POLITECNICO DI MILANO

Scuola di Ingegneria dei Sistemi

Corso di Laurea in INGEGNERIA MATEMATICA

Tesi di Laurea Specialistica

Modelli statistici per la gestione dei rischi operativi

Relatore:
Prof. Anna Maria PAGANONI

Candidato:
Stefano ZILLER Matr. 745864

Anno Accademico 2010-2011

«Sembra difficile dare una sbirciata alle carte di Dio.
Ma che Egli giochi a dadi e usi metodi telepatici
è qualcosa a cui non posso credere nemmeno per un attimo»

Albert Einstein

Ringraziamenti

Desidero innanzitutto ringraziare la Professoressa Anna Maria Paganoni, che non solo mi ha saggiamente guidato e incoraggiato durante i lavori di tesi triennale e specialistica, ma è stata un punto di riferimento durante questi cinque anni di università. Un grazie anche perché mi ha introdotto e spinto verso la statistica, trasmettendomi la stessa passione che mostra nell'insegnamento. Grazie anche a Francesca Ieva, che prima da correlatrice e poi da amica mi è stata vicina e mi ha consigliato in questi ultimi anni di università.

Un ringraziamento particolare per quanto riguarda non solo questa tesi, ma probabilmente anche il mio futuro lavorativo, va a Vittorio Vecchione e Cinzia Fumagalli, che mi hanno fatto conoscere il fantastico mondo dei rischi operativi. Un grazie anche ad Alessia Treglia, che non riesce a far partire R senza chiamarmi, ma mi ha rallegrato giornate altrimenti interminabili e ora c'è sempre quando ho bisogno di consigli.

Tante sono le persone che mi sono state vicine in questi anni, ma prima di tutti ci sono loro, gli amici di sempre, con cui fa sempre bene bere una "birretta easy" o fare un comodo viaggio di 6000km in due settimane. Quindi grazie Giulia, per la presenza costante e le infinite serate, nonostante il muro che fa continuamente con la sua eterna indecisione, Petra, per la sua dolcissima ingenuità e per averci fornito innumerabili aneddoti da raccontare, Burly, perché dopo "si, va bene" un "grazie" ci sta benissimo, e Massi, Pino, Pò, Laura, Michi, Ale, Paola, che mi permettono di non scervellarmi per trovare una frase simpatica per ognuno, ma sapete quanto contate per me!

In questi cinque anni ogni tanto mi è capitato di passare anche anche dal Poli, dove per fortuna ho conosciuto persone fantastiche, senza le quali non sarei mai sopravvissuto a cinque anni di ingegneria: grazie Checco, Noemi, Trilli, Mirco, Andrea, il Maestro, Tesio, Paolino, Novi, Stasi, Obelix e tutti gli altri!! Un grazie in particolare va a Nico, perché mi ha sopportato durante tutta la stesura della tesi e soprattutto perché appoggia sempre le mie proposte per festeggiare e/o dimenticare.

Grazie di cuore anche agli ASPer, per le notti insonni, le mitiche serate al Tabata e ogni tanto anche qualche conversazione impegnata. Come non citare in particolare Ambra e Serena, per avermi fatto scoprire le mie doti da designer, e Ste e Fra, senza i quali non sarei mai sopravvissuto al progetto.

Non posso inoltre non ringraziare tutti i miei compagni di squadra, passati e presenti, per avermi tenuto compagnia (ma soprattutto sopportato) quattro sere a settimana negli ultimi dodici anni. In particolare vorrei ringraziare Lollo, Diego, Natz, Estey e relative donne per avere allietato innumerevoli serate al kebabbaro e al focacciaro.

Un ringraziamento dovuto, ma che non sarà mai sufficiente per quello che fanno e hanno fatto per me, va ai miei genitori, mio fratello Luca e le mie nonne, che mi hanno sempre sostenuto moralmente, economicamente e psicologicamente.

Tante altre sono le persone da ringraziare, troppi i motivi, ma poco lo spazio a disposizione, quindi a tutti quelli che non sono stati nominati, ma sanno di meritarselo... GRAZIE!!!

Indice

Sommario

Le banche, e in generale tutte le istituzioni finanziare, sono obbligate ad accantonare quote di capitale per fare fronte ai rischi operativi. Tali rischi includono tutte le perdite derivanti da disastri naturali, guasti al sistema, errori umani o frodi. Questa tesi si concentra su un modello statistico, basato sull'analisi di serie storiche, appartenente alla categoria degli AMA (Advanced Measurement Approach) che ha come obiettivo finale il calcolo del capitale a rischio.

Gli eventi operativi vengono innanzitutto suddivisi in sette ET (Event Types), a seconda della causa. L'idea è quella di modellizzare separatamente la distribuzione di perdita di ogni singolo ET, per poi aggregarle ed ottenere un'unica distribuzione. A partire da essa è quindi possibile calcolare il capitale da accantonare: esso viene determinato con il VaR (Value at Risk), che è infatti definito come il quantile di livello 99.9% della distribuzione aggregata. L'approccio che si utilizza per modellizzare la singola classe di rischio è di tipo attuariale: si differenzia la probabilità di accadimento degli eventi operativi (distribuzione di frequency) e l'impatto economico del singolo evento (distribuzione di severity), per poi ottenere una distribuzione aggregata dopo convoluzione delle due.

Un primo problema sorge in quanto perdite di piccolo importo spesso non vengono contabilizzate, o comunque non vengono considerate affidabili. Per tale motivo, per la modellizzazione della severity, si utilizzano le distribuzioni troncate al di sotto di una data soglia, determinata a priori dalla banca. Inoltre, a causa della sensibilità del capitale a rischio rispetto ai quantili elevati, la coda destra della distribuzione, ovvero i valori oltre ad una soglia da determinare, vengono modellizzati con la GDP (Generalized Pareto Distribution), che risulta la più adatta nell'ambito della teoria dei valori estremi.

La distribuzione di frequency, invece, viene modellizzata con una Poisson, tenendo conto nella stima della frequenza delle sole perdite al di sopra della soglia di raccolta. L'approccio attuariale prevede quindi la convoluzione tra frequency e severity per ottenere la distribuzione ag-

gregata annua per classe di ET: ciò è portato avanti, sotto l'ipotesi di indipendenza, tramite simulazione Monte Carlo. Per quanto riguarda infine l'aggregazione tra le diverse classi di rischio, per la determinazione della distribuzione multivariata su cui si andrà a calcolare il VaR, si utilizzano le copule, particolari distribuzioni multivariate che permettono di aggregare le distribuzioni marginali mantenendo la struttura di correlazione desiderata.

L'elaborato è suddiviso in cinque Capitoli. Nel Capitolo 1 vengono introdotti i rischi operativi e le relative normative di Banca d'Italia, descrivendo poi il modello in considerazione e presentando i vari passi dell'analisi che vengono poi discussi nei Capitoli successivi. Il Capitolo 2 è invece focalizzato sulla trattazione teorica delle distribuzioni troncate, presentando i casi particolari di lognormale e Weibull, introducendone le principali proprietà e proponendo opportuni algoritmi di stima dei parametri. Nel Capitolo 3 si presenta la teoria dei valori estremi, focalizzando l'attenzione sulla scelta della soglia e la stima dei parametri della GPD. Il Capitolo 4 tratta la modellizzazione delle distribuzioni multivariate con le copule: in particulare, dopo la presentazione teorica, ci si concentra sulla struttura di dipendenza e si confrontano le principali famiglie di copule (ellittiche e Archimedee). Nel Capitolo 5, infine, vengono presentati i risultati ottenuti applicando il modello ai dati di una delle maggiori banche italiane, che rimarrà anonima a causa della sensibilità dei dati, per arrivare alla calcolo del capitale a rischio.

Il risultato principale della tesi è dato non solo dallo studio e l'integrazione di avanzate tecniche statistiche in un unico modello, ma soprattutto dall'innovazione in alcuni aspetti critici dello stesso, tra cui la determinazione degli stimatori MLE per particolari distribuzioni troncate e i metodi di scelta della soglia dei valori estremi.

Abstract

This work presents a statistical model for operational risk management. Such risk includes losses deriving from natural disaters, system failures, human errors or frauds. All financial institutions have to set a provision up, in order to face such losses. The thesis is focused on AMA (Advanced Measurement Approach) models, aimed at computing the capital at risk. These statistical models are based on the analysis of operational losses time series.

First of all, seven operational ET (Event Types) can be distinguished, according to the different causes. The idea of the model is to fit each risk class separately and then aggregate them to obtain a single distribution. Hence, the provision can be computed through the VaR (Value at Risk) indicator, defined as the 99.9% quantile of the aggregated distribution. The approach proposed is an actuarial one: the probability of event occurrence (the frequency distribution) and the economic impact of the single event (the severity distribution) are treated separately, and then an aggregated distribution is obtained through convolution of frequency and severity, for each ET.

A first problem arises since losses with a small economical impact are often neglected, hence they can rarely be trusted. Thus, the severity distribution is fitted with truncated distribution, above a threshold, which is fixed by the bank. Moreover, due to the sensibility of the capital at risk with respect to high level quantiles, the right tail of the severity distribution, which includes losses above a certain threshold, which has to be estimated, is fitted with the GPD (Generalized Pareto Distribution), which is the most appropriate in extreme values theory.

On the other hand, the frequency distribution is modeled with Poisson distribution, considering only losses above the lower threshold for the estimation. Thus, according to the actuarial approach, each ET aggregated annual loss distribution is obtained through convolution, , via Monte Carlo simulation, under the appropriate independence hypothesis. Finally, the ETs multivariate distribution, which the VaR is computed

on, is obtained exploiting copulas, which allows to aggregate marginal distributions maintaining the desired dependence structure.

The work is divided into five main Chapters. In Chapter 1, operational risk and Bank of Italy main regulations are introduced, then the considered model is described, highlighting the main steps of the analysis, which are deeply discussed in the next few Chapters. Chapter 2 presents truncated distribution, focusing on lognormal and Weibull distributions, presenting the relative properties and proposing algorithms for parameters estimate. In Chapter 3 the theory of extreme values is presented, with particular attention devoted to the threshold selection and the estimation of GPD parameters. Chapter 4 is devoted to the modeling of multivariate distributions with copulas: after a brief theoretical discussion, the attention is set on dependence structures and on the comparison of the main copulas families (elliptical and Archimedean). Finally, in Chapter 5, we present the results obtained by applying the model to the dataset of one of the major Italian banks, anonymous due to the sensitivity of the data. The conclusion is the determination of the capital at risk for that bank.

The fundamental result of this thesis is not only the study and the integration of advanced statistical techniques in a single model, but also the innovation in some of its critical aspects, including the determination of MLEs for particular truncated distributions and methods to choose the correct threshold for extreme values.

Capitolo 1

Introduzione

L'obiettivo di questo elaborato è quello di presentare un modello statistico per la previsione del capitale da accantonare in una banca per far fronte alle perdite operative. In particolare, dopo un'introduzione ai rischi operativi, si passerà alla presentazione del modello stesso, per poi approfondire dal punto di vista teorico le principali tecniche statistiche adottate.

1.1 I rischi operativi

Il *rischio operativo* viene definito come "il rischio di perdite derivanti dalla inadeguatezza o dalla disfunzione di procedure, risorse umane e sistemi interni, oppure da eventi esogeni; è compreso il rischio legale" [1]. La gestione e la prevenzione di tale tipologia di rischio è regolata da Banca d'Italia, che ha disposto norme per la raccolta delle perdite operative e per la determinazione del capitale a rischio.

La caratteristica principale dei rischi operativi è che, salvo introduzione di mitigazione e assicurazione, non possono essere evitati dalla banca: si differenziano dalle altre tipologie di rischio cui sono soggette banche e assicurazioni proprio perché non sono conseguenza di operazioni appunto "rischiose", quindi volte ad ottenere profitti maggiori, ma derivano dalla normale gestione dell'operatività.

I rischi operativi sono costituiti da una causa (ad es, mancata revisione degli impianti di sicurezza), un evento operativo (ad es, incendio) e una serie di effetti (ad es, perdite monetarie derivanti dall'evento). Essi vengono suddivisi, da normativa, in 7 classi ("Event Type", ET), a seconda della causa dell'evento operativo:

1. **Frode interna**: perdite dovute ad attività non autorizzata, frode, appropriazione indebita o violazione di leggi, regolamenti o direttive aziendali che coinvolgano almeno una risorsa interna della banca.

2. **Frode esterna:** perdite dovute a frode, appropriazione indebita o violazione di leggi da parte di soggetti esterni alla banca.

3. **Rapporto di impiego e sicurezza sul lavoro:** perdite derivanti da atti non conformi alle leggi o agli accordi in materia di impiego, salute e sicurezza sul lavoro, dal pagamento di risarcimenti a titolo di lesioni personali o da episodi di discriminazione o di mancata applicazione di condizioni paritarie.

4. **Clientela, prodotti e prassi professionali:** perdite derivanti da inadempienze relative a obblighi professionali verso clienti ovvero dalla natura o dalle caratteristiche del prodotto o del servizio prestato.

5. **Danni da eventi esterni:** Perdite derivanti da eventi esterni, quali catastrofi naturali, terrorismo, atti vandalici.

6. **Interruzioni dell'operatività e disfunzioni dei sistemi:** perdite dovute a interruzioni dell'operatività, a disfunzioni o a indisponibilità dei sistemi.

7. **Esecuzione, consegna e gestione dei processi:** perdite dovute a carenze nel perfezionamento delle operazioni o nella gestione dei processi, nonché alle relazioni con controparti commerciali, venditori e fornitori.

1.2 Metodi avanzati per i rischi operativi

Banca d'Italia propone diversi metodi per la quantificazione di questa tipologia di rischi: base, standardizzato e avanzato. Tali approcci, sebbene finalizzati tutti alla definizione del capitale regolamentare da allocare a fronte di tali rischi, presentano un grado crescente di complessità. Le banche possono adottare il metodo più adatto alle proprie dimensioni e contesto operativo.

Nel metodo base, il requisito patrimoniale è calcolato semplicemente applicando un coefficiente al margine di intermediazione, un indicatore della grandezza del business della banca. Nel metodo standardizzato, il capitale da allocare si determina applicando al margine di intermediazione differenti coefficienti per ogni linea di business in cui è suddivisa

l'attività aziendale. Nei metodi avanzati (o AMA, *Advanced Measurement Approach*), l'ammontare del requisito patrimoniale viene calcolato dalla banca attraverso modelli basati su dati di perdita operativa ed altri elementi di valutazione.

L'approccio più interessante dal punto di vista statistico, e che sarà oggetto di questo elaborato, è proprio quello dei modelli AMA. Essi offrono il vantaggio di una più puntuale misurazione dell'esposizione al rischio operativo, in quanto costruiti ad hoc per la singola banca. Gli AMA raccolgono infatti un'ampia gamma di metodologie, caratterizzate da un elevato grado di rigore statistico e sensibilità verso il rischio, basati appunto sulla stima delle perdite operative attraverso l'analisi di serie storiche.

Per ottenere l'autorizzazione di Banca d'Italia ad adottare metodi AMA, le banche, oltre a dover dimostrare un adeguato sistema di raccolta delle perdite operative e organi di sviluppo e validazione del modello interno di calcolo dei rischi, devono avere una dimensione minima del business, basata sul *patrimonio di vigilanza* [1].

Più in particolare, per costruire un adeguato modello, la banca deve tenere in considerazione dati interni di perdita operativa, dati esterni ed analisi di scenario.

I dati interni di perdita operativa costituiscono base per la costruzione di sistema di misurazione dei rischi operativi. La banca solitamente definisce delle soglie minime di perdita, al di sotto delle quali non vengono considerate le perdite. Tali soglie, che possono essere definite differentemente per ogni classe di rischio, non devono comportare l'esclusione di significativi eventi operativi e, allo stesso tempo, non devono creare stime distorte, includendo dati non affidabili. Come vedremo, l'esclusione di tali eventi rende necessario l'utilizzo di tecniche statistiche adeguate, ovvero per distribuzioni troncate, come vedremo nel Capitolo 2.

Al fine di garantire la robustezza della stima del capitale, deve essere utilizzato un dataset di perdite interne provenienti da un periodo storico di una durata minima di cinque anni. Spesso, tuttavia, i dati interni non sono sufficientemente numerosi, soprattutto per perdite di grande importo e bassa frequenza. Per questo motivo è necessario integrare tale dataset con dati esterni e di scenario.

Le principali fonti di dati esterni di perdite operative sono di natura consortile, ovvero fornite da un insieme di banche e altri intermediari finanziari, come il DIPO (Database Italiano Perdite Operative), o di mercato, cioè archivi acquisiti da fornitori del settore, oppure elaborati internamente alla banca basandosi su giornali specializzati.

I dati di scenario, ovvero corrispondenti al verificarsi di plausibili eventi operativi "estremi", ovvero di grande impatto e bassa frequenza, sono invece essere generati da un adeguato processo interno, che deve tener conto della grandezza del business della banca e del contesto socioeconomico in cui si trova. Tale processo deve coinvolgere esperti, interni o esterni, e includere un confronto con gli altri dati a disposizione per valutarne l'affidabilità.

Al fine di garantire la robustezza delle stime basate sui dati esterni e di scenario, quindi non direttamente legati alla banca, tali osservazioni vanno incorporati nel modello con di misurazione con adeguate metodologie (ad esempio, procedure di riscalamento), come vedremo nel seguito.

I dati che si utilizzeranno per la costruzione del modello per la gestione dei rischi operativi, per ogni classe di rischio, sono:

- Perdite interne, riscontrate in seguito ad eventi relativi alla banca in esame.
- Perdite esterne superiori ad una determinata soglia, come verrà spiegato più avanti, provenienti dal database DIPO e da dati pubblici.
- Perdite prospettiche derivate da analisi di scenario (superiori alla medesima soglia).

Il metodo di quantificazione dei rischi operativi che viene descritto nel seguito si adotta separatamente per ciascuna classe di rischio. L'approccio che si utilizza nella modellizzazione delle perdite è quello attuariale: il concetto fondamentale è infatti l'analisi separata di frequency (frequenza di accadimento di un particolare evento) e severity (distribuzione dell'impatto monetario di un singolo evento operativo). Il capitale a rischio totale viene poi quantificato aggregando le diverse classi attraverso una copula, che utilizza la struttura di correlazione empirica tra le diverse classi per aggregarle. Il capitale da allocare viene definito secondo alcune statistiche sulla distribuzione aggregata degli ET, come vedremo nel seguito.

1.3 La distribuzione di severity

Come accennato, la distribuzione di severity rappresenta la densità di probabilità dell'impatto monetario derivante da un singolo evento operativo.

Essa viene divisa in due parti: il corpo viene modellizzato sulla base delle perdite interne alla banca, mentre la coda superiore, ovvero le perdite

superiori alla soglia che verrà indicata con u, viene modellizzata in modo differente per dare più peso alle perdite estreme (in ottica conservativa), sfruttando quindi anche i dati di scenario e le perdite delle altre banche. Questo tipo di approccio è molto comune nel risk management, in quanto, molto spesso, le distribuzioni tradizionalmente usate per la modellizzazione di perdite (lognormale e Weibull) sottostimano le perdite nella coda destra della distribuzione, portando ad una conseguente sottostima del capitale totale.

Il valore u, oltre al quale si considerano gli importi appartenenti alla coda, viene fissato dal management della banca. Questo è un punto critico, in quanto tutte le analisi di bontà del fit, ad esempio, sono influenzate molto da questa scelta di carattere qualitativo. Una limitazione, a tal proposito, sta nel fatto che si hanno a disposizione dati esterni e di scenario unicamente superiori alla soglia. Questo, nella pratica, implica che, volendo scegliere una soglia inferiore, non si hanno dati esterni a disposizione. In ogni caso per ovviare a tali problemi, si presenteranno nel Capitolo 3 diversi metodi per determinare la soglia dei valori estremi sulla base dei dati a disposizone.

La distribuzione di severity di una singola perdita interna sarà quindi una mistura di due distribuzioni, una per il corpo e una per la coda. Quindi, sia X la variabile aleatoria che rappresenta la severity di un singolo evento operativo; allora, la sua densità sarà data da:

$$f(x) = \begin{cases} \omega \cdot f^*_{body}(x) & \text{per } x \leq u \\ (1-\omega) \cdot f^*_{tail}(x) & \text{per } x > u \end{cases} \tag{1.1}$$

dove

- $f^*_{body}(x) = f_{body}(x)/F_{body}(u)$ è la distribuzione del corpo troncata ai valori inferiori ad u;
- $f^*_{tail}(x) = f_{tail}(x)/\left[1 - F_{tail}(u)\right]$ è la distribuzione della coda troncata ai valori superiori ad u;
- $\omega = F_{body}(u)$ è il peso del corpo della distribuzione.

Per la stima dei parametri del corpo della distribuzione di severity vengono solitamente utilizzati i dati interni degli ultimi 5 anni, che risultano sufficienti per ottenere stime robuste. Il problema dei dati raccolti, tuttavia, è che le perdite inferiori ad una prestabilita soglia H non sono considerate affidabili (spesso non vengono contabilizzati eventi di piccolo importo). Anche la scelta di questo parametro è qualitativa e fatta a priori,

in quanto si basa sulla confidenza nel processo di raccolta delle perdite. Tale tematica, insieme alla trattazione teorica delle distribuzioni troncate, è affrontata nel Capitolo 2.

Solitamente la procedura di stima utilizzata è quella di massima verosimiglianza, e si usa quindi la verosimiglianza troncata sopra H. Sia perciò $\boldsymbol{\vartheta}$ il vettore dei parametri da stimare, allora:

$$f_{body}(x;\boldsymbol{\vartheta}|x>H) = \frac{f_{body}(x;\boldsymbol{\vartheta})}{1-F_{body}(H;\boldsymbol{\vartheta})}\mathbb{1}_{[H,\infty)}(x) \tag{1.2}$$

Per la distribuzione del corpo dei singoli ET si utilizzano solitamente distribuzioni lognormale o Weibull, le cui densità di probabilità troncate relative sono le seguenti:

$$g_{LN}(x;\mu,\sigma^2) = \frac{1}{\Phi\left(\frac{\mu-\log H}{\sqrt{\sigma^2}}\right)}\frac{1}{x\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2\sigma^2}(\log x-\mu)^2}\mathbb{1}_{[H,\infty)}(x) \tag{1.3}$$

$$g_{Wei}(x;\theta,k) = \frac{k}{\theta}x^{k-1}e^{-\frac{1}{\theta}\left(x^k-H^k\right)}\mathbb{1}_{[H,\infty)}(x) \tag{1.4}$$

$$\tag{1.5}$$

La distribuzione più adatta a descrivere i dati viene individuata utilizzando test di buon adattamento, quali Kolmogorov-Smirnov o Anderson-Darling. Inoltre, se si ottengono risultati di bontà simili con diverse distribuzioni, si preferisce mantenere la scelta dell'anno precedente per mantenere la continuità.

Per quanto riguarda la coda, si sfrutta la teoria dei valori estremi, in particolare metodi Peaks Over Threshold (POT), trattati nel Capitolo 3: essi sfruttano le proprietà asintotiche delle eccedenze dei valori empirici rispetto ad una determinata soglia al fine di effettuare una stima parametrica della coda da cui ricavare le misure di esposizione al rischio della banca.

É possibile quindi assumere che la parte di severity superiore alla soglia u sia distribuita secondo una Pareto Generalizzata (GPD), con funzione di ripartizione:

$$F_{GDP}(u,\beta,\xi)(x) = 1-\left(1+\xi\frac{x-u}{\beta}\right)^{-1/\xi} \tag{1.6}$$

e densità:

$$f_{GDP}(u,\beta,\xi)(x) = \frac{1}{\beta}\left(1+\xi\frac{x-u}{\beta}\right)^{-1/\xi-1} \tag{1.7}$$

entrambe definite per $x > u$ e $\beta > 0$, dove β che rappresenta il parametro di scala. La distribuzione viene ristretta al caso $\xi > 0$ (parametro di forma) per ottenere una distribuzione con densità positiva per dati estremi.

Si noti che la distribuzione scelta per la coda ha peso unicamente per $x > u$ e quindi si avrà:

$$f^*_{tail} = f_{GDP}$$

I parametri della GPD vengono solitamente stimati tramite il metodo dei *Probability Weighted Moments*. I dati che si utilizzano in questo caso sono le perdite esterne, di scenario e interne superiori alla soglia u. Ciò, come detto in precedenza, viene fatto perché le perdite estreme sono caratterizzate da una frequenza di accadimento bassa, e quindi non si ha a disposizione un numero sufficiente di dati.

Tuttavia, per imporre la continuità della distribuzione, l'unico parametro della coda che viene stimato con questo metodo è ξ, in quanto β si ottiene imponendo la condizione di continuità in u:

$$\omega \cdot f^*_{body}(u) = (1 - \omega) \cdot f^*_{tail}(u) \tag{1.8}$$

Dall'espressione della densità della GPD si ha che

$$f^*_{tail}(u) = \frac{1}{\beta}$$

e quindi

$$\beta = \frac{1 - \omega}{\omega \cdot f^*_{body}(u)} = \frac{1 - F_{body}(u)}{f_{body}(u)} \tag{1.9}$$

Il motivo di questa scelta è che i dati esterni devono essere riscalati sulla base della grandezza del business della banca: quello che importa quindi non è tanto l'ordine di grandezza quanto la forma della coda, determinata da ξ, come verrà approfondito nel Capitolo 3.

1.4 La distribuzione di frequency

La densità di frequency è definita come la distribuzione di probabilità del numero di perdite operative nell'arco di un anno. Essa viene stimata sulla base delle sole perdite interne, ritenute come le uniche rilevanti per la stima della frequenza, in quanto riflettono maggiormente le caratteristiche della banca.

In generale, tutte le classi di rischio vengono modellizzate con una Poisson, di densità:

$$f_{Pois}(k;\lambda) = \frac{\lambda^k e^{-\lambda}}{k!}$$

dove il parametro $\lambda > 0$ viene stimato con lo stimatore dei momenti sulla base del numero di perdite annue n_i, con $i = 1, \ldots, M$, dove M è il numero di anni per cui si hanno a disposizione dati riguardanti le perdite interne (nel nostro caso $M = 5$). Quindi, lo stimatore sarà:

$$\lambda = \frac{1}{M} \sum_{i=1}^{M} n_i$$

Per coerenza con la stima della severity, e comunque per utilizzare solamente dati validati, si ripresenta il problema della scarsa affidabilità dei dati sotto la soglia H. L'idea è quindi quella di introdurre una stima del parametro λ condizionata alle sole perdite sopra soglia: si utilizza quindi il seguente stimatore:

$$\lambda_{sample} = \sum_{i=1}^{M} n_i'$$

dove n_i' rappresenta il numero di perdite operative nell'anno i-esimo il cui impatto economico è superiore alla soglia H. È quindi possibile trovare la stima del parametro λ della distribuzione di tutte le perdite con impatto positivo (quindi anche al di sotto della soglia) ricordando la normalizzazione data dalla formula di Bayes, ottenendo:

$$\lambda = \frac{\lambda_{sample}}{1 - F_{body}(H)} \tag{1.10}$$

dove F_{body} è la funzione di ripartizione del corpo della severity introdotta in precedenza.

Come alternativa alla distribuzione di Poisson è possibile modellizzare la frequency con una binomiale negativa, fermo restando un goodness of fit (ottenuto con Q-Q plot e test di bontà del fit) decisamente migliore, e cercando di mantenere continuità rispetto agli anni precedenti.

1.5 Convoluzione

Secondo l'approccio attuariale, assumendo (e verificando sulla base dei dati) l'indipendenza delle distribuzioni di frequency e severity, la stima

del capitale a rischio è ottenuta tramite convoluzione delle due distribuzioni tramite algoritmo Monte Carlo.

Supponendo infatti di conoscere le due densità di probabilità di frequency (Poisson) e severity (ottenuta dalla mistura di corpo e coda), ogni simulazione di perdita aggregata annua si ottiene come segue:

$$S = \sum_{i=1}^{N} s_i + s' \tag{1.11}$$

dove N è la il numero di perdite simulato dalla distribuzione di frequency, mentre s_i è l'importo di ogni singola perdita (simulato sopra la soglia H) e s' è la media empirica delle perdite aggregate annue sotto la soglia H. Generalmente si preferisce non campionare al di sotto della soglia H, ma aggiungere semplicemente quest'ultimo termine, in quanto prima di tutto non impatta più di tanto sul valore finale, ed inoltre poiché la stima delle distribuzioni troncate generalmente si adatta poco alla distribuzione dei dati al di sotto della soglia.

Per questo tipo di simulazione si possono prendere due strade diverse, sfruttando la proprietà della Poisson per cui la somma di due Poisson indpendenti è ancora una Poisson, il cui parametro è pari alla somma dei due parametri. Il primo metodo consiste quindi nel campionare N da una Poisson di parametro λ_{sample} e successivamente simulare N perdite dalla mistura di distribuzioni della frequency. Per il secondo si calcolano le frequenze annue relativamente al corpo e alla coda della distribuzione:

$$\lambda_{body} = \lambda \left[F_{body}(u) - F_{body}(H) \right] \tag{1.12}$$

$$\lambda_{tail} = \lambda \left[1 - F_{body}(u) \right] \tag{1.13}$$

e quindi si simulano le due Poisson di parametri λ_{body} e λ_{tail}, e le successive estrazioni da corpo e coda, separatamente.

Quello che si ottiene quindi è la distribuzione aggregata annua di perdite per ogni ET. A partire da queste è possibile fornire delle stime (parziali) del capitale a rischio, separatamente per ogni classe:

- VaR (*Value at Risk*): è definito (da disposizioni di Banca d'Italia) come il 99.9% percentile della distribuzione empirica; rappresenta il capitale da allocare per un ET.
- EL (*Expected Loss*): è definito come il valore medio della distribuzione; rappresenta le perdite attese per l'anno successivo, è utile ai fini manageriali per avere un'idea del rischio cui è sottoposta la banca, ma non si può utilizzare come indicatore del capitale da allocare.

1.6 Aggregazione delle classi di rischio

A questo punto, si hanno a disposizione le distribuzioni (empiriche) delle perdite aggregate annue di ogni classe di rischio. L'approccio più conservativo consiste nello stimare il capitale totale come somma delle singole classi. In questo modo si assume una correlazione lineare perfetta tra ogni coppia di ET.

In realtà, Banca d'Italia autorizza l'utilizzo di altre tecniche di aggregazione, assumendo strutture di correlazione diverse, purché si portino analisi di robustezza delle stesse. In generale, il calcolo della matrice di correlazione viene fatto con i metodi di Tau di Kendall o Rho di Spearman, entrambi basati sui ranghi delle singole osservazioni dei diversi ET. A rigore, la stima della matrice di correlazione andrebbe fatta sui dati aggregati annualmente. In pratica, tuttavia, tale serie ha numerosità non sufficientemente alta da produrre stime adeguate, e si preferisce usare le perdite aggregate su base mensile secondo la data di contabilizzazione (cercando un tradeoff tra il numero di dati a disposizione e un arco temporale sufficientemente alto).

Una volta stabilita la struttura di correlazione tra classi di rischio, l'aggregazione viene portata avanti attraverso una copula: questa è una particolare distribuzione multivariata che permette di aggregare le distribuzioni marginali mantenendo la struttura di correlazione desiderata.

Una tecnica differente, proposta recentemente, consiste nel modellizzare la dipendenza tra i vari ET assumendo l'evoluzione stocastica nel tempo dei profili di rischio, permettendo maggiore flessibilità alle dipendenze tra distribuzioni di severity e frequency delle diverse classi di rischio. L'utilizzo di questo metodo suppone tuttavia che le distribuzioni marginali siano modellizzate con l'approccio bayesiano accennato in precedenza.

Alla base dell'approccio "standard", per riscrivere una distribuzione multivariata utilizzando una copula, vi è l'idea di usare una semplice trasformazione delle marginali in modo che queste abbiano distribuzione uniforme. A questo punto, è possibile applicare la struttura di dipendenza sulle uniformi ottenute: una copula è proprio una distribuzione multivariata uniforme con marginali uniformi. La teoria delle copule è trattata nel capitolo 4

Definiti quindi la matrice di correlazione, il tipo di copula ed i relativi parametri, è possibile passare alla simulazione del capitale totale (la parte computazionalmente più onerosa). Anche questa si ottiene con simulazione Monte Carlo, date le 7 distribuzioni simulate delle classi di rischio, simulando successivamente un'osservazione (vettoriale) dalla cop-

ula e i relativi quantili delle distribuzioni marginali. Sommando le 7 osservazioni così generate, si ottiene dunque la distribuzione empirica annua dei diversi ET aggregati. Su questa distribuzione si andrà a calcolare il vero e proprio capitale a rischio (VaR), ovvero il percentile 99.9% calcolato sulla distribuzione empirica risultante, con relativi intervalli di confidenza.

Capitolo 2

Distribuzioni troncate

Nel modello di previsione del capitale da accantonare per far fronte ai rischi operativi, sorge in problema della modellizzazione di dati troncati. In particolare, le perdite storiche sono considerate affidabili solo sopra una certa soglia di raccolta. Per questo motivo, si utilizzano distribuzioni troncate per modellizzare il corpo della distribuzione di severity.

In questo capitolo si affronterà quindi questo problema: dopo un'introduzione alle variabili aleatorie troncate, ci si concentrerà sulle due distribuzioni utilizzate nella modellizzazione delle perdite operative (lognormale e Weibull), con particolare attenzione alla stima dei parametri e ai test di buon adattamento.

2.1 Variabili aleatorie troncate

Nell'analisi statistica, le variabili aleatorie troncate risultano dalla restrizione del dominio di una distribuzione di probabilità. In particolare, il problema dei dati troncati si ha quando il processo di raccolta dei dati è limitato, nel senso che non si raccolgono (o non è possibile conoscere) dati al di fuori di un certo intervallo.

In tal senso il concetto di troncamento è diverso dalla censura, che viene spesso riscontrata in problemi di affidabilità e tempo di vita di sistemi. Nel caso di censura, infatti, il campionamento è tale da registrare le osservazioni al di fuori dall'intervallo considerato, ma senza conoscerne il vero valore: i dati censurati sono quindi presenti nel dataset, ma portano meno informazioni di quelli completi, e solitamente indicano solamente se il limite inferiore (o superiore) è stato superato.

Nella pratica, esistono diversi tipi di troncamento, a seconda della forma dell'intervallo in cui si raccolgono le osservazioni. Si consideri quindi

una variabile aleatoria X con supporto in $\mathbb{R}^+$, e due valori $0 < a < b < \infty$. Se le osservazioni vengono raccolte solo in $[a, b]$ si parla di troncamento intervallare, in $[a, \infty]$ di troncamento a sinistra e in $[0, b]$ di troncamento a destra.

Nel modello in considerazione, come abbiamo detto, non vengono considerati affidabili le perdite operative al di sotto della soglia H, e per tale motivo nelle analisi si utilizzano solo le osservazioni maggiori o uguali ad H. Si tratterà quindi di un problema di troncamento a sinistra. D'ora in poi ci limiteremo a questo caso.

Sia dunque X una variabile aleatoria continua con funzione di ripartizione F e supporto in $\mathbb{R}^+$. La legge dei valori di X troncati in $[H, \infty)$ si modellizza dal punto di vista statistico con la legge condizionata di $X|X \geq H$.

Per determinare le leggi troncate si sfrutta la definizione di probabilità condizionata:

$$\mathrm{P}\left[A|B\right] = \frac{\mathrm{P}\left[A \cap B\right]}{\mathrm{P}\left[B\right]}$$

con $\mathrm{P}\left[B\right] > 0$.

A questo punto, è quindi possibile scrivere la funzione di ripartizione della legge di X in $[H, \infty)$ troncata come segue:

$$G(x) = \mathrm{P}\left[X \leq x | X \geq H\right] = \frac{\mathrm{P}\left[H \leq X \leq x\right]}{\mathrm{P}\left[X \geq H\right]},$$

ovvero:

$$G(x) = \begin{cases} 0 & \text{per } x < H \\ \frac{F(x)-F(H)}{1-F(H)} & \text{per } x \geq H \end{cases} \tag{2.1}$$

La densità di probabilità per distribuzioni troncate è quindi ottenibile derivando la funzione di ripartizione:

$$g(x) = \frac{\mathrm{d}}{\mathrm{d}x} G(x) = \frac{f(x)}{1 - F(H)} \mathbb{1}_{[H,\infty)}(x) \tag{2.2}$$

dove $f(x)$ è la densità di probabilità della distribuzione non troncata.

La media e la varianza della distribuzione troncata sono quindi definite come segue:

$$\mathrm{E}\left[X|X \geq H\right] = \frac{1}{1 - F(H)} \int_H^\infty x f(x) \mathrm{d}x \tag{2.3}$$

$$\mathrm{Var}\left[X|X \geq H\right] = \frac{1}{1 - F(H)} \left[\int_H^\infty x^2 f(x) \mathrm{d}x - \left(\int_H^\infty x f(x) \mathrm{d}x\right)^2\right] \tag{2.4}$$

Vediamo nelle prossime sezioni come si comportano le variabili aleatorie troncate in caso di distribuzione lognormale o Weibull.

2.2 La distribuzione lognormale troncata

Una variabile aleatoria X con distribuzione lognormale di parametri (μ, σ^2) è definita a partire dalla distribuzione normale come $X = \log Y$, dove Y è una $N(\mu, \sigma^2)$. Per tale motivo la densità di probabilità e la funzione di ripartizione si ricavano da quelle della gaussiana attraverso le leggi di trasformazione delle variabili aleatorie.

In particolare, la densità di una variabile aleatoria con legge lognormale di parametri (μ, σ^2) è la seguente:

$$f_{LN}(x; \mu, \sigma^2) = \frac{1}{x\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(\log x - \mu)^2} \mathbb{1}_{[0,\infty)}(x) \tag{2.5}$$

con $\mu \in \mathbb{R}$ e $\sigma^2 > 0$.

Come nel caso della distribuzione gaussiana, non è possibile determinare l'espressione analitica della funzione di ripartizione, che si può però scrivere come:

$$F_{LN}(x; \mu, \sigma^2) = \Phi\left(\frac{\log x - \mu}{\sqrt{\sigma^2}}\right) \mathbb{1}_{[0,\infty)}(x) \tag{2.6}$$

dove $\Phi(\cdot)$ è la funzione di ripartizione della normale standard, ovvero:

$$\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2} \mathrm{d}y$$

.

Inoltre, media e varianza di $X \sim LN(\mu, \sigma^2)$ hanno le seguenti espressioni:

$$\mathrm{E}[X] = e^{\mu + \frac{\sigma^2}{2}} \tag{2.7}$$

$$\mathrm{Var}[X] = e^{2\mu+\sigma^2}\left(e^{\sigma^2} - 1\right) \tag{2.8}$$

A partire dalla distribuzione sul supporto intero e dalle espressioni (2.1) e (2.2), si possono ricavare densità e funzione di ripartizione per la lognormale troncata, ovvero per $X|X \geq H$:

$$g_{LN}(x; \mu, \sigma^2) = \frac{1}{\Phi\left(\frac{\mu - \log H}{\sqrt{\sigma^2}}\right)} \frac{1}{x\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(\log x - \mu)^2} \mathbb{1}_{[H,\infty)}(x) \tag{2.9}$$

$$G_{LN}(x; \mu, \sigma^2) = \frac{\Phi\left(\frac{\log x - \mu}{\sqrt{\sigma^2}}\right) - \Phi\left(\frac{\log H - \mu}{\sqrt{\sigma^2}}\right)}{\Phi\left(\frac{\mu - \log H}{\sqrt{\sigma^2}}\right)} \mathbb{1}_{[H,\infty)}(x) \tag{2.10}$$

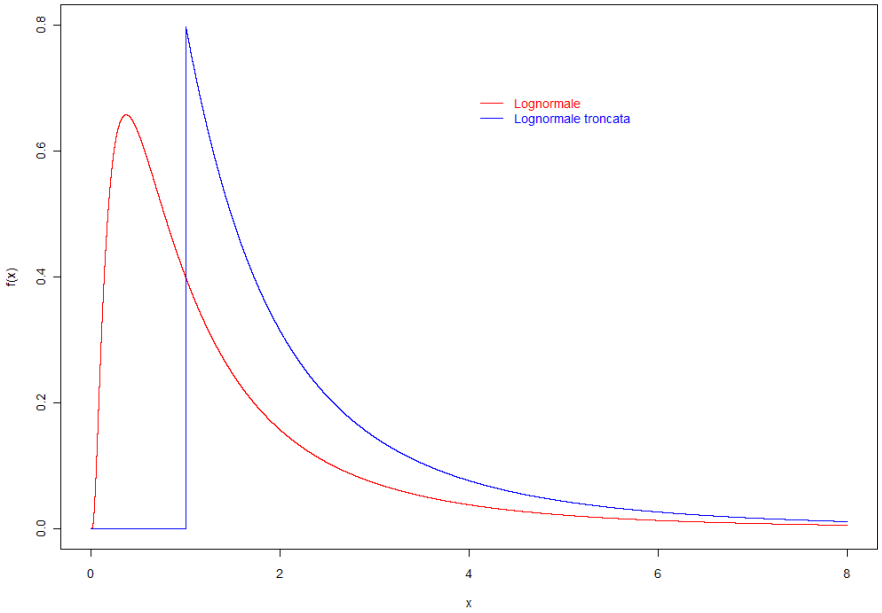

Figura 2.1: Confronto delle densità di probabilità per la distribuzione lognormale troncata e non, con $\mu = 0$, $\sigma^2 = 1$ e $H = 1$.

Si riportano in Figura 2.1 le densità di probabilità della lognormale standard ($\mu = 0$ e $\sigma^2 = 1$)su tutto il supporto e troncata sopra H.

È inoltre possibile [18] calcolare le espressioni di media e varianza nel caso di troncamento a partire dalle formule (2.3) e (2.4):

$$\begin{aligned} \mathrm{E}\left[X|X \geq H\right] &= e^{\mu+\frac{\sigma^2}{2}} \frac{\Phi\left(\frac{\log H+\sigma^2-\mu}{\sqrt{\sigma^2}}\right)}{\Phi\left(\frac{\log H-\mu}{\sqrt{\sigma^2}}\right)} \\ \mathrm{Var}\left[X|X \geq H\right] &= e^{2\mu+\sigma^2} \frac{e^{\sigma^2}\Phi\left(\frac{\log H+2\sigma^2-\mu}{\sqrt{\sigma^2}}\right) - \Phi^2\left(\frac{\log H+\sigma^2-\mu}{\sqrt{\sigma^2}}\right)}{\Phi\left(\frac{\log H-\mu}{\sqrt{\sigma^2}}\right)} \end{aligned}$$

2.3 La distribuzione Weibull troncata

L'altra distribuzione che viene generalmente utilizzata per modellizzare il corpo della distribuzione di severity delle perdite operative è la Weibull. Tale distribuzione ha supporto in $\mathbb{R}^+$ ed è caratterizzata da due parametri: $\theta > 0$, parametro di scala, e $k > 0$, parametro di forma.

Densità e funzione di ripartizione di una variabile $X \sim Wei(\lambda, k)$ sono definite come segue:

$$f_{Wei}(x;\theta,k) = \frac{k}{\theta}x^{k-1}e^{-\frac{1}{\theta}x^k}\mathbb{1}_{[0,\infty)}(x) \tag{2.11}$$

$$F_{Wei}(x;\theta,k) = \left[1-e^{-\frac{1}{\theta}x^k}\right]\mathbb{1}_{[0,\infty)}(x) \tag{2.12}$$

Inoltre, media e varianza della Weibull hanno le seguenti espressioni:

$$\mathrm{E}\left[X\right] = \theta^{\frac{1}{k}}\Gamma\left(1+\frac{1}{k}\right) \tag{2.13}$$

$$\mathrm{Var}\left[X\right] = \theta^{\frac{2}{k}}\left[\Gamma\left(1+\frac{2}{k}\right)+\Gamma^2\left(1+\frac{1}{k}\right)\right] \tag{2.14}$$

dove Γ è la funzione gamma, definita come:

$$\Gamma(\alpha) = \int_0^\infty t^{\alpha-1}e^{-t}\mathrm{d}t \tag{2.15}$$

Come per la lognormale, si ricavano densità e funzione di ripartizione per la variabile troncata dalle definizioni (2.1) e (2.2), ottenendo:

$$g_{Wei}(x;\theta,k) = \frac{k}{\theta}x^{k-1}e^{-\frac{1}{\theta}\left(x^k-H^k\right)}\mathbb{1}_{[H,\infty)}(x) \tag{2.16}$$

$$\begin{aligned} G_{Wei}(x;\theta,k) &= e^{\frac{1}{\theta}H^k}\left[e^{-\frac{1}{\theta}H^k} - e^{-\frac{1}{\theta}x^k}\right]\mathbb{1}_{[H,\infty)}(x) \\ &= \left[1 - e^{-\frac{1}{\theta}\left(x^k-H^k\right)}\right]\mathbb{1}_{[H,\infty)}(x) \end{aligned} \tag{2.17}$$

Si riportano in Figura 2.2 le densità della Weibull sull'intero supporto e troncata.

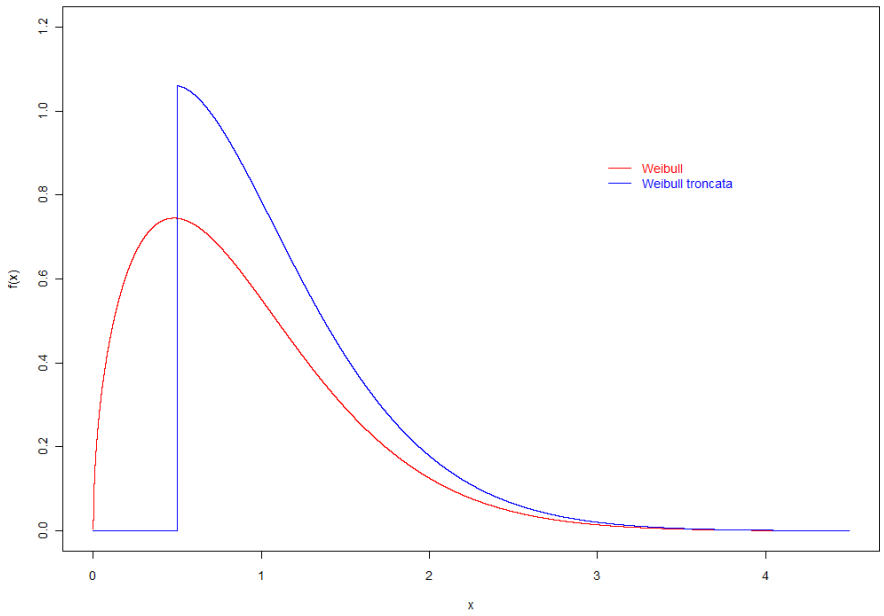

Figura 2.2: Confronto delle densità di probabilità per la distribuzione Weibull troncata e non, con $\theta = 1$, $k = 1.5$ e $H = 0.5$.

Media e varianza nel caso della Weibull troncata sono invece:

$$\mathrm{E}\left[X|X \geq H\right] = e^{\frac{1}{\theta}H^k}\theta^{\frac{1}{k}}\gamma\left(1+\frac{1}{k},\frac{1}{\theta}H^k\right)$$

$$\mathrm{Var}\left[X|X \geq H\right] = e^{\frac{1}{\theta}H^k}\theta^{\frac{2}{k}}\left[\gamma\left(1+\frac{2}{k},\frac{1}{\theta}H^k\right) + \gamma^2\left(1+\frac{1}{k},\frac{1}{\theta}H^k\right)\right]$$

dove γ è la funzione gamma incompleta, definita come:

$$\gamma(\alpha, z) = \int_z^\infty t^{\alpha-1} e^{-t} \mathrm{d}t$$

Dopo aver introdotto le distribuzioni troncate di interesse, si introdurranno nelle prossime sezioni i metodi di stima dei parametri.

2.4 Stimatori di massima verosimiglianza

Per la stima dei parametri delle distribuzioni troncate, il metodo solitamente utilizzato è quello della massima verosimiglianza (MLE); gli stimatori così ottenuti rappresentano i parametri del modello che più verosimilmente hanno generato le osservazioni a disposizione. Questo tipo di stima gode di diverse utili proprietà, discusse in [5]. Inoltre, in questo ambito, il metodo MLE è preferito al metodo dei momenti, in quanto per distribuzioni troncate è raramente possibile invertire le espressioni dei momenti per determinare analiticamente lo stimatore, come si può osservare dalle espressioni trovate nel caso di distribuzione lognormale e Weibull.

Sia quindi $y_i, \ldots, y_n$ un set di osservazioni generate da un campione i.i.d. con densità di probabilità $f(\cdot; \boldsymbol{\vartheta})$ (dove $\boldsymbol{\vartheta}$ è il vettore dei parametri del modello). Si definisce la verosimiglianza del modello come:

$$L(\boldsymbol{\vartheta}) = \prod_{i=1}^{k} f(y_i; \boldsymbol{\vartheta}).$$

Lo stimatore MLE per il campione si trova quindi massimizzando questa quantità, ovvero:

$$\widehat{\boldsymbol{\vartheta}} = \arg\max_{\boldsymbol{\vartheta}} L(\boldsymbol{\vartheta}).$$

Nella pratica, al posto della verosimiglianza è possibile massimizzare la log-verosimiglianza, definita come:

$$l(\boldsymbol{\vartheta}) = \log L(\boldsymbol{\vartheta}) = \sum_{i=1}^{k} \log f(y_i; \boldsymbol{\vartheta})$$

in quanto il massimo del logaritmo è raggiunto nello stesso valore $\widehat{\boldsymbol{\vartheta}}$ e l'espressione è solitamente più comoda dal punto di vista operativo. Per ricavare lo stimatore si pongono le derivate di questa quantità $= 0$ e, se possibile, si ottiene un'espressione esplicita. Altrimenti, è necessario ricorrere a metodi di ottimizzazione numerica per trovare il massimo.

Tornando al problema delle distribuzioni troncate, si supponga che $x_i, \ldots, x_n$ siano osservazioni troncate sopra la soglia H, generate da un campione i.i.d. con densità troncata $g(\cdot; \boldsymbol{\vartheta})$. L'espressione della log-verosimiglianza da massimizzare sarà quindi, dalla (2.2):

$$\begin{aligned} l(\boldsymbol{\vartheta}) &= \sum_{i=1}^{n} \log g(x_i; \boldsymbol{\vartheta}) = \sum_{i=1}^{n} \log \left[\frac{f(x_i; \boldsymbol{\vartheta})}{1 - F(H; \boldsymbol{\vartheta})} \right] \\ &= \sum_{i=1}^{n} \log f(x_i; \boldsymbol{\vartheta}) - n \log \left[1 - F(H; \boldsymbol{\vartheta}) \right] \end{aligned} \tag{2.18}$$

supponendo tutte le osservazioni all'interno del dominio (quindi $\geq H$). $F(\cdot; \boldsymbol{\vartheta})$ e $f(\cdot; \boldsymbol{\vartheta})$ sono rispettivamente funzione di ripartizione e densità delle variabili non troncate.

Un metodo alternativo per la stima dei parametri MLE del corpo della severity delle perdite operative è stato proposto in [2], attraverso l'uso dell'algoritmo *Expectation-Maximization* (EM), che mostra diversi vantaggi nel caso di distribuzione lognormale troncata delle perdite.

Ci concentreremo per ora sul metodo di massima verosimiglianza, vedendo come si applica nel problema dell'individuazione degli stimatori per le distribuzioni troncate utilizzate in questo ambito (lognormale e Weibull).

2.4.1 MLE per la distribuzione lognormale troncata

Per quanto riguarda la lognormale, la log-verosimiglianza per un campione $x_i, \ldots, x_n$ generato da una distribuzione lognormale troncata di parametri μ e σ^2, troncato sopra H, è la seguente:

$$\begin{aligned} l_{LN}(\mu, \sigma^2) &= \sum_{i=1}^{n} \log f_{LN}(x_i; \mu, \sigma^2) - n \log \left[1 - F_{LN}(H; \mu, \sigma^2) \right] \\ &= \sum_{i=1}^{n} \log \left[\frac{1}{x_i \sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(\log x_i - \mu)^2} \right] - n \log \Phi \left(\frac{\mu - \log H}{\sqrt{\sigma^2}} \right) \\ &\propto -\frac{1}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (\log x_i - \mu)^2 - n \log \Phi \left(\frac{\mu - \log H}{\sqrt{\sigma^2}} \right) \end{aligned} \tag{2.19}$$

dove nell'ultima espressione sono stati tralasciati i termini indipendenti da μ e σ^2, in quanto ininfluenti nella massimizzazione. Come si nota immediatamente, il problema di questa espressione è la presenza dell'integrale derivante dalla definizione della Φ. Esso rimane anche derivando

rispetto ai parametri, quindi non è possibile trovare un'espressione esplicita degli stimatori MLE nel caso della lognormale troncata. Per questo motivo, gli stimatori dei parametri della lognormale troncata vengono trovati attraverso metodi di ottimizzazione numerica bidimensionale.

Si riporta il codice R implementato per il calcolo degli stimatori MLE per μ e σ^2. Si è scelto a tal proposito l'utilizzo del metodo Nelder-Mead [24], uno degli algoritmi più utilizzati per l'ottimizzazione multidimensionale senza l'uso delle derivate:

```
MLE_LN<-function(x,H)
{
  n<-length(x)
  loglik<-function(t)
  {
    m<-t[1]
    s2<-t[2]
    phi<-pnorm((m-log(H))/sqrt(s2))
    l=-1/2*n*log(s2)-1/(2*s2)*sum((log(x)-m)^2)-n*log(phi)
    l
  }
  res<-constrOptim(c(0,1), loglik, grad=NULL, ui=c(0,1), 0,
        mu = 1e-04, control = list(fnscale=-1), "Nelder-Mead")
  res$par
}
```

2.4.2 MLE per la distribuzione Weibull troncata

Lo stesso ragionamento può essere fatto per determinare gli stimatori MLE per i parametri θ e k della distribuzione Weibull troncata sopra la soglia H. La verosimiglianza del modello, assumento le x_i generate da variabili aleatorie i.i.d. Weibull troncate, sarà:

$$\begin{aligned} l_{Wei}(\theta,k) &= \sum_{i=1}^{n} \log f_{Wei}(x_i;\theta,k) - n\log\left[1 - F_{Wei}(H;\theta,k)\right] \\ &= \sum_{i=1}^{n} \log\left[\frac{k}{\theta} {x_i}^{k-1} e^{-\frac{1}{\theta}x_i^k}\right] - n\log e^{-\frac{1}{\theta}H^k} \\ &= n\log k - n\log\theta + (k-1)\sum_{i=1}^{n}\log x_i - \frac{1}{\theta}\sum_{i=1}^{n} x_i^k + \frac{n}{\theta}H^k \end{aligned} \tag{2.20}$$

La particolarità del calcolo degli stimatori MLE per la distribuzione Weibull troncata è che è possibile combinare le due equazioni derivanti dall'imposizione delle derivate parziali uguali a 0, in modo da ottenere un'equazione in un'incognita da risolvere numericamente e un'equazione la cui radice è invece individuabile analiticamente. Questa caratteristica è stata osservata in [7] per la Weibull non troncata, ma può essere facilmente estesa al modello in considerazione.

Calcolando infatti la derivata della log-verosimiglianza rispetto a k si ha:

$$\begin{aligned}\frac{\partial}{\partial k} l_{Wei}(\theta, k) &= \frac{n}{k} + \sum_{i=1}^{n} \log x_i - \frac{1}{\theta} \sum_{i=1}^{n} x_i^k \log x_i + \frac{n}{\theta} H^k \log H \\ &= \frac{1}{\theta} \left(\frac{n}{k}\theta + \theta \sum_{i=1}^{n} \log x_i - \sum_{i=1}^{n} x_i^k \log x_i + nH^k \log H \right)\end{aligned}$$

Derivando invece rispetto a θ, si ottiene:

$$\begin{aligned}\frac{\partial}{\partial \theta} l_{Wei}(\theta, k) &= -\frac{n}{\theta} + \frac{1}{\theta^2} \sum_{i=1}^{n} x_i^k - \frac{n}{\theta^2} H^k \\ &= \frac{n}{\theta^2} \left(-\theta + \frac{1}{n} \sum_{i=1}^{n} x_i^k - H^k \right)\end{aligned}$$

Imponendo l'annullamento delle due derivate parziali, si arriva al sistema:

$$\frac{n}{k}\theta + \theta \sum_{i=1}^{n} \log x_i - \sum_{i=1}^{n} x_i^k \log x_i + nH^k \log H = 0 \tag{2.21}$$

$$-\theta + \frac{1}{n} \sum_{i=1}^{n} x_i^k - H^k = 0 \tag{2.22}$$

la cui soluzione $(\widehat{k}, \widehat{\theta})$ rappresenta lo stimatore MLE per questo modello.

È ora possibile ridurre il sistema ricavando le espressioni di θ da entrambe le equazioni ed eguagliandole:

$$\frac{\sum_{i=1}^{n} x_i^k \log x_i - nH^k \log H}{\frac{n}{k} + \sum_{i=1}^{n} \log x_i} = \frac{1}{n} \sum_{i=1}^{n} x_i^k - H^k,$$

ovvero, separando i termini dipendenti da k:

$$\frac{\sum_{i=1}^{n} x_i^k \log x_i - nH^k \log H}{\sum_{i=1}^{n} x_i^k - nH^k} - \frac{1}{k} = \frac{1}{n} \sum_{i=1}^{n} \log x_i. \tag{2.23}$$

A questo punto, lo stimatore per k si ottiene trovando numericamente le radici di questa espressione, solitamente attraverso algoritmi iterativi di ricerca locale. Si noti che da un problema di ottimizzazione numerico bidimensionale si è passati ad un problema di ricerca delle radici monodimensionale, quindi di complessità computazionale minore.

Dopo aver ottenuto in questo modo $\widehat{k}$, è possibile trovare lo stimatore MLE per θ dall'equazione (2.22):

$$\widehat{\theta} = \frac{1}{n} \sum_{i=1}^{n} x_i^{\widehat{k}} - H^{\widehat{k}} \tag{2.24}$$

Si riporta il codice R sviluppato per il calcolo degli stimatori MLE della Weibull troncata:

```
MLE_Wei<-function(x,H)
{
  n=length(x)
  find_k<-function(k)
  (sum(x^k*log(x))-n*H^k*log(H))/(sum(x^k)-n*H^k)-
                                                \1/k-1/n*sum(log(x))
  k<-uniroot(find_k, lower=10^(-5), upper=10, tol=1e-10)$root
  t=(1/n*sum(x^k)-H^k)
  c(k,t)
}
```

Dopo aver stimato i parametri del modello scelto, è necessario valutare il buon adattamento dei dati al modello. Questo viene fatto con metodi grafici (Q-Q plot) e test non parametrici.

2.5 Q-Q plot per distribuzioni troncate

Il metodo grafico più utilizzato per valutare l'adattamento dei dati alla distribuzione scelta, con i parametri stimati nel nostro caso con il metodo MLE, è il *Quantile-Quantile Plot* (Q-Q plot).

In statistica il Q-Q plot è un metodo per confrontare due distribuzioni statistiche, rappresentando i rispettivi quantili su un grafico. Nella pratica, esso viene utilizzato per verificare se una distribuzione empirica (i dati a disposizione) si adattano alla distribuzione teorica scelta per il fit.

L'idea è quella di rappresentare su un asse i valori dei quantili empirici (le osservazioni stesse) e su un altro i rispettivi quantili teorici. Se le due

distribuzioni sono simili (ovvero i dati si adattano alla distribuzione teorica con i parametri stimati), i punti che rappresentano le coppie quantile empirico-quantile teorico si disporranno approssimativamente lungo la retta $y = x$, che implica che i quantili teorici assumono valori simili a quelli empirici. Dall'osservazione del grafico, inoltre, è possibile valutare se ad esempio la distribuzione teorica sottostimi o sovrastimi la distribuzione dei dati reali.

Si noti tuttavia che tale metodo non è sufficiente per garantire un buon adattamento dei dati: esso viene solitamente affiancato ad altri metodi quantitativi, su cui ci si soffermerà nel seguito.

Sia ora $x_1, \dots, x_n$ un set di osservazioni e $F(\cdot; \boldsymbol{\vartheta})$ la funzione di ripartizione teorica scelta per il fit dei dati, con $\boldsymbol{\vartheta}$ stimato con un dato metodo. I quantili empirici, solitamente rappresentati sull'asse delle ascisse, sono semplicemente i dati a disposizione ordinati $x_{(1)}, \dots, x_{(n)}$.

Per determinare i corrispondenti quantili teorici, si definisce la funzione di ripartizione empirica come segue:

$$F_n(x) = \frac{\#\, osservazioni \leq x}{n+1} = \frac{1}{n+1} \sum_{i=1}^{n} \mathbb{1}_{[x_{(i)}, \infty)}(x), \tag{2.25}$$

dove la divisione per $n-1$ viene usata come correzione nelle applicazioni, per non avere la funzione di ripartizione esattamente pari a 1.

In particolare, si possono determinare le frequenze cumulate empiriche, corrispondenti alle osservazioni ordinate, come segue:

$$p_i^n = F_n(x_{(i)}) = \frac{i}{n+1} \qquad \text{per } i = 1, \dots, n \tag{2.26}$$

A questo punto, i quantili teorici si ottengono invertendo la funzione di ripartizione teorica e valutandola in questi valori:

$$q_i = F^{-1}(p_i^n; \boldsymbol{\vartheta})$$

ovvero calcolando i quantili corrispondenti ai livelli p_i^n. Il Q-Q plot si ottiene semplicemente rappresentando su un grafico i punti $\left\{ x_{(i)}, q_i \right\}$.

Tornando al fit di dati con distribuzioni troncate sopra la soglia H, il problema è che spesso tali funzioni di ripartizione non sono facilmente invertibili. Invece di determinare i quantili teorici invertendo la funzione di ripartizione troncata G, si passa dalla funzione di ripartizione non troncata F. Ricordando infatti la definizione della distribuzione troncata (2.1):

$$G(x) = \frac{F(x) - F(H)}{1 - F(H)}$$

i quantili teorici si possono scrivere come segue:

$$\widetilde{q}_i = F^{-1}(\widetilde{p}_i^n)$$

dove i $\widetilde{p}_i^n$ sono le frequenze cumulate della distribuzione troncata riscalate tra $F(H)$ e 1:

$$\widetilde{p}_i^n = F(H) + p_i^n \left[1 - F(H)\right] \tag{2.27}$$

A questo punto il Q-Q plot per distribuzioni troncate si ottiene rappresentando le coppie $\left\{x_{(i)}, \widetilde{q}_i\right\}$ e osservando quanto esse si dispongano lungo la retta $y = x$.

Si noti, tuttavia, che la parte più d'interesse di questo grafico, per la modellizzazione del corpo della severity delle perdite operative, sta nel tratto $[H, u]$, dove u è la soglia oltre la quale si assume che i dati siano distribuiti secondo le distribuzioni dei valori estremi, su cui ci si soffermerà nel seguito.

Nella pratica, i Q-Q plot vengono costruiti per le distribuzioni lognormale e Weibull, pertanto i quantili teorici, dall'inversione delle definizioni (2.6) e (2.12), hanno la seguente forma:

$$F_{LN}^{-1}(\widetilde{p}_i^n; \mu, \sigma^2) = \exp\left\{\Phi^{-1}(\widetilde{p}_i^n)\sqrt{\sigma^2} + \mu\right\} \tag{2.28}$$

$$F_{Wei}^{-1}(\widetilde{p}_i^n; k, \theta) = \left(\theta \log \frac{1}{1 - \widetilde{p}_i^n}\right)^{\frac{1}{k}} \tag{2.29}$$

Si riporta la funzione R per il calcolo dei quantili empirici e teorici da rappresentare nel Q-Q plot per generiche distribuzioni troncate:

```
qplot_build<-function(x,pfn,qfn,par,H)
{
  y<-sort(x)
  n<-length(x)
  z=(1:n)/(n+1)
  F<-pfn(H,par[1],par[2])
  val<-z*(1-F)+F
  w<-qfn(val,par[1],par[2])
  cbind(y,w)
}
```

2.6 Test di buon adattamento

Come accennato in precedenza, per verificare il buon adattamento dei dati alla distribuzione scelta con i parametri stimati, oltre ai metodi grafi-

ci, è opportuno portare avanti anche test non parametrici. Nell'ambito delle distribuzioni troncate, diversi test sono stati proposti in [6]. L'idea generale di questo tipo di analisi è quella di confrontare la funzione di ripartizione empirica con quella teorica corrispondente.

Sia quindi G_n la funzione di ripartizione empirica delle osservazioni $x_1, \dots, x_n$, troncate sopra la soglia H, definita nella (2.25), e G la funzione di ripartizione teorica scelta, con i parametri stimati ad esempio con il metodo MLE. I test di buon adattamento andranno a verificare l'ipotesi secondo cui i dati siano o meno distribuiti secondo la distribuzione teorica, ovvero:

$$\mathrm{H}_0 : G_n(x) \approx G(x) \qquad \text{vs} \qquad \mathrm{H}_1 : G_n(x) \not\approx G(x)$$

Quindi, una volta scelta la statistica D che identifica il test, si definisce il p-value come la probabilità che la statistica sia maggiore del valore $\overline{D}$ corrispondente alle osservazioni. Nella pratica, questa probabilità è calcolata tramite simulazione Monte Carlo, seguendo la seguente procedura:

1. Si simulano N set di campioni di numerosità n, pari al numero di osservazioni, dalla distribuzione teorica G.

2. Si calcola la statistica del test D_j per ogni set $j = 1, \dots, N$.

3. Si determina il p-value del test come la percentuale di volte in cui le statistiche simulate superano la statistica corrispondente alle osservazioni $\overline{D}$.

In definitiva, il p-value indica la percentuale di simulazioni per cui si è ottenuta una distanza (misurata in termini della statistica D) maggiore rispetto a quella del set di osservazioni. Fissato quindi il livello del test α, si rifiuterà l'ipotesi nulla di buon adattamento dei data alla distribuzione teorica scelta nel caso in cui il p-value sia inferiore ad α.

Esistono diversi tipi di statistiche per questo test, le più utilizzate nell'ambito delle distribuzioni troncate sono quella di Kolmogorov-Smirnov e quella di Anderson-Darling.

2.6.1 Kolmogorov-Smirnov

Sia d'ora in poi $x_1, \dots, x_n$ un campione generato da variabili aleatore i.i.d. troncate sopra la soglia H, con funzione di ripartizione empirica G_n. Si denota con $x_{(1)}, \dots, x_{(n)}$ il campione disposto in ordine crescente. Questo può essere il set di osservazioni utilizzate per la stima dei parametri oppure simulate nel test di buon adattamento. Si assuma poi che, G sia la

funzione di ripartizione troncata come da ipotesi nulla del test di buon adattamento e F la relativa distribuzione non troncata.

La statistica del test di Kolmogorov-Smirnov (KS) è quindi definita come segue:

$$KS = \sup_x |G_n(x) - G(x)| \tag{2.30}$$

Essa misura la distanza massima tra la funzione di ripartizione empirica e quella teorica, dando uguale peso a ogni osservazione.

Nella pratica, denotando con $z_i = F(x_{(i)})$ e $z_H = F(H)$ questa quantità è calcolata come segue, ricordando le espressioni (2.25) e (2.1):

$$\begin{aligned} KS &= \max_i \left| G_n(x_{(i)}) - G(x_{(i)}) \right| \\ &= \max_i \left| \frac{i}{n+1} - \frac{z_i - z_H}{1 - z_H} \right| \end{aligned} \tag{2.31}$$

2.6.2 Anderson-Darling

Il test di Anderson-Darling (AD) è un'estensione del test KS: nella relativa statistica, infatti, ad ogni osservazione viene assegnato un peso pari a

$$[G(x)\,(1 - G(x))]^{-1/2}\,.$$

Ovvero, viene assegnato un peso maggiore alle osservazioni nelle due code della distribuzione (lontano dalla mediana). La statistica AD è definita come segue:

$$AD = \sup_x \left| \frac{G_n(x) - G(x)}{\sqrt{G(x)\,(1 - G(x))}} \right| \tag{2.32}$$

Essa misura la distanza massima tra la funzione di ripartizione empirica e quella teorica, dando uguale peso a ogni osservazione.

Operativamente, la statistica si calcola con la seguente espressione:

$$\begin{aligned} AD &= \max_i \left| \frac{G_n(x_{(i)}) - G(x_{(i)})}{\sqrt{G(x_{(i)}) \left(1 - G(x_{(i)})\right)}} \right| \\ &= \max_i \left| \frac{1 - z_H}{\sqrt{(z_i - z_H)\,(1 - z_i)}} \left\{ \frac{i}{n+1} - \frac{z_i - z_H}{1 - z_H} \right\} \right| \\ &= \max_i \left| \frac{\frac{i}{n+1}\,(1 - z_H) - z_i + z_H}{\sqrt{(z_i - z_H)\,(1 - z_i)}} \right| \end{aligned} \tag{2.33}$$

Una variante del test AD, molto utilizzata nell'analisi di serie storiche riguardanti dati di perdite, è l'*Upper Tail* Anderson-Darling (ADup). In questo caso, ad ogni osservazione, rispetto alla statistica KS, viene assegnato un peso pari a

$$(1 - G(x))^{-1},$$

ovvero si è più preoccupati delle discrepanze tra distribuzione empirica e teorica nella coda destra della distribuzione, corrispondente ai valori estremi.

La relativa statistica è definita come segue:

$$AD_{up} = \sup_x \left| \frac{G_n(x) - G(x)}{1 - G(x)} \right| \tag{2.34}$$

ovvero:

$$\begin{aligned} AD_{up} &= \max_i \left| \frac{G_n(x_{(i)}) - G(x_{(i)})}{1 - G(x_{(i)})} \right| \\ &= \max_i \left| \frac{\frac{i}{n+1}(1 - z_H) - z_i + z_H}{1 - z_i} \right| \end{aligned}$$

I test di buon adattamento per distribuzioni troncate sono implementate nel pacchetto [29] del software R.

Capitolo 3

Teoria dei valori estremi

Nell'ambito della modellizzazione delle perdite operative, un ruolo fondamentale è giocato dalla coda della distribuzione di severity, che è quella che maggiormente influenza i quantili di ordine elevato della distribuzione. Essa viene generalmente modellizzata con una *Generalized Pareto Distribution* (GPD) che deriva dalla teoria dei valori estremi. Per questo motivo, si andrà ad approfondire questo argomento, per poi arrivare alla distribuzione di interesse e soffermarsi sul problema della stima dei parametri.

La teoria dei valori estremi ([22] e [28]) ha come obiettivo quello di modellizzare i dati appartenenti alla coda destra di una distribuzione, quindi che si discostano molto dalla mediana e per questo considerati estremi, attraverso approssimazioni asintotiche. Il punto di partenza è la modellizzazione del massimo di una serie di distribuzioni, attraverso la famiglia di distribuzioni *Generalized Extreme Value* (GEV), per arrivare infine alla GPD, come estensione della GEV per stimare il comportamento dei dati che superano una determinata soglia.

3.1 Block maxima

L'approccio alla teoria dei valori estremi che ha come scopo l'individuazione della legge del massimo prende il nome di *block maxima*: l'idea è quella di suddividere il campione in blocchi (ad esempio, osservazioni annuali) e studiare le proprietà dei massimi dei diversi blocchi, considerate realizzazioni dei valori "estremi" della distribuzione.

Si consideri la sequenza di variabili aleatorie reali i.i.d. $X_1, X_2, \ldots, X_n$ con la stessa funzione di ripartizione F. L'obiettivo di questa sezione è

determinare la legge del massimo di queste, ovvero:

$$M_n = \max\{X_1, X_2, \ldots, X_n\}$$

In generale, la funzione di ripartizione di un massimo M_n è definibile analiticamente come segue, sfruttando l'indipendenza delle X_i:

$$P[M_n \leq z] = P[X_1 \leq z, X_2 \leq z, \ldots, X_1 \leq z] = F^n(z), \tag{3.1}$$

con $z \in \mathbb{R}$.

Tuttavia, questo approccio non è utilizzabile in pratica se la funzione di ripartizione F non è nota. In tal caso sarebbe possibile utilizzare una stima della funzione di ripartizione delle singole v.a. X_i per determinare la legge del massimo. Tale metodo però porta a distorsioni notevoli nel caso di un alto numero di osservazioni n.

Per questo motivo, l'approccio della teoria dei valori estremi è quello di cercare la distribuzione teorica che meglio approssimi il comportamento asintotico di F^n, sfruttando il Teorema Centrale del Limite (TCL). Il vantaggio principale di questo metodo è che la stima di questa distribuzione è ottenibile utilizzando unicamente realizzazioni delle v.a. X_i estreme, senza formulare ipotesi circa la funzione di ripartizione F.

Il risultato principe della teoria dei valori estremi è il teorema di Fisher-Tippett, che permette, sotto le opportune ipotesi, di individuare tre famiglie di distribuzioni cui appartengono le leggi del massimo di una successione di variabili aleatorie.

Prima di procedere all'enunciato, è necessario però ragionare sul comportamento asintotico dei massimi. Questi valori, per n crescente, andranno ad avvicinarsi sempre di più verso il limite superiore del supporto della distribuzione F, e quindi la distribuzione dei massimi dipenderà dal comportamento di F nei pressi di questo valore. Esso viene definito come il minimo valore per cui la funzione di ripartizione ha valore 1, ovvero:

$$z_F = \min\{z \in \mathbb{R} : F(z) \geq 1\} = \sup\{z \in \mathbb{R} : F(z) < 1\} \tag{3.2}$$

Se $z_F < \infty$, all'aumentare di n si ha la convergenza di $F^n(z)$ a 0 per ogni $z < z_F$, ottenendo quindi una massa di probabilità concentrata in z_F, ovvero $M_n \to z_F$ in probabilità. Per evitare questa eventualità, di scarso interesse in questo ambito, si consideri la normalizzazione del massimo, come segue:

$$M_n^* = \frac{M_n - b_n}{a_n} \tag{3.3}$$

dove $a_n > 0$ e $b_n \in \mathbb{R}$ sono due successioni di costanti, scelte in modo tale da evitare che $M_n^* \xrightarrow{P} z_F$.

A questo punto, è possibile enunciare il teorema di Fisher-Tippett, che permette di individuare tre possibili classi di distribuzioni limite per il massimo in caso di convergenza della distribuzione di M_n^*.

Teorema 3.1. (Fisher-Tippett) *Sia (X_n) una successione di variabili aleatorie i.i.d. Se esistono due successioni di costanti $\{a_n > 0\}$ e $\{b_n\}$ tali che:*

$$\mathrm{P}\left[\frac{M_n - b_n}{a_n} \leq z\right] \to G(z) \qquad \text{con } n \to \infty$$

dove G è una funzione di ripartizione non degenere, allora G appartiene a una delle seguenti famiglie:

$$\begin{aligned}
\textit{Gumbel:} \qquad G(z) &= \exp\left\{-\exp\left[-\left(\frac{z-\mu}{\sigma}\right)\right]\right\} \\
\textit{Frechét:} \qquad G(z) &= \begin{cases} 0 & z \leq \mu \\ \exp\left\{-\left(\frac{z-\mu}{\sigma}\right)^{-\alpha}\right\} & z > \mu \end{cases} \\
\textit{Weibull:} \qquad G(z) &= \begin{cases} \exp\left\{-\left[-\left(\frac{z-\mu}{\sigma}\right)\right]^{\alpha}\right\} & z < \mu \\ 1 & z \geq \mu \end{cases}
\end{aligned}$$

con $\sigma > 0$, $\mu \in \mathbb{R}$ e $\alpha > 0$.

Questo teorema ci permette di individuare le tre classi di famiglie di distribuzioni che approssimano il comportamento del massimo quando $n \to \infty$. In particolare, le tre distribuzioni sono gli unici possibili limiti della distribuzione di M_n^*, opportuna standardizzazione del massimo.

3.2 La distribuzione Generalized Extreme Value

Le tre distribuzioni limiti per la legge del massimo hanno caratteristiche diverse, modellizzando quindi i valori estremi in modo diverso. Si riportano in Figura 3.1 le densità di probabilità per le tre famiglie, per particolari scelte dei parametri.

Come si vede, il limite destro del supporto z_F della Weibull è finito (pari a μ), mentre per le altre due distribuzioni è infinito. Inoltre, anche il comportamento di queste due è differente: la densità della Gumbel decresce esponenzialmente, mentre quella della Frechét polinomialmente (quest'ultima ha anche supporto limitato a sinistra da μ). Un'analisi dettagliata delle differenze tra le distribuzioni limite e la definizione dei bacini di attrazione delle stesse (ovvero delle caratteristiche delle distribuzioni

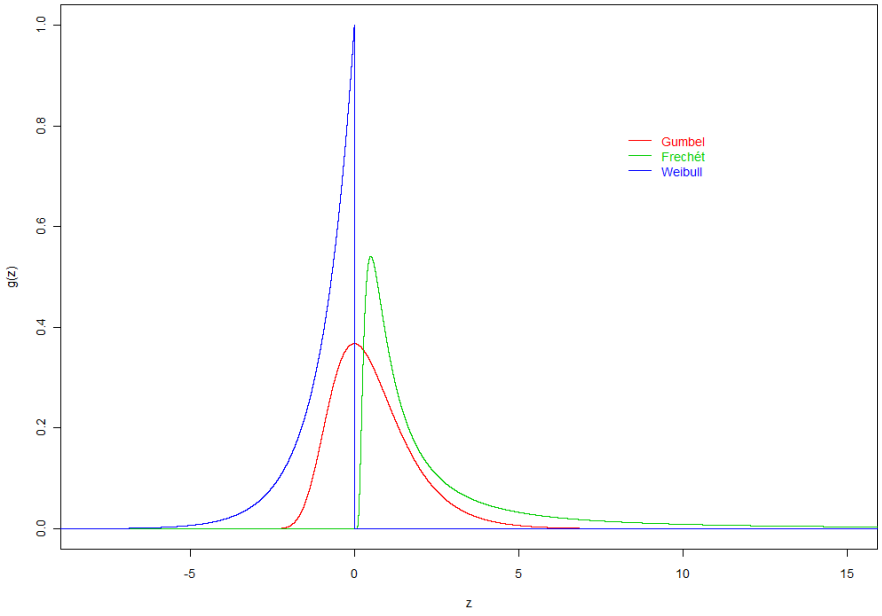

Figura 3.1: Confronto delle densità di probabilità per le famiglie Gumbel, Frechét e Weibull, con $\mu = 0$, $\alpha = 1$ e $\sigma = 1$.

di F che determinano una differente distribuzione del massimo) è riportata in [9].

Queste differenze sono un primo limite dell'utilizzo delle tre distribuzioni separatamente: nella pratica, la scelta della distribuzione limite va fatta a priori, e questa incertezza va ad incrementare quella determinata dall'inferenza per la stima dei parametri delle diverse distribuzioni.

Per questo motivo, è utile unire le tre distribuzioni per il massimo in un'unica, la famiglia di distribuzioni *Generalized Extreme Value* (GEV). Ciò è possibile introducendo un nuovo parametro ξ, che determina l'appartenenza a una delle tre classi precedenti; infatti, la funzione di ripartizione della GEV è definita come segue:

$$G(z) = \exp\left\{-\left[1+\xi\left(\frac{z-\mu}{\sigma}\right)^{-1/\xi}\right]\right\}, \tag{3.4}$$

con $\sigma > 0$, $\mu \in \mathbb{R}$ e $\xi > 0$. Il supporto della distribuzione GEV è l'insieme $\{z \in \mathbb{R} : 1+\xi(z-\mu)/\sigma > 0\}$.

I parametri della distribuzione hanno un significato fisico: μ è il parametro di posizione, σ è il parametro di scala, mentre ξ è il parametro di forma. Come accennato, è immediato osservare che diverse scelte di questo parametro implicano l'identificazione delle distribuzioni precedenti con la GEV:

$\xi = \alpha^{-1} > 0$	corrisponde alla distribuzione Frechét
$\lim \xi \to 0$	corrisponde alla distribuzione Gumbel
$\xi = -\alpha^{-1} < 0$	corrisponde alla distribuzione Weibull

Potendo parametrizzare le tre famiglie di distribuzioni limite in un'unica espressione, è quindi possibile fare inferenza sull'appartenenza della legge del massimo ad una specifica famiglia stimando, direttamente dai dati, il parametro ξ, senza dover fare ipotesi a priori.

Di conseguenza, il Teorema 3.1 può essere riscritto come segue.

Teorema 3.2. *Sia (X_n) una successione di variabili aleatorie i.i.d. Se esistono due successioni di costanti $\{a_n > 0\}$ e $\{b_n\}$ tali che:*

$$\mathrm{P}\left[\frac{M_n - b_n}{a_n} \leq z\right] \to G(z) \qquad \textit{con } n \to \infty$$

dove G è una funzione di ripartizione non degenere, allora G appartiene alla famiglia GEV

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

con $\sigma > 0$, $\mu \in \mathbb{R}$ *e* $\alpha > 0$ *e supporto* $\{z \in \mathbb{R} : 1 + \xi(z - \mu)/\sigma > 0\}$.

A questo punto, sfruttando questo teorema, è possibile determinare la distribuzione del massimo di blocchi di dati, supponendo valide le ipotesi. A questo proposito, è necessario sottolineare che le costanti di normalizzazione b_n ed a_n non sono note a priori, né possono essere stimate dai dati.

Si può però sfruttare il fatto che, supponendo che valga

$$\mathrm{P}\left[\frac{M_n - b_n}{a_n} \leq z\right] \approx G(z)$$

per valori di n sufficientemente grandi, si ha che:

$$\mathrm{P}\left[M_n \leq z\right] = \mathrm{P}\left[M_n^* \leq \frac{z - b_n}{a_n}\right] \approx G\left(\frac{z - b_n}{a_n}\right) = G^*\left(z\right)$$

dove G^* appartiene ancora alla famiglia GEV.

Grazie a questo risultato e al Teorema 3.2, la distribuzione del massimo può essere approssimata da G^* e non è necessario passare dalla normalizzazione di M_n. A partire dai dati, quindi, sarà possibile stimare i parametri di questa nuova distribuzione, pur non avendo informazioni sui coefficienti a_n e b_n. Questi saranno i parametri di interesse, non quelli della distribuzione di M_n^*.

In definitiva, con il metodo dei block maxima, a partire da sequenze di dati $X_1, \ldots, X_n$ (divisi in blocchi), è possibile approssimare la distribuzione dei massimi dei singoli blocchi con una GEV. I metodi generalmente utilizzati per la stima dei parametri sono la massima verosimiglianza e il metodo dei momenti.

Il problema del metodo presentato, tuttavia, è che è richiesto un alto numero di osservazioni X_i, tale da poterle dividere in blocchi che siano di numerosità sufficiente perché il massimo abbia significatività, ma allo stesso tempo averne abbastanza da poter fare inferenza sui parametri della GEV.

Inoltre, scegliendo un unico elemento di ogni blocco, si perdono informazioni date da altre osservazioni estreme: per questo motivo è stato introdotto il metodo dei *Peaks Over Threshold* (POT), che permette di utilizzare le osservazioni sopra una data soglia per fare inferenza sul comportamento dei valori molto grandi della distribuzione.

3.3 Peaks over threshold

Si consideri nuovamente una serie di variabili aleatorie i.i.d. $X_1, \ldots, X_n$ con funzione di ripartizione F. È lecito considerare estremi quei valori di X_i che superano una determinata soglia u (si approfondirà il problema della scelta di questa soglia nel seguito).

Si può descrivere il comportamento di questi valori estremi, che chiameremo "eccessi", attraverso probabilità condizionate, come segue:

$$\mathrm{P}\left[X_i > u + x | X_i > u\right] = \frac{1 - F(u+x)}{1 - F(u)}, \qquad x \geq 0, \tag{3.5}$$

ovvero la funzione di ripartizione di $X_i - u$, condizionato al fatto che X_i sia maggiore di 0, vale:

$$\mathrm{P}\left[X_i - u \leq x | X_i > u\right] = \frac{F(u+x) - F(u)}{1 - F(u)}, \qquad x \geq 0. \tag{3.6}$$

Allo stesso modo della distribuzione dei massimi, la legge degli eccessi oltre la soglia u è nota se si conosce la funzione di ripartizione delle singole variabili aleatorie. Tuttavia, come nella discussione precedente, spesso si ricorre ad approssimazioni di questa distribuzione, basandosi ancora una volta sulla teoria dei valori estremi. Il punto cardine di questo metodo è il legame tra la distribuzione dei massimi e quella degli eccessi, come mostrato dal seguente teorema.

Teorema 3.3. *Sia (X_n) una successione di variabili aleatorie i.i.d. con funzione di ripartizione F, e*

$$M_n = \max\left\{X_1, X_2 \ldots, X_n\right\}.$$

Si supponga che F soddisfi il Teorema 3.2, ovvero:

$$\mathrm{P}\left[\frac{M_n - b_n}{a_n} \leq z\right] \to G(z) \qquad \text{con } n \to \infty$$

dove

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

con $\sigma > 0$, $\mu \in \mathbb{R}$ e $\alpha > 0$. Allora, per un u sufficientemente grande, la funzione di ripartizione di $(X_i - u)$ condizionatamente a $X_i > u$ si può approssimare con:

$$H(x) = 1 - \left(1 + \xi\frac{x-u}{\beta}\right)^{-1/\xi} \tag{3.7}$$

definita su $\{x > 0 : 1 + \xi(x-u)/\beta > 0\}$, dove

$$\beta = \sigma + \xi(u - \mu). \tag{3.8}$$

3.4 La Generalized Pareto Distribution

La famiglia di distribuzioni descritte dalla (3.7) è detta *Generalized Pareto Distribution* (GPD). Il legame tra questa e la famiglia GEV deriva dal teorema appena enunciato: se la legge del massimo di una successione di variabili aleatorie i.i.d. appartiene (asintoticamente) alla GEV, allora gli eccessi oltre la soglia u della stessa successione sono approssimativamente distribuiti come una GPD.

Il forte legame delle due distribuzioni è evidenziato anche dal significato dei parametri della distribuzione degli eccessi e dalla loro dipendenza dai parametri della GEV. Infatti, ξ è lo stesso della GEV, e assume quindi il significato di parametro di forma della distribuzione. Il parametro di scala β, invece, si ottiene da una relazione con tutti i parametri della GEV, ma in particolare è una trasformazione lineare del parametro di scala di essa (σ). u assume ovviamente il ruolo di parametro di posizione della distribuzione.

Si noti inoltre che il caso $\xi = 0$ si può intendere come limite della (3.7), ottenibile tramite limite notevole, come segue:

$$H(x)|_{\xi=0} = \lim_{\xi \to 0} \left[1 - \left(1 + \xi \frac{x-u}{\beta} \right)^{-1/\xi} \right] = 1 - \exp\left\{ -\frac{x-u}{\beta} \right\}$$

A questo punto la distribuzione GPD di parametri $u \in \mathbb{R}$, $\beta > 0$ e $\xi \in \mathbb{R}$ si può descrivere con la funzione di ripartizione:

$$F_{GPD}(x) = \begin{cases} 1 - \left(1 + \xi \frac{x-u}{\beta} \right)^{-1/\xi} & \xi \neq 0 \\ 1 - \exp\left\{ -\frac{x-u}{\beta} \right\} & \xi = 0 \end{cases} \qquad x \in D(u, \beta, \xi) \quad (3.9)$$

e densità di probabilità:

$$f_{GPD}(x) = \begin{cases} \frac{1}{\beta} \left(1 + \xi \frac{x-u}{\beta} \right)^{-1/\xi - 1} & \xi \neq 0 \\ \frac{1}{\beta} \exp\left\{ -\frac{x-u}{\beta} \right\} & \xi = 0 \end{cases} \qquad x \in D(u, \beta, \xi) \quad (3.10)$$

dove $D(u, \beta, \xi)$ è il supporto della distribuzione, definito come:

$$D(u, \beta, \xi) = \begin{cases} x \geq u & \xi \geq 0 \\ u \leq x \leq u - \frac{\beta}{\xi} & \xi < 0 \end{cases} \quad (3.11)$$

Soffermandosi ulteriormente sui parametri della GPD, un ruolo fondamentale è svolto da ξ: esso determina il comportamento qualitativo della

distribuzione. Per $\xi < 0$, infatti, il supporto è limitato a destra, mentre per $\xi \geq 0$ no. Inoltre, un parametro di forma più grande determina una densità di probabilità molto più concentrata verso valori estremi.

Si riporta in Figura 3.2 la densità di probabilità della GPD per diversi valori di ξ positivi, osservando come, all'aumentare del parametro di forma, essa si concentri maggiormente verso valori estremi, ovvero nella coda destra.

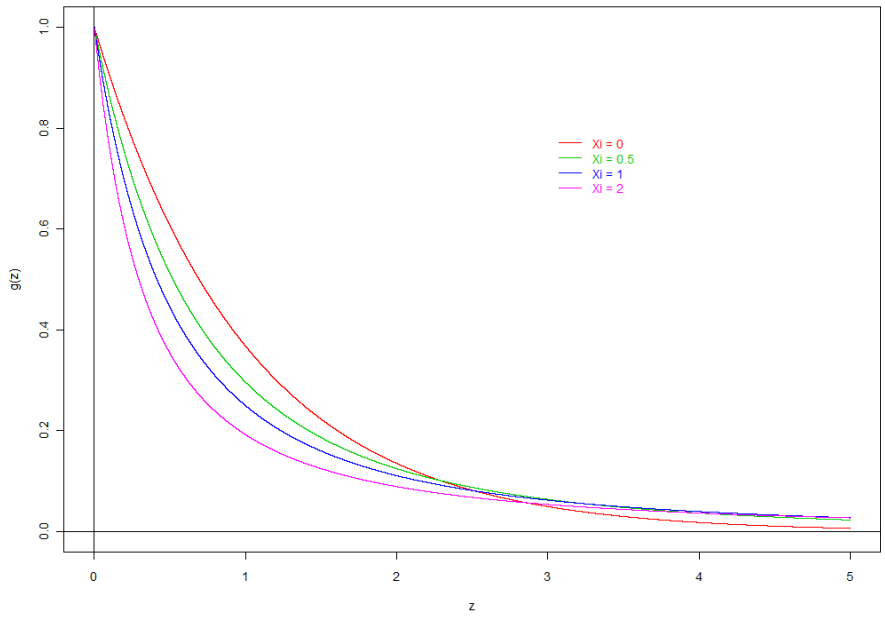

Figura 3.2: Confronto delle densità di probabilità della GPD con $u = 0$, $\beta = 1$ e e diversi valori di $\xi \geq 0$.

Analogamente, in Figura 3.3 si riportano le densità della GPD per valori di ξ negativi, osservando in particolare come tutte queste distribuzioni abbiamo supporto limitato, sempre più stretto all'aumentare del modulo di ξ.

Com'è evidente dall'analisi del comportamento della distribuzione GPD al variare del parametro di forma ξ, nel caso della modellizzazione delle perdite estreme derivanti da rischi operativi, non ha senso limitare il supporto della distribuzione: ipoteticamente, non ci sono limiti di perdite di questo tipo. Per questo motivo, la scelta di questo parametro verrà limitata al caso $\xi \geq 0$; si manterrà questa ipotesi per tutte le analisi successive.

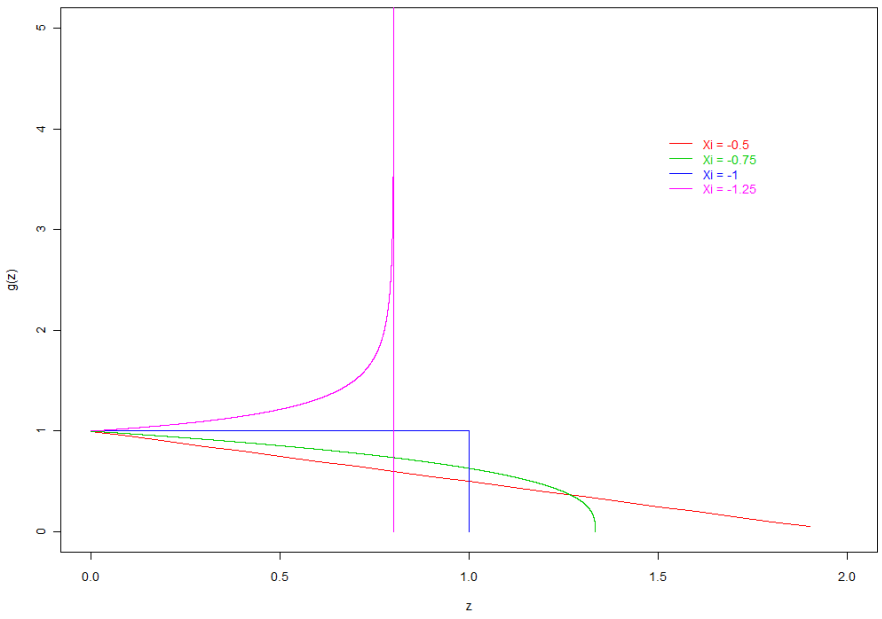

Figura 3.3: Confronto delle densità di probabilità della GPD con $u = 0$, $\beta = 1$ e diversi valori di $\xi < 0$.

3.5 Stima dei parametri

S affronterà adesso il problema della scelta dei parametri della Generalized Pareto Distribution, a partire dalla scelta della soglia u, per poi soffermarsi sulle diverse tecniche di stima dei parametri.

3.5.1 Scelta della soglia u

Un primo problema che riguarda la modellizzazione degli eventi estremi con il metodo dei POT è proprio la scelta della soglia u. In particolare, questa quantità dev'essere grande abbastanza in modo da poter considerare estreme le osservazioni superiori ad essa: scegliere u troppo piccolo implicherebbe stime distorte degli altri parametri [8]. D'altra parte, una scelta di u troppo alta implicherebbe un minor numero di osservazioni con cui poi fare inferenza sui parametri della GPD, e quindi maggiore variabilità delle stime. La scelta di questo parametro ricade, come gran parte dei problemi della statistica, nella ricerca di un tradeoff *bias-variance* (distorsione vs varianza della stima).

Nella pratica, si utilizzano due metodi per la scelta della soglia u: un primo approccio si basa sull'analisi esplorativa del dataset, mentre il secondo su analisi di robustezza delle stime degli altri parametri al variare della soglia.

Il primo metodo per la scelta di u si basa sull'espressione della media della distribuzione GPD. In particolare, se X è una variabile aleatoria che rispetta le ipotesi del Teorema 3.3, e che quindi oltre la soglia u si può approssimare con una GPD di parametri (u, β, ξ), si può verificare (lo vedremo nel dettaglio nel seguito) che la media degli eccessi è:

$$\mathrm{E}\left[X - u \middle| X > u\right] = \frac{\beta}{1-\xi}. \tag{3.12}$$

In particolare, la media è finita se $\xi < 1$. Ci restringeremo per il momento a questo caso.

A questo punto, per ogni $v > u$ vale la seguente formula:

$$\mathrm{E}\left[X - v | X > v\right] = \frac{\beta + \xi(v-u)}{1-\xi}. \tag{3.13}$$

Si noti che l'espressione (3.12) può essere ottenuta ponendo $u = v$.

Per ottenere l'equazione (3.13), è sufficiente ricordare l'espressione della media condizionata:

$$\mathrm{E}\left[X - v | X > v\right] = \frac{1}{1 - F_X(v)} \int_v^\infty (X - v) f_X(x) \mathrm{d}x$$

Ricordando che ci si trova sopra la soglia u e quindi $f_x = f_{GPD}$ e $F_x = F_{GPD}$, si avrà:

$$\mathrm{E}\left[X - v | X > v\right] = \frac{1}{\left(1+\xi\frac{v-u}{\beta}\right)^{-1/\xi}} \int_v^\infty (x-v)\frac{1}{\beta}\left(1+\xi\frac{x-u}{\beta}\right)^{-1/\xi-1} \mathrm{d}x$$

$$= \frac{1}{\left(1+\xi\frac{v-u}{\beta}\right)^{-1/\xi}} \left[\frac{\xi(v-u)+\beta+v-x}{\xi-1}\left(1+\xi\frac{x-u}{\beta}\right)^{-1/\xi}\right]_v^\infty = \frac{\beta+\xi(v-u)}{1-\xi}$$

Il fatto importante, per il nostro scopo, è che la media degli eccessi è lineare rispetto alla soglia considerata. In questo senso, l'idea è quella di trovare una soglia, tale per cui, al variare di v al di sopra di essa, la media degli eccessi segua un andamento lineare.

In particolare, si supponga di disporre di una serie di osservazioni $x_1, \ldots, x_n$ da un campione $X_1, \ldots, X_n$ i.i.d.; denotiamo con $x_1^*, \ldots, x_{n_v}^*$ gli elementi dell'insieme $\{x : x \geq v\}$, con v generico. Il metodo qualitativo in considerazione consiste nel determinare un valore u al di sopra del quale il grafico di

$$\left\{v, \frac{1}{n_v}\sum_{j=1}^{n_v}\left(x_j^* - v\right)\right\}$$

abbia un andamento lineare. Questo perché la quantità appena presentata è la media empirica degli eccessi sopra una soglia v. Se si osserva un trend lineare all'aumento della soglia, significa che ci si trova al di sopra della soglia minima per cui X può essere approssimata con una GPD.

Si noti però che, per valori di v troppo alti, questo metodo non è efficace, in quanto le stime delle media possono essere distorte in quanto non si dispone di un numero di dati sufficiente oltre la soglia considerata. Per questo motivo, si cercherà un tratto lineare nel grafico, ma fino ad un certo punto, oltre al quale la media comincerà ad avere una variabilità molto elevata. Questo metodo grafico viene utilizzato in pratica, insieme ad altre stime di buon adattamento, per determinare la correttezza de modello scelto e dei parametri stimati.

Il secondo metodo che viene usato per determinare il valore corretto di u presuppone la stima degli altri due parametri β e ξ per diversi valori della soglia, ed osservarne l'andamento.

In particolare, si può sfruttare un'ulteriore proprietà della GPD: se X sopra la soglia u si può approssimare con una GPD di parametri (v, β', ξ'), allora, per ogni $v \geq u$, ogni $X|X \geq v$ si approssima con una GPD di parametri:

$$\beta' = \beta + \xi(v-u) \tag{3.14}$$

$$\xi' = \xi \tag{3.15}$$

Scrivendo infatti la funzione di ripartizione di X condizionata a $X \geq v$ si ha:

$$\begin{aligned} F(x|X \geq v) &= \frac{F_{GPD}(x) - F_{GPD}(v)}{1 - F_{GPD}(v)} \\ &= \frac{1 - \left(1 + \xi \frac{x-u}{\beta}\right)^{-1/\xi} - \left[1 - \left(1 + \xi \frac{v-u}{\beta}\right)^{-1/\xi}\right]}{\left(1 + \xi \frac{v-u}{\beta}\right)^{-1/\xi}} \\ &= 1 - \left[\frac{\beta + \xi(x-u)}{\beta + \xi(v-u)}\right]^{-1/\xi} \\ &= 1 - \left[1 + \xi \frac{x-v}{\beta + \xi(v-u)}\right]^{-1/\xi} \end{aligned}$$

che è proprio l'espressione della distribuzione GPD di parametri (v, β', ξ') espressi in precedenza.

Grazie a questa proprietà, quindi, il secondo metodo per determinare u prevede la stima dei parametri di scala e forma della GPD per diversi valori della soglia (con uno dei metodi che verranno descritti nelle sezioni successive). A questo punto, si potrà assumere valida l'approssimazione con la GPD sopra una soglia u tale per cui, variando $v \geq u$, la stima del parametro β è lineare in v (dalla (3.14)), mentre la stima di ξ rimane pressochè costante (dalla (3.15)). Anche questo metodo perde validità per valori di v troppo elevati, in quanto la variabilità degli stimatori aumenta a causa dei pochi dati a disposizione.

Una volta individuata la soglia u, si passa alla stima dei parametri della GPD: si presentano quindi nelle sezioni successive diverse tecniche per la stima. I test di buon adattamento per verificare la bontà della stima che vengono solitamente utilizzati nella teoria dei valori estremi sono gli stessi introdotti nell'ambito delle distribuzioni troncate (Kolmogorov-Smirnov e Anderson-Darling).

3.5.2 Metodo di massima verosimiglianza

Un primo metodo di stima che può essere utilizzato per fare inferenza sui parametri della GPD è il metodo di massima verosimiglianza (MLE). Come detto nel capitolo precedente, l'obiettivo è quindi scrivere la log-verosimiglianza del modello e trovare i parametri che la massimizzano.

Si consideri la soglia u fissata, e, coerentemente con la notazione precedente, $x_1^*, \ldots, x_k^*$ le osservazioni che superano la soglia. Per semplicità di

calcolo, d'ora in poi considereremo nelle analisi le quantità $y_i = x_i^* - u$, ovvero gli eccessi.

Supponendo che le variabile X_i che hanno generato le osservazioni x_i^* si possano approssimare con una GPD(u, β, ξ), le quantità y_j sono le osservazioni di Y_j, distribuite a loro una volta come una GPD di parametri $(0, \beta, \xi)$. Questo riscalamento è possibile in quanto basta porre il parametro di posizione, ovvero la soglia u, a 0.

La densità delle Y_i sarà quindi:

$$f(y; \beta, \xi) = \frac{1}{\beta}\left(1 + \xi\frac{y}{\beta}\right)^{-1/\xi - 1} \mathbb{1}_{[0,\infty)}(y)$$

dove il supporto è ricavato dalla (3.11), con restrizione al caso $\xi \geq 0$, come supposto in precedenza. Si ricorda che si dovrà avere anche $\beta > 0$.

La verosimiglianza del vettore di osservazioni $\boldsymbol{y} = (y_1, \ldots, y_n)$ sarà quindi data da:

$$L(\beta, \xi) = \prod_{j=1}^{k} f(y_i; \beta, \xi) = \prod_{j=1}^{k} \frac{1}{\beta}\left(1 + \xi\frac{y_j}{\beta}\right)^{-1/\xi - 1} \mathbb{1}_{[0,\infty)}(y_j)$$

e la relativa log-verosimiglianza (omettendo i termini indipendenti dai parametri):

$$l(\beta, \xi) = \log L(\beta, \xi) = -k \log \beta - \left(\frac{1}{\xi} + 1\right) \sum_{j=1}^{k} \log\left(1 + \xi\frac{y_j}{\beta}\right) \tag{3.16}$$

Derivando rispetto a β o a ξ e ponendo le derivate uguali a 0, non si ottengano soluzioni analitiche per i parametri MLE. Per questo motivo, sono necessari metodi di ottimizzazione numerica per trovare il massimo (eventualmente locale) della log-verosimiglianza. Inoltre, è stato osservato in [17] che in alcuni casi non si arriva a convergenza con gli algoritmi di ottimizzazione, specialmente se il numero di osservazioni non è sufficientemente alto.

A questo metodo, nell'ambito dei valori estremi, solitamente vengono preferiti i due successivi, in quanto non richiedono difficoltà computazionali come per gli MLE e sono dati da espressioni analitiche.

3.5.3 Metodo dei momenti

Un primo metodo alternativo agli stimatori MLE per i parametri β e ξ è quello dei momenti. L'idea è quella di trovare le espressioni dei momenti primo e secondo (in particolare, media e varianza) della GPD e porle

uguali a media e varianza campionarie. Da queste si ricavano le espressioni dei due parametri da cui dipendono i momenti, in funzione degli stimatori campionari di media e varianza.

In particolare, mantenendo la notazione precedente e chiamando Y una generica variabile GPD$(0, \beta, \xi)$, si deve risolvere, in funzione di β e ξ, il sistema:

$$\begin{cases} \overline{y} &= \mathrm{E}[Y] \\ s^2 &= \mathrm{Var}[Y] \end{cases} \tag{3.17}$$

dove $\overline{y}$ e s^2 sono la media e varianza campionaria delle osservazioni y_j, con $j = 1, \ldots, k$:

$$\begin{aligned} \overline{y} &= \frac{1}{k}\sum_{j=1}^{k} y_j \\ s^2 &= \frac{1}{k}\sum_{j=1}^{k} (y_j - \overline{y})^2 \end{aligned}$$

Per trovare l'espressione dei momenti della GPD si può dimostrare che vale la seguente formula:

$$\mathrm{E}\left[\left(1 + \xi\frac{Y}{\beta}\right)^r\right] = \frac{1}{1 - r\xi} \tag{3.18}$$

Essa vale solo per $1 - r\xi < 0$, ovvero $\xi > 1/r$.

Infatti:

$$\begin{aligned} \mathrm{E}\left[\left(1 + \xi\frac{Y}{\beta}\right)^r\right] &= \int_0^\infty \frac{1}{\beta}\left(1 + \xi\frac{y}{\beta}\right)^{-1/\xi - 1 + r} \mathrm{d}y \\ &= \left[\frac{1}{\xi}\frac{1}{-1/\xi + r}\left(1 + \xi\frac{y}{\beta}\right)^{-1/\xi + r}\right]_0^\infty \\ &= \frac{1}{1 - r\xi} \end{aligned}$$

A questo punto, la formula della media μ si trova facilmente con $r = 1$:

$$\mathrm{E}\left[1 + \xi\frac{Y}{\beta}\right] = 1 + \xi\frac{\mu}{\beta} = \frac{1}{1 - \xi}$$

da cui si ritrova la (3.12):

$$\mathrm{E}[Y] = \frac{\beta}{1 - \xi}.$$

Allo stesso modo, il momento secondo m^2 si determina ponendo $r = 2$:

$$\begin{aligned} \mathrm{E}\left[\left(1+\xi\frac{Y}{\beta}\right)^2\right] &= \mathrm{E}\left[1+2\xi\frac{Y}{\beta}+\left(\xi\frac{Y}{\beta}\right)^2\right] \\ &= 1+2\xi\frac{\mu}{\beta}+\xi^2\frac{m^2}{\beta^2} \\ &= 1+2\frac{\xi}{1-\xi}+\xi^2\frac{m^2}{\beta^2} \end{aligned}$$

da cui, ponendo questa quantità uguale a $1/(1-2\xi)$, si ottiene:

$$m^2 = \frac{2\beta^2}{(1-\xi)(1-2\xi)} \tag{3.19}$$

e, per la varianza σ^2:

$$\sigma^2 = m^2 - \mu^2 = \frac{2\beta^2}{(1-\xi)(1-2\xi)} - \left(\frac{\beta}{1-\xi}\right)^2 = \frac{\beta^2}{(1-\xi)^2(1-2\xi)} \tag{3.20}$$

Dalle espressioni di media e varianza, ponendole uguali alle stime campionarie come da (3.17) e risolvendo in funzione di β e ξ si ricavano quindi gli stimatori dei parametri con il metodo dei momenti, come segue:

$$\widehat{\beta} = \frac{1}{2}\overline{y}\left(\frac{\overline{y}^2}{s^2}+1\right) \tag{3.21}$$

$$\widehat{\xi} = \frac{1}{2}\left(1-\frac{\overline{y}^2}{s^2}\right). \tag{3.22}$$

Tuttavia, nella stima dei parametri della GPD, al metodo dei momenti viene generalmente preferito il metodo dei momenti generalizzato (Probability Weighted Moments), che verrà presentato nella prossima sezione.

3.5.4 Probability weighted moments

Un ultimo metodo che può essere utilizzato per stimare i parametri della GPD è quello dei *Probability Weighted Moments* (PWM). Esso, come vedremo, è un'estensione del metodo dei momenti, ed è stato introdotto in [15].

Sia dunque X una variabile aleatoria con funzione di ripartizione F. Si definiscono innanzitutto le quantità:

$$M_{p,r,s} = \mathrm{E}\left[X^p\{F(X)\}^r\{1-F(X)\}^s\right] \tag{3.23}$$

con $p, r, s \in \mathbb{R}$. Questi sono appunto i p-esimi momenti della distribuzione pesati con potenze della funzione di ripartizione e della funzione di sopravvivenza.

Il metodo consiste, seguendo l'idea del metodo dei momenti, nel'eguagliare queste quantità, funzioni dei parametri della distribuzione, ai momenti pesati empirici, per particolari valori di p, r e s. Per alcune distribuzioni, tra cui la GPD, il metodo dei PWM risulta più efficace rispetto metodo dei momenti classico (che coincide con il caso $M_{p,0,0}$).

Per la GPD, in [9] e [17] è stato mostrato come si possano trovare questi stimatori scegliendo $p = 1$ e $r = 0$, ovvero, riprendendo la notazione precedente, con $Y \sim \text{GPD}(0, \beta, \xi)$:

$$M_{1,0,s} = \text{E}\left[Y \left\{1 - F_{GPD}(Y)\right\}^s\right]$$

Si può trovare, tramite la definizione, un'espressione semplice per queste quantità, al variare di s:

$$\begin{aligned} M_{1,0,s} &= \int_0^\infty y \left(1 + \xi \frac{y}{\beta}\right)^{-s/\xi} \frac{1}{\beta} \left(1 + \xi \frac{y}{\beta}\right)^{-1/\xi - 1} \mathrm{d}y \\ &= \left[\frac{\beta + sy + y}{(s+1)(\xi - s - 1)} \left(1 + \xi \frac{y}{\beta}\right)^{-\frac{s+1}{\xi}}\right]_0^\infty \\ &= \frac{\beta}{(s+1)(1 + s - \xi)} \end{aligned} \tag{3.24}$$

Definendo quindi gli stimatori momenti generalizzati di questo tipo $\omega_{1,0,s}$, per trovare l'espressione delle stime di ξ e β, si dovrà risolvere:

$$\begin{cases} \omega_{1,0,0} &= \frac{\beta}{1-\xi} \\ \omega_{1,0,1} &= \frac{\beta}{2(2-\xi)} \end{cases} \tag{3.25}$$

È evidente come la prima delle due coincida con l'espressione del metodo dei momenti utilizzando il momento primo (la media).

Risolvendo il sistema in funzione di ξ e β si ottengono le formule per li stimatori PWM a partire dalle quantità empiriche:

$$\widehat{\beta} = \frac{2\omega_{1,0,0}\omega_{1,0,1}}{\omega_{1,0,0} - 2\omega_{1,0,1}} \tag{3.26}$$

$$\widehat{\xi} = 2 - \frac{\omega_{1,0,0}}{\omega_{1,0,0} - 2\omega_{1,0,1}} \tag{3.27}$$

Un aspetto fondamentale è che il parametro ξ stimato con il metodo PWM risulta < 1, portando quindi ad una media della GPD finita, aspetto

molto utile nelle applicazioni. Infatti lo stimatore si può riscrivere come segue:

$$\widehat{\xi} = \frac{\omega_{1,0,0} - 4\omega_{1,0,1}}{\omega_{1,0,0} - 2\omega_{1,0,1}}$$

Si verifica facilmente che esso è < 1, in quanto $\omega_{1,0,1}$ è per definizione positivo, in quanto media campionaria di quantità positive, come vedremo.

A questo punto, rimane da determinare l'espressione degli stimatori empirici dei momenti primi generalizzati $\omega_{1,0,0}$ e $\omega_{1,0,1}$. Per quanto riguarda $\omega_{1,0,0}$, è chiaro come esso coincida con la media campionaria degli eccessi $\overline{y}$, quindi:

$$\omega_{1,0,0} = \frac{1}{k}\sum_{j=1}^{k} y_j$$

Invece, per determinare ω_1, l'idea è di determinare lo stimatore come segue:

$$\omega_{1,0,1} = \frac{1}{k}\sum_{j=1}^{k} y_j \left[1 - F_{GPD}(y_j)\right]$$

Si denoti quindi con $y_{(1)}, \ldots, y_{(k)}$ le osservazioni degli eccessi disposti in ordine crescente. A questo punto, si può sfruttare il fatto che le quantità $\left[1 - F_{GPD}(y_{(j)})\right]$ si possono considerare come osservazioni di una uniforme in $[0, 1]$.

Ciò si ottiene facilmente determinando la funzione di ripartizione di $Z = F(X)$, con X variabile aleatoria con funzione di ripartizione F:

$$F_Z(z) = \mathrm{P}\left[Z \leq z\right] = \mathrm{P}\left[F(X) \leq z\right] = \mathrm{P}\left[X \leq F^{-1}(z)\right] = F\left(F^{-1}(z)\right) = z, \tag{3.28}$$

che è proprio la funzione di ripartizione di un'uniforme. Lo stesso discorso vale ovviamente anche per la funzione di sopravvivenza, in quanto se U è uniforme, lo è anche $1 - U$.

Perciò, si può fare la seguente approssimazione:

$$F_{GPD}(y_{(j)}) \approx \frac{j}{k-1} \qquad \text{per } j = 1, \ldots, k$$

dove i termini a destra rappresentano gli stimatori dei quantili dell'uniforme per k osservazioni. Quindi:

$$1 - F_{GPD}(y_{(j)}) \approx \frac{k-j}{k-1} \qquad \text{per } j = 1, \ldots, k$$

A questo punto, lo stimatore $\omega_{1,0,1}$ per $M_{1,0,1}$ è dato dalla seguente espressione:

$$\omega_{1,0,1} = \frac{1}{k} \sum_{j=1}^{k} \frac{k-j}{k-1} y_{(j)} \tag{3.29}$$

Sia questo che lo stimatore della media sono non distorti.

Un metodo alternativo per determinare gli stimatori dei momenti pesati è stato proposto in [21], per cui l'espressione degli $\omega_{1,0,1}$ è data da:

$$\widetilde{\omega}_{1,0,1} = \frac{1}{j} \sum_{j=1}^{k} (1 - p_{j,k}) y_{(j)}$$

con

$$p_{j,k} = \frac{j+\gamma}{k+\delta}$$

dove γ e δ sono delle oppurtune costanti. La scelta che viene fatta in letteratura è $\gamma = -0.35$ e $\delta = 0$, in modo da ottenere un'approssimazione migliore degli stimatori dei momenti generalizzati.

In ogni caso, grazie a queste espressioni dei quantili empirici della distribuzione, dalle (3.26) si possono trovare le espressioni degli stimatori PWM per i parametri della GPD. Un vantaggio di questo metodo, rispetto ai precedenti, è che si ottengono delle stime adeguate anche in presenza di pochi dati, come solitamente accade per gli eventi estremi, che hanno bassa frequenza. Inoltre, questi stimatori appaiono più robusti al variare della soglia u.

Le proprietà dei diversi stimatori sono state discusse in [17]. In tale lavoro è stato osservato che, per $\xi < 0$, il metodo più efficace è quello dei momenti. Se, come nel nostro caso, ξ risulta positivo, il metodo PWM dev'essere preferito sia al metodo dei momenti che al metodo di massima verosimiglianza, che, considerando anche i problemi computazionali che implica, appare giustificato solo per $\xi < 0.2$, ipotesi troppo restrittiva nel caso della modellizzazione delle perdite operative, dove si preferiscono valori di ξ più elevati che, come osservato in precedenza, implicano una distribuzione più concentrata verso le code destre.

3.6 Test di buon adattamento

Il pacchetto Tutti i metodi descritti sono implementati con il software R nel pacchetto [30]. Esso, oltre alla stima dei parametri, permette di verificare qualitativamente la bontà del fit attraverso metodi grafici. In particolare, dopo aver stimato i parametri della distribuzione GPD con gli

eccessi rispetto alla soglia u, si analizza la distribuzione dei residui, osservandone lo scatterplot (si richiede assenza di trend dei residui) e il Q-Q plot, considerando come distribuzione teorica l'esponenziale [30]. Si vedrà l'utilizzo pratico di tali metodi nel Capitolo 5.

È possibile anche considerare il Q-Q plot, introdotto nel Capitolo 2), considerando ovviamente come distribuzione teorica la GPD e confrontando i diversi metodi di stima, per individuare il più adatto.

Altri metodi di confronto per supportare la scelta di un metodo di stima sfruttano le proprietà dei parametri della GPD al variare della soglia (introdotte nella paragrafo 3.5.1). In particolare, si preferirà il metodo di stima per cui si osserva un andamento costante del parametro ξ e lineare per quanto riguarda β. Come si vedrà, l'unico parametro che si utilizzerà nella simulazione è il parametro di forma ξ, e quindi maggiore attenzione sarà rivolta ad esso.

Tali metodi qualitativi devono essere portati avanti parallelamente a metodi quantitativi. In particolare, si utilizzano test di bontà del fit quali Kolmogorov-Smirnov e Anderson-Darling introdotti nella sezione 2.6 nel caso di distribuzioni troncate.

Capitolo 4

Copule

Le copule rappresentano una metodologia utile nella modellizzazione della dipendenza e le relazioni tra variabili casuali, per determinare la distribuzione congiunta del relativo vettore aleatorio. Il nome è stato per la prima volta introdotto in da Sklar in [27] e deriva dal termine latino *copulare*, ovvero connettere, unire.

In generale, ogni distribuzione multivariata contiene informazioni sulle marginale e sulla struttura di dipendenza tra di esse. Le copule vengono utilizzate per isolare la descrizione della struttura di dipendenza tra le variabili aleatorie e le loro leggi monodimensionali. Esse risultano particolarmente utili nel momento in cui le marginali sono definite empiricamente (ovvero da simulazione), come nel caso dei rischi operativi, in cui la distribuzione di ogni classe di rischio è ottenuta dalla convoluzione tramite metodi Monte Carlo tra frequency e severity.

L'approccio di costruzione di un vettore aleatorio tramite copule è del tipo *"bottom-up"*: si vuole determinare la distribuzione congiunta a partire dalle marginali e dalla struttura di correlazione tra di esse, ed è opposto all'approccio delle distribuzioni congiunte notevoli, da cui si va a determinare le marginali a partire dalla distribuzione congiunta.

In questo capitolo si introdurrà quindi il concetto di copula, insieme alle proprietà fondamentali. Si presenteranno quindi le principali misure di correlazione utilizzate in quest'ambito, e le famiglie di copule principali. Infine, l'interesse sarà sulla stima dei parametri delle copule e i test di buon adattamento ai dati.

4.1 Definizione e proprietà principali

Una copula d-dimensionale C è definita come una distribuzione multivariata con marginali uniformemente distribuite in $[0,1]$.

Si denoterà d'ora in poi la funzione di ripartizione di una generica copula con $C(\mathbf{u}) = C(u_1, u_2, \ldots, u_d)$. Dunque, C è una funzione

$$C : [0,1]^d \to [0,1].$$

Dalla definizione di copula conseguono le seguenti proprietà [26]:

1. C è non decrescente in ogni componente u_i.
2. Le marginali C_i sono tali che

   $$C_i(\mathbf{u}) = C(1, \ldots, 1, u, 1, \ldots, 1) = u \qquad \forall u \in [0,1]$$

   dove l'unico valore diverso da 1 è in corrispondenza della posizione i-esima.
3. Per ogni $\mathbf{a}, \mathbf{b} \in [0,1]^d$, con $a_i \leq b_i$, vale:

   $$\sum_{i_1 \in \{1,2\}} \cdots \sum_{i_d \in \{1,2\}} (-1)^{i_1 + \cdots + i_d} C\left(u_{1,i_1}, \ldots, u_{d,i_d}\right) \geq 0 \qquad (4.1)$$

   dove $u_{j,1} = a_j$ e $u_{j,2} = b_j$.

La prima e la seconda proprietà derivano dal fatto che le marginali di C sono distribuzioni di probabilità (e quindi non decrescenti), ed in particolare uniformi. La terza proprietà, esplicitata con la disuguaglianza rettangolare (4.1), garantisce che la quantità $\mathrm{P}[a_1 \leq U_1 \leq b_1, \ldots, a_d \leq U_d \leq b_d]$ sia non negativa, dove $\mathbf{U}$ è il vettore aleatorio con densità di probabilità C. Si noti inoltre che una qualunque funzione C che rispetta le propietà appena elencate è una copula.

Infine, si osserva che $C(1, u_1, \ldots, u_{d-1})$ è ancora una copula, e allo stesso modo lo sono tutte le marginali k-dimensionali, con $2 \leq k \leq d$. Per questo motivo, nelle trattazioni teoriche, è possibile concentrarsi sulle copule bivariate, per poi estendere il risultato al caso d-dimensionale.

4.1.1 Legame con le distribuzioni marginali

Il risultato principale nell'ambito delle copule è il Teorema di Sklar, che dimostra come tutte le distribuzioni multivariate siano copule e, viceversa,

come sia possibile usare le copule insieme alle marginali per determinare la distribuzione congiunta.

Prima di arrivare all'enunciato, è necessario ricordare la proprietà riassunta nella (3.28): se X è una variabile aleatoria con funzione di ripartizione F, allora la variabile $Y = F^{-1}(X)$ ha distribuzione uniforme in $[0,1]$. Si assumerà d'ora in poi l'esistenza dell'inversa delle funzioni di ripartizione.

Teorema 4.1. (Sklar) *Sia F una funzione di ripartizione di un vettore aleatorio, con marginali $F_1, \ldots, F_d$. Allora esiste una copula $C : [0,1]^d \to [0,1]$ tale che, per ogni $x_1, \ldots, x_d$:*

$$F(x_1, \ldots, x_d) = C\left(F_1(x_1), \ldots, F_d(x_d)\right). \tag{4.2}$$

Se le marginali sono continue, allora C è unica; altrimenti C è univocamente determinata su $RanF_1 \times RanF_1 \times \cdots \times RanF_d$, dove $RanF_i$ denota l'intervallo di variabilità di F_i.

Viceversa, se C è una copula e $F_1, \ldots, F_d$ funzioni di ripartizione univariate, allora F definita nella (4.2) è funzione di ripartizione multivariata con marginali $F_1, \ldots, F_d$.

Per una dimostrazione dettagliata si rimanda a [25]. Limitandosi al caso di funzioni di ripartizioni continue e invertibili, si può osservare che

$$F(x_1, \ldots, x_1) = \mathrm{P}\left[F_1(X_1) \leq F_1(x_1), \ldots, F_d(X_d) \leq F_d(x_d)\right].$$

Essendo le $F_i(X_i)$ distribuzioni uniformi, l'espressione appena trovata coincide con la definizione di copula.

Valutando invece la (4.2) in $x_i = F_i^{-1}(u_i)$, si ottiene:

$$C(\mathbf{u}) = F\left(F_1^{-1}(u_1), \ldots, F_d^{-1}(u_d)\right), \tag{4.3}$$

che da una rappresentazione esplicita della copula C in funzione delle marginali e della cumulata, ed è quindi unica.

In ogni caso, il Teorema di Sklar dimostra la corrispondenza biunivoca tra copule e funzioni di ripartizione multivariate. L'espressione (4.2) mostra come si possano formare distribuzioni congiunte a partire dalle marginali attraverso la copula C. Viceversa, la (4.3) evidenzia come le copule possano essere estratte da distribuzioni multivariate con marginali continue e invertibili. Un altro concetto importante di questa espressione è che le copule esprimono la dipendenza tra le variabili aleatorie in termini di quantili, il che sarà utile nel seguito,quando si andrà ad analizzare le strutture di dipendenza.

In definitiva, è possibile definire la nozione di copula di una generica distribuzione multivariata: se un vettore aleatorio $\mathbf{X}$ ha distribuzione congiunta F e marginali $F_1, \ldots, F_d$, allora la copula di F è la funzione di ripartizione C di $(F_1(X_1), \ldots, F_d(X_d))$.

4.1.2 Esempi di copule

Per comprendere il significato pratico delle copule, si considerano due esempi base di copule, che tuttavia non vengono quasi mai usate nelle applicazioni, in quanto si basano su ipotesi troppo restrittive.

La copula più immediata da considerare è la *copula di indipendenza*, che si definisce come segue:

$$\Pi(\mathbf{u}) = \prod_{i=1}^{d} u_i \tag{4.4}$$

Essa si basa sull'ipotesi di indipendenza tra le variabili aleatorie per cui si vuole trovare la struttura multivariata. Considerando quindi, con la notazione utilizzata precedentemente, la funzione di ripartizione multivariata del vettore $\mathbf{X}$, si ha:

$$F(x_1, \ldots, x_d) = \Pi(F_1(x_1), \ldots, F_d(x_d)) = \prod_{i=1}^{d} F_i(x_i). \tag{4.5}$$

Si può quindi osservare che le componenti del vettore $\mathbf{X}$ sono indipendenti se e solo se la copula del vettore è data dalla (4.4).

Un altro esempio di copula notevole è la *copula comonotona*, definita come:

$$M(\mathbf{u}) = \min\{u_1, \ldots, u_d\} \tag{4.6}$$

Si può facilmente notare come questa sia la funzione di ripartizione congiunta del vettore $(U, \ldots, U)$, dove $U \sim U(0,1)$. Infatti, questo tipo di copula viene utilizzato nel caso di perfetta dipendenza (positiva) tra le variabili aleatorie $X_1, \ldots, X_d$. Ovvero, si può scrivere $X_i = T_i(X_1)$, dove T_i sono funzioni crescenti. Quindi, le funzioni di ripartizione delle singole variabili si possono scrivere come

$$F_i(X_i) = F_1 \circ T_i^{-1}(X_1) = F_1 \circ T_i^{-1} \circ T_i(X_1)$$

dove il simbolo $\circ$ indica la composizione delle due funzioni. Sotto queste ipotesi, la copula del vettore X sarà:

$$F(x_1, \ldots, x_d) = M(F_1(x_1), \ldots, F_d(x_d)) = \min\{F_1(x_1), \ldots, F_d(x_d)\} \tag{4.7}$$

4.1.3 Densità e distribuzioni condizionali delle copule

Avendo identificato le copule come funzioni di ripartizione multivariate, è naturale introdurre la densità delle copule e le funzioni di ripartizioni marginali condizionate.

Supponendo che C sia sufficientemente derivabile, si definisce quindi la densità della copula:

$$c(\mathbf{u}) = \frac{\partial^d C(u_1, \ldots, u_d)}{\partial u_1 \cdots \partial u_d} \tag{4.8}$$

Scrivendo la copula nella forma (4.3), è inoltre possibile esprimere la densità in funzione delle densità marginali f_i e della congiunta f. Sfruttando quindi le regole di derivazione di funzioni composte e delle inverse si ottiene:

$$\begin{aligned} c(\mathbf{u}) &= \frac{\partial^d F\left(F_1^{-1}(u_1), \ldots, F_d^{-1}(u_d)\right)}{\partial u_1 \cdots \partial u_d} \\ &= f\left(F_1^{-1}(u_1), \ldots, F_d^{-1}(u_d)\right) \frac{\mathrm{d}F_1^{-1}(u_1)}{\mathrm{d}u_1} \cdots \frac{\mathrm{d}F_d^{-1}(u_d)}{\mathrm{d}u_d} \\ &= \frac{f\left(F_1^{-1}(u_1), \ldots, F_d^{-1}(u_d)\right)}{f_1(F_1^{-1}(u_1)) \cdots f_d(F_d^{-1}(u_d))}. \end{aligned} \tag{4.9}$$

A partire dalla definizione della copula stessa, è anche possibile definire le distribuzioni condizionali. Considerando il vettore (U_1, U_2) con copula bidimensionale C, si può scrivere la distribuzione di $U_2|U_1 = u_1$ come segue:

$$\begin{aligned} \mathrm{P}\left[U_2 \leq u_2 | U_1 = u_1\right] &= \lim_{\delta \to 0} \frac{\mathrm{P}\left[U_2 \leq u_2, u_1 - \delta \leq U_1 \leq u_1 + \delta\right]}{\mathrm{P}\left[u_1 - \delta \leq U_1 \leq u_1 + \delta\right]} \\ &= \lim_{\delta \to 0} \frac{C(u_1 + \delta, u_2) - C(u_1 + \delta, u_2)}{2\delta} \\ &= \frac{\partial}{\partial u_1} C(u_1, u_2) \end{aligned} \tag{4.10}$$

Dunque, le funzioni di ripartizione condizionate possono essere calcolate a partire dall'espressione della copula stessa.

4.2 Misure di correlazione

Come accennato in precedenza, le copule si ottengono a partire dalle distribuzioni marginali e dalla struttura di dipendenza. L'attenzione di ques-

ta sezione è rivolta a quest'ultimo concetto: si introdurranno quindi le principali misure di dipendenza tra variabili aleatorie. In particolare, si presenteranno la correlazione lineare e misure basate sul rango dei dati (Tau di Kendall e Rho di Spearman). Infine, ci si soffermerà sul concetto di dipendenza tra le code delle variabili aleatorie nell'ambito delle copule.

4.2.1 Correlazione lineare

Il concetto di correlazione lineare è l'indicatore tipico di dipendenza tra variabili aleatorie. In particolare, date due variabili aleatorie X e Y, la correlazione lineare $\mathrm{Cor}(X, Y)$ è una quantità scalare compresa tra -1 e 1 che misura la dipendenza lineare tra di esse. Essa si definisce come:

$$\rho_{X,Y} = \mathrm{Corr}(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}[X]\mathrm{Var}[Y]}} \tag{4.11}$$

dove $\mathrm{Cov}(X, Y) = \mathrm{E}[XY] - \mathrm{E}[X]\mathrm{E}[Y]$ è la covarianza tra le due variabili aleatorie.

In particolare, se le due variabili sono indipendenti, allora la correlazione lineare tra le due è nulla, mentre il viceversa è valido unicamente per distribuzioni gaussiane. Quindi, tale coefficiente è in grado di identificare unicamente correlazioni lineari tra le variabili. Ciò è un primo limite del suo utilizzo nello sviluppo di una copula appropriata per il modello multivariato.

In ogni caso, per stimare il coefficiente di correlazione lineare a partire da $x_1, \ldots, x_k$ e $y_1, \ldots, y_k$, osservazioni indipendenti e identicamente distribuite delle variabili X e Y, si utilizza la seguente espressione:

$$\widehat{\rho}_{X,Y} = \frac{\sum_{i=1}^{k}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{k}(x_i - \bar{x})^2 \sum_{i=1}^{k}(y_i - \bar{y})^2}} \tag{4.12}$$

dove $\bar{x}$ e $\bar{y}$ rappresentano le medie campionarie delle due variabili.

Per le caratteristiche appena sottolineate, nello sviluppo delle copule per aggregare distribuzioni non gaussiane (come nel caso delle perdite derivanti da rischi operativi) vengono utilizzati altre misure di dipendenza, basate sul rango, che verranno ora illustrate.

4.2.2 Tau di Kendall

Come accennato, l'uso della correlazione lineare è limitante nell'ambito delle copule. Le misure del grado di relazione tra due variabili solita-

mente usate si basano invece sul rango dei dati, ovvero sulla loro cardinalità nel vettore di osservazioni ordinato in ordine crescente. Dato un vettore $x_1, \ldots, x_n$, il rango della i-esima osservazione è definito come:

$$\text{rank}(x_i) = \#\, osservazioni \leq x_i = \sum_{j=1}^{d} \mathbb{1}_{[-\infty, x_i)}(x_j) \tag{4.13}$$

assumendo per semplicità che non esistano ripetizioni tra i dati.

Le misure basate sui ranghi sono invarianti rispetto alla scala utilizzata, il che è molto utile nella modellizzazione di distribuzioni multivariate con le copule. Una prima misura di correlazione per ranghi è il *Tau di Kendall* (τ_K).

Per arrivare alla definizione, si considerino due variabili X' e Y', indipendenti da X e Y e identicamente distribuite ad esse. L'idea è che se esiste correlazione positiva, il prodotto $(X - X')(Y - Y')$ avrà segno positivo, ovvero le due variabili sono concordanti (all'aumentare dell'una si ha un aumento dell'altra), mentre in caso di correlazione negativa sarà minore di zero. Proprio per il fatto che il coefficiente si basa sui segni del prodotto e non sul suo valore assoluto, il coefficiente dipende solo dai ranghi delle osservazioni, e non dal loro vero valore.

Il Tau di Kendall si definisce quindi come segue:

$$\tau_K(X, Y) = \mathrm{E}\left[\text{sign}\left\{(X - X')(Y - Y')\right\}\right] \tag{4.14}$$

Tale espressione si può riscrivere come:

$$\tau_K(X, Y) = \mathrm{P}\left[(X - X')(Y - Y') > 0\right] - \mathrm{P}\left[(X - X')(Y - Y') < 0\right]. \tag{4.15}$$

Da essa si osserva intuitivamente che il Tau di Kendall sarà nullo se le due probabilità sono uguali, ovvero se non esiste un segno ben definito per il prodotto in considerazione, ovvero non c'è concordanza o discordanza tra le variabili.

Per quanto riguarda la stima di questo coefficiente dai dati, mantenendo la notazione usata in precedenza, si utilizza la seguente espressione:

$$\widehat{\tau}_K = \frac{(\#\, coppie\ concordanti) - (\#\, coppie\ discordanti)}{\frac{1}{2}n(n-1)} \tag{4.16}$$

dove il denominatore rappresenta il numero di coppie totali.

Una prima proprietà interessante dello stimatore del Tau di Kendall [19] è il fatto che, quando il numero n di osservazione delle coppie di variabili è superiore a 10, esso può essere approssimato con una normale di

media pari al valore vero di τ_k e varianza:

$$\sigma^2_{\widehat{\tau_K}} = \frac{2(2n+5)}{9n(n-1)} \tag{4.17}$$

Ciò è utile innanzitutto per testare la dipendenza tra variabili aleatorie, conoscendo la distribuzione dello stimatore, e, nell'ambito della modellizzazione con le copule, è possibile effettuare dei test di "stress" su τ_K (prendendone ad esempio un percentile elevato) e osservare l'effetto sull'aggregazione tra le variabili casuali.

Sebbene l'interesse maggiore nell'ambito della modellizzazione con le copule sia arrivare a definire la struttura multivariata a partire dalla correlazione e dalle marginali, è anche possibile ricavare il Tau di Kendall a partire dalla copula. Supponendo infatti che X e Y abbiano legge congiunta data dalla copula C, allora:

$$\tau_K(X,Y) = 4\int_0^1\int_0^1 C(u_1,u_2)\mathrm{d}C(u_1,u_2) - 1. \tag{4.18}$$

Per la dimostrazione, riprendendo la definizione alternativa di τ_k (4.15), si ha che:

$$\begin{aligned}
\tau_K(X,Y) &= \mathrm{P}\left[(X-X')(Y-Y')>0\right] - \mathrm{P}\left[(X-X')(Y-Y')<0\right] \\
&= 2\,\mathrm{P}\left[(X-X')(Y-Y')>0\right] - 1 \\
&= 2\left\{\mathrm{P}\left[X>X', Y>Y'\right] + \mathrm{P}\left[X>X', Y>Y'\right]\right\} - 1 \\
&= 4\,\mathrm{P}\left[X>X', Y>Y'\right] - 1
\end{aligned} \tag{4.19}$$

sfruttando il fatto che le coppie (X,Y) e (X',Y') hanno la stessa legge. Inoltre, indicando con F e G le funzioni di ripartizione marginali di X e Y (e quindi di X' e Y') e ricordando che la copula C è la funzione di ripartizione del vettore:

$$\begin{aligned}
\mathrm{P}\left[X>X', Y>Y'\right] &= \int\int_{\mathbb{R}^2} \mathrm{P}\left[X'<x, Y'<y\right]\mathrm{d}C(F(x),G(y)) \\
&= \int\int_{\mathbb{R}^2} C(F(x),G(y))\mathrm{d}C(F(x),G(y)).
\end{aligned}$$

Con queste considerazioni e le trasformazioni $u_1 = F(x)$ e $u_2 = G(y)$ si arriva alla (4.18).

4.2.3 Rho di Spearman

Un altra misura di correlazione per ranghi, che viene usata alternativamente a τ_K, è il Rho di Spearman (ρ_S). Essa è fortemente legata alle copule,

in quanto misura la correlazione tra i livelli dei quantili corrispondenti alle osservazioni.

Siano infatti X e Y due variabili aleatorie continue con funzioni di ripartizione (supposte invertibili) F e G. Allora, il coefficiente di correlazione di Spearman è definito come segue:

$$\rho_S = \text{Corr}\left(F(X), G(Y)\right) \tag{4.20}$$

Anche in questo caso, un valore nullo del coefficiente indica indipendenza.

Per il calcolo di ρ_S a partire dai dati, si sfrutta ancora una volta la proprietà (3.28), per cui le funzioni di ripartizione valutate nelle osservazioni della variabile hanno distribuzione uniforme. Quindi, date le osservazioni $x_1, .., x_d$ e $y_1, \ldots, y_d$, il coefficiente si calcola a partire dalla (4.11), utilizzando al posto delle osservazioni i ranghi definiti nella (4.13). Siano quindi $r_i = \text{rank}(x_i)$ e $s_i = \text{rank}(y_i)$ per $i = 1, \ldots, n$. Allora:

$$\widehat{\rho_S}(X, Y) = \frac{\sum_{i=1}^{k}(r_i - \bar{r})(s_i - \bar{s})}{\sqrt{\sum_{i=1}^{d}(r_i - \bar{r})^2 \sum_{i=1}^{d}(s_i - \bar{s})^2}}, \tag{4.21}$$

dove $\bar{r}$ e $\bar{s}$ sono le medie dei ranghi delle osservazioni, pari a $n(n+1)/2$ nel caso in cui non ci siano osservazioni con lo stesso valore.

Il Rho si Spearman, così come τ_K, può essere ricavato dalla distribuzione della copula:

$$\rho_S(X, Y) = 12 \int_0^1 \int_0^1 C(u_1, u_2)\mathrm{d}u_1\mathrm{d}u_2 - 3 = 12 \int_0^1 \int_0^1 u_1, u_2 \mathrm{d}C(u_1, u_2) - 3. \tag{4.22}$$

Per mostrare questo legame, si consideri innanzitutto che $F(X)$ e $G(X)$ hanno distribuzione uniforme, quindi di media e varianza rispettivamente pari a $1/2$ e $1/12$. Allora

$$\begin{aligned} \rho_S(X, Y) &= \frac{\text{E}\left[F(X)G(Y)\right] - \text{E}\left[F(X)\right]\text{E}\left[G(Y)\right]}{\sqrt{\text{Var}[F(X)]\text{Var}[F(Y)]}} \\ &= 12\,\text{E}\left[F(X)G(Y)\right] - 3 \\ &= 12 \int \int_{\mathbb{R}^2} F(x)G(y)\mathrm{d}C(F(x), G(y)) - 3, \end{aligned} \tag{4.23}$$

da cui si ottiene la (4.22).

Come vedremo, le copule ellittiche (vedi sezione 4.3) si adattano molto bene all'utilizzo nei modelli di previsione dei rischi operativi, in quanto sono in grado di modellizzare dipendenze di coda differenti per ogni coppia di variabili che compongono il vettore multivariato.

4.2.4 Tail dependence

Un'altra misura della dipendenza legata alle copule è la cosiddetta *tail dependence* (dipendenza di coda). Essa viene usata per osservare la correlazione tra coppie di variabili aleatorie nelle code della relativa distribuzione, ovvero per valori estremi, molto distanti dalla media.

Nel caso di correlazione dei valori delle code sinistre delle distribuzioni, la dipendenza di coda inferiore si trova calcolando la probabilità che la variabile Y sia minore del q-esimo quantile, condizionata al fatto che X è inferiore al relativo quantile e considerandone poi il limite di q che tende a 1. Ovvero, mantenendo la notazione precedente:

$$\lambda_l(X,Y) = \lim_{q\to 0^+} \mathrm{P}\left[Y < G^{-1}(q) | X < F^{-1}(q)\right]. \tag{4.24}$$

supponendo che il limite esista. Ovviamente, è possibile scambiare X e Y nella definizione. Quindi, se $\lambda_l \in (0,1]$ le due variabili hanno dipendenza nella coda inferiore, mentre se il limite è zero esse sono asintoticamente indipendenti nella coda inferiore.

Una definizione analoga si ha per la dipendenza nella coda superiore:

$$\lambda_u(X,Y) = \lim_{q\to 1^-} \mathrm{P}\left[Y > G^{-1}(q) | X > F^{-1}(q)\right] \tag{4.25}$$

L'interesse principale nell'ambito della gestione dei rischi operativi va nella correlazione delle code superiori delle distribuzioni, in quanto influisce maggiormente sul capitale da allocare, individuato tramite il VaR, che è appunto un quantile elevato della distribuzione congiunta delle classi di rischio.

L'espressione delle correlazioni di coda può essere riscritta in funzione della copula bivariata C. Per quanto riguarda la coda inferiore:

$$\begin{aligned} \lambda_l(X,Y) &= \lim_{q\to 0^+} \frac{\mathrm{P}\left[Y < G^{-1}(q), X < F^{-1}(q)\right]}{\mathrm{P}\left[X < F^{-1}(q)\right]} \\ &= \lim_{q\to 0^+} \frac{C(q,q)}{q}. \end{aligned} \tag{4.26}$$

Analogamente, per λ_u:

$$\begin{aligned}\lambda_u(X,Y) &= \lim_{q\to 1^-} \frac{\mathrm{P}\left[Y > G^{-1}(q), X > F^{-1}(q)\right]}{\mathrm{P}\left[X > F^{-1}(q)\right]} \\ &= \lim_{q\to 1^-} \frac{1-\mathrm{P}\left[X \le F^{-1}(q) \cup Y \le G^{-1}(q)\right]}{1-\mathrm{P}\left[X \le F^{-1}(q)\right]} \\ &= \lim_{q\to 1^-} \frac{1-2q+C(q,q)}{1-q}. \end{aligned} \tag{4.27}$$

Si noti infine che per calcolare λ_l e λ_u tra due variabili specifiche appartenenti a una copula d-dimensionale, è sufficiente usare le espressioni (4.26) e (4.27) sostituendo con il valore 1 i quantili delle distribuzioni marginali delle restanti $d-2$ variabili.

Le correlazioni di coda, come le misure di correlazione per ranghi, dipendono solo sulla struttura della copula. Tuttavia, queste quantità non sono stimabili esplicitamente dai dati, quindi non vengono usate per la costruzione della copula, bensì per valutarne la bontà del fit. L'idea è quella di confrontare i valori di dipendenza di coda con il comportamento agli estremi della distribuzione empirica. Nella pratica, ad esempio per la coda superiore, si fissa un quantile sufficientemente alto e si considera la percentuale di coppie di osservazioni in cui entrambe le variabili sono superiori al quantile prescelto, confrontando questa quantità con il valore teorico, che dipende dalla struttura multivariata scelta. Alternativamente, variando il livello del quantile, si valuta la convergenza di questa frequenza empirica a zero.

4.3 Copule ellittiche

La prima famiglia di copule che si andrà a considerare è quella delle copule *ellittiche*, ottenute a partire da distribuzioni ellittiche, ovvero ottenute da una generalizzazione della distribuzione normale multivariata [25]. Di questa famiglia si andranno a trattare la copula gaussiana e la copula t di Student, le più interessanti dal punto di vista dei rischi operativi. Questa classe è molto utile nelle applicazioni, in quanto permette di simulare facilmente dalla distribuzione multivariata, ed inoltre permette di mantenere la struttura di correlazione dei dati.

Per quanto riguarda la struttura di correlazione, sia per la copula gaussiana che per la t di Student è possibile determinare un legame tra le misure di correlazione per ranghi e la correlazione lineare. Se infatti la distribuzione congiunta del vettore $X_1, \ldots, X_d$ appartiene alla famiglia delle

copule ellittiche, valgono le seguenti relazioni:

$$\tau_K(X_i, X_j) = \frac{2}{\pi} \arcsin R_{ij} \tag{4.28}$$

$$\rho_S(X_i, Y_j) = \frac{6}{\pi} \arcsin \frac{R_{ij}}{2} \tag{4.29}$$

dove $R_{ij} = \mathrm{Corr}(X_i, X_j)$. R è quindi la matrice di correlazione del vettore, che come vedremo è un parametro di questa classe di copule. Si mostreranno solo tracce di dimostrazione limitatamente al caso gaussiano; per una trattazione dettagliata si rimanda a [10].

Per quanto riguarda il Tau di Kendall, sia (Z_1, Z_2) è un gaussiano con correlazione ρ. È possibile passare alla coordinate sferiche, come segue:

$$(Z_1, Z_2) = R\left(\cos\theta, \rho\cos\theta + \sqrt{1-\rho^2}\sin\theta\right)$$

dove R è una variabile positiva (di cui, come si vedrà, non interessa la distribuzione in questo ambito), indipendente da $\theta \sim U[-\pi, \pi]$. Quindi, indicando con $\psi = \arcsin\rho$, vale:

$$\begin{aligned} \mathrm{P}\left[Z_1 > 0, Z_2 > 0\right] &= \mathrm{P}[\cos\theta > 0, \rho\cos\theta + \sqrt{1-\rho^2}\sin\theta > 0] \\ &= \mathrm{P}[\cos\theta > 0, \sin\psi\cos\theta + \cos\psi\sin\theta > 0] \\ &= \mathrm{P}[\cos\theta > 0, \cos(\theta + \psi) > 0] \\ &= \frac{1/2\pi + \psi}{2\pi} \end{aligned} \tag{4.30}$$

L'ultimo passaggio si verifica ricordando la distribuzione uniforme di θ, unica variabile in gioco. A questo punto, ricordando la (4.19):

$$\tau_K(X, Y) = 4\,\mathrm{P}\left[X - X' > 0, Y - Y' > 0\right] - 1$$

si osserva che il vettore $(X - X', Y - Y')$ è normale con matrice di correlazione $2R$. Quindi, standardizzando, la correlazione lineare tra le due variabili è ρ, da cui la (4.28).

Considerando il Rho si Spearman, ricordando la (4.22), si ha:

$$\begin{aligned} \rho_S(X, Y) &= 12\int_0^1\int_0^1 \mathrm{P}\left[\Phi(X) \le u_1, \Phi(Y) \le u_2\right] \mathrm{d}u_1\mathrm{d}u_2 - 3 \\ &= 12\int_0^1\int_0^1 \mathrm{P}\left[X \le \Phi^{-1}(u_1), Y \le \Phi^{-1}(u_2)\right] \mathrm{d}u_1\mathrm{d}u_2 - 3 \\ &= 12\int_0^1\int_0^1 \mathrm{P}\left[X \le x, Y \le y\right] \Phi(x)\Phi(y)\mathrm{d}x\mathrm{d}y - 3 \end{aligned}$$

dove Φ è funzione di ripartizione della normale standard, $x = \Phi^{-1}(u_1)$ e $x = \Phi^{-1}(u_2)$. Introducendo quindi due variabili Z_1 e Z_2, normali standard, indipendenti tra di loro e da X e Y, si può osservare che:

$$\begin{aligned}\rho_S(X,Y) &= 12\,\mathrm{P}\left[X \leq Z_1, Y \leq Z_2 | Z_1, Z_2\right] - 3 \\ &= 12\,\mathrm{P}\left[X \leq Z_1, Y \leq Z_2\right] - 3 \\ &= 12\,\mathrm{P}\left[Z_1 - X > 0, Z_2 - Y > 0\right] - 3\end{aligned}$$

Da questa relazione, sapendo che $(Z_1 - X, Z_2 - Y)$ è un vettore normale bivariate con matrice $R + I_2$, dove I_2 è la matrice identità 2×2, è possibile sfruttare la (4.30) e ottenere quindi la (4.29).

4.3.1 La copula gaussiana

La copula gaussiana, come suggerisce il nome, trae origine dalla distribuzione multivariata gaussiana. Essa si definisce come segue:

$$C_R^{Ga}(\mathbf{u}) = \Phi_R^d\left(\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_d)\right), \tag{4.31}$$

dove $\Phi_R^d(\cdot)$ è la funzione di ripartizione di un vettore gaussiano d-dimensionale standard con matrice di correlazione R ed è definita come segue:

$$\Phi_R^d(\mathbf{z}) = \int_{-\infty}^{z_1} \cdots \int_{-\infty}^{z_d} \frac{1}{(2\pi)^{\frac{d}{2}}|R|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}\mathbf{y}' R^{-1}\mathbf{y}\right\} \mathrm{d}y_d \cdots \mathrm{d}y_1 \tag{4.32}$$

dove $|R|$ indica il determinante di R. Nel caso di vettore gaussiano (X_1, X_2) bivariato con correlazione ρ si avrà quindi:

$$\Phi_\rho^2(x_1, x_2) = \int_{-\infty}^{x_1}\int_{-\infty}^{x_2} \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left\{-\frac{s^2 - 2\rho st + t^2}{2(1-\rho^2)}\right\} \mathrm{d}t\,\mathrm{d}s \tag{4.33}$$

Analogamente, $\Phi^{-1}\cdot$ è l'inversa della funzione di ripartizione della normale standard univariata.

In Figura 4.3.1 è riportato il grafico della densità della copula gaussiana con correlazione lineare pari a 0.3.

Si noti che l'unico parametro della copula gaussiana è la matrice di correlazione R, che può essere ricavato dai coefficienti di correlazione per ranghi τ_K e ρ_S, a partire dalle relazioni (4.28) e (4.29). Questo, come vedremo, è un primo vantaggio di questo tipo di copule, in quanto non è richiesto alcun costo computazionale per la stima dei parametri, oltre alla stima di una misura di correlazione.

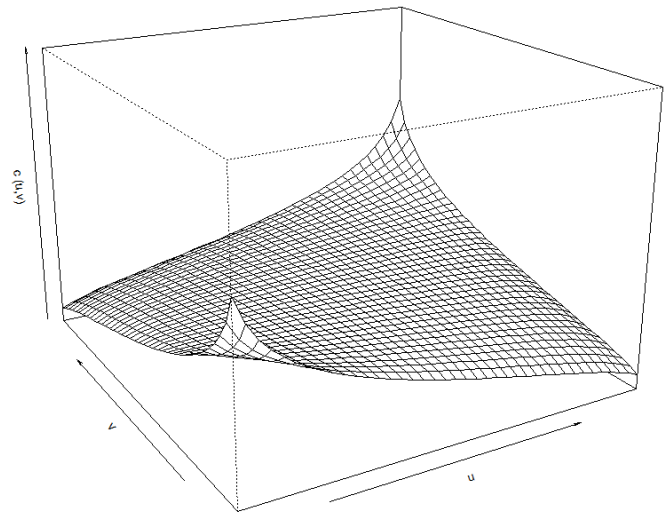

Figura 4.1: Densità di probabilità della copula gaussiana bivariata, con $\rho = 0.3$.

È inoltre possibile calcolare i parametri di dipendenza di coda della copula gaussiana. In primo luogo, si nota che la copula ha il medesimo comportamento nella coda superiore ed inferiore, in quanto la distribuzione gaussiana gode di simmetria radiale. Si denota quindi con λ la dipendenza di coda della copula. Usando la regola de L'Hôpital si ha:

$$\begin{aligned}\lambda &= \lim_{q\to 0^+} \frac{C(q,q)}{q} = \lim_{q\to 0^+} \frac{\mathrm{d}C(q,q)}{\mathrm{d}q} \\ &= \lim_{q\to 0^+} \mathrm{P}\left[U_2 \leq q | U_1 = q\right] + \lim_{q\to 0^+} \mathrm{P}\left[U_1 \leq q | U_2 = q\right].\end{aligned}$$

Inoltre, sfruttando il fatto che la copula gaussiana è scambiabile, ovvero che le marginali condizionate hanno la stessa distribuzione, si ottiene:

$$\lambda = 2 \lim_{q\to 0^+} \mathrm{P}\left[U_2 \leq q | U_1 = q\right]. \tag{4.34}$$

Sia quindi $(U_1, U_2) \sim C_R^{Ga}$. Definendo $(X_1, X_2) = (\Phi^{-1}(U_1), \Phi^{-1}(U_2))$, vettore aleatorio che risulta essere distribuito come una normale bivariata standard con correlazione lineare ρ, dalla (4.34) si ha:

$$\begin{aligned}\lambda &= 2 \lim_{q\to 0^+} \mathrm{P}\left[\Phi^{-1}(U_2) \leq \Phi^{-1}(q) | \Phi^{-1}(U_1) = \Phi^{-1}(q)\right] \\ &= 2 \lim_{x\to -\infty} \mathrm{P}\left[X_2 \leq x | X_1 = x\right]\end{aligned}$$

dove $x = \Phi^{-1}(q)$.

Sapendo poi che $X_2|X_1 = x$ è distribuito come una gaussiana di parametri $(\rho x, 1 - \rho^2)$:

$$\lambda = 2 \lim_{x \to -\infty} \Phi\left(\frac{x - \rho x}{\sqrt{1 - \rho^2}}\right) = 2 \lim_{x \to -\infty} \Phi\left(\frac{x\sqrt{1 - \rho}}{\sqrt{1 + \rho}}\right) = 0 \qquad (4.35)$$

se $\rho < 1$. Ovvero, la copula gaussiana gode di indipendenza asintotica in entrambe le code, ovvero non c'è correlazione tra i quantili estremi delle distribuzioni marginali.

Nella modellizzazione di dati multivariati con le copule, è spesso utile simulare pseudo-osservazioni dalla copula selezionata. Nell'ambito del modello per la gestione dei rischi operativi, si simula dalla copula per ottenere la distribuzione di perdite aggregate annue, come somma delle perdite delle sette categorie di rischi.

Per quanto riguarda le copule ellittiche, l'idea di partenza è quella di simulare campioni dalla distribuzione multivariata relativa, per ottenere poi le marginali uniformi calcolando i relativi livelli dei quantili. Questo è un vantaggio delle copule ellittiche, in quanto il campionamento è più semplice in quanto si può partire da distribuzioni multivariate note, per cui esistono metodi di simulazione.

Per simulare dalla distribuzione normale d-dimensionale [14] si considera innanzitutto la *decomposizione di Cholesky* della matrice di correlazione R: la matrice di Cholesky di R è l'unica matrice A triangolare inferiore tale che $R = AA^T$. È quindi possibile dimostrare che, se le variabili Z_1, Z_d sono normali standard indipendenti, allora $\mathbf{Z} = A\mathbf{Z}$ ha distribuzione normale d-dimensionale con matrice di correlazione R. Data la matrice di correlazione R, l'algoritmo per simulare un'osservazione dalla copula gaussiana è quindi il seguente:

1. Si trova la matrice di Cholesky A di R;
2. Si simulano $z_1, \ldots, z_d$ i.i.d. $\sim N(0, 1)$;
3. Si definisce $\mathbf{x} = A\mathbf{z}$;
4. Si determinano le quantità $u_i = \phi(x_i)$, con $i = 1, \ldots, d$;
5. Il vettore $\mathbf{u}$ è un campione dalla copula C_R^{Ga}.

4.3.2 La copula t di Student

Un'altra copula appartenente alla famiglia delle distribuzioni ellittiche è copula t di Student. Come la copula gaussiana, anche questa trae origine dalla relativa distribuzione univariata. In particolare, una variabile Z con distribuzione t con ν gradi di libertà si può sempre rappresentare come:

$$Z = \frac{X_1}{\sqrt{\xi/\nu}}$$

dove X_1 è una normale standard, e ξ (indipendente da X_1) ha distribuzione χ^2 a ν gradi di libertà, costruita come somma di ν gaussiane standard al quadrato, indipendenti e identicamente distribuite. La funzione di ripartizione di Z è la seguente:

$$t_\nu(z) = \int_{-\infty}^{z} \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\Gamma\left(\frac{\nu}{2}\right)} \left[1 + \frac{s^2}{\nu}\right]^{-\frac{\nu+1}{2}} \mathrm{d}s \tag{4.36}$$

dove Γ è la funzione gamma definita nella (2.15).

Analogamente, il vettore $\mathbf{Z}$ d-dimensionale ha distribuzione t con ν gradi di libertà e matrice di correlazione R se:

$$\mathbf{Z} = \left(\frac{X_1}{\sqrt{\xi/\nu}}, \frac{X_2}{\sqrt{\xi/\nu}}, \cdots, \frac{X_d}{\sqrt{\xi/\nu}}\right)$$

dove $\mathbf{X} \sim N^d(0, R)$ e $\xi \sim \chi^2(\nu)$, con $\mathbf{X}$ e ξ indipendenti.

La funzione di ripartizione $t_{\nu,R}$ della variabile $\mathbf{Z}$ si ricava dalle relative distribuzioni di $\mathbf{X}$ e ξ, ed ha la seguente espressione:

$$t^d_{\nu,\rho}(\mathbf{z}) = \int_{-\infty}^{z_1} \cdots \int_{-\infty}^{z_d} \frac{\Gamma\left[(\nu+d)/2\right]}{\Gamma\left[\nu/2\right](\nu\pi)^{d/2}|R|^{1/2}} \left[1 + \frac{1}{2}\mathbf{y}' R^{-1}\mathbf{y}\right]^{-\frac{\nu+d}{2}} \mathrm{d}y_d \cdots \mathrm{d}y_1 \tag{4.37}$$

Nel caso di vettore 2-dimensionale con correlazione ρ si ha:

$$t^2_{\nu,\rho}(z_1, z_2) = \int_{-\infty}^{z_1}\int_{-\infty}^{z_2} \frac{1}{2\pi\sqrt{1-\rho^2}} \left[1 + \frac{s^2 - 2\rho st + t^2}{\nu(1-\rho^2)}\right]^{-\frac{\nu+2}{2}} \mathrm{d}t\,\mathrm{d}s \tag{4.38}$$

È a questo punto possibile definire la copula t di Student d-dimensionale:

$$C^t_{\nu,R}(\mathbf{u}) = t_{\nu,R}\left(t_\nu^{-1}(u_1), \ldots, t_\nu^{-1}(u_d)\right), \tag{4.39}$$

In Figura 4.3.2 si riporta il grafico della densità della copula t con 2 gradi di libertà e correlazione e $\rho = 0.3$. Come si può notare immediatamente, tale copula, rispetto alla gaussiana, ha una maggiore massa di probabilità concentrata nelle code inferiore e superiore (rispettivamente $u_i \to 0$ e $u_i \to 1$, con $i = 1, 2$). Ciò, come si vedrà nel seguito, si traduce nella presenza di correlazione di coda.

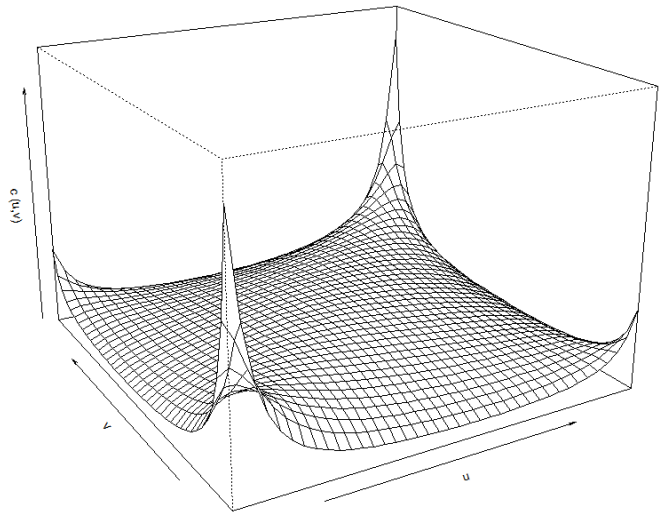

Figura 4.2: Densità di probabilità della copula t bivariata, con $\rho = 0.3$ e 2 gradi di libertà.

Dalla definizione si osserva come la copula t di Student sia determinata da due parametri: la matrice di correlazione R e i gradi di libertà ν. R, come nel caso della copula gaussiana, si può ricavare dal coefficiente di correlazione per ranghi scelto. Per quanto riguarda ν, invece, il metodo solitamente utilizzato è la stima di massima verosimiglianza (MLE). Per implementarlo si utilizzano algoritmi di ottimizzazione numerica per trovare il massimo della log-verosimiglianza delle osservazioni, dove la densità della singola osservazione è data dalla combinazione di (4.8) e (4.39). Tale metodo di stima è implementato nel software R ([20] e [31]).

Per quanto riguarda invece le dipendenze di coda della copula t, è possibile innanzitutto notare che la distribuzione t, come la normale, gode della proprietà di simmetria radiale, e quindi le misure di dipendenza di coda superiore e inferiore saranno uguali. Inoltre, se si considera un vettore (Z_1, Z_2) con distribuzione t bivariata con ν gradi di libertà e correlazione

ρ, allora [23] $Z_1|Z_2 = z$ ha a sua volta distribuzione t (non standard) con $\nu + 1$ gradi di libertà e

$$\mathrm{E}\left[Z_1|Z_2 = z\right] = \rho z, \qquad \mathrm{Var}\left[Z_1|Z_2 = z\right] = \left(\frac{\nu + z^2}{\nu + 1}\right)(1 - \rho^2).$$

Si può poi verificare che

$$\left(\frac{\nu + 1}{\nu + z^2}\right)^{1/2} \frac{Z_2 - \rho x}{1 - \rho^2} \sim t_{\nu+1}.$$

L'analogo vale anche per $Z_1|Z_2$.

Da ciò si osserva che anche la copula t di Student gode della proprietà di scambiabilità, e quindi le dipendenze di coda λ si possono esprimere con la (4.34). Sia quindi $(U_1, U_2) \sim C^t_{\nu,R}$ e si consideri il vettore $(Z_1, Z_2) = (t_\nu^{-1}(U_1), t_\nu(U_2))$, vettore aleatorio che risulta essere distribuito come una t di Student con ν gradi di libertà e correlazione lineare ρ. Quindi:

$$\begin{aligned}
\lambda &= 2 \lim_{z \to -\infty} \mathrm{P}\left[Z_2 \leq z | Z_1 = z\right] \\
&= 2 \lim_{z \to -\infty} \left(\left(\frac{\nu + 1}{\nu + z^2}\right)^{1/2} \frac{z - \rho z}{1 - \rho^2} \right) \\
&= 2 t_{\nu+1} \left(-\sqrt{\frac{(\nu + 1)(1 - \rho)}{1 + \rho}} \right)
\end{aligned} \tag{4.40}$$

Da ciò si evince come, diversamente dalla copula gaussiana, la copula t abbia dipendenza alle code, crescente con ρ e decrescente con ν. In particolare, tende a zero quando ν tende all'infinito. La dipendenza da ν evidenzia quindi come la copula t tende a comportarsi come quella gaussiana quando il numero dei gradi di libertà è molto elevato; questo comportamento deriva dal fatto che la distribuzione t converge ad una $N(0, 1)$ quando $\nu \to \infty$.

In definitiva, le due copule appena mostrate hanno comportamento simile nei pressi delle mediane delle distribuzioni, con struttura determinata dalla matrice di correlazione e proprietà di simmetria radiale. La differenza principale, come appena visto, sta nel comportamento agli estremi: le componenti della copula gaussiana sono asintoticamente indipendenti, mentre per la t si ha dipendenza nelle code. Per questo motivo, nell'ambito dei rischi operativi, solitamente si preferisce la t di Student, in quanto, basando la stima del capitale da accantonare sui valori estremi (tramite il VaR), essa permette di mantenere la struttura di correlazione

osservata dalle serie storiche. D'altro canto, come osservato, la scelta di tale copula implica la necessità di costo computazionale per la stima dei gradi di libertà.

L'algoritmo per simulare osservazioni dalla copula t con ν gradi di libertà e matrice di correlazione R è simile a quello per la copula gaussiana:

1. Si trova la matrice di Cholesky A di R;
2. Si campionano $z_1, \ldots, z_d$ i.i.d. $\sim N(0,1)$;
3. Si simula $\xi \sim \chi^2(\nu)$, indipendente da $\mathbf{z}$;
4. Si definiscono $\mathbf{x} = A\mathbf{z}$ e $\mathbf{y} = \mathbf{x}/\sqrt{\xi/\nu}$;
5. Si determinano le quantità $u_i = t_\nu(y_i)$, con $i = 1, \ldots, d$;
6. Il vettore $\mathbf{u}$ è un campione dalla copula $C^t_{R,\nu}$.

4.4 Copule Archimedee

Una differente famiglia di copule è quella delle copule *Archimedee*. La caratteristica principale è il fatto che le relative densità hanno un'espressione esplicita, contrariamente alle copule gaussiane, le cui distribuzioni sono definite tramite integrali. Uno svantaggio è invece che esse non vengono definite a partire da distribuzioni multivariate tramite il Teorema di Sklar, ed è quindi necessario assicurare che soddisfino le proprietà delle copule.

Per introdurre tale famiglia, si consideri una funzione decrescente ψ da $[0,1]$ a $[0,\infty]$. Essa viene chiamata *generatrice* delle copule Archimede; tutte le copule di questa famiglia sono infatti della forma:

$$C(\mathbf{u}) = \psi^{-1}\left(\psi(u_1) + \psi(u_2) + \cdots + \psi(u_d)\right) \tag{4.41}$$

La correttezza delle copule così definite è garantita dal seguente teorema:

Teorema 4.2. *Si consideri la funzione ψ continua e non decrescente, tale che $\psi : [0,1] \to [0,\infty]$, e $\psi(1) = 0$. Allora*

$$C(\boldsymbol{u}) = \begin{cases} \psi^{-1}\left(\psi(u_1) + \cdots + \psi(u_d)\right) & \text{se } \psi(u_1) + \cdots + \psi(u_d) \leq \psi(0) \\ 0 & \text{altrimenti} \end{cases}$$

è una copula se e solo se ψ è convessa.

Per la dimostrazione si rimanda a [25]. Si noti che, se $\psi(0) = \infty$, non vi sono restrizioni sul dominio della copula, che coincide con $\mathbb{R}^d$.

Si illustreranno ora le principali copule appartenenti alla famiglia delle copule Archimedee: Gumbel, Clayton e Frank.

Per la copula *Gumbel* si usa come funzione generatrice:

$$\psi_\theta^{Gu}(u) = [-\log(u)]^\theta$$

con $\theta \in [1, \infty)$. La copula bivariata relativa è quindi la seguente:

$$C_\theta^{Gu}(u_1, u_2) = \exp\left\{-\left([-\log(u_1)]^\theta + [-\log(u_2)]^\theta\right)^{\frac{1}{\theta}}\right\} \tag{4.42}$$

Quando $\theta = 1$ tale copula coincide con la copula di indipendenza espressa nella (4.4), mentre quando $\theta \to \infty$ essa tende alla copula comonotona (4.6). In tal senso, la copula Gumbel rappresenta una sorta di interpolazione, tramite il parametro θ, tra indipendenza e perfetta dipendenza positiva.

In Figura 4.4 è riportato il grafico della densità della copula Gumbel, con $\theta = 2$. Si osserva immediatamente come tale copula goda di dipendenza nella coda superiore.

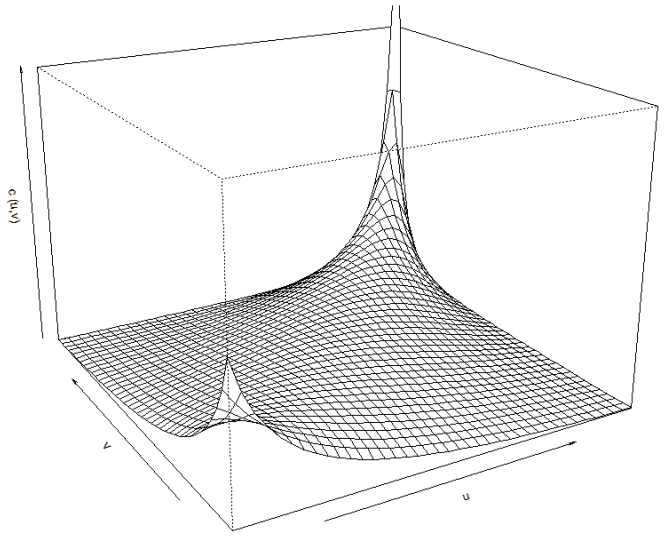

Figura 4.3: Densità di probabilità della copula Gumbel bivariata, con $\theta = 2$.

Per la Gumbel, come per tutte le copule appartenenti alla famiglia Archimedea, il parametro θ si può ricavare, nel caso di copule bidimensionali, a partire dal Tau di Kendall, attraverso la relazione (4.18). Essa,

per tale famiglia, diventa:

$$\tau_K = 4\int_0^1 \int_0^1 C(u_1, u_2)\mathrm{d}C(u_1, u_2) - 1 = 1 + 4\int_0^1 \frac{\psi(t)}{\psi(t)}\mathrm{d}t \tag{4.43}$$

Per la dimostrazione si rimanda a [25].

Per la copula Gumbel, la (4.43) diventa:

$$\tau_K = 1 + 4\int_0^1 \frac{t\log t}{\theta}\mathrm{d}t = 1 + \frac{4}{\theta}\left(\left[\frac{t^2}{2}\log t\right]_0^1 - \int_0^1 \frac{t}{2}\mathrm{d}t\right) = 1 - \frac{1}{\theta} \tag{4.44}$$

da cui si ricava il valore di θ a partire dal valore di τ stimato dai dati. In alternativa, θ si può stimare con il metodo MLE, come per la t di Student, ma è preferibile il primo metodo, in quanto ha minor costo computazionale, ma soprattutto permette di mantenere informazioni sulla correlazione tra le due variabili. Lo stesso discorso varrà anche per le altre copule della famiglia in considerazione.

Il problema, per questa come per le altre copule Archimedee, è il fatto che, passando a $d \geq 3$ dimensioni, si perde la relazione tra il parametro θ e la correlazione, che non è più rappresentata da un singolo valore, ma da una matrice. In questi casi, il parametro θ dev'essere stimato con il metodo MLE, implementato nel software R ([20] e [31]). In pratica, ciò significa non solo perdere qualsiasi informazione sulla struttura di correlazione, ma anche non essere in grado di differenziare i comportamenti delle singole marginali, che risultano quindi identicamente distribuite. Questo è un grosso limite delle copule Archimedee a più di due dimensioni, ed è il principale motivo per cui tale famiglia viene preferita quasi unicamente in presenza di dati bivariati.

Per quanto riguarda la famiglia di copule in considerazione, il fatto di avere un'espressione esplicita facilita notevolmente anche il calcolo delle dipendenze di coda dalle (4.26) e (4.27). Per la Gumbel si ha, per la coda inferiore:

$$\begin{aligned} \lambda_l^{Gu} &= \lim_{q\to 0^+} \frac{C_\theta^{Gu}(q,q)}{q} \\ &= \lim_{q\to 0^+} \frac{q^{2^{1/\theta}}}{q} = 0 \end{aligned}$$

mentre per la coda superiore, utilizzando la regola de L'Hôpital:

$$\begin{aligned}\lambda_u^{Gu} &= \lim_{q\to 1^-} \frac{1-2q+C_\theta^{Gu}(q,q)}{1-q} \\ &= \lim_{q\to 1^-} \frac{1-2q+q^{2^{1/\theta}}}{1-q} \\ &= \lim_{q\to 1^-} \frac{-2+2^{1/\theta}q^{2^{1/\theta}-1}}{-1} = 2-2^{1/\theta} \end{aligned} \tag{4.45}$$

Quindi, tale copula gode di indipendenza nella coda inferiore, mentre per la coda superiore si ha dipendenza positiva per $\theta > 1$.

Una seconda copula Archimedea è la *Frank*, la cui funzione generatrice è:

$$\psi_\theta^{Fr}(u) = \log\left(e^{-\theta}-1\right) - \log\left(e^{-\theta u}-1\right)$$

con $\theta \in [-1,\infty)\setminus 0$, e l'espressione della copula è data da:

$$C_\theta^{Fr}(u_1,u_2) = -\frac{1}{\theta}\log\left[1+\frac{(e^{-\theta u_1}-1)(e^{-\theta u_2}-1)}{e^{-\theta}-1}\right] \tag{4.46}$$

Nel caso della copula Frank, quando $\theta \to \infty$ si ottiene la copula di indipendenza, mentre quando $\theta \to 0$ la copula comonotona (4.6).

Si riporta in Figura 4.4 la densità di probabilità della copula Frank, con $\theta = 2$. DDal grafico si può osservare come questa copula abbia un comportamento simile alla gaussiana, ovvero indipendenza alle code.

Anche per tale copula si può ricavare il valore di θ a partire dal Tau di Kendall tramite la (4.43). La relazione in questo caso è più complicata, per la dimostrazione si rimanda quindi a [12]:

$$\tau_K = 1 - \frac{4}{\theta}\left[1 - D_1(\theta)\right] \tag{4.47}$$

dove $D_k(x)$ è la funzione di Debye:

$$D_k(x) = \frac{k}{x^k}\int_0^x \frac{t^k}{e^t-1}\mathrm{d}t$$

La copula di Frank gode di simmetria radiale, quindi è possibile calcolare un'unica misura di dipendenza λ^{Fr} per le code superiore e inferiore.

$$\begin{aligned}\lambda^{Fr} &= \lim_{q\to 0^+} \frac{-\frac{1}{\theta}\log\left[1+\frac{(e^{-\theta q}-1)^2}{e^{-\theta}-1}\right]}{q} \\ &= \lim_{q\to 0^+} \left[1+\frac{(e^{-\theta q}-1)^2}{e^{-\theta}-1}\right]^{-1} 2e^{-\theta q}\left(e^{-\theta q}-1\right) = 0\end{aligned}$$

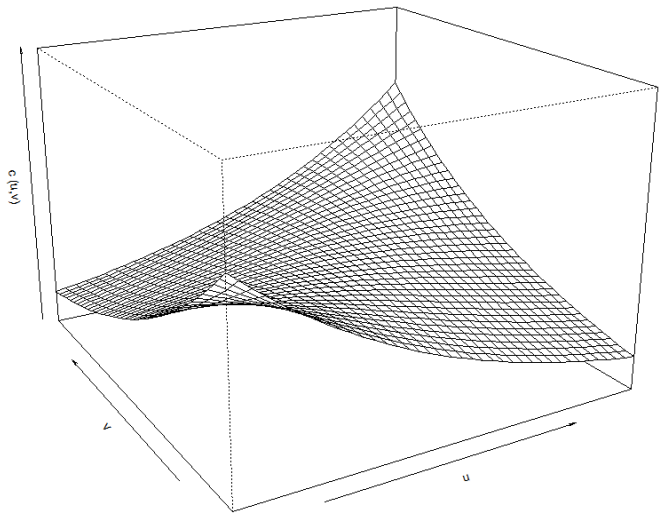

Figura 4.4: Densità di probabilità della copula Frank bivariata, con $\theta = 2$.

dove si è usata ancora una volta la regola de L'Hôpital. In definitiva, questa copula gode di indipendenza in entrambe le code.

Un'ultima copula appartenente alla famiglia è la *Clayton*, generata da:

$$\psi_\theta^{Cl}(u) = \frac{1}{\theta}\left(u^{-\theta} - 1\right)$$

con $\theta \in \mathbb{R} \setminus 0$. La relativa copula bivariata è la seguente:

$$C_\theta^{Cl}(u_1, u_2) = \left(u_1^{-\theta} + u_2^{-\theta} - 1\right)^{-\frac{1}{\theta}} \tag{4.48}$$

In Figura 4.4 è riportato il grafico della densità della Frank, con $\theta = 2$. Il comportamento di tale copula appare simmetrico rispetto alla Gumbel, mostrando una dipendenza nella coda inferiore (quando u_1 e u_2 tendono a 0).

La relazione di θ con il tau di Kendall si può ricavare dalla (4.43):

$$\tau_K = 1 + 4\int_0^1 \frac{t^{\theta+1} - t}{\theta}\mathrm{d}t = 1 + \frac{4}{\theta}\left(\frac{1}{\theta+2} - \frac{1}{2}\right) = \frac{\theta}{\theta+2} \tag{4.49}$$

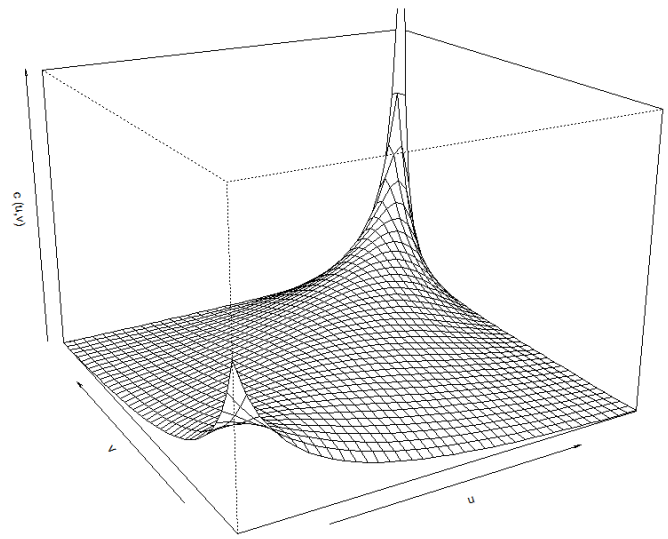

Figura 4.5: Densità di probabilità della copula Gumbel bivariata, con $\theta = 2$.

Per quanto riguarda le dipendenze di coda si ha:

$$\begin{aligned} \lambda_l^{Cl} &= \lim_{q\to 0^+} \frac{\left(2q^{-\theta}-1\right)^{-\frac{1}{\theta}}}{q} \\ &= \lim_{q\to 0^+} \left(2-q^{\theta}\right)^{-\frac{1}{\theta}} = 2^{-1/\theta} \end{aligned} \tag{4.50}$$

e

$$\begin{aligned} \lambda_u^{Cl} &= \lim_{q\to 1^-} \frac{1-2q+\left(2q^{-\theta}-1\right)^{-\frac{1}{\theta}}}{1-q} \\ &= \lim_{q\to 1^-} \frac{-2+2q^{-\theta-1}\left(2q^{-\theta}-1\right)^{-\frac{1}{\theta}-1}}{-1} = 0 \end{aligned}$$

La copula di Clayton, quindi, ha un comportamento simmetrico rispetto alla Gumbel, mostrando indipendenza nella coda superiore e dipendenza positiva in quella inferiore.

In definitiva, tutte le copule Archimedee risultano spesso comode in quanto le distribuzioni hanno espressioni esplicite. Generalmente esse vengono preferite nel caso di copule bidimensionali in quanto, dipenden-

do da un unico parametro θ, non riflettono le correlazioni tra tutte le coppie di variabili se $d \geq 3$. Un vantaggio è però la possibilità di modellizzare comportamenti differenti agli estremi. La copula Gumbel viene quindi preferita nel momento in cui si osserva correlazione tra i dati nella coda superiore, mentre una minore nella coda inferiore; nelle applicazioni, ciò accade ad esempio quando si ha a che fare con dati di perdita (come nel caso dei rischi operativi) in cui si è più interessati alle correlazioni tra perdite di grande intensità. Analogamente, la Clayton è in grado di modellizzare correlazione per valori bassi dei dati, mantenendo indipendenza nella coda superiore. Infine, la Frank gode di indipendenza tra le marginali in entrambe le code, ed è quindi preferibile rispetto alla copula gaussiana, che gode della medesima proprietà alle code, solo se si vogliono modellizzare dati bidimensionali.

4.4.1 Campionamento dalle copule Archimedee

Per ottenere presudo-osservazioni da copule Archimedee, differentemente dalle copule ellittiche, per cui era sufficiente campionare dalla relativa distribuzone multivariata, il procedimento è meno immediato.

Sia quindi G una generica funzione di ripartizione, si denoti la sua trasformata di Laplace con $\widehat{G}$, ovvero:

$$\widehat{G} = \int_0^\infty e^{-tx} \mathrm{d}G(x)$$

Si osserva innanzitutto che tale funzione è continua e strettamente decrescente. L'idea è quella di simulare un'osservazione $V \sim G$ e $X_1, \ldots, X_d$ indipendenti con distribuzione uniforme. È possibile dimostrare che, definendo

$$U_i = \widehat{G}\left(-\frac{\log X_i}{V}\right),$$

il vettore $\mathbf{U}$ ha distribuzione multivariata data dalla copula Archimedea con funzione generatrice $\psi = \widehat{G}^{-1}$.

Infatti si osserva che:

$$\begin{aligned} \mathrm{P}\left[U_i \leq u | V = v\right] &= \mathrm{P}\left[\widehat{G}\left(-\frac{\log X_i}{v}\right) \leq u\right] \\ &= \mathrm{P}\left[-\frac{\log X_i}{v} \geq \widehat{G}^{-1}(u)\right] \\ &= \mathrm{P}\left[X_i \leq e^{-v\widehat{G}^{-1}(u)}\right] = e^{-v\widehat{G}^{-1}(u)} \end{aligned}$$

e quindi si ha:

$$\begin{aligned}
\mathrm{P}\left[U_1 \leq u_1, \ldots, U_d \leq u_d\right] &= \int_0^\infty \mathrm{P}\left[U_1 \leq u_1, \ldots, U_d \leq u_d | V = v\right] \mathrm{d}G(v) \\
&= \int_0^\infty \prod_{i=1}^{d} \mathrm{P}\left[U_i \leq u_i | V = v\right] \mathrm{d}G(v) \\
&= \int_0^\infty e^{-v\left(\widehat{G}^{-1}(u_1)+\cdots+\widehat{G}^{-1}(u_d)\right)} \mathrm{d}G(v) \\
&= \widehat{G}\left(\widehat{G}^{-1}(u_1) + \cdots + \widehat{G}^{-1}(u_d)\right)
\end{aligned}$$

che è proprio l'espressione di una copula con generatrice $\widehat{G}^{-1}$.

L'algoritmo per il campionamento da una copula Archimedea con funzione generatrice ψ è quindi il seguente:

1. Si campiona un'osservazione $v \sim G$, con G tale che $\psi = \widehat{G}^{-1}$;
2. Si simulano $x_1, \ldots, x_d$ i.i.d. $\sim U(0,1)$;
3. Si determinano le quantità $u_i = \widehat{G}\left(-\frac{\log x_i}{v}\right)$, con $i = 1, \ldots, d$;
4. Il vettore $\mathbf{u}$ è un campione dalla copula Archimedea con generatrice ψ.

Per quanto riguarda le copule Gumbel, Frank e Clayton, è possibile inoltre trovare distribuzioni G notevoli [23]. Per quanto riguarda la Gumbel, infatti, si campiona V dalla distribuzione *stabile*, la cui funzione caratteristica è:

$$\varphi^{st}_{\mu,c,\alpha,\beta}(t) = \mathrm{E}\left[e^{itV}\right] = \exp\left\{it\mu - |ct|^\alpha(1 - i\beta\,\mathrm{sign}(t)\mathrm{tg}(\pi\alpha/2))\right\}$$

I relativi parametri per ottenere la ψ desiderata sono $(1/\theta, 1, \delta, 0)$, dove $\delta = \left[\cos(\pi/(2\theta))\right]^\theta$. Da questa si ottiene $\widehat{G}(t) = \exp\left\{-t^{1/\theta}\right\}$.

Per campionare dalla copula Frank la distribuzione di V è discreta con massa di probabilità

$$p(k) = \mathrm{P}[V = k] = \frac{\left(1 - e^{-\theta}\right)^k}{k\theta}.$$

Infine, per la Clayton, V ha distribuzione gamma, che ha densità di probabilità

$$g_{k,\lambda}(v) = v^{k-1}\frac{e^{-v/\lambda}}{\lambda^k \Gamma(k)},$$

di parametri $(1/\theta, 1)$. Da questa si ottiene $\widehat{G}(t) = (1+t)^{-1/\theta}$.

4.5 Test di buon adattamento

Abbiamo fin'ora visto qualitativamente quale copula scegliere sulla base del comportamento empirico delle osservazioni, basandosi ad esempio sul comportamento agli estremi, e come è possibile stimare i parametri delle diverse copule, anche a seconda della misura di correlazione scelta. Un ultimo problema che entra in gioco nella modellizzazione di vettori di osservazioni con le copule è quindi la verifica del buon adattamento della copula, con i parametri stimati secondo un determinato metodo, ai dati.

I test solitamente utilizzati nell'ambito delle copule sono una generalizzazione dei test non parametrici nel caso di distribuzioni multivariate [3], che si basano quindi sul confronto tra la distribuzione multivariata teorica scelta e quella empirica. Prima di definirne le statistiche, è quindi necessario introdurre la copula *empirica* $C^n(\mathbf{u})$, che rappresenta la distribuzione con cui si andrà a confrontare la copula teorica. Essa viene definita coerentemente con la distribuzione empirica (2.25). Sia $\mathbf{x}^1, \ldots, \mathbf{x}^n$, con $\mathbf{x}^j = \left(x_1^j, x_2^j, \ldots, x_d^j\right)$, il set di dati a disposizione. Siano quindi $x_i^{(1)}, \ldots, x_i^{(d)}$, con $i = 1, \ldots, n$ le componenti i-esime disposte in ordine crescente. Si può perciò scrivere la funzione di ripartizione empirica di della i-esima componente di rango j_i:

$$F\left(x_i^{(j_i)}\right) = \frac{j_i}{n+1}$$

con $j_i = 1, \ldots, n$ e $i = 1, \ldots, d$.

La copula empirica si definisce quindi come segue:

$$C^n\left(\frac{j_1}{n+1}, \cdots, \frac{j_d}{n+1}\right) = \frac{1}{n+1} \sum_{j=1}^{n} \prod_{i=1}^{d} \mathbb{1}_{\left\{\text{rank}\left(x_i^j\right) \leq j_i\right\}} \tag{4.51}$$

ovvero si contano in numero di coppie per cui ogni componente i è minore o uguale della relativa componente di $x_i^{(j_i)}$.

In questo ambito sono usati come test di buon adattamento estensioni dei metodi illustrati nella sezione 2.6, ovvero di Kolmogorov-Smirnov e Anderson-Darling, per cui è possibile riscriverne le statistiche (2.30) e (2.32) nell'ambito delle copule:

$$KS = \sup_{\mathbf{u}} \left| C^n(\mathbf{u}) - C(\mathbf{u}) \right| \tag{4.52}$$

$$AD = \sup_{\mathbf{u}} \left| \frac{C^n(\mathbf{u}) - C(\mathbf{u})}{\sqrt{C(\mathbf{u})\left(1 - C(\mathbf{u})\right)}} \right| \tag{4.53}$$

Tuttavia, nel caso di test di buon adattamento per le copule, a tali test è generalmente preferito il test di *Cramer-von Mises* [13]. Innanzitutto, la statistica di tale test per le copule è la seguente:

$$CvM = \int_{[0,1]^d} \left[C^n(\mathbf{u}) - C(\mathbf{u})\right]^2 \mathrm{d}C(\mathbf{u}) \tag{4.54}$$

Per il calcolo pratico di questa quantità, si rimanda a [3]. Si noti comunque che la distribuzione della statistica dipencde dalla copula scelta, e quindi il p-value è solitamente ottenuto tramite simulazione Monte Carlo, come è implementato nel software R ([20] e [31]).

La caratteristica di questa statistica è che non viene calcolata su un'unica osservazione (quella in cui si realizza la distanza massima, come nei casi precedenti), bensì viene valutata su tutto il dominio, pesando ogni elemento con la relativa densità teorica $C(\mathbf{u})$. Il motivo principale per cui viene scelto il test di Cramer-von Mises è quindi la sua stabilità, in quanto hanno molto più peso le aree con alta probabilità. D'altro canto, viene data meno importanza le code della distribuzione, che però nell'ambito delle copule sono tenute sotto controllo tramite le dipendenze di coda.

Capitolo 5

Applicazione del modello

L'obiettivo di questo capitolo è quello di mostrare l'applicazione dell'intero modello di calcolo del capitale da accantonare a fronte dei rischi operativi. I dati che verranno utilizzati sono relativi a una delle maggiori banche italiane che, per motivi di privacy, rimarrà anonima, così come i dati sono stati tutti riscalati per un comune fattore.

Si metteranno quindi in pratica tutte le metodologie di stima e simulazione descritte nei capitoli precedenti di questo elaborato, che servono quindi come supporto teorico all'applicazione. Tutte le analisi sono state effettuate con il software R; i codici sono eventualmente disponibili. Dopo una descrizione dei dataset a disposizione, presentando i risultati unicamente per una classe di rischio, si descriverà nel dettaglio il processo di modellizzazione del corpo della distribuzione di severity con le distribuzioni troncate (Capitolo 2), per poi passare alla coda della stessa, tramite l'utilizzo della teoria dei valori estremi (Capitolo 3). Quindi, si procederà con la convoluzione di frequency e severity per arrivare alla distribuzione aggregata di perdite annue. Infine, si utilizzeranno i risultati della modellizzazione di tutte le classi di rischio per effettuarne l'aggregazione tramite copule (Capitolo 4), e arrivare quindi alla determinazione del capitale a rischio.

5.1 Descrizione dei dataset

Come anticipato nell'introduzione, per la costruzione del modello si utilizzano tre diversi dataset:

- *Interni*: perdite interne degli anni 2005-2009.

- *DIPO* e *Pubblici*: perdite esterne degli anni 2005-2009 superiori alla soglia u_0, provenienti dal database Database Italiano Perdite Operative e da dati pubblici.

- *Scenario*: perdite prospettiche ottenute da analisi di scenario, superiori alla soglia u_0.

Il processo di raccolta dei rischi operativi adottato nella banca prevede l'istituzione di processi e controlli di tipo qualitativo paralleli al modello qui descritto. Per questo motivo non tutte le informazioni verranno utilizzate, come vedremo. I record a disposizione per ogni perdita operativa sono infatti:

- Classe di rischio

- Data di accadimento dell'evento

- Data di contabilizzazione

- Impatto economico

- Recuperi assicurativi

- Business line a cui si riferisce l'evento

Ovviamente le date e i recuperi assicurativi non sono presenti nei dati di scenario, in quanto prospettici.

Suddividendo innanzitutto i dati per classe di rischio, di queste informazioni si sono utilizzate l'impatto economico (indipendentemente dalla data) per quanto riguarda la distribuzione di severity, e il numero di eventi operativi annui (relative solo alle perdite interne) per la frequency.

Riguardo il dataset delle perdite interne, si riportano in Tabella 5.1 le numerosità dei dati divisi per ET e per impatto economico (inferiori alla soglia minima di raccolta H, superiori alla soglia di raccolta di dati esterni e di scenario u_0 e compresi tra le due).

Come si può notare immediatamente, la scarsa numerosità dei dati sopra la soglia u_0 implica la necessità dell'introduzione di ulteriori dati (esterni e di scenario) per la modellizzazione della coda della severity. Riguardo a questi ultimi due dataset, se ne riportano le numerosità, divise per ET, in Tabella 5.2.

La stima dei parametri e i test di robustezza per la costruzione del modello perdite annue ottenute dalla convoluzione tra severity e frequency sono state ovviamente portate avanti per tutte e 7 le classi di rischio. Tuttavia, nel seguito si presenta l'analisi unicamente per l'ET 1 (perdite

Interni	$(0, H)$	$[H, u_0)$	$[u_0, \infty)$	TOT
ET 1	226	238	7	471
ET 2	52874	2071	3	54945
ET 3	424	362	3	789
ET 4	7661	5092	34	12787
ET 5	3106	343	0	3449
ET 6	3057	369	0	3426
ET 7	71881	4708	29	76618

Tabella 5.1: Numerosità del dataset di perdite interne.

	DIPO	Pubblici	Scenario	TOT
ET 1	63	478	24	565
ET 2	20	140	24	184
ET 3	12	68	9	89
ET 4	63	724	26	813
ET 5	3	4	23	30
ET 6	3	11	14	28
ET 7	35	100	20	155

Tabella 5.2: Numerosità dei dataset di perdite esterne (DIPO e pubblici) e di scenario.

corrispondenti a frodi interne), in quanto le conclusioni per tale classe di rischio possono essere facilmente estese alle altre. Infine, si utilizzano le stime di tutti gli ET per la loro aggregazione tramite copule.

5.2 Modellizzazione del corpo della severity

Come anticipato, il corpo della distribuzione modellizza le perdite interne superiori alla soglia di raccolta H ed inferiori alla soglia dei valori estremi, fissata a u_0 dalla banca, ma, come vedremo, può essere modificata per ottenere un fit migliore. I dati che vengono utilizzati per tale stima sono tutte le perdite interne superiori ad H: si modellizzano provvisoriamente anche le perdite estreme, per garantire continuità alla distribuzione. Si riporta in Figura 5.1 il box-plot (in scala logaritmica e non) di tali dati. Si osserva immediatamente come vi sia una forte asimmetria destra: la distribuzione di severity dei dati interni appare quindi molto più concentrata sui perdite elevate, giustificando la necessità di modellizzare le perdite elevate tramite la teoria dei valori estremi.

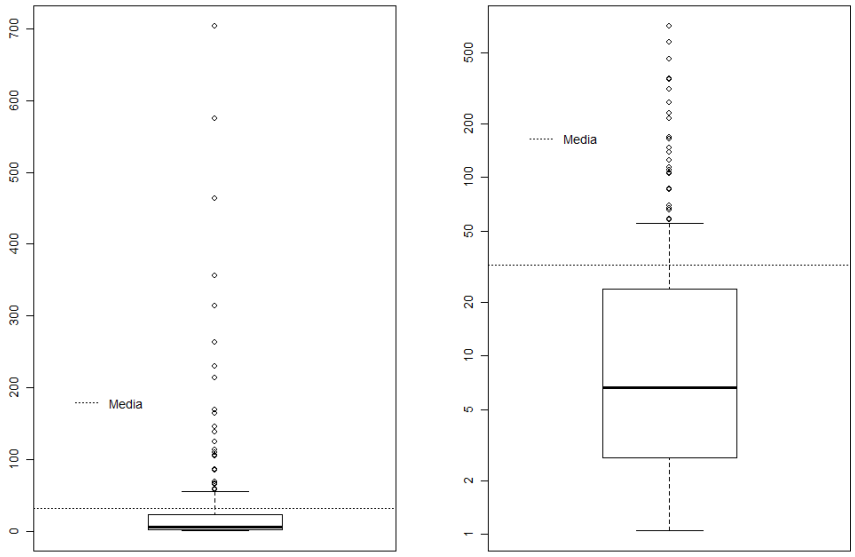

Figura 5.1: Box-plot (sinistra) e box-plot in scala logaritmica (destra) dei dati di perdita interni superiori alla soglia H.

Prima di passare alla modellizzazione vera e propria del corpo della distribuzione di severity, è necessario innanzitutto verificare le ipotesi secondo cui i dati provenienti dalla serie storica di perdite interne siano i.i.d. (ovvero indipendenti ed indenticamente distribuiti). Per tale motivo, si introducono l'autocorrelation plot e il test di Box-Jenkins, metodi rispettivamente qualitativo e quantitativo.

L'*autocorrelation plot* è volto a verificare l'ipotesi di indipendenza tra gli elementi di una serie storica: essa è testata rappresentando la funzione di autocorrelazione per differenti ritardi temporali. È innanzitutto necessario introdurre la funzione di autocorrelazione.

Sia $(x_1, \ldots, x_n)$ la serie storica di cui si vuole testare l'ipotesi i.i.d. Allora, si definisce innanzitutto l'autocovarianza con ritardo k come:

$$\gamma_k = \frac{1}{n} \sum_{t=1}^{n-k} (x_t - \overline{x}) (x_{t+k} - \overline{x}) \tag{5.1}$$

dove $\overline{x}$ è il valor medio della serie storica. Si noti che l'autocovarianza corrisponde alla varianza campionaria S^2 quando $k = 0$. L'autocorrelazione è quindi definita come segue:

$$\rho_k = \frac{\gamma_k}{S^2} \tag{5.2}$$

Misura la correlazione tra x_1 e x_k, al netto della correlazione lineare, analizzando l'andamento della serie storica negli elementi intermedi $(x_2, \ldots, x_k)$ rispetto al primo valore.

Quindi, l'autocorrelation plot si costruisce rappresentando su un grafico i valori di ρ_k per diversi valori del ritardo k. Inoltre, è utile inserire nel grafico i livelli corrispondenti a intervalli di confidenza del 95% [4]. Essi sono centrati in 0, mentre gli estremi sono:

$$\pm \frac{z_{1-\alpha/2}}{\sqrt{n}}$$

dove $z_{1-\alpha/2}$ è il quantile di ordine $1 - \alpha/2$ (nel nostro caso pari a 0.025) della distribuzione normale standard. L'identificazione di tali estremi è possibile in quanto si assume una distribuzione normale di media nulla e varianza pari a $1/n$ per i valori dell'autocorrelazione.

In definitiva, se i valori dell'autocorrelazione rimangono all'interno di tale range, l'ipotesi di indipendenza della serie storica è verificata. Ciò significa che i coefficienti di autocorrelazione sono significativamente vicini a 0.

Si è quindi effettuata tale analisi per i dati di perdite interne nella categoria ET 1 con importo superiore alla soglia H, che saranno poi utilizzati per modellizzare il corpo della distribuzione severity. In Figura 5.2 si riporta l'autocorrelation plot di tale serie storica, dove i dati sono ordinati per data di contabilizzazione dell'evento. Si osserva chiaramente come l'ipotesi di indipendenza dei dati sia rispettata, in quanto i valori dell'autocorrelazione sono all'interno dell'intervallo di confidenza di livello 95% per ogni valore di k.

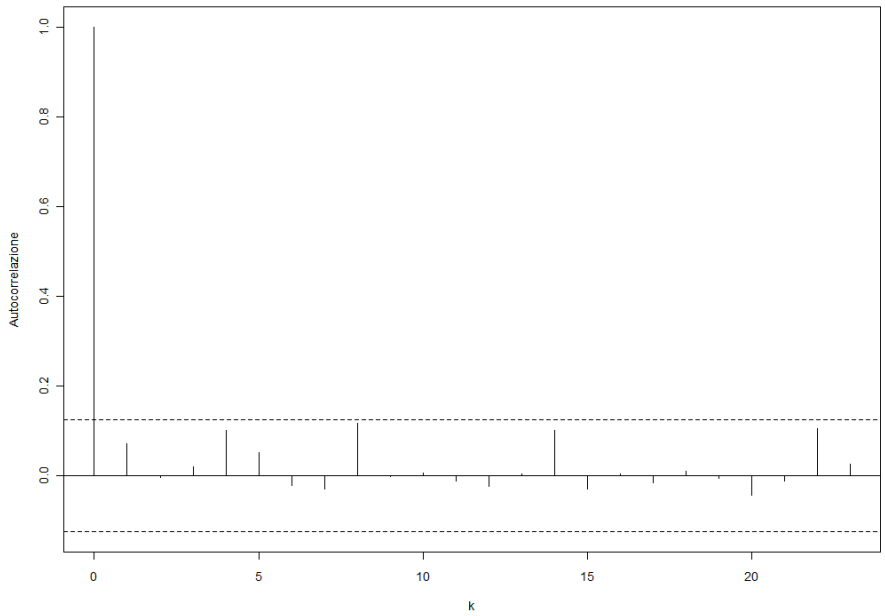

Figura 5.2: Autocorrelation plot dei dati di perdita interni superiori alla soglia H.

Il metodo quantitativo che si utilizza parallelamente per testare l'ipotesi di i.i.d. della serie in esame è il test di *Ljung-Box* [4]. Questo test sfrutta il concetto di autocorrelazione appena presentato, infatti la statistica utilizzata è una combinazione lineare dei coefficienti di autocorrelazione valutati a differenti valori di k:

$$LB = n\,(n+2) \sum_{k=1}^{L} \frac{\rho_k^2}{n-k} \tag{5.3}$$

dove i ρ_k sono stati definiti nella (4.11), e L è il ritardo massimo che si vuole considerare nel test. In assenza di ulteriori ipotesi sulla serie storica (ad esempio, periodicità dei dati) si fissa $L = \log n$.

Sotto l'ipotesi di indipendenza, è possibile dimostrare che tale statistica è distribuita come una χ^2 a L gradi di libertà, e quindi la regione di rifiuto dell'ipotesi nulla (ovvero di popolazione i.i.d.) sarà:

$$\left\{ z_1, \ldots, z_n : LB > \chi^2_{1-\alpha}(L) \right\}$$

dove α è il livello desiderato del test e $\chi^2_{1-\alpha}(L)$ il quantile di livello $1-\alpha$ della χ^2.

Applicando quindi il test di Ljung-Box al dataset delle perdite interne (n=245) si ottiene $LB = 4.49$, con conseguente p-value di 0.548. Questo risultato non permette perciò di rifiutare l'ipotesi di non correlazione nella serie storica in esame. Ovvero, è possibile applicare tutte le tecniche di stima presentate, in quanto esse sono valide solo assumendo indipendenza e stazionarietà per la serie storica in esame.

Dopo aver verificato l'ipotesi di i.i.d. dei dati interni sopra la soglia H, è quindi possibile passare alla stima dei parametri del corpo della severity.

5.2.1 Lognormale troncata

Si è innanzitutto portato avanti il fit dei dati interni con la distribuzione lognormale troncata: i parametri μ e σ^2 sono stati quindi stimati con l'algoritmo descritto nel paragrafo 2.4.1 applicato ai 245 dati interni superiori alla soglia H. Le stime risultanti sono le seguenti:

$$\begin{aligned} \widehat{\mu} &= 1.08 \\ \widehat{\sigma}^2 &= 2.11 \end{aligned}$$

Come visto nel Capitolo 2, per verificare il buon adattamento dei dati possono essere usati sia metodi qualitativi (Q-Q plot) che quantitativi (test di Kolmogorov-Smirnov e Anderson-Darling).

Si riporta innanzitutto in Figura (5.3) il Q-Q plot e relativo al fit del corpo della distribuzione di severity con i soli dati interni, dove i parametri sono stati stimati con il metodo MLE. Come si può osservare, il fit appare qualitativamente buono per i dati con perdite minori, mentre dalla soglia u_0 l'imprecisione della stima aumenta. Ciò non preoccupa dal punto di vista del modello, in quanto tali dati verranno modellizzati nella coda della distribuzione con la GPD. È comunque utile osservare il comportamento della stima del corpo della distribuzione anche per valori superiori a u_0 in

quanto, come anticipato, tale soglia è scelta a priori ma può essere modificata, come effettivamente accadrà, per ottenere un adattamento migliore della distribuzione GPD per i dati estremi.

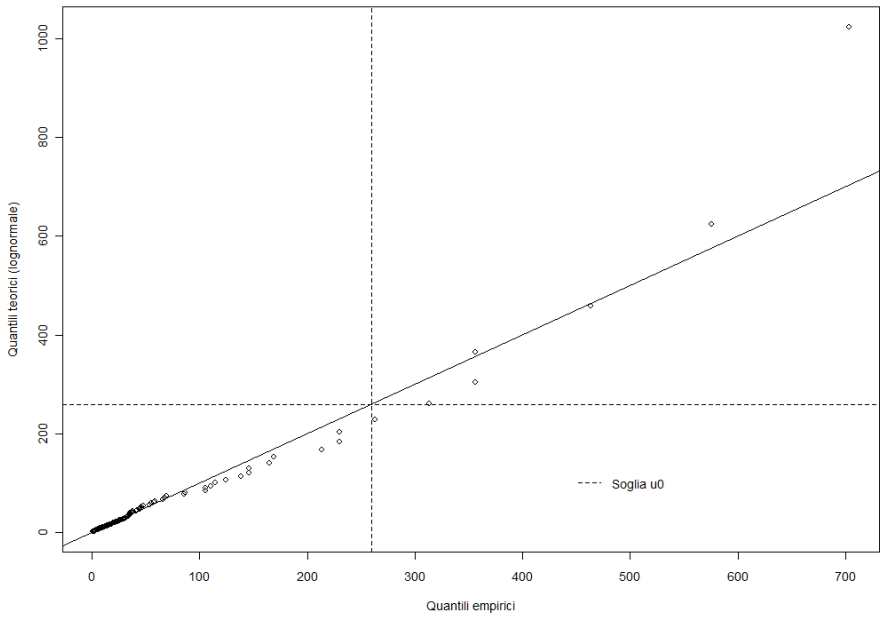

Figura 5.3: Q-Q plot relativo alla modellizzazione del corpo dell'ET 1 con la distribuzione lognormale troncata.

Conseguentemente, è possibile passare alla quantificazione della bontà del fit tramite i test per distribuzioni troncate introdotti in precedenza. I risultati sono riportati in Tabella 5.3. Come si osserva, i p-value dei test sono sufficientemente alti da non rifiutare l'ipotesi nulla, ovvero che la distribuzione dei dati sia lognormale, di parametri stimati con il metodo MLE. In particolare, il test con la maggiore significatività è quello di Anderson-Darling, con un p-value di 0.8. Ciò si spiega con il fatto che la statistica del test ADup dà maggiore peso alla coda superiore, che in questo caso gode di un fit peggiore, mentre il test AD pesa egualmente coda superiore ed inferiore, per cui si ha un fit decisamente migliore. Anche con il test di KS, che assegna uguale peso a tutti dati, ha un p-value sufficientemente alto, da poter garantire un buon adattamento della distribuzione ai dati a disposizione.

Test	Statistica	P-value
KS	0.562	0.575
AD	1.65	0.807
ADup	15.6	0.491

Tabella 5.3: Risultati dei test di bontà del fit del corpo della severity con la distribuzione lognormale troncata .

5.2.2 Weibull troncata

Lo stesso tipo di analisi è stato portato avanti anche assumendo una distribuzione Weibull troncata per il corpo della severity.

I parametri della Weibull, ovvero k e θ, ottenuti alla convergenza dell'algoritmo descritto nel paragrafo 2.4.2 sui dati a disposizione sono i seguenti:

$$\begin{aligned} \widehat{k} &= 0.298 \\ \widehat{\theta} &= 1.11 \end{aligned}$$

Si riporta in Figura 5.4 il Q-Q plot risultante da tali stime. Come osservato anche nella stima con la lognormale troncata, il fit appare buono soprattutto per le perdite inferiori alla soglia u_0, oltre il quale vi è una sottostima dei quantili teorici della Weibull. La scelta di questa distribuzione risulta ancora una volta giustificata dal fatto che i valori al di sopra di tale soglia saranno poi modellizzati con la teoria dei valori estremi.

Passando ai test quantitativi di bontà del fit, dai risultati riportati in Tabella 5.4 si può osservare come anche nel caso della Weibull non si possa rifiutare l'ipotesi per cui i dati sono distribuiti secondo tale distribuzione. In questo caso, tuttavia, il p-value più significativo si ha per il test di Kolmogorov-Smirnov: ciò è dovuto al fatto che nel caso della Weibull si ha un fit migliore nella parte centrale della distribuzione, piuttosto che alle code, per cui si ha un fit peggiore in particolare in quella destra. Le due varianti del test di Anderson-Darling, infatti, assegnano un peso maggiore ai dati estremi.

In ogni caso, la decisione presa per l'ET 1 nell'ambito di questo elaborato (e anche in sede di sviluppo all'interno della banca) sarà quella di scegliere la distribuzione Weibull per tale classe di rischio. Infatti, come già sottolineato, nel caso in cui siano accettabili entrambe le distribuzioni per il fit delle osservazioni, si preferisce mantenere quella scelta negli anni precedenti, in questo caso appunto la Weibull.

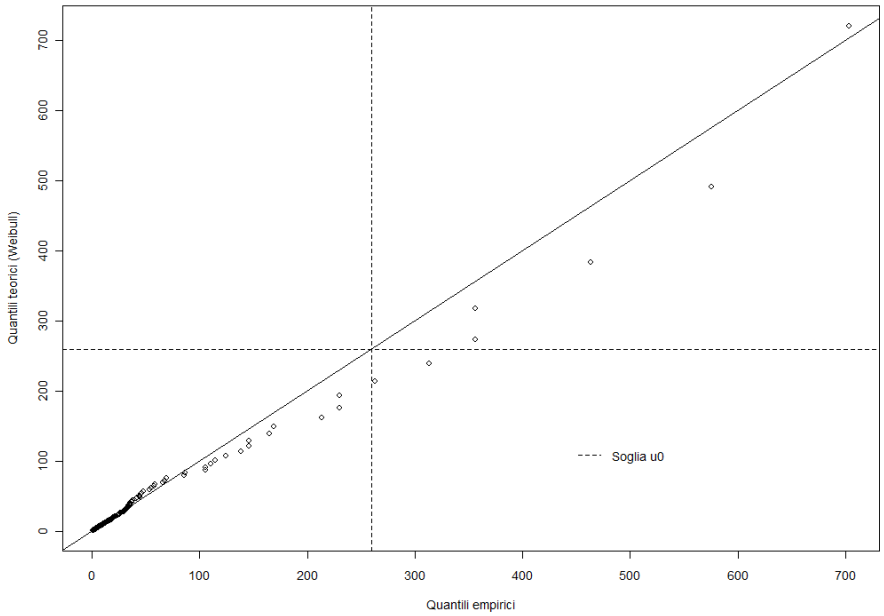

Figura 5.4: Q-Q plot relativo alla modellizzazione del corpo dell'ET 1 con la distribuzione Weibull troncata.

Test	Statistica	P-value
KS	0.479	0.836
AD	1.79	0.684
ADup	15.6	0.514

Tabella 5.4: Risultati dei test di bontà del fit del corpo della severity con la distribuzione Weibull troncata.

5.3 Modellizzazione della coda della severity

Una volta modellizzato il corpo con la distribuzione più adatta, e valutata la bontà del fit, è possibile quindi passare alla stima dei parametri della coda della severity dell'ET 1. Come discusso nel Capitolo 3, si sfrutta la teoria dei valori estremi per modellizzare con una distribuzione più adatta le perdite superiori ad una determinata soglia.

Inoltre, nel caso delle perdite operative, a causa della scarsità di dati estremi, è possibile integrare il dataset di perdite interne con i dati di altre banche (DIPO e dati pubblici) e con i dati di scenario prospettici, mantenendo unicamente il parametro ξ, che come visto nel paragrafo 3.5.1 è invariante rispetto al cambiamento di scala.

Si riportano innanzitutto in Figura 5.5 i box-plot (in scala logaritmica e non) di tali dati. Anche nel caso delle perdite estreme si nota una forte asimmetria destra: ciò era prevedibile in quanto ci si trova già nella coda destra della distribuzione, ed inoltre la distribuzione GPD si utilizza proprio perché è più concentrata ai valori estremi.

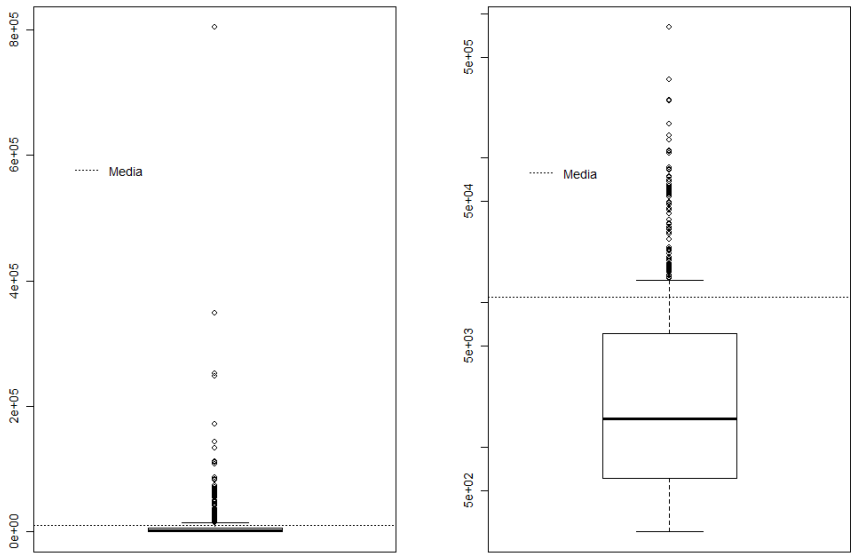

Figura 5.5: Box-plot (sinistra) e box-plot in scala logaritmica (destra) dei dati di perdita interni, esterni e di scenario superiori alla soglia u_0.

In questa sezione si discuterà quindi l'adattamento di tale set di dati alla Generalized Pareto Distribution (GPD). Prima di passare alla stima dei parametri di tale distribuzione, è però necessario valutare l'adeguatezza della scelta del parametro u_0 fatta a priori ($u_0 = 259$) come soglia di raccolta per i dati esterni. È possibile farlo innanzitutto sfruttando la proprietà (3.13), ovvero la linearità della media degli eccessi $(x_i - v | x_i \geq v)$ facendo variare la soglia $v \geq u_0$. Per individuare la soglia ottimale, ovvero per cui si evidenzia maggiormente tale andamento, per ogni valore di v è stata effettuata una regressione lineare per la media degli eccessi al variare della soglia al di sopra del singolo valore di v. La bontà della regressione è stata quantificata con il valore R^2. Come si osserva in Figura 5.6, che riporta i diversi valori di R^2 al variare di v, la regressione lineare migliore si ha in corrispondenza del valore $u_1 = 2798$.

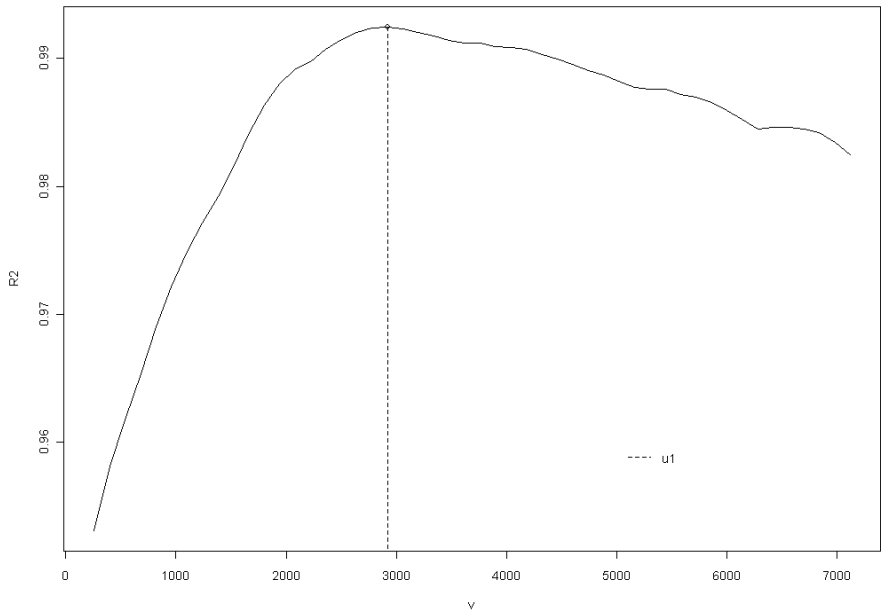

Figura 5.6: Identificazione della corretta soglia per i valori estremi, corrispondente al massimo valore di R^2 relativo alla regressione tra le medie degli eccessi e le diverse soglie.

Per valutare ulteriormente la bontà della regressione in corrispondenza della soglia u_1, in Figura 5.7 si riportano le analisi dei relativi residui: dallo scatterplot dei residui non si osserva alcun trend apparente e la me-

dia mobile sembra assestarsi intorno al valore zero, così come il Q-Q plot conferma la distribuzione normale dei residui.

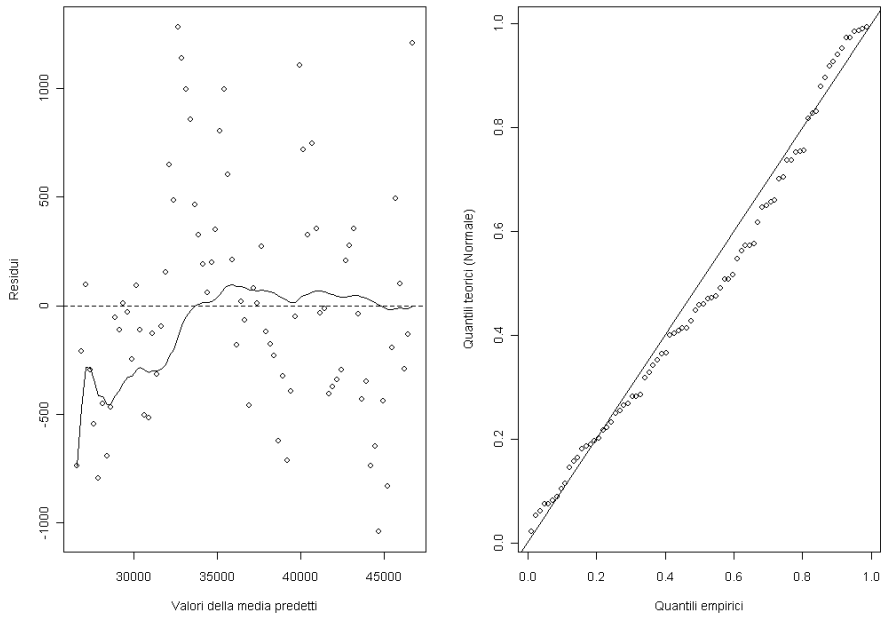

Figura 5.7: Analisi dei residui della regressione tra le medie degli eccessi e la variazione della soglia al di sopra del valore u_1. Sinistra: scatterplot dei residui. Destra: Q-Q plot per verificarne la normalità.

Si riporta quindi in Figura 5.8 il grafico delle medie empiriche al variare della soglia. Da questo si osserva come l'andamento lineare non si ottenga esattamente per tutti i valori superiori a u_0, bensì si ha un andamento lineare a partire dalla soglia u_1, corrispondente come detto al massimo valore R^2.

Per quanto riguarda la stima dei parametri della GPD, come abbiamo visto nella sezione 3.5, esistono diversi metodi. Si applicheranno in questo ambito solo il metodo di massima verosimiglianza (MLE) e dei probability weighted moments (PWM), essendo quest'ultimo una generalizzazione del metodo dei momenti. Sfruttando gli algoritmi presentati, si sono portate avanti le stime dei parametri ξ e β per diversi valori di v, per valutare ancora una volta l'adeguatezza di tale soglia. Anche tali analisi rafforzano la conclusione precedente, ovvero la necessità di un aumento della soglia da u_0 a u_1.

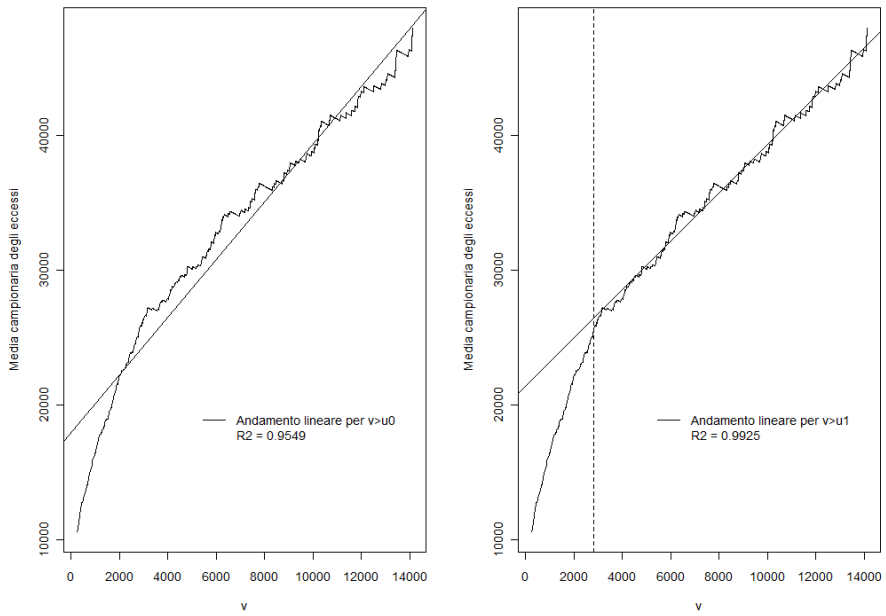

Figura 5.8: Grafico delle medie empiriche degli eccessi al variare della soglia al di sopra di u_0, con l'andamento lineare rispettivamente valutato dalla soglia di raccolta dei dati esterni e da u_1.

In Figura (5.9) si riporta l'andamento della stima del parametro di scala β al variare della soglia (ovvero stimando il parametro solo con i dati superiori a tale soglia), utilizzando i metodi MLE e PWM. Si ricorda innanzitutto che dalla (3.14) si è dimostrato un rapporto lineare tra β e v, quando v viene fatto variare al di sopra della "vera" soglia. Il grafico mostra chiaramente come si abbia un comportamento lineare solo al di sopra della soglia u_1, con entrambi i metodi. Ciò è confermato dal fatto che i valori R^2 delle regressioni lineari siano vicini a 1.

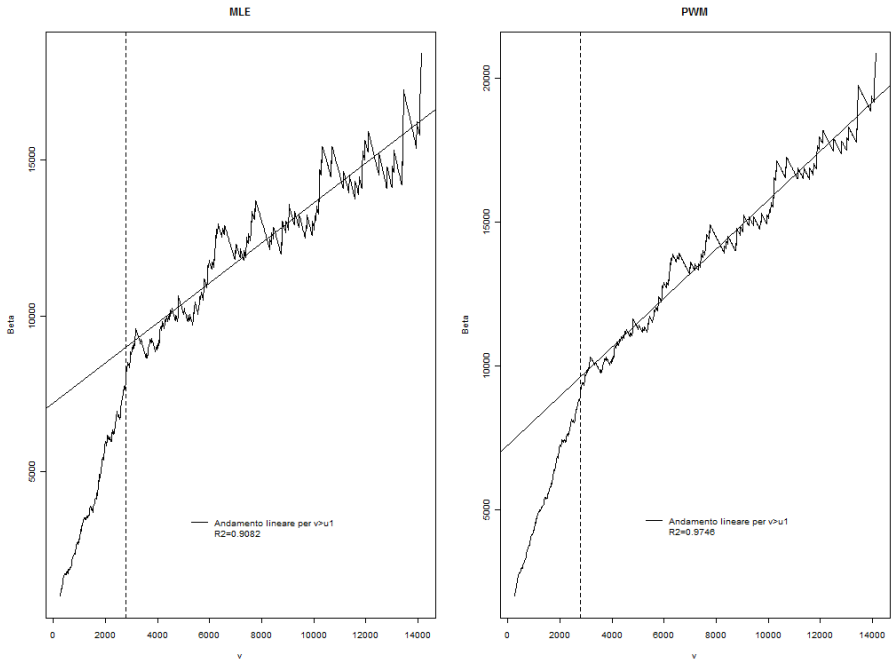

Figura 5.9: Grafico delle stime di β con MLE e PWM al variare della soglia.

La bontà della regressione lineare tra la stima di β e le diverse soglie è confermata dalle analisi sui residui riportate in Figura 5.10 per il metodo MLE e in Figura 5.11 per il metodo PWM: sia lo scatterplot che il Q-Q plot supportano l'ipotesi di normalità dei residui.

Per quanto riguarda il parametro di forma ξ, con la relazione (3.14) si era mostrato come questo parametro fosse costante rispetto alla variazione della soglia. In Figura (5.12) sono quindi riportate le stime di tale parametro per diverse soglie, con i due metodi. In questo caso, l'andamento con il metodo MLE mostra maggiormente come la soglia più adatta sia u_1, mentre non lo si può concludere dal grafico del metodo PWM.

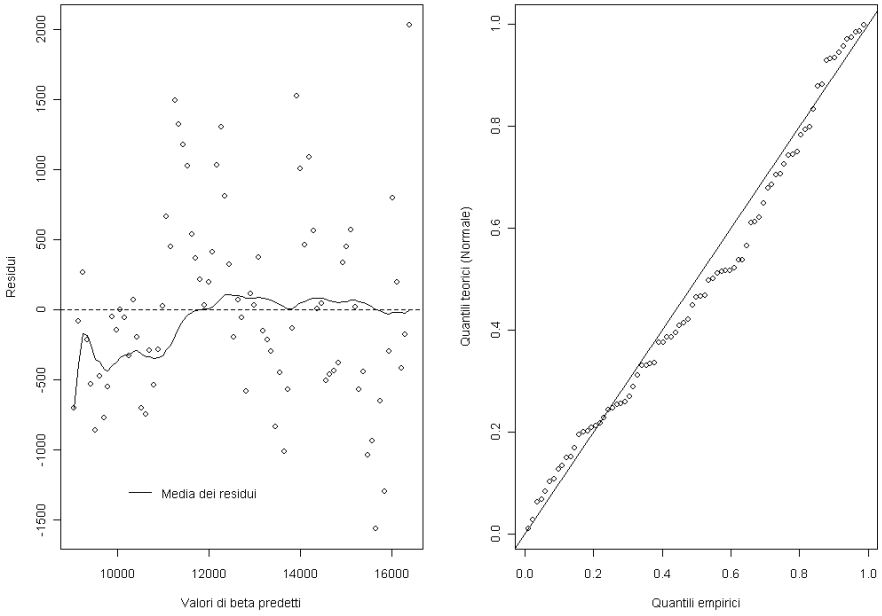

Figura 5.10: Analisi dei residui della regressione tra le stime di β con il metodo MLE e la variazione della soglia al di sopra del valore u_1. Sinistra: scatterplot dei residui. Destra: Q-Q plot per verificarne la normalità.

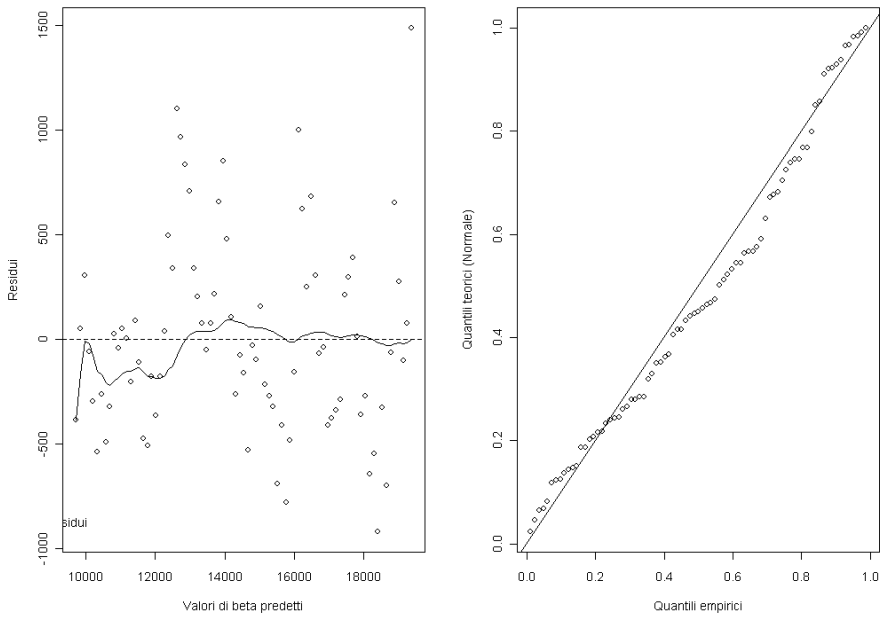

Figura 5.11: Analisi dei residui della regressione tra le stime di β con il metodo PWM e la variazione della soglia al di sopra del valore u_1. Sinistra: scatterplot dei residui. Destra: Q-Q plot per verificarne la normalità.

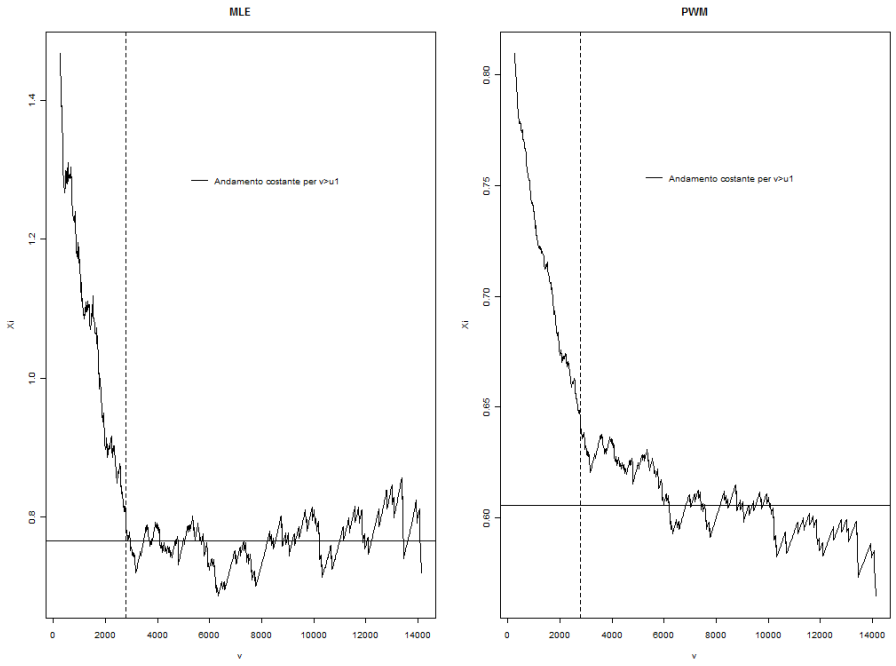

Figura 5.12: Grafico delle stime di ξ con MLE e PWM al variare della soglia.

Anche l'analisi dell'andamento della stima dei parametri con i due diversi metodi di stima conferma la necessità di una variazione della soglia u, supportata anche dai risultati di bontà del fit dei dati con la soglia u_0 e i due diversi metodi di stima, riportati in Tabella 5.5.

Metodo di stima	KS	AD
MLE	0.005	0.17
PWM	0.01	0.14

Tabella 5.5: Risultati dei test di bontà del fit della della severity con la GPD e diversi metodi di stima dei parametri, con la soglia fissata a u_0

Le analisi condotte sulla variazione della soglia supportano quindi la decisione di adottare la nuova soglia u_1 come punto discriminante tra corpo e coda della distribuzione di severity dell'ET1. Si noti che ciò non implica che tale soglia sarà valida per tutte le altre tipologie di rischio: è infatti possibile adottare una soglia differente per ciascuna, nonostante la soglia

di raccolta dei dati esterni e di scenario scelta dalla banca sia la medesima per tutti gli ET.

Passando quindi alla stima dei parametri della GPD con i due diversi metodi presentati (MLE e GPD). I valori ottenuti con il metodo di massima verosimiglianza sono i seguenti:

$$\begin{aligned} \widehat{\xi}_{MLE} &= 0.779 \\ \widehat{\beta}_{MLE} &= 8248 \end{aligned}$$

Si riportano inoltre in Figura 5.13 i grafici relativi ai residui del fit dei dati a disposizione utilizzando il metodo MLE. Per valutare qualitativamente il buon adattamento dei dati utilizzati alla GPD, dallo scatterplot dei residui non si osserva nessun trend particolare (considerando ad esempio la media mobile), mentre dal Q-Q plot degli stessi si può assumere valido un loro fit con la distribuzione esponenziale.

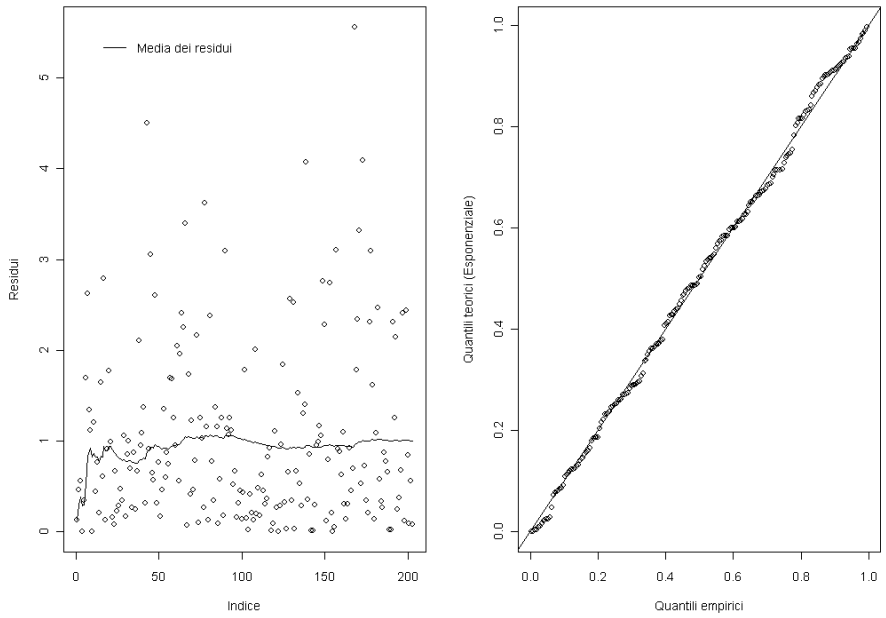

Figura 5.13: Analisi dei residui relativi al fit della coda della distribuzione di severity con il metodo MLE.

I parametri stimati con il metodo PWM sono invece i seguenti:

$$\begin{aligned} \widehat{\xi}_{PWM} &= 0.639 \\ \widehat{\beta}_{PWM} &= 9230 \end{aligned}$$

e in Figura 5.14 sono riportati i relativi grafici dei residui. Anche da questi, come nel caso della stima con il metodo MLE, sembra verificato il buon adattamento dei dati. Non è quindi possibile scartare a priori uno dei due metodi di stima, saranno quindi discriminanti i test quantitativi e gli altri metodi qualitativi.

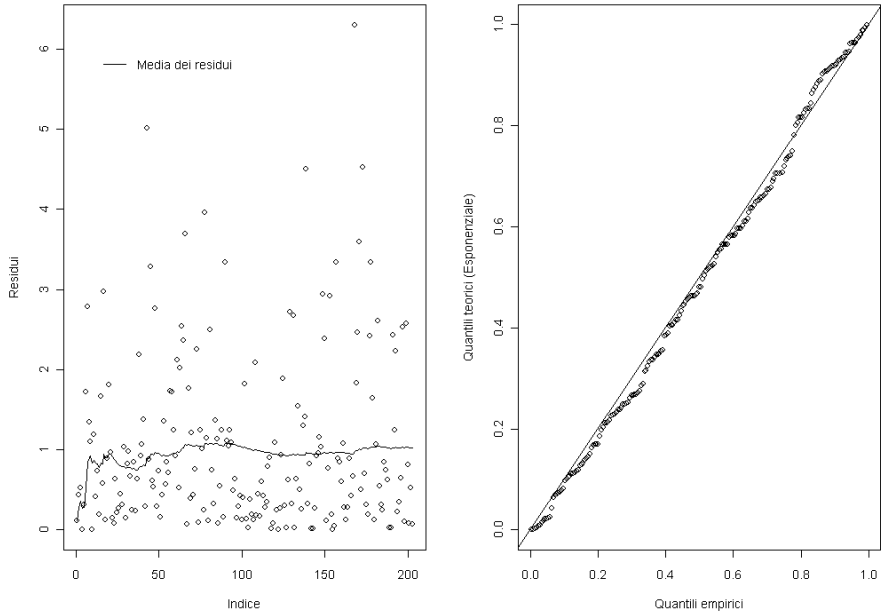

Figura 5.14: Analisi dei residui relativi al fit della coda della distribuzione di severity con il metodo PWM.

Sono stati quindi portati avanti i test di bontà del fit di Kolmogorov-Smirnov e Anderson-Darling con parametri stimati con i due diversi metodi; i p-value risultanti sono riportati in Tabella 5.6. Si può osservare come il fit con entrambi i metodi di stima sia praticamente perfetto.

Ciò è confermato anche dai Q-Q plot, riportati in Figura 5.15.

In definitiva, entrambi i metodi di stima mostrano un fit ottimo con il valore della soglia fissato a u_1, il che mostra nuovamente come la scelta della nuova soglia sia migliore rispetto a quella iniziale (dal confronto con i risultati in Tabella 5.5). Per scegliere quale stima utilizzare la decisione viene presa sulla base della stabilità della stima dei parametri (Figure 5.9 e 5.12). In particolare, per questa classe di rischio, la scelta ricade sul metodo MLE, in quanto, soprattutto per quanto riguarda ξ, l'andamento è molto

Metodo di stima	KS	AD
MLE	0.99	0.95
PWM	0.98	0.97

Tabella 5.6: Risultati dei test di bontà del fit della della severity con la GPD e diversi metodi di stima dei parametri, con la soglia fissata a u_1

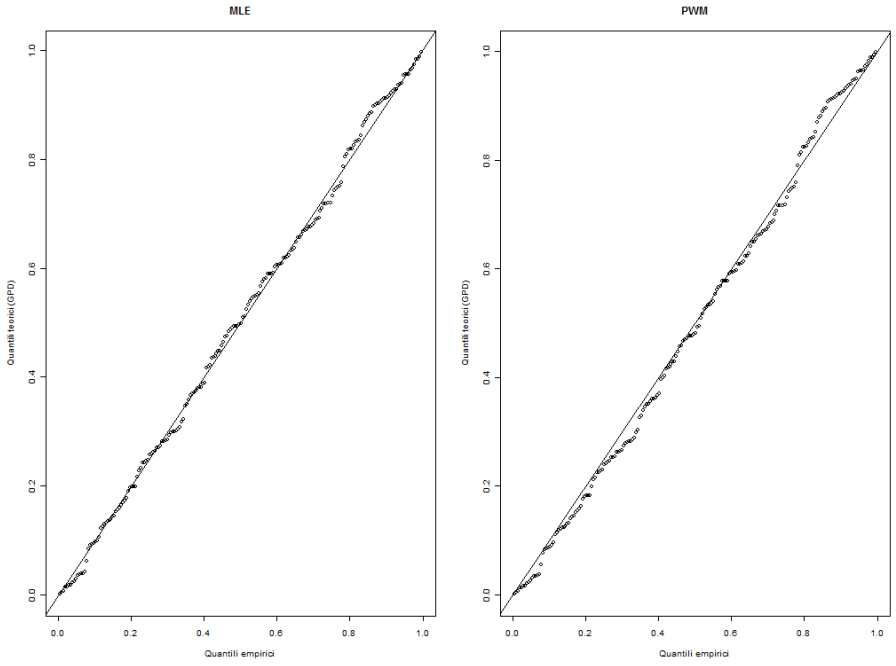

Figura 5.15: Confronto dei Q-Q plot relativi al fit della coda della severity con i metodi MLE e PWM.

più regolare, e rispetta le ipotesi teoriche di andamento costante al variare della soglia.

Tuttavia, come accennato in precedenza, la stima di β ottenuta utilizzando, oltre ai dati interni con importo superiore a u_1, anche dati esterni e di scenario è distorta. Questo perché tale parametro rappresenta la scala della GPD, è un indicatore dell'ordine di grandezza dei dati estremi, che dev'essere quindi riscalato sulla base dei dati esterni, in modo da avere continuità tra corpo e coda della distribuzione di severity, come imposto dalla (1.8). L'unica stima che si utilizza sarà quindi quella del parametro di forma ξ, che come abbiamo visto nel Capitolo 3 è invariante rispetto a cambiamenti di scala. Dalla (1.9) si ottiene quindi la nuova stima di β.

$$\widehat{\beta}_{body} = \frac{1 - F_{body}(u_1)}{f_{body}(u_1)} = 976.9$$

Si può osservare come tale stima di β sia inferiore di un'ordine di grandezza rispetto alle stime ottenute con i dati esterni e di scenario: ciò significa che questi ultimi rappresentano perdite molto più estreme di quelle in cui è incorsa la banca in esame, da cui la necessità di riscalare il parametro di scala della GPD.

Con la modellizzazione della coda e applicando la condizione di continuità si arriva dunque alla definizione di una distribuzione di severity data dalla mistura di Weibull (per valori tra H e u_1) e GPD (per valori superiori a u_1). È quindi possibile passare alle analisi sulla distribuzione di frequency, ovvero del numero di perdite operative annue per classe di rischio.

5.4 Convoluzione tra severity e frequency

Come accennato nell'introduzione, la frequenza di accadimento degli eventi operativi viene modellizzata con una Poisson. Il parametro λ di tale distribuzione viene stimato con il metodo dei momenti. Una prima stima di tale parametro si ottiene considerando la media del numero di eventi operativi annui della classe di rischio considerata per i cinque anni di cui si hanno a disposizione i dati: n_i, per i=1,...,5. Da questa:

$$\widehat{\lambda} = \frac{1}{5}\sum_{i=1}^{5} n_i = \frac{471}{5} = 94.2$$

Tale valore corrisponde alla stima per la distribuzione di frequency delle perdite operative di qualunque importo. Tuttavia, come visto in prece-

denza, le perdite sotto la soglia H non sono considerate affidabili, e quindi anche la stima di λ appena vista.

Per tale motivo si utilizzano unicamente le frequenze annue di accadimento di eventi operativi con perdite superiori alla soglia H (ovvero n_i') In particolare si ottiene:

$$\widehat{\lambda}_{sample} = \sum_{i=1}^{M} n_i' = \frac{245}{5} = 49$$

Per costruzione, tale valore non tiene conto della frequenza delle perdite sotto la soglia di raccolta. Il loro contributo verrà poi considerato separatamente nella simulazione Monte Carlo. In ogni caso, per avere una stima più affidabile della frequenza di eventi di qualunque importo, è possibile utilizzare la (1.10) utilizzando la distribuzione di severity individuata in precedenza:

$$\widehat{\lambda} = \frac{\widehat{\lambda}_{sample}}{1 - F_{body}(H)} = \frac{49}{0.317} = 120.49$$

Chiaramente, tale stima è superiore alla precedente in quanto presuppone l'esistenza di eventi operativi di importo inferiore ad H non contabilizzati nel dataset utilizzato.

A partire dalla stima della frequenza annua, è possibile definire la frequenza di perdite per il corpo e per la coda, come segue:

$$\begin{aligned} \widehat{\lambda}_{body} &= \widehat{\lambda}\left[F_{body}(u_1) - F_{body}(H)\right] \\ \widehat{\lambda}_{tail} &= \widehat{\lambda}\left[1 - F_{body}(u_1)\right] \end{aligned}$$

Il passo successivo è la costruzione della distribuzione delle perdite aggregate annue, dalla convoluzione tra frequency e severity. Prima di passare alla simulazione, è necessario verificare l'ipotesi di indipendenza tra le due distribuzioni, ipotesi necessaria per adottare l'approccio attuariale per la convoluzione. Per farlo, si utilizzano l'aggregazione mensile delle perdite, per avere un numero di dati sufficientemente alto da avere significatività per i test utilizzati.

In particolare, si vuole testare l'indipendenza tra il numero di perdite e l'impatto medio di tali perdite, aggregate mensilmente. Si riporta innanzitutto in Figura 5.16 il grafico delle coppie frequenza-perdita media per tutti i mesi a disposizione, con in evidenza la possibile retta di regressione tra le due serie storiche. Come si può osservare, non vi è alcun trend apparente nei dati, ed inoltre la stima del coefficiente di regressione è praticamente nulla. L'assenza di dipendenza lineare è supportata anche dal valore di R^2, pari a 0.009.

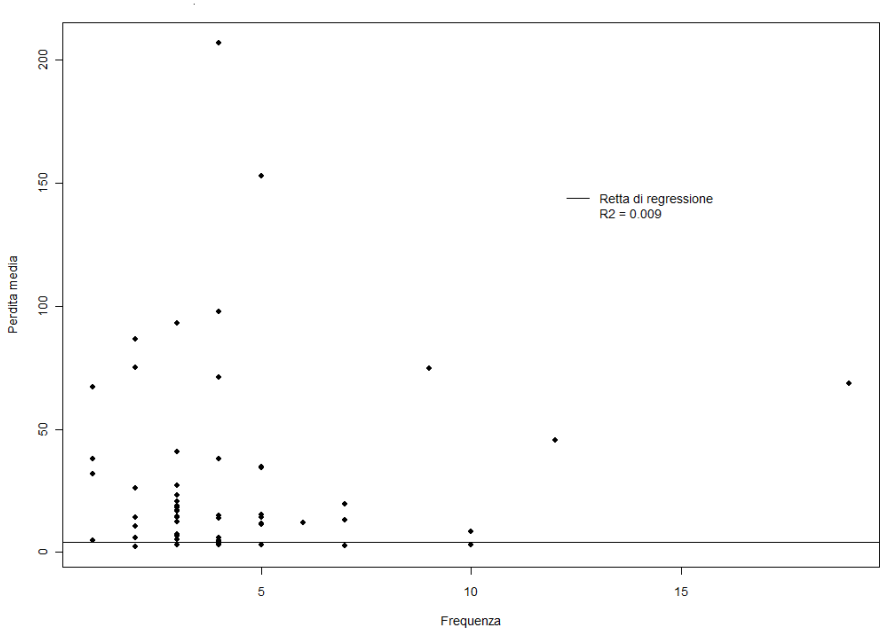

Figura 5.16: Scatterplot e retta di regressione tra frequenze e perdite medie mensili.

Per escludere qualsiasi tipo di dipendenza tra frequency e severity, è stata portata avanti un'analisi sfruttando il Tau di Kendall, e in particolare la proprietà, presentata nella sezione 4.2.2, secondo cui τ_K può essere approssimato con una normale di media nulla e varianza data dalla (4.17). Si riporta quindi la stima del Tau di Kendall tra frequenze e perdite medie mensili, insieme agli estremi dell'intervallo di confidenza di livello 95%:

$$\begin{aligned} \widehat{\tau_k} &= -0.0192 \\ IC_{95\%}(\tau_k) &= [-0.287; 0.168] \end{aligned}$$

Come si può notare, lo 0 è compreso nell'intervallo, perciò, fissando il livello del test al 5%, non è possibile rifiutare l'ipotesi di indipendenza tra frequency e severity.

Un ulteriore test non parametrico che può essere effettuato per testare l'indipendenza tra le due distribuzioni è il test di Hoeffding. Date due variabili aleatorie X e Y con funzioni di ripartizione marginali F_X e F_Y e distribuzione congiunta F_{XY}, tale test non-parametrico ha come ipotesi nulla l'indipendenza tra le due variabili, e come ipotesi alternativa la presenza di una qualche dipendenza. Per determinare la statistica del test, Hoeffding parte dall'idea che vale la seguente relazione:

$$D(x,y) = F_{XY}(x,y) - F_X(x)F_Y(y) = 0$$

se e solo se le due variabili sono indipendenti. Siano quindi $x_1, \ldots, x_n$ e $y_1, \ldots, y_n$ n realizzazioni i.i.d. delle due variabili X e Y. La statistica del test di è una stima non-paremetrica della quantità $\int D^2(x,y)\mathrm{d}F_{XY}(x,y)$, ottenibile come segue [16]:

$$D_n = \frac{Q - 2(n-2)R + (n-2)(n-3)S}{n(n-1)(n-2)(n-3)(n-4)} \tag{5.4}$$

dove

$$\begin{aligned} Q &= \sum_{i=1}^{n}(r_i - 1)(r_i - 2)(s_i - 1)(s_i - 2) \\ R &= \sum_{i=1}^{n}(r_i - 2)(s_i - 2)c_i \\ S &= \sum_{i=1}^{n}(c_i - 1)c_i \end{aligned}$$

dove, per $i = 1, \ldots, n$, r_i e s_i sono rispettivamente i ranghi delle osservazioni x_i e y_i, mentre i c_i sono il numero di coppie di osservazioni (x_j, y_j)

tali che $x_j \leq x_i$ e $y_j \leq y_i$. Hoeffding ha determinato l'esatta distribuzione (sotto l'ipotesi nulla) di D_n nel caso in cui $n = 5, 6, 7$, e tramite simulazione per diverse numerosità del campione.

Tale test è stato quindi applicato, utilizzando il software R [11], per verificare l'indipendenza tra le distribuzioni di frequency e severity, oltre che nel caso di aggregazione mensile, anche per l'aggregazione annua, in quanto la significatività del test è garantita dalla conoscenza della distribuzione esatta della statistica (in questo caso infatti $n = 5$). In Tabella 5.7 si riportano i valori delle statistiche e dei p-value asintotici restituiti dal test. Come si osserva, in entrambi i casi il p-value asintotico è pari a 1, che porta quindi all'accettazione dell'ipotesi nulla, ovvero di indipendenza tra le due distribuzioni.

Hoeffding	Statistica	P-value
Mensile	-0.012	1
Annuale	-0.375	1

Tabella 5.7: Risultati dei test di Hoeffding per valutare l'indipendenza tra le distribuzioni di frequency e severity, per aggregazione mensile e annuale delle frequenze e delle perdite medie.

Una volta verificata l'indipendenza tra frequency e severity, è possibile passare alla convoluzione tra le due distribuzioni. Come anticipato nell'introduzione, l'algoritmo di simulazione Monte Carlo per la convoluzione prevede il campionamento successivo dalla distribuzione di frequency e di severity (corpo e coda separatamente) per poi ottenere la distribuzione aggregata. Sia quindi N il numero di simulazioni Monte Carlo da effettuare (generalmente dell'ordine di 10^6 per ottenere una distribuzione empirica adeguata).

Quindi, per $i = 1, \ldots, N$, l'algoritmo di simulazione è il seguente:

1. Si campiona il numero di perdite annue per il corpo

$$m_i \sim Poi(\widehat{\lambda}_{body});$$

2. Si campiona l'ammontare di ogni singola perdita dalla distribuzione del corpo:

$$b_{ij} \sim Wei_H(\widehat{\theta}, \widehat{k})$$

per $j = 1, \ldots, m_i$;

3. Si campiona il numero di perdite annue per la coda

$$n_i \sim Poi(\widehat{\lambda}_{tail});$$

4. Si campiona l'ammontare di ogni singola perdita dalla distribuzione della coda:

$$t_{ij} \sim GPD(u_1, \widehat{\beta}_{body}, \widehat{\xi})$$

per $j = 1, \ldots, n_i$;

5. Si ottiene la perdita totale annua:

$$S_i = \sum_{j=1}^{m_i} b_{ij} + \sum_{j=1}^{n_i} t_{ij} + s'$$

dove s' è la media empirica delle perdite sotto la soglia H aggregate annualmente.

In definitiva, le quantità $S_1, \ldots, S_N$ rappresentano la distribuzione empirica delle perdite aggregate annue per ogni singolo ET. La simulazione è stata fatta con N pari a 5 milioni.

Si riporta in Figura 5.17 l'istogramma normalizzato relativo all'ET1. Come ci si aspettava, la distribuzione appare asimmetrica, ovvero la densità è più concentrata nei valori estremi, a causa dell'utilizzo della GPD per il fit della coda della distribuzione di severity.

Da tale distribuzione è anche possibile calcolare l'Expected Loss e il Value at Risk della singola classe, definiti nella sezione 1.5 rispettivamente come la media e il quantile di livello 99.9% della distribuzione aggregata simulata:

$$\begin{aligned} EL_1 &= 704.44 \\ VaR_1 &= 6715.59 \end{aligned}$$

Come detto in precedenza, tutte le analisi descritte sono state portate avanti per le 7 categorie di rischio operativo, ottenendo quindi per ognuna una distribuzione empirica di perdite annue. Il passo successivo per la determinazione del capitale a rischio è quindi l'aggregazione degli ET, per determinare quindi la distribuzione aggregata totale.

Lo strumento principale per l'aggregazione delle diverse classi è la teoria delle copule, presentata nel Capitolo 4. In particolare, si andrà innanzitutto a determinare la struttura di correlazione, basandosi sulle serie storiche, per poi determinare il tipo di copula più adatto.

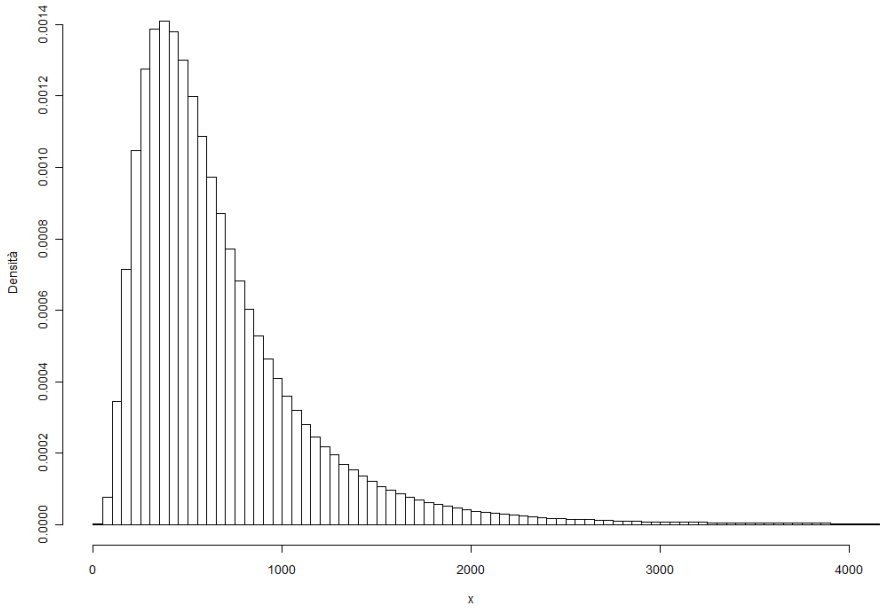

Figura 5.17: Istogramma normalizzato relativo alla simulazione delle perdite aggregate annue dell'ET1.

5.5 Identificazione della struttura di dipendenza

In primo luogo, per per determinare la struttura dipendenza, è necessario usare i dati dei singoli ET aggregati: dal punto di vista teorico, la scelta immediata è calcolare le dipendenze sulla base delle perdite annue delle singole classi di rischio, utilizzando uno dei metodi di stima presentati nella sezione 4.2. Tuttavia, a causa dei pochi dati a disposizione, tali stime verrebbe portata avanti su un numero troppo basso di coppie da confrontare (5 perdite annue per ET). Per tale motivo è pratica comune utilizzare i dati di perdita aggregati mensilmente per data di contabilizzazione. Così facendo, infatti, si otterranno 60 osservazioni di perdite mensili per ogni ET. l'analisi verrà portata avanti parallelamente con i due periodi di aggregazione, per individuare la scelta migliore.

Si riportano innanzitutto le stime ottenute con l'aggregazione mensile, utilizzando i due metodi presentati nella trattazione teorica delle copule. Innanzitutto, per il Tau di Kendall, utilizzando la formula (4.16) si ottiene:

$$\widehat{\tau}_K = \begin{bmatrix} 1 & -0.046 & 0.309 & 0.322 & 0.121 & -0.037 & 0.326 \\ & 1 & 0.091 & -0.047 & -0.033 & 0.183 & -0.042 \\ & & 1 & 0.476 & 0.181 & 0.073 & 0.368 \\ & & & 1 & 0.14 & 0.069 & 0.395 \\ & & & & 1 & 0.051 & 0.293 \\ & & & & & 1 & 0.005 \\ & & & & & & 1 \end{bmatrix}$$

mentre, per la stima con il Rho di Spearman, sfruttando la (4.21):

$$\widehat{\rho}_S = \begin{bmatrix} 1 & 0.125 & -0.084 & -0.044 & 0.274 & -0.061 & -0.089 \\ & 1 & 0.662 & 0.252 & 0.096 & 0.534 & 0.459 \\ & & 1 & 0.222 & 0.092 & 0.568 & 0.471 \\ & & & 1 & 0.087 & 0.416 & 0.171 \\ & & & & 1 & 0.008 & -0.043 \\ & & & & & 1 & 0.447 \\ & & & & & & 1 \end{bmatrix}$$

Come si può osservare immediatamente dal confronto tra le due matrici, e come anticipato in precedenza, i due metodi di stima giustamente non differiscono dal punto di vista del segno della correlazione.

Tuttavia, supponendo per il momento che la distribuzione di perdita delle 7 classi di rischio si possa descrivere con una copula ellittica, è possibile sfruttare le relazioni di dipendenza di tali parametri con la matrice

di correlazione lineare riportate nelle equazioni (4.28) e (4.29), per confrontare la stima di R con i due metodi, ottenendo, per la stima con il Tau di Kendall:

$$\widehat{R}_\tau = \begin{bmatrix} 1 & 0.131 & -0.088 & -0.046 & 0.286 & -0.064 & -0.093 \\ & 1 & 0.679 & 0.263 & 0.101 & 0.552 & 0.476 \\ & & 1 & 0.231 & 0.096 & 0.586 & 0.487 \\ & & & 1 & 0.091 & 0.432 & 0.178 \\ & & & & 1 & 0.008 & -0.045 \\ & & & & & 1 & 0.464 \\ & & & & & & 1 \end{bmatrix}$$

mentre, per il Rho di Spearman:

$$\widehat{R}_\rho = \begin{bmatrix} 1 & 0.142 & -0.074 & -0.051 & 0.284 & -0.066 & -0.072 \\ & 1 & 0.681 & 0.281 & 0.115 & 0.547 & 0.467 \\ & & 1 & 0.218 & 0.108 & 0.582 & 0.484 \\ & & & 1 & 0.078 & 0.44 & 0.188\,4 \\ & & & & 1 & 0.007 & -0.058 \\ & & & & & 1 & 0.491 \\ & & & & & & 1 \end{bmatrix}$$

Si osserva come le due stime siano consistenti. Infatti, calcolando l'errore tra le due, si osserva tra le due stime una differenza massima di 0.026, sufficientemente piccola da poter usare indifferentemente una delle due stime, almeno nel caso delle copule ellittiche. Si è inoltre calcolata la norma spettrale della differenza tra le due matrici di correlazione stimate; essa, per una matrice simmetrica (come nel caso delle matrici di correlazione), è definita come il valore assoluto dell'autovalore massimo in modulo. Tale norma, per la matrice differenza tra le stime con tau_K e ρ_S, risulta pari a 0.061.

Utilizzando invece l'aggregazione si base annua, si ottengono le seguen-

ti stime:

$$\widehat{\tau}_K = \begin{bmatrix} 1 & 0 & -0.4 & -0.2 & -0.2 & -0.2 & -0.6 \\ & 1 & 0.6 & 0 & -0.8 & 0.4 & 0.4 \\ & & 1 & 0.4 & -0.4 & 0.8 & 0.8 \\ & & & 1 & -0.2 & 0.6 & 0.6 \\ & & & & 1 & -0.6 & -0.2 \\ & & & & & 1 & 0.6 \\ & & & & & & 1 \end{bmatrix}$$

$$\widehat{\rho}_S = \begin{bmatrix} 1 & -0.1 & -0.6 & -0.3 & -0.2 & -0.3 & -0.8 \\ & 1 & 0.7 & 0 & -0.9 & 0.6 & 0.4 \\ & & 1 & 0.6 & -0.6 & 0.9 & 0.9 \\ & & & 1 & -0.2 & 0.8 & 0.7 \\ & & & & 1 & -0.7 & -0.3 \\ & & & & & 1 & 0.8 \\ & & & & & & 1 \end{bmatrix}$$

In questo caso, dalle due matrici si può notare come il segno sia preservato in tutti i casi, tranne quando i (pochi) dati utilizzati fanno supporre un'indipendenza tra coppie di ET. Tuttavia, rispetto alle stime con l'aggregazione mensile, che sono da considerare più significative per il maggior numero di dati utilizzati, sono diversi i casi in cui non si preserva neanche il segno della correlazione. Questo è preoccupante soprattutto nel momento in cui si utilizza una correlazione negativa invece che positiva, portando a una notevole sottostima del capitale a rischio.

Passando quindi alla matrice di correlazione lineare, supponendo sempre di essere nel caso di copule ellittiche, si ha:

$$\widehat{R}_\tau = \begin{bmatrix} 1 & 0 & -0.588 & -0.309 & -0.309 & -0.309 & -0.809 \\ & 1 & 0.809 & 0 & -0.951 & 0.588 & 0.588 \\ & & 1 & 0.588 & -0.588 & 0.951 & 0.951 \\ & & & 1 & -0.309 & 0.809 & 0.809 \\ & & & & 1 & -0.809 & -0.309 \\ & & & & & 1 & 0.809 \\ & & & & & & 1 \end{bmatrix}$$

$$\widehat{R}_\rho = \begin{bmatrix} 1 & -0.105 & -0.618 & -0.313 & -0.209 & -0.313 & -0.813 \\ & 1 & 0.717 & 0 & -0.908 & 0.618 & 0.416 \\ & & 1 & 0.618 & -0.618 & 0.908 & 0.908 \\ & & & 1 & -0.209 & 0.813 & 0.717 \\ & & & & 1 & -0.717 & -0.313 \\ & & & & & 1 & 0.813 \\ & & & & & & 1 \end{bmatrix}$$

Nel caso di aggregazione annuale, la differenza massima tra le matrici di correlazione lineare stimate con i due diversi metodi è di 0.17, che appare troppo alta rispetto ai piccoli valori di correlazione lineare ottenuti. Allo stesso modo, la norma spettrale della matrice differenza è pari a 0.31, di un'ordine di grandezza superiore al caso di aggregazione mensile.

Inoltre, è stata portata avanti un'analisi di stabilità delle stime sfruttando la proprietà di normalità del Tau di Kendall utilizzata anche nella sezione precedente per dimostrare l'indipendenza tra frequency e severity. Si nota immediatamente che, nel caso di aggregazione annuale, l'intervallo di confidenza per i singoli valori di τ_K comprende in molti casi l'intero dominio $[-1;1]$, confermando ancora una volta la scarsa afffidabilità delle stime ottenute con tale aggregazione.

Per tali motivi, basandosi unicamente sull'analisi di correlazione, appare preferibile la scelta di aggregare mensilmente le perdite per determinare la distribuzione multivariata. In ogni caso, si porterà avanti la modellizzazione tramite copule con entrambe le aggregazioni. Si utilizzerà la stima con il Tau di Kendall sia per quanto riguarda l'aggregazione mensile (anche se, come si è visto, la differenza tra i due metodi è quasi nulla) che per l'aggregazione annuale, in quanto appare più conservativa (ovvero, assume dipendenze più elevate).

5.6 Determinazione della copula

Una volta determinata la struttura di correlazione tra le classi di rischio, è possibile passare alla modellizzazione delle stesse con la teoria delle copule (Capitolo 4). Si procederà quindi con il fit dei dati di perdita aggregati mensilmente e annualmente con diverse tipologie di copule, individuando poi la più adatta a rappresentare la struttura multivariata dei rischi operativi.

5.6.1 Copule ellittiche

Per quanto riguarda la famiglia delle copule ellittiche (sezione 4.3), si è considerata innanzitutto la copula t di Student. Come visto nella trattazione teorica, un primo parametro di tale distribuzione è la matrice di correlazione R, che, come visto, è possibile stimare tramite il Tau di Kendall.

Un ulteriore parametro da stimare è rappresentato dai gradi di libertà ν. Come detto, lo stimatore utilizzato in questo ambito è quello di massima verosimiglianza. Applicando tale stima usando l'aggregazione mensile e il Tau di Kendall per la determinazione della matrice di correlazione, non si osserva un massimo limitato per i gradi di libertà, come si può osservare dal grafico della log-verosimiglianza, riportato in Figura 5.18. Nella parte destra della stessa, si può invece notare come, nel caso di aggregazione annuale delle perdite dei vari ET, il massimo della log-verosimiglianza sia finito, in particolare la stima MLE dei gradi di libertà è $\widehat{\nu} = 4.02$.

Si noti che, come discusso nel Capitolo 4, la copula t si approssima alla copula gaussiana quando i gradi di libertà tendono all'infinito. Per tale motivo, all'interno della famiglia di copule ellittiche, per quanto riguarda l'aggregazione mensile, la scelta ricade sulla copula gaussiana, mentre per l'aggregazione annuale sulla copula t con 4 gradi di libertà.

Il passo successivo è quindi la valutazione della bontà di adattamento dei dati (mensili e annuali) alle copule considerate. In particolare, è stato portato avanti il test di Cramer-von Mises presentato nella sezione 4.5, i cui risultati sono riportati in Tabella 5.8. Come si può notare, entrambi i p-value sono sufficientemente alti da non scartare l'ipotesi nulla (ovvero l'appartenenza dei dati multivariati alla distribuzione teorica scelta). Sulla base di questi test, non è quindi possibile scegliere quale delle due aggregazioni utilizzare per il calcolo del capitale a rischio.

Come trattato nella sezione 4.2.4, è importante nell'ambito delle copule tenere conto delle dipendenze di coda. Dato che nella pratica non è possibile calcolare la tail dependence è utile confrontare il comportamento della

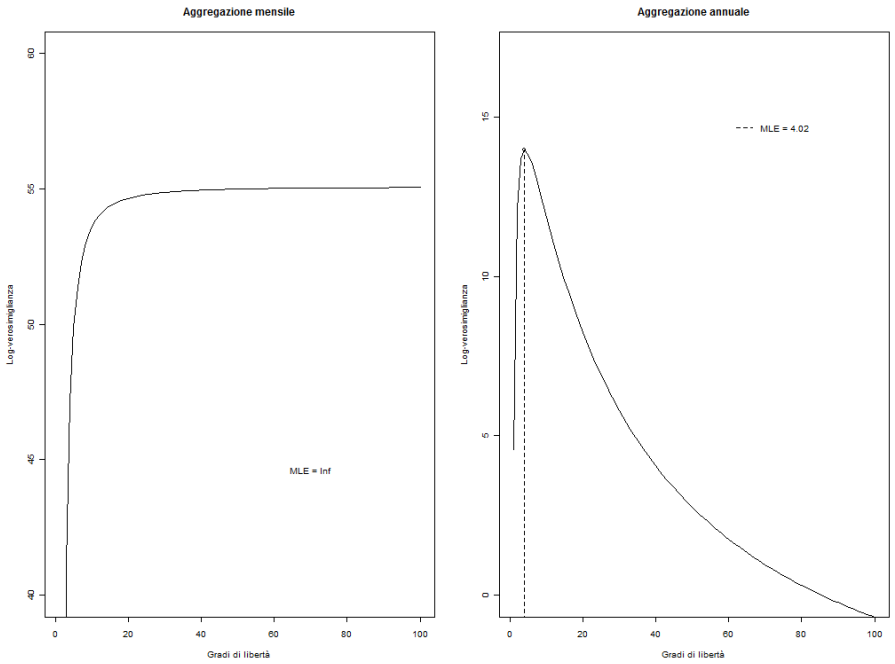

Figura 5.18: Log-verosimiglianza della copula t di Student, utilizzando aggregazione mensile e annuale, al variare dei gradi di libertà.

CvM	Statistica	P-value
Mensile	0.0254	0.697
Annuale	0.199	0.581

Tabella 5.8: Risultati dei test di Cramer-von Mises per il buon adattamento della copula gaussiana ai dati aggregati mensilmente e della copula t con 4 gradi di libertà ai dati aggregati annualmente.

distribuzione teorica rispetto a quella empirica, in particolare per la coda superiore, che sarà quella in cui si andrà a calcolare il capitale a rischio. Quindi, fissando come quantili "estremi" quelli di livello 90% e 95%, per ogni coppia di ET si calcola la percentuale di coppie di osservazioni in cui entrambe le variabili sono superiori al quantile prescelto, confrontando questa quantità con il valore teorico, che dipende dalla copula scelta.

Riguardo alle probabilità empiriche, per l'aggregazione mensile si sono riportati i seguenti risultati:

$$P_{.9}^{emp} = \begin{bmatrix} 0.1 & 0.05 & 0.017 & 0.05 & 0.033 & 0.033 & 0.033 \\ & 0.1 & 0.05 & 0.05 & 0.017 & 0.033 & 0.033 \\ & & 0.1 & 0.017 & 0.017 & 0.017 & 0.05 \\ & & & 0.1 & 0 & 0.017 & 0.017 \\ & & & & 0.1 & 0.017 & 0.017 \\ & & & & & 0.1 & 0.05 \\ & & & & & & 0.1 \end{bmatrix}$$

$$P_{.95}^{emp} = \begin{bmatrix} 0.05 & 0.017 & 0.017 & 0 & 0.017 & 0.017 & 0.017 \\ & 0.05 & 0.033 & 0 & 0 & 0.017 & 0.017 \\ & & 0.05 & 0 & 0 & 0.017 & 0.017 \\ & & & 0.05 & 0 & 0 & 0 \\ & & & & 0.05 & 0 & 0 \\ & & & & & 0.05 & 0.017 \\ & & & & & & 0.05 \end{bmatrix}$$

Le medesime quantità sono state calcolate per la copula gaussiana, con matrice di correlazione calcolata con aggregazione mensile:

$$P_{.9}^{Gauss} = \begin{bmatrix} 0.1 & 0.015 & 0.008 & 0.008 & 0.021 & 0.008 & 0.008 \\ & 0.1 & 0.045 & 0.021 & 0.014 & 0.035 & 0.03 \\ & & 0.1 & 0.018 & 0.014 & 0.038 & 0.031 \\ & & & 0.1 & 0.013 & 0.029 & 0.017 \\ & & & & 0.1 & 0.01 & 0.008 \\ & & & & & 0.1 & 0.032 \\ & & & & & & 0.1 \end{bmatrix}$$

$$P_{.95}^{Gauss} = \begin{bmatrix} 0.05 & 0.004 & 0.002 & 0.002 & 0.007 & 0.002 & 0.002 \\ & 0.05 & 0.019 & 0.007 & 0.004 & 0.014 & 0.011 \\ & & 0.05 & 0.006 & 0.004 & 0.015 & 0.012 \\ & & & 0.05 & 0.003 & 0.011 & 0.005 \\ & & & & 0.05 & 0.003 & 0.002 \\ & & & & & 0.05 & 0.012 \\ & & & & & & 0.05 \end{bmatrix}$$

Come si può notare, le strutture di dipendenza di coda teoriche sembrano riflettere quelle empiriche: non vi sono marcate differenze tra le stime delle probabilità bivariate di coda, nè si osserva una sottostima per ogni coppia, che potrebbe portare a una consecutiva sottostima del capitale a rischio. In particolare, per quanto riguarda le probabilità relative al quantile di livello 90%, il valore massimo della matrice differenza è 0.041, mentre la norma spettrale è 0.093. Per quanto riguarda invece il livello 95% la differenza massima è 0.015 e la norma spettrale è 0.045.

Passando quindi alla matrice di probabilità di coda bivariate per l'aggregazione mensile, tutti i valori risultano nulli, in quanto il numero di dati a disposizione (5 per classe di rischio) è talmente basso per cui che nessuna osservazione può superare il quantile di livello 80%. Per tale motivo non ha senso il confronto di tali matrici con quelle teoriche.

In definitiva, a causa della sensibilità della stima della correlazione con i dati mensili e l'impossibilità di osservare il comportamento alle code, tra le copule ellittiche appare preferibile l'utilizzo della copula gaussiana stimata con aggregazione mensile, con un p-value del test di buon adattamento sufficientemente alto e un comportamento aderente ai dati nella coda destra della distribuzione di ogni classe di rischio.

5.6.2 Copule Archimedee

Passando alla modellizzazione dei dati multivariati con le copule Archimedee, si svolgerà quindi l'analisi utilizzando unicamente l'aggregazione mensile e si confronterà il buon adattamento di tali copule rispetto alla copula gaussiana.

Come trattato nella sezione 4.4, una caratteristica di tale famiglia di copule è la rappresentazione della matrice di dipendenza tra gli ET con un unico parametro, senza differenziare le varie coppie. Tra le possibili copule, si valuterà il fit con la copula Gumbel, che viene preferita nel caso di distribuzioni di perdite in quanto gode di dipendenza solo nella coda superiore e con la copula Frank, che ha lo stesso comportamento alle code della copula gaussiana, ovvero gode di indipendenza.

Si noti che la stima dei singoli parametri delle copule Archimedee non possono essere ottenuti dalla relazione con il Tau di Kendall, data dalla (4.43), in quanto essa è valida solo per copule bivariate. In alternativa, si utilizza la stima con il metodo MLE. La stime risultanti per le relative copule sono le seguenti:

$$\begin{aligned} \widehat{\theta}^{Gu} &= 1.239 \\ \widehat{\theta}^{Fr} &= 1.518 \end{aligned}$$

Si noti che, in quanto si stanno trattando copule a più di due dimensioni, tali parametri non assumono un significato pratico e utile ai fini della valutazione del fit.

Per quanto riguarda i test di buon adattamento delle perdite aggregate mensilmente con le due copule della famiglia Archimedea, si riportano in Tabella 5.9 le statistiche e i relativi p-value del test di Cramer-von Mises. Anche in questo caso, in entrambi i casi sembrano adatte a rappresentare la struttura multivariata dei dati a disposizione, in particolare la copula Gumbel, con un p-value superiore a 0.66.

CvM	Statistica	P-value
Gumbel	0.0298	0.665
Frank	0.0388	0.324

Tabella 5.9: Risultati dei test di Cramer-von Mises per il buon adattamento delle copule Gumbel e Frank ai dati di perdita aggregati mensilmente.

Passando quindi alle dipendenze di coda, ovvero le probabilità, per ogni coppia di ET, che entrambe le osservazioni superino un relativo quantile, si può notare come le copule Archimedee, in quanto la struttura di dipendenza è descritta con un unico parametro, siano rappresentate da un unica probabilità bivariata per ogni coppia.

Per quanto riguarda la copula Gumbel si ha:

$$\begin{aligned} p_{.9}^{Gum} &= 0.0316 \\ p_{.95}^{Gum} &= 0.0142 \end{aligned}$$

Il confronto con le relative matrici empiriche si porta quindi avanti con le matrici composta da tali valori, eccezion fatta per la diagonale principale, in cui i valori sono 0.1 e 0.05, ovvero le probabilità univariate per la singola classe di rischio. Per le probabilità relative al quantile di livello 90%, il differenza massima tra stime empiriche e valori teorici è di 0.0316, mentre la norma spettrale della matrice differenza è 0.053. Passando al livello 95% la differenza massima è 0.014 e la norma spettrale è 0.046. Confrontando tali valori con quelli relativi della copula gaussiana, non si osservano grandi differenze, sebbene la copula Gumbel non distingua il comportamento delle varie coppie di ET.

Gli stessi valori sono stati calcolati per la copula Frank:

$$\begin{aligned} p_{.9}^{Fr} &= 0.0169 \\ p_{.95}^{Fr} &= 0.0045 \end{aligned}$$

A riguardo, si sono calcolati rispettivamente errore massimo di 0.017 e norma spettrale di 0.092 per il quantile 90%, mentre 0.004 e 0.058 per il 95%. Tali valori risultano più bassi rispetto alla copula Gumbel per il fatto che, come detto in precedenza, la Frank gode di indipendenza asintotica, come la copula gaussiana, che è la distribuzione multivariata apparsa più adatta tra le copule ellittiche.

In definitiva, anche con le copule Archimedee l'adattamento delle distribuzioni teoriche ai dati sembra essere buono; tuttavia, la scelta ricade sulla copula gaussiana, coerentemente con le scelte degli anni precedenti, ma soprattutto perché, a parità di bontà del fit, è preferibile rispecchiare la struttura di correlazione empirica con la copula teorica, come invece non viene fatto con le copule Archimedee.

5.7 Calcolo del capitale a rischio

Una volta individuata la copula più adatta a rappresentare la distribuzione multivariata delle perdite delle sette classi di rischio, è quindi possibile passare al calcolo del capitale a rischio. Per farlo, bisogna prima determinare la distribuzione aggregata tramite simulazione. In particolare, per $i = 1, \ldots, N$, l'algoritmo è il seguente:

1. Si campiona un'osservazione $\mathbf{u}_i = (u_{i1}, \ldots, u_{i7})$ dalla copula scelta, ovvero:
$$\mathbf{u}_i \sim C_R^{Ga};$$

2. Per ogni $j = 1, \ldots, 7$ si detrmina l_{ij}, il quantile di livello u_{ij} sulla distribuzione di perdita empirica dell'ET j;

3. Si determina la perdita totale, come:
$$L_i = \sum_{j=1}^{7} l_{ij}.$$

La distribuzione $L_1, \ldots, L_N$ ottenuta rappresenterà quindi le perdite annue aggregate totali per la banca. È stato sviluppato tale algoritmo per un numero di simulazioni N pari a 5 milioni.

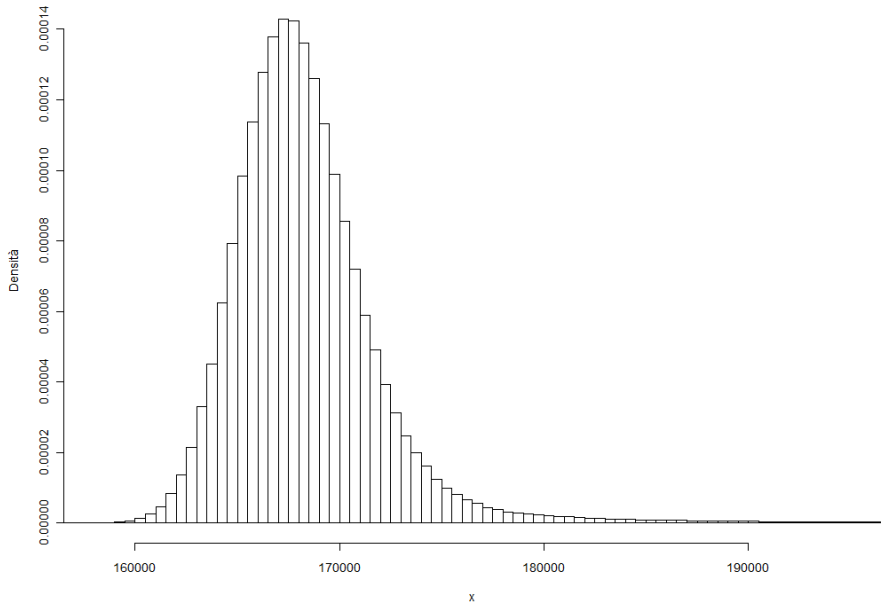

Figura 5.19: Istogramma normalizzato relativo alla simulazione delle perdite dei vari ET aggregate.

Si riporta in Figura 5.19 l'istogramma normalizzato risultante dalla simulazione.

Oltre alla stima del capitale a rischio, è utile calcolare anche l'Expected Loss totale, quantità che come detto in precedenza non potrà essere usata come capitale da accantonare, ma è un indicatore utile al management della banca. È quindi possibile determinare la perdita annua media dalla distribuzione empirica appena simulata:

$$EL = 168676$$

Passando quindi al VaR, ovvero il quantile 99.9% della distribuzione, si ottiene la seguente stima:

$$VaR = 253244$$

Esso, come detto in precedenza, rappresenta il capitale che la banca considerata deve allocare per far fronte ai rischi operativi, che è proprio lo scopo del modello presentato.

Bibliografia

[1] Banca d'Italia (2006) *"Nuove Disposizioni di Vigilanza Prudenziale per le Banche"*, Circolare n. 263 del 27 dicembre 2006.

[2] Bee, M. (2005) *"On Maximum Likelihood Estimation of Operational Loss Distributions"*. Discussion Paper No. 3, Dipartimento di Economia, Universitá degli Studi di Trento.

[3] Berg, D. and Bakken, H (2006) *"Copula Goodness-of-fit Tests: a Comparative Study"*. http://www.danielberg.no/publications/CopulaGOF.

[4] Box, G.E.P. and Jenkins, G. (1976) *"Time Series Analysis: Forecasting and Control"*. Holden-Day.

[5] Casella, G. and Berger, R.L. (2002) *"Statistical Inference"*. Second Edition. Duxbury Press, Belmont, CA.

[6] Chernobai, A., Rachev, S. and Fabozzi, F. (2005) *"Composite Goodness-of-Fit Tests for Left-Truncated Loss Samples"*. Technical report, University of California Santa Barbara.

[7] Cohen, A.C. (1965) *"Maximum Likelihood Estimation in the Weibull Distribution on Complete and on Censored Samples"*. JSTOR, Technometrics, Vol. 7, No. 4.

[8] Coles, S. (2001) *"An Introduction to Statistical Modelling of Extreme Values"*. Springer Series in Statistics.

[9] Embrechts, P., Klüppelberg, C. and Mikosch, T. (1997) *"Modelling Extremal Events"*. Springer Series: Stochastic Modelling and Applied Probability, Vol. 33.

[10] Embrechts, P., Lindskog, F. and McNeil, A. (2003) *"Modelling Dependence with Copulas and Application to Risk Management"*. In: Handbook for heavy Tailed Distributions in Finance, Elveiser, pp. 331-385.

[11] Frank, E. H., with contributions from many other users (2010) *"Hmisc: Harrell Miscellaneous"*. R package version 3.8-3.

[12] Genest, C. (1987) *"Frank's Family of Bivariate Distributions"*. Biometrika, Vol. 74, No. 3, pp. 549-555

[13] Genest, C., Remillard, B. and Beaudoin D. (2009). *"Goodness-of-fit Tests for Copulas: a Review and a Power Study"*. Insurance: Mathematics and Economics, 44, 199-214.

[14] Genz, A. (1992) *"Numerical Computation of Multivariate Normal Probabilities"*. Journal of Computational and Graphical Statistics 1 141-150.

[15] Greenwood, J. A., Landwehr, J. M., Matalas, N. C., and Wallis, J. R. (1979) *"Probability Weighted Moments: Definition and Relation to Parameters of Several Distributions Expressable in Inverse Form"* Water Resources Research, 15,1049-1054.

[16] Hoeffding, (1948) *"A Non-Parametric Test of Independence"*. The Annals of Mathematical Statistics, Vol. 19, No. 4, pp. 546-557.

[17] Hosking, J.R.M. and Wallis, J.R. (1987) *"Parameter and Quantile Estimation for the Generalized Pareto Distribution"*. JSTOR, Technometrics, 29(3): 339-349.

[18] Jawitz, J.W. (2004) *"Moments of Truncated Continuous Univariate Distributions"*. Water Resources Research, 27, 269-281.

[19] Kendall, M. G. (1938)*"A New Measure of Rank Correlation"*. Biometrika, Vol. 30, No. 1/2, pp. 81-93.

[20] Kojadinovic, I. and Yan, J. (2010) *"Modeling Multivariate Distributions with Continuous Margins Using the copula R Package"*. Journal of Statistical Software, 34(9), 1-20.

[21] Landwehr, J. M., Matalas, N. C. and Wallis, J. R. (1979) *"Estimation of Parameters and Quantiles of Wakeby Distributions"* Water Resources Research, 15, 1361-1379.

[22] McNeil, A. J. (1999) *"Extreme Value Theory for Risk Managers"*. In *Internal Modelling and CAD II*, pages 93-113. RISK Books.

[23] McNeil, A.J., Frey, R. and Embrechts, P. (2005)*"Quantitative Risk Management: Concepts, Techniques and Tools"*. Princeton Series in Finance.

[24] Nelder, J.A. and Mead, R. (1965) *"A Simplex Method for Function Minimization"*. Comput. J., 7, pp. 308-313.

[25] Nelsen, A., Frey R. and Embrechts, P. (1999) *"An Introduction to Copulas"*. Lecture notes in Statistics, Springer, Vol. 139.

[26] Schmidt, T. (2007) *"Coping with Copulas"*. In: Copulas - From Theory to Application in Finance, Risk Books, pp. 3-34.

[27] Sklar, A. (1959) *"Fonctions de ripartition in dimensions et leurs marges"*. Publications de l'Institut Statistique de l'Université de Paris, pp. 229-231.

[28] Smith, R.L. (1987) *"Estimating Tails of Probability Distributions"*. Ann. Statist. 15, 1174-1207.

[29] Wolter, T. (2008). *"truncgof: GoF Tests Allowing for Left Truncated Data"*. R package version 0.5-2.

[30] Wuertz, D., many others and see the SOURCE file (2009) *"fExtremes: Rmetrics - Extreme Financial Market Data"*. R package version 2100.77.

[31] Yan, J. (2007) *"Enjoy the Joy of Copulas: With a Package copula"*. Journal of Statistical Software, 21(4), 1-21.