzati dalla numerosità degli individui né dalla struttura per età della popolazione. La differenza tra i vari gruppi umani considerati risiede unicamente nella velocità di

Tabella 2.1: Cause di morte classificate in base alla lista breve (HCD).

N. Causa	Descrizione
0	Tutte le cause
1	Alcune malattie infettive
2	Tumori
3	Malattie del sangue e degli organi responsabili della produzione del sangue
4	Malattie endocrine, nutrizionali e metaboliche
5	Disturbi mentali e comportamentali
6	Malattie del sistema nervoso e degli organi di senso
7	Malattie cardiache
8	Malattie cerebrovascolari
9	Altri e non specificati disturbi del sistema circolatorio
10	Malattie respiratorie acute
11	Altre malattie respiratorie
12	Malattie del sistema digestivo
13	Malattie della pelle e del tessuto sottocutaneo, del sistema muscolo-scheletrico e del tessuto connettivo
14	Malattie del sistema genitourinario e complicazioni della gravidanza, del parto e del puerperio
15	Alcune condizioni originarie del periodo perinatale e malformazioni/anomalie congenite
16	Cause esterne

eliminazione dei suoi componenti, la quale è funzione della forza della mortalità alle varie età. Poiché tutti muoiono, l'*intensità* del fenomeno è sempre pari a 1, mentre a variare è solo la *cadenza*, ovvero la distribuzione degli eventi secondo l'età (Livi Bacci, 1983).

È possibile distinguere due tipi di tavole di mortalità: quelle *per coorte o generazioni*, costruite considerando una generazione di nati e seguendola fino alla sua completa estinzione, e quelle *per periodo o contemporanei*, ottenute a partire da un gruppo di individui di diverse età che convivono in un dato intervallo di tempo. Il calcolo delle tavole per coorte non è molto agevole, dal momento che bisognerebbe seguire ogni componente fino al decesso, pertanto solitamente si preferisce utilizzare le tavole di mortalità per contemporanei. In particolare, una tavola di mortalità per contemporanei descrive il processo di eliminazione per morte di una popolazione fittizia di individui, appartenenti a diverse generazioni e che coesistono nello stesso intervallo temporale, se sottoposti per l'intera la durata della loro vita alle condizioni di mortalità di quel determinato periodo (Preston et al., 2001). Il punto di partenza per costruirla è dunque l'insieme dei tassi di mortalità specifici per età osservati in quel periodo (M_x), che vengono convertiti nelle probabilità di morte (q_x), sufficienti per ricavare tutte le altre funzioni base della tavola.

Il contingente iniziale sottoposto a eliminazione si chiama *radice* della tavola, l_0, un'arbitraria potenza di 10 di solito pari a 100 000. Una tavola di mortalità classica è costituita da diverse colonne, di cui una è sempre l'età (x) e le altre contengono delle funzioni relative allo studio della mortalità, chiamate anche *funzioni biometriche*. Queste colonne sono strettamente connesse tra loro; ad esempio, le tre funzioni base principali sono:

$$
\begin{aligned}
{}_nq_x &= \text{probabilità di morire tra le età } x \text{ e } x+n \\
l_x &= \text{numero di sopravviventi all'età } x \\
{}_nd_x &= \text{numero di decessi tra le età } x \text{ e } x+n
\end{aligned}
$$

e la relazione che le lega è

$$ {}_nd_x = l_x - l_{x+n} = {}_nq_x \cdot l_x \tag{2.1} $$

Alcune funzioni, come l_x, si riferiscono a un singolo preciso anno d'età, mentre altre, quali ${}_nd_x$ e ${}_nq_x$, fanno riferimento ad intervalli d'età che iniziano all'età esatta x e si estendono per n anni.

Per ulteriori dettagli sulla costruzione e struttura di una tavola di mortalità si veda Preston et al. (2001).

2.3 Tavole di mortalità a decremento multiplo

In demografia esistono processi in cui gli individui hanno più possibili modalità di uscire da un certo stato. Negli anni '50 e '60 sono state introdotte le cosiddette tavole *a decremento multiplo*, le quali risultano essere di grande utilità e applicabilità, ad esempio, nello studio della mortalità per causa, permettendo agli individui di uscire definitivamente dalla popolazione ma con diversi "tipi" di morte.

Una tavola a decremento multiplo viene ricavata a partire da una tavola a decremento singolo come quella descritta nel paragrafo precedente, dalla quale si prendono le serie ${}_nq_x$ e l_x e si utilizzano per ottenere le probabilità di morire e il numero di decessi specifici per ciascuna causa. Inoltre, tutte le colonne che abitualmente costituiscono una tavola standard si trovano anche in una tavola a decremento multiplo, dove fanno riferimento a "tutte le cause di morte combinate".

Si definiscono due funzioni, corrispondenti a quelle riportate per la tavola classica, che appartengono esclusivamente a particolari cause di morte, qui indicate con i, cioè a specifiche modalità di uscita dalla tavola:

${}_nd_x^i$ = numero di decessi per la causa i nell'intervallo d'età da x a $x+n$

${}_nq_x^i$ = probabilità di lasciare la tavola per la causa i tra le età x e $x+n$

per un individuo che ha raggiunto l'età x

e per le quali vale una relazione analoga alla (2.1)

$$ {}_nq_x^i = {}_nd_x^i / l_x \tag{2.2} $$

Le serie dei d_x^i per tutte le 16 cause di morte da noi considerate rappresentano i dati che serviranno per stimare il modello mistura nel prossimo capitolo, vediamo allora come ricavarle.

Come già spiegato in precedenza, i dati estratti dallo HCD, che indicheremo come D_x^i, sono i conteggi dei decessi avvenuti in Francia nella popolazione reale suddivisi in classi di età e in 16 gruppi di cause. Per costruire le tavole di mortalità a decremento multiplo per periodo sono state seguite le procedure standard descritte da Preston et al. (2001), di cui si riportano solo i passi fondamentali per ottenere i dati a noi necessari:

1. Costruire una tavola di mortalità per tutte le cause di morte combinate.

2. Calcolare la probabilità di morte per la causa i nell'intervallo d'età da x a $x+n$ come:

   $$ {}_nq_x^i = {}_nq_x \cdot \frac{{}_nD_x^i}{{}_nD_x} \tag{2.3} $$

   dove ${}_nD_x^i$ è il numero osservato di decessi per la causa i tra le età x e $x+n$ nella popolazione e ${}_nD_x$ è il numero totale dei decessi per tutte le cause combinate osservato nell'intervallo. In questo modo, le probabilità di morte ${}_nq_x$ vengono ripartite tra le varie cause proporzionalmente alle morti avvenute nella popolazione reale. Dalla serie dei ${}_nq_x^i$ è agevole risalire alle altre funzioni della tavola.

3. Calcolare il numero di decessi per la causa i nell'intervallo d'età da x a $x+n$ come:

   $$ {}_nd_x^i = {}_nq_x^i \cdot l_x \tag{2.4} $$

   dove l_x qui si riferisce al numero di individui che raggiungono l'età x, ovvero che sono sopravvissuti a *tutte* le cause di morte prima dell'età x.

Dal momento che, come già detto, le colonne della tavola a decremento singolo di partenza si riferiscono a tutte le cause di morte combinate, per ogni intervallo d'età i decessi della tavola a decremento multiplo sommati rispetto a tutte le cause i devono essere uguali al numero totale di persone che escono dalla popolazione:

$$\sum_i {}_nd_x^i = {}_nd_x \tag{2.5}$$

Ragionamento analogo vale anche per le probabilità di morte e per tutte le altre funzioni della tavola:

$$\sum_i {}_nq_x^i = \sum_i \frac{{}_nd_x^i}{l_x} = \frac{{}_nd_x}{l_x} = {}_nq_x \tag{2.6}$$

Inoltre, dalla combinazione delle relazioni (2.1), (2.3) e (2.4) si deduce che per una generica causa i a una qualsiasi età x vale la seguente formula:

$${}_nd_x^i = {}_nd_x \cdot \frac{{}_nD_x^i}{{}_nD_x} \tag{2.7}$$

che equivale a dire che i morti d_x della tavola di mortalità della popolazione fittizia si distribuiscono tra le varie cause proporzionalmente ai morti D_x della popolazione reale.

Per ottenere i dati a noi necessari, il passo 1 è stato evitato recuperando le tavole di mortalità già esistenti per la Francia per ogni anno dal 2000 al 2013, elaborate e pubblicate in *The Human Mortality Database* (HMD).[2] Abbiamo utilizzato delle tavole di mortalità abbreviate con classi d'età quinquennali, coerenti con il formato dei dati estratti per le cause di morte. L'unica differenza è che le tavole dello HMD hanno come ultima classe aperta 110+ e quindi, poiché l'informazione disponibile nello HCD arriva invece solo fino a 95+, sono state opportunamente modificate così da avere anche in queste l'ultimo intervallo d'età pari a 95+, in modo da operare con serie della stessa lunghezza e rendere possibili i passi 2 e 3.

[2] *Human Mortality Database* (2017). University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany). URL: `http://www.mortality.org`

Alla fine di questa procedura, applicata per tutti gli anni e distintamente per genere, sono state ottenute delle nuove serie di dati d_x^i, contenenti ciascuna i decessi della tavola di mortalità per la causa i al variare delle età, con $i = 1, \ldots, 16$. Prima di passare alla fase di stima dei modelli, cerchiamo di ricavare qualche informazione preliminare da tali dati.

2.4 La mortalità per causa: alcune analisi descrittive

Fin dalle prime analisi esplorative dei dati della popolazione reale estratti dallo HCD è emerso che i decessi che avvengono per la causa 15, ovvero per alcune condizioni originarie del periodo perinatale e malformazioni/anomalie congenite, si concentrano soprattutto all'età 0, per poi diminuire drasticamente in quelle successive; inoltre, tale causa è responsabile di quasi il 90% delle morti complessive che accadono in quel primo intervallo d'età. Questo fa sì che per la causa 15 la forma della distribuzione dei decessi al variare dell'età sia completamente diversa da quella di tutte le altre cause. A dimostrazione di ciò, nella Figura 2.1 si riportano dei diagrammi a barre costituiti dai dati grezzi estratti dallo HCD e che rappresentano per le varie classi d'età il numero di decessi avvenuti per questa causa nel 2013 (D_x^{15}), rispettivamente per gli uomini a sinistra e per le donne a destra.

Per questi motivi e dal momento che tali decessi fanno riferimento principalmente alla mortalità infantile, mentre invece il nostro interesse è studiare l'aumento della mortalità prematura in Francia, si è deciso di escludere dalle analisi tutti i decessi avvenuti per la causa 15 (operazione ritenuta di poca influenza anche perché rappresentano meno dell'1% dei decessi totali) e i decessi appartenenti alla prima classe di età (d_0^i) di tutte le cause rimanenti. Una conseguenza di ciò sarà eliminare la prima distribuzione che costituisce il modello mistura introdotto nel Capitolo 1 (la semi-normale) e quindi anche il primo

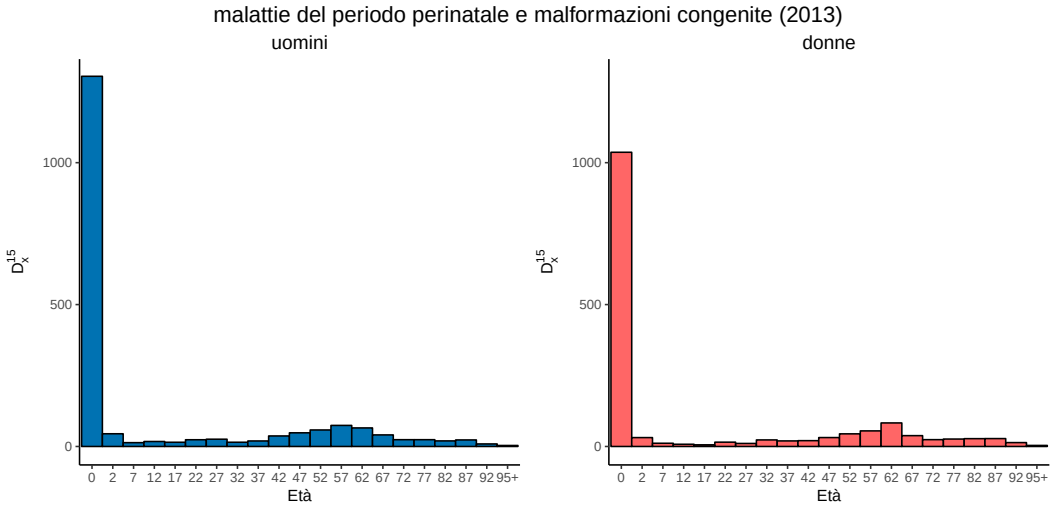

Figura 2.1: Distribuzione dei decessi maschili e femminili per età per la causa 15 nel 2013.

parametro di mistura η, come sarà descritto meglio nel prossimo capitolo. I dati considerati da qui in poi e che saranno usati in questa tesi per stimare i vari modelli sono, pertanto, i decessi delle tavole di mortalità a decremento multiplo avvenuti dall'età 1 a 95+ per tutte le cause dalla 1 alla 16, tranne la 15.

Si riportano ora alcune analisi descrittive sull'evoluzione della mortalità per causa nel periodo d'interesse.

Per prima cosa, è interessante capire qual è l'incidenza dei decessi per le varie cause di morte sul totale dei decessi. Nelle Tabelle 2.2 e 2.3 si possono osservare gli andamenti dal 2000 al 2013 delle proporzioni dei decessi maschili e femminili, rispettivamente, per ciascuna causa.

Risulta subito evidente che per entrambi i sessi e in tutti gli anni le prime due cause di morte responsabili della maggioranza dei decessi sono le cause 2 e 7, cioè i tumori e le malattie cardiache. Più precisamente, per gli uomini i tumori determinano circa il 33% dei decessi complessivi, con percentuale che aumenta nel periodo centrale e diminuisce leggermente negli ultimi anni, mentre le malattie cardiache riguardano quasi il 22% delle morti nel 2000, valore che si abbassa progressivamente fino al 2013 arrivando a poco più del 20%. La

Tabella 2.2: Evoluzione delle proporzioni dei decessi maschili in Francia per tutte le cause nel periodo 2000-2013.

Anno	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013
Causa	(%)	(%)	(%)	(%)	(%)	(%)	(%)	(%)	(%)	(%)	(%)	(%)	(%)	(%)
1	2.06	1.95	2.02	2.04	1.93	1.94	2.02	2.02	2.07	2.10	2.17	2.28	2.22	2.06
2	32.91	33.17	33.10	32.65	33.66	33.41	33.62	33.66	33.20	33.35	33.25	33.29	33.09	33.17
3	0.40	0.41	0.42	0.39	0.42	0.39	0.41	0.43	0.41	0.43	0.42	0.42	0.38	0.39
4	3.26	3.34	3.34	3.71	3.50	3.45	3.43	3.42	3.57	3.51	3.48	3.50	3.57	3.59
5	2.83	2.97	2.95	3.05	2.85	2.94	2.89	2.99	3.07	3.04	3.22	3.44	3.73	3.86
6	3.50	3.78	3.88	4.24	4.19	4.58	4.72	4.82	5.08	5.17	5.32	5.57	5.73	5.64
7	21.83	21.69	21.59	21.18	21.83	21.27	21.45	21.51	21.30	20.84	20.86	20.46	20.15	20.18
8	6.84	6.79	6.61	6.50	6.27	6.23	6.11	6.00	5.80	5.74	5.74	5.83	5.65	5.56
9	2.94	2.91	2.84	2.67	2.70	2.56	2.57	2.48	2.46	2.31	2.27	2.04	1.98	1.94
10	3.41	2.93	3.19	3.25	2.88	3.37	2.88	3.08	3.00	3.31	2.93	2.94	3.29	3.35
11	4.71	4.60	4.59	4.84	4.64	4.82	4.71	4.72	4.76	4.70	4.74	5.22	5.27	5.24
12	4.60	4.75	4.65	4.72	4.61	4.58	4.66	4.53	4.64	4.57	4.62	4.47	4.43	4.45
13	0.81	0.90	0.87	0.81	0.84	0.79	0.81	0.75	0.84	0.86	0.83	0.75	0.84	0.87
14	1.66	1.66	1.77	1.78	1.88	1.91	2.00	1.94	2.08	2.18	2.24	1.95	1.95	2.02
16	8.26	8.15	8.17	8.18	7.80	7.75	7.72	7.64	7.72	7.88	7.92	7.85	7.74	7.68
Tutte	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00

Tabella 2.3: Evoluzione delle proporzioni dei decessi femminili in Francia per tutte le cause nel periodo 2000-2013.

Anno	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013
Causa	(%)	(%)	(%)	(%)	(%)	(%)	(%)	(%)	(%)	(%)	(%)	(%)	(%)	(%)
1	2.13	2.06	2.06	2.22	2.02	2.04	2.03	2.03	2.22	2.29	2.26	2.46	2.41	2.29
2	22.32	22.51	22.60	22.18	23.09	23.02	23.30	23.08	23.33	23.43	23.59	23.78	23.64	23.77
3	0.55	0.54	0.51	0.54	0.52	0.52	0.56	0.52	0.53	0.54	0.56	0.52	0.48	0.49
4	4.81	4.93	5.00	5.66	4.91	4.81	4.83	4.76	4.74	4.82	4.80	4.46	4.51	4.59
5	4.68	4.79	4.86	4.92	4.65	4.66	4.70	4.70	4.67	4.74	4.76	5.35	5.70	5.99
6	4.63	4.97	5.36	5.96	6.03	6.60	7.05	7.51	7.91	8.11	8.40	8.75	9.12	8.98
7	25.97	26.05	25.72	25.05	25.89	25.39	25.65	25.67	25.19	24.71	24.56	24.13	23.25	23.30
8	9.85	9.75	9.29	9.04	9.13	8.73	8.64	8.30	8.28	8.14	8.12	8.07	7.93	7.83
9	2.70	2.75	2.56	2.41	2.41	2.33	2.28	2.30	2.22	2.10	2.11	1.78	1.68	1.71
10	4.43	3.68	3.99	3.96	3.58	4.34	3.50	3.65	3.62	3.82	3.24	3.45	4.08	3.83
11	3.39	3.19	3.26	3.42	3.43	3.58	3.37	3.49	3.46	3.54	3.50	3.96	4.14	4.05
12	4.53	4.59	4.58	4.42	4.54	4.38	4.48	4.40	4.39	4.36	4.37	4.16	4.06	4.11
13	1.88	1.81	1.84	1.77	1.72	1.62	1.67	1.61	1.56	1.51	1.46	1.34	1.36	1.39
14	1.59	1.77	1.83	1.84	1.89	1.96	1.98	2.07	2.07	2.07	2.26	1.88	1.92	1.91
16	6.55	6.60	6.55	6.62	6.19	6.03	5.96	5.93	5.81	5.83	6.02	5.90	5.73	5.75
Tutte	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00

situazione è un po' diversa per le donne: sono le malattie cardiache a causare il numero più elevato di decessi, pari al 26% del totale nel 2000 e con una riduzione al 23,3% nel 2013, mentre i tumori figurano sempre al secondo posto nella graduatoria delle principali cause di morte tranne che negli ultimi due anni, con una quota circa del 22% nei primi anni 2000 che poi aumenta fino al 23,77% a fine periodo, raggiungendo la prima posizione anche per le donne. Un'altra differenza rilevante che emerge è che per gli uomini la terza causa di morte sono sempre le cause esterne (causa 16), con percentuale che diminuisce nel tempo da 8,26% a 7,68%, invece per le donne questa causa rimane attorno al 6% e in posizione più bassa (quarta nel 2000 e sesta nel 2013). La terza causa di morte femminile nel 2000 sono le malattie cerebrovascolari (causa 8), rappresentative di quasi il 10% di tutti i decessi, che poi scendono a 7,83% nel 2013 diventando la quarta causa; dal 2010 in poi il terzo posto per le donne è infatti occupato dalle malattie del sistema nervoso e degli organi di senso (causa 6), che negli anni considerati assumono via via maggiore importanza e si affermano essere la causa con la più alta crescita percentuale, quasi raddoppiando da 4,63% a 8,98%. Le cause di morte 6 e 8 sono frequenti anche per gli uomini, seppur con quote di decessi più basse rispetto a quelle femminili e che si aggirano attorno al 3-5% e 5-6%, rispettivamente. Infine, la causa che occupa l'ultimo posto sia per le morti maschili che femminili è la numero 3, ovvero le malattie del sangue e degli organi emopoietici, con un'incidenza pari a 0,40% e 0,50% per i due sessi.

Già da queste prime osservazioni generali si intuisce che la mortalità è diversa tra uomini e donne. Tuttavia, finora non abbiamo tenuto in considerazione un fattore fondamentale: l'età dei decessi. Infatti, queste statistiche descrittive ci nascondono che cosa avviene proprio all'interno delle singole classi d'età. Poiché ci interessa focalizzare l'attenzione sulla mortalità prematura e distinguerla da quella adulta, abbiamo costruito delle analisi più complesse ed esaminato per ciascuna causa la forma della distribuzione dei decessi per età.

Nelle Figure 2.2, 2.3, 2.4 e 2.5 sono rappresentati dei diagrammi a barre

costituiti dalle frequenze dei decessi delle serie d_x^i di tutte e 15 le cause analizzate e relativi all'anno 2013, scelto a titolo di esempio; in ascissa ci sono gli intervalli d'età, aventi come etichetta il valore centrale, mentre l'altezza di ciascun rettangolo indica il numero di morti avvenute in quella classe d'età. Per facilitare i confronti e far percepire in modo visibile ed immediato le differenze di genere, sono state riportate nello stesso grafico le serie sia degli uomini che delle donne, colorate rispettivamente di blu e di rosa.

Si nota immediatamente che la forma della mortalità per le diverse cause varia molto tra uomini e donne. Innanzitutto, le distribuzioni dei decessi delle donne sono sempre più strette e spostate verso destra rispetto a quelle degli uomini, dimostrando che le donne vivono in media più a lungo. Un risultato importante che emerge da questi grafici è che, nel complesso, la mortalità prematura, compresa circa tra i 50 e 65 anni così come definita nel Capitolo 1, è più frequente e concentrata nella popolazione maschile.

Le differenze più significative tra uomini e donne per le morti premature riguardano le cause 2, 5, 9, 12 e 16. Negli anni centrali della vita, i tumori (causa 2), che sono una delle prime cause di morte in assoluto, colpiscono quasi il doppio degli uomini rispetto alle donne e sembrano essere la prima causa anche della mortalità prematura maschile. Anche i disturbi mentali e comportamentali (causa 5), gli altri disturbi del sistema circolatorio (causa 9) e le malattie del sistema digestivo (causa 12) registrano molti più decessi maschili tra i 50 e 70 anni, di gran lunga superiori rispetto a quelli femminili alle stesse età. Nel grafico relativo alle cause esterne di morte (causa 16), le quali comprendono incidenti stradali, omicidi, suicidi e avvelenamenti, è presente un elevato ammontare di decessi di uomini tra circa i 20 e i 65 anni che avvengono al di fuori della curva della mortalità adulta. Questo risultato non stupisce, dal momento che è noto che spesso gli uomini hanno comportamenti più a rischio rispetto alle donne. Inoltre, questa è l'unica causa che per entrambi i sessi registra anche delle cosiddette morti *accidentali* attorno ai 20-40 anni, che invece per tutte le altre cause sono praticamente assenti.

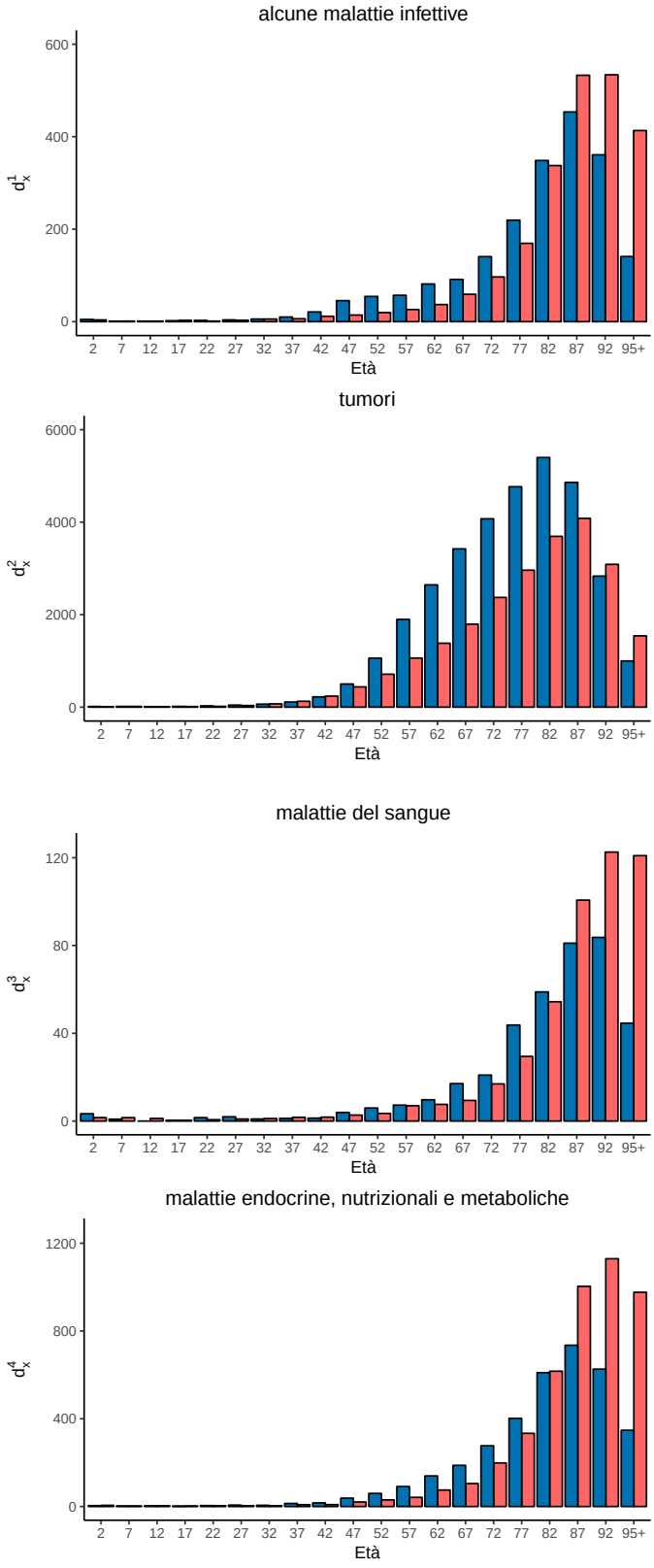

Figura 2.2: Distribuzione dei decessi maschili (blu) e femminili (rosa) per età per le cause 1, 2, 3 e 4 nel 2013.

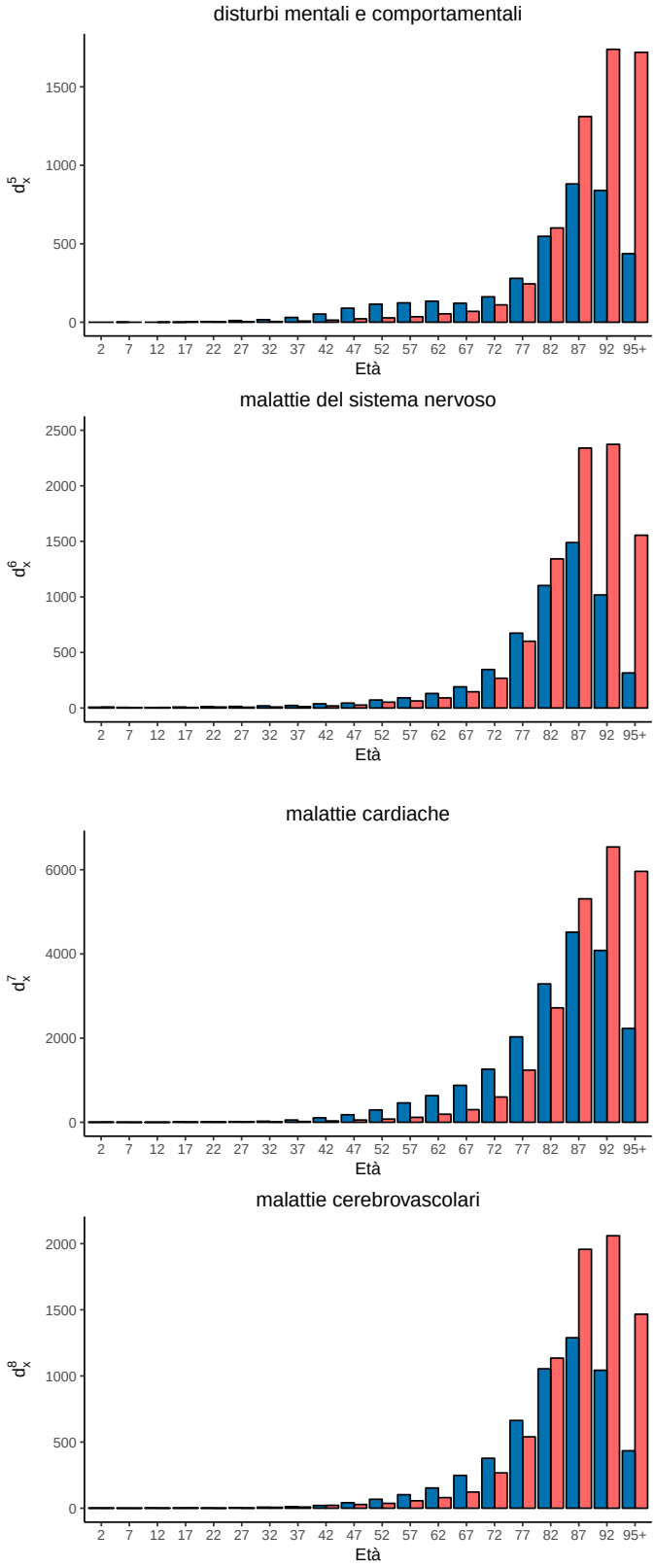

Figura 2.3: Distribuzione dei decessi maschili (blu) e femminili (rosa) per età per le cause 5, 6, 7 e 8 nel 2013.

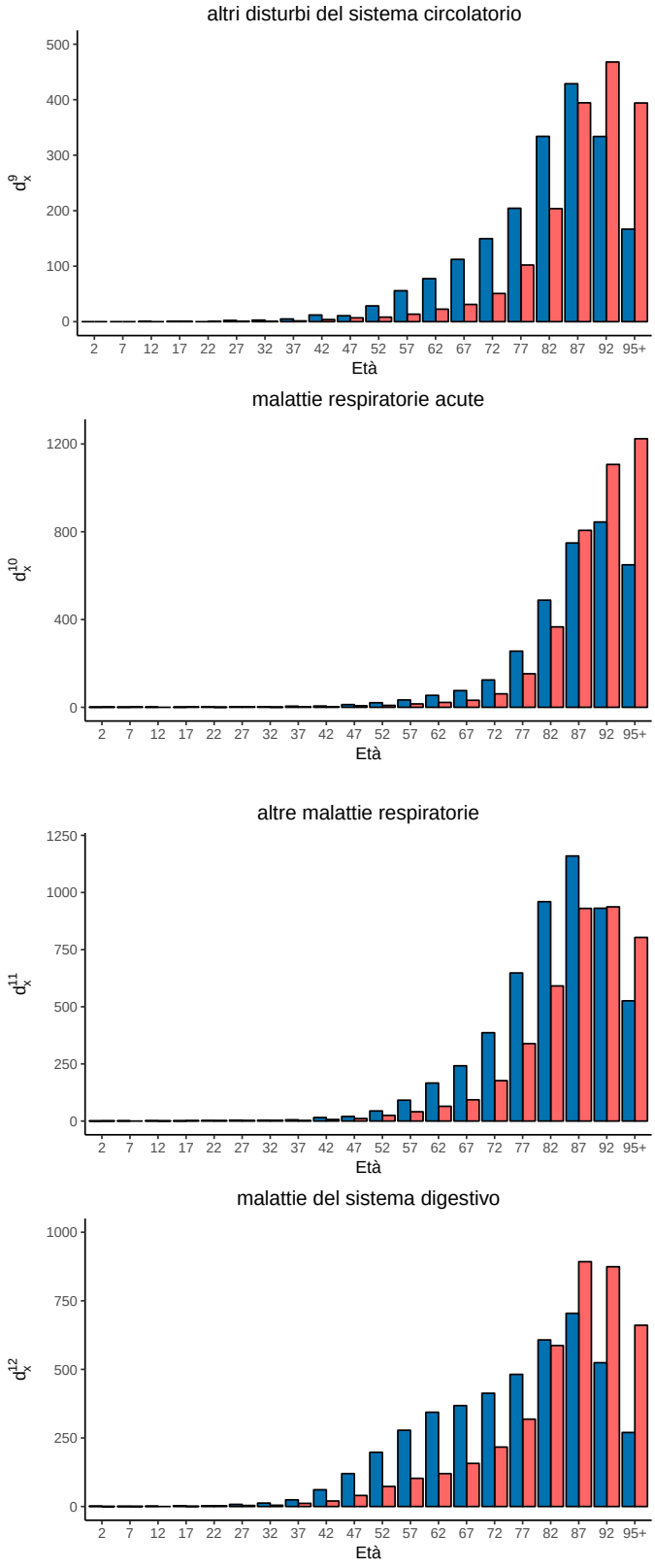

Figura 2.4: Distribuzione dei decessi maschili (blu) e femminili (rosa) per età per le cause 9, 10, 11 e 12 nel 2013.

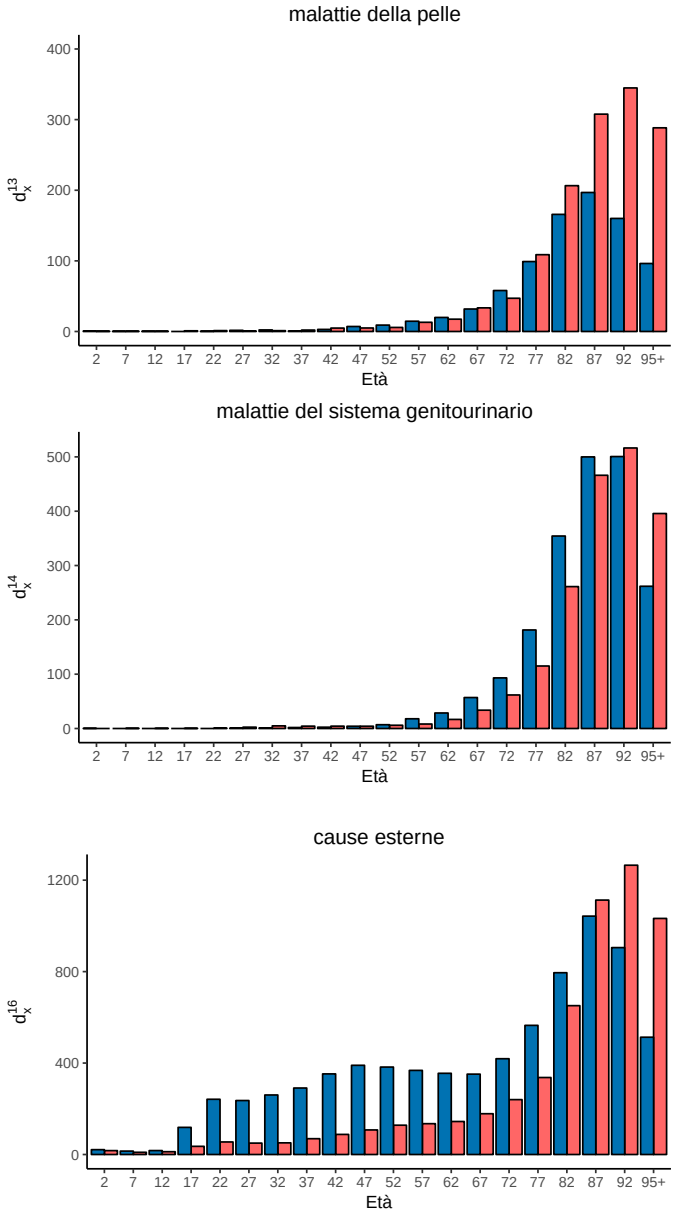

Figura 2.5: Distribuzione dei decessi maschili (blu) e femminili (rosa) per età per le cause 13, 14 e 16 nel 2013.

Le malattie cardiache (causa 7), che dalle analisi descrittive precedenti risultano essere la prima causa di morte femminile e la seconda per i maschi, sono responsabili principalmente dei decessi che avvengono alle età più avanzate (oltre i 70 anni) e non sembrano avere un ruolo di particolare rilievo nelle morti premature, almeno per quanto riguarda quelle femminili.

Un altro aspetto importante già accennato nel Capitolo 1 e che emerge in maniera chiara da questi grafici è che non necessariamente la mortalità prematura appare staccata da quella adulta. Nella distribuzione dei decessi maschili in alcuni casi, come ad esempio per i disturbi mentali e comportamentali (causa 5) e in maniera estrema per le cause esterne (causa 16), è ben visibile e accentuata la "gobba" della curva della mortalità prematura, centrata a circa 50 anni e separata dalle morti delle fasce d'età anziane; per altre cause, invece, di cui il caso più emblematico è quello dei tumori, non si riesce a distinguere un netto confine tra la curva della mortalità prematura e quella adulta, le quali sembrano piuttosto sovrapporsi e mischiarsi insieme. Quest'ultimo fatto giustifica ancora di più la nostra scelta di ricorrere ad un modello mistura di due distribuzioni per stimare questa parte dell'intera curva dei decessi.

Anche per le donne risulta esserci mortalità prematura nella fascia 50-70 anni, visibile soprattutto per le cause 2, 12 e 16, pur se in misura decisamente inferiore rispetto agli uomini. In tutti gli altri casi, i decessi prematuri femminili sono quasi trascurabili in confronto a quelli adulti.

Alla luce di tutte queste osservazioni, si deduce che la mortalità prematura è un fenomeno che colpisce principalmente gli uomini. Di conseguenza, per studiarne i cambiamenti e un eventuale aumento in Francia nel periodo 2000-2013, si è deciso di concentrare l'attenzione soprattutto sui maschi. Nei prossimi capitoli i modelli saranno, pertanto, stimati solo per la popolazione maschile. Infine, nell'ultimo capitolo di questa tesi si riportano brevemente alcune considerazioni e risultati ottenuti replicando tutte le analisi anche per le donne.

Capitolo 3

La stima del modello mistura per le singole cause di morte

Prima di passare alla stima del modello introdotto nel Capitolo 1 per i dati delle varie cause di morte, si presenta brevemente la distribuzione normale asimmetrica ed alcuni suoi aspetti peculiari, che saranno utili in seguito.

3.1 La distribuzione normale asimmetrica

3.1.1 Definizione

La distribuzione normale è una delle più utilizzate per descrivere l'andamento di fenomeni in diverse discipline. Tuttavia, esistono dei casi in cui la normale non risulta essere appropriata, soprattutto se il campione di dati oggetto di studio è caratterizzato da una più o meno marcata asimmetria. Nasce allora l'esigenza di una famiglia di distribuzioni più generale e flessibile, che riesca a comprendere e rappresentare anche queste situazioni.

La famiglia di distribuzioni normali asimmetriche, ideata ed elaborata per la prima volta da Azzalini (1985) nel caso scalare, si basa sul seguente utile risultato.

Lemma 1. *Sia f_0 una funzione di densità di probabilità unidimensionale simmetrica in 0, e sia G una funzione di ripartizione assolutamente continua tale che G' è simmetrica in 0. Allora*

$$f(y) = 2f_0(y)G(\lambda y) \qquad (-\infty < y < \infty) \tag{3.1}$$

è una funzione di densità per ogni λ reale.

Questo lemma permette di modificare una funzione di densità simmetrica f_0 attraverso una funzione di "perturbazione" $G(\lambda y)$ per costruire una nuova densità f che sia *asimmetrica*. In base alla scelta di f_0 e G si può ottenere un ampio insieme di distribuzioni "perturbate", il quale include sempre la densità di partenza f_0, che corrisponde al caso in cui $\lambda = 0$.

La distribuzione *normale asimmetrica* è il più semplice e tipico esempio di funzione che appartiene a questo insieme. Facendo riferimento al Lemma 1, Azzalini (1985) sceglie $f_0 = \phi$ e $G = \Phi$, ovvero la funzione di densità e la funzione di ripartizione della normale standard, rispettivamente, e definisce la funzione di densità

$$\phi(z;\lambda) = 2\phi(z)\Phi(\lambda z) \qquad (-\infty < z < \infty) \tag{3.2}$$

Una variabile casuale continua Z con funzione di densità (3.2) è chiamata *normale asimmetrica* con parametro di forma $\lambda \in \mathbb{R}$. Seguendo l'usuale notazione, si scrive che $Z \sim SN(\lambda)$, dove SN deriva dall'acronimo inglese di *Skew-Normal*. λ è chiamato parametro di *forma* perché è il parametro che regola, appunto, la "forma" della densità ed è legato all'asimmetria. Tuttavia, esso non coincide con il coefficiente di asimmetria della distribuzione, come sarà più chiaro in seguito. Per valori positivi di λ si ottiene una distribuzione asimmetrica a destra, per valori negativi di λ una distribuzione asimmetrica a sinistra; per valori opposti rispetto allo 0 si hanno forme della densità speculari. Nella Figura 3.1 si può vedere come si distribuisce Z per alcuni valori di λ positivi e negativi.

Analogamente alla distribuzione normale, è possibile generalizzare la densità (3.2) introducendo dei parametri di posizione e scala attraverso una trasformazione lineare.

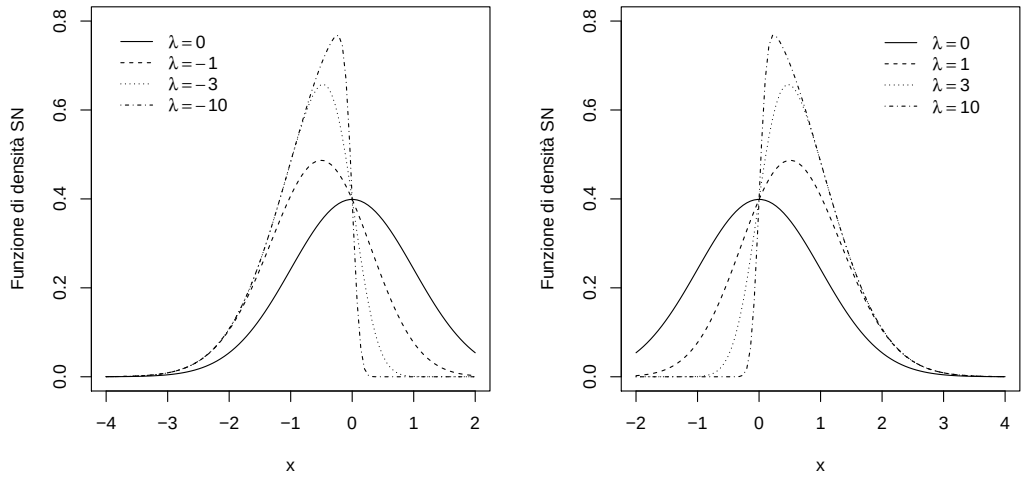

Figura 3.1: Funzione di densità di una $SN(\lambda)$ al variare di λ.

Data una variabile casuale $Z \sim SN(\lambda)$, $\xi \in \mathbb{R}$ e $\omega \in \mathbb{R}^+$, allora la variabile

$$Y = \xi + \omega Z \tag{3.3}$$

è chiamata normale asimmetrica con parametro di posizione ξ, parametro di scala ω e parametro di forma λ. La funzione di densità di Y è

$$f(y; \xi, \omega, \lambda) = \frac{2}{\omega}\, \phi\left(\frac{y-\xi}{\omega}\right) \Phi\left(\lambda\, \frac{y-\xi}{\omega}\right) \qquad (-\infty < y < \infty) \tag{3.4}$$

e si può scrivere che

$$Y \sim SN(\xi, \omega^2, \lambda)$$

dove il quadrato di ω è usato per analogia con la notazione $N(\mu, \sigma^2)$. Quando $\xi = 0$ e $\omega = 1$ si torna alla densità (3.2) e si dice che la distribuzione è "normalizzata". Scrivere $Z \sim SN(\lambda)$ equivale a $Z \sim SN(0, 1, \lambda)$.

La funzione di densità nella forma (3.4) corrisponde a quella già usata nel Capitolo 1 per definire le due distribuzioni normali asimmetriche che descrivono la mortalità accidentale e prematura e quella adulta nel modello mistura.

3.1.2 Proprietà e momenti

Dal punto di vista teorico, la classe di distribuzioni normali asimmetriche ha il vantaggio di essere matematicamente trattabile e di godere di buone proprietà. Se ne riportano alcune da Azzalini e Capitanio (2014).

Proposizione 1. *Sia Z una variabile casuale $SN(0,1,\lambda)$ con funzione di densità $\phi(x;\lambda)$. Allora valgono le seguenti proprietà:*

1. $\phi(x;0) = \phi(x), \quad \forall x;$
2. $\phi(0;\lambda) = \phi(0), \quad \forall \lambda;$
3. $-Z \sim SN(0,1,-\lambda)$, *equivalentemente* $\phi(-x;\lambda) = \phi(x;-\lambda) \quad \forall x;$
4. $\lim_{\lambda\to+\infty} \phi(x;\lambda) = 2\phi(x)I_{[0,\infty)}(x), \quad \forall x;$
5. $Z^2 \sim \chi^2_1, \quad \forall \lambda;$
6. *se* $Z' \sim SN(0,1,\lambda')$ *con* $\lambda' < \lambda$, *allora* $Z' <_{st} Z$.

Da queste proprietà risulta evidente la relazione che esiste tra la distribuzione normale e quella normale asimmetrica; più precisamente, la proprietà 1 implica che la classe delle distribuzioni normali costituisce un caso particolare della famiglia delle normali asimmetriche e corrisponde alla scelta di $\lambda = 0$. Inoltre, dalla proprietà 4 emerge che per λ tendente all'infinito si raggiunge come caso limite una distribuzione completamente asimmetrica, che converge alla cosiddetta funzione di densità semi-normale (o "mezza" normale).

Per il calcolo dei momenti di $Z \sim SN(\lambda)$, Azzalini (1985) utilizza il noto risultato seguente. Per la dimostrazione, si veda, ad esempio, Zacks, 1981, pp. 53-54.

Lemma 2 (Azzalini, 1985)**.** *Se U è una variabile casuale N(0,1), allora*

$$\mathbb{E}\{\Phi(hU+k)\} = \Phi\left(\frac{k}{\sqrt{1+h^2}}\right), \qquad \forall\, h,k \in \mathbb{R} \tag{3.5}$$

Sfruttando questo risultato, segue che la funzione generatrice dei momenti di Z è

$$M_Z(t) = 2\exp\left(\frac{t^2}{2}\right)\Phi(\delta t) \tag{3.6}$$

dove $\delta = \lambda/\sqrt{1+\lambda^2}$, e la funzione generatrice dei cumulanti è

$$K_Z(t) = \log\Big(M_Z(t)\Big) = \left(\frac{t^2}{2}\right) + \log\Big(2\Phi(\delta t)\Big) \tag{3.7}$$

Derivando la (3.7) fino al quarto ordine, con un po' di algebra si ottengono:

$$\mathbb{E}(Z) = b\delta \tag{3.8}$$

$$\mathbb{V}ar(Z) = 1 - (b\delta)^2 \tag{3.9}$$

$$\gamma_1(Z) = \frac{4-\pi}{2}\,\mathrm{sign}(\lambda)\left[\frac{\{\mathbb{E}(Z)\}^2}{\mathbb{V}ar(Z)}\right]^{3/2} \tag{3.10}$$

$$\gamma_2(Z) = 2(\pi-3)\left[\frac{\{\mathbb{E}(Z)\}^2}{\mathbb{V}ar(Z)}\right]^{2} \tag{3.11}$$

dove $b = \sqrt{\frac{2}{\pi}}$, e γ_1 e γ_2 indicano il terzo e quarto cumulante standardizzato, ovvero i comuni coefficienti di asimmetria e curtosi.

È importante sottolineare che questi due indici, per costruzione, possono assumere valori compresi in un intervallo numerico limitato:

$$(-\gamma_1^{max}, \gamma_1^{max}) \qquad (0, \gamma_2^{max}) \tag{3.12}$$

dove, rispettivamente, $\gamma_1^{max} \approx 0.995$ e $\gamma_2^{max} \approx 0.869$.
In particolare, questo vincolo sul campo di variazione di γ_1 indica che la distribuzione SN è appropriata per modellare solo asimmetrie basse o moderate; un limite che ne deriva è che dunque tale distribuzione non sempre risulta adeguata nel rappresentare campioni aventi asimmetria molto forte. Inoltre, valori di asimmetria con lo stesso segno, diversi ma elevati determinano una forma della curva sostanzialmente identica.

Per ricavare i momenti di $Y \sim SN(\xi, \omega^2, \lambda)$ si procede in modo analogo attraverso la corrispondente funzione generatrice dei momenti o, equivalentemente, tramite la funzione generatrice dei cumulanti. Per le espressioni risultanti si veda il prossimo paragrafo.

3.1.3 La parametrizzazione centrata

Come riportato in un articolo di Arellano-Valle e Azzalini (2008), i metodi basati sulla verosimiglianza presentano lati problematici quando utilizzati per fare inferenza sui parametri (ξ, ω^2, λ), specialmente in un intorno di $\lambda = 0$. I problemi riguardano soprattutto gli aspetti statistici, in quanto è stato dimostrato che la funzione di verosimiglianza per un generico campione di dati proveniente da una normale asimmetrica presenta un comportamento non regolare nei pressi di $\lambda = 0$, valore di particolare interesse dal momento che la distribuzione SN si riduce ad una normale. Più precisamente, esiste un punto di stazionarietà nella log-verosimiglianza profilo per λ in corrispondenza di $\lambda = 0$, pertanto la curva non manifesta più la tipica forma quadratica. In relazione a ciò, in $\lambda = 0$ la matrice di informazione attesa di Fisher del modello diventa singolare, nonostante i parametri siano ancora identificabili, violando le assunzioni standard che sono alla base delle proprietà asintotiche degli stimatori di massima verosimiglianza, le quali non sono quindi più garantite.

In sintesi, l'origine di tutti questi aspetti critici è stata attribuita alla parametrizzazione adottata, ritenuta "non particolarmente adatta per la stima". Proprio per questi motivi, Azzalini (1985) ha proposto una parametrizzazione alternativa per la normale asimmetrica $Y \sim SN(\xi, \omega^2, \lambda)$. Si parte dall'identità

$$Y = \xi + \omega Z = \mu + \sigma Z_0 \tag{3.13}$$

dove $Z = (Y - \xi)/\omega$ è una variabile casuale "normalizzata" con distribuzione $Z \sim SN(0, 1, \lambda)$, avente media e varianza definite dalle equazioni (3.8) e (3.9) del paragrafo precedente:

$$\mu_z = \mathbb{E}(Z) = b\delta \qquad \sigma_z^2 = \mathbb{V}ar(Z) = 1 - (b\delta)^2$$

con $b = \sqrt{2/\pi}$ e $\delta = \lambda/\sqrt{1+\lambda^2}$, e dove $Z_0 = \sigma_z^{-1}(Z - \mu_z)$ è la versione standardizzata di Z. La nuova parametrizzazione è costituita dai parametri

$(\mu, \sigma^2, \gamma_1)$, le cui espressioni esplicite, in funzione dei parametri originali, sono:

$$\mu = \mathbb{E}(Y) = \xi + \omega\mu_z \tag{3.14}$$

$$\sigma^2 = \mathbb{V}ar(Y) = \omega^2(1 - \mu_z^2) \tag{3.15}$$

$$\gamma_1 = \frac{\mathbb{E}\{(Y - \mathbb{E}(Y))^3\}}{\mathbb{V}ar(Y)^{3/2}} = \frac{4-\pi}{2}\frac{\mu_z^3}{(1-\mu_z^2)^{3/2}} \tag{3.16}$$

dove γ_1 è l'indice di asimmetria (3.10) già ottenuto nel precedente calcolo dei momenti di Z, ovvero $\gamma_1(Y) = \gamma_1(Z)$.

Quest'ultima parametrizzazione caratterizzata dai parametri $(\mu, \sigma^2, \gamma_1)$ in letteratura è chiamata *parametrizzazione centrata* (CP), poiché è ottenuta attraverso la variabile "centrata" Z_0, mentre i parametri originali (ξ, ω^2, λ) costituiscono la *parametrizzazione diretta* (DP), che è il punto di partenza della riparametrizzazione e quella impiegata fino ad ora per scrivere la funzione di densità della normale asimmetrica come definita nella (3.4).

Come descritto da Arellano-Valle e Azzalini (2008), ricorrere alla parametrizzazione centrata è estremamente vantaggioso perché permette di eliminare i problemi precedentemente menzionati che affliggono quella diretta. Con questa riparametrizzazione del modello, ad esempio, le forme delle verosimiglianze hanno andamenti più regolari e quadratici, senza punti di stazionarietà in $\gamma_1 = 0$. Nei pressi di tale valore la matrice di informazione attesa di Fisher è non singolare, pertanto tornano ad essere validi i risultati asintotici standard ed è quindi possibile applicare gli usuali metodi di inferenza basati sulla verosimiglianza. Inoltre, i vantaggi della CP sulla DP non riguardano solo il lato teorico, ma anche quello pratico; infatti, una forma più regolare della log-verosimiglianza porta ad una convergenza più rapida delle procedure di massimizzazione numerica usate per calcolare le stime di massima verosimiglianza (Azzalini e Capitanio, 1999).

Infine, un beneficio importante della CP è sicuramente legato ad una più semplice ed intuitiva interpretazione dei parametri: μ, σ^2 e γ_1 sono esattamente la media, la varianza e l'indice di asimmetria della distribuzione.

Per cercare di rendere meno problematica possibile l'inferenza sui parametri delle due normali asimmetriche che costituiscono il modello mistura che vogliamo stimare, nelle nostre analisi useremo questa più conveniente parametrizzazione. Tuttavia, come sarà ben presto chiaro dai prossimi paragrafi, questa soluzione non permetterà di evitare tutti i problemi, poiché esistono dei casi in cui le stime di massima verosimiglianza sono affette da difficoltà non rimovibili dal cambio di parametrizzazione.

3.2 L'adattamento del modello mistura ai nostri dati

Passiamo ora alla stima del modello mistura introdotto nel Capitolo 1 marginalmente per ciascuna causa di morte. A tal fine, è stato necessario apportare alcune modifiche al modello proposto da Zanotto (2016) in funzione dei dati che abbiamo a disposizione. Per i motivi spiegati all'inizio del paragrafo 2.4, si è deciso di escludere dalle analisi tutti i decessi avvenuti per la causa 15 e quelli appartenenti all'età 0 per tutte le altre cause, rimanendo pertanto con i conteggi delle morti dall'età 1 a 95+. Avendo in questo modo di fatto eliminato tutti i decessi associati alla primissima infanzia, la prima delle tre distribuzioni del modello mistura, ovvero la semi-normale f_I utilizzata per descrivere proprio la mortalità infantile nella parte iniziale della curva di mortalità, non è più necessaria in quanto i decessi rimanenti nelle prime fasi della vita sono quasi del tutto assenti, come si può vedere dai grafici riportati alla fine del Capitolo 2. Lo stesso ragionamento vale anche per il primo parametro di mistura η, che è stato rimosso mantenendo quindi solo il parametro α. Il modello risultante è una mistura di due componenti: le normali asimmetriche f_m e f_M che descrivono la mortalità accidentale e prematura e quella adulta. Un'altra differenza rispetto al modello di Zanotto (2016) è che, come anticipato alla fine del paragrafo precedente, per queste due funzioni di densità si è deciso

di adottare la parametrizzazione centrata al posto di quella diretta.

Il modello mistura che abbiamo utilizzato per approssimare la distribuzione dei decessi per età per le singole cause di morte è costituito da 7 parametri ed ha la seguente forma:

$$f(x;\theta) = \alpha f_m(x;\theta_m) + (1-\alpha) f_M(x;\theta_M) \tag{3.17}$$

dove in questo caso $\theta = (\alpha, \theta_m, \theta_M)$ con $\theta_m = (\mu_m, \sigma_m, \gamma_m)$ e $\theta_M = (\mu_M, \sigma_M, \gamma_M)$. Il parametro di asimmetria γ_1 da qui in poi sarà indicato solo con il simbolo γ senza il pedice "1" per evitare di appesantire la notazione con gli altri pedici m e M.

Le stime dei parametri in θ sono state ottenute sempre con il metodo della massima verosimiglianza seguendo il procedimento descritto nel paragrafo 1.3, con l'unica accortezza di calcolare $p(x;\theta)$ non più negli intervalli $[x, x+1)$, ma in funzione dell'ampiezza delle classi di età che abbiamo a disposizione. Pertanto, nella verosimiglianza multinomiale (1.6) i decessi d_x vengono sostituiti dalla serie dei d_x^i specifici della causa i che si sta analizzando, e $p(x;\theta)$ corrisponde alla probabilità di morire nell'intervallo d'età $[x, x+n)$, che diventa

$$p(x;\theta) = \int_x^{x+n} f(t;\theta)\, dt \tag{3.18}$$

Nella (3.18) x e n sono, rispettivamente, l'età esatta di inizio e l'ampiezza di ogni classe d'età, con $n = 4$ nel primo intervallo 1-4 anni, $n = 5$ in tutti gli intervalli quinquennali da 5-9 a 90-94 anni, e $n = 15$ nell'ultima classe 95-Ω, poiché l'ultima età Ω è stata posta pari a 110 anni. Procediamo allora alla stima del modello per i nostri dati.

3.3 Le stime di massima verosimiglianza

Il modello (3.17) è stato stimato separatamente per i dati delle 15 cause di morte analizzate per i decessi maschili negli anni 2000-2013. Nelle Tabelle 3.1 e 3.2 si riportano le stime di massima verosimiglianza dei 7 parametri in θ ottenute per il primo e l'ultimo anno del periodo considerato.

Tabella 3.1: Stime di massima verosimiglianza dei parametri del modello mistura per i decessi maschili nel 2000, specifiche per cause di morte.

		f_m			f_M		
Causa	α	μ_m	σ_m	γ_m	μ_M	σ_M	γ_M
1	0.140	41.393	11.167	-0.482	79.316	10.985	-0.773
2	0.004	1.000	16.344	-0.992	73.281	12.809	-0.667
3	0.142	1.000	42.810	-0.995	81.749	10.546	-0.800
4	0.019	27.993	21.453	-0.995	79.959	11.499	-0.808
5	0.434	68.112	22.472	0.662	85.677	6.294	-0.634
6	0.301	66.892	23.069	-0.948	81.256	7.558	-0.639
7	0.408	72.375	13.498	-0.626	84.740	7.632	-0.516
8	0.247	73.183	15.397	-0.738	83.066	7.997	-0.605
9	0.039	49.944	9.953	-0.748	80.205	10.160	-0.690
10	0.056	57.226	12.109	-0.994	86.129	7.805	-0.637
11	0.014	35.867	15.016	-0.995	80.867	10.023	-0.657
12	0.278	67.058	17.895	0.811	75.352	13.617	-0.914
13	0.092	63.653	20.014	-0.995	83.788	8.667	-0.693
14	0.369	80.883	14.424	-0.790	84.307	7.017	-0.703
16	0.655	53.438	27.644	0.850	81.674	8.655	-0.993

Tabella 3.2: Stime di massima verosimiglianza dei parametri del modello mistura per i decessi maschili nel 2013, specifiche per cause di morte.

		f_m			f_M		
Causa	α	μ_m	σ_m	γ_m	μ_M	σ_M	γ_M
1	0.590	75.707	16.725	-0.994	86.381	5.751	-0.431
2	0.167	74.898	15.118	0.900	76.127	12.746	-0.867
3	0.081	1.000	46.406	-0.995	84.487	9.953	-0.942
4	0.187	67.976	16.885	-0.995	84.588	9.724	-0.829
5	0.229	60.001	13.578	0.016	88.324	6.512	-0.449
6	0.187	65.380	18.779	-0.995	85.674	7.070	-0.621
7	0.409	74.727	13.209	-0.756	89.051	6.304	-0.353
8	0.479	79.446	13.907	-0.993	85.629	6.640	-0.484
9	0.849	80.939	12.661	-0.992	86.178	3.214	-0.027
10	0.047	55.229	12.186	-0.995	88.441	7.807	-0.668
11	0.113	70.823	14.594	-0.995	84.702	9.286	-0.710
12	0.568	67.286	13.268	-0.267	87.732	6.420	-0.162
13	0.357	79.431	16.036	-0.988	85.603	7.193	-0.474
14	0.039	60.060	8.937	-0.993	87.159	7.520	-0.654
16	0.706	70.969	36.380	0.847	85.433	6.205	-0.993

Per quanto riguarda la mortalità adulta, dalle tabelle si può notare che in 14 anni le medie μ_M della seconda normale asimmetrica per le varie cause sono aumentate nel tempo, passando da esser comprese tra gli 80 e 85 anni circa nel 2000 fino ad assumere valori tra gli 84 e 89 anni nel 2013, dimostrando quindi che l'età media dei decessi in età anziane si è spostata in avanti. In aggiunta ad una traslazione verso destra, sembra esser avvenuta anche una compressione di questa curva, poiché le deviazioni standard σ_M si sono ridotte per quasi tutte le cause. Le stime del parametro di asimmetria γ_M si mantengono pressoché stabili nel periodo osservato; in particolare, come ci aspettavamo, assumono valori sempre negativi per ciascuna causa di morte. Questo risultato conferma che è opportuno ricorrere ad una distribuzione asimmetrica a sinistra per descrivere la mortalità adulta, a supporto della teoria di Pearson (1897).

Nel complesso, le stime di massima verosimiglianza dei parametri della componente f_M sembrano essere coerenti con quanto affermato da Zanotto (2016) e non presentano differenze rilevanti tra le varie cause, specialmente nel 2013. Un'eccezione è rappresentata dai tumori (causa 2): sono l'unica causa di morte per cui la media della curva della mortalità adulta è sempre molto inferiore rispetto a quella di tutte le altre cause, fermandosi massimo tra i 73-76 anni. Questo è un risultato interessante, dal momento che dalle analisi esplorative i tumori sono sembrati essere la principale causa di mortalità maschile, sia prematura che adulta.

Tuttavia, la nostra attenzione si concentra soprattutto sull'analisi delle caratteristiche e dei cambiamenti della mortalità prematura per le singole cause. Da una prima osservazione delle stime dei parametri α nelle tabelle, associati all'intensità della mortalità accidentale e prematura, sembra che tale componente assuma particolare importanza per i decessi legati ai disturbi mentali e comportamentali (causa 5), alle malattie cardiache (causa 7), alle malattie del sistema digestivo (causa 12) e alle cause esterne (causa 16), sia nel 2000 che nel 2013, in linea con le nostre analisi esplorative. Da un'analisi più approfondita, ci si accorge, però, di alcuni risultati anomali, in contrasto

con quanto emerso in precedenza: i tumori (causa 2) presentano un α stimato pari a 0.004 e 0.167 nei due anni mostrati, quando invece, proprio perché erano sembrati essere la prima causa di morte prematura maschile, ci si aspettava un valore molto più elevato. Al contrario, compaiono, ad esempio, delle stime sospette anche per i parametri di mistura di cause come la 14 nel 2000 e della 1, 8, 9 e 13 nel 2013, di molto superiori rispetto alle previsioni. Inoltre, spiccano tra le altre alcune medie μ_m stimate uguali a 1.000, ovvero l'età minima dei decessi da noi considerata, e diverse stime di γ_m esattamente pari, o in certi casi molto vicine, a -0.995, cioè il più piccolo valore ammissibile per il coefficiente di asimmetria della distribuzione normale asimmetrica.

Tutte queste osservazioni preliminari ci hanno spinto ad indagare in maggior profondità, alla ricerca di possibili spiegazioni sull'origine di tali valori, come sarà descritto in dettaglio nel prossimo paragrafo.

3.3.1 Problemi nella stima di massima verosimiglianza

Si precisa fin da subito che la stima di massima verosimiglianza del modello si è rivelata essere, purtroppo, molto problematica. Le difficoltà maggiori sono associate soprattutto alla componente relativa alla mortalità prematura; esse riguardano specialmente le situazioni in cui la mortalità prematura è molto bassa, quasi assente, rendendo difficoltoso localizzare la posizione della prima normale asimmetrica ed identificarne i parametri, e i casi ben noti in letteratura e accennati nel Capitolo 1 in cui le curve della mortalità prematura e adulta si sovrappongono e non si riescono a distinguere in modo chiaro, sembrando quasi un'unica distribuzione.

Si ritiene che in parte ciò potrebbe essere dovuto anche alla tipologia e struttura dei dati a nostra disposizione, con decessi raggruppati in classi d'età di ampiezza quinquennale; conoscendo l'ammontare di morti per ogni singolo anno, invece, si potrebbero avere dei dati più precisi, che forse permetterebbero di cogliere degli aspetti specifici della distribuzione della mortalità complessiva in alcune età.

Il calcolo delle stime di massima verosimiglianza è stato effettuato tramite massimizzazione numerica della funzione di log-verosimiglianza. Innanzitutto, in generale sono stati riscontrati problemi di identificabilità del modello, legati all'esistenza di più possibili valori dei parametri, anche molto diversi, associati ad un valore della verosimiglianza approssimativamente simile; tuttavia, se sostituiti nella funzione di densità del modello, i vari set di parametri portano a curve con forme sostanzialmente equivalenti, senza differenze significative nell'adattamento ai dati. Questo fenomeno riguarda soprattutto i parametri della prima normale asimmetrica e di conseguenza il parametro di mistura, mentre le stime della seconda normale si mantengono praticamente identiche. Inoltre, la log-verosimiglianza del nostro modello con 7 parametri è risultata complessa da massimizzare; spesso l'algoritmo numerico ha fatto fatica a raggiungere il massimo globale e per valori diversi di inizializzazione dei parametri ha portato a stime differenti, convergendo a massimi locali. Per ovviare a ciò, il procedimento di stima è stato ripetuto più volte, partendo da vari punti iniziali distanti tra loro, e alla fine sono state scelte come stime di massima verosimiglianza dei parametri quelle che hanno permesso di ottenere il valore della verosimiglianza più alto.

Nonostante questi spiacevoli episodi, in alcuni casi il modello stimato ha portato a dei buoni risultati. A dimostrazione di ciò, nella Figura 3.2 si riporta graficamente l'adattamento del modello alla distribuzione dei decessi per due cause di morte dell'anno 2013: in alto i disturbi mentali e comportamentali (causa 5) e in basso le malattie del sistema digestivo (causa 12). Per entrambe le cause, nei grafici di sinistra è mostrata la distribuzione dei decessi per età, a cui è stata sovrapposta una curva rossa che rappresenta la funzione di densità del modello mistura (3.17), dove a θ sono state sostituite le stime di massima verosimiglianza dei parametri ottenute per quelle cause. Si precisa che per costruire tali grafici, poiché la curva rossa indica il valore della densità $f(x;\theta)$ per ogni singola età x, mentre noi conosciamo, invece, la *somma* dei conteggi dei decessi nei vari intervalli d'età, ad esempio quinquennali, è stato necessario

redistribuire i dati mediamente per ciascuna età. L'ammontare dei decessi ${}_nd_x^i$ di ogni classe d'età è stato diviso per l'ampiezza n della classe stessa, ed infine tutti i dati son stati standardizzati rispetto al totale, in modo da poter sovrapporre alla distribuzione così ottenuta la funzione di densità della mistura che, per definizione, integra ad 1 sul supporto in cui è definita. Questa procedura risulta essere poco verosimile per l'ultima classe aperta 95+, i cui decessi sono stati suddivisi equamente da 95 a 110 anni, mentre probabilmente questi si concentrano soprattutto nella prima parte dell'intervallo, ma in mancanza di informazioni più precise non è stato possibile fare altrimenti. Nei grafici di destra sempre della Figura 3.2, viene riportata di nuovo la curva rossa della funzione di densità del modello complessivo e vengono evidenziate in verde e in blu le curve corrispondenti alla mortalità prematura e adulta, rispettivamente, che unite al parametro α permettono di ottenerla, anche in questo caso sostituendo ai parametri le relative stime di massima verosimiglianza. Questa scomposizione si dimostra essere particolarmente utile per capire come le due componenti si combinano nella mistura ed eventualmente quale delle due è responsabile di problemi, come sarà più chiaro tra poco.

Come si può notare, per queste due cause il modello stimato sembra adattarsi in modo molto soddisfacente alla distribuzione dei decessi. Da tali grafici si potrebbe ipotizzare che il modello funzioni bene quando è possibile riconoscere la mortalità prematura parzialmente staccata da quella adulta, come in questi due esempi. Tuttavia, ciò non è necessariamente vero, ed il caso che lo dimostra è quello delle cause esterne. Dalle analisi esplorative è emerso che la causa 16 ha una pronunciata "gobba" di decessi prematuri; come si vede dai valori nelle Tabelle 3.1 e 3.2, la deviazione standard σ_m stimata per la prima normale asimmetrica di questa causa è enorme, pari a 27 nel 2000 e a 36 nel 2013, e l'α è superiore a 0.65 in entrambi gli anni. Andando ad esplorare i grafici della Figura 3.3 relativi alle cause esterne nel 2013, si vede che la prima componente della mistura ha varianza talmente ampia da arrivare a comprendere, in modo errato, anche i decessi che avvengono dopo i 95 anni. Questo fa sì che la

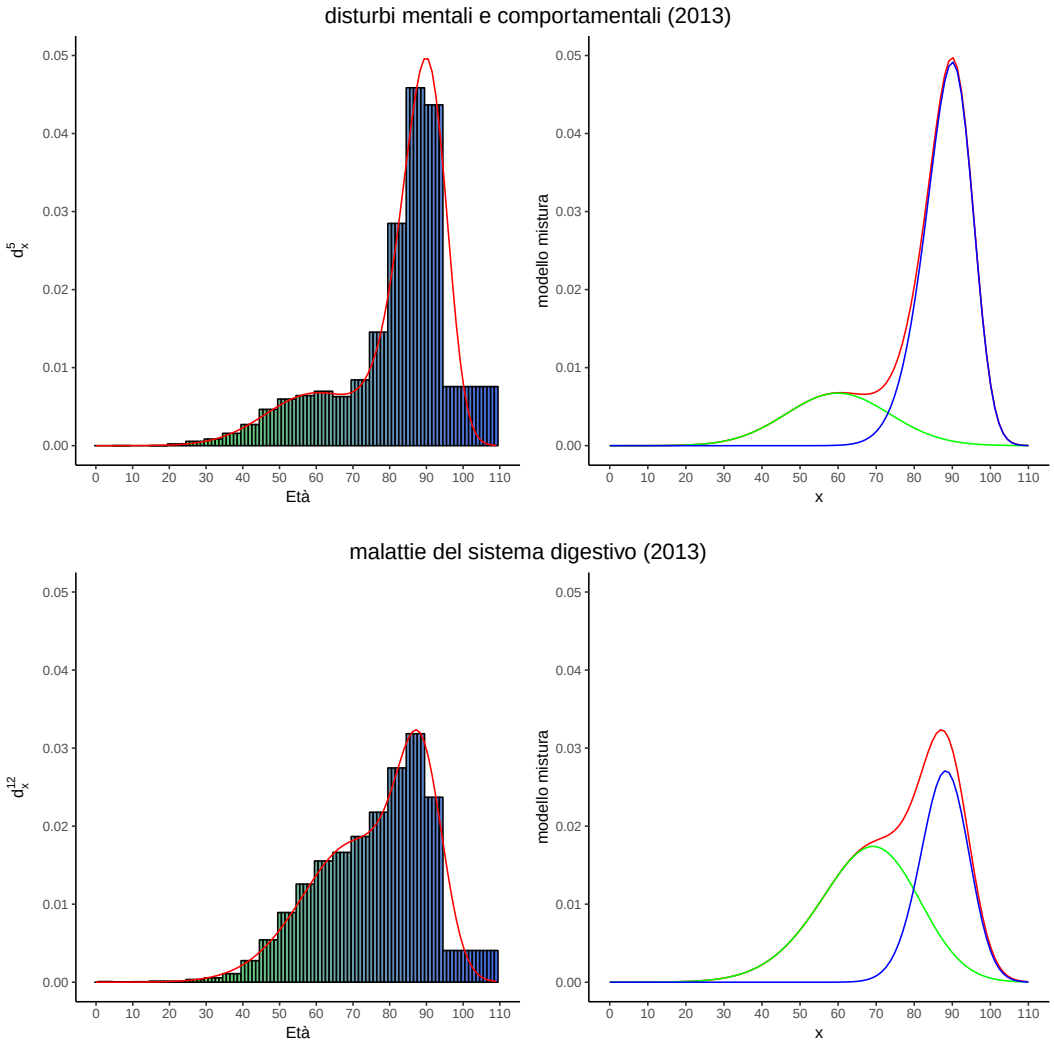

Figura 3.2: Funzione di densità del modello mistura stimato per diverse cause di morte, sovrapposta alla distribuzione dei decessi (a sinistra) e scomposta nelle sue componenti (a destra). Focus sulle cause 5 e 12 nel 2013.

stima della media della prima componente sia circa 71, decisamente troppo in avanti rispetto a quanto si osserva dal grafico a sinistra, e costringe la seconda componente ad avere un'asimmetria negativa che arriva quasi al limite inferiore (-0.993), la quale diventa quindi una "mezza" normale che non coglie correttamente la forma della mortalità adulta.

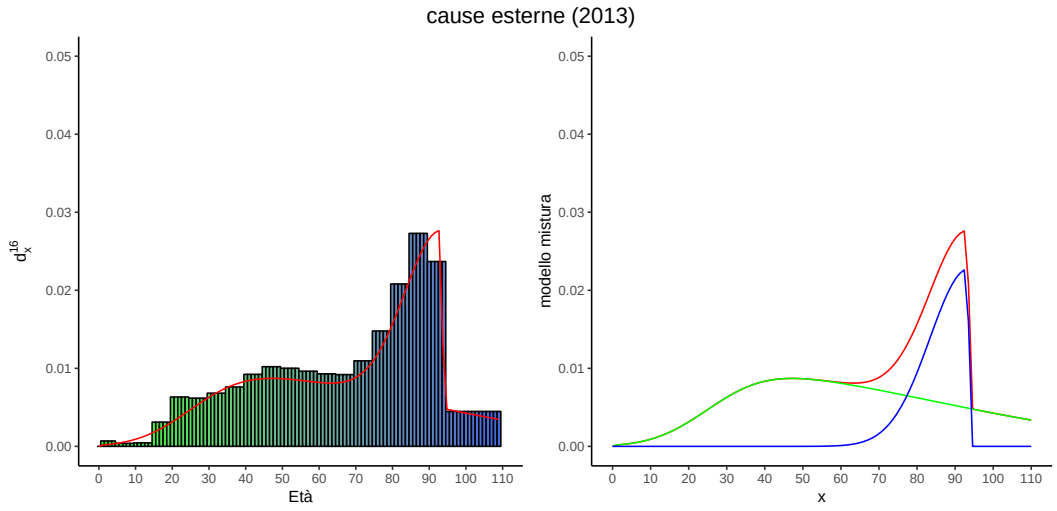

Figura 3.3: Funzione di densità del modello mistura stimato per diverse cause di morte, sovrapposta alla distribuzione dei decessi (a sinistra) e scomposta nelle sue componenti (a destra). Focus sulla causa 16 nel 2013.

Sempre rimanendo in ambito di varianza troppo estesa, le malattie del sangue e degli organi emopoietici (causa 3) hanno la stima di massima verosimiglianza più elevata per σ_m, pari a 43 e 46, che se elevate al quadrato corrispondono a varianze che esplodono. Di conseguenza la curva della mortalità accidentale e prematura risulta essere estremamente ampia, con gli altri due parametri che non riescono ad essere identificati; infatti, la media viene stimata pari ad 1 (l'età minima dei decessi) e l'asimmetria raggiunge il limite inferiore (-0.995). Questo comportamento è riportato in Figura 3.4 per l'anno 2000: l'ampissima curva verde non riesce a catturare sia un piccolo ammontare di decessi nella classe 1-4 anni sia quelli prematuri fino ai 65 anni, determinando uno spostamento all'indietro della curva blu della mortalità adulta. Infatti, nel tentativo di

comprendere la totalità dei decessi, il modello complessivo è troppo influenzato dalle morti che avvengono nella prima fase della vita, con l'effetto risultante che la seconda normale asimmetrica finisce per avere anch'essa deviazione standard elevata (superiore a 10) e quindi essere meno concentrata e più bassa del dovuto, non riuscendo più a descrivere in modo adeguato il picco di decessi attorno agli 85 anni. Nel grafico di destra si può inoltre osservare un "salto" nella curva rossa in corrispondenza del punto di giunzione delle due componenti della mistura, causato dalla massima asimmetria negativa della prima normale asimmetrica. Tuttavia, la causa 3 è quella che si trova all'ultimo posto nella graduatoria delle cause di morte maschili, associata a solo lo 0,4% dei decessi totali, pertanto l'adattamento insoddisfacente del modello potrebbe essere attribuibile anche all'esigua quantità di osservazioni.

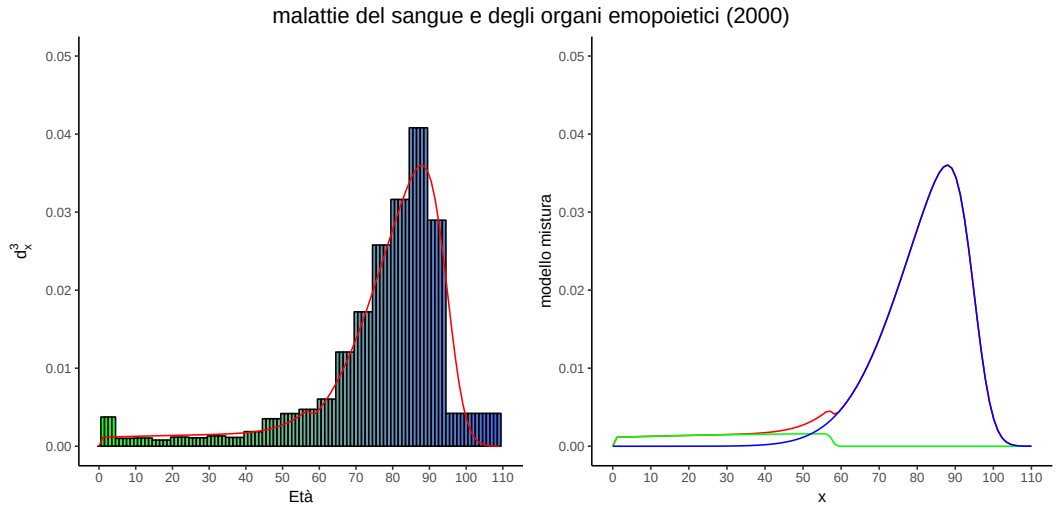

Figura 3.4: Funzione di densità del modello mistura stimato per diverse cause di morte, sovrapposta alla distribuzione dei decessi (a sinistra) e scomposta nelle sue componenti (a destra). Focus sulla causa 3 nel 2000.

Un altro fenomeno avvenuto di frequente durante la procedura di stima è che, nonostante dai grafici delle analisi descrittive per alcune cause di morte sia visibile un ammontare relativamente consistente di decessi prematuri, il modello non è in grado di distinguere la curva della mortalità prematura da quella

adulta. Il risultato è che, invece della combinazione delle due distribuzioni per formare la mistura, *tutti* i decessi vengono considerati causati soltanto dalla mortalità adulta e la curva della mortalità prematura non viene identificata, diventando praticamente ininfluente. Questo accade, ad esempio, nel 2000 per i tumori (causa 2), le malattie endocrine e metaboliche (causa 4), gli altri disturbi del sistema circolatorio (causa 9) e le altre malattie respiratorie (causa 11), tutte cause di morte per le quali dalla Tabella 3.1 si nota un valore di α stimato quasi uguale a 0. Di conseguenza la seconda funzione f_M è caratterizzata da una deviazione standard molto elevata, superiore a 10, per essere sufficientemente ampia da adattarsi da sola a tutti i decessi. Si mostra anche un esempio di questo fenomeno per le cause 2 e 9 nella Figura 3.5. Nei grafici in alto relativi ai tumori l'adattamento globale della curva rossa sembra essere molto buono; tuttavia, ciò nasconde che la curva verde associata alle morti premature è ridotta soltanto ad una linea e la totalità dei decessi, quindi anche quelli tra i 40 e 65 anni, sta sotto ad un'unica curva, quella blu. Questo inconveniente è gravemente dannoso per le nostre analisi e ostacola il nostro obiettivo principale, poiché ci impedisce di capire qual è l'effettivo ruolo svolto dai tumori e da altre cause nella mortalità prematura maschile.

Nelle stime di massima verosimiglianza non sono mancati alcuni problemi tipici dei modelli mistura. È capitato che si siano verificati episodi di *label switching*, in cui le due componenti della mistura sono state permutate (Redner e Walker, 1984). Questa complicazione, di minor importanza rispetto ad altre incontrate, è stata facilmente risolta, dapprima scambiando a mano i parametri stimati delle componenti e prendendo invece dell'α il suo complementare $1 - \alpha$, e in seguito, per comodità, imponendo il vincolo che la media della prima normale asimmetrica fosse minore della media della seconda ($\mu_m < \mu_M$). Quest'ultima è una delle strategie comunemente usate nella pratica per affrontare il problema del *label switching* nei modelli mistura, ovvero imporre un *vicolo di identificabilità* sui parametri delle componenti.

Un altro inconveniente che può accadere lavorando con modelli mistura è

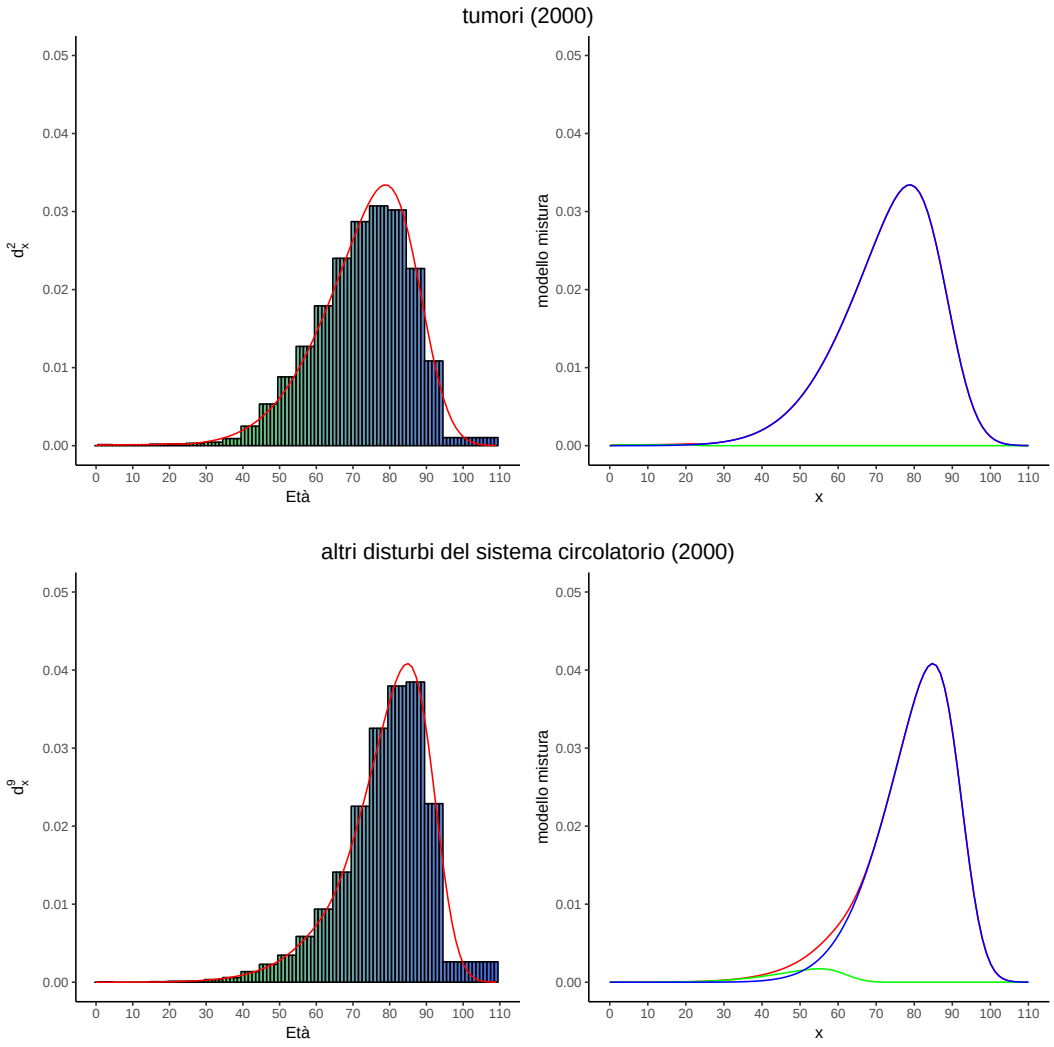

Figura 3.5: Funzione di densità del modello mistura stimato per diverse cause di morte, sovrapposta alla distribuzione dei decessi (a sinistra) e scomposta nelle sue componenti (a destra). Focus sulle cause 2 e 9 nel 2000.

che durante la fase di massimizzazione due componenti *collassino* in un'unica distribuzione. Questo fatto si è verificato anche nelle nostre stime, in particolare per le malattie respiratorie acute (causa 10) e per le malattie del sistema genitourinario e complicazioni della gravidanza (causa 14). Queste sono due cause di morte per le quali la distribuzione dei decessi è fortemente concentrata e spostata verso le età più avanzate; essendo dunque caratterizzate da un bassissimo numero di morti premature prima dei 65 anni, per approssimare l'intera forma della mortalità basterebbe soltanto la curva della mortalità adulta. Infatti, è proprio in questa componente che collassano entrambe le normali asimmetriche, restituendo gli stessi valori stimati per i parametri delle due distribuzioni. Ripetendo la procedura di stima diverse volte, alla fine si è riusciti ad evitare questo problema; infatti nelle Tabelle 3.1 e 3.2 i parametri α per la causa 10 sia nel 2000 che 2013 e per la causa 14 nel 2013 sono attorno a 0.05, dimostrando giustamente la quasi totale assenza della mortalità prematura. Tuttavia, si possono vedere ancora alcune tracce del fenomeno appena descritto nelle stime di massima verosimiglianza dei parametri del modello stimato per la causa 14 nell'anno 2000: la media della prima normale asimmetrica è pari a circa 81 anni e la media della seconda a 84, dunque molto vicine, con parametri di asimmetria quasi identici per entrambe (in prossimità di -0.7). La deviazione standard è, invece, diversa per le due distribuzioni e vale 14 e 7, rispettivamente. Nel grafico di destra della Figura 3.6 relativa a questa situazione si può notare che la curva verde e la curva blu sono molto simili e che la varianza di quella verde è più elevata per descrivere i decessi che si trovano appena prima e subito dopo quella blu. In ogni caso, una distribuzione come questa prima normale asimmetrica così spostata in avanti e con media attorno agli 80 anni non corrisponde alla nostra definizione di mortalità "prematura".

Infine, un ultimo problema di notevole rilevanza che affligge le nostre stime è che spesso la stima di massima verosimiglianza del parametro di asimmetria γ_m della prima normale asimmetrica è molto vicina al più piccolo valore ammissibile oppure si trova proprio sulla frontiera dello spazio parametrico: -0.995. Questo

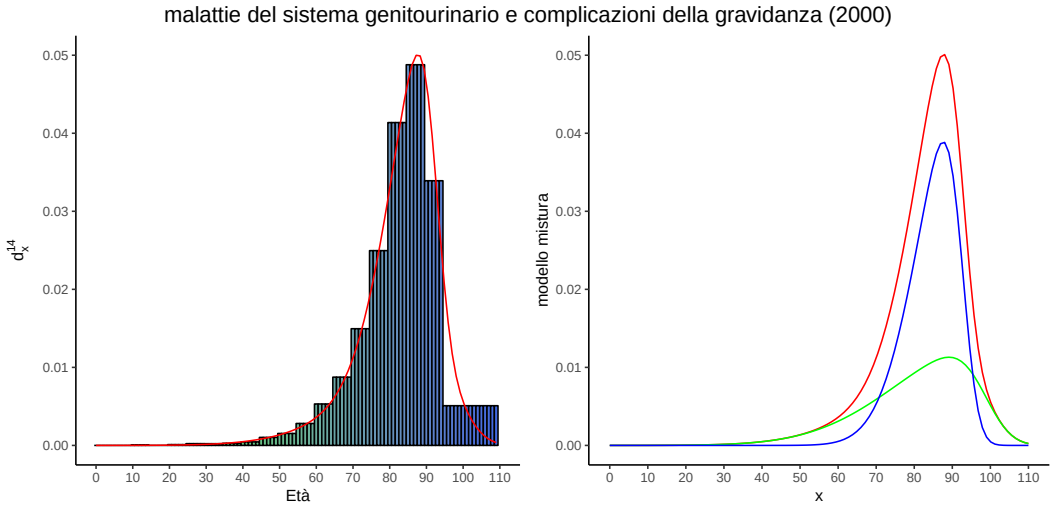

Figura 3.6: Funzione di densità del modello mistura stimato per diverse cause di morte, sovrapposta alla distribuzione dei decessi (a sinistra) e scomposta nelle sue componenti (a destra). Focus sulla causa 14 nel 2000.

succede in concomitanza con i problemi di identificabilità del modello, quando la mortalità prematura incide molto poco sul totale, come appena descritto per le cause 10 e 14, oppure quando la prima componente è assai ampia e i suoi parametri vengono stimati con grande difficoltà, determinando una distribuzione estremamente asimmetrica verso sinistra per riuscire a catturare tutti i decessi che avvengono prima dei 65 anni, come nel caso delle malattie del sangue già discusso in precedenza e mostrato nella Figura 3.4. In altre situazioni, però, accade qualcosa di diverso; ne sono un esempio nel 2013 le malattie infettive (causa 1), le malattie cardiovascolari (causa 8) e gli altri disturbi del sistema circolatorio (causa 9). Dall'osservazione della distribuzione dei decessi per età di queste cause di morte è evidente che c'è una presenza non trascurabile di decessi prematuri; nonostante ciò, il parametro γ_m viene comunque stimato in prossimità di -0.995, come si può leggere dalla Tabella 3.2. Da quest'ultima emergono anche dei valori di α insolitamente elevati, come 0.59 per la causa 1 e addirittura 0.85 per la causa 9; lo stesso vale per le medie della prima normale asimmetrica, che per le tre cause indicate registrano un valore di μ_m prossimo a 80 anni, decisamente troppo alto per essere l'età media della

mortalità prematura. Per capire meglio a che cosa siano dovuti questi valori anomali, si è provato ancora una volta ad evidenziare il ruolo svolto delle singole componenti della mistura, come riportato nei grafici di destra della Figura 3.7 per le malattie infettive (in alto) e per gli altri disturbi del sistema circolatorio (in basso) nel 2013.

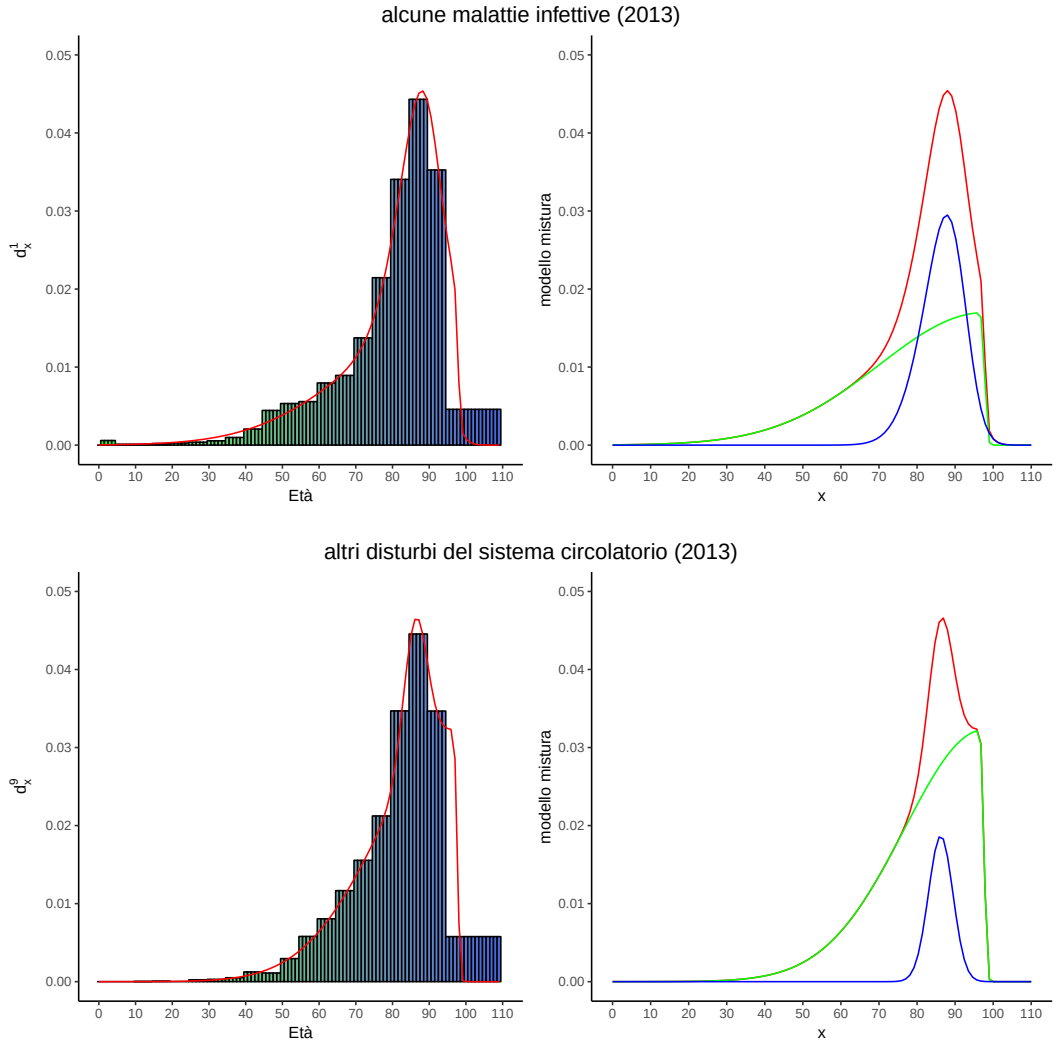

Figura 3.7: Funzione di densità del modello mistura stimato per diverse cause di morte, sovrapposta alla distribuzione dei decessi (a sinistra) e scomposta nelle sue componenti (a destra). Focus sulle cause 1 e 9 nel 2013.

In entrambe le distribuzioni dei decessi per età di sinistra c'è un consistente

ammontare di decessi prematuri prima dei 70 anni, ben riconoscibile rispetto alle morti adulte, molto concentrate verso le età più avanzate. Dalla scomposizione della funzione di densità del modello mistura stimato per le malattie infettive e in modo ancora più estremo da quella relativa agli altri disturbi del sistema circolatorio, emerge che, nel tentativo della prima normale asimmetrica di approssimare la forma della mortalità prematura, il corrispondente parametro di asimmetria tende a valori quasi sulla frontiera dello spazio parametrico ammissibile, determinando perciò una prima componente troppo spostata verso destra (da cui deriva la stima di μ_m pari a circa 80), ovvero una semi-normale con massima asimmetria negativa. Il modello stimato sembra aderire in modo eccellente ai dati nella prima parte della distribuzione dei decessi; nonostante ciò, il problema è chiaramente visibile a destra, con un'anomalia nella curva dopo i 90 anni. Si ritiene opportuno ricordare che le morti attribuite a queste due cause nel 2013 incidono solo per circa il 2% sulla totalità dei decessi, pertanto una ridotta numerosità campionaria potrebbe aver influito negativamente sulle stime.

È noto in letteratura che nonostante la distribuzione normale asimmetrica goda di buone proprietà formali dal punto di vista matematico-probabilistico, nelle applicazioni pratiche c'è la possibilità che la stima di massima verosimiglianza del parametro legato all'asimmetria diverga. La probabilità che insorga questo problema inferenziale diminuisce all'aumentare della dimensione campionaria, ma lavorando con campioni finiti tale probabilità diventa non trascurabile, causando spiacevoli conseguenze nella fase di inferenza. Sono stati proposti vari metodi per ovviare a questa situazione, sia in ambito frequentista che bayesiano (Azzalini e Capitanio, 1999; Sartori, 2006; Azzalini e Arellano-Valle, 2013; Liseo e Loperfido, 2006; Bayes e Branco, 2007).

Alla luce di tutte le problematiche descritte in questo paragrafo emerse dalla stima di massima verosimiglianza del modello mistura per le varie cause di morte, si ritiene che i risultati ottenuti siano qualitativamente insoddisfacenti. Pertanto, non si procede oltre nel fare confronti o provare a trarre conclusioni

per quanto riguarda la mortalità prematura maschile per le diverse cause tra il 2000 e il 2013, in quanto sarebbero di dubbia affidabilità.

Il modello stimato in questo capitolo è stato usato come punto di partenza per cercare un metodo di analisi alternativo. Seguendo questa strada, siamo giunti alla costruzione di un modello gerarchico bayesiano, argomento su cui si concentra il prossimo capitolo di questa tesi.

Capitolo 4

Un modello gerarchico bayesiano per l'analisi delle cause di morte

4.1 Il modello proposto

Date le problematiche riscontrate nelle stime di massima verosimiglianza dei parametri del modello mistura (3.17) stimato marginalmente per ogni causa di morte, soprattutto per quanto riguarda la prima normale asimmetrica, abbiamo provato a costruire un modello gerarchico. Si è ritenuto fosse opportuno, in primo luogo, utilizzare un modello che tenesse in considerazione la struttura gerarchica dei nostri dati: dei decessi raggruppati in categorie di cause di morte. I singoli decessi diventano quindi per noi le unità di primo livello, mentre le cause di morte fungono da gruppi o unità di secondo livello.

Il modello gerarchico rappresenta un compromesso tra due estremi: il modello *pooling* e quello *no-pooling*. La stima *pooling* consiste nell'accorpare insieme tutte le osservazioni in un unico campione, senza distinguere tra i vari gruppi e dunque ignorandone le differenze; la stima *no-pooling*, invece, è di fatto quello che abbiamo fatto finora, ovvero stimare dei modelli separatamente entro

ciascuna categoria di dati, operando come se le osservazioni provenissero da campioni differenti. Il primo caso non rientra nei nostri scopi poiché siamo interessati proprio a capire quali siano le differenze della mortalità prematura tra le varie cause; nel secondo caso, tuttavia, il modello tende a sovra-adattarsi ai dati di ogni gruppo, a scapito della precisione delle stime, con la conseguenza di accentuare la variabilità e le differenze che esistono tra questi e con il rischio di dare eccessiva enfasi a gruppi con bassa numerosità. Inoltre, un modello *no-pooling* può portare anche a dei problemi di identificabilità, come di fatto è capitato nelle stime effettuate nel Capitolo 3. I modelli gerarchici sono estensioni di modelli classici utili quando si hanno delle osservazioni strutturate in gruppi, con la caratteristica che alcuni dei parametri possono variare tra questi gruppi. Ai parametri che variano viene assegnata una comune distribuzione di probabilità, i cui (iper)parametri vengono stimati da tutti i dati; in questo modo, in ogni gruppo di osservazioni avviene un compromesso tra i due tipi di stima precedenti e per questo tali modelli vengono anche detti *partial-pooling*.

Un aspetto attrattivo dei modelli gerarchici, spesso invocato per motivarne l'utilizzo, è la possibilità di beneficiare del cosiddetto effetto "borrowing strength", attraverso il quale è possibile migliorare l'efficienza delle stime di alcuni parametri di interesse "prendendo in prestito" l'informazione dagli altri dati e parametri del modello. Avendo a disposizione dei dati raggruppati in più categorie di cause di morte, l'idea a cui auspichiamo è che la stima dei parametri che appartengono ad un gruppo possa trarre vantaggio dalle informazioni provenienti da altri gruppi. Spesso sono infatti i parametri di gruppi costituiti da un numero esiguo di osservazioni che nelle nostre analisi necessitano di miglioramenti. Desideriamo inoltre sfruttare questa proprietà per riuscire ad identificare la mortalità prematura nelle distribuzioni dei decessi per le cause in cui non viene localizzata in modo corretto.

Riteniamo, pertanto, che dal punto di vista teorico ricorrere ad un modello gerarchico rappresenti una possibile soluzione per le nostre analisi, riservandoci

di capire se nella pratica questa strategia aiuti effettivamente ad ottenere delle stime più stabili e robuste e a risolvere (o almeno attenuare) certi problemi. A tal scopo, abbiamo deciso di seguire un approccio di tipo bayesiano. Dato il modello molto complesso e problematico con cui abbiamo lavorato finora, si pensa che in una situazione come la nostra l'utilizzo dell'inferenza bayesiana rispetto a quella frequentista possa risultare conveniente. Infatti, la stima di un modello in ambito bayesiano ha il vantaggio di poter ricorrere a metodi come *Markov Chain Monte Carlo* (MCMC) che funzionano bene per tante classi di modelli, ad esempio quelli gerarchici come il nostro, che potrebbero essere complicati per un'analisi classica. Inoltre, l'inferenza bayesiana è utile quando si desidera introdurre dell'informazione aggiuntiva in un modello. A tal proposito, un altro motivo che ci ha spinti verso questa strada è che vogliamo sfruttare tutte le informazioni che abbiamo a disposizione sui nostri dati, sia quelle emerse dalle analisi esplorative sia quelle dedotte dalle stime di massima verosimiglianza, e utilizzarle nel nostro modello gerarchico per elicitare le distribuzioni a priori dei parametri. Idealmente le distribuzioni a priori andrebbero specificate in modo da sintetizzare quello che si conosce sull'argomento *prima* di aver osservato i dati. Tuttavia, come accade spesso in realtà, le nostre distribuzioni a priori esprimono quanto ci è noto complessivamente fino ad ora su come variano i parametri tra le diverse cause, informazioni che potrebbero contribuire all'analisi e rivelarsi utili per ottenere delle stime più accurate ed efficienti. Proprio per questo, in alcuni casi abbiamo deciso di essere anche molto informativi.

Rappresentare l'incertezza dei parametri tramite distribuzioni di probabilità è uno dei tratti distintivi dell'inferenza bayesiana; in particolare, tutte le quantità che sono ignote, dunque anche i parametri, sono considerate variabili casuali e perciò ammettono una propria distribuzione a priori. Il punto di arrivo è trovare la cosiddetta distribuzione a posteriori, ovvero la distribuzione dei parametri condizionata ai dati osservati.

A partire da un vettore di parametri ignoti θ con distribuzione a priori $\pi(\theta)$ e da dei dati $y = (y_1, \ldots, y_n)$ con densità congiunta la verosimiglianza $L(\theta|y)$,

la distribuzione a posteriori $\pi(\theta|y)$ è ottenuta attraverso il teorema di Bayes, che aggiorna l'informazione iniziale contenuta in $\pi(\theta)$ alla luce dei dati in y:

$$\pi(\theta|y) = \frac{\pi(\theta)L(\theta|y)}{\int_\Theta \pi(\theta)L(\theta|y)\,d\theta} \tag{4.1}$$

$$\propto \pi(\theta)L(\theta|y) \tag{4.2}$$

Nei prossimi paragrafi verranno presentati i passaggi che hanno portato a specificare la distribuzione a posteriori e tutti gli elementi che costituiscono il nostro modello gerarchico.

4.1.1 La preparazione dei dati

Volendo partire dal modello mistura (3.17) per costruire un modello gerarchico, è emersa subito una complicazione: i parametri di interesse in θ su cui si vuole fare inferenza sono contenuti nelle probabilità di morire $p(x;\theta)$, ovvero si trovano all'interno di un integrale (si veda la formula 3.18). In tale modello i dati osservati d_x^i che costituiscono la verosimiglianza multinomiale sono i conteggi dei decessi avvenuti in ogni intervallo d'età nel periodo considerato. Dal momento che non si hanno informazioni sugli anni esatti che possiede ciascun individuo alla morte e i dati che abbiamo a disposizione sono in forma aggregata, in seguito si è cercato di "disaggregarli", in modo che le osservazioni che entreranno nel nuovo modello siano le singole età x, seppur approssimate, dei soggetti al momento della morte.

Poiché i dati che abbiamo sono raggruppati in classi d'età, si è deciso di procedere creando per ciascuna causa di morte dei vettori contenenti tante osservazioni quanti sono i decessi registrati nei vari intervalli, alle quali è stato attribuito come valore l'età media del corrispondente intervallo. Ad esempio, se nella tavola di mortalità a decremento multiplo relativa all'anno 2000 nella classe d'età 50-54 anni ci sono 1500 decessi maschili per tumori (${}_5d_{50}^2$), allora nel vettore relativo alla causa 2 nel 2000 sono state create, tra le altre, 1500 osservazioni di età pari a 52.5. Per tutte le classi di ampiezza 5 anni è stato

utilizzato il valore centrale, allo stesso modo per la prima classe (1-4) si è scelto il valore 3, mentre per l'ultima classe aperta (95+) 98. Per quanto riguarda quest'ultima, anche se nelle precedenti analisi erano stati considerati i decessi compresi tra 95 e 110 anni, si ritiene tuttavia che la maggioranza di queste morti si concentrino soprattutto nella prima parte dell'intervallo e che gli uomini ancora in vita dopo i 100 anni siano ben pochi; dunque, in mancanza di informazioni più dettagliate, si pensa che, dovendo scegliere un'unica età che sintetizzi tutti i decessi di questa classe, 98 possa essere un valore plausibile. Siamo consapevoli che i dati ottenuti con questo procedimento siano meno accurati rispetto a quelli originali; essi risulterebbero sicuramente più precisi se si conoscesse il numero di decessi per intervalli di singole età anziché, ad esempio, di ampiezza cinque, senza dover ricorrere ad una età media approssimativa per ogni classe. Tuttavia, ora la distribuzione dei decessi continua ad avere la stessa forma di prima, ma con il vantaggio che i dati così trasformati provengono direttamente dalla mistura di due normali asimmetriche, evitando di ricorrere alla distribuzione multinomiale e soprattutto il passaggio per l'integrale.

Dato uno specifico anno di studio, questo metodo è stato ripetuto per tutte le cause di morte; ad ogni causa j, con $j = 1, \ldots, J$ e $J = 15$, corrisponde un vettore di nuove osservazioni $y^j = (y_1, \ldots, y_{n_j})$, dove n_j è il numero di decessi avvenuti in quell'anno per la causa j. Infine, unendo tutti i vari vettori y^j in un unico vettore y, il campione finale di osservazioni per ogni anno dal 2000 al 2013 è del tipo $y = (y_1, \ldots, y_n)$, dove n è il numero totale dei decessi della tavola di mortalità avvenuti tra le età 1 e 95+ in quell'anno per tutte le 15 cause analizzate, ovvero $\sum_{j=1}^{J} n_j = n$.

In seguito indicheremo i singoli decessi (unità di primo livello) con i, per $i = 1, \ldots, n$, mentre per i gruppi di cause (unità di secondo livello) verrà usato l'indice j. I gruppi che abbiamo a disposizione sono 15 e corrispondono alle cause dalla 1 alla 14 più la causa 16 che diventa dunque il quindicesimo gruppo. Ciascuna osservazione y_i rappresenta l'età a cui è avvenuto il decesso i per la causa j.

Infine, insieme ad y è stato creato un altro vettore chiamato "*causa*" della stessa dimensione n, contenente dei valori da 1 a 15 che specificano a quale causa di morte appartiene la corrispondente osservazione di posizione i in y e che sarà utile in seguito per la stima del modello.

4.1.2 Specificazione del modello

Una parte fondamentale nella costruzione di un modello gerarchico bayesiano è scegliere delle opportune distribuzioni a priori per i parametri e stabilire quali parametri far variare tra i gruppi. Per quanto riguarda quest'ultimo aspetto, dopo numerosi tentativi siamo giunti alla formulazione di un modello "ibrido". Ricordiamo che il modello mistura da cui siamo partiti ha 7 parametri: il parametro di mistura α, i parametri $(\mu_m, \sigma_m, \gamma_m)$ della prima normale asimmetrica e $(\mu_M, \sigma_M, \gamma_M)$ della seconda. Dal momento che le stime di massima verosimiglianza dei parametri di f_M non avevano presentato particolari problemi e sono risultate essere nel complesso molto simili tra tutte le cause, per la seconda normale asimmetrica si è deciso di effettuare un *pooling* completo dei dati. Pertanto, per modellare la curva della mortalità adulta le varie cause sono considerate equivalenti, i parametri μ_M, σ_M e γ_M sono gli stessi per tutte le osservazioni y_i e saranno stimati usando la totalità dei dati. Operando in questo modo, l'intenzione è quella di dare maggior stabilità al modello, evitare un numero eccessivo di parametri e potersi concentrare sull'analisi della mortalità prematura, più problematica e per noi di maggiore interesse. Inoltre, poiché le caratteristiche della curva f_m della mortalità prematura dipendono da quelle della curva f_M della mortalità adulta, riuscendo ad individuare una mortalità adulta più o meno comune per tutte le cause, tutti i decessi che si trovano fuori da questa sono considerati provenire dalla mortalità prematura, che può così essere oggetto di confronti coerenti tra le varie cause.

I parametri del modello che sono stati fatti variare tra i gruppi sono il parametro di mistura α e i tre parametri della prima normale asimmetrica μ_m, σ_m e γ_m. In particolare, ad α e σ_m sono state assegnate delle distribuzioni a

priori comuni ma con iperparametri fissati, dunque la loro stima corrisponde, di fatto, a quella di un modello *no-pooling*; la motivazione è che questi parametri sono molto diversi tra le cause, pertanto si è deciso di mantenere il loro comportamento separato in modo da non influenzare troppo le stime tra i gruppi, soprattutto per quanto riguarda α. Inoltre, per modellarli è stata usata una distribuzione uniforme e, dopo alcune prove iniziali, mettere o meno degli iperparametri comuni non ha portato a differenze significative nelle loro stime. I parametri che rendono, appunto, gerarchico il nostro modello sono μ_m e γ_m: variano tra le cause ma appartengono tutti alla medesima distribuzione a priori, che in entrambi i casi è una normale la cui deviazione standard, rispettivamente σ_{μ_m} e σ_{γ_m}, è un iperparametro non noto che deve essere stimato dai dati.

Definiamo ora il nostro modello gerarchico bayesiano. Esistono diversi modi equivalenti per scrivere un modello gerarchico; la notazione che abbiamo utilizzato è quella prediletta da Gelman e Hill (2007).

Date le osservazioni y_i con $i = 1, \ldots, n$ corrispondenti ad unità suddivise nei gruppi $j = 1, \ldots, J$, indichiamo con $j[i]$ il gruppo j che contiene l'unità i-esima. Ad esempio, j[33]=1 significa che il 33^o decesso dei dati ($i = 33$) appartiene alla causa 1.

Si può supporre che ogni singola osservazione derivi da una diversa distribuzione, una mistura di due normali asimmetriche con parametri specifici per il suo gruppo di appartenenza. La distribuzione di probabilità per ogni osservazione y_i è:

$$y_i \sim f(y_i|\alpha_{j[i]}, \mu_{m\,j[i]}, \sigma_{m\,j[i]}, \gamma_{m\,j[i]}, \mu_M, \sigma_M, \gamma_M) \quad \text{per } i = 1, \ldots, n \tag{4.3}$$

dove

$$\begin{aligned} &f(y_i|\alpha_{j[i]}, \mu_{m\,j[i]}, \sigma_{m\,j[i]}, \gamma_{m\,j[i]}, \mu_M, \sigma_M, \gamma_M) = \\ &\qquad \alpha_{j[i]} f_m(y_i|\mu_{m\,j[i]}, \sigma_{m\,j[i]}, \gamma_{m\,j[i]}) + (1 - \alpha_{j[i]}) f_M(y_i|\mu_M, \sigma_M, \gamma_M) \end{aligned} \tag{4.4}$$

La (4.3) rappresenta il modello per i dati e, se consideriamo tutte le n osservazioni, dà origine alla verosimiglianza.

Le distribuzioni a priori scelte per i parametri, per $j = 1, \ldots, J$, sono:

$$\alpha_j \sim U(0, 0.9) \tag{4.5}$$

$$\mu_{m\,j} \sim N(60, \sigma^2_{\mu_m})\ T[-\infty, 75] \tag{4.6}$$

$$\sigma_{\mu_m} \sim U(0, 2.5) \tag{4.7}$$

$$\sigma_{m\,j} \sim U(0, 20) \tag{4.8}$$

$$\gamma_{m\,j} \sim N(0, \sigma^2_{\gamma_m})\ T[-0.8, 0.995] \tag{4.9}$$

$$\sigma_{\gamma_m} \sim U(0, 0.2) \tag{4.10}$$

$$\mu_M \sim N(87, 2^2) \tag{4.11}$$

$$\sigma_M \sim U(0, 9) \tag{4.12}$$

$$\gamma_M \sim SN(-1, 0.5, 1)\ T[-0.995, 0.995] \tag{4.13}$$

dove con $T[a, b]$ si intende una distribuzione troncata nell'intervallo $[a, b]$. Moltiplicando queste densità di probabilità indipendenti si ottiene la distribuzione a priori congiunta del modello.

Come già accennato all'inizio di questo capitolo, per elicitare le distribuzioni a priori dei parametri abbiamo sfruttato le informazioni che provengono dall'osservazione delle forme della distribuzione dei decessi per le varie cause di morte e dalle stime di massima verosimiglianza del modello mistura ottenute nel Capitolo 3. Tenendo ben presente tutte le problematiche riscontrate in precedenza, nel tentativo di porvi rimedio in alcuni casi abbiamo cercato di essere molto informativi. Innanzitutto, per modellare le medie delle due normali asimmetriche abbiamo usato delle distribuzioni normali, per i parametri di asimmetria della mortalità prematura abbiamo specificato delle normali e una normale asimmetrica per il coefficiente di asimmetria della mortalità adulta, mentre per tutte le deviazioni standard presenti nel modello siamo stati abbastanza non informativi, ricorrendo a delle distribuzioni uniformi, come suggerito da Gelman e Hill (2007). Anche per i parametri di mistura α_j abbiamo scelto delle priori uniformi, con l'accortezza di restringere il loro naturale supporto tra $[0, 1]$ ad essere compreso tra $[0, 0.9]$ a causa di un problema insorto durante

le stime, in quanto la distribuzione a posteriori marginale per l'α della causa 2 (tumori) in alcuni anni presentava una piccola moda locale sulla frontiera dello spazio parametrico in 1 che distorceva leggermente le altre stime.

Scendendo nei dettagli, per quanto riguarda le priori dei parametri della curva della mortalità adulta, gli stessi per tutte le cause di morte, sono stati fissati degli iperparametri che riflettessero quanto emerso dalle stime di massima verosimiglianza, che per questa curva si erano rivelate essere abbastanza soddisfacenti. In particolare, la priori normale (4.11) per μ_M è centrata a 87 anni con deviazione standard pari a 2; la priori (4.12) della deviazione standard σ_M è un'uniforme con estremo superiore 9, in modo da stringere la varianza della seconda componente della mistura impedendole di assumere valori troppo elevati, così da prevenire uno dei problemi principali che hanno afflitto le stime di massima verosimiglianza per alcune cause, ovvero che un'ampia seconda curva possa descrivere da sola tutti i decessi e la prima non riesca ad essere riconosciuta. Infine, abbiamo modellato γ_M con una distribuzione normale asimmetrica (4.13) troncata tra -0.995 e 0.995, cioè il supporto del coefficiente di asimmetria, i cui iperparametri sono stati selezionati in modo da attribuire maggior probabilità ai valori negativi, in accordo con la teoria di Pearson (1897) che prevede che per descrivere la mortalità adulta è opportuno ricorrere ad una distribuzione asimmetrica a sinistra.

Le distribuzioni a priori dei parametri della prima componente della mistura e che abbiamo fatto variare tra i gruppi sono state più complesse da elicitare, come del resto ci aspettavamo memori delle difficoltà incontrate nelle stime di massima verosimiglianza. Dopo numerosi tentativi non andati a buon fine, siamo giunti alle seguenti considerazioni. Per prima cosa, per riuscire ad eliminare alcuni dei problemi precedenti che, purtroppo, continuavano a persistere anche in certe stime del modello gerarchico, sono stati imposti due vincoli: la distribuzione normale (4.6) per i $\mu_{m\,j}$ è stata troncata ad assumere massimo valore 75, mentre la distribuzione normale (4.9) per i $\gamma_{m\,j}$ è stata vincolata tra $(-0.8, 0.995)$ invece che tra $(-0.995, 0.995)$. Nel primo caso, il

vincolo a 75 serve per evitare che si verifichino fenomeni di *label switching* e soprattutto il collasso delle due normali asimmetriche della mistura in un'unica distribuzione. Inoltre, in questo modo impediamo alla prima media di assumere valori troppo in avanti, nei pressi degli 80 anni, e costringiamo dunque la prima componente ad essere identificata prima dei 75 anni, limite superiore che riteniamo essere appropriato in quanto la mortalità prematura, per come l'abbiamo definita nel Capitolo 1, quando è presente dovrebbe essere compresa tra circa i 50 e i 65 anni. Il secondo vincolo ha, invece, un'origine un po' più travagliata. Un altro inconveniente capitato di frequente nelle stime di massima verosimiglianza è che il parametro γ_m ha raggiunto la frontiera dello spazio parametrico ammissibile, ovvero è stato spesso stimato nei pressi di -0.995, comportando le spiacevoli conseguenze di varia natura già ampiamente spiegate nel Capitolo 3. Quest'ultime hanno continuato a comparire anche in alcune stime del modello gerarchico bayesiano, arrivando addirittura a causare gravi difficoltà nella convergenza del metodo di stima utilizzato, il quale sarà descritto nel prossimo paragrafo. Siamo riusciti a porre rimedio a questa situazione bloccando l'asimmetria della prima normale asimmetrica ad assumere come valore minimo -0.8 invece di -0.995. Probabilmente esistono anche altre strade che avremmo potuto seguire per ovviare al problema; tuttavia, dal momento che stiamo parlando di casi in cui l'importanza della mortalità prematura di solito è relativamente bassa, l'incidenza del nostro vincolo è limitata poiché, lavorando con la parametrizzazione centrata, quando si passa da -0.995 a -0.8 la forma della distribuzione SN sostanzialmente non cambia di molto. Inoltre, un'elevata asimmetria negativa della curva della mortalità prematura non ha nemmeno un significato demografico particolarmente interessante, a differenza di quella della mortalità adulta. Nel complesso, riteniamo quindi che il vincolo che abbiamo aggiunto abbia creato pochi "danni", con il vantaggio di permetterci di riuscire ad identificare adeguatamente la curva della mortalità prematura per le varie cause e ottenere delle stime più precise per i suoi parametri. Infatti, come sarà più chiaro dai risultati che saranno presentati nel prossimo capitolo, la

media a posteriori del parametro $\gamma_{m\,j}$ di nessuna causa di morte è pari al valore limite -0.8, provando che in realtà la frontiera dello spazio parametrico veniva raggiunta principalmente a causa dei vari problemi di identificabilità del modello di partenza.

Sulla base di queste premesse, le medie $\mu_{m\,j}$ si distribuiscono come delle normali (4.6) centrate in 60, aventi come deviazione standard un iperparametro comune σ_{μ_m} che dev'essere stimato dai dati. Per quest'ultimo è stata scelta un'iperpriori (o priori di secondo livello) uniforme (4.7) nell'intervallo $[0, 2.5]$, in modo da localizzare l'età media della mortalità prematura per tutte le cause indicativamente tra i 50 e 70 anni. Per quanto riguarda le deviazioni standard $\sigma_{m\,j}$, ci siamo mantenuti non informativi specificando delle distribuzioni uniformi (4.8) in un ampio range tra $[0, 20]$, così da ammettere una curva f_m molto estesa per coprire tutti i decessi precedenti ai 65 anni, ma allo stesso tempo tenendone sotto controllo la varianza e impedendole di esplodere. Infine, come già anticipato, i parametri $\gamma_{m\,j}$ hanno come distribuzione a priori una normale (4.9) centrata in 0 e troncata nell'intervallo $[-0.8, 0.995]$ e sono governati da una comune deviazione standard σ_{γ_m} con una distribuzione a iperpriori (4.10) che è ancora una volta un'uniforme tra [0, 0.2]. Questo campo di variazione limitato è stato scelto per comprimere le priori, dando maggior probabilità ai valori di $\gamma_{m\,j}$ compresi approssimativamente tra -0.5 e 0.5 e bassa probabilità a quelli vicini alla frontiera.

Con il modello gerarchico specificato come appena illustrato, abbiamo provato ad ottenere per i parametri delle stime migliori rispetto a quelle precedenti, confidando nell'effetto "borrowing strength" sulle stime dei $\mu_{m\,j}$ e $\gamma_{m\,j}$ specificati a livello di gruppo. Le proprietà delle distribuzioni a posteriori per ogni anno dal 2000 al 2013 sono state analizzate tramite simulazione, stimando il modello con il software Stan, come sarà spiegato nel prossimo paragrafo. Un modello specificato in linguaggio di programmazione Stan deve includere una distribuzione di probabilità per ogni singola osservazione e per ciascun parametro e iperparametro non noto, che nel nostro caso corrispondono

alle funzioni di densità dalla (4.3) alla (4.13). Queste hanno permesso di arrivare alla costruzione della distribuzione a posteriori congiunta del nostro modello, costituita dai seguenti parametri:

$$\pi(\alpha, \mu_m, \sigma_m, \gamma_m, \mu_M, \sigma_M, \gamma_M, \sigma_{\mu_m}, \sigma_{\gamma_m} | y) \tag{4.14}$$

dove $\alpha = (\alpha_1, \ldots, \alpha_J)$, $\mu_m = (\mu_{m\,1}, \ldots, \mu_{m\,J})$, $\sigma_m = (\sigma_{m\,1}, \ldots, \sigma_{m\,J})$ e $\gamma_m = (\gamma_{m\,1}, \ldots, \gamma_{m\,J})$, con $J = 15$, per un totale di 65 parametri, che nel prossimo paragrafo indicheremo per comodità con il vettore θ. In Appendice B è riportato un esempio di codice scritto con questo linguaggio e che è stato usato per specificare e stimare il nostro modello.

4.2 La stima del modello con Stan

Per stimare il modello gerarchico bayesiano proposto abbiamo usato il software Stan. Stan è un linguaggio di programmazione probabilistico moderno, potente, flessibile e open-source per l'inferenza statistica, utilizzato per specificare modelli statistici, soprattutto in ambito bayesiano. Un programma scritto in Stan definisce in modo imperativo il logaritmo di una funzione di densità di probabilità $\pi(\theta|y)$, come ad esempio una posteriori, a meno di costanti di proporzionalità, dove θ è una sequenza di parametri ignoti da modellare e y sono delle variabili osservate. Il programma consiste in una serie di dichiarazioni e istruzioni suddivise in appositi blocchi di codice che specificano tra gli altri, ad esempio, quali sono le variabili osservate, i parametri o variabili casuali non osservate, e il modello, ovvero la log-posteriori. Stan è molto efficiente perché compila i programmi in linguaggio C++ ed inoltre può essere utilizzato in modo interattivo sfruttando diverse interfacce, tra cui R, Python, Julia e molti altri. Per le nostre analisi abbiamo utilizzato RStan, l'interfaccia R a Stan, sfruttando il pacchetto `rstan` (Stan Development Team, 2017a).

Tra i suoi vari utilizzi, Stan è stato sviluppato per supportare l'inferenza bayesiana attraverso metodi *Markov Chain Monte Carlo* (MCMC) (Metropolis

et al., 1953) che si basano su algoritmi Hamiltonian Monte Carlo (HMC) (Duane et al., 1987; Neal, 1994, 2011). Un algoritmo HMC, come tutti i metodi MCMC, permette di simulare da una distribuzione π da cui sarebbe altrimenti difficile simulare tramite altri metodi, costruendo una catena markoviana che abbia π come distribuzione stazionaria (detta anche distribuzione limite o target). La densità π è tipicamente una distribuzione a posteriori $\pi(\theta|y)$. I campioni della catena generati dopo un periodo di *warmup*, ovvero dei vettori contenenti un valore per ogni parametro, possono essere trattati come una sequenza di campioni non-indipendenti e identicamente distribuiti $\theta^{(1)}, \theta^{(2)}, \ldots$, ognuno con distribuzione marginale π. Quando si lavora con modelli con posteriori molto complesse, l'algoritmo Hamiltonian Monte Carlo è una tecnica molto più efficiente e robusta rispetto a metodi come Metropolis-Hastings o Gibbs Sampling (Neal et al., 2011; Hoffman et al., 2014). Un altro vantaggio rispetto a questi due metodi, ad esempio, è che in confronto al primo non necessita la scelta di una distribuzione *proposta* da cui generare un nuovo valore, mentre a differenza del secondo non richiede di specificare le densità condizionate (*full conditionals*), non sempre facili da ricavare. Più in generale, la famiglia HMC si differenzia dai classici algoritmi MCMC perché sfrutta l'evoluzione di un sistema noto come "dinamica Hamiltoniana" che permette di ridurre notevolmente l'autocorrelazione tra successivi stati della catena, evitando l'inefficiente tipico comportamento "random walk" quando c'è un'elevata correlazione nella posteriori che degrada la performance di certe tecniche di simulazione, e soprattutto consente di convergere alla distribuzione target più velocemente rispetto ai più semplici metodi citati. Questa caratteristica risulta essere molto vantaggiosa nell'analisi di modelli gerarchici come il nostro. Infatti, l'HMC esplora in modo più efficiente le intere distribuzioni target, specialmente se complicate e in elevate dimensioni, perché può proporre valori (quasi) ovunque nella posteriori partendo da un qualsiasi punto, sfruttando l'informazione che deriva dall'utilizzo del gradiente della log-posteriori per generare delle transizioni "guidate" verso regioni ad elevata probabilità.

Nel prossimo paragrafo verrà presentato a grandi linee in che cosa consiste l'algoritmo usato da Stan, l'Hamiltonian Monte Carlo (HMC) e la sua variante adattiva No-U-Turn Sampler (NUTS). Per ulteriori dettagli, si vedano l'articolo di riferimento (Carpenter et al., 2017) e il dettagliato manuale di Stan (Stan Development Team, 2017b).

4.2.1 L'algoritmo HMC e la sua variante NUTS

L'algoritmo Hamiltonian Monte Carlo (HMC) è un metodo MCMC che sfrutta le derivate parziali di una funzione di densità dalla quale si desidera simulare per generare delle transizioni efficienti che attraversino l'intera posteriori (Betancourt et al., 2015; Neal et al., 2011). Si supponga da qui in poi di indicare solo con $\pi(\theta)$ la distribuzione target in uno spazio parametrico costituito dal vettore θ dei parametri del modello, detto anche *posizione*. Per prima cosa, l'HMC introduce delle variabili ausiliarie ρ chiamate *momentum* ed estrae valori dalla densità congiunta

$$\pi(\theta, \rho) = \pi(\rho|\theta)\pi(\theta) \tag{4.15}$$

Nella maggior parte delle applicazioni dell'HMC, compreso Stan, la densità ausiliaria è una normale multivariata indipendente dai parametri θ, ovvero

$$\rho \sim N_k(0, \Sigma) \tag{4.16}$$

dove k indica la dimensione del vettore ρ e la matrice di covarianza Σ permette di ruotare e ridimensionare la distribuzione target, in modo da avere una geometria più semplice per il campionamento.

La densità congiunta $\pi(\theta, \rho)$ definisce il cosiddetto *Hamiltoniano* $H(\theta, \rho)$, dove

$$\pi(\theta, \rho) = \exp\{-H(\theta, \rho)\} \tag{4.17}$$

$$H(\theta, \rho) = -\log \pi(\theta, \rho) \tag{4.18}$$

$$= -\log \pi(\rho|\theta) - \log \pi(\theta) \tag{4.19}$$

$$= \quad K(\rho, \theta) + V(\theta) \tag{4.20}$$

Il termine

$$K(\rho, \theta) = -\log \pi(\rho|\theta)$$

è chiamato *energia cinetica* e il termine

$$V(\theta) = -\log \pi(\theta)$$

è chiamato *energia potenziale* e corrisponde all'opposto della log-posteriori. L'Hamiltoniano, ovvero l'energia totale somma dei due tipi di energia precedenti, viene preservato durante ogni transizione, dunque le traiettorie sono limitate lungo certi livelli di energia.

L'algoritmo Hamiltonian Monte Carlo comincia da un set di valori iniziali per i parametri in θ, specificati dall'utente o generati da Stan in modo casuale. A partire da questi valori nella prima iterazione, o dai valori correnti dei parametri in θ in quelle successive, viene generata una transizione verso un nuovo stato attraverso vari passi. Per prima cosa, viene estratto dalla distribuzione (4.16) un vettore di valori casuali per il momentum, indipendentemente da θ; in seguito, il sistema congiunto (θ, ρ) costituito dai valori correnti dei parametri θ e dal nuovo momentum ρ si evolve attraverso le cosiddette equazioni di Hamilton:

$$\frac{d\theta}{dt} = +\frac{\partial H}{\partial \rho} = +\frac{\partial K}{\partial \rho} \tag{4.21}$$

$$\frac{d\rho}{dt} = -\frac{\partial H}{\partial \theta} = -\frac{\partial K}{\partial \theta} - \frac{\partial V}{\partial \theta} \tag{4.22}$$

Con la densità del momentum indipendente dalla distribuzione target, cioé $\pi(\rho|\theta) = \pi(\rho)$, il primo termine della (4.22) è uguale a zero rendendo le derivate pari a

$$\frac{d\theta}{dt} = +\frac{\partial K}{\partial \rho} \tag{4.23}$$

$$\frac{d\rho}{dt} = -\frac{\partial V}{\partial \theta} \tag{4.24}$$

dove $\frac{\partial V}{\partial \theta}$ nella (4.24) è il gradiente del logaritmo della distribuzione target, che dà informazioni sulla geometria della posteriori.

Per risolvere il sistema precedente, Stan utilizza il metodo del *leapfrog* per l'integrazione numerica delle equazioni differenziali. Il leapfrog è il più

semplice algoritmo appartenente alla classe dei cosiddetti algoritmi *simplettici*, specificamente adattato per fornire risultati stabili nello studio di sistemi Hamiltoniani di equazioni. L'algoritmo inizia generando il nuovo vettore per il momentum, indipendentemente dal valore dei parametri θ e dal momentum precedente, che in questo modo non viene conservato attraverso le iterazioni, per poi alternare l'aggiornamento di ρ sulla base di θ e di θ sulla base di ρ, secondo la dinamica Hamiltoniana del sistema. Le soluzioni ottenute dopo un numero L di passaggi leapfrog, ognuno della durata di un piccolo intervallo di tempo discreto ϵ, sono indicate come (θ^*, ρ^*). Per tener conto di possibili errori numerici che potrebbero avvenire durante l'integrazione e correggerli, l'ultimo stadio della procedura consiste nell'applicazione di un passo di accettazione Metropolis (Metropolis et al., 1953; Hastings, 1970), dove la probabilità di tenere la proposta (θ^*, ρ^*) generata dalla transizione a partire da (θ, ρ) corrente è

$$\min(1, \exp(H(\theta, \rho) - H(\theta^*, \rho^*))) \tag{4.25}$$

Sulla base di questa probabilità si decide se accettare la nuova proposta oppure conservare lo stato di partenza. In entrambi i casi, i valori della variabile ausiliaria ρ vengono sempre scartati e si tengono solo quelli del vettore θ, che diventano così il punto di partenza dell'iterazione successiva.

Tutti questi passi vengono ripetuti per un certo numero prescelto di iterazioni, al termine delle quali si ottengono dei campioni di θ provenienti dalla distribuzione target.

L'algoritmo HMC base appena descritto comporta tre parametri che devono essere calibrati: la matrice di covarianza Σ, il tempo di discretizzazione ϵ e il numero di passaggi L dell'algoritmo leapfrog. Nella pratica, l'efficienza dell'algoritmo dipende sensibilmente dalla scelta di questi tre parametri di regolazione, che diventa quindi una fase critica.

I metodi di campionamento implementati in Stan comprendono l'algoritmo HMC base (o statico) e una sua forma adattiva chiamata NUTS (No-U-Turn Sampler). Nelle nostre analisi abbiamo usato la variante NUTS, che è anche

quella implementata di default da Stan, la quale è in grado di selezionare in modo automatico dei valori ottimali per i parametri Σ, ϵ e L (Hoffman et al., 2014). In particolare, il NUTS determina un appropriato numero L di step del leapfrog per ciascuna iterazione attraverso un sofisticato algoritmo ricorsivo ad albero binario bilanciato, usato per costruire un insieme di probabili punti candidati che attraversano un ampio spettro della distribuzione target, fermandosi automaticamente quando la direzione della traiettoria comincia a ritornare sui suoi passi. La procedura precedentemente illustrata subisce di conseguenza alcune modifiche. Tuttavia, in questo modo la variante NUTS ha una performance pari a quella di un algoritmo HMC standard ben regolato, talvolta arrivando ad essere anche più efficiente, e soprattutto gode del vantaggio di non richiedere interventi di *tuning* da parte dell'utente.

Tuttavia, le tecniche HMC base e NUTS sono simili agli altri algoritmi MCMC per un importante aspetto: la validità dell'inferenza è condizionata alla convergenza della catena (o delle catene) alla distribuzione target. Le diagnostiche utilizzate nelle nostre analisi per assicurarci di aver raggiunto la convergenza sono le stesse dei comuni metodi MCMC (Gelman e Rubin, 1992).

Nel prossimo capitolo saranno presentati i risultati che abbiamo ottenuto dalla stima del modello gerarchico bayesiano con Stan.

Capitolo 5

I risultati del modello gerarchico

Il modello gerarchico bayesiano introdotto nel Capitolo 4 è stato stimato con Stan per tutti gli anni dal 2000 al 2013. Per cercare di esplorare l'intero supporto delle distribuzioni a posteriori abbiamo utilizzato più catene indipendenti, inizializzate a partire da valori diversi per i parametri. Son state simulate 3 catene parallele di 5000 iterazioni ciascuna e con un periodo di *warmup* pari a 2500, ovvero scartando la prima metà di ogni sequenza. Le simulazioni sono state eseguite sfruttando il cluster di calcolo "Calculus" installato presso il Dipartimento di Scienze Statistiche. L'accesso al server ha permesso di parallelizzare le simulazioni e velocizzare i tempi di calcolo.

I campioni finali generati dalla stima del modello per ogni anno e che saranno utilizzati per descrivere la distribuzione target sono costituiti da una sequenza di 7500 valori non-indipendenti per ciascun parametro della distribuzione a posteriori e sono stati ottenuti dopo aver eliminato da ogni catena la prima metà di iterazioni corrispondenti al *warmup* e unito assieme in un unico campione i rimanenti valori di tutte e tre le catene. Per verificare di aver raggiunto la convergenza di tutte le catene alla comune distribuzione target sono state utilizzate varie tecniche. Al termine della procedura di stima, Stan fornisce diversi indici riassuntivi che caratterizzano le distribuzioni a posteriori tra cui medie, deviazioni standard e quantili principali per ciascuna variabile, e riporta anche alcune diagnostiche di convergenza tipiche dei metodi MCMC,

sia grafiche che statistiche. In primo luogo, sono stati analizzati i *trace plot* di tutti i parametri, nei quali le traiettorie delle tre catene hanno manifestato un buon *mixing*, con oscillazioni molto veloci e senza mostrare particolari pattern. Per ogni parametro Stan riporta il valore della statistica $\hat{R}$ e quello dell'*effective sample size*. Il *potential scale reduction factor* $\hat{R}$ (Gelman e Rubin, 1992) è un indice che fornisce informazioni sulla convergenza dell'algoritmo, basato sull'analisi delle differenze tra le catene multiple simulate e costruito confrontando la varianza stimata dentro le catene (within) e quella tra le catene (between). Se le catene hanno raggiunto la convergenza alla distribuzione target, la statistica $\hat{R}$ dovrebbe essere vicina a 1. In ognuna delle nostre simulazioni per tutti i parametri del modello l'$\hat{R}$ è risultato essere sempre uguale a 1, dimostrando l'avvenuta convergenza. Poiché i metodi MCMC generano campioni dipendenti, l'*effective sample size*, n_{eff}, indica il numero effettivo di valori indipendenti simulati, cioè quanta informazione è contenuta nel campione finale prodotto dalle catene in confronto ad un campione $i.i.d.$ della stessa dimensione. In particolare, l'*effective sample size* specifica qual è il livello di autocorrelazione delle catene: meno autocorrelate sono le catene, maggiore è il valore di questa statistica e dunque migliore è il *mixing*. Se i campioni fossero stati indipendenti, l'n_{eff} sarebbe stato pari a 7500, ovvero il numero di valori generati in totale da tutte e tre le catene dopo averne scartato la prima metà; tuttavia, nelle nostre stime l'n_{eff} si è mantenuto in un range all'incirca tra 3000 e 7500, escluse alcune rare eccezioni solo per la media e la deviazione standard della seconda normale asimmetrica in cui si arriva anche a 1800, mostrando nell'insieme un'autocorrelazione relativamente contenuta. Anche l'analisi dei grafici di autocorrelazione delle catene per ogni componente ha confermato una dipendenza piuttosto breve nel tempo.

Nel complesso, dal monitoraggio delle diagnostiche di convergenza è emerso che i risultati ottenuti sembrano essere soddisfacenti. Mostriamo dunque le stime prodotte da Stan.

Nelle Tabelle 5.1, 5.2, 5.3, 5.4 e 5.5 sono riportate come statistiche riassuntive

le medie e le deviazioni standard a posteriori per tutti i parametri del modello gerarchico. Per brevità, in tali tabelle sono presenti solo le stime relative agli anni 2000, 2007 e 2013, mentre quelle corrispondenti a tutti gli altri anni del periodo analizzato si trovano in Appendice A. Procediamo ora con alcuni commenti su questi risultati.

Tabella 5.1: Medie e deviazioni standard a posteriori dei parametri relativi alla mortalità adulta nel modello gerarchico stimato per gli anni 2000, 2007 e 2013.

	μ_M		σ_M		γ_M	
Anno	media	s.d.	media	s.d.	media	s.d.
2000	83.096	0.133	8.529	0.092	-0.562	0.014
2007	84.923	0.109	8.186	0.085	-0.579	0.015
2013	86.109	0.099	7.978	0.082	-0.653	0.013

Tabella 5.2: Medie e deviazioni standard a posteriori degli iperparametri del modello gerarchico stimato per gli anni 2000, 2007 e 2013.

	σ_{μ_m}		σ_{γ_m}	
Anno	media	s.d.	media	s.d.
2000	2.476	0.023	0.198	0.002
2007	2.476	0.024	0.197	0.002
2013	2.478	0.021	0.198	0.002

Tabella 5.3: Medie e deviazioni standard a posteriori dei parametri di mistura e dei parametri relativi alla mortalità prematura nel modello gerarchico stimato per l'anno 2000.

	α_j		μ_{mj}		σ_{mj}		γ_{mj}	
Causa	media	s.d.	media	s.d.	media	s.d.	media	s.d.
1	0.356	0.025	57.543	1.319	17.338	0.619	-0.283	0.139
2	0.816	0.026	70.185	0.405	12.830	0.106	-0.645	0.014
3	0.220	0.033	56.641	2.264	19.505	0.454	-0.386	0.150
4	0.231	0.025	63.964	1.590	14.579	0.750	-0.579	0.102
5	0.224	0.011	51.262	0.795	11.382	0.604	-0.182	0.132
6	0.274	0.021	62.273	1.163	18.309	0.560	-0.785	0.016
7	0.216	0.016	65.159	0.905	13.139	0.335	-0.487	0.076
8	0.169	0.024	66.638	1.642	14.186	0.486	-0.701	0.090
9	0.291	0.033	67.727	1.288	12.402	0.486	-0.590	0.110
10	0.034	0.006	57.362	2.292	17.678	1.679	-0.086	0.157
11	0.204	0.025	67.726	1.329	13.097	0.517	-0.701	0.068
12	0.531	0.020	62.723	0.645	12.901	0.364	-0.201	0.071
13	0.074	0.016	59.367	2.197	18.187	1.286	-0.194	0.165
14	0.050	0.010	59.543	2.251	14.944	1.933	-0.056	0.185
16	0.566	0.008	44.351	0.442	19.232	0.339	0.459	0.053

Tabella 5.4: Medie e deviazioni standard a posteriori dei parametri di mistura e dei parametri relativi alla mortalità prematura nel modello gerarchico stimato per l'anno 2007.

	α_j		μ_{mj}		σ_{mj}		γ_{mj}	
Causa	media	s.d.	media	s.d.	media	s.d.	media	s.d.
1	0.311	0.023	60.497	1.311	16.511	0.654	-0.368	0.118
2	0.850	0.027	72.419	0.405	12.885	0.096	-0.701	0.012
3	0.217	0.032	58.868	2.252	19.280	0.613	-0.388	0.148
4	0.264	0.023	65.253	1.334	14.803	0.582	-0.612	0.087
5	0.230	0.011	52.602	0.729	11.366	0.561	-0.271	0.143
6	0.187	0.016	62.754	1.425	18.480	0.556	-0.770	0.038
7	0.155	0.010	62.420	0.931	12.557	0.434	-0.330	0.075
8	0.199	0.026	68.455	1.588	13.884	0.444	-0.711	0.090
9	0.248	0.028	67.815	1.311	12.074	0.554	-0.412	0.128
10	0.048	0.006	56.680	2.228	17.320	1.687	-0.161	0.160
11	0.214	0.021	69.625	1.071	12.562	0.473	-0.734	0.060
12	0.506	0.018	63.930	0.604	13.179	0.320	-0.144	0.060
13	0.076	0.016	59.696	2.250	17.094	1.695	-0.115	0.180
14	0.041	0.008	60.018	2.085	11.693	1.764	-0.063	0.185
16	0.524	0.009	46.755	0.484	18.638	0.356	0.203	0.062

Tabella 5.5: Medie e deviazioni standard a posteriori dei parametri di mistura e dei parametri relativi alla mortalità prematura nel modello gerarchico stimato per l'anno 2013.

	α_j		μ_{mj}		σ_{mj}		γ_{mj}	
Causa	media	s.d.	media	s.d.	media	s.d.	media	s.d.
1	0.258	0.023	62.399	1.474	15.891	0.690	-0.453	0.117
2	0.844	0.023	73.269	0.361	12.754	0.095	-0.721	0.012
3	0.157	0.028	59.227	2.199	18.923	0.867	-0.320	0.155
4	0.256	0.022	66.397	1.273	14.624	0.541	-0.632	0.086
5	0.160	0.007	53.483	0.625	10.321	0.466	-0.385	0.132
6	0.188	0.015	65.173	1.183	17.841	0.530	-0.787	0.014
7	0.145	0.009	62.234	0.963	12.022	0.478	-0.405	0.087
8	0.250	0.024	71.160	1.089	13.468	0.373	-0.769	0.031
9	0.230	0.023	66.165	1.332	12.160	0.701	-0.329	0.115
10	0.042	0.006	57.639	2.073	16.307	1.811	-0.221	0.165
11	0.223	0.022	70.751	1.085	12.245	0.426	-0.688	0.083
12	0.495	0.018	65.070	0.605	12.970	0.334	-0.283	0.063
13	0.119	0.020	62.359	2.037	15.278	1.418	-0.261	0.162
14	0.032	0.007	60.363	2.091	12.647	1.980	-0.087	0.190
16	0.494	0.009	48.876	0.482	18.390	0.326	0.054	0.046

5.1 La mortalità adulta

Per i motivi già discussi nel Capitolo 4 riguardo le scelte che hanno portato alla specificazione del modello gerarchico, si è deciso di stimare un'unica curva per la mortalità adulta, comune per tutte le cause. Effettuando un *pooling* completo dei dati, le stime relative ai parametri della seconda normale asimmetrica sono in linea con quanto affermato da Zanotto (2016) e confermano anche ciò che era emerso dalle precedenti stime di massima verosimiglianza del modello mistura, che per questa componente non erano risultate affette da particolari problemi ed erano sembrate piuttosto soddisfacenti. Le medie a posteriori dei parametri di f_M riportate nella Tabella 5.1 per gli anni 2000, 2007 e 2013 dimostrano ancora una volta che nel periodo in analisi la curva della mortalità adulta ha subito una traslazione verso destra associata ad una compressione. Infatti, la media μ_M aumenta progressivamente il suo valore passando da circa 83 a 86, mentre la deviazione standard σ_M si riduce da 8.53 a 7.98. Anche le stime del parametro γ_M sono diminuite nel tempo, accentuando l'asimmetria negativa da -0.56 a -0.65 e confermando la teoria di Pearson (1897), poiché il numero dei decessi che si verificano nelle età più avanzate e dunque la forma della distribuzione sono influenzati dall'incidenza delle morti avvenute nelle età precedenti.

I tre grafici della Figura 5.1 rappresentano come variano, rispettivamente, i parametri μ_M, σ_M e γ_M nel periodo d'interesse. In particolare, è stato evidenziato in blu l'andamento delle medie a posteriori dei tre parametri in questione dal 2000 al 2013, insieme alla banda del corrispondente intervallo di credibilità HPD (Highest Posterior Density) al 95%.

Nel complesso, tutti i risultati emersi dalla stima del modello gerarchico bayesiano relativi alla componente della mortalità adulta mostrano come in 14 anni la durata della vita degli uomini in Francia si sia leggermente allungata e che in media i decessi adulti si siano spostati e concentrati verso età più avanzate.

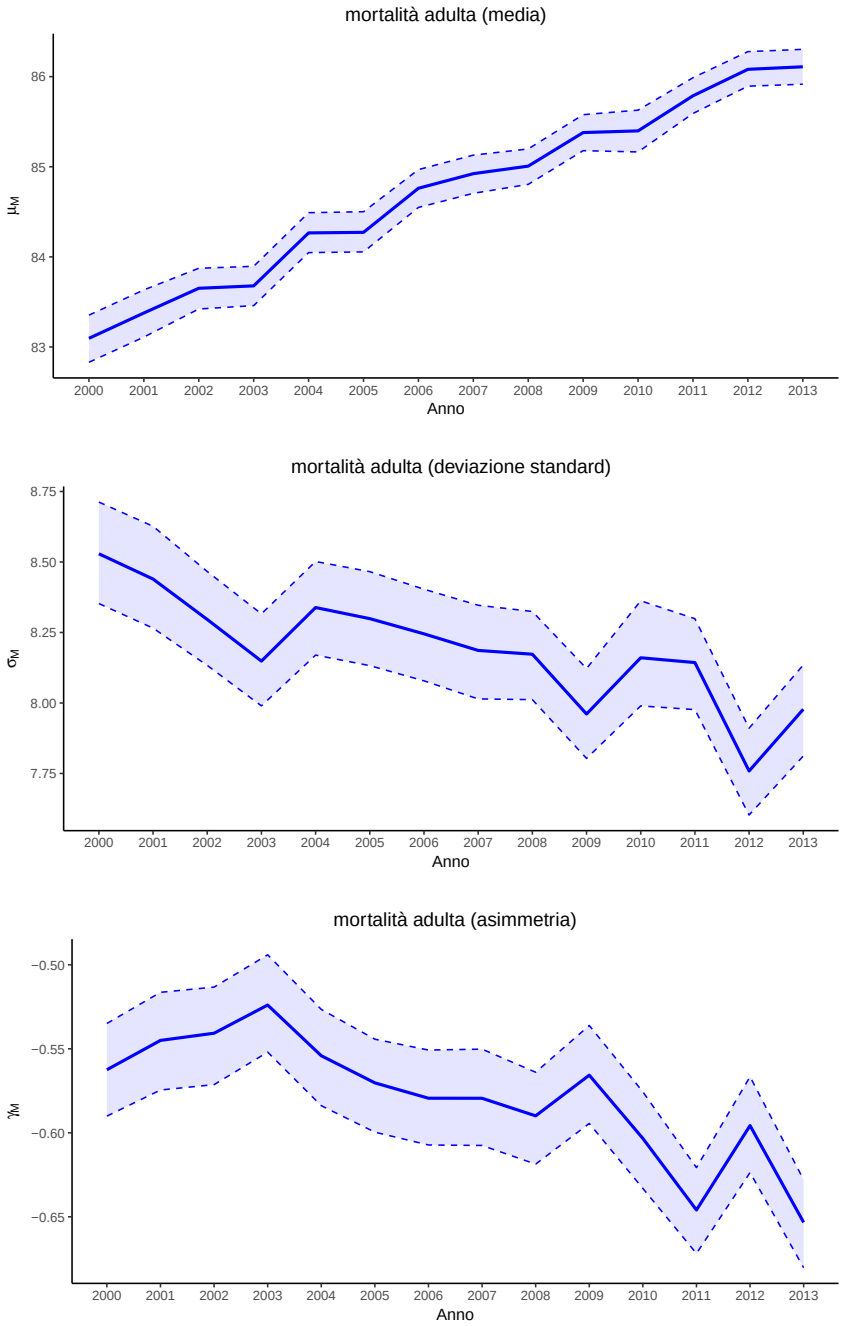

Figura 5.1: Medie a posteriori e intervalli di credibilità HPD al 95% dei parametri μ_M, σ_M e γ_M dal 2000 al 2013.

5.2 La mortalità prematura

Ci concentriamo ora sull'aspetto per noi rilevante: l'analisi della mortalità prematura maschile per diverse cause di morte. Il modello gerarchico proposto in questa tesi ha origine dalle problematiche riscontrate nelle stime di massima verosimiglianza del modello mistura (3.17) descritte nel Capitolo 3, a causa delle quali non è stato possibile fare confronti né trarre conclusioni affidabili riguardo a questa parte della distribuzione dei decessi. Prima di procedere nel commentare i risultati ottenuti con il nuovo modello, si descrive brevemente come l'utilizzo del modello gerarchico abbia permesso di risolvere i problemi incontrati in precedenza.

5.2.1 Risoluzione dei precedenti problemi di stima

Per dimostrare come il modello gerarchico proposto abbia prodotto delle stime qualitativamente migliori rispetto a quelle di massima verosimiglianza, riproponiamo uno ad uno gli aspetti critici dell'inferenza emersi nel Capitolo 3 e mostriamo come questi ora non siano invece più presenti, anche con l'aiuto di alcuni grafici analoghi a quelli precedentemente costruiti dai quali è possibile vedere come l'adattamento del nuovo modello ai dati sia decisamente cambiato e migliorato.

Un primo problema riscontrato nel modello mistura stimato marginalmente per ciascuna causa di morte è che la componente associata alla mortalità prematura a volte era caratterizzata da una varianza troppo ampia, quasi esplosiva, impedendo agli altri parametri di essere identificati correttamente e degradando l'adattamento complessivo del modello alla distribuzione dei decessi. Un esempio di ciò era stato rappresentato in Figura 3.3 per le cause esterne nel 2013. Si riportano ora gli stessi grafici in Figura 5.2, con alcuni adattamenti. A sinistra vi è un istogramma che raffigura la distribuzione delle morti per la causa 16, costituito dai nuovi dati del vettore y^{16} ottenuti attraverso il procedimento descritto nel paragrafo 4.1.1, a cui è stata sovrapposta

una curva rossa che rappresenta la funzione di densità della mistura di due normali asimmetriche con parametri specifici per il gruppo di appartenenza dei dati, mentre a destra sono evidenziate le componenti da cui è costituita. Quest'ultima è la distribuzione di probabilità che possiedono le osservazioni y_i che appartengono alla causa 16 (il quindicesimo gruppo) ed è stata ottenuta sostituendo nella (4.4) le stime del modello gerarchico nel 2013, ovvero le medie a posteriori del parametro di mistura α_{15}, dei parametri $\mu_{m\,15}$, $\sigma_{m\,15}$ e $\gamma_{m\,15}$ nella prima componente e dei parametri μ_M, σ_M e γ_M comuni a tutte le cause nella seconda componente (per i corrispondenti valori si vedano le Tabelle 5.1 e 5.5).

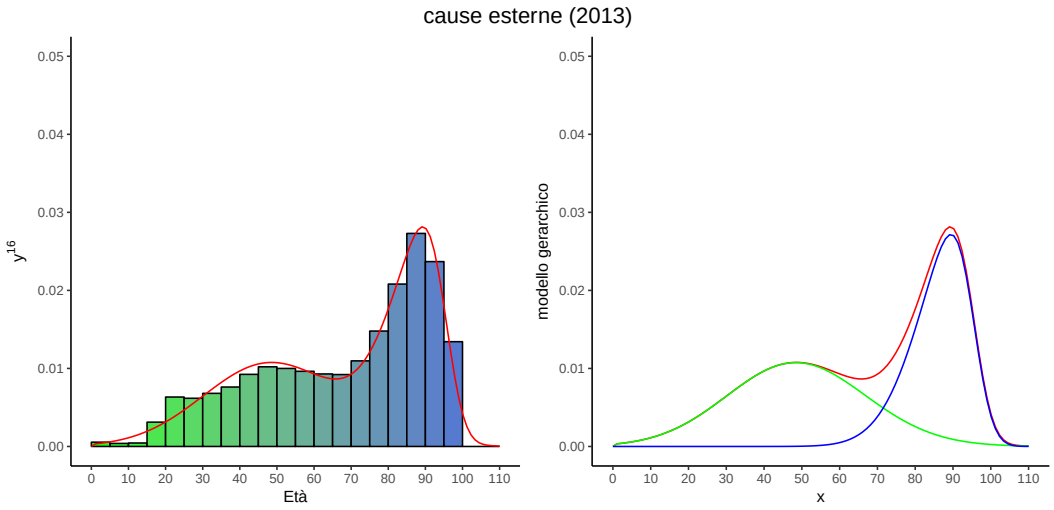

Figura 5.2: Il modello gerarchico: un focus sulla causa 16 nel 2013. Funzione di densità della mistura di due normali asimmetriche con parametri specifici per il gruppo di appartenenza dei dati, sovrapposta alla distribuzione dei decessi (a sinistra) e scomposta nelle sue componenti (a destra).

Da tali grafici si vede chiaramente come sia la "gobba" di decessi prematuri sia l'ammontare di decessi adulti avvenuti per cause esterne ora vengano colti in modo corretto dal modello gerarchico. Nella specificazione della distribuzione a priori per la deviazione standard $\sigma_{m\,j}$ era stata scelta infatti un'uniforme nell'intervallo $[0, 20]$, con lo scopo di tenere sotto controllo la varianza della curva della mortalità prematura e impedirle di assumere valori troppo elevati. A

differenza della stima di massima verosimiglianza per la deviazione standard σ_m che per questa causa nel 2013 era pari a 36, ora per lo stesso anno $\sigma_{m\,15}$ ha valore 18.39 (media a posteriori), dunque è decisamente più contenuta.

Uno dei fenomeni più critici verificatosi durante le stime di massima verosimiglianza è che spesso il modello utilizzato non è stato in grado di distinguere la curva della mortalità prematura da quella adulta, considerando la totalità dei decessi provenienti da un'unica distribuzione, ovvero quella adulta, senza riuscire a riconoscere la prima componente, quindi stimata con un'incidenza quasi pari a 0. Il caso più emblematico e di grande interesse per noi è quello dei tumori nel 2000, mostrato nel Capitolo 3 nei grafici in alto della Figura 3.5, dove la curva verde dei decessi prematuri era ridotta soltanto ad una linea. In contrasto a ciò, si riporta in Figura 5.3 qual è ora l'adattamento della distribuzione di probabilità dei dati nella stessa situazione, sostituendo ai parametri le stime ottenute dal modello gerarchico per la causa 2 nel 2000 come descritto nell'esempio precedente.

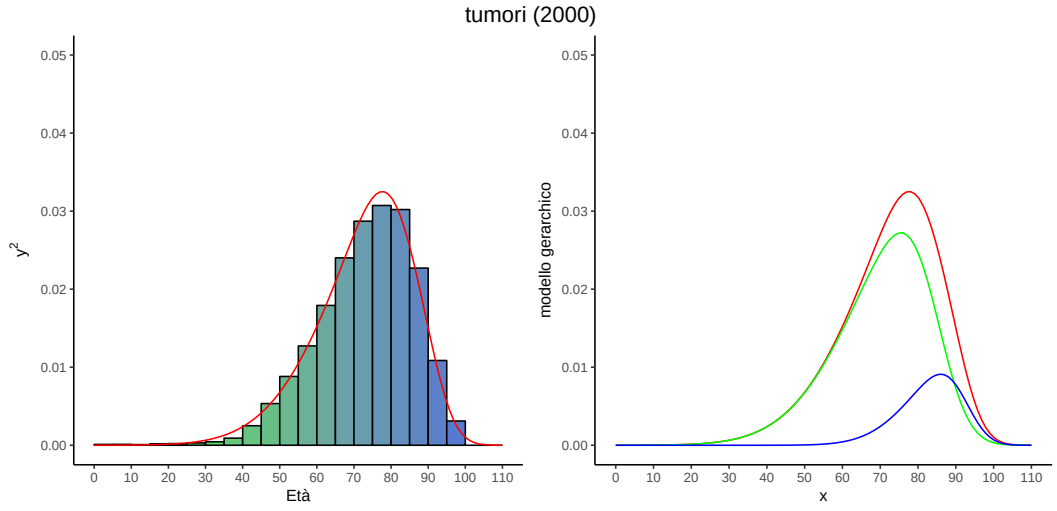

Figura 5.3: Il modello gerarchico: un focus sulla causa 2 nel 2000. Funzione di densità della mistura di due normali asimmetriche con parametri specifici per il gruppo di appartenenza dei dati, sovrapposta alla distribuzione dei decessi (a sinistra) e scomposta nelle sue componenti (a destra).

La distribuzione dei decessi per tumore, la causa che finora è sembrata essere la principale responsabile delle morti premature maschili, adesso è descritta da una mistura di due distribuzioni, di cui il ruolo maggiore è attribuito alla prima componente che approssima la consistente parte di decessi che avviene prima dei 70-75 anni, mentre la curva blu della mortalità adulta coglie i pochi altri rimanenti in età più avanzate. Questi grafici sono più coerenti con le nostre aspettative e da essi si evince l'importanza di questa causa di morte per i decessi prematuri, come sarà più chiaro in seguito, con un parametro di mistura α_2 avente media a posteriori pari a 0.82 nel 2000.

Nelle stime del modello gerarchico per i vari anni sia i problemi di *label switching* sia quelli di collasso delle due componenti della mistura in un'unica distribuzione sono stati completamente eliminati; questo è stato possibile grazie all'introduzione di informazioni a priori che hanno permesso di chiarire le posizioni delle due componenti della mistura e attraverso il vincolo di identificabilità imposto sulla distribuzione a priori delle medie $\mu_{m\,j}$ della prima normale asimmetrica troncata a 75. A conferma di ciò, si mostra la Figura 5.4 in opposizione alla Figura 3.6 del Capitolo 3 che evidenziava il secondo dei due fenomeni appena citati per le malattie del sistema genitourinario nel 2000. L'inconveniente del collasso in un'unica distribuzione (quella adulta) è scomparso ed ora il parametro di mistura α_{14} per questa causa nel 2000 è uguale a 0.05, riflettendo correttamente la quasi totale assenza di morti premature per le malattie del sistema genitourinario, come rappresentato dalla curva verde nel grafico di destra.

Infine, l'ultimo spiacevole episodio avvenuto in precedenza è che spesso la stima di massima verosimiglianza del parametro di asimmetria γ_m della prima normale asimmetrica è risultata essere prossima a -0.995, il più piccolo valore ammissibile che cade proprio sulla frontiera dello spazio parametrico. Come già motivato nel Capitolo 4, per ovviare a questo problema nel modello gerarchico proposto è stato introdotto un vincolo che impedisce a tale parametro di assumere valori inferiori a -0.8. Questo accorgimento ha permesso di identificare

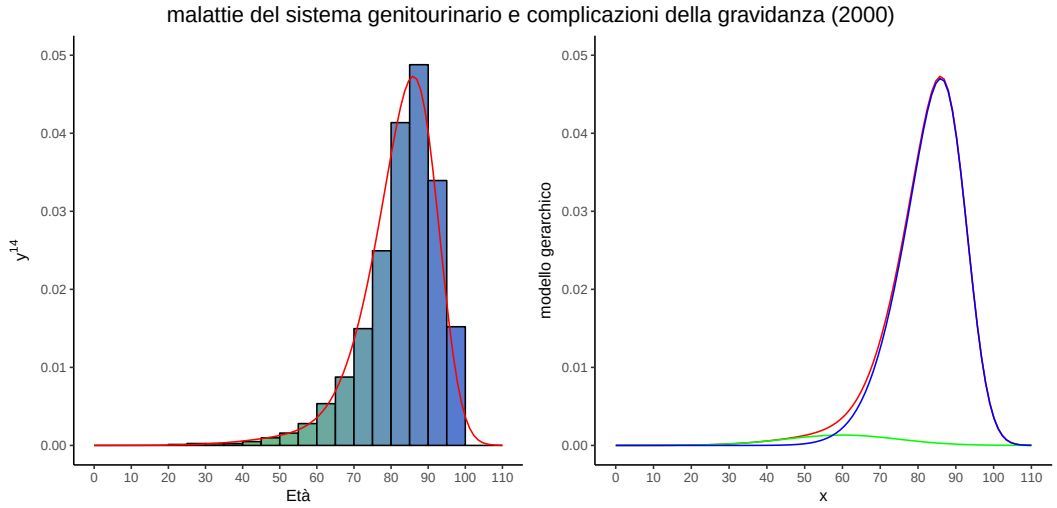

Figura 5.4: Il modello gerarchico: un focus sulla causa 14 nel 2000. Funzione di densità della mistura di due normali asimmetriche con parametri specifici per il gruppo di appartenenza dei dati, sovrapposta alla distribuzione dei decessi (a sinistra) e scomposta nelle sue componenti (a destra).

adeguatamente la curva della mortalità prematura per diverse cause per le quali prima non era possibile farlo. A dimostrazione di ciò, si riporta nella Figura 5.5 l'adattamento del modello ai dati degli altri disturbi del sistema circolatorio, in contrapposizione ai grafici in basso della Figura 3.7 del Capitolo 3 nei quali per la stessa causa 9 nel 2013 si evidenziava all'estremo la massima asimmetria negativa della prima componente. Il consistente ammontare di decessi prematuri prima dei 70 anni è descritto in modo adeguato dalla curva verde della mortalità prematura, la quale adesso possiede una forma molto più simmetrica, e l'anomalia dopo i 90 anni ben visibile in precedenza causata da un'asimmetria vicina a -0.995 ora è del tutto scomparsa. Infatti, come si può notare dalla Tabella 5.5, la media a posteriori di $\gamma_{m\,9}$ nel modello gerarchico stimato per il 2013 è pari a -0.329, dunque nettamente inferiore.

Gli esempi appena commentati dimostrano come la costruzione di un modello gerarchico sia stata una soluzione valida ai nostri problemi poiché, di fatto, sono stati eliminati. L'introduzione dell'informazione che avevamo a disposizione attraverso l'elicitazione di opportune distribuzioni a priori per i parametri, la

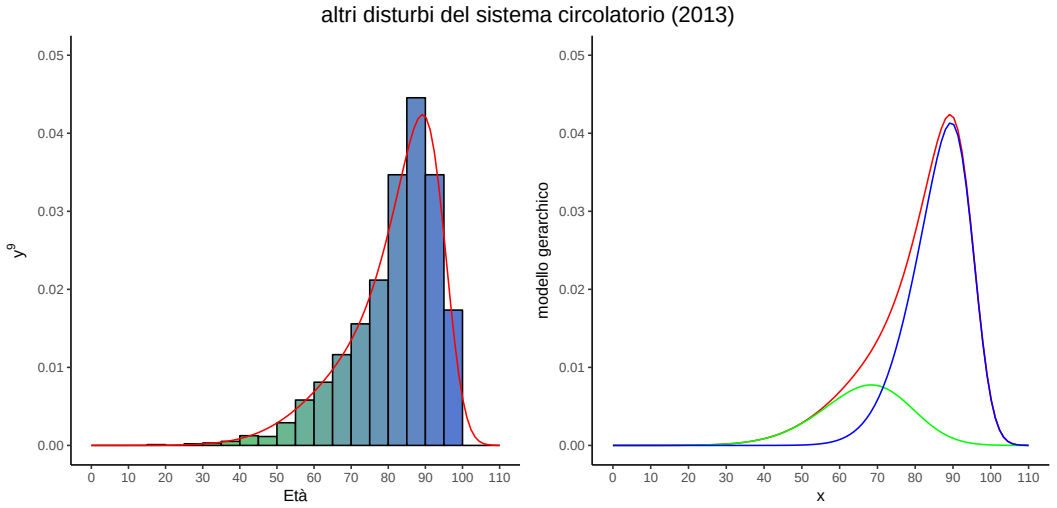

Figura 5.5: Il modello gerarchico: un focus sulla causa 9 nel 2013. Funzione di densità della mistura di due normali asimmetriche con parametri specifici per il gruppo di appartenenza dei dati, sovrapposta alla distribuzione dei decessi (a sinistra) e scomposta nelle sue componenti (a destra).

scelta di un'unica distribuzione per la mortalità adulta comune per tutte le cause e l'effetto "borrowing strength" sulle stime dei μ_{mj} e γ_{mj} specificati a livello di gruppo hanno permesso nella pratica di migliorare significativamente la stabilità e l'efficienza delle stime.

Alla luce di quanto appena affermato, si ritiene che i risultati ottenuti e riportati per il modello gerarchico siano affidabili. È dunque finalmente possibile procedere con l'analisi e il confronto della mortalità prematura maschile per le diverse cause di morte tra il 2000 e il 2013.

5.2.2 Un focus sulle principali cause della mortalità prematura maschile

Ricordiamo che nel modello gerarchico proposto i parametri che sono stati fatti variare tra i gruppi sono il parametro di mistura α e i tre parametri della componente per la mortalità prematura μ_m, σ_m e γ_m. In particolare, la stima di α e σ_m corrisponde a quella di un modello *no-pooling*, con delle distribuzioni a

priori uniformi comuni tra le cause ma con iperparametri fissati, mentre μ_m e γ_m sono i parametri che rendono gerarchico il modello, i quali variano tra i gruppi e appartengono tutti alla medesima distribuzione a priori ma possiedono anche lo stesso iperparametro non noto e che deve essere stimato dai dati. Infatti, le priori per μ_{mj} e γ_{mj} sono delle normali aventi come deviazione standard l'iperparametro σ_{μ_m} e σ_{γ_m}, rispettivamente. Dalla Tabella 5.2 è possibile notare che le medie a posteriori di tali iperparametri si mantengono pressoché costanti nell'intero periodo oggetto di studio e assumono valore 2.476 per σ_{μ_m} e 0.198 per σ_{γ_m}.

Analizziamo ora con attenzione i risultati relativi alla mortalità prematura riportati nelle Tabelle 5.3, 5.4 e 5.5 per tutte e 15 le cause per gli anni 2000, 2007 e 2013. Innanzitutto, ci soffermiamo sull'osservazione dei cambiamenti nel tempo delle medie a posteriori dei parametri di mistura α_j che indicano l'importanza della mortalità accidentale e prematura nella distribuzione complessiva dei decessi rispetto a quella adulta. Si evince che le prime tre cause della mortalità prematura maschile in tutti i 14 anni sono i tumori (causa 2), le malattie del sistema digestivo (causa 12) e le cause esterne (causa 16), poiché hanno la più alta incidenza attorno a 0.84, 0.50 e 0.52, rispettivamente. Le cause che invece non hanno effetto sulla mortalità prematura sono le malattie respiratorie acute (causa 10) e le malattie del sistema genitourinario e complicazioni della gravidanza (causa 14), con un α quasi pari a 0 e costante nel tempo.

Nel complesso, nel periodo 2000-2013 il livello della mortalità prematura si è abbassato per le malattie infettive (causa 1) passando da 0.36 a 0.26 e per le cause esterne da 0.57 a 0.49. Anche per le malattie cardiache (causa 7), che dalle analisi esplorative iniziali erano state classificate come seconda causa di morte maschile (21% del totale), il parametro α è diminuito da 0.22 a 0.14, suggerendo che questa causa è responsabile principalmente delle morti in età più anziane. Al contrario delle nostre aspettative, non si rilevano invece innalzamenti significativi del ruolo assunto da qualche specifica causa di morte nella mortalità prematura; le rare eccezioni riguardano le malattie cerebrovascolari (causa 8)

per le quali la media a posteriori di α aumenta da 0.17 a 0.25 e le malattie della pelle e del sistema muscolo-scheletrico (causa 13), con un valore di α che si mantiene stabile a 0.07 in tutti gli anni e cresce a 0.12 solo negli ultimi due. Per quanto riguarda i tumori, si possono notare delle oscillazioni di tale parametro, come sarà spiegato a breve. Per tutte le altre cause l'incidenza della mortalità prematura non subisce variazioni di rilievo e si mantiene all'incirca sullo stesso livello nel tempo.

Si riportano alcune osservazioni generali in merito ai tre parametri che determinano le caratteristiche della prima normale asimmetrica per le varie cause. Dall'analisi delle stime a posteriori contenute nelle tabelle mostrate in questo capitolo emerge che tutte le medie μ_{mj} assumono valori compresi tra circa 55 e 70 e nella quasi totalità dei casi aumentano tutte leggermente di anno in anno. Questo fenomeno conferma quanto già sostenuto da Zanotto (2016), ovvero che in contemporanea ad una traslazione verso destra della curva della mortalità adulta è avvenuto uno spostamento in avanti anche della curva della mortalità accidentale e prematura, a dimostrazione della forte relazione esistente tra queste due componenti poiché l'evoluzione della mortalità prematura segue e dipende dai cambiamenti di quella adulta. Le deviazioni standard σ_{mj} più elevate (attorno a 18-19) appartengono alle malattie del sangue (causa 3) e alle cause esterne (causa 16), come era già emerso dalle stime di massima verosimiglianza, con la differenza che ora nel modello gerarchico le varianze di tali distribuzioni molto estese sono tenute sotto controllo. Infatti, i valori dei σ_{mj} sono sempre inferiori a 20, dunque il nuovo modello consente di stimare in modo più efficiente tutti i parametri della prima normale asimmetrica per le cause dove prima questo non era possibile. La medesima considerazione vale anche per le stime dei parametri γ_{mj}: in seguito al vincolo imposto a -0.8 sul minimo valore ammissibile per il coefficiente di asimmetria della prima normale asimmetrica, per nessuna causa di morte la media a posteriori di tale parametro è pari al valore limite, a prova del fatto che nelle stime di massima verosimiglianza la frontiera dello spazio parametrico (-0.995) era raggiunta a

causa della combinazione dei diversi problemi di identificabilità che affliggevano quel modello di partenza.

Dopo aver individuato quali sono le cause che incidono di più sulla mortalità prematura per gli uomini, ci concentriamo sulle prime tre di queste, ovvero i tumori, le cause esterne e le malattie del sistema digestivo.

In analogia a quanto fatto nel paragrafo precedente per descrivere le variazioni dei parametri che costituiscono la curva della mortalità adulta f_M, nelle Figure 5.6, 5.7, 5.8 e 5.9 sono rappresentati i cambiamenti nel tempo dei parametri $\alpha_{m\,j}$, $\mu_{m\,j}$, $\sigma_{m\,j}$ e $\gamma_{m\,j}$ per le tre cause di morte d'interesse. In particolare, sono riportati gli andamenti delle medie a posteriori dal 2000 al 2013 assieme ai corrispondenti intervalli di credibilità HPD al 95%. In ciascuna figura sono presenti tre grafici: quelli in alto e colorati di rosso si riferiscono ai tumori (causa 2), quelli viola al centro sono relativi alle malattie del sistema digestivo (causa 12) ed infine quelli che si trovano in basso e colorati di verde appartengono alle cause esterne (causa 16).

Al terzo posto per ordine di importanza troviamo le malattie del sistema digestivo, le quali sono tra l'altro responsabili solo del 4,5% delle morti maschili complessive. Il parametro di mistura α_{12} per questa causa è pari a 0.53 nel 2000, scende a 0.48 nel 2006 per poi mantenersi attorno a questo valore negli anni successivi. La media $\mu_{m\,12}$ aumenta nel tempo da 62.72 fino a 65.07, la deviazione standard $\sigma_{m\,12}$ rimane pari a circa 13, mentre il coefficiente di asimmetria $\gamma_{m\,12}$, esclusi due picchi a -0.1 nel 2002 e nel 2011, si aggira su -0.2, per diminuire infine a -0.3 nel 2013.

La seconda causa di morte prematura maschile sono le cause esterne, le quali occupano anche il terzo posto nella graduatoria delle principali cause di morte in totale e comprendono incidenti di trasporto, omicidi, suicidi e avvelenamenti. Nei grafici che rappresentano la distribuzione dei decessi avvenuti per questa causa è ben visibile in tutti gli anni un'accentuata "gobba" associata alla mortalità prematura, staccata dalle morti delle fasce d'età più anziane. È infatti questa l'unica causa tra quelle a nostra disposizione che registra

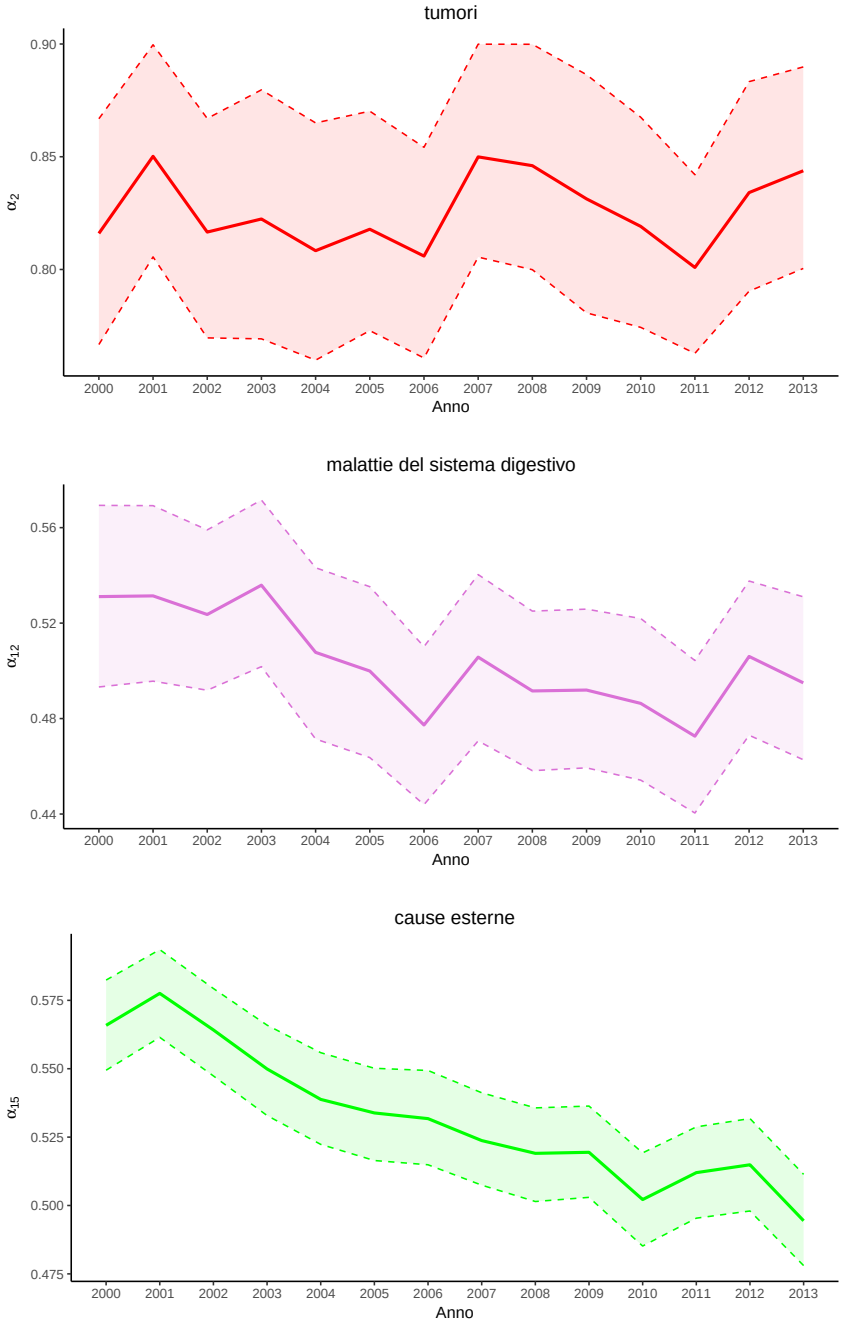

Figura 5.6: Medie a posteriori e intervalli di credibilità HPD al 95% dei parametri α_j per le cause 2, 12 e 16 dal 2000 al 2013.

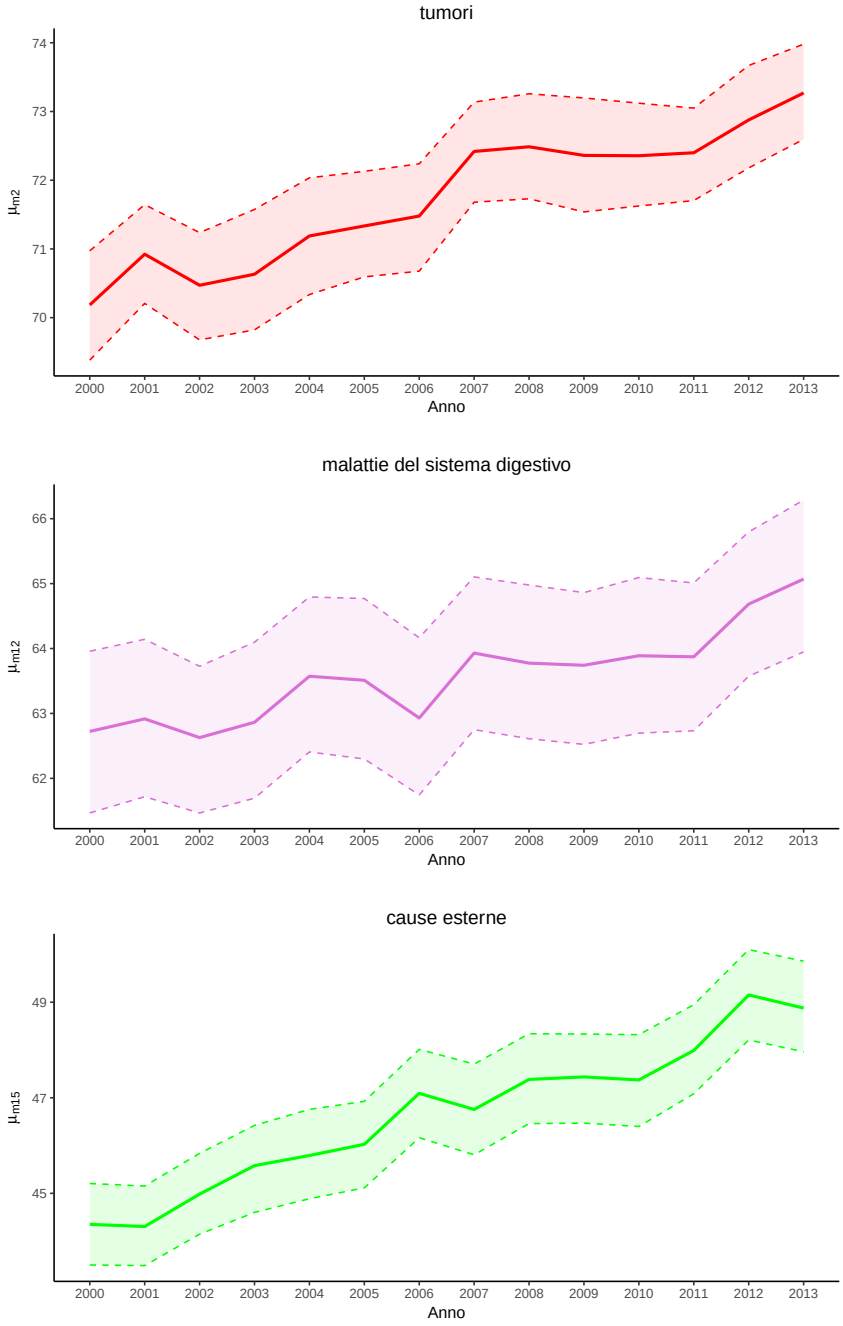

Figura 5.7: Medie a posteriori e intervalli di credibilità HPD al 95% dei parametri μ_{mj} per le cause 2, 12 e 16 dal 2000 al 2013.

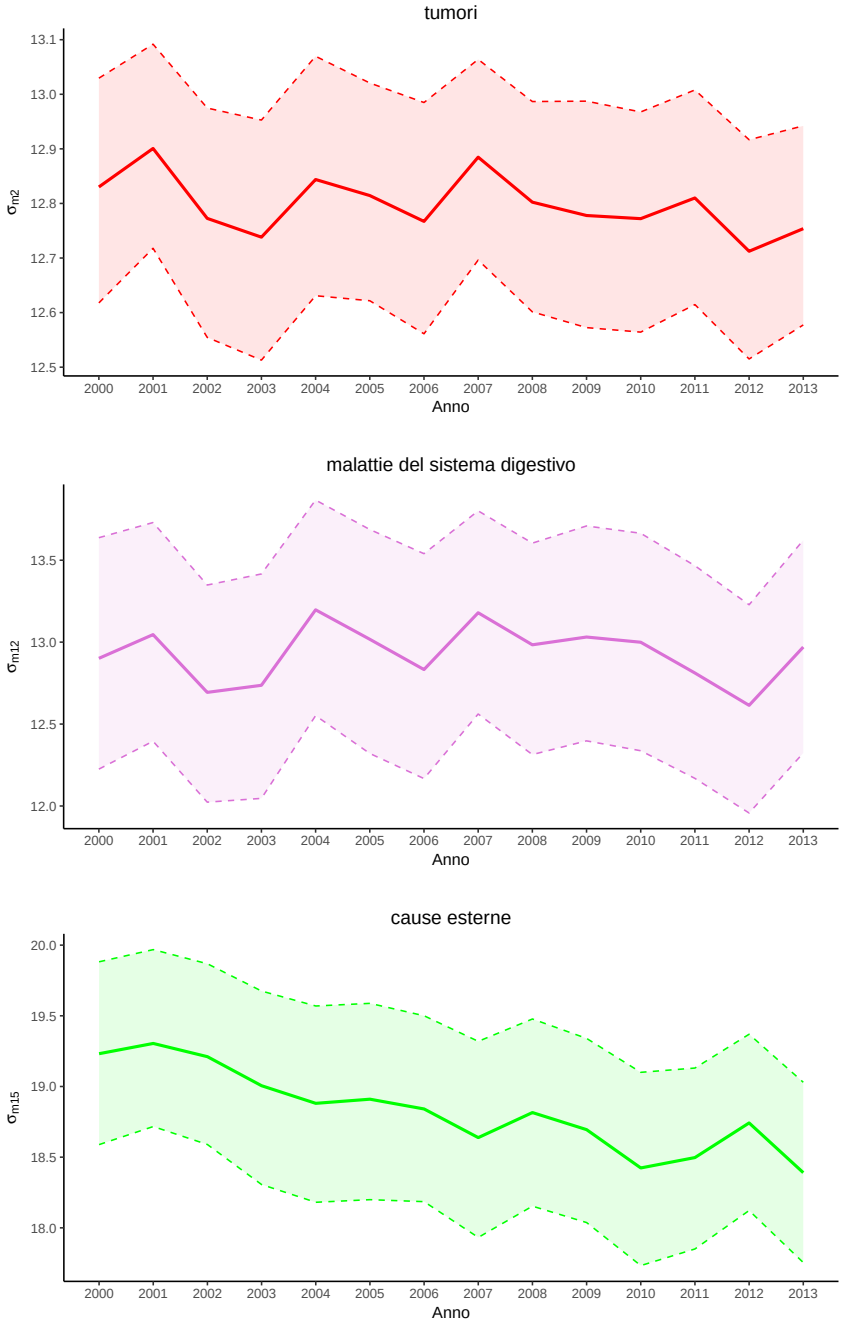

Figura 5.8: Medie a posteriori e intervalli di credibilità HPD al 95% dei parametri σ_{mj} per le cause 2, 12 e 16 dal 2000 al 2013.

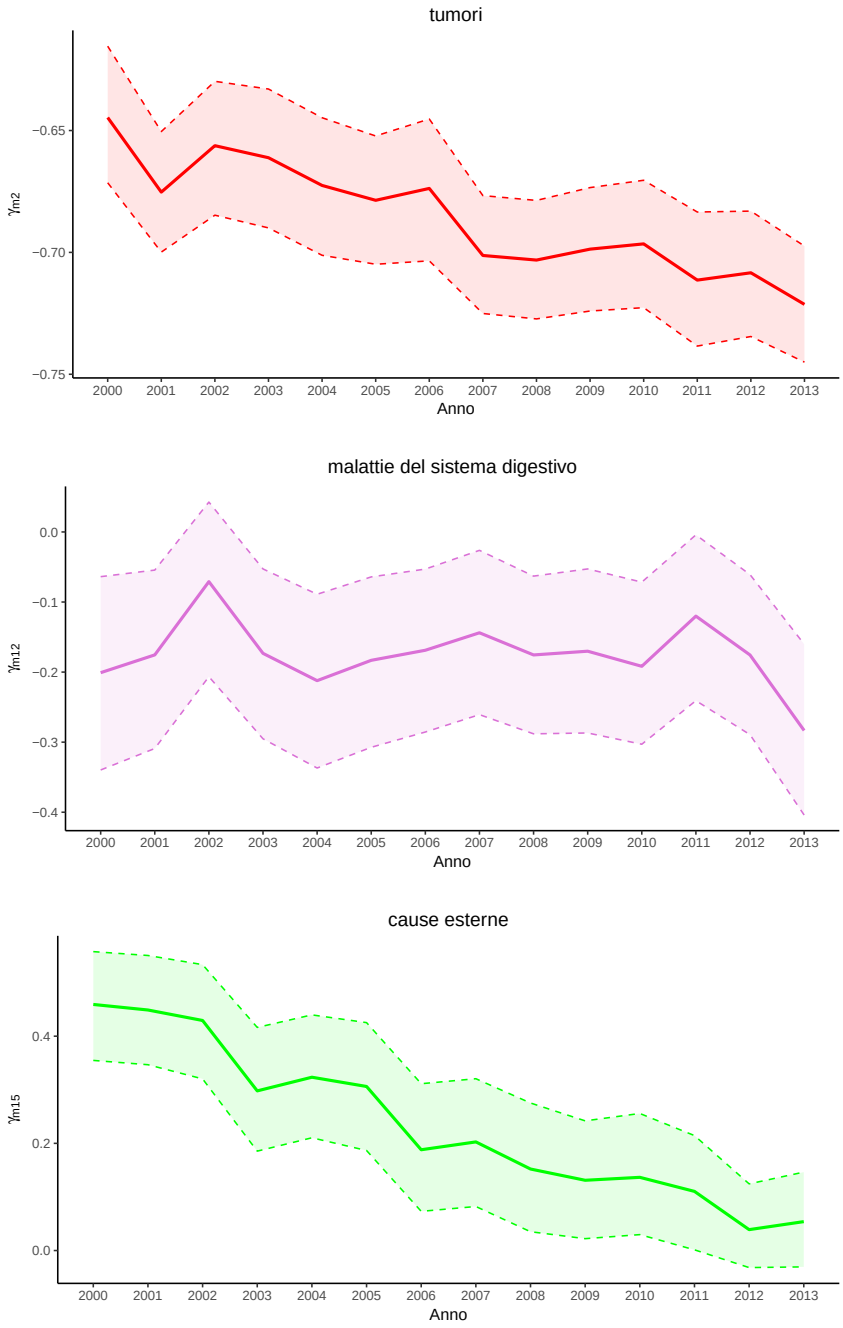

Figura 5.9: Medie a posteriori e intervalli di credibilità HPD al 95% dei parametri γ_{mj} per le cause 2, 12 e 16 dal 2000 al 2013.

anche delle morti accidentali attorno ai 20-40 anni. Tuttavia, in 14 anni la forma della mortalità accidentale e prematura per cause esterne sembra essere molto cambiata: è avvenuto un abbassamento notevole del contributo di tale componente accompagnato da una concentrazione e simmetrizzazione della corrispondente distribuzione. Infatti, dalle tabelle e dai grafici riportati si osserva che α_{15} diminuisce drasticamente da 0.57 a 0.49. La media $\mu_{m\,15}$ aumenta da 44.35 a quasi 49; appartiene proprio a questa causa l'età media di morte più bassa in assoluto, come prevedibile dal momento che le morti per cause esterne spesso si concentrano soprattutto nella prima metà della vita, mentre le medie di tutte le altre cause partono da oltre i 55 anni. Inoltre, la deviazione standard $\sigma_{m\,15}$ decresce leggermente nel tempo da 19.23 a 18.39, ma ciò che colpisce è l'enorme abbassamento del valore di $\gamma_{m\,15}$ da 0.46 a circa 0. Il parametro di asimmetria di tale causa è l'unico ad essere positivo, mentre tutti gli altri hanno segno negativo. Questa differenza è dovuta ad un consistente ammontare di decessi maschili che avvengono anche attorno ai 20 anni in Francia per cause esterne nei primi anni 2000, il quale poi si riduce nel tempo con picco verso i 45-50 anni. Per riuscire a descrivere anche questi cosiddetti decessi accidentali che capitano nelle età giovanili, la prima componente del modello a partire dal 2000 è caratterizzata da un coefficiente di asimmetria positivo e diventa dunque asimmetrica a destra; negli anni successivi le morti a 20 anni si abbassano gradualmente e si concentrano attorno ai 45-50, pertanto il parametro di asimmetria scende via via fino ad arrivare a zero negli ultimi due anni. Questo valore che fa sì che la distribuzione della mortalità accidentale e prematura per cause esterne a fine periodo si avvicini a quella di una normale simmetrica avente la tipica forma a campana, come si può anche vedere per l'anno 2013 dalla Figura 5.2.

Come era emerso fin da subito dalle analisi esplorative condotte all'inizio di questa tesi, con una prevalenza del 33% e quasi un terzo dei decessi totali la principale causa di morte maschile sono i tumori, i quali ora si confermano essere anche la prima causa responsabile della mortalità prematura in Francia

dal 2000 al 2013. Nella distribuzione dei decessi per età relativa ai tumori nell'intero periodo si osserva sempre un elevato numero di morti che avvengono prima dei 70 anni, molto maggiore in confronto a quello delle altre cause; tuttavia, uno degli svantaggi già incontrati in precedenza è che per questa causa non si riesce a distinguere in modo chiaro il confine tra quelli che sono considerati decessi prematuri e quelli adulti, poiché sembrano sovrapporsi e mescolarsi assieme. Questo problema controproducente per i nostri scopi è stato ovviato attraverso il modello gerarchico come mostrato all'inizio di questo paragrafo. Infatti, riuscendo ad individuare una componente per la mortalità adulta comune a tutte le cause, tutti i restanti eventi che capitano fuori da questa vengono considerati prematuri per ciascuna specifica causa. Una delle conseguenze di questa scelta è che per i tumori si ottiene che più dell'80% dei decessi è associato alla mortalità prematura. Infatti, come mostrano i grafici, il parametro di mistura α_2 è molto elevato, pari a 0.82 nel 2000, cresce a 0.85 l'anno successivo per poi tornare al livello di partenza e stabilizzarsi fino al 2006; nel 2007 avviene un decisivo innalzamento a 0.85, valore che poi cala ancora via via fino ad arrivare a 0.80 nel 2011, per infine risalire di nuovo a 0.84 nel 2013. Questo andamento sembra essere molto simile a quello dell'evoluzione delle proporzioni dei decessi maschili per tumore riportato nella Tabella 2.2 delle analisi descrittive del Capitolo 2, il quale era caratterizzato da un picco nel 2007 e poi una graduale decrescita accompagnata alla fine da variazioni altalenanti. La media $\mu_{m\,2}$ della prima normale asimmetrica è la più alta registrata tra tutte le cause: è uguale a 70.18 nel 2000 e poi è caratterizzata da un trend sempre crescente fino a quota 73.27 nel 2013. La deviazione standard $\sigma_{m\,2}$ si mantiene costante attorno a 12.80, mentre il coefficiente di asimmetria $\gamma_{m\,2}$ si riduce durante l'intero periodo da -0.64 a -0.72, accentuando dunque l'asimmetria a sinistra della prima distribuzione del modello.

Un'incidenza della mortalità prematura oltre l'80%, valore di gran lunga superiore a quello di tutte le altre cause, fa apparire i tumori come l'unica causa per cui la mortalità prematura ha un ruolo predominante rispetto a

quella adulta. Tuttavia, 70-73 anni sembrerebbe un'età media elevata per essere considerata, appunto, "prematura", specialmente se confrontata con il range entro cui si trova la quasi totalità delle medie di tutte le altre cause di morte, che di fatto coincide con la nostra definizione iniziale di mortalità prematura confinata tra circa i 50-65 anni.

Questi risultati ci lasciano qualche dubbio. Nel tentativo di dare un'interpretazione più precisa a quanto appena descritto, siamo andati ad analizzare con attenzione i grafici mostrati nel Capitolo 2 nelle Figure 2.2, 2.3, 2.4 e 2.5 riguardanti le forme delle distribuzioni dei decessi per le diverse cause di morte nel 2013 confrontate tra uomini e donne. Per tutte le cause il picco più alto di decessi maschili si osserva nella classe d'età 85-89 e in rari casi anche 90-94, ad eccezione dei tumori che sono l'unica causa per cui la moda della distribuzione si trova, invece, nella classe immediatamente precedente, ovvero 80-84 anni. Pertanto, chi muore per tumore sembra morire mediamente prima rispetto alle altre cause. Un altro indizio a sostengo di questa ipotesi deriva dalle stime di massima verosimiglianza ottenute nel Capitolo 3: come avevamo già sottolineato in quell'occasione, nei risultati relativi alla curva della mortalità adulta i tumori si sono distinti per essere la sola causa di morte per cui la media μ_M è risultata essere sempre molto inferiore in confronto a quella di tutte le altre, fermandosi al massimo tra i 73-76 anni.

Se consideriamo ogni causa separatamente, sia dall'osservazione delle analisi esplorative che dal commento delle stime di massima verosimiglianza, si evince che volendo approssimare la distribuzione dei decessi per tumore con una mistura di due componenti sembra che la curva della mortalità adulta per questa causa si trovi prima, spostata più indietro rispetto a quella delle altre cause. Questa deduzione potrebbe apparire ragionevole da un lato, tuttavia dall'altro riteniamo che la mortalità adulta debba essere una sola. È proprio questo uno dei motivi che ci ha spinto a specificare nel nostro modello gerarchico un'unica distribuzione per questa componente, uguale per tutte le cause, in modo da far variare tra i gruppi solo la mortalità prematura e rendere quindi possibili e coerenti eventuali

confronti. Questa (discutibile) scelta fa sì che stimando il modello gerarchico per i dati che abbiamo a disposizione i decessi che si trovano al di fuori di quella che viene identificata come l'effettiva comune mortalità adulta siano considerati prematuri, dunque anche quelli che capitano in età relativamente avanzate come 70-73 anni, perché di fatto non rientrano nella stretta e concentrata curva adulta centrata sugli 83-86 anni. Ecco da che cosa ha origine quel valore superiore a 0.80 per la prevalenza della mortalità prematura per i tumori. Tale valore potrebbe sembrare di per sé anomalo se le morti per tumore venissero considerate da sole, ma dal momento che stiamo utilizzando un modello gerarchico i decessi per tumore vengono uniti a quelli di tutte le altre cause e dunque quella stima di α_2 acquista un significato dalla sfumatura differente: più dell'80% degli uomini che muoiono per tumore muore prematuramente rispetto a tutti gli altri, ovvero in confronto ai decessi definiti "adulti" congiuntamente da tutte le cause messe insieme. Questa interpretazione ci sembra molto più sensata, la quale ammette e riflette tutte le varie considerazioni emerse più volte dalle nostre analisi, cioè che i tumori assumono una posizione di rilievo nella mortalità prematura ed inoltre che per tumore in media si muore effettivamente prima che per altre cause.

5.3 La nostra interpretazione

Giunti a questo punto delle analisi riteniamo opportuno provare a dare una risposta alla domanda iniziale che ha motivato e guidato fino a qui il nostro studio: poiché la durata media della vita si sta allungando, perché dal lavoro di Zanotto (2016) è emerso che in Francia negli ultimi anni si osserva un aumento del numero di decessi associati alla mortalità accidentale e prematura?

Alla ricerca di una possibile spiegazione, il nostro tentativo è stato quello di esplorare la forma e i cambiamenti delle distribuzioni dei decessi per età per diverse cause di morte. Dall'applicazione del modello gerarchico bayesiano proposto nel complesso la mortalità prematura non sembra essere aumentata

in modo particolarmente significativo per qualche causa specifica nel periodo analizzato dal 2000 al 2013. Infatti, per quanto riguarda le tre cause di morte più rilevanti, per le malattie del sistema digestivo e le cause esterne l'incidenza delle morti accidentali e premature si è addirittura abbassata, mentre per i tumori è stato riscontrato un innalzamento considerevole della percentuale di decessi associati a morti premature nel 2007, che poi è gradualmente diminuita nel tempo e risalita negli ultimi due anni. Nonostante ciò, le nostre analisi hanno contribuito a far luce su alcuni lati della questione che prima non erano chiari.

Le conclusioni a cui siamo arrivati sulla base dei risultati del modello elaborato per fornire una risposta alla domanda precedente si fondano su due aspetti principali. Il primo di questi fa riferimento al punto di partenza delle analisi, ovvero al grafico di Zanotto (2016) riportato nel Capitolo 1 di questa tesi nella Figura 1.5 in cui viene rappresentato l'aumento della percentuale dei decessi accidentali e prematuri maschili in Francia. Tale grafico comprende gli anni dal 1990 al 2013 e l'accelerazione della serie avviene dal 1990 a circa il 2006, per poi rallentare fino al 2013. Purtroppo i dati che abbiamo a disposizione sulla mortalità per causa estratti dallo *Human Cause-of-Death Database* per la Francia sono relativi solamente agli anni 2000-2013. Sospettiamo pertanto che questo sia un periodo forse troppo breve per riuscire a distinguere dei cambiamenti significativi, soprattutto perché, di fatto, gli anni rilevanti per l'obiettivo diventano solo quelli dal 2000 al 2006, quindi la metà di quelli con cui abbiamo lavorato finora, i quali corrispondono alla fase finale della crescita. Riteniamo potrebbe essere d'aiuto risalire alle serie di dati sulla mortalità per causa anche degli anni precedenti, almeno fino al 1990, in modo da poter avere una visione più generale e completa di quanto accaduto nell'intero periodo in cui si è verificato il fenomeno d'interesse. Sfortunatamente queste informazioni non sono reperibili dallo HCD, pertanto bisognerebbe ricorrere ad altre basi di dati ed effettuare un lavoro non banale per riuscire a ricostruire delle serie temporali continue per la mortalità specifica per causa e con cause di morte classificate

in base a liste omogenee e costanti. Questa tesi può quindi rappresentare un eventuale punto di partenza per ulteriori ricerche future in questa direzione.

L'altro aspetto da tenere in considerazione riguarda la corretta interpretazione di quel grafico. Poiché dalle nostre analisi non è emersa una percentuale di decessi prematuri cresciuta in modo significativo per qualche specifica causa di morte, ci siamo chiesti se ci sia stato un aumento vero e proprio della mortalità prematura in Francia nel ristretto periodo che ci è stato possibile analizzare. Crediamo che probabilmente la mortalità prematura in totale sia un po' aumentata, ma che forse ciò non sia da attribuire appunto all'incremento di un qualche tipo di mortalità per causa e sia dovuto semplicemente allo spostamento in avanti della mortalità adulta, il quale potrebbe aver fatto risaltare di più soprattutto i decessi per tumore. Come più volte è apparso dalle nostre analisi, negli ultimi anni è avvenuta non solo una traslazione verso destra della curva della mortalità adulta ma anche una sua compressione. Una possibile conseguenza di questo fenomeno è che passando da una situazione in cui c'era più variabilità, e dunque la curva dei decessi adulti era molto ampia, ad un'altra in cui la mortalità adulta si è spostata e "schiacciata", potrebbe essere emersa in maniera più evidente parte della mortalità prematura che era già presente anche prima. Per quanto riguarda i tumori, questi sono sicuramente la causa più responsabile dei decessi maschili sia prematuri che adulti ma, poiché la durata media della vita si è allungata, le morti che avvengono per la maggioranza delle cause si verificano in età molto avanzate e di conseguenza i decessi per tumore, che invece in media continuano a capitare in età precedenti, vengono isolati e risaltano di più come morti premature rispetto alle altre.

In conclusione, per il breve periodo temporale che abbiamo potuto analizzare, si suppone che l'aumento della mortalità prematura in Francia non dipenda dalla maggiore incidenza di una qualche causa di morte ma sia piuttosto il risultato del cambiamento di forma della distribuzione dei decessi per età avvenuto negli ultimi anni.

Capitolo 6

L'analisi della mortalità per causa nelle donne

Le analisi condotte finora si sono concentrate sulla mortalità maschile. Infatti, l'obiettivo principale di questa tesi è quello di dare una risposta alla questione lasciata aperta dal lavoro di Zanotto (2016) sui motivi dell'aumento delle morti premature in Francia. Poiché quest'ultimo è stato osservato per la popolazione maschile e inoltre dalle nostre analisi esplorative è emerso che la mortalità prematura è un fenomeno che colpisce principalmente gli uomini, abbiamo preferito in primo luogo stimare i vari modelli soltanto per i maschi. Una volta giunti a fornire una possibile interpretazione al problema iniziale sulla base del nuovo modello gerarchico elaborato, ci è sembrato interessante fare un ulteriore approfondimento e provare a replicare le stesse analisi anche per le donne, con lo scopo di capire come è variata la mortalità per causa nella popolazione femminile in Francia nello stesso periodo ed eventualmente fare dei confronti di genere. Pertanto, il modello mistura presentato nel Capitolo 3 è stato stimato marginalmente per ciascuna causa di morte ed in seguito abbiamo provato a costruire un modello gerarchico anche per la popolazione femminile. Tuttavia, in entrambi i casi i risultati non sono stati soddisfacenti; a differenza di quanto accaduto per gli uomini, per le donne anche il modello gerarchico sviluppato ha manifestato diversi problemi che non ci hanno permesso di trarre

le conclusioni sperate.

Nei prossimi paragrafi di questo capitolo saranno presentati brevemente alcuni dei risultati ottenuti applicando entrambi i metodi di stima. Prima di procedere, ricordiamo alcune considerazioni emerse dalle analisi descrittive del Capitolo 2 per le donne. La prima causa di morte femminile sono le malattie cardiache (causa 7), con un 26% nel 2000 e un graduale abbassamento al 23,3% nel 2013 che le porta a scendere al secondo posto a fine periodo, sorpassate dai tumori (causa 2). Quest'ultimi si trovano sempre in seconda posizione nella graduatoria delle principali cause di morte femminile con una percentuale del 22% rispetto al totale dei decessi, eccetto che per gli ultimi due anni; infatti, con una crescita progressiva durante tutti i 14 anni di studio, nel 2012 e 2013 i tumori diventano i maggiori responsabili anche delle morti femminili. La terza causa di morte per le donne nel 2000 sono le malattie cerebrovascolari (causa 8), mentre dal 2010 in poi lo stesso posto è occupato dalle malattie del sistema nervoso (causa 6). Già da queste statistiche iniziali era sembrato chiaro fin da subito che la mortalità femminile possiede caratteristiche differenti da quella maschile, intuizione supportata anche dai grafici delle Figure 2.2, 2.3, 2.4 e 2.5 che mostrano come variano le distribuzioni dei decessi per età marginalmente per ciascuna delle 15 cause nel 2013 confrontate tra i due sessi. Quest'ultimi hanno confermato che in generale le morti femminili sono sempre più concentrate e spostate verso età più avanzate rispetto a quelle degli uomini e soprattutto che l'incidenza della mortalità prematura per le donne è molto bassa, in certi casi addirittura quasi trascurabile. Dall'osservazione dalla forma delle distribuzioni di tali grafici sembra che le principali cause di mortalità accidentale e prematura siano anche per le donne i tumori, le malattie del sistema digestivo e le cause esterne, seppur con una prevalenza molto inferiore rispetto a quella riscontrata per gli uomini. Inoltre, si può notare che anche la distribuzione dei decessi femminili per tumore è mediamente spostata più indietro in confronto a quella delle altre cause. Per capire se queste ipotesi preliminari trovino un fondamento concreto nella pratica, siamo passati alla

fase di stima dei modelli.

6.1 La stima del modello mistura per le donne

Il modello mistura (3.17) presentato nel Capitolo 3 e adattato ai nostri scopi è stato stimato con il metodo della massima verosimiglianza separatamente per tutte le 15 cause di morte per i decessi femminili negli anni 2000-2013. Pur essendo ben consapevoli delle varie difficoltà già incontrate nell'inferenza per gli uomini, abbiamo voluto provare comunque a stimare questo modello anche per le donne. Come sospettavamo, non siamo giunti a conclusioni molto diverse. Infatti, non riportiamo i risultati ottenuti in quanto sono affetti dalle medesime problematiche che non rendevano attendibili le stime di massima verosimiglianza del modello per i maschi, già ampiamente descritte in precedenza. Ci limitiamo a commentare brevemente solo alcuni aspetti principali.

Per quanto riguarda la curva f_M della mortalità adulta, come accaduto per gli uomini le stime di massima verosimiglianza dei parametri sono state soddisfacenti e molto simili tra le varie cause di morte. In particolare, anche questa volta è stata riscontrata una traslazione verso destra annessa ad una compressione di tale curva nell'intero periodo. Infatti, le medie μ_M per le varie cause nel 2000 assumono valori tra circa 86 e 88 e poi aumentano nel tempo fino ad essere comprese tra 88 e 91 nel 2013, confermando che in media le donne vivono più a lungo. Le deviazioni standard σ_M si sono ridotte per quasi tutte le cause, mentre i coefficiente di asimmetria γ_M sono sempre negativi e rimangono abbastanza inalterati nel tempo. Nonostante ciò, anche per le donne il nostro interesse è rivolto soprattutto all'analisi della mortalità prematura tra le varie cause. Come già anticipato, anche in questo caso la stima di massima verosimiglianza del modello mistura è stata complessa, specialmente per quanto riguarda la prima componente, a causa delle stesse problematiche già incontrate e descritte nel paragrafo 3.3.1. Inoltre, i problemi di identificabilità in certe occasioni sono stati addirittura più frequenti, accentuati dal fatto che

la mortalità prematura nelle donne è un fenomeno molto più raro, pertanto per il modello mistura è stato ancora più complicato riuscire ad individuare in modo corretto la posizione e la forma della prima normale asimmetrica. Inoltre, proprio perché la distribuzione dei decessi femminili per la maggioranza delle cause è caratterizzata da un numero di morti premature molto basso, le curve ottenute per questa parte sono estremamente ampie e il coefficiente di asimmetria γ_m è stato stimato quasi sempre nei pressi della frontiera inferiore dello spazio parametrico (-0.995).

Sulla base di tutto ciò, per i motivi già spiegati nei capitoli precedenti, abbiamo valutato la possibilità di procedere anche in questo caso alla costruzione di un modello gerarchico bayesiano, con la speranza di ottenere delle stime più efficienti come accaduto per i dati relativi agli uomini.

6.2 Il modello gerarchico bayesiano per l'analisi della mortalità femminile

Dopo diversi tentativi e adattamenti, il modello gerarchico bayesiano sviluppato per studiare la mortalità per causa nelle donne è identico a quello presentato nel Capitolo 4 per gli uomini e specificato attraverso le equazioni dalla (4.3) alla (4.13). L'unica differenza riguarda il vincolo introdotto per evitare che si verifichino fenomeni di *label switching* e di collasso in un'unica componente della mistura: la distribuzione a priori normale (4.6) scelta per i parametri μ_{mj} è stata troncata ad assumere massimo valore 80 invece di 75, per riflettere il fatto che la distribuzione dei decessi per le varie cause per le donne è spostata mediamente più in avanti rispetto a quella degli uomini e dunque ammettere la possibilità di una mortalità prematura localizzata leggermente più a destra.

Anche per le donne il modello gerarchico è stato stimato con Stan per tutti gli anni dal 2000 al 2013, sfruttando il server "Calculus". Tuttavia, utilizzando

nelle simulazioni sempre 3 catene parallele di 5000 iterazioni ciascuna come in precedenza, questa volta ci sono state gravi difficoltà nella convergenza di tutte le catene alla comune distribuzione target. Lo stesso fenomeno si è verificando anche provando ad aumentare il numero di iterazioni. Le diagnostiche di convergenza hanno confermato l'inadeguatezza dei risultati; infatti, analizzando i *trace plot* è emerso che per alcune componenti le catene convergevano in posti differenti e manifestavano un'elevata autocorrelazione. Inoltre, per la maggior parte dei parametri i valori della statistica $\hat{R}$ erano molto superiori a 1 e quelli dell'n_{eff} estremamente bassi. Uno dei risultati che ha destato maggiori sospetti è stato che il coefficiente di asimmetria γ_M della curva della mortalità adulta comune per tutte le cause veniva stimato sempre pari a -0.995, ovvero sulla frontiera dello spazio parametrico e corrispondente ad un parametro di forma divergente. Questo fatto è sembrato molto strano, poiché tale problema si era già presentato nelle stime di massima verosimiglianza ma solo per quanto riguarda l'asimmetria della prima normale asimmetrica e mai per la seconda.

Dopo varie prove, siamo riusciti a raggiungere la convergenza nelle simulazioni introducendo un ulteriore vincolo nel modello gerarchico. In modo analogo a quanto effettuato per risolvere i problemi legati al parametro di asimmetria della componente della mortalità prematura, abbiamo provato a vincolare tra $(-0.8, 0.995)$ invece che tra $(-0.995, 0.995)$ anche la distribuzione a priori (4.13) per γ_M. Il risultato è stato che per tutti gli anni dal 2000 al 2013 il metodo di stima è arrivato a convergenza e le tecniche diagnostiche non hanno evidenziato particolari problemi. Le stime del modello gerarchico dal 2000 al 2003 sembrano essere soddisfacenti, ma non si può dire lo stesso per quelle degli anni successivi fino al 2013. Infatti, nonostante le diagnostiche di convergenza siano nel complesso molto buone, abbiamo notato che la stima di γ_M dal 2004 in poi è sempre pari al limite inferiore imposto, ovvero -0.8, e dunque il sospetto è che le stime della media e deviazione standard della medesima normale asimmetrica per la mortalità adulta si "aggiustino" di conseguenza. In particolare, la distribuzione a posteriori marginale di γ_M è quasi degenere in -0.8. A causa

di questo spiacevole inconveniente e dal momento che per tale parametro viene raggiunta di nuovo la frontiera dello spazio parametrico ammissibile, appurato che lo stesso problema continua a persistere sia che si tratti di -0.8 sia di -0.995, temiamo di non poterci fidare dei risultati ottenuti. Proprio per questo, anche se abbiamo a disposizione le stime del modello gerarchico per le donne per tutti gli anni, non le riportiamo e non riteniamo opportuno commentarle né fare confronti tra le cause o tra i due sessi. Il nostro interesse si concentra soprattutto sulla mortalità prematura ma, poiché stiamo lavorando con una mistura di due distribuzioni e ricordando che la forma e i cambiamenti della mortalità prematura dipendono fortemente da quelli della mortalità adulta, i valori stimati per i parametri della prima componente risentono dell'influenza negativa di quelli problematici che abbiamo ottenuto per la seconda. In sintesi, riteniamo che tutti questi aspetti critici ci impediscano di giungere a delle conclusioni attendibili.

Ragionando su una possibile interpretazione del problema, dal punto di vista teorico una distribuzione normale asimmetrica con coefficiente di asimmetria uguale a -0.995 nella parametrizzazione centrata corrisponde ad una funzione di densità con parametro di forma tendente a $-\infty$ nella parametrizzazione diretta e converge al caso limite di una semi-normale con massima asimmetria negativa. Alla ricerca di un'eventuale spiegazione del perché questo si verifichi nelle stime, un'ipotesi che abbiamo formulato è che tale problema non abbia origine dalla specificazione del modello, ma sia piuttosto legato ai dati che abbiamo utilizzato per stimare il modello gerarchico per le donne. Infatti, sempre dall'osservazione delle forme delle distribuzioni dei decessi per età delle varie cause di morte riportate nelle Figure 2.2, 2.3, 2.4 e 2.5, abbiamo notato che per le donne la maggioranza dei decessi si concentra soprattutto nelle ultime classi d'età; in particolare, diversamente dagli uomini, il numero di morti femminili nella classe 95+ per molte cause è estremamente elevato. Quest'ultimo fatto indica che non abbiamo alcuna informazione sulle età di morte per una consistente parte di osservazioni del nostro campione, ma sappiamo solo che

si trovano dopo i 95 anni. Se prima questo non era stato un problema per gli uomini, ora invece lo diventa dal momento che la mortalità femminile raggiunge l'apice di eventi soprattutto in quelle età, dunque la posizione della moda e quindi la forma della curva dei decessi nella parte finale a destra non risultano sempre riconoscibili né definite in maniera chiara. Inoltre, il modo in cui abbiamo "disaggregato" i dati originali attraverso la procedura descritta nel paragrafo 4.1.1 per creare le nuove osservazioni che entrano nel modello gerarchico e che rappresentano l'età approssimata a cui sono avvenuti i decessi, non è adatto per l'ultima classe aperta. In realtà, in questo caso diventa assolutamente riduttivo e poco preciso scegliere un unico valore come 98 (o qualsiasi altro) che sintetizzi l'età di morte dell'elevato ammontare di decessi femminili dopo i 95 anni. Tuttavia, in mancanza di ulteriori informazioni non è stato possibile redistribuire diversamente le morti che capitano in questo ampio intervallo finale.

Una delle implicazioni pratiche che comporta la tipologia di osservazioni con cui stiamo lavorando è che la curva che descrive la mortalità femminile adulta nel modello gerarchico sia caratterizzata da una forte asimmetria negativa, in quanto la forma complessiva della distribuzione dei decessi marginale per certe cause assomiglia effettivamente molto a quella di una semi-normale completamente asimmetrica a sinistra. Questo fatto è ulteriormente accentuato dalla scelta di effettuare un *pooling* di tutte le osservazioni per stimare i parametri della seconda componente della mistura nel modello gerarchico. Pertanto, poiché i dati che abbiamo non sempre permettono di vedere l'effettivo naturale declino della distribuzione dei decessi femminili dopo i 95 anni, il metodo di stima utilizzato fallisce perché non riesce ad identificare in modo preciso la reale forma della curva della mortalità adulta per le donne; di conseguenza, il parametro di asimmetria γ_M tende a valori sulla frontiera dello spazio parametrico, restituendo come stima il minimo valore ammissibile, più basso del dovuto, e determinando così una curva troppo ripida.

Sulla base di queste considerazioni, riteniamo che le stime che abbiamo

a disposizione per il modello gerarchico femminile siano da rivedere perché nutriamo il sospetto che lavorando con questi dati la parte di informazione mancante per le ultime età in cui la mortalità per le donne è ancora elevata abbia, di fatto, un peso rilevante e quindi non permetta di ottenere delle stime appropriate. Il nostro suggerimento per possibili sviluppi futuri in questo ambito è quello di recuperare da altre fonti delle serie di dati più accurate e complete per la mortalità femminile per causa, specialmente per quanto riguarda i decessi che accadono nella fase finale della vita, dato che le informazioni estratte dallo *Human Cause-of-Death Database* sono disponibili per classi d'età di ampiezza quinquennale e solo fino all'ultima classe aperta 95+. Inoltre, sottolineiamo ancora una volta che nel complesso la mortalità possiede caratteristiche molto diverse tra uomini e donne, in particolare per quest'ultime la mortalità prematura ha una bassa incidenza, dunque per analizzare i decessi femminili per causa potrebbe essere opportuno valutare la possibilità di elaborare un modello anche abbastanza diverso rispetto a quello costruito in precedenza per gli uomini.

Conclusioni

In questa tesi è stato proposto un nuovo modello gerarchico bayesiano per l'analisi delle cause di morte. In particolare, l'obiettivo principale è stato quello di fornire una possibile spiegazione all'aumento del numero di decessi associati alla mortalità prematura osservato in Francia negli ultimi anni (Zanotto, 2016). Per raggiungere tale scopo, sono state utilizzate delle serie di dati estratte dallo *Human Cause-of-Death Database* (HCD) relative alla mortalità per causa in Francia dal 2000 al 2013 e sono state intraprese due strade. In entrambe è stata analizzata la forma della distribuzione dei decessi per età e i cambiamenti nel tempo delle sue componenti per diverse cause di morte, focalizzando l'attenzione soprattutto sulla mortalità prematura. Nella prima strada è stato stimato, dopo opportuni adattamenti, il modello mistura proposto da Zanotto. Tuttavia, questo primo tentativo non è andato a buon fine a causa di numerosi problemi che hanno afflitto e reso poco affidabili le stime di massima verosimiglianza. La seconda strada ha portato allo sviluppo di un modello gerarchico bayesiano. Quest'ultimo si è dimostrato rappresentare una valida soluzione per risolvere tutte le difficoltà precedenti e ottenere delle stime più efficienti.

Sulla base dei risultati raggiunti attraverso il modello gerarchico elaborato, questo lavoro di tesi ha contribuito a far luce su diversi aspetti che prima non erano chiari e soprattutto ha permesso di trovare una possibile risposta alla domanda che ha motivato e guidato le analisi. Dall'applicazione del modello proposto sui dati che avevamo a disposizione è emerso che la mortalità prematura maschile è associata soprattutto ai tumori, alle malattie del sistema digestivo e alle cause esterne; nonostante ciò, nel complesso non sembra essere aumentata

in modo particolarmente significativo per qualche specifica causa negli anni considerati. È opportuno premettere che, poiché il fenomeno in questione era stato osservato in Francia principalmente tra gli anni 1990-2006 circa, il sospetto è che forse il periodo che ci è stato possibile analizzare dal 2000 al 2013 con i dati dello HCD sia troppo breve per riuscire a scorgere degli effetti rilevanti; pertanto, a tal fine potrebbe essere utile recuperare da altre fonti delle informazioni relative alla mortalità per causa anche per gli anni precedenti al 2000. Tuttavia, si ritiene che per quanto riguarda l'intervallo temporale oggetto del nostro studio l'aumento della mortalità prematura in Francia non sia dovuto alla maggiore incidenza di una qualche causa di morte ma sia piuttosto la conseguenza di un progressivo cambiamento di forma della distribuzione dei decessi avvenuto nel tempo, il quale, in seguito ad uno spostamento in avanti e ad una compressione della mortalità adulta, potrebbe aver isolato e fatto emergere in modo più evidente parte della mortalità prematura già presente e messo in risalto soprattutto i decessi causati dai tumori, che in media si verificano prima rispetto a quelli dovuti ad altre cause di morte.

Appendice A

Risultati del modello gerarchico per gli anni 2001-2006 e 2008-2012

Tabella A.1: Medie e deviazioni standard a posteriori dei parametri relativi alla mortalità adulta nel modello gerarchico stimato per gli anni 2000-2013.

	μ_M		σ_M		γ_M	
Anno	media	s.d.	media	s.d.	media	s.d.
2000	83.096	0.133	8.529	0.092	-0.562	0.014
2001	83.377	0.134	8.439	0.093	-0.545	0.015
2002	83.651	0.117	8.296	0.085	-0.541	0.015
2003	83.678	0.112	8.149	0.083	-0.524	0.015
2004	84.266	0.114	8.338	0.085	-0.554	0.015
2005	84.272	0.114	8.299	0.085	-0.570	0.014
2006	84.761	0.107	8.245	0.083	-0.579	0.014
2007	84.923	0.109	8.186	0.085	-0.579	0.015
2008	85.007	0.101	8.173	0.080	-0.590	0.014
2009	85.379	0.103	7.961	0.082	-0.566	0.015
2010	85.398	0.120	8.160	0.094	-0.603	0.015
2011	85.787	0.101	8.143	0.082	-0.646	0.013
2012	86.081	0.098	7.759	0.080	-0.596	0.015
2013	86.109	0.099	7.978	0.082	-0.653	0.013

Tabella A.2: Medie e deviazioni standard a posteriori degli iperparametri del modello gerarchico stimato per gli anni 2000-2013.

	σ_{μ_m}		σ_{γ_m}	
Anno	media	s.d.	media	s.d.
2000	2.476	0.023	0.198	0.002
2001	2.476	0.024	0.198	0.002
2002	2.476	0.022	0.198	0.002
2003	2.475	0.024	0.198	0.002
2004	2.477	0.022	0.198	0.002
2005	2.476	0.023	0.198	0.002
2006	2.475	0.024	0.198	0.002
2007	2.476	0.024	0.197	0.002
2008	2.476	0.023	0.197	0.002
2009	2.476	0.024	0.198	0.002
2010	2.473	0.026	0.197	0.003
2011	2.477	0.023	0.198	0.002
2012	2.478	0.021	0.198	0.002
2013	2.478	0.021	0.198	0.002

Tabella A.3: Medie e deviazioni standard a posteriori dei parametri di mistura e dei parametri relativi alla mortalità prematura nel modello gerarchico stimato per l'anno 2001.

	α_j		μ_{mj}		σ_{mj}		γ_{mj}	
Causa	media	s.d.	media	s.d.	media	s.d.	media	s.d.
1	0.358	0.023	57.499	1.274	17.280	0.636	-0.229	0.119
2	0.850	0.027	70.924	0.402	12.901	0.097	-0.675	0.013
3	0.219	0.034	58.679	2.216	19.141	0.748	-0.297	0.153
4	0.263	0.026	64.561	1.499	15.386	0.639	-0.655	0.081
5	0.221	0.010	51.103	0.646	10.807	0.482	-0.307	0.132
6	0.258	0.019	61.639	1.192	18.838	0.519	-0.785	0.017
7	0.203	0.015	64.619	0.914	13.283	0.344	-0.455	0.082
8	0.175	0.026	66.657	1.733	14.265	0.478	-0.699	0.099
9	0.254	0.028	66.439	1.327	12.534	0.556	-0.508	0.118
10	0.043	0.007	56.507	2.481	15.923	2.142	-0.028	0.187
11	0.214	0.025	68.183	1.231	13.121	0.499	-0.728	0.058
12	0.531	0.019	62.916	0.616	13.046	0.341	-0.175	0.066
13	0.061	0.014	59.645	2.266	16.709	1.821	-0.110	0.179
14	0.043	0.009	59.487	2.240	15.315	2.016	-0.047	0.181
16	0.578	0.008	44.305	0.434	19.304	0.329	0.449	0.052

Tabella A.4: Medie e deviazioni standard a posteriori dei parametri di mistura e dei parametri relativi alla mortalità prematura nel modello gerarchico stimato per l'anno 2002.

	α_j		μ_{mj}		σ_{mj}		γ_{mj}	
Causa	media	s.d.	media	s.d.	media	s.d.	media	s.d.
1	0.338	0.021	57.345	1.223	17.064	0.651	-0.165	0.106
2	0.817	0.025	70.472	0.398	12.773	0.107	-0.656	0.014
3	0.229	0.033	57.714	2.164	19.324	0.607	-0.356	0.144
4	0.265	0.025	64.736	1.440	15.602	0.618	-0.687	0.067
5	0.225	0.010	51.526	0.701	11.092	0.517	-0.239	0.114
6	0.249	0.018	61.939	1.207	18.613	0.531	-0.781	0.021
7	0.203	0.014	64.491	0.877	13.242	0.340	-0.425	0.070
8	0.203	0.023	67.913	1.340	13.739	0.432	-0.735	0.060
9	0.275	0.029	67.423	1.264	12.425	0.536	-0.542	0.118
10	0.039	0.006	55.430	2.485	15.025	2.290	-0.063	0.202
11	0.224	0.025	68.666	1.172	12.979	0.472	-0.740	0.051
12	0.524	0.017	62.628	0.577	12.693	0.335	-0.071	0.062
13	0.065	0.014	59.692	2.249	15.161	2.008	-0.051	0.180
14	0.051	0.010	59.924	2.125	13.550	1.800	-0.067	0.185
16	0.564	0.008	44.984	0.432	19.211	0.333	0.429	0.054

Tabella A.5: Medie e deviazioni standard a posteriori dei parametri di mistura e dei parametri relativi alla mortalità prematura nel modello gerarchico stimato per l'anno 2003.

	α_j		μ_{mj}		σ_{mj}		γ_{mj}	
Causa	media	s.d.	media	s.d.	media	s.d.	media	s.d.
1	0.352	0.026	58.446	1.431	17.806	0.668	-0.392	0.135
2	0.822	0.028	70.632	0.450	12.738	0.115	-0.661	0.015
3	0.210	0.031	57.840	2.135	19.162	0.712	-0.293	0.155
4	0.242	0.024	64.436	1.553	15.350	0.649	-0.591	0.091
5	0.227	0.010	51.929	0.684	11.119	0.525	-0.275	0.128
6	0.242	0.018	62.341	1.206	18.231	0.527	-0.781	0.020
7	0.205	0.013	64.341	0.841	13.055	0.345	-0.414	0.071
8	0.224	0.023	68.768	1.165	13.816	0.395	-0.768	0.033
9	0.316	0.034	68.443	1.248	12.138	0.463	-0.587	0.105
10	0.044	0.006	57.464	2.191	17.979	1.416	-0.128	0.151
11	0.198	0.021	67.850	1.201	12.396	0.528	-0.674	0.087
12	0.536	0.018	62.864	0.611	12.737	0.353	-0.173	0.063
13	0.075	0.016	60.334	2.215	15.545	1.880	-0.083	0.185
14	0.056	0.011	60.079	2.099	14.052	1.717	-0.121	0.181
16	0.550	0.008	45.580	0.464	19.005	0.347	0.298	0.059

Tabella A.6: Medie e deviazioni standard a posteriori dei parametri di mistura e dei parametri relativi alla mortalità prematura nel modello gerarchico stimato per l'anno 2004.

	α_j		μ_{mj}		σ_{mj}		γ_{mj}	
Causa	media	s.d.	media	s.d.	media	s.d.	media	s.d.
1	0.322	0.022	58.186	1.328	17.261	0.722	-0.198	0.108
2	0.808	0.027	71.189	0.436	12.844	0.111	-0.673	0.014
3	0.224	0.033	58.606	2.173	19.137	0.723	-0.347	0.150
4	0.264	0.025	65.500	1.389	14.933	0.586	-0.650	0.079
5	0.230	0.011	51.868	0.773	11.404	0.605	-0.194	0.144
6	0.206	0.018	61.595	1.507	18.891	0.541	-0.762	0.047
7	0.175	0.012	64.197	0.885	13.232	0.350	-0.391	0.074
8	0.199	0.022	68.526	1.247	13.833	0.450	-0.748	0.049
9	0.299	0.032	68.796	1.262	12.632	0.483	-0.658	0.098
10	0.035	0.006	56.156	2.329	15.200	2.120	0.025	0.197
11	0.200	0.024	68.874	1.287	12.810	0.489	-0.701	0.075
12	0.508	0.018	63.573	0.617	13.197	0.337	-0.212	0.063
13	0.067	0.014	59.688	2.211	17.124	1.713	-0.109	0.178
14	0.042	0.008	59.035	2.184	13.447	1.906	-0.042	0.182
16	0.539	0.009	45.792	0.474	18.881	0.355	0.323	0.060

Tabella A.7: Medie e deviazioni standard a posteriori dei parametri di mistura e dei parametri relativi alla mortalità prematura nel modello gerarchico stimato per l'anno 2005.

	α_j		μ_{mj}		σ_{mj}		γ_{mj}	
Causa	media	s.d.	media	s.d.	media	s.d.	media	s.d.
1	0.321	0.021	58.753	1.276	16.559	0.672	-0.230	0.107
2	0.818	0.025	71.335	0.399	12.814	0.102	-0.679	0.013
3	0.218	0.033	59.082	2.165	19.006	0.809	-0.321	0.150
4	0.279	0.026	65.840	1.390	14.937	0.593	-0.677	0.075
5	0.235	0.010	51.706	0.634	11.047	0.476	-0.313	0.128
6	0.189	0.017	61.485	1.538	18.904	0.543	-0.761	0.052
7	0.177	0.012	63.818	0.862	13.119	0.349	-0.346	0.071
8	0.189	0.026	67.879	1.676	14.030	0.452	-0.710	0.093
9	0.278	0.031	67.928	1.315	12.441	0.527	-0.528	0.129
10	0.038	0.006	57.067	2.177	17.780	1.544	-0.118	0.154
11	0.210	0.023	69.471	1.088	12.377	0.464	-0.734	0.058
12	0.500	0.018	63.511	0.637	13.017	0.351	-0.183	0.063
13	0.091	0.018	60.358	2.227	17.653	1.458	-0.242	0.165
14	0.041	0.008	59.257	2.160	12.584	1.797	-0.016	0.186
16	0.534	0.009	46.026	0.472	18.910	0.355	0.306	0.062

Tabella A.8: Medie e deviazioni standard a posteriori dei parametri di mistura e dei parametri relativi alla mortalità prematura nel modello gerarchico stimato per l'anno 2006.

	α_j		μ_{mj}		σ_{mj}		γ_{mj}	
Causa	media	s.d.	media	s.d.	media	s.d.	media	s.d.
1	0.300	0.020	58.847	1.250	16.348	0.675	-0.265	0.106
2	0.806	0.024	71.478	0.393	12.767	0.109	-0.674	0.015
3	0.229	0.033	59.203	2.184	18.974	0.850	-0.325	0.151
4	0.262	0.023	65.416	1.356	14.978	0.593	-0.656	0.080
5	0.225	0.010	51.606	0.665	10.981	0.519	-0.293	0.130
6	0.205	0.016	62.146	1.307	18.986	0.505	-0.779	0.026
7	0.161	0.010	62.803	0.918	12.958	0.404	-0.353	0.067
8	0.205	0.025	68.614	1.484	13.839	0.433	-0.723	0.087
9	0.315	0.032	69.445	1.176	12.251	0.468	-0.644	0.101
10	0.041	0.006	56.776	2.286	16.157	1.988	-0.105	0.174
11	0.189	0.021	68.565	1.216	13.095	0.539	-0.716	0.069
12	0.477	0.017	62.929	0.624	12.833	0.357	-0.169	0.060
13	0.080	0.017	60.433	2.227	17.205	1.649	-0.190	0.178
14	0.038	0.008	59.534	2.202	15.651	1.886	-0.101	0.178
16	0.532	0.009	47.092	0.474	18.841	0.338	0.188	0.061

Tabella A.9: Medie e deviazioni standard a posteriori dei parametri di mistura e dei parametri relativi alla mortalità prematura nel modello gerarchico stimato per l'anno 2008.

	α_j		μ_{mj}		σ_{mj}		γ_{mj}	
Causa	media	s.d.	media	s.d.	media	s.d.	media	s.d.
1	0.302	0.023	60.266	1.365	16.600	0.672	-0.365	0.125
2	0.846	0.028	72.487	0.421	12.802	0.100	-0.703	0.013
3	0.201	0.031	58.718	2.146	19.181	0.687	-0.346	0.151
4	0.257	0.023	65.890	1.324	14.779	0.573	-0.654	0.083
5	0.231	0.009	52.164	0.589	10.925	0.440	-0.420	0.122
6	0.164	0.016	61.808	1.659	18.806	0.552	-0.750	0.062
7	0.143	0.009	61.453	0.986	12.575	0.469	-0.384	0.080
8	0.215	0.022	69.122	1.213	13.821	0.430	-0.755	0.046
9	0.246	0.024	66.610	1.233	12.166	0.580	-0.370	0.116
10	0.047	0.006	56.786	2.139	16.995	1.678	-0.114	0.166
11	0.209	0.023	69.674	1.181	12.847	0.466	-0.728	0.063
12	0.492	0.017	63.775	0.606	12.984	0.334	-0.176	0.058
13	0.077	0.016	60.493	2.148	15.738	1.906	-0.138	0.182
14	0.046	0.008	60.062	2.042	14.743	1.752	-0.177	0.156
16	0.519	0.009	47.384	0.485	18.816	0.338	0.152	0.061

Tabella A.10: Medie e deviazioni standard a posteriori dei parametri di mistura e dei parametri relativi alla mortalità prematura nel modello gerarchico stimato per l'anno 2009.

	α_j		μ_{mj}		σ_{mj}		γ_{mj}	
Causa	media	s.d.	media	s.d.	media	s.d.	media	s.d.
1	0.299	0.024	61.181	1.400	16.594	0.685	-0.438	0.117
2	0.831	0.027	72.361	0.433	12.778	0.108	-0.699	0.013
3	0.225	0.031	58.519	2.101	19.181	0.698	-0.332	0.149
4	0.249	0.023	65.831	1.424	14.735	0.645	-0.613	0.092
5	0.237	0.011	53.060	0.793	11.794	0.628	-0.185	0.141
6	0.175	0.016	62.130	1.578	18.742	0.560	-0.755	0.055
7	0.159	0.010	62.238	0.967	12.531	0.458	-0.406	0.077
8	0.213	0.024	69.020	1.345	13.533	0.432	-0.715	0.077
9	0.293	0.031	68.588	1.309	12.615	0.527	-0.552	0.114
10	0.052	0.007	56.944	2.128	15.402	1.818	-0.268	0.165
11	0.222	0.024	70.170	1.207	12.629	0.446	-0.718	0.065
12	0.492	0.017	63.742	0.601	13.031	0.336	-0.170	0.061
13	0.078	0.015	60.873	2.111	16.631	1.695	-0.235	0.163
14	0.041	0.008	60.268	2.064	14.053	1.869	-0.194	0.166
16	0.519	0.008	47.436	0.478	18.695	0.334	0.131	0.057

Tabella A.11: Medie e deviazioni standard a posteriori dei parametri di mistura e dei parametri relativi alla mortalità prematura nel modello gerarchico stimato per l'anno 2010.

	α_j		μ_{mj}		σ_{mj}		γ_{mj}	
Causa	media	s.d.	media	s.d.	media	s.d.	media	s.d.
1	0.252	0.019	60.533	1.365	15.968	0.747	-0.365	0.105
2	0.819	0.024	72.356	0.387	12.772	0.102	-0.697	0.013
3	0.194	0.030	58.701	2.145	19.101	0.752	-0.314	0.152
4	0.233	0.020	65.230	1.300	14.274	0.609	-0.560	0.097
5	0.221	0.010	52.700	0.685	11.377	0.514	-0.359	0.133
6	0.162	0.017	61.724	1.830	18.893	0.552	-0.738	0.079
7	0.130	0.010	60.474	1.105	11.740	0.542	-0.391	0.096
8	0.169	0.029	66.510	2.126	13.964	0.499	-0.552	0.165
9	0.237	0.023	66.016	1.291	12.400	0.647	-0.351	0.114
10	0.046	0.007	58.264	2.166	15.279	1.866	-0.214	0.163
11	0.240	0.025	70.883	1.112	12.763	0.427	-0.752	0.044
12	0.486	0.017	63.889	0.618	12.999	0.341	-0.192	0.060
13	0.069	0.015	60.831	2.180	14.306	2.001	-0.102	0.180
14	0.035	0.007	59.694	2.059	12.350	1.824	-0.089	0.188
16	0.502	0.009	47.373	0.490	18.424	0.349	0.137	0.058

Tabella A.12: Medie e deviazioni standard a posteriori dei parametri di mistura e dei parametri relativi alla mortalità prematura nel modello gerarchico stimato per l'anno 2011.

	α_j		μ_{mj}		σ_{mj}		γ_{mj}	
Causa	media	s.d.	media	s.d.	media	s.d.	media	s.d.
1	0.251	0.021	62.118	1.395	15.699	0.677	-0.402	0.112
2	0.801	0.021	72.400	0.343	12.810	0.100	-0.711	0.014
3	0.193	0.029	58.869	2.123	18.661	1.016	-0.263	0.158
4	0.284	0.024	66.793	1.275	15.434	0.515	-0.716	0.064
5	0.176	0.008	52.987	0.573	9.954	0.428	-0.401	0.136
6	0.160	0.014	62.972	1.443	18.699	0.556	-0.769	0.036
7	0.135	0.009	61.031	1.028	12.039	0.504	-0.422	0.088
8	0.205	0.024	69.624	1.434	13.880	0.435	-0.739	0.072
9	0.280	0.027	67.587	1.217	12.216	0.581	-0.413	0.114
10	0.050	0.007	57.068	2.121	16.515	1.748	-0.185	0.164
11	0.221	0.021	70.235	1.064	12.800	0.434	-0.737	0.056
12	0.473	0.016	63.872	0.587	12.812	0.333	-0.120	0.061
13	0.085	0.018	61.298	2.152	14.659	1.860	-0.145	0.177
14	0.029	0.006	59.462	2.247	16.797	1.818	-0.131	0.175
16	0.512	0.009	47.990	0.477	18.497	0.329	0.110	0.056

Tabella A.13: Medie e deviazioni standard a posteriori dei parametri di mistura e dei parametri relativi alla mortalità prematura nel modello gerarchico stimato per l'anno 2012.

	α_j		μ_{mj}		σ_{mj}		γ_{mj}	
Causa	media	s.d.	media	s.d.	media	s.d.	media	s.d.
1	0.294	0.024	63.722	1.349	15.744	0.634	-0.549	0.111
2	0.834	0.024	72.877	0.385	12.712	0.104	-0.708	0.013
3	0.194	0.029	58.419	2.144	19.225	0.665	-0.358	0.150
4	0.282	0.023	66.229	1.314	15.017	0.566	-0.647	0.078
5	0.173	0.008	53.464	0.626	10.531	0.469	-0.388	0.129
6	0.189	0.015	64.718	1.204	17.744	0.521	-0.779	0.023
7	0.165	0.010	63.562	0.857	12.461	0.401	-0.360	0.068
8	0.222	0.025	69.651	1.395	13.316	0.418	-0.700	0.086
9	0.278	0.026	67.696	1.217	12.164	0.597	-0.433	0.119
10	0.039	0.006	57.619	2.098	14.864	1.919	-0.188	0.186
11	0.257	0.023	71.911	0.980	11.618	0.379	-0.708	0.072
12	0.506	0.017	64.684	0.572	12.614	0.328	-0.176	0.059
13	0.147	0.023	63.172	1.883	13.926	1.338	-0.267	0.165
14	0.042	0.008	59.956	2.097	13.438	1.821	-0.097	0.182
16	0.515	0.009	49.151	0.478	18.742	0.320	0.039	0.040

Appendice B

Codice Stan utilizzato per la stima del modello gerarchico

```
data {
int<lower=0> N; // numero di osservazioni
int<lower=1> J; // numero di gruppi
real y[N];
int causa[N];
}

parameters {
// iperparametri
real<lower=0, upper=2.5> sigma_mu_m;
real<lower=0, upper=0.2> sigma_gamma_m;

// parametri
real<lower=0,upper=0.90> alpha[J];
real mu_m_tilde[J];
real<lower=0, upper=20> sigma_m[J];
real<lower=-0.8,upper=0.995> gamma_m[J];
real mu_M_tilde;
real<lower=0, upper=9> sigma_M;
```

```
real<lower=-0.995,upper=0.995> gamma_M;
}

transformed parameters {
real<upper=75> mu_m[J];
real mu_M;
for(j in 1:J){
mu_m[j]=60+sigma_mu_m*mu_m_tilde[j];
}
mu_M=87+2*mu_M_tilde;
}

model {
// Riparametrizzazione (parametrizzazione centrata SN)

// variabili locali
real c_m[J];
real muz_m[J];
real csi_m[J];
real omega_m[J];
real lambda_m[J];

real c_M;
real muz_M;
real csi_M;
real omega_M;
real lambda_M;

for(j in 1:J){
if(gamma_m[j]<0){
c_m[j] = ((-1)*pow((2*(-gamma_m[j]))*inv(4-pi()), 0.3333333));
}
```

```
else{
c_m[j] = ((1)*pow((2*(gamma_m[j]))*inv(4-pi()), 0.3333333));
}

muz_m[j] = c_m[j]*inv_sqrt(1+square(c_m[j]));
lambda_m[j] = (muz_m[j]*sqrt(pi()*0.5))*inv_sqrt(1-square(muz_m[j])*pi()
    *0.5);
omega_m[j] = sigma_m[j]*inv_sqrt(1-square(muz_m[j]));
csi_m[j] = mu_m[j]-omega_m[j]*muz_m[j];
}

if(gamma_M<0){
c_M = ((-1)*pow((2*(-gamma_M))*inv(4-pi()), 0.3333333));
}
else{
c_M = ((1)*pow((2*(gamma_M))*inv(4-pi()), 0.3333333));
}

muz_M = c_M*inv_sqrt(1+square(c_M));
lambda_M = (muz_M*sqrt(pi()*0.5))*inv_sqrt(1-square(muz_M)*pi()*0.5);
omega_M = sigma_M*inv_sqrt(1-square(muz_M));
csi_M = mu_M-omega_M*muz_M;

// MODELLO GERARCHICO

// DISTRIBUZIONI A PRIORI
//iperparametri comuni
sigma_mu_m ~ uniform(0, 2.5);
sigma_gamma_m ~ uniform(0, 0.2);

//parametri che variano tra i gruppi
for(j in 1:J){
```

```
alpha[j] ~ uniform(0, 0.90);


//mu_m[j] ~ normal(60, sigma_mu_m) T[ , 75];
mu_m_tilde[j] ~ normal(0, 1);
sigma_m[j] ~ uniform(0, 20);
gamma_m[j] ~ normal(0, sigma_gamma_m) T[-0.8, 0.995];
}


// parametri comuni tra i gruppi (seconda SN)
//mu_M ~ normal(87, 2);
mu_M_tilde ~ normal(0, 1);
sigma_M ~ uniform(0, 9);
gamma_M ~ skew_normal(-1, 0.5, 1) T[-0.995, 0.995];


// MODELLO MISTURA DI DUE SN PER I DATI in y
for (n in 1:N) {
target += log_sum_exp( log(alpha[causa[n]])
             + skew_normal_lpdf(y[n] | csi_m[causa[n]], omega_m[causa[n]],
                 lambda_m[causa[n]]),
                       log1m(alpha[causa[n]])
             + skew_normal_lpdf(y[n] | csi_M, omega_M, lambda_M) );
}


}
```

Bibliografia

Arellano-Valle, R. B. e A. Azzalini (2008). The centred parametrization for the multivariate skew-normal distribution. *Journal of Multivariate Analysis*, 99 (7), 1362–1382.

Azzalini, A. (1985). A class of distributions which includes the normal ones. *Scandinavian journal of statistics*, 171–178.

Azzalini, A. e R. B. Arellano-Valle (2013). Maximum penalized likelihood estimation for skew-normal and skew-t distributions. *Journal of Statistical Planning and Inference*, 143 (2), 419–433.

Azzalini, A. e A. Capitanio (1999). Statistical applications of the multivariate skew normal distribution. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61 (3), 579–602.

Azzalini, A. e A. Capitanio (2014). *The skew-normal and related families. Institute of Mathematical Statistics Monographs*.

Bayes, C. L. e M. D'E. Branco (2007). Bayesian inference for the skewness parameter of the scalar skew-normal distribution. *Brazilian Journal of Probability and Statistics*, 141–163.

Betancourt, M. e M. Girolami (2015). Hamiltonian Monte Carlo for hierarchical models. *Current trends in Bayesian methodology with applications*, 79, 30.

Carpenter, B., Gelman A., Hoffman M. D. et al. (2017). Stan: A Probabilistic Programming Language. *Journal of Statistical Software, Articles*, 76 (1), 1–32. ISSN: 1548-7660.

Duane, S., A. D. Kennedy, B. J. Pendleton e D. Roweth (1987). Hybrid Monte Carlo. *Physics letters B*, 195 (2), 216–222.

Gelman, A. e J. Hill (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Vol. 1. Cambridge University Press New York, NY, USA.

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

Gompertz, B. (1825). On the nature of the function expressive of the law of human mortality, and on a new mode of determining the value of life contingencies. *Philosophical transactions of the Royal Society of London*, 115, 513–583.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57 (1), 97–109.

Heligman, L. e J. H. Pollard (1980). The age pattern of mortality. *Journal of the Institute of Actuaries*, 107, 49–80.

Hoffman, M. D. e A. Gelman (2014). The No-U-turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15 (1), 1593–1623.

Human Cause-of-Death Database (2017). French Institute for Demographic Studies (France) and Max Planck Institute for Demographic Research (Germany). URL: `http://www.causeofdeath.org`.

Human Mortality Database (2017). University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany). URL: `http://www.mortality.org`.

Istat (2014). *Le principali cause di morte in Italia*. URL: `https://www.istat.it`.

Istat (2017). *L'evoluzione della mortalità per causa: le prime 25 cause di morte*. URL: `https://www.istat.it`.

Lexis, W. H. R. A. (1879). *Sur la durée normale de la vie humaine et sur la théorie de la stabilité des rapports statistiques*. Vve. F. Henry.

Liseo, B. e N. Loperfido (2006). A note on reference priors for the scalar skew-normal distribution. *Journal of Statistical Planning and Inference*, 136 (2), 373–389.

Livi Bacci, M. (1983). *Introduzione alla demografia*. Loescher Editore.

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller e E. Teller (1953). Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21 (6), 1087–1092.

Neal, R. M. (1994). An improved acceptance procedure for the hybrid Monte Carlo algorithm. *Journal of Computational Physics*, 111 (1), 194–203.

Neal, R. M. et al. (2011). MCMC using Hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, 2 (11).

Pearson, K. (1897). *Chances of Death, and Other Studies in Evolution*. Vol. 1. CUP Archive.

Preston, S. H., P. Heuveline e M. Guillot (2001). *Demography: Measuring and Modeling Population Processes*. Wiley-Blackwell.

Redner, R. A. e H. F. Walker (1984). Mixture densities, maximum likelihood and the EM algorithm. *SIAM review*, 26 (2), 195–239.

Sartori, N. (2006). Bias prevention of maximum likelihood estimates for scalar skew normal and skew t distributions. *Journal of Statistical Planning and Inference*, 136 (12), 4259–4275.

Stan Development Team (2017a). *RStan: the R interface to Stan*. R package version 2.16.2. URL: `http://mc-stan.org`.

Stan Development Team (2017b). *Stan Modeling Language Users Guide and Reference Manual*. Version 2.16.0. URL: `http://mc-stan.org`.

Zacks, S. (1981). *Parametric Statistical Inference*. Pergamon Press, Oxford.

Zanotto, L. (2016). A mixture model to distinguish mortality components. Tesi di dottorato. Università degli Studi di Padova.

Ringraziamenti

Arrivata alla fine di questa tesi e soprattutto di questo percorso universitario, credo sia giunta l'ora di dire Grazie a chi, a suo modo, mi ha accompagnata fino a qui.

Grazie al Professor Mazzuco per l'aiuto che mi ha dato e il tempo che mi ha dedicato durante questo lavoro, per essere stato sempre disponibile e paziente con me, per tutto quello che ho avuto la possibilità e il piacere di imparare in questi mesi e per avermi ricordato nei momenti più bui che a volte "i problemi possono diventare opportunità!".

Grazie a Lucia per la sua gentilezza e i suoi preziosi consigli, e perché senza il suo lavoro di dottorato questa tesi non sarebbe esistita.

Grazie ai miei genitori per avermi sostenuto economicamente e soprattutto moralmente in tutti questi anni di studio, per aver gioito con me dei miei successi e per essere stati comprensivi quando mi sono trovata in difficoltà, senza farmi pesare mai nulla. Grazie Mamma per l'esempio di Donna che sei per me, per dimostrarmi ogni giorno la tua forza e il tuo coraggio nell'affrontare tutte le sfide della vita.

Grazie a Giulia, mia sorella gemella, che nonostante le nostre diversità c'è sempre e comunque. Grazie per il nostro legame speciale, so che qualunque direzione prenderanno le nostre vite potremo sempre contare l'una sull'altra.

Grazie ai miei nonni Maria e Fiorenzo per farmi sentire sempre tanto amata. Grazie anche per tutte le preghiere recitate prima di ogni mio esame!

Grazie ad Angela, Davide e Fede, miei compagni d'avventura dal primo all'ultimo giorno, per aver condiviso con me tutti i momenti più importanti, sia

belli che brutti, di questo percorso universitario e non solo. Grazie ad Angela, per essermi stata vicina e avermi saputo capire forse più di chiunque altro in questi anni, e che da essere una semplice compagna di studi è diventata una delle migliori amiche che io potessi desiderare. Grazie a Davide, che con la sua spensieratezza mi mostra che nella vita c'è anche altro oltre allo studio e al lavoro. Grazie a Fede, fonte di saggezza, per essere sempre stato un mio grande punto di riferimento. Grazie mille amici, siete stati e siete tuttora fondamentali per me, senza di voi questi anni non sarebbero stati la stessa cosa.

Grazie anche a tutti gli altri compagni e amici che hanno reso più piacevole il tempo passato in "facoltà" e quello lontano dai libri.

Grazie agli insegnanti incontrati durante i miei studi che sono riusciti a trasmettermi la passione per l'affascinante mondo della Statistica, facendomi scoprire e apprezzare sempre nuovi particolari.

Grazie a tutte le persone che mi hanno sopportato, incoraggiato e mi son state accanto durante questi lunghi anni e che sanno che cosa significhi davvero per me aver raggiunto questo traguardo. Grazie a chi ha creduto in me molto più di quanto spesso abbia fatto io.

Infine, a costo di andare controcorrente, un ultimo Grazie va alla parte più testarda e determinata di me, che nonostante tutti i miei dubbi, insicurezze e difficoltà in fondo non ha mai smesso di credere che prima o poi questo momento sarebbe arrivato. Grazie alla forza e alla motivazione che pian piano sono riuscita a trovare fuori e dentro di me che mi hanno spinto a continuare a scalare, anche se a piccoli passi, questa montagna che tante, troppe volte mi è sembrata così insormontabile, ed ora che finalmente ce l'ho fatta ad arrivare in cima mi stanno concedendo la possibilità di guardare al futuro con curiosità e speranza e al passato con nuovi occhi, felice e orgogliosa di tutto questo.

"È il tempo che hai perduto per la tua rosa
che ha fatto la tua rosa così importante."
– Antoine de Saint-Exupéry