UNIVERSITÀ DEGLI STUDI DEL PIEMONTE ORIENTALE
FACOLTÀ DI SCIENZE MATEMATICHE FISICHE E NATURALI

CORSO DI LAUREA IN INFORMATICA

CARATTERIZZAZIONE DEL CARICO PER SISTEMI GRID

Relatore: Prof. COSIMO ANGLANO
Correlatore: Dr. MASSIMO CANONICO

Candidato: MARCO GUAZZONE

ANNO ACCADEMICO 2006–2007

Copyright e Licenza

Licenza

- *Condividi allo stesso modo*. Se alteri o trasformi quest'opera, o se la usi per crearne un'altra, puoi distribuire l'opera risultante solo con una licenza identica o equivalente a questa.

Ogni volta che usi o distribuisci quest'opera, devi farlo secondo i termini di questa licenza, che va comunicata con chiarezza.

In ogni caso, puoi concordare col titolare dei diritti utilizzi di quest'opera non consentiti da questa licenza.

Questa licenza lascia impregiudicati i diritti morali.

Limitazione di responsabilità

Le utilizzazioni consentite dalla legge sul diritto d'autore e gli altri diritti non sono in alcun modo limitati da quanto sopra.

A Lorenza,
per la sua infinita pazienza
e per la felicità che mi dona ogni giorno.

Al Prof. C. Anglano e al Dr. M. Canonico,
per tutto l'aiuto che mi hanno dato,
l'entusiasmo che mi hanno trasmesso,
e l'umanità che mi hanno dimostrato.

Al Prof. A. Bobbio,
per i suoi utilissimi consigli sulla Statistica.

Ai miei genitori.

Indice

Notazioni e Nomenclatura

$\mathbf{A}$	Matrice.
a_{ij}	Elemento della matrice $\mathbf{A}$ posto nella riga i e colonna j.
$[a_{ij}]$	Forma compatta per rappresentare la matrice $\mathbf{A}$ di elementi a_{ij}.
$\mathbf{A}_{nm}$	Matrice di dimensione $n \times m$.
$\mathbf{A}_n$	Matrice quadrata di ordine n.
$\mathbf{I}$	Matrice identità (tutti gli elementi sono uguali a zero tranne quelli sulla diagonale principale che sono uguali a uno).
$\mathbf{0}$	Matrice nulla (tutti gli elementi sono uguali a zero).
$\mathbf{J}$	Matrice con tutti gli elementi uguali a uno.
$\mathbf{A}^T$	Matrice trasposta della matrice $\mathbf{A}$.
$\bar{\mathbf{A}}$	Matrice complessa coniugata della matrice $\mathbf{A}$.
$\mathbf{A}^H$	Matrice coniugata trasposta (o aggiunta, o Hermitiana aggiunta, o Hermitiana trasposta) della matrice $\mathbf{A}$.
$\mathbf{A} \otimes \mathbf{B}$	Prodotto di Kronecker tra le matrici $\mathbf{A}$ e $\mathbf{B}$.
$\mathbf{A} \oplus \mathbf{B}$	Somma di Kronecker tra le matrici $\mathbf{A}$ e $\mathbf{B}$.
$\vec{\mathbf{v}}$	Vettore riga.
v_i	Elemento i-esimo del vettore riga $\vec{\mathbf{v}}$.
$[v_i]$	Forma compatta per rappresentare il vettore $\vec{\mathbf{v}}$ di elementi v_i.
$\vec{\mathbf{1}}$	Vettore riga in cui tutti gli elementi valgono 1.
$\vec{\mathbf{0}}$	Vettore riga in cui tutti gli elementi valgono 0.
$\vec{\mathbf{v}}^T$	Vettore (colonna) trasposto del vettore riga $\vec{\mathbf{v}}$.
$\vec{\mathbf{v}}^H$	Vettore (colonna) coniugato trasposto del vettore $\vec{\mathbf{v}}$.

$\mathbb{N}$	Insieme dei numeri naturali.
$\mathbb{R}$	Insieme dei numeri reali.
$\mathbb{R}^*$	Insieme dei numeri reali positivi, incluso il numero 0.
$\mathbb{R}^+$	Insieme dei numeri reali positivi, escluso il numero 0.
$\mathbb{C}$	Insieme dei numeri complessi.
$\Im(x)$	Parte immaginaria del numero $x \in \mathbb{C}$.
$\Re(x)$	Parte reale del numero $x \in \mathbb{C}$.
$\gcd\{x_1, \ldots, x_n\}$	Massimo Comune Divisore tra $x_1, \ldots, x_n$.
$\nabla\varphi(\vec{\mathbf{x}})$	Gradiente della funzione $\varphi(\cdot)$.
x_+	$\max\{x, 0\}$ (massimo tra x e zero).
$f(x) \sim g(x),\ t \to \infty$	Significa che $\lim_{x\to\infty} f(x)/g(x) = 1$.

PDF	Funzione di densità di probabilità.
PMF	Funzione di massa di probabilità.
CDF	Funzione di distribuzione cumulativa.
CCDF	Complemento della funzione di distribuzione cumulativa (cioè $1 -$ CDF).
$\mathrm{E}[X]$	Valore atteso della variabile aleatoria X.
$\mathrm{Var}[X]$	Varianza della variabile aleatoria X.
$\mathrm{MAD}[X_1, \ldots, X_n]$	Mediana della Deviazione Assoluta dalla Mediana del campione $X_1, \ldots, X_n$.
$\mathrm{IQR}[X_1, \ldots, X_n]$	Intervallo Inter-Quartile del campione $X_1, \ldots, X_n$.
$\mathrm{CI}_\alpha(l, u)$	Intervallo di confidenza al $(1-\alpha)\%$ con estremo inferiore l ed estremo superiore u.
CV	Coefficiente di Variazione.
EDF	Funzione di distribuzione cumulativa empirica.
MGF	Funzione Generatrice dei Momenti.
$M(\theta)$	Funzione Generatrice dei Momenti.
PGF	Funzione Generatrice delle Probabilità (o Trasformata-z).
$G(z)$	Funzione Generatrice delle Probabilità (o Trasformata-z).
LST	Trasformata di Laplace-Stieltjes.
$\{\mathcal{L}^*\varphi\}(s)$	Trasformata di Laplace-Stieltjes della funzione $\varphi(\cdot)$.
LT	Trasformata di Laplace.
$\{\mathcal{L}\varphi\}(s)$	Trasformata di Laplace della funzione $\varphi(\cdot)$.

$\mathbb{I}(x)$	Funzione Indicatore (o Funzione Caratteristica) che vale 1 se l'evento x si verifica, e 0 altrimenti.
$\Phi(z)$	Funzione di distribuzione di una Normale standardizzata.
$\mathcal{L}(\theta; x_1, \ldots, x_n)$	Funzione di Verosimiglianza per il parametro ignoto θ rispetto al campione fissato $x_1, \ldots, x_n$.
$\mathrm{rank}(x)$	Rango di x all'interno di un certo insieme di osservazioni.
$X_n \xrightarrow{\mathcal{D}} X$	La variabile aleatoria X_n converge in distribuzione alla variabile aleatoria X.
$X_n \xrightarrow{\mathcal{P}} X$	La variabile aleatoria X_n converge in probabilità alla variabile aleatoria X.
CTMC	Catena di Markov a Tempo Continuo.
DTMC	Catena di Markov a Tempo Discreto.
EMC	Catena di Markov (a Tempo Discreto) Embedded.
MC	Catena di Markov.
MMPP	Processo di Poisson modulato da una Catena di Markov.
$\mathcal{PH}$	Insieme delle distribuzioni Phase-Type.
PH	Generica distribuzione Phase-Type.
APH	Distribuzione Phase-Type Aciclica.
DPH	Distribuzione Phase-Type Discreta.
CPH	Distribuzione Phase-Type Continua.
QBD	Processo Quasi Birth-Death (o Non Birth-Death).
BoT	Bag-of-Task.
VO	Organizzazione Virtuale.

Capitolo 1

Introduzione

L'uso del calcolatore ha notevolmente semplificato lo svolgimento di operazioni computazionalmente complesse sia dal punto di vista matematico sia da quello temporale: al giorno d'oggi, si è in grado di effettuare esperimenti matematicamente molto complicati, "rendering" di immagini ad alta definizione, simulazioni di sistemi complessi, previsioni a lungo termine, ... Ciò che accomuna questo tipo di operazioni è la richiesta di un'elevata potenza computazionale da utilizzare in un arco di tempo piuttosto limitato.

La velocità dello sviluppo tecnologico e l'abbassamento dei relativi costi ha diffuso l'utilizzo dei computer pressoché in ogni angolo del Pianeta. Tuttavia, mentre sempre più persone utilizzano e possiedono un computer, l'uso che ne fanno è di solito ristretto ad applicazioni caratterizzate da una bassa richiesta computazionale; da ciò ne segue che la maggior parte dei computer accesi, sparsi in tutto il mondo, risulta inutilizzata.

L'idea del *Grid Computing* è quella di utilizzare questa, teoricamente enorme, potenza computazionale, attraverso la condivisione dinamica, su larga scala, di risorse eterogenee e indipendenti. Per poter ottenere questi risultati, occorre che i sistemi Grid siano dotati di una logica in grado di *gestire le risorse* in modo efficiente, trasparente, distribuito e coordinato.

Uno dei componenti fondamentali di un sistema Grid è lo *scheduler*, il quale si occupa di assegnare un job su una macchina remota (*nodo di computazione*) che sia in grado di eseguirlo. Nella letteratura scientifica sono state presen-

tate diverse strategie di scheduling per sistemi Grid (ad es., si veda [71, 21]); purtroppo sembra non esserci una particolare convergenza su quali di queste risultino più adatte nel contesto del Grid Computing. Lo sviluppo e la scelta di una strategia di scheduling dipende, in generale, dal tipo di informazioni che si hanno a disposizione e dalle caratteristiche del carico generato nei sistemi Grid; mentre è possibile individuare un insieme limitato di informazioni che possono essere utilizzate (ad es., potenza computazionale, tempo medio di vita di una macchina, durata e dimensione di un job, ...), vi è una scarsa conoscenza sulle caratteristiche che queste informazioni possono avere nel contesto del Grid Computing e sulla loro influenza nei confronti della caratterizzazione del carico; ciò, può essere dovuto a un impiego, piuttosto recente, dei sistemi Grid in ambienti di produzione e, quindi, alla ancora scarsa disponibilità di tracce. Lo scopo di questo lavoro è quello di ottenere maggiori informazioni sulle caratteristiche del carico nei sistemi Grid, attraverso l'analisi statistica di alcune tracce.

1.1 Grid Computing

Il *Grid Computing* nasce come conseguenza della necessità di condivisione di risorse (computazionali e di memorizzazione) al fine di rendere possibile e migliorare collaborazioni scientifiche su larga scala. Il termine *Grid*, coniato nella metà degli anni '90, deriva da una metafora con il sistema di trasmissione dell'energia elettrica: l'accesso alle risorse (computazionali e di memorizzazione) deve essere ugualmente facile e disponibile come quando si inserisce una presa nella spina della corrente elettrica.

Per *Grid Computing*, si intende un'infrastruttura hardware/software che permetta di condividere, in modo dinamico, trasparente, sicuro e su larga scala, delle risorse eterogenee tra entità differenti [46]. Per condivisione *dinamica* si intende un tipo di condivisione che può mutare nel tempo. La *trasparenza* permette di mantenere nascosta agli utenti la natura distribuita ed eterogenea delle risorse. La *sicurezza* è necessaria al fine di garantire che le risorse condivise non vengano compromesse e siano utilizzate secondo specifiche politiche

di condivisione. La condivisione *su larga scala* permette di utilizzare risorse remote sparse in tutto il mondo.

Tra i fattori che hanno contribuito allo sviluppo del *Grid Computing* vi si trova:

- *Rapido Sviluppo Tecnologico*. L'incredibile velocità con cui la tecnologia sta evolvendo ha fatto sì di avere, al giorno d'oggi, dei computer molto più potenti dei "mainframe" di una decina di anni fa.

- *Abbattimento dei Costi Tecnologici*. I costi della tecnologia tendono a descescere, secondo la *legge di Moore*, con la stessa velocità con cui evolve la tecnologia stessa; ciò ha reso possibile la diffusione dei calcolatori tra una vasta massa della popolazione mondiale.

- *Aumento della Richiesta Computazionale*. Parallelamente alla crescita della disponibilità computazionale ottenuta dallo sviluppo tecnologico, sembra crescere altrettanto velocemente la richiesta di potenza computazionale da parte delle applicazioni; mentre ciò può essere dovuto, in parte, a una necessità di mercato che impone, per motivi di concorrenza, lo sviluppo di applicazioni, anche poco ottimizzate, nel più breve tempo possibile, dall'altra parte la disponibilità di una più grande potenza computazionale rende possibile lo sviluppo di applicazioni che fino a pochi anni fa erano impensabili da sviluppare, come la simulazione della nascita dell'universo, la previsione di eventi geosismici, ...

- *Aumento della Disponibilità di Potenza non utilizzata*. Se, da un lato, sono aumentate le applicazioni che richiedono una grande potenza computazionale, il loro utilizzo è ristretto a poche comunità di individui (ad es., scienziati) e, anche in questi casi, la richiesta di potenza computazionale è generalmente *on-demand*: discontinua e sbilanciata; per esempio, un ricercatore può aver bisogno di una potenza computazionale estremamente elevata solo per condurre una particolare simulazione e per un tempo limitato; per il resto del tempo il computer potrebbe rimanere inutilizzato. Anche per quanto concerne i computer "personali", cioè destinati ad un uso casalingo o per fini di "office automation", risulta

che, per la maggior parte del tempo, il calcolatore è pressoché inutilizzato (*idle*). Inoltre, il costo relativamente basso dell'energia elettrica, ha fatto diffondere la tendenza a lasciare acceso un computer anche quando deve rimanere non utilizzato per diverse ore o giorni [1].

- *Condivisione dei Dati su Larga Scala*. La crescente diffusione di Internet ha reso più semplice la comunicazione e collaborazione di entità poste in zone geograficamente opposte del Pianeta; ciò ha permesso anche il diffondersi della condivisione dei dati, in particolare di informazioni e di risultati di esperimenti all'interno della comunità scientifica. La condivisione dei dati ha anche il vantaggio di evitare, nella maggior parte dei casi, una replicazione di dati: se i dati possono essere disponibili in qualsiasi momento è inutile farsene una copia (salvo per motivi di "back-up"); quando la mole dei dati è molto grande questo fatto non è trascurabile, in quanto permette di ottenere un notevole risparmio sui costi.

- *Reti di Comunicazione ad alta velocità*. La velocità delle reti di comunicazione, comprese quelle su scala geografica, che si è in grado di raggiungere al giorno d'oggi, permette di scambiare dei dati in un tempo molto breve.

Dai punti appena elencati si evince che il rapido sviluppo tecnologico e l'abbattimento dei costi hanno reso possibile la diffusione di "personal computer" molto potenti su scala mondiale; questi computer sono, nella maggior parte dei casi, sotto utilizzati per la quasi totalità del tempo in cui rimangono accesi. La diffusione di Internet e il basso costo dell'energia rende potenzialmente disponibile in qualsiasi momento della giornata un'enorme capacità computazionale. La crescita tecnologica ha reso possibile lo sviluppo e la diffusione di applicazioni, molto intense dal punto di vista computazionale, per la risoluzioni di problemi che, fino a qualche anno fa, era impossibile pensare di risolvere in breve tempo; ne segue quindi una domanda sempre più crescente di potenza computazionale; tale domanda segue in genere un andamento a picchi:

[1] Nei prossimi anni questa tendenza è probabilmente destinata a sparire a causa delle conseguenza sul clima del Pianeta.

è necessaria una potenza computazionale molto elevata ma per un periodo di tempo relativamente limitato. La "globalizzazione" a livello scientifico ha fatto aumentare la domanda di condivisione di informazioni e di risultati di esperimenti. Tutto ciò ha reso possibile la diffusione del *Grid Computing*.

Il *Grid Computing* tenta di raggiungere il suo obiettivo cercando di sfruttare l'enorme capacità inutilizzata dei computer sparsi in tutto il mondo. La condivisione delle risorse (potenza di calcolo, supporti di memorizzazione, dati, ...) viene effettuata attraverso la definizione di *Organizzazioni Virtuali* (*VO*). Una VO è un insieme di individui o organizzazioni accomunati da una serie di regole di condivisione di risorse [47]; un tipico caso in cui vi è l'esigenza di definire delle VO, è la conduzione di diversi esperimenti fisici, a partire da un enorme insieme di dati, in vari laboratori sparsi in tutto il mondo; per esempio, il progetto LCG del CERN è un sistema Grid nato con lo scopo di mettere a disposizione a diversi laboratori fisici (molto sparsi geograficamente) i dati prodotti dal LHC (Large Hadron Collider) e, al tempo stesso, condividere la potenza computazionale e di memorizzazione per l'elaborazione dei dati e il salvataggio dei risultati. Nei sistemi Grid, l'unità minima di esecuzione è un *job* [2]. L'esecuzione di un *job*, in un sistema Grid, viene effettuata su una macchina remota (*nodo di computazione*); la scelta di quale job eseguire e di quale macchina assegnargli viene presa dallo *scheduler* di un sistema Grid.

1.2 Scheduling nei Sistemi Grid

L'esecuzione di un job in un sistema Grid è influenzata da diversi fattori, tra cui:

- la frazione di CPU che il proprietario del nodo di computazione decide di dedicare alle computazioni Grid: la percentuale di CPU messa a disposizione su un nodo di computazione varia a seconda delle necessità computazionali del proprietario;

[2]Quando si parla di *Bag-of-Task*, invece, con *job* si intende un insieme di *task* paralleli e indipendenti; in tal caso, l'unità minima di esecuzione è un *task*.

- la disponibilità del nodo di computazione: una macchina remota non è facilmente controllabile come una locale; per esempio potrebbe essere spenta in qualsiasi momento, oppure il proprietario del nodo di computazione potrebbe decidere di "disabilitare" temporaneamente il componente "client" che si interfaccia al sistema Grid;

- la strategia di scheduling utilizzata per scegliere un job da eseguire e un nodo di computazione a cui assegnare il job.

Per quanto riguarda la frazione di CPU dedicata alle computazioni Grid, di solito si cerca di incentivare un proprietario di un nodo di computazione tramite un sistema di crediti: più è alta la percentuale di potenza e di tempo della CPU messa a disposizione, maggiore è il guadagno di crediti da parte di un utente; la gestione dei crediti varia di solito da sistema a sistema; per esempio, tali crediti possono consentire di ottenere dei privilegi nel caso in cui si abbia la necessità di eseguire un job attraverso il sistema Grid.

Per quanto concerne la disponibilità dei nodi di computazione, la maggior parte dei sistemi Grid cercano di risolvere il problema dotando il *middleware* [3] di un sistema di *checkpointing*, tramite il quale lo stato dell'esecuzione di un job viene periodicamente salvato, in modo da riprendere l'esecuzione del job in un secondo momento, oppure, quando possibile, di spostarla su un nodo di computazione differente, qualora il nodo di computazione corrente diventi temporaneamente indisponibile.

Riguardo alle strategie di scheduling, esse dipendono da:

- il *tipo* di informazioni che hanno a disposizione: possono variare a seconda del middleware utilizzato, delle politiche adottate dalle VO coinvolte e dalla loro disponibilità nel fornire informazioni;

- *se* un'informazione è necessaria: dipende da quali caratteristiche influenzano il carico;

- *come* un'informazione viene utilizzata: dipende sostanzialmente dalla natura delle informazioni che vengono utilizzate;

[3] Strato software che nasconde la natura distribuita di un sistema Grid agli utenti.

- *quando* utilizzare un'informazione: dipende dal tipo di euristica che si vuole usare; in un'euristica *batch*, l'assegnazione dei job alle macchine avviene per gruppi di job, in corrispondenza dello scattare di particolari eventi (*mapping event*); in un'euristica *online*, l'assegnazione dei job alle macchine avviene considerando un job alla volta.

Il tipo di informazioni tipicamente utilizzate comprende:

- i tempi di interarrivo dei job;
- la durata dell'esecuzione di un job;
- la dimensione di un job;
- la tipologia del job (job singolo o Bag-of-Task);
- la dimensione della fase di *stage-in* e *stage-out*.

Esiste una vasta letteratura su come e quando utilizzare le informazioni in una strategia di scheduling (ad es., si veda [71, 21]); per esempio, l'euristica *Min-Min* [71] assegna il job più veloce alla macchina che, al momento, è in grado di completarlo nel più breve tempo possibile; l'euristica $WQRxN$ assegna un job (scelto a caso) alle prime N macchine disponibili; l'euristica *Sufferage* [71] assegna una certa macchina al task che "soffrirebbe" maggiormente (in termini di tempo di completamento) nel caso non gli venisse assegnata; l'euristica *XSufferage* [21], estende l'idea dell'euristica *Sufferage*, confrontando, per il calcolo della "sofferenza" di un job, i tempi di completamento a livello cluster .

Fra tutte le euristiche proposte, non ne esiste una migliore delle altre; per esempio c'è chi sostiene che l'euristica migliore sia *WQRx1* o *WQRx2*, grazie alla sua semplicità, all'efficienza (in senso computazionale) e all'assenza di particolari assunzioni sulle caratteristiche dei job e delle macchine. In molti degli articoli che propongono nuove euristiche, di solito vi è una parte relativa agli esperimenti effettuati e alla presentazione dei risultati; questi esperimenti, sono generalmente creati ad-hoc e condotti effettuando una serie di confronti con le più conosciute euristiche per sistemi Grid. Il problema principale di questo

approccio è che questi esperimenti sono difficilmente ripetibili, se non da chi li ha originariamente effettuati, a causa della mancanza di scenari standard; questo può risultare in uno svantaggio sia per chi effettua la revisione di un lavoro proposto in un articolo, in quanto non ha modo di verificare i risultati ottenuti, sia per chi propone la nuova euristica, in quanto i risultati sono generalmente confutabili variando semplicemente qualche parametro che caratterizza lo scenario dell'esperimento. L'ideale quindi sarebbe la creazione di una serie di scenari standard in cui verificare le prestazioni di una certa strategia di scheduling. Per creare questi scenari è necessario avere una conoscenza accurata delle caratteristiche del carico presente nei sistemi Grid. Mentre esiste una numerosa letteratura per sistemi paralleli (ad es., si veda [19, 40]), purtroppo ancora poco è stato scritto per sistemi Grid.

1.3 Caratterizzazione del Carico

L'analisi delle proprietà del carico si pone principalmente i seguenti obiettivi:

1. *Standardizzazione degli scenari di esecuzione dei propri esperimenti*. Quando si sviluppa una nuova soluzione, ad es. una nuova strategia di scheduling, per verificarne e dimostrarne l'efficacia occorre condurre una serie di esperimenti. In questo caso, l'esistenza di scenari standard in cui effettuare i propri esperimenti ha il vantaggio di permettere il confronto del proprio operato con lavori già esistenti e approvati, e, quindi, di valutare con una migliore criticità le relative prestazioni.

2. *Costruzione di modelli realistici*. La creazione di scenari standard per l'esecuzione di esperimenti non è sufficiente: il supporto di una base teorica, benché permetta di stabilire una serie di regole e metriche comuni per la conduzione degli esperimenti e la valutazione dei risultati, potrebbe non essere utilizzabile nella pratica. Uno scopo ancor più ambizioso è quello di creare dei modelli che siano il più possibile fedeli al comportamento reale dei sistemi Grid; un modello teorico che non ha riscontro

con la realtà (o che le si avvicina solo marginalmente) è di scarsa utilità e sicuramente è destinato ad avere una vita breve.

3. *Valutazione delle prestazioni*. Un sistema Grid rappresenta un sistema molto complesso: la presenza di macchine eterogenee e autonome potrebbe rendere inadatte le misure delle prestazioni tradizionali. Per esempio, la classica assunzione dei tempi di interarrivo modellati seconda una distribuzione Esponenziale potrebbe rappresentare una semplificazione non ammissibile. Il consolidamento di una teoria per la misura delle prestazioni permette di studiare il comportamento dei sistemi Grid da un punto di vista quantitativo e di costruire modelli generativi che ne permettano una simulazione attinente alla realtà..

4. *Sviluppo di soluzioni più mirate*. La conoscenza di quali caratteristiche influenzano maggiormente il carico e delle relative proprietà consente lo sviluppo di soluzioni più ottimizzate. Per esempio, se si scoprisse che nei sistemi Grid la dimensione di un job è strettamente correlata positivamente alla relativa durata dell'esecuzione, si potrebbero progettare delle strategie di scheduling in grado di rispecchiare tale relazione; oppure, se si notasse che la dimensione di un job non influisce sulla caratterizzazione del traffico, si potrebbero progettare strategie di scheduling che evitino di tenere in considerazione questo attributo, in modo da ottenere un guadagno sul tempo di esecuzione dell'euristica.

5. *Confronto di differenti soluzioni a uno stesso problema*. La possibilità di verificare i risultati del proprio operato con altri lavori, basati sullo stesso modello e approvati dalla comunità scientifica, permette, a chi ha effettuato il lavoro, di valutarne in ogni istante la relativa bontà e, a chi ne effettua la revisione, di renderne più semplice la valutazione ed esprimere una critica più precisa.

Esiste una vasta letteratura riguardo la caratterizzazione del carico per sistemi paralleli (ad es., si veda [19, 40]); tuttavia il carico generato in questi sistemi dipende, in generale, da caratteristiche differenti da quelle che posso-

no influenzare il carico nei sistemi Grid. Fra le differenze più significative si ricorda:

- I sistemi paralleli sono di solito costituiti da macchine omogenee, mentre i sistemi Grid sono caratterizzati dall'alto tasso di eterogeneità delle macchine.
- La disponibilità di una macchina in un sistema parallelo è una caratteristica facilmente ipotizzabile; nei sistemi Grid, le macchine su cui un job viene eseguito sono, di solito, fuori dal controllo di chi ha sottomesso il job.
- La scala di comunicazione relativa ai sistemi paralleli non supera di solito una rete locale; al contrario, i sistemi Grid sono sistemi a larga scala, in cui le macchine sono lascamente connesse tramite una rete geografica.
- La tipologia di job in esecuzione sui sistemi paralleli è di solito costituita da job paralleli composti da task fra loro dipendenti. Nei sistemi Grid, invece, i job tendono a essere dei Bag-of-Task, cioè gruppi di task, riguardanti una stessa applicazione, fra loro indipendenti.

Per realizzare un modello del carico che riproduca nel modo più fedele possibile il comportamento di un sistema Grid, è necessario raccogliere una serie di tracce reali ed effettuarne un'accurata analisi statistica. Dato che il problema della caratterizzazione del carico nel contesto del Grid Computing costituisce un campo di ricerca recente, la disponibilità di tracce per questo tipo di sistemi risulta ancora limitata.

In questo lavoro vengono analizzate due tracce:

- *LCG*, prelevata dal sito *Parallel Workload Archive* [68] e relativa all'omonimo sistema Grid sviluppato per il progetto LHC (Large Hadron Collider)
- *TeraGrid*, ottenuta dal sistema TeraGrid [4].

L'analisi di ogni traccia si pone come obiettivo lo studio di due caratteristiche del carico. il *tempo di interarrivo* di un job e la relativa *durata dell'esecuzione*, e la ricerca di un modello statistico che possa descriverne il comportamento.

[4] http://www.teragrid.org

1.4 Organizzazione della Tesi

Il presente documento è diviso in due parti. Nella Parte I vengono presentati i principali concetti e le tecniche usate per l'analisi di una traccia e la creazione di un modello statistico; in particolare:

- il Cap. 2 descrive la metodologia seguita per effettuare l'analisi delle tracce;
- il Cap. 3 presenta, in modo dettagliato, le principali nozioni riguardanti l'adattamento di una distribuzione a un insieme di dati, e le tecniche per effettuarne la stima dei parametri;
- il Cap. 4 illustra le principali distribuzioni di probabilità prese in esame per l'adattamento ai dati;
- il Cap. 5 descrive, in dettaglio, una particolare famiglia di distribuzioni molto flessibile, chiamata *Phase-Type*;
- il Cap. 6 presenta una particolare tipologia di distribuzioni, detta *heavy-tailed*, le cui caratteristiche principali sono la maggior influenza degli eventi estremi (cioè, quelli delle code della distribuzione) rispetto alle tradizionali distribuzioni di probabilità, come la Normale, e un conseguente allontanamento dalla normalità, anche dal punto di vista asintotico.

La Parte II è dedicata alla presentazione dell'analisi delle tracce e dei relativi risultati:

- il Cap. 7 illustra l'analisi statistica della traccia LCG;
- il Cap. 8 presenta l'analisi statistica della traccia TeraGrid.
- il Cap. 9 ha lo scopo di descrivere i risultati ottenuti da questo lavoro, i risultati ottenuti in lavori simili e le possibili direzioni da prendere per eventuali sviluppi futuri.

L'ultima parte del documento, Parte III, contiene alcune appendici:

- il Cap. A presenta la struttura del codice sorgente sviluppato per la conduzione degli esperimenti;
- il Cap. B contiene alcuni richiami ai principali concetti della Probabilità e della Statistica;
- il Cap. C descrive una trasformazione statistica utilizzata, in particolare, nel contesto dei test sulla bontà dell'adattamento di una distribuzione ai dati;
- il Cap. D presenta alcune trasformazioni utilizzate, specialmente, nell'ambito dell'analisi dei processi Markoviani.
- il Cap. E contiene un riepilogo dei principali concetti associati ai processi Markoviani, con un riferimento particolare alle Catene di Markov;
- il Cap. F presenta la definizione dell'esponenziale di una matrice e alcune tecniche per calcolarlo;
- il Cap. G descrive due operazioni matriciali: la somma e il prodotto di Kronecker.

Parte I

Metodologia

Capitolo 2

Analisi dei Dati

In questo capitolo viene presentata la metodologia generale utilizzata per affrontare la costruzione di un modello statistico; saranno illustrati i vari passi seguiti per l'analisi statistica dei dati, partendo dalla descrizione degli obiettivi che si intende raggiungere §2.2, fino ad arrivare alla costruzione §2.7 e validazione del modello §2.8. L'ultima sezione §2.9 descrive in maniera piè specifica le scelte personali effettuate per l'applicazione dei metodi statistici all'analisi delle tracce di sistemi Grid.

2.1 Introduzione

Il processo di costruzione di un modello statistico varia a seconda del criterio seguito; esistono, sostanzialmente, tre approcci:

- approccio classico, detto *Confirmatory Data Analysis* (*CDA*);
- approccio orientato ai dati, chiamato *Explanatory Data Analysis* (*EDA*);
- approccio basato sulla conoscenza/esperienza, detto *Approccio Bayesiano*.

I suddetti approcci non sono di per sé delle tecniche statistiche, bensì delle linee guida utilizzabili per affrontare l'analisi dei dati e giungere alla costruzione di un modello statistico: ad ogni passo della costruzione del modello, esse suggeriscono le tecniche statistiche più appropriate da utilizzare.

L'approccio CDA è l'approccio classico, guidato dall'ipotesi. Si basa sull'aspetto deduttivo e quindi utilizza tecniche della statistica inferenziale; l'analisi statistica viene effettuata attraverso la costruzione e verifica di ipotesi; spesso, vengono assunte come valide delle ipotesi sui dati effettuate prima di iniziare l'analisi (ad es., ipotesi di normalità). Il vantaggio principale di questo approccio è che, nel caso in cui le ipotesi effettuate siano effettivamente corrette, i risultati ottenuti sarebbero molto precisi. Al contrario, nel caso in cui una o più ipotesi non siano valide, si rischierebbe che l'intero processo di analisi diventi privo di utilità.

L'approccio EDA [110] è un approccio più moderno, basato sui dati. Si appoggia sull'aspetto induttivo della statistica, cercando di trarre il maggior numero di informazioni dai dati stessi; esso, quindi, sfrutta le tecniche che rientrano nella statistica descrittiva. I vantaggi principali sono il numero ridotto di assunzioni e l'utilizzo delle sole informazioni contenute nei dati; questi vantaggi, però, si devono pagare al prezzo di informazioni meno precise e raramente discriminanti, e di una maggiore difficoltà nell'interpretare i risultati ottenuti o nel combinare quelli ricavati da diversi metodi statistici (ad es., da diversi test sulla bontà di adattamento §3.2.6).

Infine, l'approccio Bayesiano tenta di incorporare, all'interno dell'analisi, la conoscenza e l'esperienza che si ha sui vari aspetti (caratteristiche) che costituiscono i dati; ciò viene effettuato, assegnando una particolare *distribuzione a priori* (*prior distribution*), indipendente dagli stessi dati, definita sui parametri del modello prescelto; la "distribuzione a priori" viene scelta in modo da rispecchiare la conoscenza che si ha a priori sul dominio di applicazione. L'analisi, quindi, consiste nell'effettuare delle inferenze e delle ipotesi statistiche sui dati sotto osservazione, condizionate alla validità della distribuzione a priori.

I tre approcci differiscono nel modo in cui viene affrontata la costruzione e validazione di un modello statistico, sia per il tipo di informazione che utilizzano sia per come la utilizzano:

- approccio CDA:

$$\text{Problema} \Rightarrow \text{Dati} \Rightarrow \text{Modello} \Rightarrow \text{Analisi} \Rightarrow \text{Conclusioni}$$

La raccolta dei dati è seguita da una scelta di un modello, effettuata secondo particolari ipotesi fatte a priori (ad es., normalità, linearità,...), e dall'analisi, stima e verifica dei relativi parametri;

- approccio EDA:

  Problema ⇒ Dati ⇒ Analisi ⇒ Modello ⇒ Conclusioni

  La raccolta dei dati è seguita da un'analisi sugli stessi, con lo scopo di trarre informazioni direttamente da essi (evitando di effettuare ipotesi a priori); in base alle informazioni ricavate, viene quindi scelto il modello più appropriato;

- approccio Bayesiano:

  Problema ⇒ Dati ⇒ Modello ⇒ Distribuzione a Priori ⇒ Analisi ⇒ Conclusioni

  La scelta di un modello viene effettuata dopo la raccolta dei dati, come in CDA, imponendo, però, su di esso una distribuzione a priori, indipendente dai dati e rispecchiante la conoscenza sul dominio di applicazione; il modello verrà quindi analizzato e verificato in maniera condizionata alla scelta della distribuzione a priori.

Lo scopo di tutti gli approcci è la creazione di un modello che imiti il più possibile le proprietà e il comportamento del fenomeno sotto osservazione. La creazione di un modello di solito comprende i seguenti passi:

1. scelta dell'obiettivo;
2. scelta delle caratteristiche da esaminare;
3. raccolta dei dati;
4. bonifica dei dati;
5. analisi delle proprietà statistiche dei dati;
6. scelta di un modello;

7. verifica della bontà di adattamento;

8. ripetizione dei passi da 6 a 7 finchè non si ottiene un modello che soddisfi le proprie esigenze.

Si noti che questi passi rappresentano solo una traccia. Nella pratica, essi non sono sempre effettuati nell'ordine mostrato, alcuni sono del tutto ignorati e altri vengono combinati; in effetti, spesso, occorre ritornare a un passo precedente una volta che si è ottenuta una maggiore conoscenza sulle proprietà statistiche dei dati; per esempio, dopo un'analisi sulla tendenza centrale e dispersione dei dati, è possibile tornare al passo di bonifica dei dati per eliminare qualche "outlier", o, magari, per reincludere delle osservazioni che in precedenza si aveva deciso di eliminare.

2.2 Scelta dell'obiettivo

L'obiettivo di questo lavoro è lo studio e la caratterizzazione del carico nei sistemi Grid. Vi sono diversi modi per cercare di raggiungere questo obiettivo; in [41] vengono proposte diverse metodologie; per esempio, ci si potrebbe concentrare sullo studio delle singole caratteristiche, oppure sulle loro relazioni; nel primo caso, l'analisi statistica da effettuare è del tipo "univariato" in quanto viene preso in esame solo una caratteristica per volta; nel secondo caso, invece, l'analisi statistica da effettuare è del tipo "multivariato". L'analisi univariata ha il vantaggio di essere semplice e di richiedere poche assunzioni sui dati [1]; lo svantaggio è che potrebbero venire escluse dalla caratterizzazione alcune proprietà salienti del carico. Come mostrato in [68], l'assunzione di indipendenza tra le caratteristiche, o di una forma di relazione non corretta, può portare a un degrado delle prestazioni del sistema; in particolare, nel caso dell'articolo citato, viene evidenziato come l'assunzione di indipendenza tra la dimensione dei job (cioè il numero di processori richiesti) e il relativo tempo di esecuzione, o di correlazione negativa (cioè, il tempo di esecuzione dei

[1] Con un approccio EDA, di solito si effettua solo un'assunzione di indipendenza tra le osservazioni di una certa caratteristica e, in alcuni casi, tale assunzione non è nemmeno necessaria.

job "grandi" è minore di quello dei job "piccoli") può portare a un degrado delle prestazioni dello "scheduler" nel caso in cui le politiche di "scheduling" utilizzate si basano su queste assunzioni.

Prima di iniziare la procedura di creazione di un modello, occorre decidere il tipo di modello che si intende creare. Un *modello descrittivo* è un modello che tenta di descrivere le caratteristiche del carico; una volta individuate le proprietà statistiche delle caratteristiche del carico e le possibili distribuzioni teoriche che le descrivono, è possibile generare carichi sintetici campionando, semplicemente, da tali distribuzioni. Spesso questo tipo di modelli viene detto "basato sulle distribuzioni". Un *modello generativo*, al contrario, è un modello che tenta di imitare il processo di generazione del carico; in generale, la distribuzione delle caratteristiche del carico non viene modellata esplicitamente, ma risulta una conseguenza del comportamento del modello (ad es., l'impiego di una Catena di Markov (Cap. E) per l'implementazione di un modello generativo implica che nel modello stesso la distribuzione dei tempi di interarrivo sia Esponenziale). Spesso questo tipo di modelli viene detto "basato sui processi".

Tra i vantaggi derivanti dall'utilizzo di un modello generativo, vi è, per esempio, la possibilità di effettuare, abbastanza facilmente, uno studio quantitativo di diverse grandezze (ad es., il tempo medio di soggiorno dei job nella coda di uno "scheduler"), non solo a livello delle singole grandezze ma anche considerando l'intero processo, e la presenza di pochi parametri da impostare. Il problema principale è che, in molti casi, alcune peculiarità del carico sono difficili da catturare oppure portano a un modello molto complesso. Per esempio, la presenza di componenti deterministiche cicliche nel carico "reale" può rendere inadeguato il modello che tenta di descriverlo [2]. Inoltre, può risultare più complesso il processo di generalizzazione, cioè di catturare con uno stesso modello il comportamento tipico dei carichi di diversi sistemi Grid, in quanto piccole variazioni nei parametri del modello possono portare a comportamenti totalmente differenti (ciò è colpisce soprattutto i modelli basati su processi frattali).

D'altro canto un modello descrittivo è in genere più semplice da costruire,

[2] Ad es., un processo Markoviano (Cap. E) sarebbe sicuramente inadeguato.

permette di descrivere con maggior precisione le caratteristiche del carico e di solito fa minor uso di assunzioni sui dati; purtroppo non permette di ottenere tutti i vantaggi che si otterrebbero utilizzando un modello generativo; in particolare, l'analisi quantitativa di grandezze derivate dalla combinazione di diverse caratteristiche potrebbe essere di difficile attuazione, a causa del fatto che non tutte le caratteristiche sono stato modellate o che la combinazione del comportamento delle varie caratteristiche porta a modelli matematici poco trattabili; anche la simulazione del comportamento congiunto di diverse caratteristiche è resa più complicata dal fatto che occorre effettuare i campionamenti da distribuzioni multivariate.

A parte il tipo di modello che si decide di utilizzare, la costruzione di un modello dovrebbe comunque essere effettuata tenendo ben in mente il principio di Occam [3], secondo il quale il modello costruito deve essere il più semplice possibile; l'applicazione di questo principio porta a numerosi vantaggi, in particolare rende il modello semplice da trattare e da utilizzare nella pratica.

Il tipo di caratterizzazione del carico che si è deciso di effettuare nel presente progetto, ha come sbocco naturale quello dell'ottimizzazione delle tecniche di schedulazione dei job e lo sviluppo di nuove politiche basate sui risultati ottenuto da questo lavoro; perciò, l'enfasi verrà posta sullo studio delle caratteristiche che possono influenzare le prestazioni dello "scheduler". Per giungere a tale obiettivo, si è deciso di concentrare gli sforzi sulla costruzione di un modello descrittivo; questa decisione è rafforzata dalla convinzione che la costruzione di un modello generativo sufficientemente generico, e nello stesso tempo rappresentativo, può solo essere effettuata dopo un accurato studio delle caratteristiche dei carichi di diversi sistemi Grid. Il presente studio è, inoltre, limitato all'analisi delle singole caratteristiche del carico; l'analisi della correlazione tra due o più caratteristiche è lasciata a lavori futuri.

[3] Il principio di Occam (Occam's razor) esprime la "legge della parsimonia" (*lex parsimoniae*): "Non aggiungere elementi quando non serve" (*Entia non sunt multiplicanda praeter necessitatem*).

2.3 Scelta delle Caratteristiche del Carico

Per *caratteristica* del carico si intende una particolare grandezza osservabile che fa parte del carico di un sistema; spesso i termini caratteristica, *attributo*, *feature* e *factor* vengono usati come sinonimi.

Non esistono tecniche particolari per decidere quali siano le caratteristiche più importanti (rappresentative) per descrivere il carico di un sistema; il criterio di scelta deve sicuramente basarsi sull'obiettivo che si intende raggiungere. Per esempio, se l'obiettivo è il miglioramento delle politiche di scheduling, occorrerà tenere in considerazione i tempi di interarrivo dei job e la durata della loro esecuzione; tuttavia, oltre a questi attributi, potrebbero essercene altri che influenzano in modo indiretto i primi.

Occorre anche far notare che, nella pratica, la scelta delle caratteristiche è spesso condizionata dai dati che si hanno a disposizione; per esempio, una caratteristica che potrebbe essere interessante al fine di ottimizzare le politiche di "scheduling" è la dimensione delle cosidette *Bag-of-Task* (*BoT*); un BoT non è nient'altro che un insieme di task, relativi a una stessa applicazione, indipendenti fra loro e che, quindi, possono essere eseguiti in modo parallelo. L'utilizzo di BoT trova spazio in diversi scenari: data-mining, simulazioni Monte Carlo, biologia computazionale, ... La maggior parte di questi scenari rappresentano anche quelli associati alle applicazioni eseguite nei sistemi Grid. Lo studio delle prestazioni e di politiche di "scheduling" adeguate per questo tipo di applicazioni, ha suscitato molto interesse nella comunità scientifica coinvolta nei sistemi Grid (ad es., si veda [71, 21, 20]). Da ciò risulta chiaro come un'accurata caratterizzazione del carico, rivolta allo "scheduling", dovrebbe tenere in considerazione anche le caratteristiche associate ai BoT (ad es., tempo di interarrivo di un BoT, tempo di completamento di tutti i task dell'intero BoT, ...); purtroppo non tutte le tracce dispongono di questa tipologia di informazioni.

2.4 Raccolta dei Dati

Uno dei problemi principali che si deve affrontare nella creazione di un modello, riguarda la disponibilità dei dati; in modo particolare, nel contesto del Grid Computing, la caratterizzazione del carico è un campo di ricerca piuttosto recente, per cui le tracce disponibili sono ancora molto limitate.

Oltre al reperimento dei dati, vi sono due altri aspetti, non meno importanti, da tenere in considerazione: la rappresentatività e la numerosità dei dati. L'utilizzo di dati non rappresentativi ha come conseguenza la creazione di modelli altrettanto non rappresentativi; per esempio se una traccia contiene dati relativi a un periodo in cui nel sistema erano presenti solo job "corti", l'effetto che ne segue è che l'eventuale presenza di job "lunghi" può provocare un degrado delle prestazioni. Per quanto riguarda la numerosità dei dati, si tratta di un aspetto importante in quanto, anche nel caso in cui i dati a disposizione siano rappresentativi, potrebbero non essere sufficienti a individuare le caratteristiche peculiari del carico, oppure il loro numero potrebbe essere troppo piccolo per applicare le tecniche statistiche tradizionali.

Quindi, dati poco rappresentativi o poco numerosi potrebbero non fare emergere aspetti del carico come:

- *ciclicità*: il carico diurno potrebbe comportarsi in modo differente da quello notturno; stesso discorso per il carico relativo alle pause pranzo, ai periodi di ferie, ...
- *carico multi-classe*: il carico potrebbe comportarsi in maniera diversa a seconda del tipo di utenti, del tipo di job (batch o interattivi), ...
- *valori estremi*: per modellare le code di una distribuzione è necessario un numero piuttosto grande di osservazioni, in modo da aumentare la probabilità che tali eventi estremi della distribuzione si verifichino;
- *autocorrelazione* e *correlazione* tra due o più attributi: più i dati sono numerosi e rappresentativi, maggiore è la precisione con cui si riesce a catturare e a modellare le correlazioni tra uno stesso attributo (come la *long range dependence*) o tra attributi differenti (*cross correlation*).

2.5 Bonifica dei Dati

Prima di procedere con la costruzione di un modello occorre esaminare i dati alla ricerca di informazioni errate, incomplete o, in generale, anomale. Questa fase dell'analisi è fondamentale in quanto la presenza di dati non corretti può portare alla costruzione di un modello con proprietà ben lontane da quelle reali. La soluzione più semplice che si può adottare in presenza di dati anomali è la loro rimozione; tuttavia, ciò non sempre rappresenta la soluzione migliore. Infatti, in alcuni casi, valori estremamente grandi o estremamente piccoli non rappresentano situazioni anomale, bensì fanno parte del comportamento intrinseco del sistema; rappresentano, cioè, situazioni eccezionali e abbastanza rare, ma non così rare da poterle escludere.

In generale, i tipi di dati anomali possono essere distinti nelle seguenti classi [43]:

- *Missing Value*. L'informazione non appare fra i dati e non è nemmeno derivabile fra quelli disponibili. Un esempio tipico è quello riguardante l'analisi dei log di un server web di un sistema in cui le richieste web sono filtrate da un proxy; in tal caso, il log del server web conterrà solo le richieste che hanno superato il proxy; quelle che sono state bloccate dal proxy o che si trovavano nella sua "cache" non sono disponibili a questo livello. Il fatto che questo costituisca un problema dipende, ovviamente, dal tipo di aspetto che si intende analizzare (ad es., carico del server web oppure numero o tipologia di richieste web). Nel contesto del Grid Computing, un fatto simile potrebbe capitare nel caso in cui si volesse esaminare il carico generato dalla fase di "stage-in": di solito, il "middleware" Grid provvede a effettuare il trasferimento dei file sui nodi di computazione solo se realmente necessario, cioè solo se questi non sono già presenti sulla macchina; anche in tal caso, ciò potrebbe rappresentare un problema allorquando si volesse studiare il carico derivante dalla fase di "stage-in" con lo scopo, per esempio, di migliorare o implementare politiche di "caching" o di "file sharing".

- *Censored Data*. L'informazione appare parzialmente specificata. Un esem-

pio tipico riguarda la durata dell'esecuzione di un job; potrebbe capitare che il log esaminato contenga l'informazione sulla schedulazione di un job ma non sulla sua terminazione in quanto, nel momento in cui la traccia è stata utilizzata per l'analisi, il job era ancora in esecuzione.

- *Granularità*. L'informazione è presente ma a una granularità diversa da quella che si intende trattare. Un caso del genere può verificarsi nell'analisi del traffico di rete; si supponga, ad esempio, che si voglia analizzare il traffico TCP di una certa rete; nel caso in cui si abbiano a disposizione delle tracce contenenti solo la registrazione di pacchetti IP, sarà necessario una fase di "pre-processing" per individuare e ricostruire le varie sessioni TCP e per scartare i pacchetti relativi ad altri protocolli, come UDP e ICMP.

- *Flurry*. I dati presentano pochi "picchi" ("burst") di grande intensità. Di solito rappresentano dei momenti di attività intensa causati da entià isolate del sistema (come utenti,...). La rimozione di questi "picchi" è abbastanza semplice: basta eliminare i dati che superano una certa soglia. Il problema reale è quello di capire se effettivamente questi dati rappresentino degli "outlier" e quindi vadano rimossi o se invece non debbano essere considerati come una componente caratteristica del carico. Per esempio, nella maggior parte dei sistemi operativi moderni, sono presenti dei processi di sistema il cui scopo è quello di ottimizzare l'esecuzione delle applicazioni installate; in particolare, in molti sistemi Linux, il processo *prelink* viene eseguito automaticamente una volta al giorno per modificare gli eseguibili e le librerie dinamiche con il fine di ridurre il tempo di caricamento e rilocazione da parte del "linker"; anche se la durata dell'esecuzione del comando "prelink" è abbastanza breve, il carico sulla CPU e sul disco generato non è trascurabile. Si tratta quindi di un tipico esempio di "flurry" che non costuisce un evento eccezionale ma piuttosto un comportamento tipico e rappresentativo del sistema.

- *Altre Anomalie*. La presenza di informazioni anomale potrebbe essere causata anche da eventi eccezionali come una caduta di corrente, la rottu-

ra di una macchina, la rete non raggiungibile, lo spazio su disco esaurito, errori software, . . . Per esempio l'evento "host irraggiungibile" causerebbe un aumento dei tempi di risposta, di trasferimento dei file e del traffico sulla rete. Un esempio ancor più ecclatante riguarda la traccia "'98 Soccer World Cup" [11]; questo file contiene, circa, 1.3 miliardi di registrazioni di richieste al sito www.france98.com, per un arco temporale pari a 3 mesi. Ogni registrazione specifica, tra le altre informazioni, la dimensione del documento trasferito [4]; eccezionalmente, circa 125 milioni di richieste riguardano il trasferimento di file di dimensione pari a $2^{32} - 1$ (circa 4GB), quando la dimensione media associata alle restanti richieste è pari a 15.5KB; si tratta in effetti di un caso di errore dovuto, probabilmente, alla scelta sbagliata del tipo di dati per rappresentare l'informazione: quasi sicuramente, l'intenzione era quella di registrare il valore -1 (dimensione sconosciuta), il quale, utilizzando tipi "unsigned" a 32 bit, viene trasformato in $2^{32} - 1$.

Per gestire le situazioni suddette occorre effettuare un'analisi accurata e avere una buona conoscenza del dominio. Per l'analisi si possono utilizzare:

- procedure di ispezione dei dati scritte ad-hoc;
- metodi grafici della statistica descrittiva;
- metodi numerici della statistica descrittiva;
- metodi numerici della statistica inferenziale.

La conoscenza del dominio, e l'eventuale possibilità di confrontarsi con gli esperti della fonte da cui provengono i dati (ad es., amministratori di sistema), permette di effettuare una selezione più accurata tra i casi di "falsi outlier" (cioè, valori estremi che sono parte integrante del comportamento del sistema) e "veri outlier" (cioè valori anomali o errati).

I punti appena elencati costituiscono soltanto delle linee guida che dovrebbero essere applicate con molta cautela; l'operazione di rimozione di un "ou-

[4] Si tratta del campo `byte` del log.

tlier" è, in generale, un'operazione pericolosa in quanto potrebbe compromettere i risultati della caratterizzazione; inoltre, in alcuni casi si ha a che fare con insiemi di dati di dimensione ridotta e quindi la rimozione di qualche osservazione renderebbe ancora più piccolo l'insieme. Di conseguenza, prima di procedere con la rimozione di uno o più "outlier", occorre tenere bene in mente le seguenti due regole:

- per ogni "outlier" che si intende rimuovere occorre trovare una ragionevole spiegazione;
- quando possibile, cercare di correggere il "outlier", eventualmente sotto la supervisione di un esperto.

2.5.1 Procedure di Ispezione Dati

Prima di verificare la qualità dei dati attraverso metodi statistici, conviene, quando possibile, implementare delle procedure di ispezione dei dati, al fine di cercare delle incongruenze o delle informazioni mancanti; il tipo di procedura da sviluppare dipende, ovviamente, dal contesto, cioè dalla tipologia di dati che si ha a disposizione.

Per esempio, per l'analisi della traccia TeraGrid (Cap. 8), è stata costruita una procedura per effettuare i seguenti controlli:

- la data di sottomissione di un job a uno "scheduler" non sia superiore alla data in cui il job viene schedulato per l'esecuzione;
- la data di schedulazione di un job non sia superiore alla data della sua terminazione;
- il numero di nodi sia un numero intero positivo;
- non vi siano informazioni incomplete o mancanti.

Una volta individuate le anomalie, prima di rimuoverle, conviene cercare di correggerle; tale operazione non è sempre facile e possibile. Per esempio, nel caso della traccia TeraGrid, le registrazioni aventi la data di sottomissione di

un job superiore a quella della sua schedulazione, possono essere facilmente corrette, impostando la data di sottomissione uguale alla data di schedulazione, oppure impostando la data di schedulazione alla data di sottomissione [5]. Invece, per quanto concerne le anomalie riguardanti il numero di nodi, non esiste una correzione diretta in quanto non è possibile sapere il numero di nodi utilizzati da un job; un'eventuale indagine presso un esperto della fonte da cui proviene la traccia potrebbe risultare molto utile in questi casi. Infine, l'assenza di qualche informazione, come il tempo di terminazione di un job, rappresenta un fatto che non può essere corretto; anche in questo caso, però, occorrerebbe capire se l'assenza di tale informazione è dovuta a un errore o, piuttosto, al fatto che il job si trovi ancora in esecuzione; in quest'ultimo caso, anzichè rimuovere l'osservazione, si potrebbe trattare come un valore estremo della coda (destra) della distribuzione sottostante.

2.5.2 Metodi Grafici della Statistica Descrittiva

I metodi grafici consentono di valutare la distribuzione delle osservazioni dal punto di vista grafico sotto vari aspetti, come centralità, dispersione e asimmetria, mettendo in grado l'analista di comprendere meglio se e quali osservazioni rappresentano dei "veri outlier". L'approccio EDA consiglia di tracciare i seguenti grafici:

- *Istogramma delle Frequenze*: permette di ottenere un andamento approssimato della forma della distribuzione in funzione delle frequenze assolute ottenute dalle osservazioni; si tratta di un grafico che divide le osservazioni in classi (*bin*), di solito equispaziate, e per ogni classe mostra il numero di osservazioni che vi appartengono; vi sono diversi modo per dividere l'intervallo delle osservazioni in classi, nessuno dei quali risulta ottimale; fra questi, quelli più utilizzati sono:
  - *Formula di Sturges*:
    $$\#\text{bins} = \lceil \log_2 n + 1 \rceil$$

[5] In questo caso, andrebbe aggiustata anche la data di terminazione in modo da rispettare la durata del job così come appare nella traccia.

dal quale si ottiene:

$$\text{bin-width} = \frac{\max_i \{x_i\} - \min_i \{x_i\}}{\log_2 n + 1}$$

Questo metodo lega, implicitamente, la dimensione delle classi al numero di osservazioni; per questo motivo è considerato poco adatto per campioni di piccola numerosità ($n < 30$) o in presenza di un numero discreto di "outlier" (i quali avrebbero l'effetto di allargare la dimensione delle classi).

– *Regola di Scott* [103]:

$$\text{bin-width} = 3.5sn^{-\frac{1}{3}}$$

dove s è la deviazione standard campionaria corretta.

– *Regola di Freedman-Diaconis* [48]:

$$\text{bin-width} = 2\,\text{IQR}\,(x)\,n^{-\frac{1}{3}}$$

dove $\text{IQR}(x)$ è l'intervallo inter-quartile del campione x; questo metodo risulta più immune agli "outlier", rispetto a quello di Scott, in quanto utilizza statistiche robuste; inoltre tende a creare delle classi più piccole rispetto ai metodi precedenti.

Una variante dell'istogramma delle frequenze è quello dell'istogramma delle frequenze relative, il quale ha la stessa forma di quello delle frequenze assolute, ma varia la scale dell'asse delle ordinate e l'area totale dell'istogramma vale 1.

- *Density Plot*: permette di ottenere una stima della densità della distribuzione in funzione delle frequenze relative ottenute dalle osservazioni. Per ricavare la stima della densità della distribuzione di un campione di osservazioni $x_1, \ldots, x_n$, si utilizza il concetto di stima *kernel-density* (o

Parzen window):

$$\hat{f}_b(x) = \frac{1}{nb} \sum_{i=1}^{n} \mathrm{K}\left(\frac{x - x_i}{b}\right)$$

dove $\mathrm{K}(\cdot)$ è una funzione *kernel* e b la *larghezza di banda* (*bandwidth*), detta anche parametro *smoothing*. Una funzione kernel $\mathrm{K}(\cdot)$ è una funzione che soddisfa le seguenti proprietà:

- $\forall t : \mathrm{K}(t) \geq 0$
- $\int_{-\infty}^{\infty} \mathrm{K}\ dt = 1$
- $\forall t : \mathrm{K}(-t) = \mathrm{K}(t)$ (simmetria)

La stima della vera densità $f(x)$ nel punto x è determinata dalla stima $\hat{f}_b(x)$, la quale varia a seconda della funzione kernel scelta e del valore della "bandwidth". La "bandwidth" b determina la larghezza del *vicinato* (*neighborhood*), o *finestra*, del punto x, al cui interno si trovano i punti che influenzano il calcolo della stima di x. Tra le funzioni kernel più utilizzate vi è [6]:

- kernel *Gaussiano*:
$$\mathrm{K}(t) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{t^2}{2}\right)$$
- kernel *Triangolare*:
$$\mathrm{K}(t) = (1 - |t|)\ \mathbb{I}(|t| \leq 1)$$
- kernel *Epanechnikov*:
$$\mathrm{K}(t) = \frac{3}{4}\left(1 - t^2\right)\ \mathbb{I}(|t| \leq 1)$$

La scelta del parametro "bandwidth" risulta tanto importante quanto lo è la scelta della larghezza dei "bin" per un istogramma delle frequen-

[6] Gli esempi di funzioni kernel sono forniti in versione "standardizzata", cioè dove la "bandwidth" b viene specificata indirettamente attraverso la variabile t; ad es., $t = \frac{x - x_i}{b}$.

ze. Il vantaggio di questo tipo di grafico rispetto all'istogramma delle frequenze, è, la possibilità di sovrapporre diversi istogrammi di densità, relativi a uno stesso insieme di dati e ricavati da differenti stimatori oppure relativi a insiemi differenti e ottenuti con lo stesso stimatore, al fine di effettuare un confronto; tale operazione sarebbe ardua nel caso si utilizzassero gli istogrammi delle frequenze. Inoltre, risulta più indicato per distribuzioni continue, in quanto l'istogramma delle frequenze, a causa della discretizzazione dovuta al "binning", produce, in generale, una stima grossolana della vera densità.

- *Box plot*: fornisce, in modo robusto, un'informazione sulla dispersione e sulla presenza di "outlier". Utilizza la cosidetta regola del *box plot* [110], in cui si sfrutta il concetto di *interquartile range* (*IQR*) per identificare i possibili "outlier": se un'osservazione x è compresa tra Q_3 e Q_1 (rispettivamente, primo e terzo quartile), non si tratta di un "outlier"; viceversa, si parla di '*mild outlier* se $x \in (Q_3 + k \cdot IQR, Q_3)$ o $x \in (Q_1, Q_1 - k \cdot IQR)$, o di *extreme outlier* negli altri casi. La costante k di solito è uguale a 1.5. Oltre alla versione originale di Tukey, ne esistono numerose varianti che differiscono principalmente per come calcolano i quartili [49].

- *Normal Q-Q plot*: fornisce un modo visivo per verificare la *normalità* dei dati, cioè se è ragionevale supporre che la distribuzione delle osservazioni sia approssimativamente Normale. Si tratta di un Q-Q plot §3.2.1 tra le statistiche d'ordine campionarie e i corrispondenti quantili di una Normale standard (cioè, di una Normale con media zero e varianza pari a uno); se i punti del grafico si distribuiscono in modo lineare, in particolare, attorno alla retta con pendenza pari a 45 gradi e passante per l'origine, si può ipotizzare che la distribuzione dei dati sia approssimativamente Normale, altrimenti no [7]. Il fatto di sapere che la distribuzione dei dati segue, approssimativamente, una Normale ha il vantaggio di poter uti-

[7] Il risultato ottenuto è, ovviamente, confinato al campione osservato; infatti potrebbe succedere che la distribuzione della popolazione sia approssimativamente Normale anche se il test sulla normalità applicato a un certo campione fallisce.

lizzare tutte le tecniche e i test della statistica che prevedono l'assunzione della normalità.

Dei quattro grafici, l'istogramma delle frequenze e il "density plot" sono più adatti per ottenere informazioni riguardo la forma della distribuzione, mentre il "Normal Q-Q plot" e il "box plot", in particolare, sono più utili all'individuazione di "outlier". La Fig. 2.1 mostra un esempio dei quattro tipi di grafici, relativo a un campione di 200 elementi, estratto da una Weibull$(1.3, 8.7)$; dal "box plot" e dal "Q-Q plot" si evince la presenza di almeno 3 possibili "outlier" nella coda destra; inoltre, dal "Q-Q plot" e dall'istogramma delle frequenze, si può notare come la coda destra sia più lunga di quella sinistra; il "density plot" fornisce un'idea della possibile forma della densità della distribuzione; infine, dal "Normal Q-Q plot" si ricava che, mentre il corpo della distribuzione potrebbe essere approssimato con una Normale o, ancora meglio, tramite una distribuzione Gamma, le sue code, in particolare quella destra, devono essere modellate in modo differente.

Oltre ai grafici suddetti, si possono costruire dei grafici ad-hoc, che dipendono dal dominio di applicazione e dalla caratteristica del carico che si sta studiando . Per esempio, per identificare i "flurry" si può costruire l'istogramma di una certa caratteristica analizzata su diverse scale temporali (minuti, ore, giorni, ...) o partizionata a seconda dell'utilizzo (come il numero di job utilizzati da ogni utente) [43]. La Fig. 2.2 mostra un esempio di carico in cui i "flurry" sono in realtà dovuti all'attività di singoli utenti; casi come questo andrebbero modellati come carico *multi-classe*. Questo tipo di grafico permette di capire se la presenza di uno o più "flurry" possa essere imputata a poche entità del sistema.

2.5.3 Metodi Numerici della Statistica Descrittiva

I metodi numerici della statistica descrittiva includono le misure di dispersione, di locazione (o centralità) e di correlazione; in particolare, le misure di dispersione e locazione permettono di ottenere una misura quantitativa sulla variabilità dei dati rispetto alla loro media; mentre quelle di correlazione per-

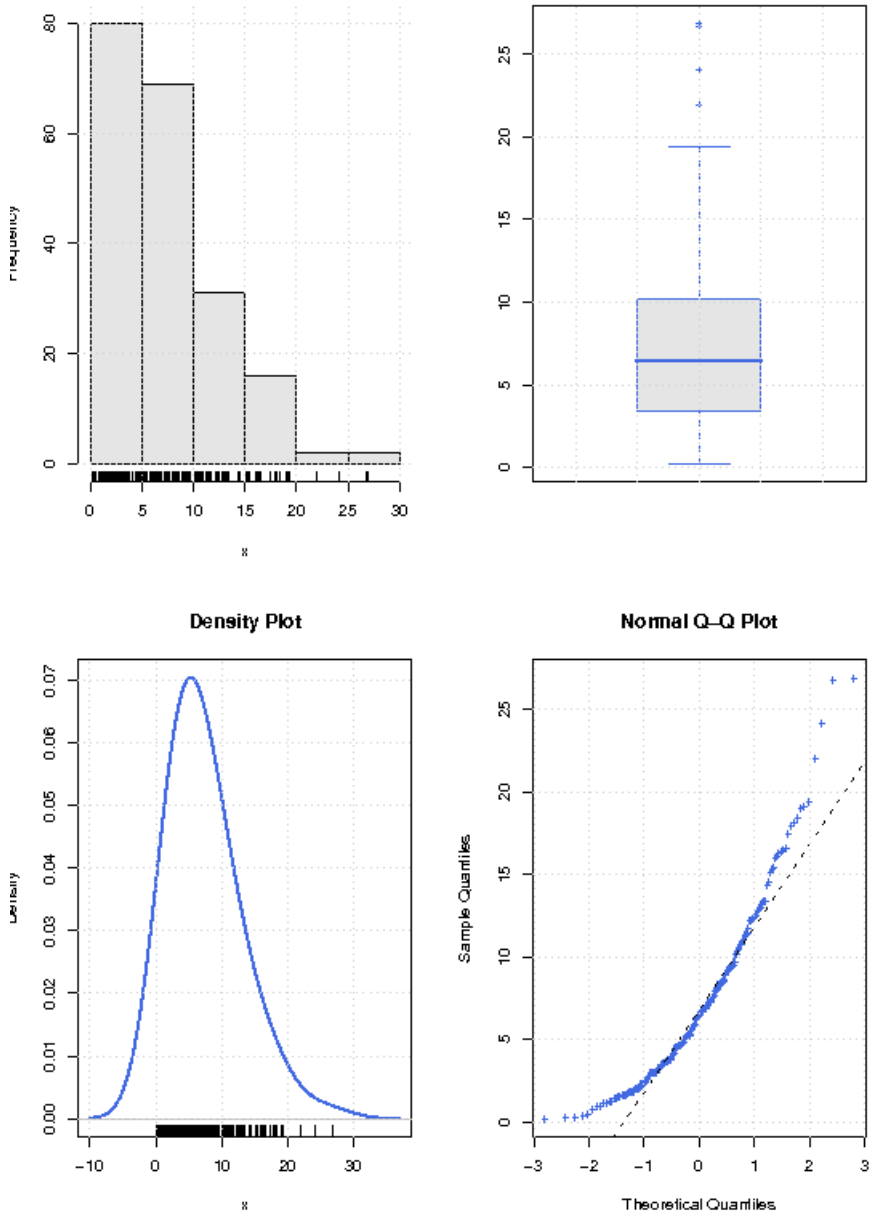
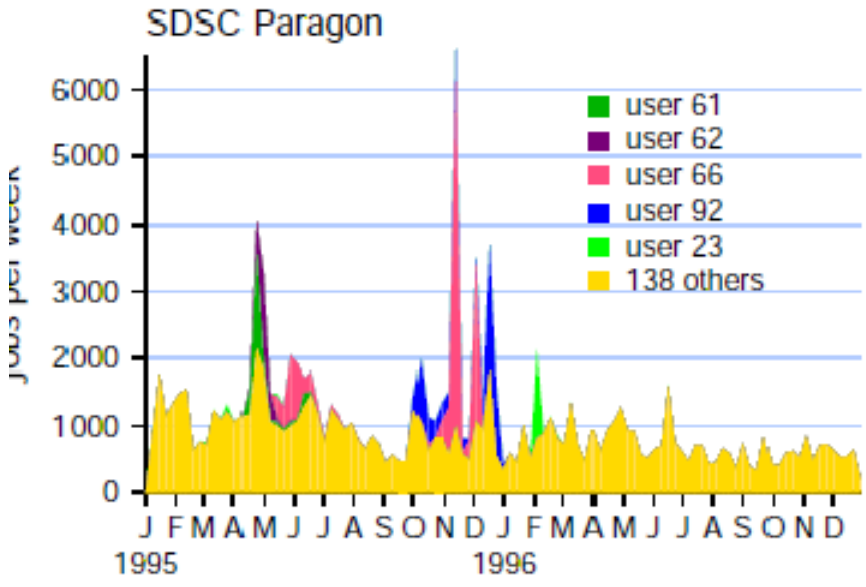


Figura 2.1: Grafico EDA per un campione Weibull$(1.3, 8.7)$ di numerosità 200.


SDSC Paragon
6000
5000
4000
3000
2000
1000
0
user 61
user 62
user 66
user 92
user 23
138 others
J F M A M J J A S O N D J F M A M J J A S O N D
1995
1996

mettono, per esempio, di avere informazioni sull'indipendenza di una o più caratteristiche sotto osservazione. Occorre far notare che l'utilizzo dei momenti campionari per l'analisi sulla variabilità non è sempre consigliato in quanto tali statistiche non sono robuste rispetto alla presenza di "outlier"; inoltre, più aumenta l'ordine dei momenti, maggiore è la dipendenza dai valori estremi. In genere, è quindi consigliato affiancare i valori ottenuti dai momenti campionari con quelli relativi alle statistiche d'ordine, come la mediana o l'intervallo inter-quartile, giacchè quest'ultime sono di solito robuste alla presenza di "outlier".

Oltre all'utilizzo delle statistiche d'ordine, è possibile utilizzare altri tipi di indicatori come la media geometrica o quella armonica, le quali risultano, di solito, meno sensibili alla presenza di "outlier"; per esempio, nel calcolo della media geometrica, le osservazioni contribuiscono con lo stesso peso, in quanto esse vengono moltiplicate fra loro; nella media aritmetica (momento campionario del primo ordine), invece, le osservazioni estreme hanno un peso maggiore in quanto l'operazione utilizzata nel calcolo è la somma.

Dato che questi metodi possono essere utilizzati sia per l'individuazione degli "outlier" sia per l'analisi della forma e delle proprietà della densità della distribuzione, la loro trattazione è rimandata alla sezione §2.6.

2.5.4 Metodo Numerici della Statistica Inferenziale

Tra i metodi più utilizzati vi è il *test di Grubbs* [54]: si tratta di un test di ipotesi dove:

$$\mathcal{H}_0 : \text{l'insieme di dati } X_1, \ldots, X_n \text{ non contiene "outlier"}$$
$$\mathcal{H}_1 : \text{c'è almeno un outlier in } X_1, \ldots, X_n$$

In pratica:

- per test a una coda, si calcola la statistica:

$$G = \frac{\max\{X_i\} - \bar{X}}{s}$$

oppure:

$$G = \frac{\bar{X} - \min\{X_i\}}{s}$$

per verificare, rispettivamente, se il massimo o il minimo valore sono "outlier';

- per test a due code, si calcola la statistica:

$$G = \frac{\max\left\{\left|X_i - \bar{X}\right|\right\}}{s}$$

per verificare se sia il minimo sia il massimo valore sono "outlier";

dove $\bar{X}$ indica la media campionaria e s lo scarto quadratico medio campionario corretto dell'insieme $X_1, \ldots, X_n$. L'ipotesi nulla viene rifiutata con significatività α se:

$$G > \frac{n-1}{\sqrt{n}} \sqrt{\frac{t^2_{\alpha/2n,n-2}}{n-2+t^2_{\alpha/2n,n-2}}}$$

dove $t_{\alpha/2n,n-2}$ rappresenta il quantile di una variabile T_{n-2} distribuita secondo una t-Student a $n-2$ gradi di libertà, tale per cui:

$$\Pr\left\{T_{n-2} > t_{\frac{\alpha}{2n},n-2}\right\} = \frac{\alpha}{2n}$$

Dato che il test riesce a individuare uno o due "outlier" per volta, per poterne individuare un numero superiore, occorre ripeterlo più volte, in modo tale che ad ogni iterazione si verifichi la presenza di "outlier" su un insieme di dati più piccolo, privato del "outlier" trovato nell'iterazione precedente. Di solito si consiglia di applicare il test di Grubbs solo quando si pensa che le osservazioni siano approssimativamente distribuite secondo una Normale, in quanto tale test assume che i dati siano normalmente distribuiti e identifica i possibili "outlier" come quelle osservazioni che si discostano eccessivamente dalla normalità.

In generale, i metodi che fanno parte di questa categoria sono adatti ai casi in cui si ha un'idea del tipo di distribuzione che governa i dati; in effetti, nel caso in cui non si conosca il tipo di distribuzione della popolazione, l'utilizzo di questi metodi può portare allo scarto di "falsi outlier", cioè di valori che, benchè estremi, non sono così rari da poter essere considerati "outlier"; questo è il caso di dati generati da una distribuzione le cui code sono governate da leggi polinomiali (come le distribuzioni *heavy-tailed*): in questo caso i valori estremi non devono essere scartati in quanto caratterizzano il tipo di distribuzione e quindi devono essere parte integrante del processo di caratterizzazione.

2.6 Analisi delle Proprietà Statistiche dei Dati

Le principali proprietà statistiche ricavabili dai dati, riguardano le misure della tendenza centrale, per individuare il centro della distribuzione e la posizione del valor medio, della dispersione, per valutare la variabilità dei dati rispetto al centro della distribuzione o al valor medio, e della correlazione, per studiare il tipo di legami tra le osservazioni di una stessa caratteristica o tra caratteristiche differenti.

2.6.1 Misure della Tendenza Centrale

Lo scopo delle misure *di tendenza centrale*, anche dette *di locazione*, è quello di fornire un'indicazione sul centro della popolazione; ciò è utile per determinare il valore atteso di un campione o dove, in media, le osservazioni estratte dalla popolazione tendano a distribuirsi.

Media

Esistono diversi tipi di media; quelli più utilizzati sono:

- *Media Aritmetica*

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

- *Media Geometrica*

$$G = \sqrt[n]{\prod_{i=1}^{n} x_i}$$

- *Media Armonica*

$$H = \left(\sum_{i=1}^{n} \frac{1}{x_i} \right)^{-1}$$

Benchè la media aritmetica sia la più utilizzata, essa risulta anche quella più vulnerabile alla presenza di "outlier"; d'altra parte, nel caso in cui nei dati non siano presenti degli "outlier", la media aritmetica rappresenta in genere uno stimatore *MVUE* della media reale §3.1. Gli altri due tipi di medie sono delle misure più resistenti agli "outlier", malgrado ne risentano anch'esse della presenza. La media geometrica è di solito utilizzata quando occorre effettuare dei confronti, come il confronto tra il comportamento medio di un nuovo e di un vecchio modello; in questo caso è più conveniente effettuare il rapporto tra le medie geometriche anzichè quello tra le medie aritmetiche. La media armonica è indicata quando occorre calcolare il valor medio di tassi.

Mediana

La *mediana campionaria* (Def. B.2.10) è quel valore che divide a metà le osservazioni di un campione. È sempre utile considerare la mediana nell'analisi della centralità in quanto si tratta di una misura robusta, rispetto alla presenza di "outlier", del centro della distribuzione.

Moda

La *moda campionaria* è il valore più frequente fra le osservazioni di un campione. Se una distribuzione è approssimativamente simmetrica, la moda fornisce un buon indicatore del centro della popolazione; invece, se la distribuzione è asimmetrica, la moda dà un'idea del lato della distribuzione in cui si concentra la maggior parte dei dati. Come la mediana, si tratta di uno stimatore robusto alla presenza di "outlier"; occorre però far notare che mentre la mo-

da campionaria esiste sempre, quella della distribuzione potrebbe non essere unica (come nelle distribuzioni multimodali) o addirittura avere infiniti valori (come nella distribuzione Uniforme); quindi, non si tratta sempre di un indicatore rappresentativo della distribuzione dei dati.

2.6.2 Misure della Dispersione

Le misure *di dispersione*, anche dette *di scala*, forniscono un'indicazione sulla variabilità di un insieme di dati.

Range

È la differenza tra il massimo e il minimo valore delle osservazioni:

$$R = \max_i \{x_i\} - \min_i \{x_i\}$$

Si tratta di una delle misure più semplici ma anche tra le più sensibili agli "outlier"; infatti se i dati contengono degli "outlier", il massimo o il minimo dei dati sarà proprio un "outlier", e il valore del "range" rispecchierà l'ordine di grandezza degli "outlier".

Varianza

La *varianza campionaria* (corretta) è una stima del momento del secondo ordine della popolazione, centrato intorno alla media reale (Def. B.2.1):

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

dove $\bar{x}$ è la media campionaria (aritmetica) delle osservazioni. In genere si preferisce utilizzare lo *scarto quadratico medio* s (anche detto *standard error* o *deviazione standard*), cioè la radice quadrata di s^2, in quanto, essendo espresso nella stessa unità dei dati, è maggiormente confrontabile con i dati e con il valore medio campionario. Come per tutte le misure che coinvolgono i momenti, si tratta di un indicatore non robusto alla presenza di "outlier", sebbene, nei

casi in cui i dati sono privi di "outlier" e sono distribuiti, approssimamente, come una Normale, rappresenti uno stimatore *MVUE* della varianza reale §3.1.

Mediana della Deviazione Assoluta dalla Mediana (MAD)

La *Mediana della Deviazione Assoluta dalla Mediana* (*Median Absolute Deviation*, in breve *MAD*) rappresenta la mediana degli scostamenti assoluti dei dati dalla loro mediana campionaria:

$$\mathrm{MAD}\left(x_1, \ldots, x_n\right) = \mathrm{median}\left(\sum_{i=1}^{n} \left|x_i - \mathrm{median}\left(x_1, \ldots, x_n\right)\right|\right)$$

dove con $\mathrm{median}\left(x_1, \ldots, x_n\right)$ si intende la mediana campionaria delle osservazioni $x_1, \ldots, x_n$. Sebbene MAD non sia uno stimatore *MVUE* dello scarto quadratico medio della popolazione, (nel caso in cui i dati siano approssimativamente Normali ed esenti da "outlier"), rappresenta un indicatore di dispersione più robusto agli "outlier", rispetto alla deviazione standard, grazie all'utilizzo della mediana.

Alcuni software statistici, come *MATLAB*, adottano come definizione di "default" per MAD, la *deviazione assoluta media dalla media* (*Mean Absolute Deviation*):

$$\mathrm{mean}\left(\sum_{i=1}^{n} \left|x_i - \mathrm{mean}\left(x_1, \ldots, x_n\right)\right|\right)$$

dove con $\mathrm{mean}\left(x_1, \ldots, x_n\right)$ si intende la media aritmetica delle osservazioni $x_1, \ldots, x_n$. Si tratta però di una misura vulnerabile agli "outlier" a causa dell'utilizzo della media campionaria.

Intervallo Inter-Quartile (IQR)

L'*Intervallo Inter-Quartile* (*Interquartile Range*, in breve *IQR*) è la differenza tra il terzo e il primo quartile delle osservazioni $x_1, \ldots, x_n$:

$$\mathrm{IQR}\left(x_1, \ldots, x_n\right) = Q_3 - Q_1$$

dove Q_3 e Q_1 rappresentano, rispettivamente, il terzo e il primo quartile (cioè, il 75-esimo e 25-esimo percentile, Def. B.1.6) dei dati. Questo indicatore è robusto rispetto agli "outlier" in quanto è influenzato solo dal 50% dei dati centrali (ossia, da quelli intorno alla mediana).

2.6.3 Misure della Forma

Le *misure della forma* consentono di ottenere delle informazioni sulla forma della distribuzione di un insieme di dati, in particolare della sua densità. Nella sezione riguardante la bonifica dei dati §2.5 sono già stati presentati alcuni metodi grafici per ricavare delle informazioni sulla forma dai dati §2.5.2; per esempio, un istogramma o un "density plot" consentono di rilevare il tipo di simmetria, oppure un "box plot" o un "Q-Q plot" permettono di ottenere informazioni sulla lunghezza della coda. Altre informazioni che si possono ottenere riguardano, il tipo di asimmetria (destra o sinistra) e il tipo del picco. La maggior parte delle misure utilizza come distribuzione di riferimento la distribuzione Normale, fornendo quindi un indice dello scostamento della forma della distribuzione dei dati da quella di una Normale.

Asimmetria

L'*asimmetria* (*skewness*) campionaria fornisce una misura del grado di asimmetria, o di simmetria, della distribuzione dei dati:

$$g_1 = \frac{m_3}{m_2^{3/2}}$$

dove $x_1, \ldots, x_n$ è un campione di osservazioni, m_k è il momento campionario centrale di ordine k (Def. B.2.2) e $\bar{x}$ è la media campionaria. Dato che g_1 è uno stimatore non corretto, di solito si preferisce utilizzare la misura nota come *indice di asimmetria di Fisher*:

$$G_1 = g_1 \frac{\sqrt{n(n-1)}}{n-2}$$

Se il valore ottenuto è negativo, i dati tendono a disperdersi verso sinistra (*asimmetria sinistra o negativa*), cioè vi sono più osservazioni nella coda sinistra di quanto vi sarebbero in una distribuzione Normale; un valore positivo, invece, indica una dispersione verso destra (*asimmetria destra o positiva*), ossia vi sono più osservazioni nella coda destra di quanto vi sarebbero in una distribuzione Normale. Come riferimento si utilizza il fatto che l'asimmetria di una distribuzione Normale è pari a zero [8]. Più il valore di asimmetria si allontana da zero, più la coda della distribuzione sarà "lunga".

Curtosi

La *curtosi* (*kurtosis*) campionaria fornisce un'informazione sul grado di ripidezza (appiattimento o allungamento) della distribuzione dei dati rispetto a una Normale:

$$g_2 = \frac{m_4}{m_2^2}$$

dove m_k rappresenta il momento campionario centrale di ordine k (Def. B.2.2) del campione di osservazioni $x_1, \ldots, x_n$. La curtosi di una distribuzione Normale è pari a 3; la quantità:

$$g_2' = g_2 - 3$$

viene chiamata *curtosi di eccesso* (*excess kurtosis*) ed è utilizzata per assegnare un valore pari a zero alla curtosi di una Normale. Al posto di g_2, spesso si preferisce utilizzare la misura detta *indice di curtosi di Fisher*:

$$G_2 = \frac{(n+1)(n-1)}{(n-2)(n-3)}\left(g_2 - \frac{3(n-1)}{n+1}\right)$$

Più il valore della curtosi è grande, più stretto e lungo sarà il picco della figura della densità della distribuzione, in corrispondenza del centro della distribuzione; viceversa, più è piccolo il valore, più largo e piatto sarà il picco rispetto al centro.

[8] Non è detto che valga l'inverso; infatti una distribuzione asimmetrica potrebbe avere un indice di asimmetria nulla.

2.6.4 Autocorrelazione

L'*autocorrelazione*, anche detta *correlazione seriale*, misura il grado di correlazione (cioè di dipendenza) tra le osservarzioni di uno stesso insieme, rispetto a una certa distanza (*lag*). Di solito viene utilizzata nel contesto delle serie temporali, in cui si analizzano le osservanzioni nell'ordine temporale con cui sono state raccolte; in tal caso, si misura la correlazione tra le osservazioni poste a una certa distanza temporale.

La verifica della presenza di autocorrelazione è importante in quanto permette di studiare il grado di indipendenza tra le osservazioni; l'eventuale mancanza di indipendenza, infatti, rende inefficace la maggior parte dei metodi tradizionali della statistica, a causa dell'utilizzo dell'assunzione di indipendenza fatta sui campioni, da essi effettuata.

La misura di autocorrelazione è espressa tramite la *funzione di autocorrelazione* (*ACF*):

$$r_h = \frac{c_h}{c_0} = \frac{\sum_{t=1}^{n-h} (x_t - \bar{x})(x_{t+h} - \bar{x})}{\sum_{t=1}^{n} (x_t - \bar{x})^2}, \quad h = 0, 1, \ldots, n-1$$

dove $x_1, \ldots, x_n$ è un campione di osservazioni, c_h è la *lag-h covarianza campionaria*:

$$c_h = \frac{1}{n} \sum_{t=1}^{n-h} (x_t - \bar{x})(x_{t+h} - \bar{x})$$

mentre c_0 è la *lag-0 covarianza campionaria* (che corrispodenza alla varianza campionaria $\hat{\sigma}^2$ non corretta Def. B.2.2). Il valore r_h è detto *lag-h autocorrelazione* ed è compreso tra -1 e 1.

Per verificare la presenza di autocorrelazione, o di qualche altra forma di dipendenza come la ciclicità, l'approccio EDA consiglia l'utilizzo di almeno due tipi di grafici:

- grafico *run sequence* (o *trend*), il quale non è altro che il grafico delle osservazioni nello stesso ordine con cui sono state raccolte: l'asse delle ascisse contiene i tempi di osservazione (*time*, indicati anche come *index* o *sequence*), mentre l'asse delle ordinate contiene le relative osservazioni;

questo grafico è utile per individuare l'eventuale presenza di forme di periodicità nelle osservazioni; inoltre, è possibile ricavare informazioni riguardo spostamenti di scala o di locazione, e sulla possibile presenza di "outlier";

- grafico *autocorrelazione*, è un grafico in cui le ascisse contengono diversi valori del *lag*, e le ordinate contengono il relativo valore dell'autocorrelazione ottenuto attraverso la ACF; questo grafico permette di verificare, oltre la presenza di autocorrelazione, anche la validità dell'ipotesi di indipendenza del campione: se l'autocorrelazione risulta prossima allo zero per ogni "lag", il campione può essere considerato casuale; viceversa, un campione non casuale sarà caratterizzato da una autocorrelazione significativamente diversa da zero per uno o più "lag".

La Fig. 2.3 mostra un esempio di grafico "run sequence" e relativo grafico "autocorrelazione", associati a un campione di 200 osservazioni, generato da una Weibull$(1.3, 8.7)$. Il grafico "run sequence" non mostra nessuna particolare periodicità; il grafico dell'autocorrelazione mostra che l'assunzione di indipendenza del campione può essere ritenuta valida al 95% di fiducia; l'intervallo di confidenza al 95% è individuabile dalle due linee orizzontali tratteggiate, le quali rappresentano i limiti superiore e inferiore dell'intervallo.

2.6.5 Dipendenza a Lungo Termine

La *Dipendenza a Lungo Termine* (*Long Range Dependence*, in breve *LRD*) rappresenta un tipo di autocorrelazione a distanza temporale (*lag*) piuttosto elevata: due osservazioni, di una stessa caratteristica, poste a distanza molto grande, sono tra loro dipendenti. Lo studio della LRD è importante in quanto, la sua presenza impedisce di effettuare particolari assunzioni e quindi previene l'utilizzo metodi statistici errati. Per esempio, la presenza di LRD implica che i campioni non possono essere considerati indipendenti; inoltre, nel contesto della simulazione di un certo sistema, se una caratteristica del sistema esibisce LRD, l'analisi statistica del comportamento del sistema non può essere effettuata ipotizzando una condizione di equilibrio operazionale, in quanto, a cau-

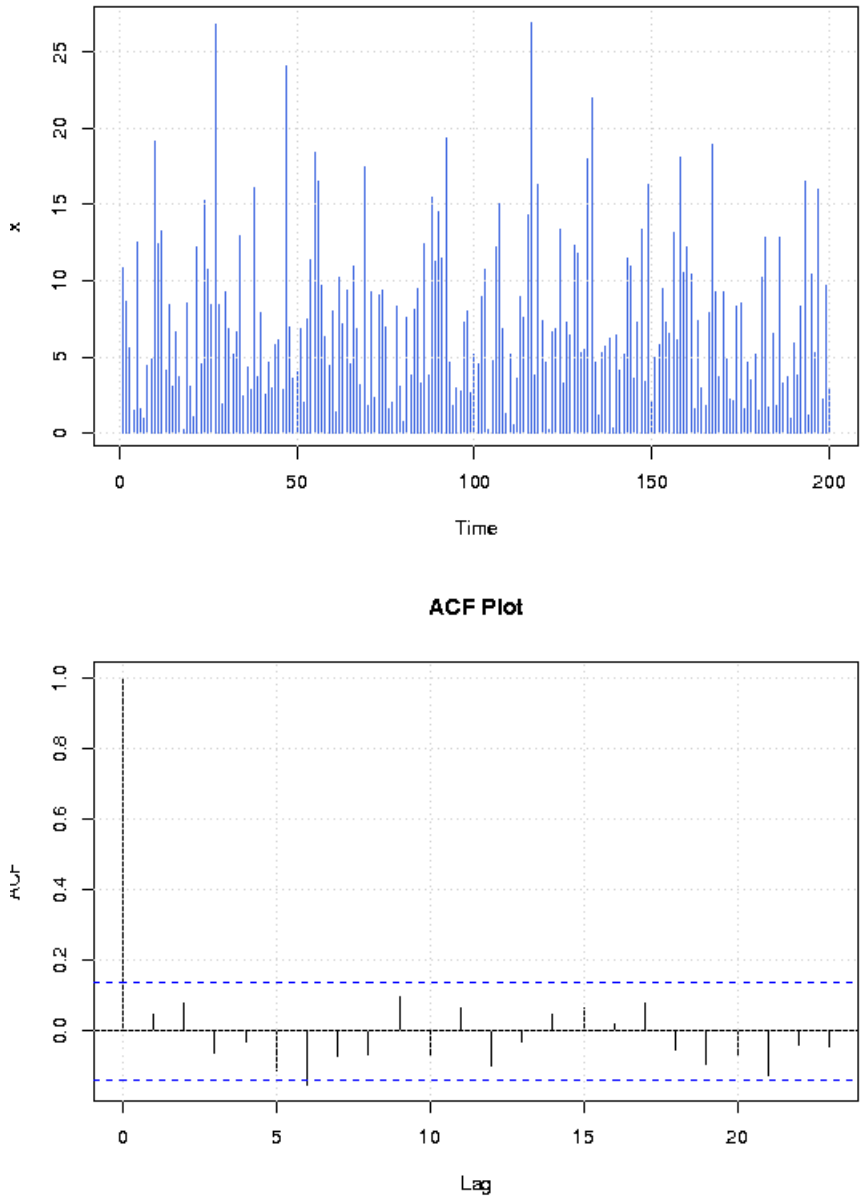


Figura 2.3: Grafico EDA per rilevare la presenza di autocorrelazione.

sa della dipendenza su intervalli di tempi molto lunghi, il funzionamento del sistema non potrà mai essere considerato a regime.

La presenza di dipendenza a lungo termine viene di solito verificata tramite la stima dell'*esponente di Hurst* H o, in modo equivalente, della *Dimensione Frattale* d.

La *dimensione frattale* d è una quantità statistica, compresa tra -0.5 e 0.5, che indica come un frattale occupa lo spazio intorno a sé, mano a mano che si effettuano ingrandimenti ("zoom-in") o rimpicciolimenti ("zoom-out") del frattale stesso. A differenza della classica dimensione euclidea, la dimensione frattale può essere un valore non intero; ciò significa, semplicemente, che il frattale non occupa tutto lo spazio che lo circonda. Esistono diverse definizioni per questa quantità; la maggior parte sfrutta la proprietà di *Auto-Similarità* (*Self-Similarity*) intrinseca nella definizione di frattale: ingrandendo o rimpicciolendo una figura frattale, si ritrova la stessa struttura; questà proprietà è a volte chiamate *Invarianza di Scala* (*Scale Invariance* o *Scale Free*).

L'*esponente di Hurst* (anche detto *indice di dipendenza*) è una misura di scala, compresa tra 0 e 1, che indica quanto una serie temporale si distribuisce intorno alla tendenza media o, al contrario, si concentra verso una particolare direzione. Più precisamente, data una serie temporale X_i, di lungheza N, e definita la serie aggregata:

$$X^{(m)}(k) = \frac{1}{m} \sum_{i=(k-1)m+1}^{km} X_i, \quad k = 1, 2, \ldots, \lceil N/m \rceil \tag{2.6.1}$$

si ha che, per N/m e m sufficientemente grandi, $X^{(m)}$ è *asintoticamente autosimilare* a $m^{H-1}S$, cioè:

$$X^{(m)} \xrightarrow{\mathcal{D}} m^{H-1}S \tag{2.6.2}$$

dove S è un processo che dipende dalla distribuzione di X (ma non da m). Nel caso di distribuzioni con varianza finita, S è il processo *Fractional Gaussian Noise* (*FGN*) e la relazione (2.6.2) diventa di uguaglianza (in distribuzione); nel caso di varianza infinita, S è il processo *Linear Fractional Stable Noise* (*LFSN*) (ad es., si veda [108]).

Le quantità d e H sono intrinsecamente legate dalla relazione:

$$H = \begin{cases} d + 1/2, & \text{se S ha varianza finita} \\ d + 1/\alpha, & \text{se S ha varianza infinita} \end{cases} \tag{2.6.3}$$

dove α, con $0 < \alpha \leq 2$, è il *tail-index* di una distribuzione α-*Stable* o di una *Heavy-Tailed* [9] (Cap. 6).

Per quanto riguarda la LRD, d ne rappresenta una misura dell'intensità, mentre H ne è una misura di scala. Si può dimostrare che la presenza di LRD in un processo è verificabile controllando se $d > 0$ o, in modo equivalente, se $H > 1/2$. Un valore di H pari a 0.5 indica assenza di dipendenza a lungo termine; un valore di H inferiore a 0.5 corrisponde a dipendenza negativa. Esistono vari metodi per stimare H, quelli maggiormente utilizzati includono [107]:

- *Aggregation of Variance*
- *Rescaled Range* (*R/S*)
- *Periodogram*

Tutti i metodi citati sono metodi grafici in cui l'esponente di Hurst viene stimato attraverso una regressione effettuata su una *reference line*, specifica per ogni metodo. Si noti che ci sono casi in cui questi metodi possono fornire una stima dell'esponente di Hurst superiore a 1; questo può essere una conseguenza della presenza di particolari "pattern" nell'insieme dei dati (ad es., dei "trend"); per maggiori informazioni si veda [63].

È inoltre possibile verificare la presenza della proprietà di *self-similarity* graficamente, tracciando opportuni grafici della distribuzioine empirica su diversi livelli di scala (tipicamente, su una granularità più fine) e verificando la presenza di "pattern" ripetitivi, approssimativamente, a ogni livello di scala. I metodi più diffusi includono:

- *Aggregated Run Sequence Plot*

[9] Si noti che una distribuzione può essere *heavy-tailed* ma non α-*Stable*; un esempio tipico è una distribuzione Pareto con parametro *shape* inferiore a 2.

- *Aggregated log-log CDF Plot*
- *Aggregated log-log CCDF Plot*

2.6.6 Code della Distribuzione

Quando una distribuzione ha una coda che decade secondo una legge esponenziale, come la distribuzione Esponenziale o la Normale, le osservazioni estreme, cioè quelle appartenenti alla coda, sono così rare da poter essere ignorate: la probabilità che esse compaiano in un campione tende a zero esponenzialmente; per tali ragioni, questo tipo di distribuzioni viene spesso utilizzato per modellare dati con supporto limitato (*bounded support*). Le distribuzioni in cui la coda è governata da una legge esponenziale sono chiamate *short tailed*.

Ci sono, tuttavia, distribuzioni la cui coda cade più lentamente di una legge esponenziale; esse sono dette *long-tailed* (o *fat-tailed*); fra queste distribuzioni si distinguono in modo particolare le cosidette distribuzioni *power-law*, le cui code cadono secondo una legge polinomiale. Esistono, poi, altre distribuzioni le cui code non seguono una legge polinomiale ma il decadimento avviene, comunque, in modo sub-esponenziale (ad es., Weibull, con parametro "scale" minore di 1, e Log-Normale).

Molte delle distribuzioni sub-esponenziali, comprese alcune della famiglia "power-law", pur avendo un decadimento delle code più lento di una Normale, si comportano asintoticamente come una Gaussiana. Tuttavia, nella classe di distribuzioni "power-law", ve n'è una di particolare interesse, nota come classe delle distribuzioni *heavy-tailed*, per cui, anche asintoticamente, il comportamento si discosta da una Normale. I motivi per cui le distribuzioni "heavy-tailed" destano molta attenzione, riguardano le proprietà che hanno associate:

- *Momenti Infiniti*: alcuni momenti della distribuzione non esistono.
- *Expectation Paradox*: la vita media residua tende a crescere al trascorrere del tempo.

- *Teorema del Limite Centrale classico non valido*: a causa della presenza di momenti infiniti, il CLT non è applicabile.
- *Mass-Count Disparity*: le osservazioni più frequenti della popolazione non sono le stesse che contribuiscono a determinarne la massa.

Il Cap. 6 è interamente dedicato alle distribuzioni "heavy-tailed"; esso ne descrive le principali caratteristiche e proprietà e presenta alcuni metodi per verificarne la presenza e stimarne i parametri, a partire da un insieme di osservazioni. Di solito la verifica della presenza di distribuzioni "heavy-tailed" viene effettuata combinando test grafici a metodi numerici; quelli utilizzati nel presente progetto, includono:

- Grafico dello *stimatore di Hill* §6.3.1.
- Grafico della *Curva di Lorenz* §6.3.3.
- Grafico della *Mass-Count disparity* §6.3.4.
- Grafico *Log-Log CCDF* in versione semplice e aggregata §6.3.2.

2.7 Scelta di un Modello

La scelta di un modello nel caso univariato consiste nella scelta di una distribuzione di probabilità, che meglio si adatti ai dati osservati; la scelta di una distribuzione dovrebbe tenere conto sia delle misure di centralità e dispersione, sia di aspetti più legati alla forma della densità, come il grado di asimmetria e il tipo di legge che governa le code di una distribuzione. Dato che si tratta di un argomento molto vasto, esso verrà trattato in un specifico capitolo (si veda Cap. 3).

2.8 Verifica di un Modello

Dopo che è stato scelto un modello per descrivere i dati osservati, occorre verificare la bontà dell'adattamento; per il caso univariato esistono diversi test sia grafici sia numerici; alcuni di questi test verranno discussi nel capitolo Cap. 3.

2.9 Approccio all'Analisi delle Tracce

L'approccio utilizzato nel presente progetto utilizza una combinazione di EDA e di CDA; questa filosofia di affrontare la costruzione di modelli statistici è spesso utilizzata nella statistica moderna, in quanto permette di sfruttare i vantaggi di entrambi gli approcci, ottenendo alla fine un modello meno sensibile a eventuali ipotesi effettuate a priori. L'approccio EDA viene utilizzato per ricercare delle relazioni o dei "pattern"; quello CDA è usato per verificare l'effettiva significatività dei risultati ottenuti dall'approccio EDA; oppure, l'approccio EDA viene utilizzato per verificare che delle ipotesi, effettuate durante l'analisi condotta tramite l'approccio CDA, abbiano un riscontro anche dal punto di vista descrittivo (ad es., grafico). Un esempio tipico di questo modo di procedere è l'analisi di una regressione lineare, in cui le ipotesi di relazione lineare tra le caratteristiche vengono validate attraverso un approccio EDA, come l'analisi dei residui [38].

Di seguito vengono brevemente descritti, in maniera generale, i vari passi effettuati durante l'analisi del carico; maggiori dettagli verranno forniti nella parte del documento relativi all'analisi delle varie tracce (si veda Parte II).

1. *Scelta dell'Obiettivo*. L'obiettivo è l'acquisizione di una maggiore conoscenza riguardo il tipo di proprietà associate alle caratteristiche del carico di alcuni sistemi Grid; in particolare, l'attenzione è rivolta allo studio delle caratteristiche del carico che possano avere un effetto sulle politiche di "scheduling".

2. *Scelte delle Caratteristiche del Carico*. Nel progetto, come caratteristiche rappresentative, si è scelto il tempo di interarrivo dei job e la durata della relativa esecuzione; per quanto concerne le caratteristiche associate ai BoT, non sono state prese in considerazione in quanto le tracce analizzate non contenevano questo tipo di informazioni e non era nemmeno possibile ricavarle in funzione di altre.

3. *Raccolta dei Dati*. Nel contesto del Grid Computing la disponibilità di tracce è al momento molto limitata; ciò è causato, principalmente, dal

fatto che la caratterizzazione del carico di sistemi Grid rappresenta una materia di studio abbastanza recente. Le tracce analizzate nel presente lavoro, sono state recuperate dal sito web *Parallel Workload Archive* [39].

4. *Bonifica dei Dati*. Per ogni traccia analizzata sono state costruite delle procedure "ad-hoc" di individuazione ed eventuale correzione o eliminazione dei dati anomali; maggiori dettagli su questo argomento saranno approfonditi nei capitoli specifci ad ogni traccia (si veda Parte II).

   Per quanto riguarda i metodi grafici della statistica descrittiva, sono stati tracciati i quattro grafici consigliati dall'approccio EDA per la rilevazione degli "outlier": istogramma delle frequenza, "density plot", "box plot" e "Normal Q-Q plot". Per il disegno dell'istogramma delle frequenze, la tecnica utilizzata per dividere l'intervallo delle osservazioni in "bin" è la formula di "Sturges", la quale è quella fornita di "default" dal linguaggio *R*. Il tipo di kernel scelto per il grafico della densità è quello Gaussiano, con una "bandwidth" pari al doppio dell'intervallo inter-quartile; il valore della "bandwidth" deriva dalla regola di "Freedman-Diaconis" §2.5.2 e tenta di fornire una stima robusta della dispersione dei dati; insieme al grafico della densità viene anche visualizzato l'istogramma delle frequenze relative, in modo da poter confrontare più chiaramente la stima della densità tramite la funzione "kernel" con quella data dall'istogramma. Il grafico "box plot" usa, come calcolo dei quartili, la variante suggerita da Tukey [110], mentre, come costante k dell'intervallo inter-quartile, il valore 1.5, il quale deriva dalla normalità asintotica della mediana e rappresenta anche il valore di "default" usato dal linguaggio *R*. Nel grafico "Normal Q-Q", come "reference line", anzichè tracciare una retta di 45 gradi passante per l'origine, si disegna una retta passante per il primo e terzo quartile; in questo modo è possibile anche individuare relazioni di normalità che differiscono solo per spostamenti di scala ("reference line" con pendenza diversa da uno) e di locazione ("reference line" con punto di intersezione con l'asse delle ordinate diverso da zero); per maggiori dettagli si veda la sezione relativa al grafico "Q-Q" §3.2.1.

Sono stati anche utilizzati dei metodi numerici della statistica descrittiva per l'analisi della centralità, della dispersione e della correlazione; questi metodi sono serviti anche all'analisi delle proprietà statistiche dei dati.

Si è deciso di non utilizzare nessun metodo numerico della statistica inferenziale, a causa delle assunzioni che richiedono per la loro applicazione; in particolare, il test di Grubbs non è stato utilizzato in quanto ipotizza la presenza di normalità nelle osservazioni.

5. *Analisi delle Proprietà Statistiche*. L'analisi delle proprietà statistiche delle varie caratteristiche fa uso della maggior parte delle misure descritte in §2.6. In particolare, uno degli interessi principali è lo studio della leggi che descrivono le code della distribuzione dei dati; infatti, la possibile presenza di code "lunghe" (cioè, aventi una legge polinomiale) influisce notevolmente sulla scelta del modello; per esempio, la presenza di code "heavy" rende non valido l'utilizzo del Teorema del Limite Centrale classico B.1.1 e, quindi, l'assunzione di comportamento normale al crescere della dimensione del campione. La presenza di code "lunghe" insieme alla dipendenza a lungo termine, rende inoltre inadeguato l'utilizzo delle Catene di Markov (Cap. E) per la costruzione di modelli generativi, in quanto, in quest'ultime, la coda avrebbe un decadimento esponenziale mentre la dipendenza tra le osservazioni di una stessa caratteristica è limitata dalla proprietà di assenza di memoria.

6. *Scelta di un Modello*. L'analisi effettuata nel presente progetto è stata limitata alle singole osservazioni, cioè all'analisi statistica univariata; la scelta del modello è stata quindi ristretta alla ricerca delle distribuzioni di probabilità che riescono a descrivere la caratteristica sotto osservazione. Sono state prese in considerazione molte distribuzioni, fra cui: Gamma, Log-Normale, Normale, Pareto Generalizzata, Phase-Type, Valori Estremi Generalizzata, Weibull, ... Altre distribuzioni, che all'inizio si aveva deciso di includere nell'analisi, sono state escluse in quanto rappresentano solamente un caso specifico di una più larga famiglia di distribuzioni; per esempio, la distribuzione Iper-Esponenziale è stata esclu-

sa in favore dell'utilizzo della distribuzione Phase-Type, una famiglia di distribuzioni di cui la Iper-Esponenziale fa parte (Cap. 5).

L'adattamento di una distribuzione ai dati, è stato mediante la stima dei parametri della distribuzione a partire dall'intero insieme di dati; questo modo di procedere ha il vantaggio di riuscire a catturare tutte le proprietà della caratteristica che i dati a disposizione permettono di rilevare e di permettere più facilmente l'utilizzo di quelle statistiche che assumono che l'insieme dei dati sia costituito da un numero minimo di osservazioni, Purtroppo vi è lo svantaggio che la verifica del modello viene resa più complicata in quanto i parametri non possono essere considerati indipendenti dalle osservazioni e quindi molti test statistici sull'adattamento non possono essere applicati o devono essere opportunamente modificati §3.2.6. Una possibile alternativa per la stima dei parametri, potrebbe essere l'utilizzo della tecnica *k-fold cross validation* [116], per cui l'insieme dei dati viene divisio in k parti (di solito di uguale dimensione) in modo casuale: a turno, una parte viene usata per effettuare il "fitting" della distribuzione, mentre le rimanenti sono utilizzate per effettuare la verifica del modello; il problema di questo approccio è la possibile "rottura" della struttura dei dati e dei legami fra le osservazioni; per esempio, se la divisione creasse un insieme di dati rappresentante il corpo della distribuzione originale e un altro rappresentante la coda più lunga, molto probabilmente la distribuzione ricavata dall'insieme utilizzato per il "fitting" non si adatterebbe all'insieme usato per la verifica (e viceversa) [10]. Un'altra possibile alternativa è l'utilizzo del metodo del *bootstrap non parametrico* [35], tramite il quale si effettuano una serie di ricampionamenti (di solito, almeno 1000) dalla distribuzione empirica dei dati; ogni ricampionamento genera un campione la cui numerosità, generalmente, coincide con quella del campione originale; per ogni campione si calcola la stima dei parametri della distribuzione; la stima finale dei parametri può quindi essere ricavata come una media effettuata sulle singole stime.

[10] Si noti che non è possibile utilizzare la *stratificazione*, come di solito accade in *Machine Learning*, in quanto i valori rappresentativi di una caratteristica non sono conosciuti.

Anche in questo caso, vi è lo svantaggio che i ricampionamenti possono "spezzare" le relazioni tra le osservazioni; tuttavia, questo problema risulta meno accentuato nel "bootstrap" grazie all'utilizzo del ricampionamento dalla distribuzione empirica, anzichè solo dall'insieme dei dati, come accade con il metodo del "cross-validation".

7. *Verifica di un Modello*. Nella fase di verifica del modello sono stati utilizzati tutti i test di adattamento descritti in §3.2; come spiegato nel precedente punto, la stima dei parametri di una distribuzione dall'intero insieme dei dati, ha portato alla revisione delle tecniche numeriche per la verifica dell'adattamento; in particolare, nel test del χ^2 §3.2.3 sono stati sottratti un numero di gradi di libertà per ogni parametro stimato, mentre nel test di *Kolmogorov-Smirnov* §3.2.4 e di *Anderson-Darling* §3.2.5 è stata utilizzata la tecnica del *bootstrap parametrico*. I test grafici, invece, sono stati usati nella versione classica. Per maggiori informazioni, si veda §3.2.

L'analisi statistica appena descritta, è stata effettuata attraverso l'implementazione di alcuni programmi, la maggior parte dei quali scritti nel linguaggio per la statistica *R* [45]; per maggiori dettagli si veda Cap. A.

Capitolo 3

Fitting di Distribuzioni

Uno dei problemi comuni in cui ci si può imbattere durante l'analisi dei dati, è il cosidetto *fitting di distribuzioni* e riguarda l'inferenza sulla possibile distribuzione teorica che possa aver generato i dati sotto osservazione. Il problema può essere affrontato sostanzialmente in due modi:

- "adattare" una distribuzione di probabilità teorica alla distribuzione dei dati osservati;
- studiare direttamente le proprietà della distribuzione dei dati senza ipotizzare l'appartenenza a una specifica distribuzione teorica.

Il primo metodo va sotto il nome di *metodo parametrico*; consiste nel considerare una o più famiglie parametriche di distribuzioni teoriche, stimare eventualmente i relativi parametri, effettuare una serie di test statistici per valutare la bontà di adattamento e quindi scegliere la distribuzione teorica che meglio si adatti ai dati. Un vantaggio di questo metodo sta nel fatto che rende possibile la costruzione di modelli generativi completamente caratterizzati, dove le proprietà delle leggi di probabilità sono totalmente note; fra gli svantaggi, vi è la possibile difficoltà nello scegliere quali distribuzioni considerare e l'eventuale complessità dei metodi numerici associati; inoltre potrebbe capitare che i dati o le statistiche analizzate non seguano una distribuzione "standard". Il secondo metodo prende il nome di *metodo non parametrico*; in questo caso ciò che si studia è la distribuzione dei dati stessi (come la mediana o altre statistiche d'ordi-

ne §B.2.2); come per il metodo parametrico, la bontà delle inferenze effettuate può essere studiata attraverso opportuni test. Uno dei vantaggi è la flessibilità e la facilità di utilizzo; infatti, i dati non devono essere necessariamente quantitativi ma possono essere anche categorici, e le procedure coinvolte sono molto semplici e veloci da eseguire; inoltre, non facendo nessuna ipotesi riguardo i meccanismi che hanno generato i dati né sulla forma della distribuzione sottostante, questi metodi sono appropriati quando i dati o le statistiche non seguono una distribuzione "standard". Uno svantaggio è la minor potenza rispetto a un test parametrico; ciò suggerisce che, quando possibile, è preferibile utilizzare un test parametrico.

Questo capitolo fornisce una panoramica sui principali metodi utilizzati nell'ambito del "fitting" di distribuzioni; dato che nel presente progetto sono stati impiegati solo metodi parametrici, questo capitolo tratterà solo tale classe di metodi, tralasciando i metodi non parametrici. La prima sezione §3.1 presenta le principali tecniche che si possono utilizzare per effettuare la stima dei parametri di una distribuzione teorica, mentre la sezione §3.2 descrive i test di verifica sulla bontà dell'adattamento di una distribuzione teorica rispetto ai dati osservati, limitandosi a illustrare solo quelli utilizzati nel progetto.

3.1 Stima dei Parametri

I *Metodi Parametrici* sono una classe di metodi utilizzati per ricavare una stima dei parametri di una distribuzione teorica, a partire da un insieme di dati.

L'adattamento di distribuzioni teoriche ai dati osservati può essere effettuato in uno dei seguenti modi:

- adattamento di una distribuzione teorica completamente nota;
- adattamento di una distribuzione teorica parzialmente specificata;
- adattamento di una distribuzione teorica con parametri completamente sconosciuti.

Il primo punto rappresenta il caso più semplice; si tratta, tuttavia, di un caso che difficilmente si trova applicato nella realtà, se non, magari, in contesti

puramente didattici o quando vi siano forti basi teoriche o pratiche che ne supportino l'utilizzo. Gli ultimi due punti, invece, riguardano i casi che si incontrano più comunemente, in cui l'adattamento a un insieme di dati viene effettuato rispetto a una famiglia parametrica di distribuzioni; in tal caso, occorre cercare di stimare i parametri ignoti, che caratterizzano la distribuzione, a partire da un insieme di dati.

La stima dei parametri di una distribuzione prevede che da un campione casuale si calcolino una o più statistiche rappresentative dei parametri di interesse (Def. B.2.3); di solito, oltre alle statistiche, vengono ricavate altre misure, come la variabilità, che permettono di descrivere con miglior accuratezza la stima del parametro calcolata. Vi sono due modi per fornire una stima: *stima puntuale* e *stima intervallare*.

Definizione 3.1.1 (Stimatore Puntuale)**.** Sia X una variabile casuale con distribuzione di probabilità $f(x)$, θ uno dei parametri sconosciuti che caratterizzano la distribuzione di X, e $X_1, \ldots, X_n$ un campione casuale di numerosità n estratto da X. La statistica $\hat{\Theta}_\theta = h(X_1, \ldots, X_n)$ è detta *stimatore puntuale* (o, semplicemente, *stimatore*) di θ. Dopo che il campione è stato selezionato (ad es., $X_1 = x_1, \ldots, X_n = x_n$), il valore $\hat{\theta}$, assunto dalla statistica $\hat{\Theta}_\theta$, prende il nome di *stima puntuale* (o, semplicemente, *stima*) di θ.

L'efficacia di uno stimatore puntuale si può misurare in termini di *Errore Quadratico Medio* (*MSE*):

$$\mathrm{MSE}\left(\hat{\Theta}_\theta\right) = \mathrm{E}\left[\left(\hat{\Theta}_\theta - \theta\right)^2\right] \tag{3.1.1}$$

In generale il MSE di uno stimatore è differente dalla sua *Varianza*. La *Varianza* di uno stimatore $\hat{\Theta}_\theta$ è definita come la deviazione quadratica media dal valor medio del parametro stimato:

$$\mathrm{Var}\left(\hat{\Theta}_\theta\right) = \mathrm{E}\left[\left(\hat{\Theta}_\theta - \mathrm{E}\left[\theta\right]\right)^2\right] \tag{3.1.2}$$

Le due quantità sono comunque in relazione fra di loro e, sotto particolari condizioni (illustrate nel prossimo paragrafo), coincidono.

Le proprietà sicuramente più importanti di uno stimatore puntuale sono:

- *Correttezza*: uno stimatore $\hat{\Theta}_\theta$ si dice *corretto*, o *non distorto* (in inglese, *unbiased*), se il suo valore atteso coincide con il parametro che deve stimare:
$$\mathrm{E}\left[\hat{\Theta}_\theta\right] - \theta = 0 \tag{3.1.3}$$
In tal caso, la varianza dello stimatore (Eq. (3.1.2)) coincide con l'errore quadratico medio (Eq. (3.1.1)). Al contrario, uno stimatore è *distorto* (*biased*) se la quantità (3.1.3):
$$b\left(\hat{\Theta}_\theta\right) = \mathrm{E}\left[\hat{\Theta}_\theta\right] - \theta \tag{3.1.4}$$
è diversa da zero; il valore $b\left(\hat{\Theta}_\theta\right)$ è detto *bias* di $\hat{\Theta}_\theta$ come stimatore di θ. In questo caso, la varianza dello stimatore (Eq. (3.1.2)) e l'errore quadratico medio (Eq. (3.1.1)) sono legati dalla relazione:
$$\mathrm{MSE}\left(\hat{\Theta}_\theta\right) = \mathrm{Var}\left(\hat{\Theta}_\theta\right) + b^2\left(\hat{\Theta}_\theta\right) \tag{3.1.5}$$

- *Efficienza*. Dati k stimatori $\hat{\Theta}_1, \ldots, \hat{\Theta}_k$ di uno stesso parametro θ, lo stimatore più *efficiente* è quello con la minima varianza e prende il nome di *stimatore MVUE* (*Minimum Variance Unbiased Estimator*).

Una stima puntuale è, in generale, poco informativa; per esempio, una volta che è stata determinata da un particolare campione, nulla si conosce sulla variabilità o sull'errore che si commette nell'utilizzarla al posto del parametro ignoto; queste informazioni, oltre a dipendere dal metodo con cui si determinano le stime, sono funzione del campione su cui la stima viene calcolata. Lo scopo di una *stima intervallare* è quello di arricchire una stima puntuale fornendo informazioni sulla probabilità che il parametro ignoto assuma un valore compreso in un intervallo costruito intorno alla stima stessa.

Definizione 3.1.2 (Stimatore Intervallare)**.** Uno *stimatore intervallare* di un parametro ignoto θ rappresenta un intervallo aleatorio di possibili (probabili) valori

che θ può assumere. Dato che l'intervallo è ricavato in funzione dei campioni, esso rappresenta una variabile aleatoria. Fissato il campione, lo specifico valore di questo intervallo è detto *stima intervallare*.

Siccome l'intervallo viene determinato solo in funzione del campione analizzato (e non dall'intera popolazione), non si può essere sicuri che contenga il vero valore del parametro ignoto; è possibile, tuttavia, costruire questi intervalli in modo tale da avere una certa "confidenza" che il vero valore del parametro vi sia effettivamente contenuto. Questo tipo di intervalli prende il nome di *intervallo di confidenza*.

Definizione 3.1.3 (Intervallo di Confidenza (CI))**.** Dato un campione $X_1, \ldots, X_n$, un *Intervallo di Confidenza* (*CI*) per un parametro θ è un intervallo della forma $l \leq \theta \leq u$, in cui gli estremi l e u sono calcolati dal campione in esame. Dato che, in generale, a campioni differenti corrisponderanno valori di u e l diversi, questi estremi possono essere visti come valori delle variabili casuali L e U, rispettivamente. Si definisce *Intervallo di Confidenza a un livello* $(1-\alpha)$, o *Intervallo di Confidenza al* $100(1-\alpha)\%$, l'intervallo i cui estremi soddisfano la seguente probabilità:

$$\Pr\{L \leq \theta \leq U\} = 1 - \alpha \tag{3.1.6}$$

dove $1-\alpha$ è detto *livello di confidenza*, con $0 \leq \alpha \leq 1$; l'espressione (3.1.6) afferma che c'è una probabilità pari a $1-\alpha$ di selezionare un campione il cui CI contenga il vero valore di θ. Le quantità l e u sono dette, rispettivamente, *estremo inferiore* e *superiore* dell'intervallo.

Osservazione. Si noti che non è corretto affermare che, a partire da un campione $X_1, \ldots, X_n$, il parametro θ è contenuto nell'intervallo $L \leq \theta \leq U$ con una probabilità pari a $1-\alpha$; infatti, dato che nell'Eq. (3.1.6) gli estremi U e L sono variabili casuali, il CI è un *intervallo casuale*, e quindi la corretta interpretazione è che, fissato α, su infiniti campionamenti casuali effettuati dalla popolazione, associata al parametro ignoto θ, se si calcolano gli intervalli di confidenza al $100(1-\alpha)\%$, $100(1-\alpha)\%$ di questi intervalli conterrà il valore vero del parametro θ.

Le definizioni date sinora non forniscono nessun modo per ottenere uno stimatore per un dato parametro ignoto; di seguito, si presentano due metodi "classici" per ricavare degli stimatori puntuali: il *Metodo dei Momenti* e il *Metodo di Massima Verosimiglianza*. Dei due, il metodo di massima verosimiglianza è di solito preferibile al metodo dei momenti in quanto fornisce, generalmente, una miglior efficienza; tuttavia il metodo dei momenti è computazionalmente meno intenso del metodo di massima verosimiglianza e può essere applicato in più contesti. Nei capitoli successivi, verranno mostrati altri metodi, per ricavare stimatori puntuali, costruiti ad-hoc per particolari tipi di parametri e distribuzioni. La sezione si chiude con alcuni esempi di stimatori puntuali, insieme ai relativi stimatori intervallari.

3.1.1 Metodo dei Momenti (MOM)

Lo scopo del *Metodo dei Momenti* è quello di fornire una stima dei *momenti della popolazione* (Def. B.1.4) in funzione dei *momenti campionari* (Def. B.2.1). Dato che i momenti della popolazione sono chiaramente funzione dei parametri θ, incogniti, della stessa popolazione, una volta trovate la stime dei momenti è possibile risolvere le equazioni che legano i momenti della popolazione ai parametri θ, utilizzando, al posto dei momenti reali, le stime calcolate dal campione.

Ricordando che il momento campionario k-esimo è definito come la media aritmetica dei campioni elevati alla k (Def. B.2.1), segue la definizione di metodo e stimatore dei momenti.

Definizione 3.1.4 (Metodo dei Momenti (MOM))**.** Sia $X_1, \ldots, X_n$ un campione casuale proveniente da una popolazione con m parametri incogniti $\theta_1, \ldots, \theta_m$. Gli *Stimatori dei Momenti* $\hat{\Theta}_1 \ldots, \hat{\Theta}_m$ si calcolano eguagliando i primi m momenti della popolazione con i primi m momenti campionari e risolvendo le equazioni risultanti rispetto ai parametri incogniti.

In generale è preferibile utilizzare altri tipi di stimatori più efficienti e robusti. I momenti, specialmente quelli di ordine elevato, sono molto influenzabili dai valori assunti dai dati e quindi possono essere poco rappresentativi della

popolazione; ad es., se il campione sotto osservazione contiene molti valori di ordini di grandezza "piccolo" e qualche valore con ordine di grandezza "più grande" rispetto ai primi, il valore dei momenti campionari sarà principalmente determinato da questi ultimi valori. Un esempio pratico ne chiarisce immediatamente il concetto. Si supponga che da una popolazione distribuita secondo la variabile aleatoria discreta X:

X	1	2	3	4	25	126
$p(X)$	$\frac{1}{3}$	$\frac{5}{24}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{12}$	$\frac{1}{24}$

si estragga il seguente campione di numerosità pari a 10:

$$1, 1, 126, 2, 2, 1, 1, 25, 1, 2$$

La media reale della popolazione è pari a $222/24 = 9.25$ mentre la media campionaria, relativamente al campione estratto, vale $162/10 = 16.2$, un valore che è quasi il doppio del valor medio reale. Anche nel caso in cui il campione estratto fosse stato:

$$1, 1, 1, 1, 126, 1, 1, 1, 1, 1$$

cioè dove appare solo un valore di ordine di grandezza maggiore, la media campionaria, uguale a $135/10 = 13.5$, sarebbe stata piuttosto lontana dal valore vero; si noti che, in questi due casi illustrati, il valore 126, cioè quello di ordine di grandezza maggiore, dà un contributo di circa il 57% e il 93%, rispettivamente, al calcolo del valore medio campionario. Il fenomeno sarebbe amplificato se si analizzassero i momenti di ordine più elevato, in quanto ogni valore viene elevato all'ordine del momento; per esempio, la varianza reale risulterebbe pari a 633.3 mentre quella campionaria (corretta) varrebbe 1543.7, per il primo campione, e 1562.5, per il secondo campione. Si noti, inoltre, che questi valori con un ordine di grandezza "elevato" non è detto che siano necessariamente degli "outlier" (nel qual caso potrebbero applicarsi opportune tecniche di rimozione); potrebbero essere invece dei valori caratterizzanti la popolazione sottostante, i quali, pur avendo associata una piccola probabilità di verificarsi, non possono essere considerati così "rari" che difficilmente appaiano in un campione (questo è il caso di distribuzioni di probabilità le cui

code sono governate da leggi polinomiali). Un altro motivo per evitare l'uso del metodo dei momenti, per ricavare la stima dei parametri ignoti, è legato al fatto che alcune distribuzioni possiedono dei momenti infiniti (ad es., le cosidette distribuzioni *Heavy-Tail*); in questo caso la stima del parametro incognito non sarebbe sicuramente rappresentativa della popolazione.

Vi sono tuttavia delle ragioni per cui conviene, a volte, utilizzare questo metodo. Innanzi tutto, si tratta di un metodo estremamente semplice (spesso utilizzabile senza l'ausilio di un calcolatore) e che richiede poche risorse computazionali; inoltre, risulta essere una buona tecnica per ricavare dei valori iniziali da utilizzare durante il passo di inizializzazione di algoritmi iterativi usati per calcolare altri stimatori più efficienti (come lo stimatore di massima verosimiglianza). Infine, vi sono casi in cui non è possibile determinare il valore di stimatori più efficienti a causa di irregolarità presenti nella funzione di probabilità della distribuzione.

3.1.2 Metodo di Massima Verosimiglianza (MLE)

Il *Metodo di Massima Verosimiglianza*, ideato da Sir. A. R. Fisher tra il 1912 e il 1922, è uno dei metodi più utilizzati per ricavare delle stime di parametri ignoti; alla base del metodo vi è il concetto di *Funzione di Verosimiglianza*.

Definizione 3.1.5 (Stimatore di Massima Verosimiglianza (MLE))**.** Sia $X_1 \ldots, X_n$ un campione casuale proveniente da una popolazione con parametro incognito θ e con funzione di probabilità $f(\cdot;\theta)$ [1]. Fissato il campione di osservazioni $X_1 = x_1, \ldots, X_n = x_n$, si denoti con $f(x_1 \ldots, x_n;\theta)$ la funzione di probabilità congiunta di $x_1, \ldots, x_n$ in funzione di θ. Si definisce *Funzione di Verosimiglianza* (*Likelihood Function*) la funzione in θ:

$$\mathcal{L}(\theta; x_1, \ldots, x_n) = f(x_1, \ldots, x_n;\theta) \tag{3.1.7}$$

I valori $x_1, \ldots, x_n$ del campione sono delle costanti, mentre il parametro θ è una variabile.

[1] La notazione $f(\cdot;\theta)$ indica una funzione di probabilità, pmf o pdf, dipendente dal parametro θ.

Lo *Stimatore di Massima Verosimiglianza* (*MLE*) di θ è quel valore $\hat{\theta}$ che, fra tutti i possibili valori per θ, massimizza la funzione di verosimiglianza $\mathcal{L}(\theta; x_1, \ldots, x_n)$:

$$\hat{\theta} = \arg\max_{\theta} \mathcal{L}(\theta; x_1, \ldots, x_n) \tag{3.1.8}$$

La definizione appena mostrata è di difficile applicazione pratica in quanto presuppone la conoscenza della distribuzione congiunta e potrebbe dar luogo a una formula piuttosto complessa da risolvere; di seguito si presentano due regole pratiche largamente utilizzate:

1. nel caso in cui non si conosca la distribuzione congiunta, si può supporre che il campione $X_1, \ldots, X_n$ sia i.i.d.; in tal caso è possibile sfruttare la proprietà degli eventi indipendenti nella definizione di funzione di verosimiglianza e quindi calcolare lo stimatore di massima verosimiglianza $\hat{\theta}$ come:

$$\begin{aligned} \hat{\theta} &= \arg\max_{\theta} \mathcal{L}(\theta; x_1, \ldots, x_n) \\ &= \arg\max_{\theta} f(x_1, \ldots, x_n; \theta) \\ &= \arg\max_{\theta} \prod_{i=1}^{n} f(x_1; \theta) \quad \text{(per definizione di eventi indipendenti)} \end{aligned} \tag{3.1.9}$$

2. spesso, per semplificare l'espressione della funzione di verosimiglianza, conviene utilizzare la cosidetta *Funzione di log-Verosimiglianza* (*log-Likelihood Function*):

$$\log \mathcal{L}(\theta; x_1, \ldots, x_n)$$

Perciò, il calcolo dello stimatore di massima verosimiglianza $\hat{\theta}$ diventa:

$$\hat{\theta} = \arg\max_{\theta} \log \mathcal{L}(\theta; x_1, \ldots, x_n) \tag{3.1.10}$$

da cui è possibile sfruttare la proprietà dei logaritmi tramite la quale il logartimo di un prodotto viene trasformato in una somma di logaritmi;

per esempio, nel caso di campioni i.i.d., si ottiene:

$$\hat{\theta} = \arg\max_{\theta} \sum_{i=1}^{n} \log f\left(x_i; \theta\right)$$

Il valore $\hat{\theta}$ ottenuto dall'Eq. (3.1.10) è lo stesso di quello ricavato dall'Eq. (3.1.8) in quanto, dato che il logaritmo è una funzione continua strettamente crescente, le funzioni $\mathcal{L}(\cdot)$ e $\log \mathcal{L}(\cdot)$ raggiungono il massimo nello stesso punto [2]. Tale proprietà è detta *Invarianza Funzionale*.

La definizione di stimatore di massima verosimiglianza implica che esso possa essere ricavato annullando la derivata prima della funzione di verosimiglianza (o, in alternativa, della funzione di log-verosimiglianza) [3]. Così, nel caso, per esempio, di una distribuzione di Bernoulli di parametro p, la funzione di verosimiglianza è

$$\mathcal{L}\left(p; x_1, \ldots, x_n\right) = \prod_{i=1}^{n} p^{x_i} \left(1-p\right)^{1-x_i} = p^{\sum_{i=1}^{n} x_i} \left(1-p\right)^{n-\sum_{i=1}^{n} x_i}$$

Passando al logaritmo, derivando rispetto a p e ponendo la derivata prima uguale a zero, si ottiene:

$$\frac{d}{dp} \log \mathcal{L}\left(p; x_1, \ldots, x_n\right) = 0 \Rightarrow \frac{1}{p} \sum_{i=1}^{n} x_i - \frac{1}{1-p} \left(n - \sum_{i=1}^{n} x_i \right) = 0$$

da cui si ricava che la stima di massima verosimiglianza $\hat{p} = \frac{1}{n} \sum_{i=1}^{n} x_i$ coincide con la media campionaria $\bar{x}$. Nel caso in cui θ sia un vettore $\vec{\theta}$ di parametri ignoti, occorre ricorrere alle derivate parziali:

$$\nabla \mathcal{L}\left(\vec{\theta}; x_1, \ldots, x_n\right) = \vec{\mathbf{0}}$$

[2] Dal punto di vista pratico, è di uso comune cercare il minimo della funzione $-\log \mathcal{L}(\cdot)$, detta *Funzione di log-Verosimiglianza Negativa* (*Negative log-Likelihood Function*), anzichè trovare il massimo di $\log \mathcal{L}(\cdot)$; i valori ottenuti sono ovviamente identici.

[3] Se la distribuzione è discreta e lo spazio dei parametri è finito, ad es., di dimensione k, è sufficiente scegliere il valore del parametro θ, fra i k valori possibili, che massimizza $\Pr\left(X_1 = x_1, \ldots, X_n = x_n; \theta\right)$.

cioè, supponendo che i parmetri ignoti siano $\theta_1, \ldots, \theta_m$, occorre risolvere le seguenti equazioni:

$$\begin{aligned} \frac{\partial}{\partial \theta_1} \mathcal{L}\left(\vec{\theta}; x_1, \ldots, x_n\right) &= 0 \\ &\vdots \\ \frac{\partial}{\partial \theta_m} \mathcal{L}\left(\vec{\theta}; x_1, \ldots, x_n\right) &= 0 \end{aligned}$$

Per esempio, nel caso di una distribuzione Normale con parametri incogniti μ e σ^2 (rispettivamente, media e varianza), occorre risolvere le seguenti due equazioni:

$$\begin{aligned} \frac{\partial}{\partial \mu} \log \mathcal{L}\left(\mu, \sigma^2; x_1, \ldots, x_n\right) &= 0 \\ \frac{\partial}{\partial \sigma} \log \mathcal{L}\left(\mu, \sigma^2; x_1, \ldots, x_n\right) &= 0 \end{aligned}$$

da cui si ricava:

$$\begin{aligned} \hat{\mu} &= \frac{1}{n} \sum_{i=1}^{n} x_i \\ \hat{\sigma}^2 &= \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{\mu})^2 \end{aligned}$$

L'utilizzo del metodo di massima verosimiglianza, rispetto al metodo dei momenti, è giustificato dalle seguenti proprietà:

Proposizione 3.1.1. *Indicando con n la numerosità del campione, il MLE $\hat{\Theta}$ del parametro ignoto θ gode delle seguenti proprietà:*

1. *$\hat{\Theta}$ è uno stimatore asintoticamente corretto per θ, cioè*

$$\lim_{n \to \infty} \mathrm{E}\left[\hat{\Theta}\right] = \theta$$

2. *$\hat{\Theta}$ è uno stimatore asintoticamente efficiente per θ, cioè, al crescere della dimen-*

sione n del campione, la sua varianza è tanto piccola quanto la varianza ricavata da qualsiasi altro stimatore [4]*;*

3. *$\hat{\Theta}$ è uno stimatore asintoticamente normale per θ, cioè la distribuzione di $\hat{\Theta}$ tende, al crescere di n, a una distribuzione Normale con media pari a θ.*

Malgrado queste proprietà siano valide all'infinito, in pratica possono essere considerate soddisfatte per n "sufficientemente grande".

Il metodo di massima verosimiglianza possiede anche dei difetti. Innanzi tutto, potrebbe non essere semplice trovare gli zeri della derivata prima oppure potrebbe essere computazionalmente intenso o poco stabile dal punto di vista numerico. Inoltre, a causa di irregolarità della funzione di verosimiglianza, il punto di massimo potrebbe non esistere, non essere unico, oppure dovrebbe essere calcolato con metodi approssimati; in quest'ultimo caso, i risultati ottenuti dal metodo numerico potrebbero essere influenzati dalla presenza di massimi locali (Fig. 3.1) o di zone piatte, dette *plateaux* o *flat* (Fig. 3.2).

3.1.3 Esempi di Stimatori

Stimatore della Media Sia X una variabile casuale di media μ ignota e varianza σ^2. Dato un campione $X_1, \ldots, X_n$ i.i.d., la *media campionaria*

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

è uno stimatore corretto per la media reale μ. Sotto ipotesi di validità del *Teorema del Limite Centrale* classico (B.1.1), la distribuzione di $\bar{X}$, per n abbastanza grande, è approssimativamente una Normale con media μ e varianza σ^2/n; si

[4] Detto in altri termini, nessuno stimatore corretto possiede un MSE più piccolo di quello associato a $\hat{\Theta}$.

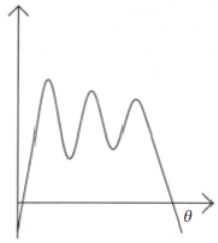


Figura 3.1: Massimi locali nella Funzione di Verosimiglianza.

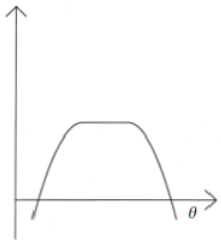


Figura 3.2: *Plateaux* nella Funzione di Verosimiglianza.

può scrivere, quindi, l'intervallo di confidenza al $100(1-\alpha)\%$:

$$\left(\bar{X} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right), \quad \text{se } \sigma^2 \text{ è nota}$$
$$\left(\bar{X} - t_{\frac{\alpha}{2},n-1}\frac{S}{\sqrt{n}}, \bar{X} + t_{\frac{\alpha}{2},n-1}\frac{S}{\sqrt{n}}\right), \quad \text{se } \sigma^2 \text{ è ignota}$$

dove:

- $z_{\alpha/2}$: è il quantile di una variabile Z, distribuita secondo una Normale standardizzata, tale per cui:

$$\Pr\left\{Z > z_{\frac{\alpha}{2}}\right\} = 1 - \Phi\left(z_{\frac{\alpha}{2}}\right) = \frac{\alpha}{2}$$

- $t_{\alpha/2,n-1}$: è il quantile di una variabile T_{n-1}, distribuita secondo una t-Student a $n-1$ gradi di libertà, tale per cui:

$$\Pr\left\{T_{n-1} > t_{\frac{\alpha}{2},n-1}\right\} = \frac{\alpha}{2}$$

Stimatore della Varianza Sia X una variabile casuale di varianza σ^2 ignota e media μ. Dato un campione $X_1, \ldots, X_n$ i.i.d., di media $\bar{X}$, la *varianza campionaria*

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$$

è uno stimatore distorto per la varianza reale σ^2. Dato che $\hat{\sigma}^2$ è uno stimatore non corretto, al suo posto si preferisce spesso utilizzare (soprattutto per valori di n piccoli) la cosidetta *varianza campionaria corretta*

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$$

la quale è uno stimatore corretto per la varianza reale σ^2. Nell'ipotesi di validità del *Teorema del Limite Centrale* classico (B.1.1), la distribuzione di $(n-1)S^2/\sigma^2$, per n abbastanza grande, è approssimativamente una χ^2 (Chi-Quadro) con $n-1$ gradi di libertà; si può scrivere, quindi, l'intervallo di confidenza al

$100(1-\alpha)\%$:

$$\left(\frac{(n-1)\,S^2}{\chi^2_{\frac{\alpha}{2},n-1}}, \frac{(n-1)\,S^2}{\chi^2_{1-\frac{\alpha}{2},n-1}}\right)$$

dove:

- $\chi^2_{\alpha/2,n-1}$: è il quantile di una variabile C_{n-1}, distribuita secondo una χ^2 con $n-1$ gradi di libertà, tale per cui:

$$\Pr\left\{C_{n-1} > \chi_{\frac{\alpha}{2},n-1}\right\} = \frac{\alpha}{2}$$

- $\chi^2_{1-\alpha/2,n-1}$: è il quantile di una variabile C_{n-1}, distribuita secondo una χ^2 con $n-1$ gradi di libertà, tale per cui:

$$\Pr\left\{C_{n-1} > \chi_{1-\frac{\alpha}{2},n-1}\right\} = 1 - \frac{\alpha}{2}$$

3.2 Test sulla Bontà di Adattamento (GoF)

Per costruire un modello è necessario capire quale, fra le distribuzioni di probabilità teoriche prese in considerazione, descriva meglio i campioni in esame. Ad esempio, la distribuzione Binomiale viene di solito utilizzata per caratterizzare la probabilità di ottenere un certo numero di successi durante l'esecuzione di una serie di esperimenti identici, ciascuno dei quali può assumere due possibili risultati (di solito denotati con "successo" e "fallimento") ed è indipendente dagli altri. Un altro esempio è dato dalla distribuzione di Poisson, la quale viene spesso utilizzata per descrivere il numero di arrivi in una coda o, in generale, il numero di nascite o morti in una popolazione. In alcuni casi, la scelta di utilizzare una particolare distribuzione, piuttosto che un'altra, deriva solo dall'esigenza di effettuare semplificazioni analitiche sul modello o da assunzioni basate su esperienze simili precedenti; è chiaro che questo non garantisce la costruzione di un modello che rispecchi la realtà e, in alcuni casi, può risultare in un lavoro totalmente inutile. Un esempio tipico, è l'uso e abuso della distribuzione Esponenziale per modellare i tempi di interarrivo in

una coda (e quindi, della distribuzione di Poisson per descrivere il numero di arrivi); ricerche abbastanza recenti (ad es., si veda [66, 93]), hanno messo in luce che, in particolari contesti, come quello della reti di computer, la miglior distribuzione che caratterizza i tempi di interarrivo è una distribuzione *Heavy-Tailed* (come una Pareto), per cui non valgono tutte le assunzioni che derivano dall'uso di una distribuzione Esponenziale. Nasce quindi l'esigenza critica di:

1. saper riconoscere e quantificare l'adattabilità di una certa distribuzione a un particolare insieme di dati;

2. poter capire quale tra due o più distribuzioni descriva meglio i dati in esame;

3. poter riconoscere se due campioni provengano da una stessa distribuzione.

Questa sezione ha lo scopo di illustrare i cosidetti *Test sulla Bontà di Adattamento* – in inglese, *Goodness-of-Fit* (*GoF*) *Tests* – i quali possono essere considerati dei test d'ipotesi (Def. B.2.11), in cui l'ipotesi nulla $\mathcal{H}_0$ prevede la presenza di adattamento, mentre quella alternativa $\mathcal{H}_1$ denota la mancanza di adattamento. Questi test possono essere divisi in due categorie:

- *test a un campione*, in cui si verifica l'adattabilità di una distribuzione teorica rispetto a un insieme di dati; quindi, dato un campione casuale Y_i, con $i = 1, \ldots, n$, i.i.d., proveniente da una distribuzione ignota Y, e una distribuzione teorica X, il test d'ipotesi consiste in:

$$\begin{aligned} \mathcal{H}_0 &: \Pr\{Y = y_i\} = p_X(y_i), \quad \text{per ogni } i = 1, \ldots, n \\ \mathcal{H}_1 &: \Pr\{Y = y_i\} \neq p_X(y_i), \quad \text{per qualche } i = 1, \ldots, n \end{aligned}$$

  dove $p_X(\cdot)$ denota la funzione di probabilità di X.

- *test a due campioni*, in cui si confrontano due insiemi di dati per verificare se provengono da una stessa distribuzione; cioè, dati due campioni casuali i.i.d. X_i, con $i = 1, \ldots, n$, proveniente da una distribuzione ignota

X con CDF $F_X(\cdot)$, e Y_j, con $j = 1, \ldots, n$, proveniente da una distribuzione ignota Y con CDF $F_Y(\cdot)$, il test d'ipotesi "bilaterale" (anche detto "a due code") consiste in:

$$\begin{aligned} \mathcal{H}_0 &: F_X(z) = F_Y(z), \quad \text{per ogni } z \\ \mathcal{H}_1 &: F_X(z) \neq F_Y(z), \quad \text{per qualche } z \end{aligned}$$

È anche possibile effettuare test "unilaterali" (anche detti "a una coda"); in questo caso l'ipotesi alternativa $\mathcal{H}_1$ suppone che una delle due CDF sia maggiore o minore dell'altra; in particolare, si ha un test "unilaterale destro" se l'ipotesi alternativa è:

$$\mathcal{H}_1 : F_X(z) > F_Y(z), \quad \text{per qualche } z$$

mentre si ha un test "unilaterale sinistro" se l'ipotesi alternativa è:

$$\mathcal{H}_1 : F_X(z) < F_Y(z), \quad \text{per qualche } z$$

I test sulla bontà di adattamento possono essere *test grafici*, in cui la bontà dell'adattamento viene valutata da un punto di vista visivo (e soggettivo), e *test numerici*, in cui l'indicazione sulla bontà dell'adattamento viene fornita in termini di p-value (cioè, di minima probabilità a partire dalla quale occorre rifiutare l'ipotesi nulla). I test grafici hanno il vantaggio di godere di un'applicabilità pressoché illimitata, ossia possono essere utilizzati con qualsiasi distribuzione e insieme di dati. Un possibile problema riguarda l'eventuale ambiguità nella loro interpretazione e la difficoltà di confronto dell'adattamento di due o più distribuzioni con uno stesso insieme di dati. I test numerici, d'altro canto, sono meno vulnerabili alle ambiguità di interpretazione e i loro risultati possono essere facilmente confrontati; tuttavia, hanno di solito un'applicabilità più ristretta, posso fornire risultati non propriamente corretti nel caso in cui i dati mostrino una particolare tendenza (ad es. code a legge polinomiale) e test diversi possono portare a conclusioni differenti. Un'ulteriore classificazione dei test consiste in test *parametrici* e test *non-parametrici*; un test si dice *non-*

parametrico se la distribuzione della statistica del test non dipende dalla distribuzione della popolazione, né esplicitamente (ad es., dalla forma) né implicitamente (ad es., da qualche parametro); viceversa, il test si dice *parametrico*. I test non-parametrici hanno il grosso vantaggio di poter essere applicati rispetto a qualsiasi distribuzione, mentre i test parametrici soffrono del problema che il valore del p-value deve essere recuperato da opportune tabelle (specifiche per la distribuzione sottoposta al test) o calcolato tramite specifiche simulazioni Monte-Carlo (anche in questo caso, parametrizzate secondo la distribuzione dell'ipotesi nulla del test); dal punto di vista dei risultati, si suppone che i test parametrici siano più sensibili alle piccole variazioni grazie al fatto di dipendere dalla distribuzione sottoposta al test. Dato che non esiste un unico modo per stabilire la bontà dell'adattamento, l'approccio consigliato è quello di utilizzare differenti test (anche grafici) e confrontare i risultati ottenuti; per motivi di flessibilità, e in seguito alla varietà di distribuzioni considerate, tutti i test numerici utilizzati nel presente lavoro saranno del tipo non-parametrico.

Il resto della sezione prende in esame i test di adattamento utilizzati nel presente lavoro; le prime due sezioni (§3.2.1 e §3.2.2) descrivono due test grafici basati sui quantili campionari, mentre le rimanenti sezioni, presentano una serie di test numerici non-parametrici. L'ultima sezione è dedicata alla descrizione di alcuni aspetti pratici che sono stati adottati nel progetto per la realizzazione dei test. Ove non specificato, si assumerà che il test, oggetto della discussione, sia di tipo "bilaterale".

3.2.1 Q-Q Plot

Il grafico *Quantile-Quantile* (*Q-Q plot*) [115] fa parte della famiglia dei cosidetti *grafici di probabilità* (*probability plot*). Può essere utilizzato sia per test a un campione sia per quelli a due campioni; nel caso di test a un campione, il campione dei dati viene confrontato con uno generato sinteticamente dalla distribuzione di probabilità teorica per cui si sta conducendo il test.

Questo grafico non è altro che un *diagramma a dispersione* (*scatter plot*) [101] relativo ai quantili dei due insiemi di dati (o della distribuzione). Per un test a un campione si utilizza il concetto di *quantile di una popolazione* (Def. B.1.6) e

di *quantile campionario* (Def. B.2.8); in un test a due campioni, invece, vengono presi in considerazione solo i quantili campionari dei due insiemi di dati.

Il calcolo dei quantili di distribuzioni teoriche può essere ricavato dalla definizione di *funzione quantile* (Def. B.1.6); nel caso in cui la funzione di distribuzione cumulativa sia invertibile, l'inversa della funzione di distribuzione cumulativa corrisponde alla funzione quantile. Per il calcolo dei quantili campionari, invece, esistono varie alternative; come spiegato in §B.2.2 (Def. B.2.8), nel presente lavoro si utilizza la forma fornita di "default" dal linguaggio *R* basata sulla seguente espressione:

$$p\left(k\right) = \frac{k-1}{n-1} \tag{3.2.1}$$

dove $p\left(k\right)$ è un valore di probabilità, k è l'ordine del quantile da stimare e n è la dimensione del campione.

La costruzione di un grafico Q-Q può essere effettuata nel seguente modo:

1. Dati due campioni di osservazioni $X_1 = x_1, \ldots, X_n = x_n$ e $Y_1 = y_1, \ldots, Y_m = y_m$, li si ordinino in modo crescente in modo da ottenere le relative *statistiche d'ordine* (Def. B.2.4):

$$\left\{x_{(i)}\right\}_{i=1}^{n} = x_{(1)}, \ldots, x_{(n)}$$
$$\left\{y_{(j)}\right\}_{j=1}^{m} = y_{(1)}, \ldots, y_{(m)}$$

2. Su un sistema di assi cartesiani, si disegnino i punti:

$$\left(x_{(i)}, y_{(i)}\right), \quad 1 \leq i \leq \min\left\{n, m\right\}$$

   Nel caso in cui $m \neq n$, l'insieme di dimensione maggiore è sottoposto a interpolazione lineare, in modo da uniformare la dimensione dei due insiemi.

3. Si tracci la *reference line*, cioè la retta passante per i 25-esimi e 75-esimi

quantili (ossia, per i *primi* e *terzi quartili*, rispettivamente):

$$y = ax + b \tag{3.2.2}$$

con:

$$\begin{aligned} a &= \frac{q^{(y)}_{0.75} - q^{(y)}_{0.25}}{q^{(x)}_{0.75} - q^{(x)}_{0.25}} \\ b &= q^{(y)}_{0.25} - a q^{(x)}_{0.25} \end{aligned} \tag{3.2.3}$$

dove $q^{(x)}_p$ e $q^{(y)}_p$ rappresentano, rispettivamente, i quantili di ordine p relativi all'insieme $\{x_i\}_{i=1}^n$ e all'insieme $\{y_j\}_{j=1}^m$.

La procedura appena descritta è adatta per effettuare un test a due campioni; per condurre un test a un campione, è sufficiente eseguire un test a due campioni, utilizzando, come uno dei due insiemi di dati, un campione generato dalla distribuzione teorica sottoposta al test; in particolare:

1. Dato un campione di osservazioni $y_1, \ldots, y_m$, lo si ordini in modo crescente in modo da ottenere le relative *statistiche d'ordine* $y_{(1)}, \ldots, y_{(m)}$.

2. Data una distribuzione teorica con funzione quantile $Q(\cdot)$, si generi il seguente campione:
$$x_{(i)} = Q\left(\frac{i - 0.5}{m}\right)$$
cioè il campione delle statistiche d'ordine relative ai *probability point* $(i - 0.5)/m$ (Def. B.2.9).

3. Procedere come il grafico Q-Q a due campioni, tracciando i punti $\left(x_{(i)}, y_{(i)}\right)$ e la "reference line" passante per i 25-esimi e 45-esimi quantili.

Se i punti risultano "vicini" alla "reference line", è ragionevole ipotizzare che i due campionni $\{x_i\}_{i=1}^n$ e $\{y_j\}_{j=1}^m$ provengano dalla stessa distribuzione (si veda, ad es., la Fig. 3.3); viceversa, se il grafico mostra un certo grado di curvatura si può sospettare che i due insiemi di dati provengano da distribuzioni differenti (si veda, ad es., la Fig. 3.4). Si noti che l'utilizzo di una retta passante per i quantili, anzichè, ad esempio, della retta di regressione lineare, come tecnica

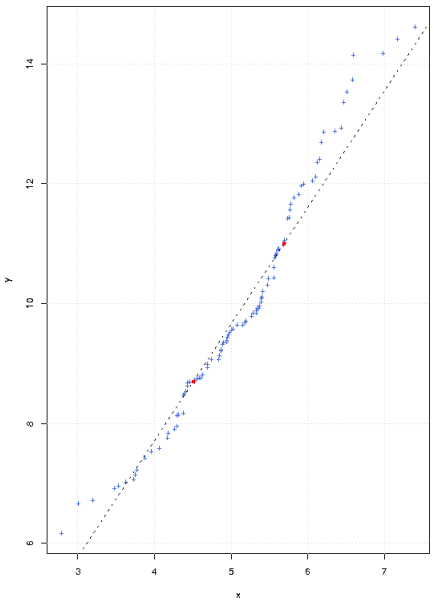


Figura 3.3: Grafico Q-Q con buona presenza di fit ($X_i \sim \mathcal{N}(5,1)$ e $Y_i \sim \mathcal{N}(10,2)$).

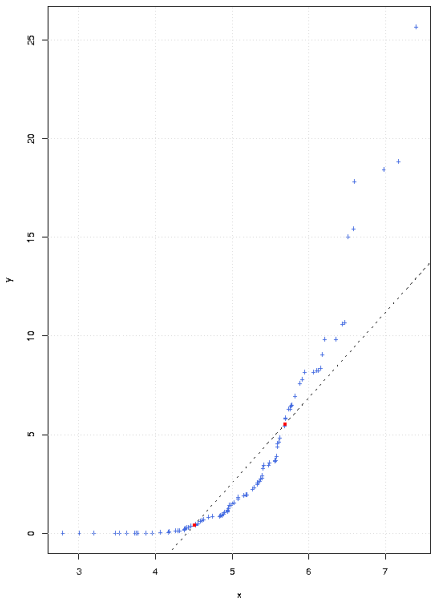


Figura 3.4: Grafico Q-Q con mancanza di fit ($X_i \sim \mathcal{N}(5,1)$ e $Y_i \sim$ Weib$(0.5,2)$).

di costruzione della "reference line", rappresenta un metodo "robusto" rispetto alla presenza di eventuali "outlier" o nel caso in cui fra i due campioni non vi sia una relazione di "normalità". Se i due insiemi di dati provengono dalla stessa distribuzione caratterizzata dagli stessi parametri, la retta sarà del tipo $y = x$ (retta a 45 gradi passante per l'origine); i due insiemi potrebbero però appartenere alla stessa famiglia di distribuzioni ma avere parametri differenti; il metodo risulta comunque "robusto" rispetto ai cambiamenti di locazione e di scala per entrambe le distribuzioni, ottenendo opportune variazioni nel punto di intersezione b e nel coefficiente angolare a: spostamenti nella locazione causeranno spostamenti dei punti del grafico a destra o a sinistra della "reference line"; differenze nella scala causeranno variazioni della pendenza della "reference line" dal valore 1. Un modo per capire se i dati si avvicinano alla retta, oltre a quello visuale, consiste, ad esempio, nel calcolare alcuni *coefficienti di correlazione campionaria*; fra i vari indici esistenti, si è deciso di utilizzare il

Coefficiente di Correlazione secondo Pearson (Def. B.2.12):

$$r = \frac{\sum_{i=1}^{N} (x_i - \bar{x})\,(y_i - \bar{y})}{(N-1)\,s_x s_y}$$

dove N è la numerosità dei due campioni $x_1, \ldots, x_N$ e $y_1, \ldots, y_N$, mentre s_x e s_y ne sono le rispettivte varianze campionarie. È bene far notare che il valore fornito dal coefficiente di Pearson, deve essere sempre giudicato con una certa criticità; infatti, tale coefficiente, ha due punti deboli non trascurabili: l'ipotesi di una relazione di normalità tra i due campioni e la sensibilità alla presenza di "outlier"; in effetti, l'uso di questo coefficiente equivale a utilizzare, come "reference line", la retta di regressione lineare, in quanto è noto che il relativo *coefficiente di determinazione* R^2 non è altro che il coefficiente di Pearson elevato al quadrato (ad es., si veda [101]). Sarebbe più opportuno utilizzare un coefficiente di tipo non-parametrico, come quello di *Kendall* o *Spearman* (Def. B.2.12), il quale, in generale, è più robusto sia alla presenza di "outlier" sia nel caso di trasformazioni monotone tra i dati; purtroppo, nei casi come questo in cui interessa solo determinare il grado di relazione lineare tra i dati, l'uso di questa classe di coefficienti non è indicato in quanto si catturerebbero anche relazioni di tipo non lineare. Ad esempio, indicando con τ il coefficiente di Kendall e con ρ quello di Spearman, con riferimento alla Fig. 3.3 si avrebbe:

$$r = 0.9827953$$
$$\tau = 1$$
$$\rho = 1$$

mentre rispetto alla Fig. 3.4:

$$r = 0.8531941$$
$$\tau = 1$$
$$\rho = 1$$

Come si può notare sia il coefficiente di Kendall sia quello di Spearman non

permettono di determinare il grado di linearità.

Il coefficiente di Pearson fornisce un valore compreso tra -1 e 1. Più il valore del coefficiente, in valore assoluto, è vicino all'unità (cioè più il grafico ha un andamento rettilineo), maggiore è la probabilità che i due insiemi di dati provengano dalla stessa distribuzione; al contrario, più il valore si avvicina a zero (cioè più marcata è la curvatura del grafico), maggiore è la probabilità che i due insiemi di dati provengano da distribuzioni differenti.

Il coeffieciente di correlazione, di per sé, non è sufficiente a capire la bontà dell'adattamento; infatti i punti del grafico potrebbero avere associato un coefficiente di correlazione piuttosto alto (vicino a 1) ma la retta intorno a cui si concentrano potrebbe non essere la "reference line"; come spiegato in precedenza, un caso del genere significherebbe che i punti sulle ascisse e quelle sulle ordinate appartengono alla stessa famiglia ma hanno diversi valori per i parametri di locazione e di scala. Quando occorre confrontare diversi grafici Q-Q si vorrebbe scegliere quello che, oltre ad avere un coefficiente di correlazione prossimo a 1, sia anche il più possibile vicino alla "reference line". Una misura che può fornire questo tipo di informazione è l'area ΔA, che di qui in seguito verrà chiamata *differenza di area assoluta*, compresa tra la "reference line" e la curva ottenuta unendo i punti del grafico Q-Q (*curva Q-Q*):

$$\begin{aligned}\Delta A &= |\langle\text{Area sottesa la Reference Line}\rangle - \langle\text{Area sottesa la Curva Q-Q}\rangle| \\ &\simeq \left| \sum_{i=2}^{n} \left| x_{(i)} y'_{(i)} - x_{(i-1)} y'_{(i-1)} \right| - \sum_{i=2}^{n} \left| x_{(i)} y_{(i)} - x_{(i-1)} y_{(i-1)} \right| \right| \\ &= \left| \sum_{i=2}^{n} \left(x_{(i)} y'_{(i)} - x_{(i-1)} y'_{(i-1)} \right) - \sum_{i=2}^{n} \left(x_{(i)} y_{(i)} - x_{(i-1)} y_{(i-1)} \right) \right|\end{aligned} \tag{3.2.4}$$

dove il calcolo dell'area è approssimato tramite la formula di integrazione del rettangolo composta [98] e $y'_{(i)}$ è il valore dell'ordinata sulla "reference line" in corrispondenza dell'ascissa $x_{(i)}$:

$$y'_{(i)} = a x_{(i)} + b, \quad i = 1, \ldots, n$$

dove a e b sono i coefficienti della "reference line", come definiti in (3.2.3). Nell'Eq. (3.2.4), l'ultimo passaggio è sempre valido in quanto si stanno consi-

derando le statistiche d'ordine, per cui vale una relazione di ordinamento in senso crescente. Inoltre, si è scelto di prendere il valore assoluto della differenza fra le due aree in quanto ciò che interessa, in questo caso, è una misura che indichi il grado di scostamento dalla "reference line"; nel caso, invece, si fosse interessati a conoscere se la curva Q-Q si trova più al di sopra piuttosto che al di sotto della "reference line", occorrerebbe eliminare l'uso del valore assoluto.

Per comodità di interpretazione e di confronto, è anche possibile definire una versione relativa della differenza di area (rispetto alla "reference line"), che di qui in seguito verrà chiamata *differenza di area relativa* ΔA_r:

$$\begin{aligned} \Delta A_r &= \frac{|\langle \text{Area sottesa la Reference Line} \rangle - \langle \text{Area sottesa la Curva Q-Q} \rangle|}{\langle \text{Area sottesa la Reference Line} \rangle} \\ &\simeq \frac{\Delta A}{\sum_{i=2}^{n} \left(x_{(i)} y'_{(i)} - x_{(i-1)} y'_{(i-1)} \right)} \end{aligned} \tag{3.2.5}$$

Più ΔA_r è prossima a 0, minore è la distanza della curva Q-Q dalla "reference line". Si noti che ΔA_r non è una misura normalizzata tra 0 e 1, in quanto può capitare che la curva Q-Q stia al di sopra della "reference line" e la differenza di area ΔA sia maggiore dell'area sottesa la "reference line"; in tal caso ΔA_r asssumerebbe un valore maggiore di 1.

Sebbene l'interpretazione di un grafico Q-Q sia di tipo soggettivo, esistono alcune linee guida, da usare in congiunzione con le misure quantitative descritte in precedenza. La Tab. 3.1 mostra l'interpretazione che si può dare ad alcuni dei "pattern" più comuni che si possono incontrare durante l'analisi di un grafico Q-Q.

I vantaggi di un test basato su di un grafico Q-Q comprendono:

- la semplicità e flessibilità di applicazione;
- l'efficienza con cui è possibile costruire il grafico;
- nel caso di test a due campioni, i due campioni analizzati non devono necessariamente avere la stessa dimensione;
- è invariante rispetto ai cambiamenti di scala e locazione (cioè una trasformazione lineare dei dati modifica solo la pendenza e il punto di interse-

Pattern	**Interpretazione**
Tutti i punti, tranne alcuni si distribuiscono intorno alla "reference line".	Presenza di "outlier" nei dati.
La parte terminale sinistra del "pattern" è al di sotto della "reference line", mentre la parte terminale destra si trova al di sopra.	Simmetria, code "lunghe" in entrambi gli estremi.
La parte terminale sinistra del "pattern" è al di sopra della "reference line", mentre la parte terminale destra si trova al di sotto.	Simmetria, code "lunghe" in entrambi gli estremi.
Andamento curvilineo convesso (cioè con pendenza crescente da sinistra verso destra).	Asimmetria a destra.
Andamento curvilineo concavo (cioè con pendenza decrescente da sinistra verso destra).	Asimmetria a sinistra.
Andamento a gradini.	I dati sono discreti o sono stati discretizzati.

Tabella 3.1: Interpretazione di alcuni "pattern" in un grafico Q-Q.

zione della "reference line"; rimane, invece, invariato il tipo di relazione che intercorre fra i due insieme di dati);

- è possibile verificare vari aspetti della distribuzione, come gli spostamenti di scala e di locazione, i cambiamenti di simmetria, la presenza di possibili "outlier". Per esempio, se i due campioni a confronto provengono da una stessa famiglia di distribuzioni e differiscono solo per spostamenti di locazione, i punti del grafico si posizioneranno intorno a una retta di 45 gradi parallela alla retta $y = x$, mentre se differiscono solo per differenze di scala, i punti si posizioneranno intorno a una retta con coefficiente angolare maggiore o minore di 45 gradi;
- è indicato per individuare differenze nelle code della distribuzione dei due campioni a confronto.

Come tutti i metodi grafici, soffre dello svantaggio di non fornire una misura quantitativa del "fitting"; tuttavia, come illustrato in precedenza, si possono utilizzare alcuni strumenti statistici (come i coefficienti di correlazione) per ottenere informazioni sulla bontà dell'adattamento. Altri svantaggi includono quello di non mettere bene in evidenza le eventuali differenze tra il "corpo" delle distribuzioni da cui i due campioni provengono, e che la sensibilità ai valori estremi (cioè ai valori provenienti dalle code della distribuzione) può risultare in un grafico con una eccessiva concentrazione di punti nella zona relativa al corpo della distribuzione [5]. In quest'ultimo caso, in [41], viene suggerito di tracciare un grafico Q-Q utilizzando una scala logaritmica, la quale ha l'effetto di ridurre la distanza tra punti con ordini di grandezza molto differente; per fare ciò, è sufficiente modificare il punto 2 della procedura di disegno di un grafico Q-Q, utilizzando una trasformazione logaritmica dei punti:

$$\left(\log x_{(i)}, \log y_{(i)}\right), \quad 1 \leq i \leq \min\{n, m\}$$

L'effetto dell'utilizzo della scala logaritmica applicato ai grafici Q-Q delle Fig. 3.3 e Fig. 3.4 è mostrato in Fig. 3.5 e in Fig. 3.6, rispettivamente. Questa trasfor-

[5] La presenza di qualche punto con ordine di grandezza molto grande, o molto piccolo, rispetto alla maggioranza dei punti influenza la scala del grafico.

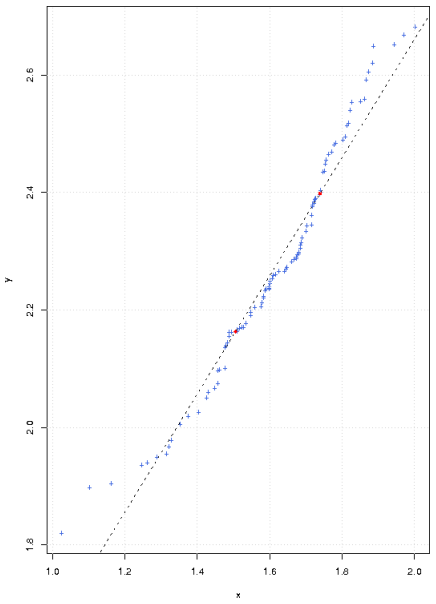


Figura 3.5: Grafico Q-Q con buona presenza di fit e scala logaritmica ($X_i \sim \mathcal{N}(5,1)$ e $Y_i \sim \mathcal{N}(10,2)$).

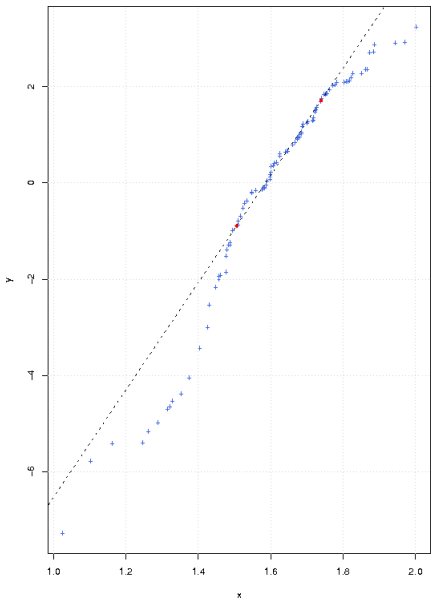


Figura 3.6: Grafico Q-Q con mancanza di fit e scala logaritmica ($X_i \sim \mathcal{N}(5,1)$ e $Y_i \sim \text{Weib}(0.5,2)$).

mazione risulta particolarmente utile quando si vuole analizzare il corpo della distribuzione e si è in presenza di distribuzioni caratterizzate da code a legge polinomiale; in alternativa, per lo studio del corpo della distribuzione, si può utilizzare un grafico P-P §3.2.2.

3.2.2 P-P Plot

Il grafico *Probabilità-Probabilità* (*P-P plot*) [115] è un altro tipo di grafico della famiglia dei "probability plot". Di solito viene utilizzato per confrontare un insieme di dati rispetto a una particolare distribuzione teorica.

La costruzione del grafico è simile a quella presentata per il grafico Q-Q §3.2.1:

1. Data una distribuzione di probabilità teorica con funzione di distribuzione $F(\cdot)$ e un campione di osservazioni $Y_1 = y_1, \ldots, Y_n = y_n$, si ordinino le

osservazioni in modo crescente in modo da ottenere le relative *statistiche d'ordine* $y_{(1)}, \ldots, y_{(n)}$ (Def. B.2.4).

2. Si costruisca l'insieme di dati $x_1, \ldots, x_n$ a partire dai punti di probabilità $(i - 0.5)/n$, con $1 \leq i \leq n$ (Def. B.2.9):

$$x_i = \frac{i - 0.5}{n}, \quad 1 \leq i \leq n$$

3. Si costruisca l'insieme di dati $u_1, \ldots, u_n$ a partire dalle statistiche d'ordine $y_{(1)}, \ldots, y_{(n]}$ e dalla funzione di distribuzione $F(\cdot)$:

$$u_i = F\left(y_{(i)}\right), \quad 1 \leq i \leq n$$

4. Su un sistema di assi cartesiani, si disegnino i punti:

$$(x_i, u_i), \quad 1 \leq i \leq n$$

5. Si tracci come "reference line" la retta passante per l'origine e con pendenza pari a 45 gradi:

$$y = x$$

È possibile utilizzare questo tipo di retta grazie alle regole di costruizione del grafico P-P [6].

Se i punti risultano "vicini" alla "reference line", è ragionevole ipotizzare che i due insiemi di dati $\{x_i\}_{i=1}^n$ e $\{u_j\}_{j=1}^n$ provengano dalla stessa distribuzione (si veda, ad es., la Fig. 3.7); viceversa, se il grafico mostra un certo grado di curvatura si può sospettare che i due insiemi di dati provengano da distribuzioni differenti (si veda, ad es., la Fig. 3.8) o dalla stessa famiglia di distribuzioni caratterizzate da valori di parametri differenti (si veda, ad es., la Fig. 3.9). Per valutare il grado di linearità tra i due insiemi di dati, si può ricorrere, come nel

[6] In presenza di fit, i punti di probabilità sui due assi si trovano alla stessa distanza.

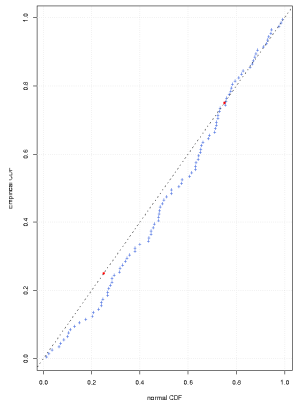


Figura 3.7: Grafico P-P con buona presenza di fit ($X_i \sim \mathcal{N}(5,1)$ e $F(\cdot) \sim \mathcal{N}(5,1)$).

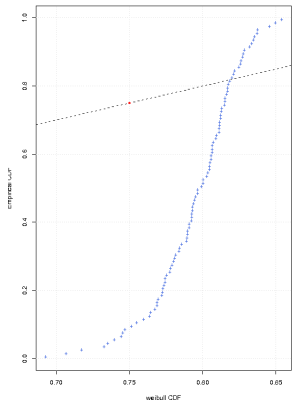


Figura 3.8: Grafico P-P con mancanza di fit: distribuzioni differenti ($X_i \sim \mathcal{N}(5,1)$ e $F(\cdot) \sim$ Weib$(0.5,2)$).

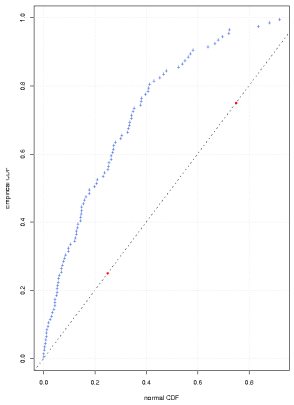


Figura 3.9: Grafico P-P con mancanza di fit: parametri differenti ($X_i \sim \mathcal{N}(5,1)$ e $F(\cdot) \sim \mathcal{N}(6,1)$).

grafico Q-Q, al *coefficiente di correlazione* r *di Pearson* e al calcolo della *differenza di area assoluta* ΔA e *relativa* ΔA_r.

Un vantaggio derivante dall'utilizzo di un grafico P-P è la più facile individuazione di eventuali differenze nel "corpo" della distribuzione, rispetto all'uso di un grafico QQ [52, 53]; ciò è essenzialmente dovuto al fatto che un grafico P-P è meno sensibile agli "outlier" rispetto a un grafico Q-Q, grazie all'utilizzo, come punti del grafico, dei valori di probabilità che, nel caso di "outlier", sono molto bassi. Inoltre in [57] vengono mostrati alcuni esempi in cui, in presenza di distribuzioni differenti, i grafici Q-Q possono risultare uguali, mentre i grafici P-P sono differenti; questo è, essenzialmente, causato dal fatto che un grafico Q-Q mette maggiormente in evidenza le differenze nelle code delle distribuzioni. Altri vantaggi, sono quelli che derivano dall'utilizzo dei metodi grafici (ossia, semplicità e flessibilità di utilizzo ed efficienza con cui è possibile costruirli) e che entrambi gli assi contengono solo valori compresi tra 0 e 1, rendendo più facile il confronto tra diversi grafici P-P.

Purtroppo una delle grosse limitazioni dei grafici P-P è quella di non essere invarianti rispetto ai cambiamenti di scala e di locazione; la Fig. 3.9 né è una prova: la distribuzione teorica e quella della popolazione da cui provie-

ne il campione appartengono alla stessa famiglia (distribuzione Normale), ma differiscono per uno spostamento della locazione. Come si può notare dalla figura, il grafico P-P non dà nessuna indicazione di questa relazione e mostra soltanto una "netta" mancanza di adattamento.

In maniera simile a quanto fatto notare per il grafico Q-Q, i punti del grafico P-P potrebbero risultare molto concentrati nelle zone relative alle code della distribuzione (in cui i valori di probabilità sono molto piccoli); per poter analizzare, con maggior precisione, tali zone mediante un grafico P-P, si può effettuare una trasformazione logaritmica dei punti di probabilità e dei valori della CDF teorica e quindi tracciarne il grafico P-P mediante la solita procedura; in alternativa, per lo studio delle code di una distribuzione si può utilizzare un grafico Q-Q.

3.2.3 Test del χ^2 secondo Pearson

Il testi del χ^2 *secondo Pearson* [95] è uno dei primi e meglio conosciuti test di adattamento numerici; fa parte della famiglia dei test del χ^2 in cui la statistica del test, sotto l'assunzione di validità dell'ipotesi nulla, segue (o approssima, nel caso di campioni piccoli) una distribuzione χ^2.

Il test prevede che i campioni siano i.i.d. e che siano divisi in k gruppi (*bin*), con $k > 1$; la versione *a un campione* verifica se le frequenze relative dei dati osservati (*frequenze osservate*) seguono una specifica distribuzione di frequenze data dall'ipotesi nulla (*frequeze attese*):

$$C_k = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \tag{3.2.6}$$

dove E_i rappresenta la frequenza osservata per il gruppo i (ricavata dal campione), mentre O_i denota la frequenza attesa per il gruppo i (ricavata dalla distribuzione dell'ipotesi nulla); le frequenze attese E_i si calcolano a partire dalla CDF $F(\cdot)$ della distribuzione dell'ipotesi nulla:

$$E_i = n\left[F\left(X_u\right) - F\left(X_l\right)\right]$$

dove X_u è il limite superiore per la classe i, mentre X_l è il limite inferiore; n è la dimensione del campione. In particolare, siano $C_1, \ldots, C_k$ le k classi e p_i la probabilità $\Pr\{X_1 \in C_i\}$ che un'osservazione del campione cada nella classe i-esima [7]; il numero di osservazioni O_i all'interno della classe i ha una distribuzione binomiale con parametri n e p_i. Si giunge quindi alla seguente espressione:

$$C_k = \sum_{i=1}^{k} \frac{(O_i - np_i)^2}{np_i} \tag{3.2.7}$$

Sotto la validità dell'ipotesi nulla (cioè che il campione provenga dalla distribuzione ipotizzata), la statistica del test C_k segue una distribuzione χ^2_{k-1} a $k-1$ gradi di libertà; dato un certo valore critico c, l'ipotesi nulla viene rifiutata per qualsiasi livello di significatività superiore al p-value p dato da:

$$p = \Pr\{C_k > c\}$$

La versione del test *a due campioni* ha un'espressione simile all'Eq. (3.2.6); per maggiori informazioni si veda, ad es., [96].

Il test del χ^2 è molto semplice da realizzare e non richiede eccessive risorse dal punto di vista computazionale; inoltre può essere applicato sia a distribuzioni discrete sia a quelle continue. Tuttavia soffre di alcune problematiche:

- il test richiede che i dati siano divisi in classi; il numero di gruppi e il modo con cui sono creati influenza, in generale, la potenza del test, cioè determina la sensibilità del test nell'individuare le variazioni dall'ipotesi nulla;

- la potenza del test, oltre a essere influenzata dal numero e dalla modalità di costruzione dei gruppi, è anche determinata dal numero dei campioni e dalla forma della distribuzione dell'ipotesi nulla;

- nel caso di dati discreti, l'appartenenza a un determinato gruppo può essere determinata senza ambiguità; al contrario, se i dati sono continui, i

[7] Per l'osservazione del campione si è utilizzato l'indice 1 in quanto i campioni $X_1, \ldots, X_n$ sono identicamente distribuiti.

gruppi devono essere definiti "a tratti", suddividendo l'intervallo continuo dei possibili valori in un certo numero di sotto-intervalli disgiunti; l'appartenenza a un gruppo è definita dai limiti di questi sotto-intervalli. Dato che questa suddivisione in gruppi è del tutto artificiale, la potenza del test ne può essere influenzata;

- per campioni di dimensione molto piccola, la statistica del test non approssima una distribuzione χ^2.

Per fortuna esistono alcune regole pratiche che permettono di compensare, in parte, alcuni dei punti deboli elencati in precedenza:

- per quanto riguarda la modalità di suddivisione di dati continui in gruppi, un modo tipico di procedere è quello di definire intervalli equispaziati (cioè, creare un istogramma dei dati in cui le barre hanno base uguale, ma, in generale, area differente); il problema di questo approccio è che il test diventa meno sensibile alle differenze in zone a bassa probabilità, come le code della distribuzione. Un approccio alternativo, e da molti considerato migliore [73, 60], è quello di costruire intervalli equiprobabili (cioè, creare un istogramma dei dati in cui le barre hanno area uguale, ma, in generale, base diversa); questa soluzione permette, al test, di guadagnare maggiore sensibilità, in particolar modo, nelle code della distribuzione;

- per quanto concerne il numero di gruppi, esistono due regole pratiche:

  1. utilizzare un numero di gruppi pari a $\lceil n^{2/5} \rceil$ [73, 81] o pari a $\lceil 1 + \log_2 n \rceil$, quest'ultimo usato da alcuni tool statistici, come *Mathwave EasyFit*; delle due tecniche, che in questo documento verranno chiamate come *metodo Moore* e *metodo Sturges*, rispettivamente, il primo tende a generare un numero maggiore di classi;

  2. assicurarsi che ogni gruppo, in cui è stata divisa la distribuzione dell'ipotesi nulla, contenga almeno 5 osservazioni, unendo, eventualmente, gruppi vicini in un unico gruppo (*bin pooling*) [104].

Nel presente lavoro sono stati implementati tutti gli accorgimenti sopra elencati; l'unico che non viene applicato automaticamente (ma viene attivato su richiesta dell'utente) è il "bin pooling", in quanto non sembra esservi un accordo nella comunità statistica sulla sua validità; molti sono a favore, in quanto viene assicurata la validità delle ipotesi effettuate durante la derivazione matematica del test (cioè, che, asintoticamente, la statistica del test segua una distribuzione χ^2); fra questi, però, la maggior parte sostiene che il numero minimo di valori in ogni classe sia 5, mentre altri affermano che tale limite inferiore sia posto uguale a 10; infine, c'è anche chi sostiene che l'utilizzao del "bin pooling" possa portare a una perdita di potenza del test, a causa del fatto che il raggruppamento tende ad avvenire nelle code delle distribuzione: molto spesso, queste zone sono quelle in cui le differenze tra l'ipotesi nulla e quella alternativa sono più importanti; l'uso del raggruppamento causerebbe quindi il mascheramento di queste differenze [24].

3.2.4 Test di Kolmogorov-Smirnov

Il test di *Kolmogorov-Smirnov* (*K-S*, in breve) [61] è un test numerico basato sulla EDF (Def. B.2.7) [105]; la relativa statistica fa parte delle cosidette *Supremum statistics* e rappresenta una misura di distanza verticale massima fra due funzioni di distribuzione; se la statistica risulta maggiore di un certo valore critico, l'ipotesi nulla del test viene rifiutata.

La versione *a un campione* misura la distanza massima tra la funzione di distribuzione cumulativa empirica $F_n(\cdot)$ di un campione e la funzione di distribuzione cumulativa teorica $F(\cdot)$ dell'ipotesi nulla, applicata allo stesso campione. La statistica D del test, a due code, è definita come

$$D_n = \sup_{-\infty \leq x \leq \infty} \{|F_n(x) - F(x)|\} \tag{3.2.8}$$

la quale può essere riscritta come:

$$D_n = \max\left\{D_n^+, D_n^-\right\}$$

dove:

$$D_n^+ = \sup_{-\infty \leq x \leq \infty} \{F_n(x) - F(x)\}$$
$$D_n^- = \sup_{-\infty \leq x \leq \infty} \{F(x) - F_n(x)\}$$

sono le statistiche da utilizzare nel caso in cui si voglia condurre un test unilaterale (destro e sinistro, rispettivamente).

La versione del test *a due campioni* è simile a quella a un campione, fatta eccezione che, al posto della distanza tra la distribuzione teorica dell'ipotesi nulla e la distribuzione empirica del campione, si calcola la distanza tra le distribuzioni empiriche dei due campioni a confronto; dati due campioni X_i, con $i = 1, \ldots, n$, e Y_j, con $j = 1, \ldots, m$, e le relative funzioni di distribuzioni empiriche $F_n(\cdot)$ e $G_m(\cdot)$, rispettivamente, la statistica del test vale:

$$D_{nm} = \sup_{-\infty \leq x \leq \infty} \{|F_n(x) - G_m(x)|\} \tag{3.2.9}$$

Applicando il PIT Cap. C alla statistica $D_{(\cdot)}$ del test, si ottiene:

$$D_N = \max_{1 \leq i \leq N} \left\{ \frac{i}{N} - u_i, u_i - \frac{j-1}{N} \right\} \tag{3.2.10}$$

dove $u_1, \ldots, u_N$ sono dei valori compresi tra 0 e 1 ottenuti attraverso l'applicazione del PIT, e N é la relativa numerosità. Per il test a un campione, indicando con x_i, con $i = 1, \ldots, n$, il campione sottoposto al test, e con $x_{(i)}$, con $i = 1, \ldots, n$, le relative statistiche d'ordine, si ha che $u_i = F\left(x_{(i)}\right)$, per $i = 1, \ldots, n$, e $N = n$. Per il test a due campioni, indicando i due campioni a confronto con x_i, con $i = 1, \ldots, n$, e y_j, con $j = 1, \ldots, m$, si ha che $u_1, \ldots, u_N$ sono le statistiche d'ordine del campione $x_1, \ldots, x_n, y_1, \ldots, y_m$ (unione dei due campioni da confrontare), ottenute attraverso un riordinamento in senso crescente, e $N = (n \times m)/(n + m)$. Quindi, sotto l'ipotesi che le condizioni di applicabilità del PIT siano soddisfatte, dall'Eq. (3.2.10) emerge che il test è indipendente dalla distribuzione (*distribution-free*) e può essere ridotto a un test per l'uniformità.

Il calcolo dei p-value della statistica D_N, per un certo valore critico d, è stato effettuato seguendo il metodo presentato in [75], da cui si ha che:

$$\begin{aligned}\Pr\{D_N > d\} &= 1 - \Pr\{D_N \leq d\} \\ &= 1 - \Pr\left\{D_N \leq \frac{k-h}{N}\right\} \quad \text{(con } k > 0 \text{ e } 0 \leq h \leq 1\text{)} \\ &= 1 - \frac{N!}{N^N} t_{kk}\end{aligned} \tag{3.2.11}$$

dove:

- N è la numerosità del campione;
- k è un numero positivo;
- $0 \leq h \leq 1$;
- t_{kk} è il k-esimo elemento della matrice $T = H^n$ e H è una matrice quadrata di ordine m, con $m = 2k - 1$, i cui elementi possiedono la seguente struttura:

$$h_{ij} = \begin{cases} \frac{\left(1-h^i\right)}{i!}, & i = 1, \ldots, (m-1) \text{ e } j = 1 \\ \frac{1}{(i-j+1)!}, & i = 1, \ldots, (m-1) \text{ e } j = 2, \ldots, (i+1) \\ 0, & i = 1, \ldots, (m-1) \text{ e } j = (i+1), \ldots, m \\ \frac{\left(1-2h^6\right)}{6!}, & i = m \text{ e } j = 1 \\ \frac{\left(1-h^{m-j+1}\right)}{(m-j+1)!}, & i = m \text{ e } j = 2, \ldots, m \end{cases}$$

La procedura per effettuare il test K-S a un campione è la seguente:

1. Dato un campione di osservazioni $x_1, \ldots, x_n$ i.i.d., lo si riordini in modo crescente per ottenere il relativo campione delle statistiche d'ordine $x_{(1)}, \ldots, x_{(n)}$.
2. Si calcolino i valori $u_i = F\left(x_{(i)}\right)$, con $i = 1, \ldots, n$, dove $F(\cdot)$ è la CDF associata all'ipotesi nulla del test.

3. Si applichi l'Eq. (3.2.8), utilizzando i valori calcolati al passo precedente, per calcolare il valore critico d del test.

4. Si calcoli il p-value relativo al valore critico d, cioè la probabilità $\Pr\{D_n > d\}$

5. Si rifiuti l'ipotesi nulla se il p-value, ottenuto al passo precedente, risulta più piccolo di un certo livello di significatività α prefissato (o se il p-value ottenuto è un valore troppo piccolo per non rfiutare l'ipotesi nulla).

Il vantaggio principale del test K-S è quello di essere, sotto l'ipotesi di validità del PIT, un test non parametrico, cioè totalmentee indipendente dalla distribuzione cumulativa sottoposta al test; inoltre, sembra che superi in potenza il test del χ^2 (cioè può rilevare differenze più piccole); tali caratteristiche hanno però un costo:

- può essere utilizzato solo per distribuzioni continue;
- tende a essere più sensibile intorno al centro della distribuzione (mediana) e a trascurare le code (a causa del fatto che vicino alle code delle distribuzioni la differenza tra le CDF è di solito molto piccola);
- la distribuzione deve essere completamente specificata; cioè, se i parametri della distribuzione dell'ipotesi nulla sono stimati direttamente dai dati sottoposti al test, il test perde di potenza.

In seguito alle suddette limitazioni, spesso si preferisce utilizzare il test di *Anderson-Darling*, il quale, fra i test basati sulla EDF, risulta uno fra i più potenti.

3.2.5 Test di Anderson-Darling

Il test di *Anderson-Darling* (*A-D*, in breve) [10, 9] è un test numerico, basato sulla EDF [105], la cui statistica fa parte delle cosidette *Quadratic statistics*; rappresenta una misura di distanza verticale quadratica pesata fra due funzioni di distribuzione; l'ipotesi nulla del test viene rifiutata quando la statistica risulta maggiore di un certo valore critico.

Si tratta di un caso particolare della famiglia di statistiche di *Cramer-von Mises*:

$$W_n^2 = n \int_{-\infty}^{\infty} \left[F_n(x) - F(x)\right]^2 \varphi\left(F(x)\right) \, dF(x) \tag{3.2.12}$$

in cui la funzione peso risulta $\varphi(t) = \left(t\left(1-t\right)\right)^{-1}$, con $t \in [0,1]$.

La versione del test *a un campione* misura la distanza quadratica pesata tra la funzione di distribuzione cumulativa empirica $F_n(\cdot)$ di un campione e la funzione di distribuzione cumulativa teorica $F(\cdot)$ dell'ipotesi nulla, applicata allo stesso campione:

$$A_n^2 = n \int_{-\infty}^{\infty} \frac{\left[F_n(x) - F(x)\right]^2}{F(x)\left(1 - F(x)\right)} \, dF(x) \tag{3.2.13}$$

In maniera simile, il test *a due campioni* [8], misura la distanza quadratica pesata tra le funzioni di distribuzione empiriche dei due campioni a confronto $X_1, \ldots, X_n$ e $Y_1, \ldots, Y_m$:

$$A_{nm}^2 = \frac{nm}{N} \int_{-\infty}^{\infty} \frac{\left[F_n\left(x\right) - G_m\left(x\right)\right]^2}{H_N\left(x\right)\left(1 - H_N\left(x\right)\right)} \, dH_N\left(x\right) \tag{3.2.14}$$

dove $N = n + m$, e $F_n\left(\cdot\right)$, $G_m\left(\cdot\right)$ e $H_N\left(\cdot\right)$ sono le EDF dei campioni $X_1, \ldots, X_n$, $Y_1, \ldots, Y_m$ e di quello combinato $X_1, \ldots, X_n, Y_1, \ldots, Y_m$; la funzione $H_N\left(\cdot\right)$ è definita come:

$$H_N\left(x\right) = \frac{nF_n\left(x\right) + mG_m\left(x\right)}{N}$$

Supponendo che la CDF della distribuzione dell'ipotesi nulla sia invertibile, è possibile applicare il PIT Cap. C alla statistica $A_{(\cdot)}^2$ del test, per il caso a un campione; ciò che si ottiene è la seguente espressione:

$$A_n^2 = -n - \sum_{i=1}^{n} \frac{2i-1}{n} \left[\ln u_i + \ln\left(1 - u_{n-i+1}\right)\right] \tag{3.2.15}$$

dove $u_1, \ldots, u_n$ sono dei valori compresi tra 0 e 1 ottenuti attraverso l'applicazione del PIT. Quindi, sotto l'ipotesi che le condizioni di applicabilità del PIT siano soddisfatte, dall'Eq. (3.2.15) emerge che il test è *non-parametrico*, cioè è indipendente dalla distribuzione (*distribution-free*), e può essere ridotto a un

test per l'uniformità.

La procedura per effettuare il test A-D a un campione è la seguente:

1. Dato un campione di osservazioni $x_1, \ldots, x_n$ i.i.d., lo si riordini in modo crescente per ottenere il relativo campione delle statistiche d'ordine $x_{(1)}, \ldots, x_{(n)}$.

2. Si calcolino i valori $u_i = F\left(x_{(i)}\right)$, con $i = 1, \ldots, n$, dove $F(\cdot)$ è la CDF associata all'ipotesi nulla del test.

3. Si applichi l'Eq. (3.2.13), utilizzando i valori calcolati al passo precedente, per calcolare il valore critico a del test.

4. Si calcoli il p-value relativo al valore critico a, cioè la probabilità $\Pr\{A_n^2 > a\}$

5. Si rifiuti l'ipotesi nulla se il p-value ottenuto al passo precedente risulta più piccolo di un certo livello di significatività α prefissato.

Per il caso a due campioni, si può ottenere un risultato analogo; tale caso non verrà preso in considerazione in quanto non è di interesse per il presente progetto; per maggiori informazioni si veda, ad esempio, [8, 97].

Il vantaggio principale di questo test è che risulta essere sensibile alle code della distribuzione; infatti la funzione peso $\varphi(u) = 1/[u(1-u)]$ è costruita in modo tale che per u "molto piccoli" ($u \to 0$) o "molto grandi" ($u \to 1$) assegni un peso maggiore; se si intendono queste due quantità come i valori restituiti da una CDF, esse, allora, rappresentano dei valori provenienti dalle code della distribuzione [8]. Quindi, quando un valore fa parte della coda della distribuzione, la "piccola" quantità ottenuta al numeratore [9] è compensata dal valore "grande" ottenuto al denominatore. Grazie a questa sensibilità per le code della distribuzione, l'applicazione di questo test è indicata specialmente nei casi in cui si sospetti che la distribuzione dell'ipotesi nulla abbia delle code che seguono una legge polinomiale. In [106], viene mostrato come il test A-D sia uno dei test più potenti fra quelli maggiormente utilizzati.

[8] Si ricordi, infatti, che il test lavora sulle statistiche d'ordine.

[9] Intorno alle code della distribuzione è più probabile che le differenze tra le due CDF siano molto piccole.

3.2.6 Considerazioni Aggiuntive

Parametri di una Distribuzione stimati dal Campione

I test numerici descritti nella precedente sezione, assumono che i parametri della distribuzione dell'ipotesi nulla siano indipendenti dal campione in esame; tuttavia in molti casi pratici, incluso il presente progetto, i parametri della distribuzione sono stimati dallo stesso insieme di dati sottoposto al test. A causa di questa dipendenza tra campione e distribuzione dell'ipotesi nulla, l'effetto che si ottiene è una perdita di potenza del test e una possibile inaffidabilità dei risultati ottenuti; per esempio, il test di Kolmogorov-Smirnov, in questi casi, risulta essere "conservativo" rispetto all'ipotesi nulla, fornendo dei valori delle statistiche il cui p-value è maggiore di quello che effettivamente dovrebbe essere [29]. Inoltre non è, in generale, nemmeno applicabile il PIT ai test basati sulla EDF (K-S, A-D, ...), in quanto i campioni uniformi ottenuti attraverso la sua applicazione non possono più essere considerati indipendenti [30]; questo implica che i test resi non-parametrici grazie all'applicazione del PIT (come K-S e A-D) diventano parametrici e quindi i valori dei p-value, dipendendo anche dalla distribuzione dell'ipotesi nulla, devono essere calcolati attraverso simulazioni Monte Carlo.

È possibile tuttavia ridurre gli effetti di questo problema utilizzando alcuni accorgimenti, specifici per ogni test.

Per il test del χ^2 una tecnica comune è quella di ridurre il numero di gradi di libertà per ogni parametro stimato dal campione sottoposto al test; così, se k è il numero di "bin" in cui viene suddiviso il campione, la statistica del test C_k, anzichè considerarla distribuita come una χ^2_{k-1} a $k-1$ gradi di libertà, si suppone che segua una distribuzione χ^2_{k-1-r} a $k-1-r$ gradi di libertà, dove r è il numero di parametri stimati dal campione sottoposto al test [23].

Per quanto concerne i test basati sulla EDF, non vi sono, al momento, delle regole generali ufficialmente riconosciute dalla comunità statistica; un'idea è quella di utilizzare la tecnica del *bootstrap* in versione *parametrica*. La tecnica del *bootstrap* è una particolare procedura di ricampionamento, introdotta per la prima volta da Efron [35], la cui idea di base può essere rias-

sunta nel seguente modo. Dato un campione casuale di osservazioni $\tilde{X} = (X_1 = x_1, \ldots, X_n = x_n)$ i.i.d., proveniente da una distribuzione di parametro ignoto θ e con CDF $F(\cdot;\theta)$, si supponga di voler analizzare la distribuzione di una certa statistica $T\left(\tilde{X}\right)$ (ad es., uno stimatore o una statistica di un test). Se la CDF $F(\cdot)$ fosse nota, sarebbe sufficiente campionare N campioni indipendenti $\tilde{X}_1, \ldots, \tilde{X}_N$ di numerosità n da $F(\cdot)$, con $\tilde{X}_i = (X_{i1} = x_{i1}, \ldots, X_{in} = x_{in})$ e $i = 1, \ldots, N$, e quindi calcolare le statistiche sugli N campioni, ottenendo il campione $T\left(\tilde{X}_1\right), \ldots, T\left(\tilde{X}_N\right)$; per N sufficientemente grande, è possibile stimare la CDF $F_T\left(\tilde{X}\right)$ di $T\left(\tilde{X}\right)$ con una buona accuratezza. Nella maggior parte dei casi, tuttavia, la funzione $F(\cdot)$ è ignota; in tal caso la si può stimare con la EDF $F_n(\cdot)$ del campione $\tilde{X} = (X_1 = x_1, \ldots . X_n = x_n)$. La tecnica del bootstrap prevede il ricampionamente casuale da $F_n(\cdot)$ ripetuto una serie di volte, ad es. N volte, in modo da ottenere i campioni $\tilde{X}_1^*, \ldots, \tilde{X}_N^*$; la distribuzione $F_T(\cdot)^*$ individuata da tale sequenza è detta *distribuzione del bootstrap* di $T\left(\tilde{X}\right)$ (mentre i campioni $\tilde{X}_i^*$ sono chiamati *campioni del bootstrap*). L'idea che sta alla base del "bootstrap" è che, supponendo che il campione originale $\tilde{X}$ rappresenti la popolazione dalla quale è stato estratto, ricampionando un numero molto alto di volte da esso (cioè dalla sua funzione di distribuzione empirica), si dovrebbe ricavare una serie di campioni molto simile a quella che si otterrebbe campionando direttamente dalla popolazione. A partire da questa considerazione, si può quindi supporre che la distribuzione ottenuta dal "bootstrap" sia una buona rappresentazione della distribuzione campionaria reale. Per quanto riguarda l'applicazione del "bootstrap" ai test di adattamento, in [13, 14] viene illustrato, sottoforma di dimostrazioni matematiche, come il metodo del "bootstrap", in un'ampia gamma di situazioni, fornisca una stima valida per i p-value di alcuni test di adattamento; in particolare, in [14] si fornisce una dimostrazione diretta per i test K-S e A-D [10]. Nel caso in cui i parametri della distribuzione dell'ipotesi nulla siano stati stimati dal campione sottoposto al test, occorre applicare il cosidetto *bootstrap parametrico*, in cui i campioni $\tilde{X}_1^*, \ldots, \tilde{X}_N^*$ sono generati dalla distribuzione i cui parametri sono

[10] Più precisamente, in [14] si dimostra come il "bootstrap" sia valido per quei test di GoF le cui statistiche tendano a un processo "browniano" $Y_n(x;\theta) = \sqrt{n}\left(F_n(x) - F(x;\theta)\right)$ (*Brownian Bridge*).

stati stimati dal campione originale $\tilde{X}$. I passi da effettuare per realizzare un test di adattamento tramite "bootstrap" parametrico sono i seguenti:

1. Dato un campione $\tilde{X} = (x_1, \ldots, x_n)$ e una distribuzione $F(\cdot;\theta)$ con parametro ignoto θ, si stima il parametro θ dal campione $\tilde{X}$, ottenendo la stima $\hat{\theta}$.

2. Fissato un certo test di adattamento T, si calcola la statistica

$$\hat{T} = T\left(\tilde{X}, F\left(\cdot;\hat{\theta}\right)\right)$$

   del test T (a un campione), in funzione di $\tilde{X}$ e della distribuzione $F(\cdot;\hat{\theta})$.

3. Si eseguono B iterazioni, in ciascuna delle quali:

   (a) Si genera un campione casuale $\tilde{X}^* = (x_1^*, \ldots, x_n^*)$ dalla distribuzione $F(;\hat{\theta})$.

   (b) Si calcola la stima $\hat{\theta}^*$ a partire dal campione $\tilde{X}_1^*, \ldots, \tilde{X}_N^*$, con lo stesso metodo usato per calcolare la stima $\hat{\theta}$ dal campione $\tilde{X}$.

   (c) Si calcola la statistica

$$\hat{T}^* = T\left(\tilde{X}^*, F\left(\cdot;\hat{\theta}^*\right)\right)$$

   del test di adattamento (a un campione), in funzione di $\tilde{X}^*$ e della distribuzione $F\left(\cdot;\hat{\theta}^*\right)$.

   (d) Si incrementa un contatore c (inizializzato a 0 prima del ciclo), nel caso in cui il valore di $\hat{T}^*$ supera quello di $\hat{T}$.

4. Si calcola la stima del p-value del test di adattamento come il rapporto tra il contatore c e il numero di iterazioni B incrementato di 1:

$$p\text{-value} = \frac{c}{B+1}$$

L'idea quindi è quella di simulare la distribuzione della statistica del test e verificare se il valore della statistica $\hat{T}$, ottenuta attraverso il campione originale

e la distribuzione con i parametri stimati $\hat{\theta}$, sia dovuto al caso (portando quindi a un non rifiuto dell'ipotesi nulla) o invece rappresenti un evento raro che si è verificato proprio sotto l'assunzione dell'ipotesi nulla (portando quindi a un rifiuto dell'ipotesi nulla). Questo controllo viene effettuato ricalcolando la statistica del test a partire da un nuovo campione $\tilde{X}^*$, ottenuto tramite ricampionamento; se il valore della nuova statistica $\hat{T}^*$ supera quello di $\hat{T}$, significa che si è verificato l'evento in cui la statistica del test supera il valore critico. Dopo aver eseguito le iterazioni del "bootstrap" si calcola la proporzione delle volte in cui la statistica del test ha superato il valore critico; tale proporzione rappresenta la stima del p-value.

Si noti che l'interpretazione dei risultati forniti da un test sulla bontà dell'adattamento, realizzato tramite "bootstrap" parametrico, dovrebbe essere effettuata con cautela; infatti, dato che ad ogni iterazione il valore di $\hat{T}^*$ viene ottenuto in funzione di un campione estratto da una distribuzione con parametro noto pari a $\hat{\theta}^*$, è chiaro che il test ne risulterà influenzato positivamente, e quindi, ciò che ci si aspetta, è un non rifiuto dell'ipotesi nulla; in tal caso il risultato del test non è da considerarsi significativo. Il risultato diventa, invece, molto significativo se porta a un rifiuto dell'ipotesi nulla, in quanto la probabilità che ciò avvegna è molto bassa. Quindi l'utilizzo del "bootstrap" parametrico in un test GoF, nel presente progetto, viene effettuato per confutare (e quindi escludere) il fatto che un certo campione provenga da una particolare distribuzione.

Potenza dei Test

Purtroppo non esistono teorie precise che dimostrino che un test di adattamento sia più "potente" (Def. B.2.11) rispetto a un altro. In generale la potenza di un test dipende, oltre che dalle sue caratteristiche intrinsiche, anche dalle proprietà della distribuzione (ad es., simmetria e tipo di coda) e dalla dimensione del campione. Lo studio della potenza di un test deve essere fatto, in generale, in maniera empirica attraverso una serie simulazioni Monte Carlo e al variare dei parametri dell'ipotesi alternativa; queste simulazioni devono essere condotte specificando la dimensione del campione, il tipo di distribuzione

dell'ipotesi nulla e, nel caso dei test non parametrici, il tipo di distribuzione per l'ipotesi alternativa. Dagli anni 50 sino a oggi sono stati prodotti molti articoli, alcuni dei quali in disaccordo tra di loro; ancora oggi, la ricerca su questo argomento della statistica matematica è ancora attiva.

Il test di Pearson χ^2 è un test che gode di una larga applicabilità e, per tale motivo, viene da molti definito come un "omnibus test". Tuttavia, a causa di questa sua generalità, è da considerarsi un test poco potente, in special modo quando i dati sottoposti al test non sono divisi in classi; in quest'ultimo caso, per poter utilizzare il test, è necessario effettuare una divisione artificiale dei dati.

I test basati sulle statistiche del massimo ("supremum statistics"), come il test di Kolmogorov-Smirnov, sono ritenuti più potenti del test di Pearson χ^2; ad es., in [62] si dimostra che per ottenere la stessa potenza raggiunta con il test di Pearson χ^2, su un campione di dimensione n, è sufficiente effettuare il test di Kolmogorov-Smirnov su un campione di numerosità pari a $2n^{2/5}$; in [105] sono mostrati altri risultati ottenuti attraverso una serie di simulazioni Monte Carlo. La maggiore potenza del test di Kolmogorov-Smirnov, rispetto a Pearson χ^2, può essere, intuitivamente, spiegata dall'assenza del vincolo di raggruppamento dei dati.

I test basati sulle statistiche quadratiche ("quadratic statistics"), come il test di Anderson Darling, si pensa che siano quelli con la maggiore potenza, in quanto ogni osservazione del campione contribuisce, in maniera pesata, al calcolo della statistica; il test di Kolmogorov-Smirnov, invece, tiene solo conto delle osservazioni che forniscono le distanze massime. Alla base di queste supposizione vi sono alcuni studi di potenza empirici effettuati tramite simulazioni Monte Carlo (ad es., si veda [105, 106]).

Interpretazione Pratica dei Risultati

I risultati ottenuti da un test dovrebbero essere valutati con cautela; si ricordi, infatti, che un test di adattamento rimane pur sempre un test, nel senso che fornisce risultati a cui occorre dare un significato probabilistico. Vi sono inoltre diversi fattori "pratici" che possono influenzare il risultato di un test:

- utilizzo di distribuzioni asintotiche per la statistica di un test; in molti casi occorre usare una distribuzione asintotica (cioè valida per un campione la cui dimensione tende all'infinito) in quanto può essere l'unica per cui si conosce una formula esplicita o per la quale sono disponibili i valori dei p-value;

- presenza di dati mancanti, parziali (*censored*) o errati;

- instabilità numerica dell'algoritmo usato per il calcolo della statistica del test; ad es., errori di cancellazione, di arrotondamento "underflow" e "overflow" possono portare a risultati completamente fuorvianti [98, 99];

- assunzioni dei test non completamente soddisfatte; ad es., la maggior parte dei test presuppone che i campione siano i.i.d.;

- inadeguatezza del test rispetto alla distribuzione dei dati; non tutti i test hanno la stessa potenza e alcuni non sono sensibili alle code della distribuzione;

- la presenza di un campione numeroso potrebbe amplificare alcuni dei problemi citati nei punti precedenti (ad es. gli errori di arrotondamento).

Nel presente progetto, sono stati realizzati tutti i test di adattamento presentati in questo capitolo. Non sembra esistere, nella teoria statistica, un modo corretto per combinare i risultati ottenuti da diversi test di adattamento, in special modo, da quelli numerici; un tentativo fu fatto da E. S. Pearson in [94], ma i risultati ottenuti sono di limitata applicabilità. Si è deciso quindi di seguire il seguente approccio empirico:

1. Si stabilisce un livello di significatività α, rappresentante la probabilità massima di commettere un errore di prima specie.

2. Si effettuano i test numerici, di Anderson-Darling, Kolmogorov-Smirnov e Pearson χ^2.

3. Sulla base delle considerazioni fatte precedentemente sulla potenza dei test, per la valutazione dei risultati ottenuti al passo precedente, si può

utilizzare la seguente regola empirica [11]:

$$\chi^2 \preccurlyeq \text{Supremum Statistics} \preccurlyeq \text{Quadratic Statistics} \tag{3.2.16}$$

Se tutti i test portano alla stessa conclusione (rifiuto o non rifiuto), relativamente al livello α prefissato, non vi sono ovviamente problemi di interpretazione. Viceversa, possono capitare casi in cui alcuni test portino a rifiutare l'ipotesi nulla, mentre altri, a non rifiutarla. Per prendere una decisione in questi casi, è stata ideata una euristica che sfrutta la definizione di potenza di un test e la regola empirica appena mostrata: dato che un test poco potente tenderà più facilmente a non rifiutare l'ipotesi nulla, mentre uno più potente, grazie alla sua maggior sensibilità, tenderà a rifiutarla, si avrà che un rifiuto da parte di un test poco potente o un non rifiuto da parte di un test potente possono essere considerati due fatti significativi. Quindi:

- un rifiuto da parte di un test poco potente e, contemporaneamente, un rifiuto da parte di un test più potente, porta a un rifiuto dell'ipotesi nulla;
- un non rifiuto da parte di un test poco potente e, contemporaneamente, un non rifiuto da parte di un test più potente, porta a un non rifiuto dell'ipotesi nulla;
- un rifiuto da parte di un test poco potente e, contemporaneamente, un non rifiuto da parte di un test più potente, rappresenta un caso ambiguo;
- un non rifiuto da parte di un test poco potente e, contemporaneamente, un rifiuto da parte di un test più potente, rappresenta un caso ambiguo.

Gli ultimi due casi sono ambigui perché rappresentano, rispettivamente, un fatto significativo e uno non significativo per entrambi i test. Si

[11] La regola empirica esclude i casi in cui i test di adattamento vengono eseguiti utilizzando la tecnica del "bootstrap" parametrico.

potrebbe decidere di scegliere come risultato finale quello del test più potente; tuttavia, si pensa che sia più ragionevole scegliere il risultato in funzione della distanza tra il livello di significatività α e il p-value del test, pesando il tutto rispetto alla potenza del test. Così, indicando con p_i il p-value del test i, con w_i il peso del test i e con r_i il risultato del test i, con $i \in \{ \chi^2, \text{K-S}, \text{A-D}\}$, si ottiene la seguente espressione:

$$\begin{aligned} \varepsilon(r_i) &= \begin{cases} w_i(r_i)(\alpha - p_i), & \text{se } r_i = \text{“rifiuto”} \\ w_i(r_i)(p_i - \alpha), & \text{se } r_i = \text{“non rifiuto”} \end{cases} \\ &= w_i(r_i)\,|\alpha - p_i| \end{aligned} \tag{3.2.17}$$

dove i pesi w_i possono essere assegnati secondo la relazione di potenza 3.2.16 e rispetto alla significatività del risultato; ad esempio:

$$\begin{aligned} w_{\chi^2}(\text{“rifiuto”}) &= 0.33, & w_{\chi^2}(\text{“non rifiuto”}) &= 0.17 \\ w_{K-S}(\text{“rifiuto”}) &= 0.33, & w_{K-S}(\text{“non rifiuto”}) &= 0.66 \\ w_{A-D}(\text{“rifiuto”}) &= 0.5, & w_{A-D}(\text{“non rifiuto”}) &= 1 \end{aligned} \tag{3.2.18}$$

L'espressione (3.2.17), in pratica, afferma che, in caso di rifiuto dell'ipotesi nulla, $\varepsilon(\cdot)$ equivale all'ampiezza dell'intervallo $[p_i, \alpha]$ (si noti che, quando avviene un rifiuto, il p-value del test è sempre minore o uguale a α); maggiore è l'ampiezza dell'intervallo, più alta è la probabilità che il rifiuto sia corretto. Invece, in caso di un non rifiuto dell'ipotesi nulla, $\varepsilon(\cdot)$ rappresenta l'ampiezza dell'intervallo $[\alpha, p_i]$ (infatti, in presenza di un non rifiuto, il p-value del test è sempre maggiore di α); maggiore è l'ampiezza dell'intervallo, più alta è la probabilità che vi sia adattamento.

Quindi per i casi ambigui, si sceglie il risultato del test i^* tale per cui:

$$i^* = \arg\max_i \left\{ \varepsilon(r_i) : i \in \left\{ \chi^2, \text{K-S}, \text{A-D} \right\} \right\} \tag{3.2.19}$$

La Tab. 3.2 riassume le considerazioni fatte in questo punto; le colonne etichettate con χ^2, K-S, A-D rappresentano i possibili risultati che si possono ottenere dai testi di Pearson χ^2, Kolmogorov-Smirnov e Anderson-

χ^2	**K-S**	**A-D**	i^*	r_{i^*}
R	R	R	χ^2	R
R	R	$\neg R$	Eq. (3.2.19)	r_{i^*}
R	$\neg R$	R	Eq. (3.2.19)	r_{i^*}
R	$\neg R$	$\neg R$	Eq. (3.2.19)	r_{i^*}
$\neg R$	R	R	Eq. (3.2.19)	r_{i^*}
$\neg R$	R	$\neg R$	Eq. (3.2.19)	r_{i^*}
$\neg R$	$\neg R$	R	Eq. (3.2.19)	r_{i^*}
$\neg R$	$\neg R$	$\neg R$	A-D	$\neg R$

Tabella 3.2: Regola pratica per combinare il risultato di diversi test di GoF: caso senza "bootstrap".

Darling, rispettivamente; ogni cella può contenere il simbolo R, rappresentante un "rifiuto", o $\neg R$, che denota un "non rifiuto". La colonna i^* riporta uno o più simboli di test che hanno contribuito, in maniera significativa, a prendere la decisione sul risultato, quando esso è determinato senza alcuna ambiguità; invece, in caso di situazioni ambigue, appare un riferimento all'Eq. (3.2.19). La colonna etichettata con r_{i^*} riporta il risultato finale della decisione; ogni sua cella può contenere il simbolo R, per un "rifiuto", $\neg R$, per un "non rifiuto", o r_{i*}, nel caso in cui il risultato dipenda dall'Eq. (3.2.19).

Come spiegato in precedenza, nel presente progetto, a causa del fatto che i parametri delle distribuzioni sono stati stimati dallo stesso insieme di dati coinvolto nel test, si è dovuta utilizzare la tecnica del "bootstrap" parametrico per eseguire i test di Kolmogorov-Smirnov e Anderson-Darling. In tal caso, la strategia di confronto tra i test deve essere modificata:

- se il risultato del test, eseguito tramite "bootstrap", è un non rifiuto, il test viene semplicemente ignorato;
- se il risultato del test, eseguito tramite "bootstrap", è un rifiuto:
  - in caso di ambiguità con i risultati degli altri test, si utilizza l'Eq. (3.2.17), impostando il peso in modo opportuno;
  - in caso di non ambiguità (cioè il risultato degli altri test non è significativo), si decide di rfiutare l'ipotesi nulla.

χ^2	K-S$_b$	A-D$_b$	i^*	r_{i^*}
R	R	R	χ^2, K-S, A-D	R
R	R	$\neg R$	χ^2, K-S$_b$	R
R	$\neg R$	R	χ^2, A-D$_b$	R
R	$\neg R$	$\neg R$	χ^2	R
$\neg R$	R	R	Eq. (3.2.19)	r_{i^*}
$\neg R$	R	$\neg R$	Eq. (3.2.19)	r_{i^*}
$\neg R$	$\neg R$	R	Eq. (3.2.19)	r_{i^*}
$\neg R$	$\neg R$	$\neg R$	χ^2	$\neg R$

Tabella 3.3: Regola pratica per combinare il risultato di diversi test di GoF: caso con "bootstrap".

La scelta dei pesi da utilizzare per questo tipo di test è del tutto arbitraria; occorre comunque tenere in considerazione che un rifiuto da parte di questo tipo di test dovrebbe rappresentare un fatto non trascurabile. Invece, in caso di non rifiuto è sufficiente assegnare ai pesi il valore 0. La Tab. 3.3 deriva dalla Tab. 3.2, dalla quale i test di Kolmogorov-Smirnov (K-S) e Anderson-Darling (A-D) sono stati sostituiti con le relative versioni eseguite tramite "bootstrap" parametrico (etichettati, rispettivamente, K-S$_b$ e A-D$_b$). Si noti che, a differenza di quanto uno ci si poteva aspettare, al test di Pearson χ^2 è stata conferita un'importanza maggiore rispetto alla Tab. 3.2; la motivazione di ciò risiede nel fatto che: il caso di non rifiuto dipende solamente dal risultato del test di Pearson χ^2 e la regola empirica sulla potenza dei test farebbe reputare poco significativo un non rifiuto per il test di Peasron χ^2; da questi due fatti, ne segue che difficilmente si riuscirebbe a ottenere un non rifiuto dell'ipotesi nulla. Quindi, per cercare di rendere più equilibrata la situazione, si è deciso di rendere "abbastanza significativo" il caso di non rifiuto da parte del test di Pearson χ^2, in modo tale che vi sia almeno un caso in cui si abbia, per certo, un non rifiuto; tuttavia, un rifiuto da parte di un test eseguito tramite "bootstrap" parametrico rimane un fatto più significativo di un non rifiuto da parte del test di Pearson χ^2.

4. In ogni caso, valutare i risultati dei test numerici attraverso uno o più

test grafici e valutare il relativo coefficiente di correlazione lineare; in generale, conviene utilizzare sia un grafico Q-Q sia un grafico P-P, in modo da sfruttarne i relativi vantaggi.

Capitolo 4

Distribuzioni di Probabilità

In questo capitolo, vengono presentate le principali distribuzioni di probabilità utilizzate per effettuare l'adattamento a un insieme di dati.

4.1 Cauchy

Si tratta di una distribuzione *heavy-tailed* su entrambe le code (Cap. 6) con media e varianza non definita; in alcuni contesti, viene anche chiamata distribuzione di *Lorentz* o di *Breit-Wigner*.

4.1.1 Caratterizzazione

Parametri La distribuzione ha due parametri:

- **Location** $x_0 \in \mathbb{R}$
- **Scale** $\gamma \in \mathbb{R}^+$

e si indica con *Cauchy*(x_0, γ).

Supporto La distribuzione è definita su tutti i valori dell'insieme dei numeri reali, cioè $x \in \mathbb{R}$.

CDF La funzione di distribuzione cumulativa ha la seguente forma:

$$F(x; x_0, \gamma) = \frac{1}{\pi} \arctan\left(\frac{x - x_0}{\gamma}\right) + \frac{1}{2}$$

PDF La funzione di densità di massa ha la seguente forma:

$$f(x; x_0, \gamma) = \frac{1}{\pi\gamma \left[1 + \left(\frac{x-x_0}{\gamma}\right)^2\right]}$$

Quantili La funzione quantile ha la seguente forma:

$$Q(p; x_0, \gamma) = x_0 + \gamma \tan\left[\pi \left(p - \tfrac{1}{2}\right)\right]$$

4.1.2 Stima dei Parametri

La stima dei parametri è ottenuta attraverso il metodo MLE, cioè la massimizzazione della stimatore di verosimiglianza.

4.1.3 Generazione di Numeri Casuali

Sia u un numero casuale uniforme in $[0, 1]$; applicando il metodo dell'inversione [32], si ottiene:

$$x = x_0 + \gamma \tan\left[\pi \left(u - \tfrac{1}{2}\right)\right]$$

dove x risulta distribuito come una $Cauchy\,(x_0, \gamma)$.

4.2 Fréchet

La distribuzione Fréchet [22] appare quando si è in presenza di distribuzioni "long-tailed", in cui la coda decade secondo una legge polinomiale ("power-law"), cioè più lentamente di una esponenziale, e "heavy-tailed" (distribuzioni "long-tailed" con momenti sono infiniti); in tali contesti il parametro "shape"

della distribuzione viene anche chiamato "tail-index", in quanto il suo valore influisce sulla decrescita della coda.

4.2.1 Caratterizzazione

Definizione 4.2.1 (Distribuzione Fréchet o Type II Extreme Value)**.** Data una sequenza di variabili casuali $Y_1, \ldots, Y_n$ i.i.d., con funzione di distribuzione F_Y, cioè:

$$F_Y(y) = F_{Y_i}(y) = \Pr\{Y_i \leq y\}), \quad 1 \leq i \leq n$$

e posto $M_n = \max\{Y_1, \ldots, Y_n\}$, la variabile

$$W = \frac{M_n - \mu}{\sigma}$$

è distribuita secondo una *Fréchet* di parametri μ, σ e α se:

$$\begin{aligned} F_W(w) &= \Pr\left\{\frac{M_n - \mu}{\sigma} \leq w\right\} \\ &= F_Y(\sigma w + \mu)^n \\ &\to H(w), \quad \text{per } n \to \infty \\ &= \begin{cases} 0 & w \leq 0 \\ \exp(-w^{-\alpha}) & w > 0 \end{cases} \\ &= \begin{cases} 0 & y \leq \mu \\ \exp\left[-\left(\frac{y-\mu}{\sigma}\right)^{-\alpha}\right] & y > \mu \end{cases}, \quad y = (\sigma w + \mu) \end{aligned}$$

La distribuzione dei minimi $m_n = \min\{Y_1, \ldots, Y_n\}$ può essere ottenuta a partire da quella dei massimi $M_n = \max\{Y_1, \ldots, Y_n\}$, ponendo $m_n = -M_n$.

Di seguito si indicherá con $\Gamma(z) = (z-1)!$, con $z \geq 1$, è la funzione *Gamma*.

Parametri La distribuzione ha tre parametri:

- **Location** $\mu \in \mathbb{R}$
- **Scale** $\sigma \in \mathbb{R}^+$

- **Shape** (o *tail-index*) $\alpha \in \mathbb{R}^+$

Quando $\mu = 0$ e $\sigma = 1$ la distribuzione viene detta *Fréchet Standard*.

Supporto La distribuzione è definita sull'intervallo $x \in (\mu - \sigma\alpha, +\infty)$.

CDF La funzione di distribuzione cumulativa ha la seguente forma:

$$F(X; \mu, \sigma, \alpha) = \exp\left[-\left(\frac{x-\mu}{\sigma}\right)^{-\alpha}\right]$$

PDF La funzione di densità di massa ha la seguente forma:

$$f(X; \mu, \sigma, \alpha) = \frac{\alpha}{\sigma}\left(\frac{x-\mu}{\sigma}\right)^{-(1+\alpha)} \exp\left[-\left(\frac{x-\mu}{\sigma}\right)^{-\alpha}\right]$$

Media La media vale:

$$E[X] = \mu - \sigma\alpha\left[1 - \Gamma\left(1 - \frac{1}{\alpha}\right)\right] < \infty \text{ iif } \alpha > 1$$

Varianza La varianza vale:

$$Var[X] = \sigma^2\alpha^2\left[\Gamma\left(1 - \frac{2}{\alpha}\right) - E[X]^2\right] < \infty \text{ iif } \alpha > 2$$

4.2.2 Stima dei Parametri

Metodo MLE

La funzione di log-verosimiglianza $\ln[\mathcal{L}(\mu,\sigma)] = \ln[f(x_1,\ldots,x_n;\mu,\sigma)]$ è:

$$\begin{aligned}
\ln[\mathcal{L}(\mu,\sigma,\alpha)] &= \ln\left\{\prod_{i=1}^{n}\frac{\alpha}{\sigma}\left(\frac{x_i-\mu}{\sigma}\right)^{-(1+\alpha)}\exp\left[-\left(\frac{x_i-\mu}{\sigma}\right)^{-\alpha}\right]\right\} \\
&= \sum_{i=1}^{n}\ln\left\{\frac{\alpha}{\sigma}\left(\frac{x_i-\mu}{\sigma}\right)^{-(1+\alpha)}\exp\left[-\left(\frac{x_i-\mu}{\sigma}\right)^{-\alpha}\right]\right\} \\
&= \sum_{i=1}^{n}\ln\left(\frac{1}{\sigma}\right)+\sum_{i=1}^{n}\left(-\frac{x_i-\mu}{\sigma}\right)+\sum_{i=1}^{n}\left[-\exp\left(-\frac{x_i-\mu}{\sigma}\right)\right] \\
&= \sum_{i=1}^{n}\ln\left(\frac{\alpha}{\sigma}\right)-(1+\alpha)\sum_{i=1}^{n}\ln\left(\frac{x_i-\mu}{\sigma}\right)-\sum_{i=1}^{n}\left(\frac{x_i-\mu}{\sigma}\right)^{-\alpha}
\end{aligned}$$

Il metodo MLE consiste nel risolvere le equazioni:

$$\begin{aligned}
\frac{\partial}{\partial\mu}\ln[\mathcal{L}(\mu,\sigma,\alpha)] &= 0 \\
\frac{\partial}{\partial\sigma}\ln[\mathcal{L}(\mu,\sigma,\alpha)] &= 0 \\
\frac{\partial}{\partial\alpha}\ln[\mathcal{L}(\mu,\sigma,\alpha)] &= 0
\end{aligned}$$

4.2.3 Generazione di Numeri Casuali

Sia u un numero casuale uniforme in $[0,1]$; applicando il metodo dell'inversione [32], si ottiene:

$$x = \mu + \sigma\left[\ln\left(\frac{1}{u}\right)\right]^{-\frac{1}{\alpha}}$$

dove x risulta distribuito come una $Frchet(\mu,\sigma,\alpha)$.

4.3 Gamma

4.3.1 Caratterizzazione

Nel seguito si indicherà con il simbolo $\gamma(\cdot)$ la funzione *gamma incompleta*:

$$\gamma(a, x) = \int_0^x t^{a-1}\, e^{-t}\, dt$$

mentre con il simbolo $\Gamma(\cdot)$ la funzione *gamma*:

$$\Gamma(z) = \int_0^\infty t^{z-1} e^{-t}\, \mathrm{d}t$$

Parametri La distribuzione ha due parametri:

- **Shape** $\alpha \in \mathbb{R}^+$
- **Scale** (o *Inverse Rate*) $\sigma \in \mathbb{R}^+$

e si indica con $Gamma(\alpha, \sigma)$.

Supporto La distribuzione è definita su tutti i valori reali non negativi, cioè $x \in \mathbb{R}^*$.

CDF La funzione di distribuzione cumulativa ha la seguente forma:

$$F(x; \alpha, \sigma) = \frac{\gamma(\alpha, x/\sigma)}{\Gamma(\alpha)}$$

PDF La funzione di densità di massa ha la seguente forma:

$$f(x; \alpha, \sigma) = x^{\alpha-1} \frac{\exp\left(-x/\sigma\right)}{\Gamma(\alpha)\, \sigma^\alpha}$$

Quantili La funzione quantile viene generata secondo l'algoritmo descritto in [15].

4.3.2 Stima dei Parametri

La stima dei parametri è ottenuta attraverso il metodo MLE, cioè la massimizzazione della stimatore di verosimiglianza. I valori iniziali da utilizzare per l'algoritmo iterativo per la massimizzazione della verosimiglianza sono ricavati utilizzando il metodo dei momenti.

4.3.3 Generazione di Numeri Casuali

Per la generazione di numeri casuali, distribuiti secondo una $Gamma(\alpha, \sigma)$, si utilizzano due algoritmi:

- per $\alpha \geq 1$, si utilizza l'algoritmo descritto in [7];
- per $0 < \alpha < 1$, si utilizza l'algoritmo descritto in [6].

4.4 Gumbel

La distribuzione *Gumbel* [50] (così chiamata in onore di Emil Julius Gumbel (1891-1966)) è utilizzata per trovare il massimo (o il minimo) di un numero di campioni. Ad esempio, potrebbe essere utilizzata per trovare il massimo livello d'acqua raggiunto da un fiume in uno specifico anno, a partire da una serie di massimi livelli per gli ultimi dieci anni. È quindi utile per predirre la probabilità che un estremo terremoto, uragano o qualsiasi altro disastro naturale avvenga.

I campioni X_i devono essere indipendenti e identicamente distribuiti (*i.i.d.*). La distribuzione Gumbel è utilizzata nella *teoria dei valori estremi* (*extreme value theory*) e rappresenta un caso limite della distribuzione *Generalized Extreme Value* (*GEV*); in questo conteso viene spesso chiamata come distribuzione *Type I Extreme Value*.

In particolare, la distribuzione Gumbel è un caso speciale della distribuzione *Fisher-Tippett* (così chiamata in onore di Sir. Ronald Aylmer Fisher (1890-1962) e di Leonard Henry Caleb Tippett (1902-1985)), nota anche come distribuzioni *log-Weibull*.

Quando intesa come distribuzione dei massimi, la distribuzione *Gumbel* è asimmetrica a destra; se $X \sim Gumbel(\mu, \sigma)$, allora $Y = -X$ è sempre distribuita come una Gumbel ma è asimmetrica a sinistra e rappresenta la distribuzione dei minimi.

La distribuzione Gumbel può rivelarsi utile nella modellazione di distribuzioni "medium-tailed" non limitate, in cui per valori grandi si ha una descrescita esponenziale; ad es. è possibile dimostrare che la distribuzione Normale ed Esponenziale sono un caso particolare della distribuzione Gumbel [1].

4.4.1 Caratterizzazione

Definizione 4.4.1 (Distribuzione Gumbel o Type I Extreme Value)**.** Data una sequenza di variabili casuali $Y_1, \ldots, Y_n$ i.i.d., con funzione di distribuzione F_Y, cioè:

$$F_Y(y) = F_{Y_i}(y) = \Pr\{Y_i \leq y\}), \quad 1 \leq i \leq n$$

e posto $M_n = \max\{Y_1, \ldots, Y_n\}$, la variabile

$$W = \frac{M_n - \mu}{\sigma}$$

è distribuita secondo una *Gumbel* di parametri μ e σ se:

$$\begin{aligned} F_W(w) &= \Pr\left\{\frac{M_n - \mu}{\sigma} \leq w\right\} \\ &= F_Y(\sigma w + \mu)^n \\ &\to H(w), \quad \text{per } n \to \infty \\ &= \exp[-\exp(-w)], \quad \forall w \in \mathbb{R} \\ &= \exp\left[-\exp\left(-\frac{y-\mu}{\sigma}\right)\right], \quad \forall y = (\sigma w + \mu) \in \mathbb{R} \end{aligned}$$

La distribuzione di

$$\rho(y) = \exp\left(-\frac{y-\mu}{\sigma}\right)$$

[1] In tale caso si dice anche che le distribuzioni Normale ed Esponenziale sono nello stesso "dominio di attrazione" della Gumbel.

è detta *parent distribution*.

La distribuzione dei minimi $m_n = \min\{Y_1, \ldots, Y_n\}$ può essere ottenuta a partire da quella dei massimi $M_n = \max\{Y_1, \ldots, Y_n\}$, ponendo $m_n = -M_n$.

Di seguito, il simbolo γ denoterà la costate di *Eulero-Mascheroni* [37], cioè $\gamma = 0.57721\ldots$

Parametri La distribuzione ha due parametri:

- **Location** $\mu \in \mathbb{R}$
- **Scale** $\sigma \in \mathbb{R}^+$

Quando $\mu = 0$ e $\sigma = 1$ la distribuzione viene detta *Gumbel Standard*.

Supporto La distribuzione è definita su tutto l'insieme $\mathbb{R}$, cioè $x \in (-\infty, +\infty)$.

CDF La funzione di distribuzione cumulativa ha la seguente forma:

$$F(x; \alpha, \mu, \sigma) = \exp\left[-\exp\left(-\frac{x-\mu}{\sigma}\right)\right]$$

PDF La funzione di densità di massa ha la seguente forma:

$$f(x; \mu, \sigma) = \frac{1}{\sigma}\exp\left(-\frac{x-\mu}{\sigma}\right)\exp\left[-\exp\left(-\frac{x-\mu}{\sigma}\right)\right]$$

Media La media vale:

$$E[X] = \mu - \sigma\gamma$$

Varianza La varianza vale:

$$Var[X] = \frac{\pi^2}{6}\sigma^2$$

4.4.2 Stima dei Parametri

Metodo dei Momenti

I parametri μ e σ possono essere stimati a partire dalla media campionaria $\bar{X}$ e dalla varianza campionaria s^2:

$$\hat{\sigma} = s\frac{\sqrt{6}}{\pi}$$
$$\hat{\mu} = \bar{X} - \gamma\hat{\sigma}$$

Metodo MLE

La funzione di log-verosimiglianza $\ln\left[\mathcal{L}\left(\mu,\sigma\right)\right] = \ln\left[f\left(x_1,\ldots,x_n;\mu,\sigma\right)\right]$ è:

$$\begin{aligned}\ln\left[\mathcal{L}\left(\mu,\sigma\right)\right] &= \ln\left\{\prod_{i=1}^{n}\frac{1}{\sigma}\exp\left(-\frac{x_i-\mu}{\sigma}\right)\exp\left[-\exp\left(-\frac{x_i-\mu}{\sigma}\right)\right]\right\}\\ &= -n\ln\left(\sigma\right) - \sum_{i=1}^{n}x_i - n\frac{\mu}{\sigma} - \sum_{i=1}^{n}\exp\left(-\frac{x_i-\mu}{\sigma}\right)\end{aligned}$$

Il metodo MLE consiste nel risolvere le equazioni:

$$\frac{\partial}{\partial\mu}\ln\left[\mathcal{L}\left(\mu,\sigma\right)\right] = 0$$
$$\frac{\partial}{\partial\sigma}\ln\left[\mathcal{L}\left(\mu,\sigma\right)\right] = 0$$

4.4.3 Generazione di Numeri Casuali

Sia u un numero casuale uniforme in $[0,1]$; applicando il metodo dell'inversione [32], si ottiene:

$$x = \mu - \sigma\ln\left[-\ln\left(u\right)\right]$$

dove x risulta distribuito come una $Gumbel\left(\mu,\sigma\right)$.

4.5 Logistica

La distribuzione *Logistica* è una distribuzione di probabilità simmetrica, caratterizzata da due code più lunghe e da un picco più alto rispetto alla distribuzione Normale; la sua CDF rappresenta una funzione *logistica* e, in particolare, una funzione *sigmoide* (anche detta funzione *S-shaped*), e viene spesso utilizzata nelle reti neurali per introdurre una componente di non linearità. La distribuzione *Logistica* è talvolta chiamata distribuzione *Sech-Squared*.

4.5.1 Caratterizzazione

Parametri La distribuzione ha due parametri:

- **Location** $\mu \in \mathbb{R}$
- **Scale** $\sigma \in \mathbb{R}^+$

e si indica con $Logistic(\mu, \sigma)$.

Supporto La distribuzione è definita su tutti i valori reali, cioè $x \in \mathbb{R}$.

CDF La funzione di distribuzione cumulativa ha la seguente forma:

$$\begin{aligned} F(x; \alpha, \sigma) &= \frac{1}{1 + \exp\left(-\frac{x-\mu}{\sigma}\right)} \\ &= \frac{1}{2}\left[1 + \tanh\left(\frac{x-\mu}{2\sigma}\right)\right] \end{aligned}$$

dove $\tanh$ rappresenta la funzione *tangente iperbolica*.

PDF La funzione di densità di massa ha la seguente forma:

$$f(x; \mu, \sigma) = \frac{\exp\left(-\frac{x-\mu}{\sigma}\right)}{\sigma\left[1 + \exp\left(-\frac{x-\mu}{\sigma}\right)\right]^2}$$

Quantili La funzione quantile ha la seguente forma:

$$Q(p; \mu, \sigma) = \mu + \sigma \ln\left(\frac{p}{1-p}\right)$$

4.5.2 Stima dei Parametri

La stima dei parametri è ottenuta attraverso il metodo MLE, cioè la massimizzazione della stimatore di verosimiglianza. I valori iniziali da utilizzare per l'algoritmo iterativo per la massimizzazione della verosimiglianza sono ricavati utilizzando il metodo dei momenti.

4.5.3 Generazione di Numeri Casuali

Sia u un numero casuale uniforme in $[0, 1]$; applicando il metodo dell'inversione [32], si ottiene:

$$x = \mu + \sigma \ln \left(\frac{u}{1-u} \right)$$

4.6 Log-Normale

La distribuzione *Log-Normale* è una distribuzione di probabilità dotata di una coda lunga (sub-esponenziale); essa è strettamente correlata alla distribuzione Normale; infatti, una variabile casuale X ha distribuzione *Log-Normale* se la variabile casuale $Y = \ln(X)$ ha distribuzione Normale di media μ e deviazione standard σ. Questa relazione si può esprimere, per esempio, attraverso la funzione di densità:

$$f(x; \mu, \sigma) = \frac{1}{x} f_{\mathcal{N}}(\ln x; \mu, \sigma) \tag{4.6.1}$$

dove $f(\cdot)$ è la funzione di densità di una Log-Normale, mentre $f_{\mathcal{N}}(\cdot)$ è la funzione di densità della relativa distribuzione Normale.

4.6.1 Caratterizzazione

Nel seguito si indicherà con il simbolo $\mathrm{erf}(\cdot)$ la *error function*:

$$\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt$$

mentre con il simbolo $\Phi(\cdot)$ la funzione di distribuzione di una distribuzione Normale standard:

$$\frac{1}{2}\left[1+\operatorname{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right]$$

Parametri La distribuzione ha due parametri:

- **Location** $\mu \in \mathbb{R}$ (media della distribuzione Normale associata)
- **Scale** $\sigma \in \mathbb{R}^*$ (deviazione standard della distribuzione Normale associata)

e si indica con $Log\mathcal{N}(\mu, \sigma)$.

Supporto La distribuzione è definita su tutti i valori reali non negativi, cioè $x \in \mathbb{R}^*$.

CDF La funzione di distribuzione cumulativa ha la seguente forma:

$$F(x;\mu,\sigma) = \frac{1}{2}\left\{1+\operatorname{erf}\left[\frac{\ln(x)-\mu}{\sigma\sqrt{2}}\right]\right\} = \Phi\left[\frac{\ln(x)-\mu}{\sigma}\right]$$

PDF La funzione di densità di massa ha la seguente forma:

$$f(x;\mu,\sigma) = \frac{1}{x\sigma\sqrt{2\pi}}\exp\left(-\frac{[\ln(x)-\mu]^2}{2\sigma^2}\right)$$

Quantili La funzione quantile ha la seguente forma:

$$Q\left(p;\mu,\sigma\right) = \exp\left[Q_{\mathcal{N}}\left(p;\mu,\sigma\right)\right]$$

dove $Q_{\mathcal{N}}(\cdot)$ è la funzione quantile della distribuzione Normale associata.

$$Q_{\mathcal{N}}\left(p;\mu,\sigma\right) = \mu + \sigma\Phi^{-1}\left(p\right) = \mu + \sigma\sqrt{2}\ \operatorname{erf}^{-1}\left(2p-1\right)$$

4.6.2 Stima dei Parametri

La stima dei parametri è ottenuta attraverso la massimizzazione dello stimatore di verosimiglianza (metodo MLE). In particolare, sfruttando l'Eq. (4.6.1),

si ottiene che:

$$\mathcal{L}(\mu, \sigma | x_1, \ldots, x_n) = \sum_k \ln x_k + \mathcal{L}_{\mathcal{N}}(\mu, \sigma | \ln x_1, \ldots, \ln x_n)$$

dove $\mathcal{L}_{\mathcal{N}}(\cdot)$ è la funzione di verosimiglianza di una Normale.

4.6.3 Generazione di Numeri Casuali

Sia u un numero casuale uniforme in $[0, 1]$; applicando il metodo dell'inversione [32], si ottiene:

$$x = \exp\left[\mu + \sigma\Phi^{-1}(u)\right] = \exp\left[\mu + \sigma\sqrt{2}\ \mathrm{erf}^{-1}(2u - 1)\right]$$

4.7 Long-Tailed e Heavy-Tailed

Una distribuzione a *Coda Lunga* (*Long-Tailed*) è una distribuzione in cui almeno una delle due code cade più lentamente delle tradizionali distribuzioni Gaussiane. Nella distribuzione Normale, e in quelle che asintoticamente tendono a essa, il decadimento della coda è di tipo esponenziale; invece, nelle distribuzioni a coda lunga la coda cade con un legge sub-esponenziale. L'effetto diretto di questa proprietà è il maggior peso che i valori estremi hanno sul comportamento dell'intera distribuzione.

Fra le distribuzioni a coda lunga, destano particolare interesse le cosidette *Power-Law*, cioè quelle in cui la coda decade secondo una legge polinomiale; tra queste, ve n'è una classe, chiamata *Heavy-Tailed*, la cui peculiarità principale è quella di avere alcuni momenti infiniti.

Data la vastità dell'argomento, il Cap. 6 è interamente dedicato alla descrizione delle caratteristiche di questa classe di distribuzioni.

4.8 Pareto Generalizzata (GPD)

4.8.1 Caratterizzazione

Parametri La distribuzione ha tre parametri:

- **Location** (*threshold*) $\mu \in \mathbb{R}$
- **Scale** $\sigma \in \mathbb{R}^+$
- **Shape** (*tail-index*) $\xi \in \mathbb{R}$

Supporto La distribuzione ha il seguente supporto:

$$\begin{cases} x \geq \mu, & \xi \geq 0 \\ \mu < x < -\frac{\sigma}{\xi}, & \xi < 0 \end{cases}$$

Quando $\xi = 0$ e $\mu = 0$, la GPD equivale a una distribuzione *Esponenziale a due parametri* (o *Esponenziale shifted*), cioè a una distribuzione Esponenziale con tasso σ e con in più il parametro "location" μ; se $\mu = 0$, la GPD corrisponde a una Esponenziale tradizionale. Quando $\xi > 0$ e $\mu = \sigma$, la GPD corrisponde alla distribuzione Pareto classica.

CDF La funzione di distribuzione cumulativa ha la seguente forma:

$$F\left(x; \xi, \mu, \sigma\right) = \begin{cases} 1 - \left(1 + \xi \frac{x-\mu}{\sigma}\right)_+^{-1/\xi}, & \xi \neq 0 \\ 1 - \exp\left(-\frac{x-\mu}{\sigma}\right), & \xi = 0 \end{cases}$$

Il caso speciale $\xi = 0$ rappresenta la CDF di una distribuzione Esponenziale a due parametri.

PDF La funzione di densità ha la seguente forma:

$$f\left(x; \xi, \mu, \sigma\right) = \begin{cases} \frac{1}{\sigma}\left(1 + \xi \frac{x-\mu}{\sigma}\right)^{-1-\frac{1}{\xi}}, & \xi \neq 0 \\ \frac{1}{\sigma}\exp\left(-\frac{x-\mu}{\sigma}\right), & \xi = 0 \end{cases}$$

Quantili La funzione quantile ha la seguente forma:

$$Q\left(p;\xi,\mu\,\sigma\right)=\begin{cases}\mu+\frac{\sigma}{\xi}\left(1-p\right)^{\xi}-1, & \xi\neq 0\\ \mu-\sigma\ln\left(1-p\right), & \xi=0\end{cases}$$

4.8.2 Stima dei Parametri

Metodo MLE

La funzione di log-verosimiglianza $\ln\left[\mathcal{L}\left(\xi,\mu,\sigma\right)\right]=\ln\left[f\left(x_1,\ldots,x_n;\xi,\mu,\sigma\right)\right]$ è:

$$\begin{aligned}\ln\left[\mathcal{L}\left(\xi,\mu,\sigma\right)\right]&=\begin{cases}\ln\left[\prod_{i=1}^{n}\frac{1}{\sigma}\left(1+\xi\frac{x_i-\mu}{\sigma}\right)^{-1-\frac{1}{\xi}}\right], & \xi\neq 0\\ \ln\left[\prod_{i=1}^{n}\frac{1}{\sigma}\exp\left(-\frac{x_i-\mu}{\sigma}\right)\right], & \xi=0\end{cases}\\ &=\begin{cases}-n\ln\left(\sigma\right)-\left(1+\frac{1}{\xi}\right)\sum_{i=1}^{n}\ln\left(1+\xi\frac{x_i-\mu}{\sigma}\right), & \xi\neq 0\\ -n\ln\left(\sigma\right)-\frac{1}{\sigma}\sum_{i=1}^{n}\left(x_i-\mu\right), & \xi=0\end{cases}\end{aligned}$$

Il metodo MLE consiste nel risolvere le equazioni:

$$\begin{aligned}\frac{\partial}{\partial\xi}\ln\left[\mathcal{L}\left(\xi,\mu,\sigma\right)\right]&=0\\ \frac{\partial}{\partial\mu}\ln\left[\mathcal{L}\left(\xi,\mu,\sigma\right)\right]&=0\\ \frac{\partial}{\partial\sigma}\ln\left[\mathcal{L}\left(\xi,\mu,\sigma\right)\right]&=0\end{aligned}$$

4.8.3 Generazione di Numeri Casuali

Sia u un numero casuale uniforme in $[0,1]$; applicando il metodo dell'inversione [32], si ottiene:

$$x=\begin{cases}\mu+\frac{\sigma}{\xi}\left(1-u\right)^{\xi}-1, & \xi\neq 0\\ \mu-\sigma\ln\left(1-u\right), & \xi=0\end{cases}$$

dove x risulta distribuito come una $GPD\left(\xi,\mu,\sigma\right)$.

4.9 Phase-Type

Una distribuzione *Phase-Type* (*PH*) rappresenta una famiglia di distribuzioni di probabilità che può essere considerata una generalizzazione della distribuzione Esponenziale, attraverso la composizione di "fasi" esponenziali; più precisamente, rappresenta la distribuzione del tempo all'assorbimento in una Catena di Markov avente uno stato di assorbimento.

Data la vastità dell'argomento, il Cap. 5 è interamente dedicato alla descrizione delle caratteristiche di questa classe di distribuzioni.

4.10 Valori Estremi Generalizzata (GEV)

4.10.1 Caratterizzazione

Definizione 4.10.1 (Distribuzione dei Valori Estremi Generalizzata (GEV))**.** La distribuzione dei *Valori Estremi Generalizzata* (*GEV*) è una famiglia di distribuzioni ricavata nel seguente modo: dato $M_n = \max\{X_1, \ldots, X_n\}$, dove X_i sono variabili aleatorie i.i.d., si supponga che esistano delle costanti di normalizzazione $a_n > 0$ e $b_n \in \mathbb{R}$ tale che:

$$\Pr\left\{\frac{M_n - b_n}{a_n} \leq x\right\} \xrightarrow{\mathcal{D}} F(x)$$

La funzione $F(\dot{})$ rappresenta proprio la CDF della distribuzione dei Valori Estremi Generalizzata, o di una delle sue possibili specializzazioni: distribuzione *Gumbel* (o *Valori Estremi Tipo I*), distribuzione *Frechét* (o *Valori Estremi Tipo II*) e distribuzione *Reverse Weibull* (o *Valori Estremi Tipo III*).

Parametri La distribuzione ha tre parametri:

- **Location** $\mu \in \mathbb{R}$
- **Scale** $\sigma \in \mathbb{R}^+$
- **Shape** $\xi \in \mathbb{R}$

Supporto La distribuzione ha il seguente supporto:

$$\begin{cases} x > \mu - \sigma/\xi, & \xi > 0 \\ x < \mu - \sigma/\xi, & \xi < 0 \\ x \in [-\infty, \infty], \xi = 0 \end{cases}$$

Quando $\xi = 0$, la GEV equivale a una distribuzione *Gumbel* (o *Valori Estremi di Tipo I*) di parametro μ e σ; se $\xi > 0$, la GEV corrisponde a una distribuzione *Frechét* (o *Valori Estremi di Tipo II*) di parametri μ, σ e $\alpha = 1/|\xi|$. Quando $\xi < 0$, la GEV corrisponde alla distribuzione *Reverse Weibull* (o *Valori Estremi di Tipo III*) di parametri μ, σ e $\alpha = 1/|\xi|$.

CDF La funzione di distribuzione cumulativa ha la seguente forma:

$$F(x; \mu, \sigma, \xi) = \exp\left\{ -\left[1 + \xi\left(\frac{x-\mu}{\sigma}\right)\right]_+^{-1/\xi} \right\}$$

con $1 + \xi(x - \mu)/\sigma > 0$, e dove la notazione x_+ sta per $\max\{x, 0\}$.

PDF La funzione di densità ha la seguente forma:

$$f(x; \mu, \sigma, \xi) = \frac{1}{\sigma}\left[1 + \xi\left(\frac{x-\mu}{\sigma}\right)\right]_+^{-1/\xi-1} \exp\left\{ -\left[1 + \xi\left(\frac{x-\mu}{\sigma}\right)\right]_+^{-1/\xi} \right\}$$

con $1 + \xi(x - \mu)/\sigma > 0$, e dove la notazione x_+ sta per $\max\{x, 0\}$.

Quantili La funzione quantile ha la seguente forma:

$$Q(p; \mu\, \sigma, \xi) = \mu - \frac{\sigma}{\xi}\left\{1 - [-\ln(1-p)]^{-\xi}\right\}$$

Nella terminologia dei valori estremi, il quantile $Q(p)$ è detto *livello di ritorno* (*return level*) associato con il periodo di ritorno $1/p$.

4.10.2 Stima dei Parametri

La stima dei parametri prevede l'utilizzo del metodo MLE, cioè della massimizzazione della funzione di verosimiglianza. Si noti, tuttavia, che per $\xi < -1$, il MLE, in generale, non esiste.

4.10.3 Generazione di Numeri Casuali

Sia u un numero casuale uniforme in $[0, 1]$; applicando il metodo dell'inversione [32], si ottiene:

$$x = \mu - \frac{\sigma}{\xi}\left\{1 - [-\ln(1-u)]^{-\xi}\right\}$$

dove x risulta distribuito come una $GEV\ (\mu, \sigma, \xi)$.

4.11 Weibull

La distribuzione *Weibull* [114] viene spesso utilizzata nella teoria dell'affidabilità per modellare il tempo di vita di un oggetto; in questo contesto, il relativo parametro "scale" prende il nome di *Vita Caratteristica*, mentre il parametro "location" rappresenta il minimo tempo di fallimento. Questa distribuzione ha la proprietà di essere molto flessibile, in quanto può riprodurre il comportamento di altre distribuzioni (come la Normale o l'Esponenziale) variando solamente il valore di qualche suo parametro. La Fig. 4.1 e la Fig. 4.2 forniscono un esempio di questa flessibilità: facendo variare solo il parametro "scale" (Fig. 4.1) o solo il parametro "shape" (Fig. 4.2), è possibile ottenere curve di densità molto differenti fra loro; per esempio, si noti, in Fig. 4.1, come, per valori del parametro "scale" strettamente inferiori a 1, le code della distribuzione diventino più lunghe.

4.11.1 Caratterizzazione

Di seguito, il simbolo $\Gamma(\cdot)$ rappresenta la funzione *Gamma* (completa), con $\Gamma(n+1) = n!$, per $n \in \mathbb{N}$.

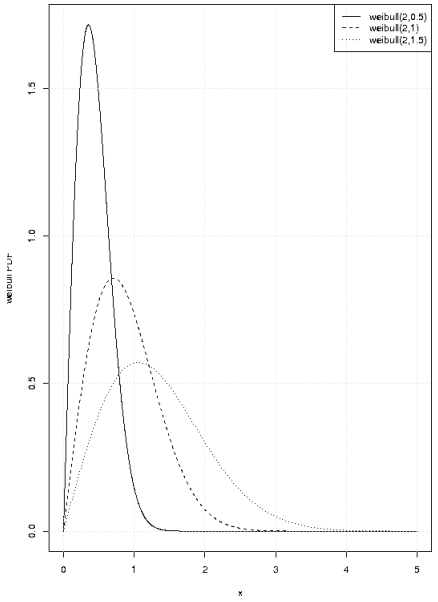


Figura 4.1: Distribuzione Weibull al variare del parametro "scale".

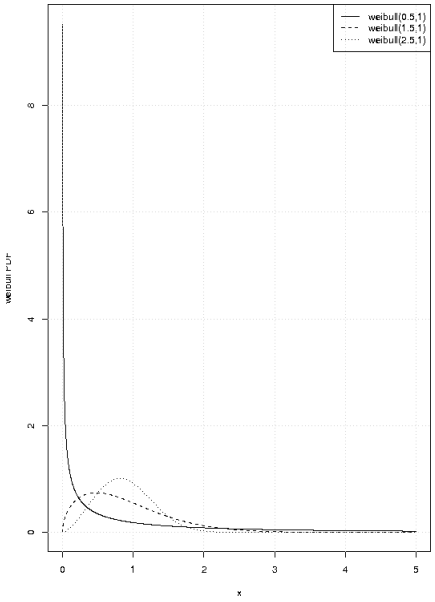


Figura 4.2: Distribuzione Weibull al variare del parametro "shape".

Parametri La distribuzione ha tre parametri:

- **Location** $\mu \in \mathbb{R}$
- **Shape** (o *tail-index*) $\alpha > 0$
- **Scale** $\sigma \in \mathbb{R}^+$

e si indica con *Weibull*(μ, α, σ). Quando $\mu = 0$, la distribuzione viene chiamata *Weibull a 2 parametri* e si indica con *Weibull*(α, σ); una distribuzione Weibull a tre parametri può sempre essere ridotta a una Weibull a due parametri, effettuando uno spostamento di locazione, cioè se X è una Weibull a tre parametri con parametro "location" pari a l, $X - l$ è una Weibull a due parametri. Quando $\mu = 0$ e $\sigma = 1$, la distribuzione viene detta *Weibull Standard*.

Supporto La distribuzione è definita su tutti i valori di x maggiore o uguali del parametro "location", cioè $x \in [\mu, +\infty)$.

CDF La funzione di distribuzione cumulativa ha la seguente forma:

$$F(x;\mu,\alpha,\sigma) = 1 - \exp\left[-\left(\frac{x-\mu}{\sigma}\right)^{\alpha}\right]$$

PDF La funzione di densità di massa ha la seguente forma:

$$f(x;\mu,\alpha,\sigma) = \frac{\alpha}{\sigma}\left(\frac{x-\mu}{\sigma}\right)^{\alpha-1}\exp\left[-\left(\frac{x-\mu}{\sigma}\right)^{\alpha}\right]$$

Quantili La funzione quantile ha la seguente forma:

$$Q\left(p;\mu,\alpha,\sigma\right) = \begin{cases} -\infty, & p < 0 \\ \sigma\left[-\ln\left(1-p\right)\right]^{\alpha} + \mu, & 0 \leq p \leq 1 \\ +\infty, & p > 1 \end{cases}$$

Media La media vale:

$$\mathrm{E}\left[X\right] = \sigma\Gamma\left(1+\frac{1}{\alpha}\right) + \mu$$

Varianza La varianza vale:

$$\mathrm{Var}\left[X\right] = \sigma^2\Gamma\left(1+\frac{2}{\alpha}\right) - \mathrm{E}^2\left[X\right]$$

4.11.2 Stima dei Parametri

Metodo MLE

La funzione di log-verosimiglianza $\ln\left[\mathcal{L}\left(\mu,\alpha,\sigma\right)\right] = \ln f\left(x_1,\ldots,x_n;\mu,\alpha,\sigma\right)$ è:

$$\begin{aligned}\ln\left[\mathcal{L}\left(\mu,\alpha,\sigma\right)\right] &= \ln\left\{\prod_{i=1}^{n}\frac{\alpha}{\sigma}\left(\frac{x_i-\mu}{\sigma}\right)^{\alpha-1}\exp\left[-\left(\frac{x_i-\mu}{\sigma}\right)^{\alpha}\right]\right\} \\ &= n\ln\left(\frac{\alpha}{\sigma}\right) + \sum_{i=1}^{n}\left[\ln\left(\frac{x_i-\mu}{\sigma}\right) - \left(\frac{x_i-\mu}{\sigma}\right)^{\alpha}\right]\end{aligned}$$

Il metodo MLE consiste nel risolvere le equazioni:

$$\frac{\partial}{\partial \mu} \ln \left[\mathcal{L} \left(\mu, \alpha, \sigma \right) \right] = 0$$
$$\frac{\partial}{\partial \alpha} \ln \left[\mathcal{L} \left(\mu, \alpha, \sigma \right) \right] = 0$$
$$\frac{\partial}{\partial \sigma} \ln \left[\mathcal{L} \left(\mu, \alpha, \sigma \right) \right] = 0$$

La stima del parametro "location" può risultare problematica in quanto può causare irregolarità nell'ottimizzazione della funzione di verosimiglianza; per queste ragioni, spesso si preferisce utilizzare la versione della Weibull a due parametri (cioè, con il parametro "location" uguale a zero) o utilizzare delle euristiche, come impostare il valore del parametro "location" al minimo valore delle osservazioni che compongono il campione sotto esame.

4.11.3 Generazione di Numeri Casuali

Sia u un numero casuale uniforme in $[0, 1]$; applicando il metodo dell'inversione [32], si ottiene:

$$x = \sigma \left(- \ln \left(1 - u \right) \right)^{\frac{1}{\alpha}} + \mu$$

dove x risulta distribuito come una $Weibull \left(\mu, \sigma \right)$.

Capitolo 5

Distribuzioni Phase-Type

Questo capitolo presenta le distribuzioni *Phase-Type* [83, 84], un'importante classe di distribuzioni di probabilità che può essere considerata, in modo informale, una generalizzazione della distribuzione esponenziale, realizzata mediante la composizione di "fasi" esponenziali.

La distribuzione esponenziale è stata, ed è ancora, largamente utilizzata nella costruzione di modelli per l'analisi delle prestazioni; la ragione principale è sicuramente l'essere estremamente facile da trattare dal punto di vista analitico; questa semplicità è in gran parte dovuta alla proprietà di assenza di memoria E.1.2, per la quale si ottiene l'indipendenza condizionale del futuro dal passato, dato il presente. Tuttavia, si è notato che, in molte situazioni pratiche (ad es. si veda [66, 93]), le semplificazioni introdotte dall'utilizzo della distribuzione esponenziale hanno reso inadeguati i modelli costruiti. Le distribuzioni Phase-Type riescono a offrire i vantaggi di una semplice trattabilità matematica uniti a una buona flessibilità e generalità.

Le distribuzioni Phase-Type devono al loro nome al fatto che possono essere considerate come l'attraversamento di una serie di *fasi* esponenziali; più precisamente, una distribuzione Phase-Type è definita come la distribuzione del tempo all'assorbimento in una *Catena di Markov* con uno stato di assorbimento: ogni stato *transiente* della catena di Markov rappresenta una fase (*phase*) e il tempo di soggiorno in ciascuno stato segue una distribuzione esponenziale.

Uno dei pionieri di questo modello di distribuzioni fu sicuramente Erlang

[36], la cui omonima distribuzione di probabilità rappresenta una caso particolare di distribuzione Phase-Type; Erlang, durante i suoi studi sul traffico telefonico e in particolare sul numero di chiamate telefoniche simultanee che possono essere effettuate a un "call-center", fu uno dei primi a estendere la famiglia di distribuzioni esponenziali introducendo il concetto di "stadi" (*stages*) [1]. Un importante contributo venne successivamente dato da Cox, nel 1955 [25], il quale generalizzò i concetti proposti da Erlang. Il lavoro di Neuts [84], tracciò invece le basi per l'approccio teorico moderno. Per alcuni contributi recenti si veda, ad esempio, [12, 17, 44, 88].

Il capitolo è diviso nel seguente modo: nella sezione §5.1 vengono date le definizioni formali di alcune classi di distribuzioni Phase-Type e vengono illustrate alcune delle loro proprietà principali. Nella sezione §5.2 vengono presentati alcuni esempi di distribuzioni Phase-Type. La sezione §5.4 descrive alcuni metodi di stima dei parametri, mentre la sezione §5.5 descrive la tecnica utilizzata per la generazione di numeri casuali distribuiti secondo una distribuzione Phase-Type.

5.1 Caratterizzazione

5.1.1 Definizioni

Definizione 5.1.1. Una distribuzione *Phase-Type* (*PH*) rappresenta la distribuzione del tempo all'assorbimento in una MC; essa è caratterizzata da un vettore $\vec{\alpha}$ e da una matrice $\mathbf{T}$, e si dice che $PH(\vec{\alpha}, \mathbf{T})$ è la *rappresentazione* della distribuzione Phase-Type PH.

In Fig. 5.1 è mostrato un esempio di distribuzione PH a tre fasi.

Definizione 5.1.2 (Distribuzione Phase-Type Discreta (DPH))**.** Si consideri una DTMC §E.2 $\{X_n | n = 0, 1, \ldots\}$ omogenea costituita da $m+1$ stati $\{1, \ldots, m+1\}$, di cui i primi m sono transienti e l'ultimo è assorbente [2]. Si supponga che il

[1] Nel lavoro di Erlang non si parla esplicitamente di "stadi" ma piuttosto di "linee".

[2] Alcuni autori preferiscono considerare gli stati $\{0, 1, \ldots, m\}$ e associare lo stato assorbente allo stato 0.

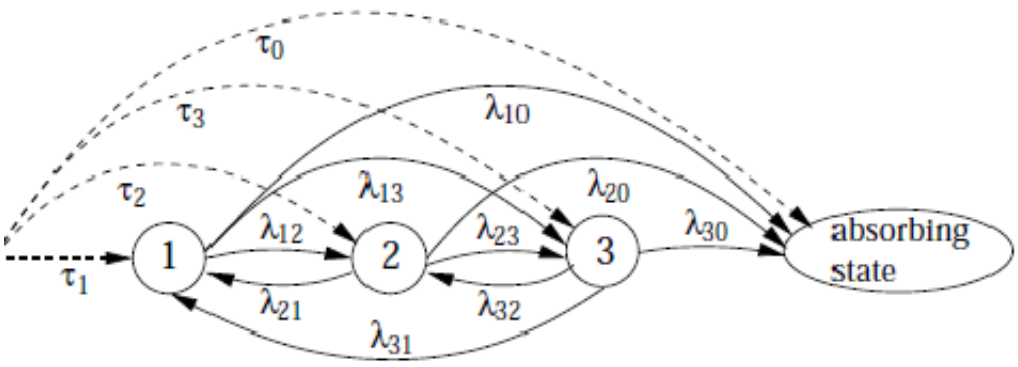


Figura 5.1: Distribuzione PH a tre fasi.

vettore delle probabilità di stato iniziali sia pari a $\vec{\mathbf{p_0}} = [\vec{\alpha}, \alpha_{m+1}]$, e la matrice delle probabilità di transizione a un passo $\mathbf{P}$ sia data da:

$$\mathbf{P} = \begin{bmatrix} \mathbf{T} & \vec{\eta}^T \\ \vec{\mathbf{0}} & 1 \end{bmatrix}$$

dove:

- $\vec{\alpha}$ è un vettore (riga) di dimensione m;
- $\mathbf{T}$ è una matrice quadrata di ordine m;
- η^T è un vettore (colonna) di dimensione m.

Sia $\tau = \inf\{n \geq 0 | X_n = m + 1\}$ la variabile casuale del tempo (numero di passi) all'assorbimento della DTMC. La distribuzione di τ è chiamata distribuzione *Phase-Type Discreta* (*DPH*) di parametri $(\vec{\alpha}, \mathbf{T})$ e si indica $\tau \sim \mathrm{DPH}(\vec{\alpha}, \mathbf{T})$.

Per le proprietà delle MC (matrice $\mathbf{P}$ stocastica e vettore $[\vec{\alpha}, \alpha_{m+1}]$ distribuzione di probabilità) si ha:

$$\vec{\eta}^T = \vec{\mathbf{1}}^T - \mathbf{T}\vec{\mathbf{1}}^T$$
$$\alpha_{m+1} = 1 - \vec{\alpha}\vec{\mathbf{1}}^T$$

In generale $\alpha_{m+1} \geq 0$; quando $\alpha_{m+1} = 0$, il tempo all'assorbimento sarà uguale a zero.

Proposizione 5.1.1. *Data una variabile casuale $\tau \sim \mathrm{DPH}(\vec{\alpha}, \mathbf{T})$, si può dimostrare che:*

- *Funzione di massa di probabilità:*

$$p_\tau(k) = \Pr\{\tau = k\} = \vec{\alpha}\vec{\mathbf{T}}^{k-1}\vec{\eta}^T, \quad k = 1, 2, \ldots$$

- *Funzione di distribuzione:*

$$F_\tau(k) = \Pr\{\tau \leq k\} = 1 - \vec{\alpha}\mathbf{T}^k\vec{\mathbf{1}}^T, \quad k = 0, 1, 2, \ldots$$

- *Momenti "fattoriali":*

$$\gamma_k = E\left[\tau\left(\tau-1\right)\ldots\left(\tau-k+1\right)\right] = k!\,\vec{\alpha}\left(\mathbf{I}-\mathbf{T}\right)^{-k}\mathbf{T}^{k-1}\vec{\mathbf{1}}^T$$

- *Trasformata z §D.2:*

$$\mathcal{F}\left(z\right) = E\left[z^{\tau}\right] = z\vec{\alpha}\left(\mathbf{I}-z\mathbf{T}\right)^{-1}\vec{\eta}^T$$

In maniera simile a quanto effetuato per le distribuzioni DPH, è possibile definire una distribuzione Phase-Type Continua a partire da una CTMC.

Definizione 5.1.3 (Distribuzione Phase-Type Continua (CPH))**.** Si consideri una CTMC §E.3 $\{X\left(t\right)|t\geq 0\}$ omogenea costituita da $m+1$ stati $\{1,\ldots,m+1\}$, di cui i primi m sono transienti e l'ultimo è assorbente. Si supponga che il vettore delle probabilità di stato iniziali sia pari a $\vec{\mathbf{p_0}} = [\vec{\alpha},\alpha_{m+1}]$, e il generatore infinitesimale $\mathbf{Q}$ sia dato da:

$$\mathbf{Q} = \begin{bmatrix}\mathbf{T} & \vec{\eta}^T\\ \vec{\mathbf{0}} & 0\end{bmatrix}$$

dove:

- $\vec{\alpha}$ è un vettore (riga) di dimensione m;
- $\mathbf{T}$ è una matrice quadrata di ordine m;
- η^T è un vettore (colonna) di dimensione m.

Sia $\tau = \inf\{t\geq 0|X\left(t\right)=m+1\}$ la variabile casuale del tempo all'assorbimento della CTMC. La distribuzione di τ è chiamata distribuzione *Phase-Type Continua* (*CPH*) di parametri $(\vec{\alpha},\mathbf{T})$ e si indica $\tau\sim\mathrm{CPH}\left(\vec{\alpha},\mathbf{T}\right)$.

Per le proprietà delle MC (matrice $\mathbf{Q}$ con righe che sommano a zero e vettore $[\vec{\alpha},\alpha_{m+1}]$ distribuzione di probabilità) si ha:

$$\vec{\eta}^T = -\mathbf{T}\vec{\mathbf{1}}^T$$
$$\alpha_{m+1} = 1-\vec{\alpha}\vec{\mathbf{1}}^T$$

Proposizione 5.1.2. *Data una variabile casuale* $\tau \sim \mathrm{CPH}\left(\vec{\alpha}, \mathbf{T}\right)$*, si può dimostrare che:*

- *Funzione di densità di probabilità:*

$$f_{\tau}\left(t\right) = \vec{\alpha}\exp\left(\mathbf{T}t\right)\vec{\eta}^{T}$$

- *Funzione di distribuzione cumulativa:*

$$F_{\tau}\left(t\right) = \Pr\left\{\tau \leq t\right\} = 1 - \vec{\alpha}\exp\left(\mathbf{T}t\right)\vec{\mathbf{1}}^{T}$$

- *Momenti "potenza":*

$$\mu_{k} = E\left[\tau^{k}\right] = \left(-1\right)^{k} k!\vec{\alpha}\mathbf{T}^{-k}\vec{\mathbf{1}}^{T} = k!\vec{\alpha}\left(-\mathbf{T}\right)^{-k}\vec{\mathbf{1}}^{T}$$

- *Trasformata di* Laplace-Stieltjes *(*LST*) §D.3 di* $F_{\tau}\left(\cdot\right)$*:*

$$\left\{\mathcal{L}^{*}F\right\}\left(s\right) = F^{*}\left(s\right) = \alpha_{m+1} + \vec{\alpha}\left(s\mathbf{I} - \mathbf{T}\right)^{-1}\vec{\eta}^{T}, \quad \Re\left(s\right) \geq 0$$

dove $\exp\left(\mathbf{A}\right)$ *rappresenta l'operazione di* esponenziale della matrice $\mathbf{A}$ *Cap. F.*

Inoltre, per una generica distribuzione $\tau \sim \mathrm{PH}(\vec{\alpha}, \mathbf{T})$ valgono i seguenti fatti:

- si tratta di una distribuzione con supporto nell'intervallo $[0, \infty)$;
- la coppia $(\vec{\alpha}, \mathbf{T})$ è detta *rappresentazione* di τ;
- il vettore $\vec{\alpha}$ prende il nome di *vettore delle probabilità iniziali* o *vettore d'entrata* (*entrance vector*) e rappresenta il vettore delle probabilità con cui ogni stato può essere quello iniziale;
- la matrice $\mathbf{T}$ viene detta *generatore* della distribuzione PH e rappresenta la matrice dei tassi (o delle probabilità, nel caso di una distribuzione DPH) di transizione fra gli stati transienti;

- il vettore $\vec{\eta}^T$ prende il nome di *vettore delle probabilità d'assorbimento* o *vettore d'uscita* (*exit vector*) e rappresenta il vettore dei tassi (o delle probabilità, nel caso di una distribuzione DPH) di transizione dagli stati transienti a quello assorbente;

- la dimensione m di $\mathbf{T}$ è detta *ordine* o *numero di fasi* della distribuzione PH;

- gli stati transienti $\{1, \ldots, m\}$ della MC associata sono detti *fasi* della distribuzione PH.

Due importanti sottoinsiemi delle distribuzioni PH sono l'insieme delle distribuzioni PH *Acicliche* e quello delle distribuzioni PH *Coxian*.

Definizione 5.1.4 (Distribuzione PH Acicliche (APH))**.** Una distribuzione PH *Aciclica* (*APH*) di parametri $(\vec{\alpha}, \mathbf{T})$ è una distribuzione PH in cui i suoi stati possono essere riordinati in modo tale che $t_{ij} = 0$, per ogni $i > j$, cioè la matrice $\mathbf{T}$ è triangolare superiore.

In Fig. 5.2 è mostrato un esempio di distribuzione PH Acliclica a tre fasi.

Definizione 5.1.5 (Distribuzione PH Coxian)**.** Una distribuzione PH *Coxian* di parametri $(\vec{\alpha}, \mathbf{T})$ a m fasi è una distribuzione APH a m fasi in cui $\alpha_i = 0$, per $2 \leq i \leq m$, e $t_{ij} = 0$, per $i + 1 < j \leq m$. La distribuzione si dice PH *Coxian*$^+$ a m fasi se è una distribuzione PH Coxian in cui $\alpha_1 = 0$.

In Fig. 5.3 è mostrato un esempio di distribuzione PH Coxian a tre fasi.

5.1.2 Alcune Proprietà

In questa sezione vengono presentate alcune delle più importanti proprietà delle distribuzioni PH; per maggiori informazioni e altre proprietà si veda, ad esempio, [65, 82, 84].

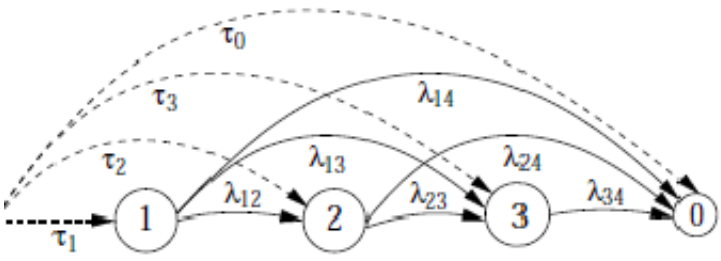


Figura 5.2: Distribuzione APH a tre fasi.

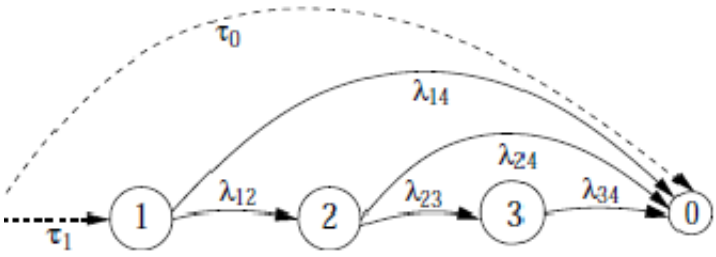


Figura 5.3: Distribuzione PH Coxian a tre fasi.

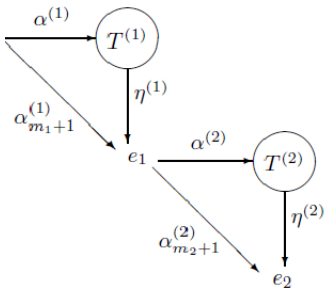


Figura 5.4: Convoluzione di due distribuzioni PH.

Proposizione 5.1.3 ($\mathcal{PH}$ è denso)**.** *L'insieme $\mathcal{PH}$, cioè l'insieme delle distribuzioni PH, è* denso [3] *rispetto all'insieme delle distribuzioni non negative (ossia quelle con supporto in $[0, \infty)$).*

Un'importante proprietà delle distribuzioni PH è quella di *chiusura* rispetto ad alcune tipologie di composizione; di seguito vengono esposte tre proprietà di chiusura rispetto alla *convoluzione* (Fig. 5.4), alla *mistura* (Fig. 5.5) e alla composizione di *Kronecker* (Fig. 5.6).

[3] Un insieme $A \subseteq B$ è *denso* in B se l'unico sottoinsieme *chiuso* di B (cioè contenente tutti i suoi *punti di frontiera*) che comprende A è B stesso; in altri termini, A è chiuso in B se la *chiusura* di A è B.

Proposizione 5.1.4 (Convoluzione di distribuzioni PH)**.** *Si considerino due distribuzioni PH:*

$$X_1 \sim \mathrm{PH}\left(\vec{\alpha}^{(1)}, \mathbf{T}^{(1)}\right) \textit{ di ordine } m_1$$
$$X_2 \sim \mathrm{PH}\left(\vec{\alpha}^{(2)}, \mathbf{T}^{(2)}\right) \textit{ di ordine } m_2$$

La convoluzione X *di* X_1 *e* X_2*,* $X = X_1 + X_2$ *è una distribuzione* $\mathrm{PH}\left(\vec{\alpha}, \mathbf{T}\right)$ *di ordine* $m = m_1 + m_2$ *tale che:*

$$\vec{\alpha} = \left[\vec{\alpha}^{(1)}, \alpha^{(1)}_{m_1+1}\vec{\alpha}^{(2)}\right]$$
$$\mathbf{T} = \begin{bmatrix} \mathbf{T}^{(1)} & \vec{\eta}^{(1)}\vec{\alpha}^{(2)} \\ \mathbf{0} & \mathbf{T}^{(2)} \end{bmatrix}$$

dove $\alpha^{(1)}_{m+1} = 1 - \vec{\alpha}^{(1)}\vec{\mathbf{1}}^T$ *e* $\vec{\eta}^{(1)} = -\mathbf{T}^{(1)}\vec{\mathbf{1}}^T$*, in caso di CPH, o* $\vec{\eta}^{(1)} = \vec{\mathbf{1}}^T - \mathbf{T}^{(1)}\vec{\mathbf{1}}^T$*, in caso di DPH.*

La Fig. 5.4 mostra come la convoluzione di due distribuzioni PH X_1 e X_2 possa essere vista come una distribuzione PH X rappresentante la somma delle due distribuzioni PH, ossia il tempo totale all'assorbimento nello stato di assorbimento della seconda PH, passando prima dallo stato di assorbimento della prima PH. Si noti che nella MC associata alla distribuzione PH X risultante dalla convoluzione, l'unico stato di assorbimento è quello di X_2; quello di X_1 viene solo utilizzato come costruzione analitica per facilitare la spiegazione; una volta che X_1 raggiunge lo stato di assorbimento (come stato iniziale o per una transizione da uno stato transiente di X_1 al suo stato assorbente), la MC associata a X passa immediatamente in uno degli stati transienti di X_2 con probabilità $\alpha^{(2)}_i$, con $1 \leq i \leq m_2$ o, eventualmente, nel suo stato assorbente. La costruizione del generatore $\mathbf{T}$ di X segue direttamente da queste ultime considerazioni: tra gli stati transienti delle due distribuzioni PH non possono avvenire transizioni dirette e l'unico modo per passare da una all'altra distribuzione PH è attraverso lo stato assorbente di X_1.

Proposizione 5.1.5 (Mistura di distribuzioni PH)**.** *Si considerino due distribuzio-*

ni PH:

$$X_1 \sim \mathrm{PH}\left(\vec{\alpha}^{(1)}, \mathbf{T}^{(1)}\right) \textit{ di ordine } m_1$$
$$X_2 \sim \mathrm{PH}\left(\vec{\alpha}^{(2)}, \mathbf{T}^{(2)}\right) \textit{ di ordine } m_2$$

e un numero reale $p \in [0, 1]$. *La* mistura X *di* X_1 *e* X_2, $X = pX_1 + (1-p)\,X_2$ *è una distribuzione* $\mathrm{PH}\left(\vec{\alpha}, \mathbf{T}\right)$ *di ordine* $m = m_1 + m_2$ *tale che:*

$$\vec{\alpha} = \left[p\vec{\alpha}^{(1)}, (1-p)\,\vec{\alpha}^{(2)}\right]$$
$$\mathbf{T} = \begin{bmatrix} \mathbf{T}^{(1)} & \mathbf{0} \\ \mathbf{0} & \mathbf{T}^{(2)} \end{bmatrix}$$

La Fig. 5.5 illustra come la mistura di due distribuzioni PH X_1 e X_2 possa essere considerata come una distribuzione PH X uguale a X_1, con probabilità p, o a X_2, con probabilità $(1-p)$; in effetti, non esiste nessuna transizione tra gli stati transienti delle due distribuzioni. Si noti, infine, come lo stato assorbente possa essere lo stato iniziale con probabilità $p\alpha^{(1)}_{m_1+1} + (1-p)\,\alpha^{(2)}_{m_2+1}$.

Proposizione 5.1.6 (Composizione di Kronecker)**.** *Si considerino due distribuzioni PH:*

$$X_1 \sim \mathrm{PH}\left(\vec{\alpha}^{(1)}, \mathbf{T}^{(1)}\right) \textit{ di ordine } m_1$$
$$X_2 \sim \mathrm{PH}\left(\vec{\alpha}^{(2)}, \mathbf{T}^{(2)}\right) \textit{ di ordine } m_2$$

La composizione di Kronecker X *di* X_1 *e* X_2, $X = \min\{X_1, X_2\}$ *è una distribuzione* $\mathrm{PH}\left(\vec{\alpha}, \mathbf{T}\right)$ *di ordine* $m = m_1 \times m_2$ *tale che:*

$$\vec{\alpha} = \vec{\alpha}^{(1)} \otimes \vec{\alpha}^{(2)}$$
$$\mathbf{T} = \mathbf{T}^{(1)} \oplus \mathbf{T}^{(2)}$$

dove $\otimes$ *e* $\oplus$ *rappresentano, rispettivamente, il* prodotto *e la* somma *di* Kronecker *Cap. G.*

La Fig. 5.6 mostra come la composizione di Kronecker di due distribuzioni PH possa essere vista come una sovrapposizione delle due distribuzioni, cioè

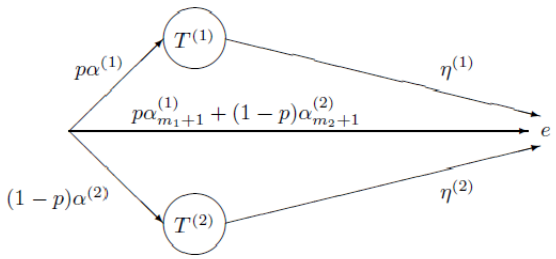


Figura 5.5: Mistura di due distribuzioni PH.

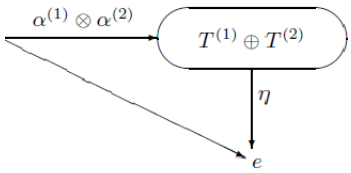


Figura 5.6: Composizione di Kronecker di due distribuzioni PH.

come la distribuzione del minimo tempo all'assorbimento tra le MC $\mathcal{X}^{(1)}$ (la MC associata alla prima distribuzione PH) e $\mathcal{X}^{(2)}$ (la MC associata alla seconda distribuzione PH), che procedono in modo concorrente. Il vettore d'uscita $\vec{\eta}$ è dato da $\vec{\eta} = \vec{\eta}^{(1)} + \vec{\eta}^{(2)}$, dove $\vec{\eta}^{(1)}$ è il vettore d'uscita della prima distribuzione PH, mentre $\vec{\eta}^{(2)}$ è il vettore d'uscita della seconda. L'utilizzo di composizioni di Kronecker risulta particolarmente utile non solo per le distribuzioni PH ma, in generale, per l'analisi di qualsiasi MC, soprattutto quando questa è costituita da un elevato numero di stati: in questi casi infatti l'enorme dimensione del generatore (o della matrice delle probabilità di transizione) della catena rappresenta un limite pratico per la soluzione della catena, a causa degli eccessivi requisiti di memoria richiesti; anche impiegando algoritmi iterativi, al

posto del tradizionale metodo di eliminazione di Gauss [98], per la soluzione dei sistemi lineari coinvolti, i problemi, dovuti ai limiti delle risorse finite, permangono. Tecniche piuttosto recenti (ad es. si veda [18]) hanno messo in luce come si possa evitare di memorizzare e generare in modo esplicito il generatore; un'idea è, per esempio, quella di rappresentare il generatore $\mathbf{Q}$ con una sottomatrice di una matrice $\mathbf{W}$ ottenuta come somma di prodotti di Kronecker tra matrici più piccole, la cui natura dipende dal formalismo ad alto livello da cui si è ricavata la catena (ad es. una rete di Petri).

5.2 Esempi di Distribuzioni PH

Di seguito vengono forniti alcuni esempi di distribuzioni di probabilità appartenenti alla famiglia delle distribuzioni PH.

Esempio 5.2.1 (Distribuzione Geometrica)**.** La distribuzione *Geometrica* è il più semplice esempio di DPH. Se $X \sim \mathrm{Geo}(\beta)$, allora $X \sim \mathrm{DPH}(\vec{\alpha}, \mathbf{T})$ tale che:

$$\vec{\alpha} = [1]$$
$$\mathbf{T} = \Big[1 - \beta\Big]$$

cioè X rappresenta il numero di passi all'assorbimento di una DTMC il cui vettore delle probabilità iniziali è $[1, 0]$ e la matrice delle probabilità di transizione a un passo è:

$$\mathbf{P} = \begin{bmatrix} 1-\beta & \beta \\ 0 & 1 \end{bmatrix}$$

Indicando con 1 lo stato transiente e con 2 quello assorbente, si ha che al passo iniziale, lo stato del sistema è lo stato 1 (stato iniziale); dopo una permanenza in tale stato per un numero di passi distribuito come una distribuzione geometrica di parametro β, il sistema passa nello stato di assorbimento 2. Ne risulta che il numero di passi all'assorbimento segue una una distribuzione geometrica di parametro β.

Esempio 5.2.2 (Distribuzione Esponenziale)**.** La distribuzione *Esponenziale* è il

più semplice esempio di CPH. Se $X \sim \text{Exp}(\lambda)$, allora $X \sim \text{CPH}(\vec{\alpha}, \mathbf{T})$ tale che:

$$\vec{\alpha} = [1]$$
$$\mathbf{T} = \begin{bmatrix} -\lambda \end{bmatrix}$$

cioè X rappresenta il tempo all'assorbimento di una CTMC il cui vettore delle probabilità iniziali è $[1, 0]$ e il generatore infinitesimale è:

$$\mathbf{Q} = \begin{bmatrix} -\lambda & \lambda \\ 0 & 0 \end{bmatrix}$$

La Fig. 5.7 mostra una possibile rappresentazione grafica di una esponenziale di parametro λ vista come PH a 1 fase. Indicando con 1 lo stato transiente e con 2 quello assorbente, si ha che all'istante iniziale, lo stato del sistema è lo stato 1 (stato iniziale); dopo una permanenza in tale stato per un tempo distribuito come una distribuzione esponenziale di parametro λ, il sistema passa nello stato di assorbimento 2. Ne risulta che il tempo all'assorbimento segue una una distribuzione esponenziale di parametro λ.

Esempio 5.2.3 (Distribuzione Erlang)**.** La distribuzione di probabilità *Erlang* fu introdotta da A. K. Erlang per esaminare il numero di chiamate telefoniche simulatenee dirette agli operatori di un "call-center" [36]; una distribuzione Erlang a n *gradi di libertà*, anche detta a n *stadi*, è la distribuzione della somma di n variabili Esponenziali di parametro λ:

$$X \sim \text{Erlang}(n, \lambda) \equiv Y_1 + \cdots + Y_n, \quad Y_i \sim \text{Exp}(\lambda) \text{ e } i = 1, 2, \ldots, n$$

Se $X \sim \text{Erlang}(n, \lambda)$, allora $X \sim \text{CPH}\,(\vec{\alpha}, \mathbf{T})$ tale che:

$$\vec{\alpha} = [1, 0, \ldots, 0]$$
$$\mathbf{T} = \begin{bmatrix} -\lambda & \lambda & 0 & \cdots & \cdots & 0 \\ 0 & -\lambda & \lambda & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & \cdots & \cdots & 0 & -\lambda \end{bmatrix}$$

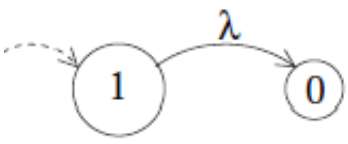


Figura 5.7: Esponenziale di parametro λ.

cioè X rappresenta il tempo all'assorbimento di una CTMC il cui vettore delle probabilità iniziali è $[1, 0, \ldots, 0]$ e il generatore infinitesimale è:

$$\mathbf{Q} = \begin{bmatrix} -\lambda & \lambda & 0 & \cdots & \cdots & \cdots & 0 \\ 0 & -\lambda & \lambda & 0 & \cdots & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & \cdots & \cdots & 0 & -\lambda & \lambda \\ 0 & \cdots & \cdots & \cdots & \cdots & \cdots & 0 \end{bmatrix}$$

La Fig. 5.8 mostra una possibile rappresentazione grafica di una distribuzione Erlang a 2 fasi e con parametro λ.

Esempio 5.2.4 (Distribuzione Iper-Esponenziale)**.** Una distribuzione *Iper-Esponenziale* può essere definita come una mistura di un numero finito di distribuzioni Esponenziali con parametri, in generale, differenti:

$$X \sim \text{HyperExp}\left(\lambda_1, \ldots, \lambda_n, p_1, \ldots, p_n\right) \Leftrightarrow f_X\left(x\right) = \sum_{i=1}^{n} p_i f_{Y_i}\left(x\right)$$

con

$$\sum_{i=1}^{n} p_i = 1 \quad \text{e} \quad Y_i \sim \text{Exp}\left(\lambda_i\right), \quad i = 1, 2, \ldots, n$$

Se $X \sim \text{HyperExp}\left(\lambda_1, \ldots, \lambda_n, p_1, \ldots, p_n\right)$, allora $X \sim \text{CPH}\left(\vec{\alpha}, \mathbf{T}\right)$ tale che:

$$\vec{\alpha} = [p_1, \ldots, p_n]$$

$$\mathbf{T} = \begin{bmatrix} -\lambda_1 & 0 & \cdots & \cdots & 0 \\ 0 & -\lambda_2 & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & \cdots & 0 & -\lambda_n \end{bmatrix}$$

cioè X rappresenta il tempo all'assorbimento di una CTMC il cui vettore delle

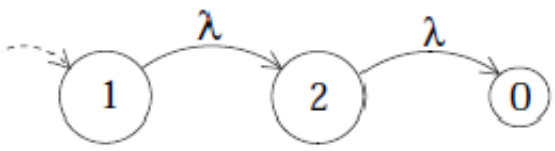


Figura 5.8: Distribuzione Erlang a 2 fasi e di parametro λ.

probabilità iniziali è $[p_1, \ldots, p_n, 0]$ e il generatore infinitesimale è:

$$\mathbf{Q} = \begin{bmatrix} -\lambda_1 & 0 & \cdots & \cdots & 0 & \lambda_1 \\ 0 & -\lambda_2 & 0 & \cdots & 0 & \lambda_2 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & \cdots & 0 & -\lambda_n & \lambda_n \\ 0 & \cdots & \cdots & \cdots & \cdots & 0 \end{bmatrix}$$

La Fig. 5.9 mostra una possibile rappresentazione grafica di una distribuzione Iper-Esponenziale a 2 fasi e con parametri λ_1, λ_2, p_1 e p_2.

Esempio 5.2.5 (Distribuzione Ipo-Esponenziale)**.** Una distribuzione *Ipo-Esponenziale*, o *Erlang generalizzata*, a n *gradi di libertà*, o a n *stadi*, è la distribuzione della somma di n variabili Esponenziali con parametri, in generale, differenti:

$$X \sim \text{HypoExp}\,(\lambda_1, \ldots, \lambda_n) \equiv Y_1 + \cdots + Y_n, \quad Y_i \sim \text{Exp}\,(\lambda_i) \text{ e } i = 1, 2, \ldots, n$$

Se $X \sim \text{HypoExp}(\lambda_1, \ldots, \lambda_n)$, allora $X \sim \text{CPH}\,(\vec{\alpha}, \mathbf{T})$ tale che:

$$\vec{\alpha} = [1, 0, \ldots, 0]$$
$$\mathbf{T} = \begin{bmatrix} -\lambda_1 & \lambda_1 & 0 & \cdots & \cdots & 0 \\ 0 & -\lambda_2 & \lambda_2 & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & \cdots & \cdots & 0 & -\lambda_n \end{bmatrix}$$

cioè X rappresenta il tempo all'assorbimento di una CTMC il cui vettore delle probabilità iniziali è $[1, 0, \ldots, 0]$ e il generatore infinitesimale è:

$$\mathbf{Q} = \begin{bmatrix} -\lambda_1 & \lambda_1 & 0 & \cdots & \cdots & \cdots & 0 \\ 0 & -\lambda_2 & \lambda_2 & 0 & \cdots & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & \cdots & \cdots & 0 & -\lambda_n & \lambda_n \\ 0 & \cdots & \cdots & \cdots & \cdots & \cdots & 0 \end{bmatrix}$$

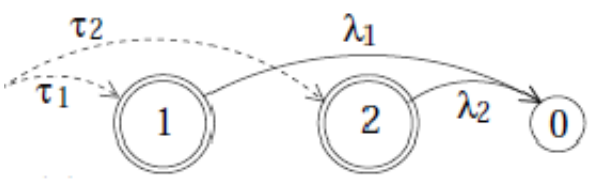


Figura 5.9: Distribuzione Iper-Esponenziale a 2 fasi e di parametri λ_1, λ_2, p_1 e p_2.

La Fig. 5.10 mostra una possibile rappresentazione grafica di una distribuzione Ipo-Esponenziale a n fasi e con parametri $\lambda_1, \ldots, \lambda_n$.

Esempio 5.2.6 (Distribuzione Coxian)**.** Come definito in 5.1.5, una distribuzione *Coxian*, o *di Cox*, è una distribuzione PH aciclica composta da più fasi sequenziali, non necessariamente identiche, in cui, per ogni fase, vi è una probabilità di passare direttamente nello stato di assorbimento.
Se $X \sim \mathrm{Cox}(\lambda_1, \ldots, \lambda_n, p_1, \ldots, p_n)$, allora $X \sim \mathrm{CPH}(\vec{\alpha}, \mathbf{T})$ tale che:

$$\vec{\alpha} = [1, \ldots, 0]$$

$$\mathbf{T} = \begin{bmatrix} -\lambda_1 & p_1\lambda_1 & \cdots & \cdots & 0 \\ 0 & -\lambda_2 & p_2\lambda_2 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & \cdots & 0 & -\lambda_n \end{bmatrix}$$

cioè X rappresenta il tempo all'assorbimento di una CTMC il cui vettore delle probabilità iniziali è $[1, \ldots, 0]$ e il generatore infinitesimale è:

$$\mathbf{Q} = \begin{bmatrix} -\lambda_1 & p_1\lambda_1 & \cdots & \cdots & 0 & (1-p_1)\lambda_1 \\ 0 & -\lambda_2 & p_2\lambda_2 & \cdots & 0 & (1-p_2)\lambda_2 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & \cdots & 0 & -\lambda_n & \lambda_n \\ 0 & \cdots & \cdots & \cdots & \cdots & 0 \end{bmatrix}$$

La Fig. 5.3 mostra una possibile rappresentazione grafica di una distribuzione Coxian a 3 fasi e con parametri λ_1, λ_2, λ_3, p_1, p_2 e p_3.

5.3 Generazione dei Quantili

La forma analitica della funzione quantile si ricava invertendo l'espressione della funzione di distribuzione; nel caso di una distribuzione PH, ciò implica il calcolo dell'inversa di diverse matrici. Per esempio, per una distribuzione

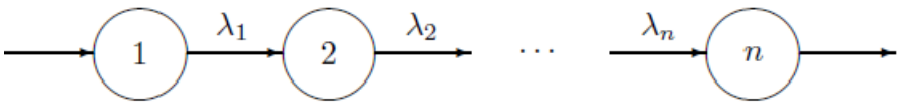


Figura 5.10: Distribuzione Ipo-Esponenziale a n fasi e di parametri $\lambda_1, \ldots, \lambda_n$.

CPH τ risulta:

$$\begin{aligned}
p &= F_\tau(t) \\
p &= 1 - \vec{\alpha} \exp(\mathbf{T}t)\, \vec{\mathbf{1}}^T \\
1 - p &= \vec{\alpha} \exp(\mathbf{T}t)\, \vec{\mathbf{1}}^T \\
\vec{\alpha}^H (1-p) &= \vec{\alpha}^H \vec{\alpha} \exp(\mathbf{T}t)\, \vec{\mathbf{1}}^T \\
\vec{\alpha}^H (1-p) &= \mathbf{A} \exp(\mathbf{T}t)\, \vec{\mathbf{1}}^T \\
\mathbf{A}^{-1} \vec{\alpha}^H (1-p) &= \exp(\mathbf{T}t)\, \vec{\mathbf{1}}^T \\
\mathbf{A}^{-1} \vec{\alpha}^H (1-p)\, \vec{\mathbf{1}} &= \exp(\mathbf{T}t)\, \mathbf{J} \\
\mathbf{A}^{-1} \vec{\alpha}^H (1-p)\, \vec{\mathbf{1}} \mathbf{J}^{-1} &= \exp(\mathbf{T}t) \\
\ln\left[\mathbf{A}^{-1} \vec{\alpha}^H (1-p)\, \vec{\mathbf{1}} \mathbf{J}^{-1}\right] &= \mathbf{T}t \\
\mathbf{T}^{-1} \ln\left[\mathbf{A}^{-1} \vec{\alpha}^H (1-p)\, \vec{\mathbf{1}} \mathbf{J}^{-1}\right] &= \mathbf{I}t
\end{aligned} \tag{5.3.1}$$

dove $\mathbf{A} = \vec{\alpha}^H \vec{\alpha}$ e $\mathbf{J}$ è una matrice con tutti gli elementi uguali a uno. Mentre l'esistenza dell'inversa della matrice $\mathbf{T}$ è assicurata dal fatto che gli stati $1, \ldots, m$ della MC sottostante sono transienti se e solo se la matrice T è non singolare [84], l'inversa della matrice $\mathbf{A}$ non è detto che esista e di certo non esiste quella della matrice $\mathbf{J}$ per ordine maggiori o uguali a due.

Per tale motivo, il calcolo dei quantili viene ottenuto tramite approssimazioni successive. Data una distribuzione di probabilità PH τ con funzione di distribuzione $F_\tau(\cdot)$, si consideri la funzione:

$$\phi(t; p) = F_\tau(t) - p$$

nella variabile t e parametro p. Il quantile della distribuzione τ relativo alla probabilità p è quel valore $\hat{t}$ (in questo caso detto *radice*) che annulla la funzione $\phi(\cdot)$, cioè tale per cui $F_\tau\left(\hat{t}\right) = p$. Si noti che l'univocità di $\hat{t}$ non è garantita in quanto la funzione di distribuzione non è in generale strettamente crescente (in particolare per distribuzioni di probabilità discrete). Questo significa che il quantile calcolato potrebbe non essere particolarmente accurato.

Fra i vari metodi numerici di ricerca degli zeri di una funzione, risulta par-

ticolarmente interessante il metodo di *Steffersen*, il quale consiste nell'applicare il metodo di *Newton* (visto come funzione di iterazione di un metodo di punto fisso) in congiunzione con l'*accelerazione di Aitken* [98]. Come punto iniziale viene utilizzato il valor medio della distribuzione PH τ.

5.4 Stima dei Parametri

Per la stima dei parametri di una distribuzione PH è stato utilizzato l'algoritmo descritto in [88], il quale rappresenta un'applicazione del metodo dei momenti (§3.1.1). L'algoritmo stima i parametri della distribuzione cercando una distribuzione *PH Aciclica* (*APH*) del tipo *Coxian-Erlang* i cui primi tre momenti corrispondano a quelli da stimare; la distribuzione PH trovata ha un numero di fasi quasi minimale. Recentemente, è stato sviluppato un algoritmo, sempre basato sul metodo dei momenti, in grado di trovare una distribuzione PH con un numero minimo di fasi [17].

Questo metodo, oltre a essere molto efficiente dal punto di vista computazionale, ha il vantaggio di essere molto flessibile, grazie alla scelta automatica del numero di fasi (quasi-)ottimale e al fatto di non dipendere dalla scelta di un valore iniziale, come di solito accade nella maggior parte dei metodi iterativi. Tuttavia, ricavando le stime dei parametri dai momenti, è molto sensibile agli "outlier" e tiene poco in considerazione le caratteristiche della coda della distribuzione.

Oltre al metodo dei momenti suddetto, ne esistono altri, la maggior parte dei quali sono algoritmi di tipo iterativo. Di seguito ne vengono presentati alcuni la cui caratteristica comune è la minimizzazione di una funzione di distanza tra la distribuzione empirica e la PH approssimata.

Il metodo di *Feldmann-Whitt* [44] è un metodo ricorsivo che adatta una particolare distribuzione tramite una mistura di distribuzioni Esponenziali (cioè tramite una Iper-Esponenziale); l'algoritmo procede ricorsivamente ricavando una distribuzione esponenziale per ogni porzione della coda non ancora sottoposta al "fitting"; questo metodo, pur essendo abbastanza semplice, è di solito utilizzato per l'adattamento di distribuzioni teoriche (gli autori mo-

strano un esempio di adattamento di una Weibull e di una Pareto). Come gli stessi autori suggeriscono, è consigliabile non usare direttamente questo metodo per effettuare l'adattamento di una distribuzione empirica; piuttosto, conviene prima cercare una distribuzione teorica che si adatti a quella empirica e, successivamente, applicare l'algoritmo di Feldmann-Whitt per trovare una Iper-Esponenziale che si adatta alla distribuzione teorica trovata. A questo punto, però, il vantaggio di utilizzare una distribuzione PH, anzichè direttamente quella teorica riguarda, probabilmente, solo una più facile trattabilità matematica. L'algoritmo, inoltre, dipende dalla scelta del numero di fasi per la distribuzione Iper-Esponenziale; tale scelta è delegata all'utente e rappresenta uno svantaggio per il tipo di analisi descritta in questo lavoro.

Un altro metodo, è quello descritto in [12], basato sulla tecnica statistica *Expectation-Maximization* (*EM*): ad ogni iterazione, l'algoritmo cerca quei valori dei parametri di una distribuzione PH che minimizzino una funzione di distanza. Il problema di questo algoritmo, per gli scopi del presente lavoro, è legato alla scelta del numero di fasi, lasciata all'utente; non è necessario specificare i valori iniziali dei parametri, in quanto l'algoritmo è in grado di generarli a caso; tuttavia, si è notato che una cattiva scelta dei valori iniziali rende la convergenza dell'algoritmo estremamente lenta e quindi diventa poco utilizzabile dal punto di vista pratico. Gli esperimenti sono stati effettuati tramite il tool *EMpht* [89], sviluppato dagli stessi autori che hanno proposto l'algoritmo in questione.

Un'evoluzione del metodo di Feldmann-Whitt è quella proposta in [58] e implementata nel programma *PhFit*. L'aspetto interessante di questo metodo è la distinzione tra l'adattamento del corpo e quello della coda della distribuzione empirica; un possibile svantaggio di questo metodo, per gli scopi del presente lavoro, è legato alla necessità di dover specificare, come parametri iniziali, il numero di fasi e il punto di "taglio" che divide il corpo della distribuzione empirica dalla relativa coda. In particolare, quest'ultima informazione purtroppo non è, in generale, semplice da reperire. Inoltre, sperimentalmente, si sono incontrati dei problemi di stabilità numerica (probabilmente dovuti a una cattiva scelta dei parametri iniziali).

Viste le considerazioni appena esposte, nel presente progetto si è deciso di utilizzare il primo metodo descritto, ossia quello basato sui momenti, consapevoli, comunque, delle sue limitazioni.

5.5 Generazione di Numeri Casuali

Per la generazione di numeri casuali distribuiti secondo una PH è stato utilizzato il metodo descritto in [85]. Questo metodo consiste nella simulazione della Catena di Markov sottostante fino al raggiungimento dello stato di assorbimento: per ogni transizione che avviene nella catena, si "ricorda" il numero di volte k in cui ogni stato viene visitato, prima dell'assorbimento, e quindi si genera il numero casuale tramite una somma di numeri casuali distribuiti come una Erlang a k stadi. Più precisamente, il tempo speso in uno stato i, prima di effettuare una transizione in uno stato j, rappresenta una quantità casuale che contribuisce al tempo totale di assorbimento; come noto, ogni soggiorno in un particolare stato i è distribuito secondo una Esponenziale di parametro $-q_{ii}$, mentre la transizione dallo stato i allo stato j avviene con una probabilità pari a $q_{ij}/(-q_{ii})$. Per generare un numero casuale distribuito secondo una PH, il metodo prevede la memorizzazione del numero di volte k_i in cui un certo stato i viene incontrato prima dell'assorbimento, e quindi la generazione della quantità x_{PH}, distribuita come una PH, tramite una somma di numeri casuali distribuiti secondo una Erlang a k_i stadi e di parametro $-q_{ii}$.

Per generare le transizioni casuali per lo stato iniziale e per le transizioni tra gli stati, si utilizza la tecnica nota come *Metodo dell'Alias* [111]; si tratta di una tecnica molto ingegnosa per la generazione di numeri casuali, a partire da una distribuzione di probabilità discreta, che richiede una tabella di dimensione M e necessita di un solo confronto. Vengono utilizzati due vettori: Q (*vettore dei cutoff*) e J; si sceglie, in maniera uniforme, un valore k compreso tra 1 e M, si genera un numero uniforme u e lo si confronta con la cella $Q[k]$; se $u < Q[k]$ allora k è il valore restituito, altrimenti si restituisce $J[k]$ (*alias*). L'implementazione utilizzata nel progetto, fa uso dell'algoritmo descritto in [64], grazie al quale si ottiene una complessità computazionale $\mathcal{O}(M)$ anzichè $\mathcal{O}(M^2)$, co-

me si otterebbe se si utilizzasse il metodo originale; questo miglioramento è ottenuto grazie all'utilizzo della tecnica *Robin-Hood* [74].

Capitolo 6

Distribuzioni Heavy-Tail

Dalla statistica di base, in particolare dal Teorema del Limite Centrale B.1.1, si apprende che per qualsiasi campione $X_1, \ldots, X_n$ i.i.d. estratto da una popolazione con media e varianza *finita*, la distribuzione Normale rappresenta il limite asintotico della distribuzione della media campionaria.

Esistono, però, distribuzioni per cui tale asserzione non è valida a causa della presenza di momenti *infiniti*; tali distribuzioni sono chiamate *heavy-tailed*. Il nome "heavy-tailed" deriva dal fatto che una o entrambe le code, di queste distribuzioni, decrescono molto più lentamente di quelle delle distribuzioni Gaussiane; come conseguenza, si ha che i valori appartenenti alla coda "heavy" non sono così rari come lo sono nelle distribuzioni Gaussiane (ad es., Esponenziale o Normale). Questo fatto ha una serie di implicazioni sia teoriche, come la non validità della Legge dei Grandi Numeri [41], sia pratiche, come l'impossibilità di effettuare simulazioni di sistemi in stato stazionario [26].

Questo capitolo ha lo scopo di fornire un'introduzione alle distribuzioni "heavy-tailed", ponendo particolare attenzione sui principali metodi utilizzabili per verificarne la presenza e per stimarne i parametri. La prima sezione §6.1 introduce alcune definizioni di base; la seconda sezione §6.2 descrive le principali proprietà delle distribuzioni "heavy-tailed"; l'ultima sezione §6.3 propone alcui metodi di verifica di presenza di code "heavy" e di stima dei parametri.

6.1 Definizioni

Quando una distribuzione ha una coda che decade secondo una legge esponenziale, come la distribuzione Esponenziale o la Normale, le osservazioni estreme, cioè quelle appartenenti alla coda, sono così rare da poter essere ignorate: la probabilità che esse compaiano in un campione tende a zero esponenzialmente; per tali motivi, questo tipo di distribuzioni viene spesso utilizzato per modellare dati con *bounded support*. Le distribuzioni in cui la coda è governata da una legge esponenziale sono chiamate *short tailed*; formalmente, se X è una distribuzione "short-tailed" con CDF $F(\cdot)$, allora:

$$\exists \beta > 0 : \lim_{x \to \infty} e^{\beta x} (1 - F(x)) = 0$$

Ci sono, tuttavia, distribuzioni la cui coda cade più lentamente di una legge esponenziale; esse sono dette *long-tailed*. Dal punto di vista formale, esistono diverse definizioni di "long-tail"; di seguito si propongono quelle maggiormente diffuse:

Definizione 6.1.1 (Distribuzione Long-Tailed)**.**

I *Definizione I*. Se X è una distribuzione *long-tailed* con CDF $F(\cdot)$, allora:

$$\forall \beta > 0 : \lim_{x \to \infty} e^{\beta x} (1 - F(x)) = \infty$$

II *Definizione II*. Data una sequenza di variabili casuali $X_1, \ldots, X_n$ i.i.d., distribuite secondo una distribuzione D, allora se D è *long-tailed* si ha:

$$\lim_{x \to \infty} \frac{\Pr\{X_1 + X_2 + \cdots + X_n > x\}}{\Pr\{\max_{1 \leq i \leq n} \{X_i\} > x\}} = 1, \quad \forall n \geq 1$$

Se una coda di una distribuzione "long-tailed" decade secondo una legge polinomiale la distribuzione è detta *power-law* [86]. In tal caso:

$$(1 - F(x)) \sim x^{-\alpha}, \quad \alpha > 0$$

dove α è detto *tail-index*. Una delle distribuzioni "power-law" più semplici,

che di solito viene usata come rappresentante di questa categoria, è la distribuzione di Pareto:

$$F(x) = 1 - \left(\frac{x}{x_{\mathrm{m}}}\right)^{-\alpha}$$

Vi sono poi distribuzioni che non sono "power-law" ma che hanno, comunque, almeno una coda che descresce secondo una legge sub-esponenziale, e rientrano quindi nella famiglia delle distribuzioni "long-tailed"; queste distribuzioni includono la Weibull (quando il parametro "scale" è strettamente minore di 1) e la Log-Normale. In certi contesti, queste distribuzioni vengono chiamate *medium tailed*.

Fra le distribuzioni "power-law" ve n'è una classe di particolare interesse, nota come classe delle distribuzioni *heavy-tailed*.

Definizione 6.1.2 (Distribuzione Heavy-Tailed)**.** Se X è una distribuzione di probabilità con con CDF $F(\cdot)$, allora:

$$X \textit{heavy-tailed} \Leftrightarrow X \textit{power-law} \text{ con } \textit{tail-index } \alpha \text{ e } 0 < \alpha \leq 2$$

6.2 Proprietà

6.2.1 Momenti Infiniti

Un'importante caratteristica associata alle distribuzioni "heavy-tailed" è che non tutti i momenti della distribuzione esistono; in particolare la varianza è infinita e, se $\alpha \leq 1$, lo è pure la media. Questo significa che, asintoticamente, le osservazioni non tenderanno a concentrarsi intorno a un valore, piuttosto saranno caratterizzate da una dispersione molto accentuata. La Fig. 6.1 mostra il "comportamento erratico" delle medie campionarie cumulative (*moving average*) calcolate su un campione di numerosità 10000, estratto da una Pareto con "tail-index" inferiore a 2; il grafico mette a confronto le "moving average", calcolate su valori crescenti della numerosità del campione (da 1 a 10000), relative a delle osservazioni estratte da una Pareto con "tail-index" pari a 1.5, da una Esponenziale con tasso 0.33 e da una Normale con media 3 e varianza

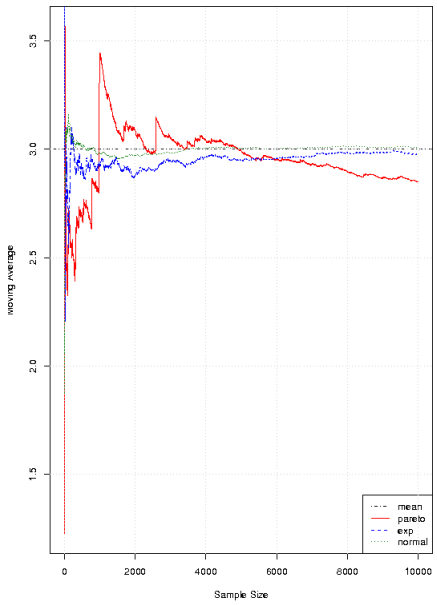


Figura 6.1: Comportamento erratico della media di una Pareto.

1; la linea orizzontale tratteggiata corrisponde al valor medio teorico (che per tutte le distribuzioni è pari a 3). Come si può notare per la Pareto, malgrado il suo valor medio teorico esista e sia uguale a 3, le "moving average" non convergono a tale valore, anzi crescono "a salti"; ciò è essenzialmente dovuto alla presenza di osservazioni "estreme", cioè appartenenti alla coda (destra) della distribuzione, le quali, avendo un ordine di grandezza maggiore rispetto alle osservazioni del corpo della distribuzione, dominano su queste ultime e determinano l'ordine di grandezza del valor medio.

L'esistenza dei primi momenti di ordine k è condizionata al valore del parametro "tail-index" α; in particolare si può dimostrare che i primi k momenti esistono solo se $k < \alpha$ [86].

6.2.2 Scale Invariance

Al contrario di quanto succede per la maggior parte delle distribuzioni tradizionali, quelle "power-law", e quindi anche quelle "heavy-tailed", si comportano nello stesso modo su scale di valori differenti; per esempio, se, nel

contesto dell'analisi del carico di un server FTP risultasse che la distribuzione della dimensione dei file trasferiti segue una distribuzione "heavy-tailed", allora risulterebbe che se i file di dimensione $1KB$ e $2KB$ sono in proporzione 4:1, cambiando la scala della dimensione, anche i file di dimensione $1MB$ e $2MB$ sarebbero, approssimativamente, nella stessa proporzione.

Definizione 6.2.1 (Scale Invariance)**.** Sia X una variabile aleatoria distribuita secondo una *power-law* con *tail-index* α e CDF $F(\cdot)$, allora:

$$1 - F(cx) = \Pr\{X > cx\} \propto (cx)^{-\alpha} = c^{-\alpha}x^{-\alpha} \propto 1 - F(x), \quad c \in \mathbb{R}$$

Questa proprietà è chiamata *scale invariance* (o *scale free*).

La proprietà di invarianza di scala non è altrettanto valida per le distribuzioni Gaussiane (ossia per quelle la cui coda segue una legge esponenziale; per esempio, se X è una variabile aleatoria con distribuzione Esponenziale di parametro λ e CDF $F(\cdot)$, si ha:

$$X \sim \text{Exp}(\lambda) \Rightarrow 1 - F(cx) = \Pr\{X > cx\} \propto e^{-c\lambda x} = e^{-(c\lambda)x} \Rightarrow \text{Exp}(c\lambda), \quad c \in \mathbb{R}$$

6.2.3 Stabilità e Teorema del Limite Centrale Generalizzato

Il Teorema del Limite Centrale classico B.1.1 afferma che la somma di n variabili aleatorie i.i.d., con media μ e varianza *finita* σ^2, tende a una distribuzione Normale:

$$X_1, \ldots, X_n \text{i.i.d. con } \mu, \sigma^2 \in \mathbb{R} \Rightarrow a_n (X_1, \ldots, X_n) - b_n \to Z \sim \mathcal{N}(0,1) \text{ per } n \to \infty$$

con

$$a_n = \frac{1}{\sigma\sqrt{n}}, b_n = \frac{\sqrt{n}}{\sigma}\mu$$

Nel caso di variabili aleatorie distribuite secondo una distribuzione "heavy-tailed", il CLT, in forma classica, non è applicabile a causa della varianza infinita. Tuttavia, si può dimostrare la validità di una forma più generale di CLT; prima di presentarla, occorre introdurre la definizione di distribuzione α-*Stable*;

vi sono vari modo per definire una distribuzione α-*Stable*; la più utilizzata è quella che fa uso della *proprietà di stabilità* [87].

Definizione 6.2.2 (Distribuzione α-Stable)**.** Una variabile aleatoria X si dice distribuita secondo una α-*Stable* di parametri α, β, γ e δ, e si indica con $S(\alpha, \beta, \gamma, \delta)$, se, prese due variabili casuali X_1, X_2 i.i.d. distribuite come X, vale:

$$\forall a, b > 0, \exists c > 0, d \in \mathbb{R} : aX_1 + bX_2 \xrightarrow{\mathcal{D}} cX + d$$

La distribuzione ha quattro parametri:

- **Characteristic Exponent** (o *Index of Stability*) $\alpha \in (0, 2]$, il quale determina le caratteristiche delle code;
- **Skewness** $\beta \in [-1, 1]$, il quale influisce sull'asimmetria della distribuzione;
- **Scale** $\gamma \in \mathbb{R}^*$, il quale determina la variabilità intorno a δ [1];
- **Location** $\delta \in \mathbb{R}$, il quale sposta a destra o a sinistra la distribuzione.

Proposizione 6.2.1 (α-Stabilità)**.** *Dalla Def. 6.2.2 segue che, data una sequenza di variabili aleatorie $X_1, \ldots, X_n$ i.i.d. distribuite secondo una α-Stable $S(\alpha, \beta, \gamma, \delta)$, la variabile aleatoria Y, ottenatua da una loro combinazione lineare, per qualche costante a_n e b_n, con $a_n > 0$ e $b_n \in \mathbb{R}$, è distribuita secondo una α-Stable $S(\alpha, \beta, \gamma', \delta')$. Cioè:*

$$\exists a_n > 0, b_n \in \mathbb{R} : Y = a_n \sum_{i=1}^{n} X_i + b_n \xrightarrow{\mathcal{D}} X_1$$

Si può inoltre dimostrare che l'unico valore possibile per a_n è pari a $n^{-(1/\alpha)}$ [87].

Le distribuzioni α-Stable sono di particolare interesse nel contesto delle distribuzioni "heavy-tailed" in quanto:

Proposizione 6.2.2. *Si può dimostrare che tutte le distribuzione α-Stable per $\alpha < 2$ sono* heavy-tailed. *Quando $\alpha = 2$ la distribuzione rientra nella famiglia delle distribuzioni Gaussiane; questo è l'unico caso in cui una distribuzione α-Stable possiede varianza finita.*

[1] Il caso con $\gamma = 0$ rappresenta una distribuzione degenere concentrata in δ.

Esempi tipici di distribuzioni α-Stable "heavy-tailed" sono la Cauchy e la Laplace; la Pareto non è una distribuzione α-Stable, in quanto risulta "heavy-tailed" solo quando il suo parametro "shape" ("tail-index") è minore o uguale a due. Esiste una classe più generica di distribuzioni chiamata classe delle distribuzioni *Stable*, la quale include, fra le altre, la distribuzione α-Stable. la Valori Estremi Generalizzata, la Log-Normale, ... Per maggior informazioni si veda, ad es., [78].

È ora possibile presentare la versione generalizzata del CLT [87]:

Teorema 6.2.3 (Teorema del Limite Centrale Generalizzato (GCLT))**.** *Data una sequenza di n variabili aleatorie $X_1, \ldots, X_n$ i.i.d., si ha:*

$$\exists a_n > 0, b_n \in \mathbb{R} : a_n \left(X_1 + \cdots + X_n\right) - b_n \xrightarrow{\mathcal{D}} Z \Leftrightarrow Z \sim S\left(\alpha, \beta, \gamma, \delta\right) \tag{6.2.1}$$

per qualche $0 < \alpha \leq 2$. Inoltre se D è la distribuzione delle variabili casuali X_i, con $i = 1, \ldots, n$, si dice che D è nel dominio di attrazione *(*domain of attraction*) di Z e si indica $D \in \mathrm{DA}(Z)$.*

Dal GCLT segue, per esempio, che la somma di variabili casuali distribuite secondo una Pareto con "tail-index" minore o uguale a due, ha come distribuzione limite una α-Stable; mentre la somma di variabili casuali distribuite come una Pareto con "tail-index" superiore a due, tende a una Normale.

6.2.4 Expectation Paradox

La presenza di code "heavy" ha implicazioni anche sulla cosidetta *conditional expectation* $\mathrm{E}\left[X | X > \tau\right]$. La *conditional expectation* è definita nel seguente modo:

Definizione 6.2.3 (Conditional Expectation)**.** Data una variabile casuale X e un valore τ, il valore atteso di X condizionato ai valori di X superiori a τ è dato da:

$$\begin{aligned}\mathrm{E}\left[X | X > \tau\right] &= \int_{\tau}^{\infty} (t - \tau) \frac{f(t)}{\int_{\tau}^{\infty} f(u)\, du}\, dt \\ &= \frac{\int_{\tau}^{\infty} t f(t)\, dt}{\int_{\tau}^{\infty} f(u)\, du} - \tau\end{aligned}$$

dove $f(\cdot)$ è la funzione di densità della variabile aleatoria X e τ è detto *valore soglia* (*threshold*).

Nella definizione appena presentata, la "conditional expectation" rappresenta una funzione in τ [2].

Il tipo di coda di una distribuzione, influisce sul comportamento della "conditional expectation". Per esempio se X è la distribuzione del tempo di esecuzione di un job, la "conditional expection" fornisce il tempo medio residuo che occorre perché uno specifico job termini la sua esecuzione. Si possono identificare tre comportamenti tipici:

- comportamento *short-tail* (ad es., distribuzione Uniforme): il tempo residuo decresce al crescere di τ;
- comportamento *memoryless* (ad es., distribuzione Esponenziale): il tempo residuo è indipendente da τ;
- comportamento *heavy-tail* (ad es., distribuzione Pareto con "tail-index" non superiore a 2): il tempo residuo cresce al crescere di τ.

Il caso per le distribuzioni "heavy-tail" dà luogo al cosidetto *expectation paradox*:

> Tra due job, è più probabile che termini dopo quello che è in esecuzione da più tempo.

In Fig. 6.2 è mostrato un esempio di quanto detto sopra; in particolare viene visualizzata la media condizionata, al variare della soglia τ, per una Uniforme in $[0, 2a]$, una Esponenziale di parametro $1/a$ e una Pareto con "tail-index" pari ad $a > 1$ [41]; dalla figura si nota chiaramente come, la media condizionata di una Pareto cresce al crescere del valore soglia τ. Una conseguenza diretta del "expectation paradox" nel contesto dello "scheduling" è che, nel caso in cui i tempi di esecuzione siano distribuiti secondo una distribuzione

[2] Esiste una definizione analoga in cui la "conditional expectation" rappresenta una variabile aleatoria; in tal caso, il condizionamento non è effettuato su un valore τ, piuttosto su una variabile aleatoria T, cioè $\mathrm{E}[X|T]$.

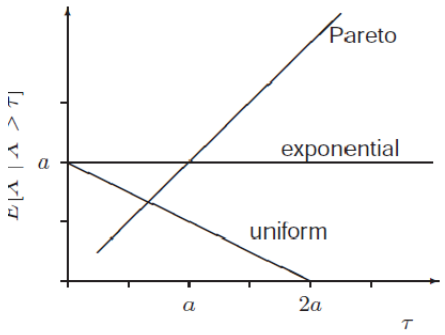


Figura 6.2: Conditional Expectation per una Uniforme, una Esponenziale e una Pareto.

"heavy-tailed", potrebbe essere più conveniente assegnare un nuovo job a una macchina su cui è in esecuzione un job da minor tempo, nel caso non siano presenti macchine "idle".

6.2.5 Mass-Count Disparity

Per una distribuzione vale, in generale, il cosidetto *Principio di Pareto* (*80/20 rule*), per il quale il 80% della massa (ad es., ricchezza) è determinato dal 20% della numerosità (ad es., popolazione); viceversa, il 80% della numerosità (ad es., popolazione) influisce solo sul 20% della massa (ad es., ricchezza).

Il significato del Principio di Pareto, applicato alle distribuzioni "power-law", e in particolare a quelle "heavy-tailed", è che nonostante i valori non estremi rappresentino la maggioranza delle osservazioni, ciò che contribuisce alla maggior parte della massa di probabilità sono i valori estremi della coda. Questo principio si esprime, matematicamente, tramite il concetto di *Mass-Count Disparity*.

Definizione 6.2.4 (Mass-Count Disparity)**.**

$$\lim_{x\to\infty} \frac{\Pr\{X_1 + X_2 + \cdots + X_n > x\}}{\Pr\{\max_{1\le i\le n}\{X_i\} > x\}} = 1, \quad \forall n \ge 1$$

La definizione appena data afferma che malgrado l'osservazione tipica sia "piccola" (cioè provenga dal corpo della distribuzione), l'unità di lavoro tipica è invece grande (cioè è determinata dalle osservazioni estreme della coda). Un esempio pratico può chiarirne meglio il concetto; se si considera il carico generato dal trasferimento di file attraverso il Web, relativamente a un particolare server HTTP, è probabile che il relativo modello segua una distribuzione "power-law" (e, in particolare, una "heavy-tailed"); il motivo di ciò, è che la maggior parte dei file trasferiti riguarda file di piccole dimensioni, come pagine HTML, immagini, script, mentre la maggior parte del carico, in termini di byte, è data dal trasferimento di pochi file grandi, come file video o audio (ad es., si veda [28]).

6.3 Presenza di Heavy-Tail e Stima del Tail-Index

È importante poter riconoscere se un insieme di dati possa provenire da una distribuzione "heavy-tailed". Esistono vari approcci per verificare la presenza di code "heavy", molti dei quali permettono anche di ottenere una stima del "tail-index". In questo capitolo verranno presentati alcuni dei metodi più diffusi; per maggiori informazioni si veda, per esempio, [5, 41]. Oltre ai metodi presentati, è sempre possibile applicare uno dei metodi standard per la stima dei parametri di una distribuzione, come il metodo dei momenti e il metodo MLE (si veda Cap. 3); mentre questi ultimi metodi hanno il vantaggio di fornire spesso stime più precise nel caso in cui la distribuzione sottoposta all'adattamento rappresenti effettivamente quella dei dati, quelli presentati in questo capitolo hanno il vantaggio di non dipendere dal tipo particolare di distribuzione "heavy-tail"; ciò è utile in quanto evita di provare a effettuare il "fitting" su diverse distribuzioni "heavy-tailed" come, la Pareto (con $a \leq 2$), la Cauchy e la Laplace, e, cosa più importante, permettono di verificare l'effettiva presenza di code "heavy-tailed" (mentre i metodi tradizionali si limitano a fornire una stima dei parametri, che può risultare grossolana se la distribuzione dei dati non è caratterizzata da code "heavy").

6.3.1 Stimatore di Hill

Lo stimatore di *Hill* [56] non è altro che una forma rielaborata del MLE di una Pareto.

Definizione 6.3.1 (Stimatore di Hill)**.** Dato un campione di osservazioni $x_1, \ldots, x_n$ i.i.d., e fissato un valore intero $1 \leq k < n$, si definisce stimatore di *Hill*, la quantità:

$$\hat{\gamma}_{n,k} = \frac{1}{k} \sum_{i=1}^{k} \left[\ln x_{(n-i+1)} - \ln x_{(n-k)} \right] \tag{6.3.1}$$

dove $\hat{\gamma}_{n,k}$ rappresenta il reciproco della stima del "tail-index" e $x_{(1)} < \cdots < x_{(n-k)} < \cdots < x_{(n)}$ sono le statistiche d'ordine del campione di osservazioni

$x_1, \ldots, x_n$. Il valore k rappresenta il numero di osservazioni estreme della coda che si vuole tenere in considerazione.

Sotto opportune ipotesi, si può dimostrare che la distribuzione di $\hat{\gamma}_{n,k}$ tende a una Gaussiana con media γ (cioè il reciproco del "tail-index" reale) e varianza $(\gamma^2 k)^{-1}$. Si può quindi definire un intervallo di confidenza per $\hat{\gamma}_{n,k}$ al $(1-\beta)\%$:

$$\hat{\gamma}_{n,k} \pm z_{\frac{\beta}{2}} \frac{\hat{\gamma}_{n,k}}{\sqrt{k}} \tag{6.3.2}$$

dove $z_{\beta/2}$ è il quantile di una Normale standard tale per cui:

$$\Phi\left(z_{\frac{\beta}{2}}\right) = \Pr\left\{z_{\frac{\beta}{2}}\right\} = 1 - \frac{\beta}{2}$$

La procedura per la verifica della presenza di code "heavy" tramite lo stimatore di *Hill* è la seguente:

1. Dato un campione di osservazioni $x_1, \ldots, x_n$, si considerino le relative statistiche d'ordine (Def. B.2.4), cioè si riordini il campione in modo crescente in modo da ottenere il nuovo campione $x_{(1)}, \ldots, x_{(n)}$, con $x_{(1)} < \cdots < x_{(n)}$.

2. Per $k = t, t+1, \ldots, n-1$, con $1 \leq t < n$:

   (a) Si calcoli $\hat{\gamma}_{n,k}$ tramite lo stimatore di Hill (6.3.1).

   (b) Si disegni su un grafico il punto:

   $$(k, \hat{\gamma}_{n,k})$$

   (c) Si calcolino gli estremi $l_{n,k,\beta}$ e $u_{n,k,\beta}$ dell'intervallo di confidenza (6.3.2) al $(1-\beta)\%$:

   $$l_{n,k,\beta} = \hat{\gamma}_{n,k} - z_{\frac{\beta}{2}} \frac{\hat{\gamma}_{n,k}}{\sqrt{k}}$$
   $$u_{n,k,\beta} = \hat{\gamma}_{n,k} + z_{\frac{\beta}{2}} \frac{\hat{\gamma}_{n,k}}{\sqrt{k}}$$

(d) Si disegnino su un grafico i punti:

$$(k, l_{n,k,\beta}) \,, \quad (k, u_{n,k,\beta})$$

I passi da 2c a 2d sono opzionali; tuttavia sono utili per verificare visivamente l'andamento dell'intervallo di confidenza al crescere di k (*curve di envelope*).

Se la distribuzione dei dati è "heavy-tailed", il grafico dovrebbe approssimativamente tendere, al crescere di k, ad una linea retta orizzontale; il valore dell'ordinata di questa retta limite, rappresenta la stima del reciproco del "tail-index". Quindi, come stima del "tail-index" si può usare il reciproco dell'ultimo valore di $\hat{\gamma}_{n,n-1}$ ritrovato, oppure effettuare una regressione lineare sul grafico di Hill e utilizzare il reciproco del valore del "intercept" (verificando prima che la pendenza della retta di regressione sia approssimativamente uguale a zero). La Fig. 6.3 mostra due grafici dello stimatore di Hill calcoli su un campione di numerosità 500 e con un valore di k fatto variare da 1 a 499: un grafico è relativo a una distribuzione Esponenziale di media 1, mentre l'altro è relativo a una Pareto con "tail-index" pari a 1.5; come si può notare, il grafico dello stimatore di Hill della Esponenziale, non mostra alcun tipo di convergenza; invece quello per la distribuzione Pareto converge verso il valore 0.65 circa, da cui si ottiene un "tail-index" pari a 1.54, un valore non molto distante da quello reale.

Questo metodo dovrebbe essere utilizzato solo per una verifica della presenza di code "heavy" e in congiunzione ad altri metodi; infatti, l'uso dello stimatore di Hill, è efficace solo quando la distribuzione sottostante è una Pareto o una molto simile [34]: in tal caso, lo stimatore di Hill tende, al crescere di k, allo stimatore di massima verosimiglianza. Negli altri casi, anziché considerare l'intera zona destra del grafico, si cerca una regione di stabilità in tutto il grafico; la presenza di una tale regione, può essere considerato un sintomo di coda "heavy" e l'ordinata del grafico, a cui la regione di stabilità tende, può essere interpretata come una stima dell'inverso del "tail-index".

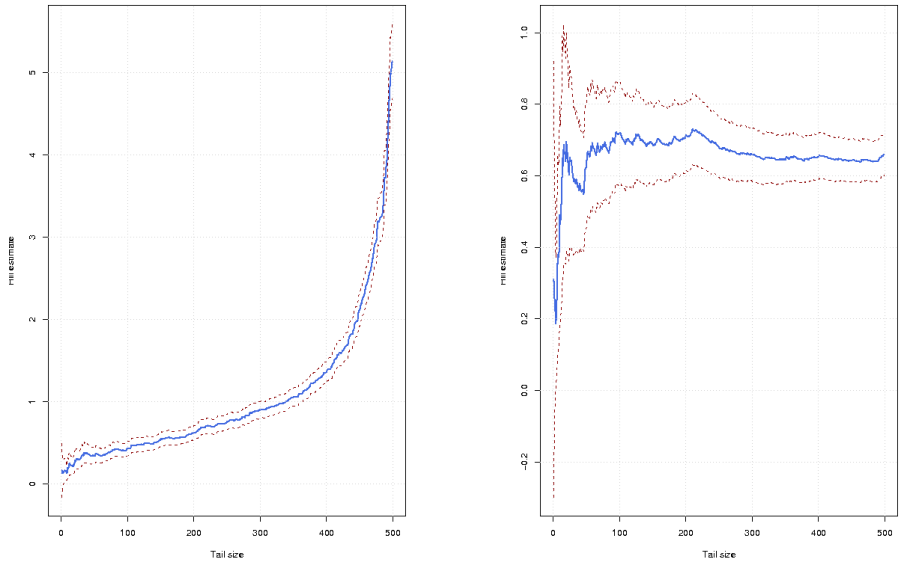


Figura 6.3: Hill plot di una Esponenziale e di una Pareto.

6.3.2 Grafico Log-Log CCDF

Metodo della Curvatura

Un grafico *log-log CCDF* consiste nel grafico, in scala log-log, della funzione di distribuzione cumulativa complementare (*CCDF*) $\bar{F}(\cdot) = 1 - F(\cdot)$, conosciuta anche con il nome di funzione di sopravvivenza, relativa a una particolare distribuzione di probabilità avente come CDF la funzione $F(\cdot)$. Il grafico è composto dai punti:

$$\left(\log x, \log \bar{F}\left(x\right)\right)$$

Se la distribuzione è "heavy-tailed" con "tail-index" pari a α, deve risultare che:

$$\forall x > x_0 : \frac{d}{d \log x} \log \bar{F}\left(x\right) \simeq -\alpha$$

con x_0 grande.

La procedura pratica per stimare il "tail-index" α è la seguente [28]:

1. Dato un campione di osservazioni $x_1, \ldots, x_n$ e una distribuzione con CDF

$\dot{F}()$, si traccia il grafico log-log CCDF:

$$\left(\log x_i, \log \bar{F}\left(x_i\right)\right), \quad i = 1, \ldots, n$$

2. Se il grafico, a partire da un certo punto, non risulta approssimativamente lineare, è inutile proseguire con i passi successivi.

3. Altrimenti, si sceglie un punto x_0 a partire dal quale il grafico log-log CCDF appare lineare.

4. Si sceglie una serie di punti equispaziati maggiori di x_0 e si effettua una regressione lineare.

5. La pendenza della retta di regressione lineare cambiata di segno fornisce la stima $\hat{\alpha}$ del "tail-index" α.

Metodo dell'Aggregazione

Un'estensione del metodo della *curvatura* è quella di tracciare una serie di grafici log-log CCDF costruiti a partire da versioni aggregate dell'insieme dei dati originale. Lo scopo è quello di verificare la presenza della proprietà di invarianza di scala (§6.2.2), utilizzando la proprietà di stabilità (§6.2.3). Dalla proprietà di stabilità (Def. 6.2.1), risulta che la somma di variabili casuali α-Stable non altera il "tail-index" α. L'idea, quindi, è quella di tracciare il grafico log-log LLCD su diversi livelli di aggregazione e verificare se la stima del "tail-index" ottenuta da ogni grafico sia approssimativamente costante.

La procedura di costruzione del grafico è la seguente [27]:

1. Dato un campione di osservazioni $x_1, \ldots, x_n$ e una distribuzione X con CDF $\dot{F}()$, si traccia il grafico log-log CCDF:

$$\left(\log x_i, \log \bar{F}\left(x_i\right)\right), \quad i = 1, \ldots, n$$

2. Se il grafico, a partire da un certo punto, non risulta approssimativamente lineare, è inutile proseguire con i passi successivi.

3. Altrimenti, per $1 < m < n$:

   (a) Si considerano gli insiemi di osservazioni m-aggregate $y_1^{(m)}, \ldots, y_{n_m}^{(m)}$, tale che:

   $$y_k^{(m)} = \sum_{i=(k-1)m+1}^{k} mx_i, \quad k = 1, \ldots, n_m = \lfloor \frac{n}{m} \rfloor$$

   (b) Si traccia il grafico log-log CCDF dei punti:

   $$\left(\log y_i^{(m)}, \log \bar{F}\left(y_i^{(m)} \right) \right), \quad i = 1, \ldots, n_m$$

4. Si stima il "tail-index" di ogni grafico log-log CCDF tramite una regressione lineare (come in §6.3.2) e si calcola la media di tutti i valori ottenuti.

5. Se la differenza tra la media e ogni "tail-index" è sufficientemente piccola (o, in alternativa, se lo scarto quadratico medio è abbastanza piccolo) si può supporre che la proprietà di invarianza di scala valga e che il valor medio ottenuto rappresenti una buona stima del "tail-index" reale.

6.3.3 Curva di Lorenz

La *Curva di Lorenz* è un modo alternativo per rappresentare graficamente la funzione di distribuzione di una variabile aleatoria; in particolare il grafico mette in relazione la distribuzione della numerosità (*count distribution*) con quella della massa aggregata (*mass distribution*) [42]. Se la distribuzione della massa è equamente distribuita, il $p\%$ della numerosità della popolazione contribuirà esattamente al $p\%$ della massa (*perfect equality*); all'altro estremo, se la distribuzione della massa è distribuita in modo completamente sbilanciato, il 1% della numerosità controlla tutta la massa (*perfect inequality*); in generale, si avrà che il $x\%$ della numerosità influisce sul $y\%$ della massa.

La presenza di code "heavy" per una certa distribuizione può essere effettuata osservando se il grafico della Curva di Lorenz segue il principio di Pareto (*80/20 rule*), tale per cui vi è un forte sbilanciamento tra numerosità e

massa: per distribuzioni "heavy-tailed" la percentuale maggiore della massa è determinata da una piccola percentuale della numerosità, e quindi la Curva di Lorenz tenderà ad avvicinarsi alla curva di perfetta iniquità.

La Curva di Lorenz per una variabile aleatoria X con CDF $F(\cdot)$, non è nient'altro che il grafico dei punti:

$$\left(F_c\left(x\right), F_m\left(x\right)\right), \quad \forall x$$

dove $F_c(\cdot)$ è la "count distribution":

$$F_c\left(x\right) = F(x), \quad \forall x$$

mentre $F_m(\cdot)$ è la "mass distribution":

$$F_m\left(x\right) = \frac{\int_{-\infty}^{x} t\, dF(t)}{\int_{-\infty}^{\infty} u\, dF(u)}$$

La costruzione della curva, a partire da un insieme di osservazioni, può essere effettuata nel seguente modo:

1. Dato un campione di osservazioni $x_1, \ldots, x_n$, si considerino le relative statistiche d'ordine $x_{(1)}, \ldots, x_{(n)}$, con $x_{(1)} < \cdots < x_{(n)}$.
2. Data una distribuzione di probabilità X con CDF $F(\cdot)$ si costruisca la *count distribution* $F_c(\cdot)$ e la *mass distribution* $F_m(\cdot)$ nella seguente maniera:
   (a) $F_c(0) = 0, F_m(0) = 0$
   (b) Per $i = 1, 2, \ldots, n$:
   $$F_c(i) = \frac{i}{n}$$
   $$F_m(i) = \frac{\sum_{i=1}^{i} x_i}{\sum_{i=1}^{n} x_i}$$
3. Si traccino su un grafico i punti:

$$\left(F_c\left(i\right), F_m\left(i\right)\right), \quad 1 \leq i \leq n$$

4. Si tracci la retta di perfetta equità:

$$y = x$$

5. Si tracci la curva di perfetta iniquità:

$$y = 0, \quad \forall x \in [0, 1]$$
$$x = 1, \quad \forall y \in [0, 1]$$

È possibile inoltre ottenere dalla Curva di Lorenz una misura quantitativa chiamata *Coefficiente di Gini*, la quale rappresenta una misura dello scostamento della Curva di Lorenz dalla retta di perfetta equità; si ottiene attraverso il calcolo dell'area compresa tra la retta di perfetta equità e la Curva di Lorenz (moltiplicata per 2 in modo da ottenere un valore compreso tra 0 e 1):

$$\begin{aligned} G &= 2\left(\langle \text{Area sottesa Retta di Perfetta Equità} \rangle - \langle \text{Area sottesa Curva di Lorenz } L(x) \rangle\right) \\ &= 2\left(0.5 - \int_0^1 L(t)\, dt\right) \end{aligned}$$

Il valore del Coefficiente di Gini moltiplicato per 100 è a volte chiamato *Gini index*.

6.3.4 Grafico Mass-Count Disparity

Il grafico *Mass-Count Disparity* deriva dal grafico della Curva di Lorenz (§6.3.3); tuttavia, anzichè mettere in relazione uno a uno la "count distribution" $F_c(\cdot)$ con la "mass distribution" $F_m(\cdot)$, traccia le due curve separatamente (ma in corrispondenza delle stesse ascisse). La presenza di una coda "heavy" può essere rilevata attraverso il calcolo dell'area compresa tra le due curve e di altre misure di distanza, come: la percentuale di elementi della coda della distribuzione necessari per determinare il 50% della massa ($N_{1/2}$), la percentuale della massa controllata dalla prima metà degli elementi del corpo della distribuzione ($W_{1/2}$), la distanza tra la mediana della "count distribution" e quella

della "mass distribution" (*m-m distance*), . . . Per maggiori informazioni, si veda [42, 41].

Parte II

Analisi dei Dati

Capitolo 7

Analisi della traccia LCG

La traccia presa in esame in questo capitolo proviene dal progetto *LHC Computing Grid* (*LCG*) [4], un sistema Grid legato al progetto *Large Hadron Collider* (*LHC*) [1], un potente acceleratore di particelle costruito dal *CERN*. I dati provenienti dagli esperimenti effettuati nel LHC, vengono resi disponibili in maniera trasparente ai ricercatori, sparsi in tutto il mondo, attraverso il sistema LCG; LCG, oltre a fornire un supporto di memorizzazione per l'enorme quantità di dati proveniente da LHC, mette a disposizione una grande potenza computazionale per l'analisi di questi dati, la quale consiste, di solito, nell'esecuzione di procedure molto intense dal punto di vista computazionale. La piattaforma di test, da cui proviene la traccia analizzata, è composta da, all'incirca, 180 siti attivi, 24515 CPU e 3 Petabyte di memoria secondaria [67].

Il capitolo è organizzato nel seguente modo: la sezione §7.1 descrive il formato interno della traccia e il significato dei vari campi che compongono ogni sua riga; la sezione §7.2 è dedicata all'analisi statistica della traccia e, in particolare, allo studio della distribuzione dei tempi di interarrivo dei job (§7.2.2 e §7.2.3) e dei relativi tempi di esecuzione (§7.2.4 e §7.2.5); infine, la sezione §7.3 fornisce un riepilogo dell'analisi effettuata sulla traccia e le relative considerazioni.

7.1 Formato

Il log analizzato è stato recuperato dal *Parallel Workload Archive* [39] su concessione del *e-Science Group of HEP, Imperial College London*; contiene 11 giorni di attività, eseguite nel periodo compreso tra il 2005-11-20 e il 2005-01-30. Ogni attività è stata registrata dal LCG Real Time Monitor [2] a livello di "resource broker". Le attività riguardano diversi esperimenti fisici, come ALICE, ATLAS, CMS e LHCb; i nomi di questi esperimenti sono stati usati per assegnare un nome alle varie organizzazioni virtuali (VO) create per il progetto LCG.

Di seguito viene mostrato il formato della traccia; ogni riga ("entry") del log si riferisce all'esecuzione di un particolare job:

```
<timestamp> <vo> <uid> <ce> <runtime>
```

dove:

`timestamp`: data di sottomissione del job;

`vo`: organizzazione virtuale che ha sottomesso il job;

`uid`: identificatore dell'utente che ha sottomesso il job;

`ce`: identificatore del nodo di computazione (*computing element*) a cui il job è stato assegnato;

`runtime`: tempo di esecuzione del job.

Come si può notare, nel formato della traccia mancano le informazioni relative alla eventuale presenza di Bag-of-Task (BoT); in effetti, quando un job arriva al "resource broker" non possiede più nessuna informazione relativa al BoT di cui, eventualmente, poteva far parte; l'unico modo per ottenere questo tipo di informazioni è l'analisi di tracce registrate da uno "scheduler" locale a una certa VO. Quindi, come effetto collaterale, si ha, per esempio, che i job sono tutti considerati come aventi dimensione pari a uno.

7.2 Analisi Statistica

Prima di concentrarsi sull'analisi statistica dei tempi di interarrivo e dei tempi di esecuzione è utile indagare sulle caratteristiche generali della traccia per comprendere, in particolare, dove e come sono distribuiti i job fra le varie VO. Dopo l'analisi delle caratteristiche generali, seguirà quindi l'analisi specifica dei tempi di interarrivo dei job (ricavato dal campo `timestamp`) e dei relativi tempi di esecuzione (campo `runtime`); essa verrà effettuata sia a livello globale, cioè dal punto di vista dell'intero sistema Grid, sia a livello di VO; in quest'ultimo caso verrà utilizzata l'informazione contenuta nel campo `vo` del log.

7.2.1 Caratteristiche Generali

La traccia è composta da 188041 osservazioni nessuna delle quali è risultata anomala rispetto a un controllo sul tipo dei dati (ad es., il tempo di esecuzione non risulta avere dei valori negativi) e su valori mancanti.

Dalla Fig. 7.1 si evince che i job presenti nella traccia sono distribuiti in modo abbastanza uniforme nell'arco temporale considerato (dal 2005-11-20 al 2005-11-30), con un minimo pari a 13650 in data 2005-11-20 e un massimo pari a 22783 in data 2005-11-30. A livello giornaliero, quindi, non si osservano particolari "pattern", anche se potrebbero essercene a una granularità più fine; in generale, nei sistemi Grid è difficile rilevare la presenza di ciclità dal punto di vista temporale a causa della dispersione degli utenti nel mondo e quindi della differenza tra fusi orari. La distribuzione del numero dei job risulta invece totalmente sbilanciata se si osserva il numero di job per Utente o per VO; infatti, sempre dalla Fig. 7.1, si osserva che il maggior numero di job proviene, sostanzialmente, da cinque VO: il 7.6% da *alice*, il 4.5 da *atlas*, il 26.2% da *cms*, il 13.8% da *dteam* e il 36.6% da *lhcb*; per quanto riguarda gli utenti, il 9.4% dei job è dovuto all'utente 17, il 7.2% appartiene all'utente 118, e il 32.5% dei job è generato dall'utente 129.

Dato che la maggior parte dei job (circa il 88.7%) proviene solo da 5 VO su 28: *alice*, *atlas*, *cms*, *dteam*, *lhcb*, l'analisi dei tempi di interarrivo e di esecuzione

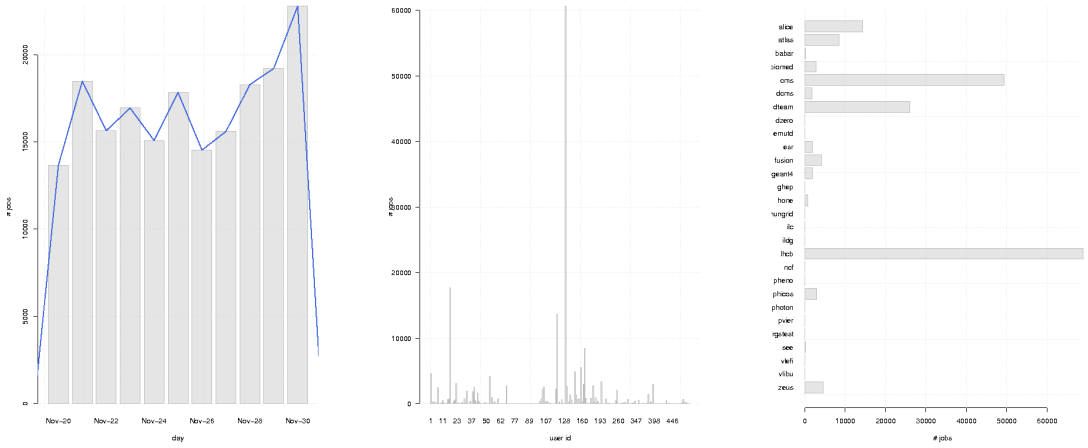


Figura 7.1: Numero di Job per Data (a sinistra), per Utente (al centro) e per VO (a destra).

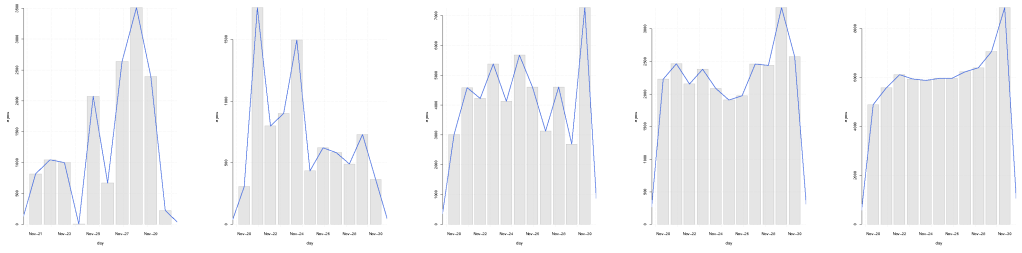


Figura 7.2: Numero di Job per Data nelle principali VO (*alice*, *atlas*, *cms*, *dteam*, *lhcb*).

a livello VO si limiterà a prendere in considerazione queste VO. In Fig. 7.2 è mostrata la distribuzione del numero dei job per data fra le cinque VO suddette; mentre per *dteam* e *lhcb* la distribuzione dei job, nei vari giorni, è abbastanza uniforme, nelle altre VO, e in particolare in *alice*, si osserva un'alta variabilità. Per esempio, nella VO *alice*, in data 2005-11-24 sono stati generati soltanto 4 job rispetto ad una media giornaliera di circa 1437 job e a un "secondo minimo" pari a 669 job (in data 2005-11-26); situazioni come queste dovrebbero essere studiate con particolare attenzione in quanto potrebbero celare delle anomalie. L'analisi di questi casi è rimandata nelle sezioni relative all'analisi del carico a livello di VO.

7.2.2 Tempi di Interarrivo – Livello Grid

Bonifica dei Dati

Un'ispezione della traccia tramite una procedura automatica scritta ad-hoc non ha rilevato la presenza di anomalie nei dati.

In Fig. 7.3 sono mostrati i quattro grafici consigliati dall'approccio EDA §2.5. La figura mostra un possibile "outlier" in corrispondenza dell'osservazione avente valore 122. La riga della traccia relativa a questa osservazione non presenta particolari anomalie:

```
1133058031 lhcb 26 ce101.grid.ucy.ac.cy 199
1133058153 lhcb 26 ce01.grid.acad.bg 137
1133058175 lhcb 26 ce01.esc.qmul.ac.uk 136
```

La riga in grassetto è quella relativa al possibile "outlier". Un grafico comparativo "box plot" tra l'insieme dei dati completo e quello senza il possibile "outlier" non mostra nessuna variazione significativa della distribuzione (Fig. 7.4; l'etichetta "with outlier" denota l'insieme contenente il possibile "outlier", mentre l'etichetta "no outlier" rappresenta l'insieme privato del possibile "outlier"). Anche dal calcolo di alcune statistiche per la centralità e la dispersione dei dati, non emerge nessuna importante differenza:

- *con outlier*:

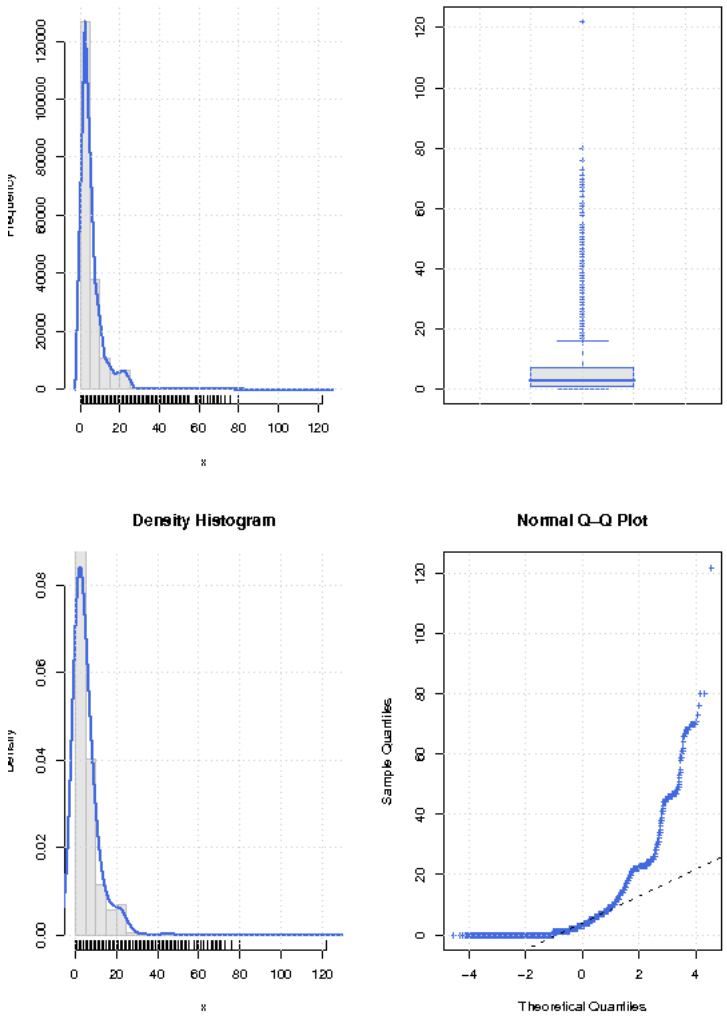


Figura 7.3: Grafico EDA per la forma della distribuzione dei Tempi di Interarrivo (livello Grid).

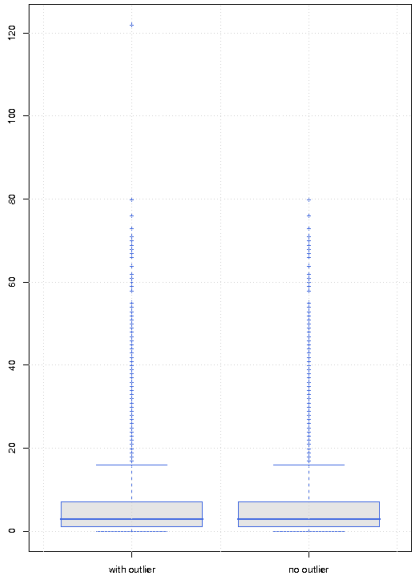


Figura 7.4: Confronto tra i Tempi di Interarrivo con e senza il possibile "outlier" (livello Grid).

  - Media: 5.054174 (CI_5 $(5.027361, 5.080988)$);
  - Mediana: 3.00 (CI_5 $(0, 19)$);
  - Scarto Quadratico Medio: 5.932414 (CI_5 $(5.913515, 5.951436)$);
  - Asimmetria: 2.436856
  - Curtosi: 13.1995

- *senza outlier*:

  - Media: 5.053552 (CI_5 $(5.026767, 5.080338)$);
  - Mediana: 3.00 (CI_5 $(0, 19)$);
  - Scarto Quadratico Medio: 5.926297 (CI_5 $(5.907417, 5.945299)$);
  - Asimmetria: 2.403873
  - Curtosi: 12.44876

Per queste ragioni, e per il fatto che non si sarebbe in grado di giustificare

Statistica	**Valore**	CI_5
Min	0	–
Primo Quartile	1	–
Mediana	3	$(0, 19)$
Terzo Quartile	7	–
Max	122	–
IQR	6	–
MAD	2.965	–
Asimmetria	2.437	–
Quartile-Asimmetria	0.333	–
Curtosi	13.199	–
Curtosi Eccesso	10.199	–
Media	5.054	$(5.027, 5.081)$
Deviazione Standard	5.932	$(5.91351, 5.95144)$
CV	1.174	–

Tabella 7.1: Riepilogo delle misure di centralità e dispersione dei Tempi di Interarrivo (livello Grid).

l'assenza di quell'osservazione, il possibile "outlier" non viene escluso dall'analisi.

Analisi delle Proprietà Statistiche

In Tab. 7.1 sono illustrate i valori delle principali misure di centralità e di dispersione. Un'analisi della tabella permette di constatare una discreta dispersione dei dati rispetto al centro della distribuzione; la mediana, pari a 3 ha associato un intervallo di confidenza la cui larghezza supera del triplo l'intervallo interquartile; la media, il valore dove tendono a concentrarsi le osservazioni, ha una distanza dal centro della distribuzione (mediana) non trascurabile.

Dalla Fig. 7.3 risulta chiara l'asimmetria destra, il picco allungato e la coda destra molto pronunciata della densità della distribuzione dei dati. Dal "box plot" si nota come malgrado la maggioranza dei dati si concentri entro l'osservazione avente valore 20, la maggior parte del tempo sembra dovuto ai valori in prossimità della coda destra. Per capire se si tratta di una coda "lunga" e, in particolare, di una coda "heavy" si sono analizzati i grafici della funzione

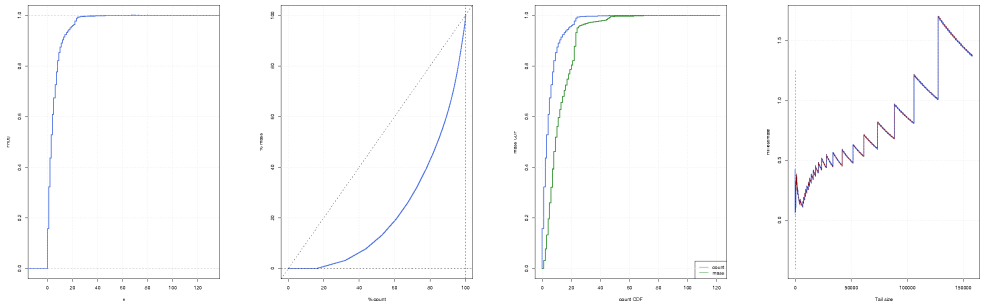


Figura 7.5: ECDF, Curva di Lorenz, Mass-Count Disparity e Hill plot dei Tempi di Interarrivo (livello Grid).

di distribuzione empirica, della curva di Lorenz, il grafico della Mass-Count Disparity e il grafico dello stimatore di Hill (Fig. 7.5); dai grafici si intuisce che, malgrado la coda si possa considerare abbastanza lunga, non sembra valere il principio della *mass-count disparity* [42] e quindi non pare esserci evidenza di una coda "heavy".

Il grafico Normal Q-Q mostra un chiaro allontanamento dalla normalità, che si accentua mano a mano che ci si sposta verso le code della distribuzione (in particolare, verso quella destra) ; ciò può essere notato anche in Tab. 7.1, dove la curtosi vale 13.199, mentre per una Normale dovrebbe valere 3, e l'asimmetria vale 2.437, mentre per una Normale dovrebbe essere uguale a 1. Questo comporta che molti dei test che fanno uso dell'assunzione di normalità non possano essere utilizzati.

Per quanto concerne l'assunzione di indipendenza delle osservazioni, dalla Fig. 7.6 si osserva una forte correlazione per tutti i valori del *lag* presi in considerazione. La presenza di autocorrelazione nei tempi di interarrivo può essere dovuta a come sono distribuiti i job; infatti, in §7.2.1, si è fatto notare come la maggior parte dei job (circa il 90%) provenga da poche VO e da un ristretto numero di utenti; si presume, quindi, che tale situazione generi dipendenza nel modo in cui i vari job giungano ai "resource broker". Questo fatto suggerisce che la classica assunzione dei tempi di interarrivo distribuiti secondo una Esponenziale o, in maniera analoga, del numero di job modellati secondo un processo di Poisson, non può essere effettuata in quanto porterebbe a

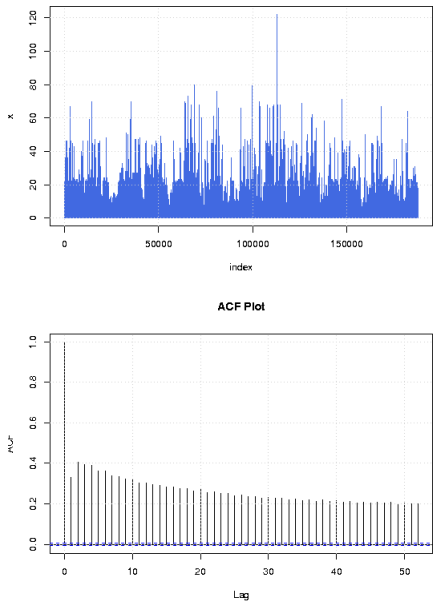


Figura 7.6: Autocorrelazione nella distribuzione dei Tempi di Interarrivo (livello Grid).

Metodo	H
Variance	0.891
R/S	0.970
Periodogram	0.961

Tabella 7.2: Esponente di Hurst per i Tempi di Interarrivo (livello Grid).

considerare indipendenti i tempi di interarrivo dei job.

Vista la presenza di autocorrelazione, è utile indagare l'eventuale esistenza di dipendenza a lungo termine; dalla Tab. 7.2 si ricava che l'esponente di Hurst risulta maggiore di 0.5, per tutti i metodi di stima, e, quindi, che i dati sono caratterizzati da una dipendenza a lungo termine; in effetti, analizzando i dati a diversi livelli di granularità (Fig. 7.7) risulta abbastanza evidente la proprietà di invarianza di scala.

La presenza di dipendenza sia a breve sia a lungo termine rende maggiormente inefficace l'utilizzo dei metodi statistici che assumono l'indipendenza fra i campioni. Ciò deve quindi essere tenuto bene in mente durante la scelta e la verifica del modello.

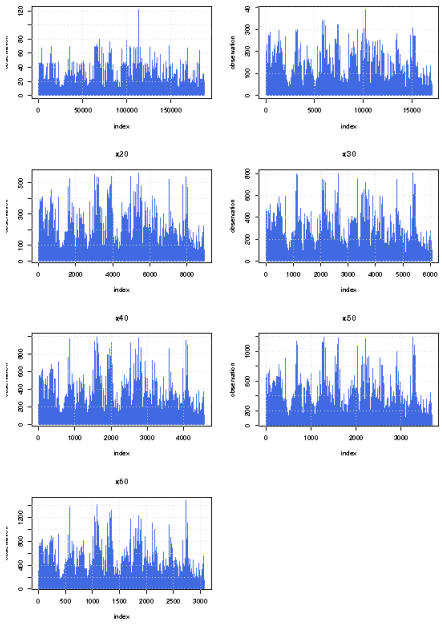


Figura 7.7: Invarianza di Scala nei Tempi di Interarrivo (livello Grid).

Scelta e Verifica del Modello

La Fig. 7.8 conferma quanto affermato sinora; essa mostra i grafici della CDF e della CCDF, in scala log-log, e della PDF della distribuzione empirica dei tempi di interarrivo (etichetta "data") insieme a quello di alcune distribuzioni i cui parametri sono stati stimati dall'intero insieme dei dati. Il primo aspetto che salta alla luce è la differenza dell'adattamento della distribuzione per le osservazioni fino al valore 20 e per quelle oltre tale valore; queste due zone potrebbero essere chiamate, rispettivamente, "corpo" e "coda" (destra) della distribuzione. Nel corpo della distribuzione, la maggior parte delle distribuzioni utilizzate per l'adattamento sembra descrivere abbastanza bene il comportamento: in particolare le distribuzioni Esponenziale, Gamma, Log-Normale, Pareto Generalizzata e Phase-Type continua sembrano descrivere meglio la distribuzione dei dati, segue quindi la Valori Estremi Generalizzata e la Logistica; le uniche distribuzioni che sembrano essere inadatte sono la Cauchy, la Normale, la Pareto e la Weibull. Nella coda destra della distribuzione, la bontà dell'adattamento è meno visibile; le distribuzioni teoriche che descrivono meglio

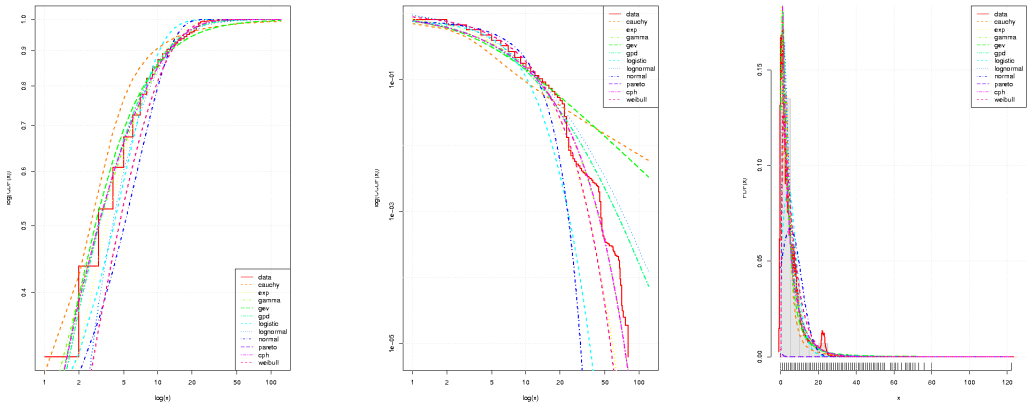


Figura 7.8: Fit per la distribuzione dei Tempi di Interarrivo (livello Grid).

questa parte della distribuzione dei dati sono la Gamma e la Phase-Type continua e, in minor misura, anche la Weibull e l'Esponenziale. Come fatto notare in precedenza, la distribuzione dei tempi di interarrivo tende ad allontanarsi dalla normalità, sebbene il suo corpo ne mostri qualche caratteristica; inoltre la coda (destra) della distribuzione empirica, benchè sia più lunga di quella di una Normale, non sembra appartenere alle code "heavy", in quanto, il grafico log-log CCDF non presenta, relativamente ai dati, alcun andamento rettilineo, tipico di una distribuzione "heavy tailed", e, inoltre, spostandosi verso osservazioni più grandi, la differenza tra la coda della distribuzione empirica e, per esempio, quella di una Pareto Generalizzata si fa sempre più marcata.

La bontà dell'adattamento nel corpo della distribuzione empirica dei tempi di interarrivo è mostrata nei grafici P-P Fig. 7.9, mentre quella nella coda della distribuzione, è visibile nei grafici Q-Q di Fig. 7.10. Da un'analisi visiva dei grafici Q-Q, P-P, log-log CDF e PDF, non è ben chiaro quale sia la distribuzione che meglio descriva il corpo e la coda della distribuzione empirica; passando, invece, all'analisi dei coefficienti di correlazione r e delle differenze di area relativa ΔA_r (Tab. 7.3), calcolati a partire dai grafici P-P e Q-Q, si possono effettuare le seguenti considerazioni:

- dalle differenze di area e dai coefficienti di correlazione dei grafici P-P, risulta che le distribuzioni Gamma, Phase-Type continua, Esponenziale e Weibull sono quelle che si adattano meglio al corpo della distribuzione

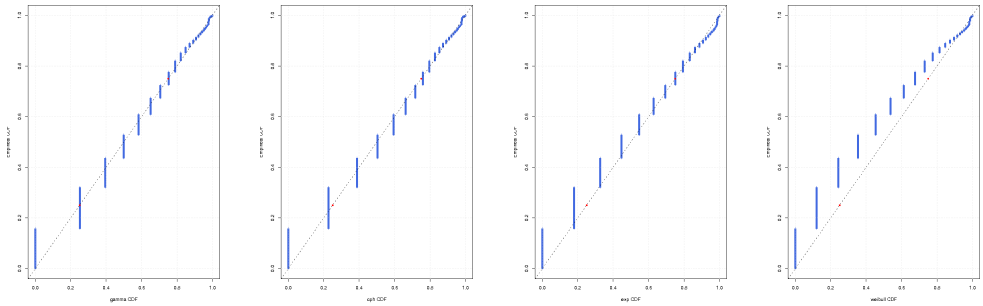


Figura 7.9: P-P plot di Gamma, Phase-Type, Esponenziale, Weibull rispetto ai Tempi di Interarrivo (livello Grid).

empirica, seguite immediatamente dalla Pareto Generalizzata e, in minor misura, dalla Log-Normale (si veda Fig. 7.9);

- dalle differenze di area e dai coefficienti di correlazione dei grafici Q-Q, si evince che tutte le distribuzioni hanno differenze di area relativa piuttosto alte, benchè abbiano un coefficiente di correlazione vicino a 1; fra queste, quelle migliori sembrano essere le distribuzioni Phase-Type continua e la Gamma, seguite dalla Log-Normale ed Esponenziale e, in minor misura, la Valori Estremi Generalizzata e la Pareto (si veda Fig. 7.10);

- complessivamente, le distribuzioni che esibiscono – sia per il corpo sia per la coda – un basso valore per le differenze di area e, contemporaneamente, un alto valore per i coefficienti di correlazione, sono, principalmente, la Phase-Type continua e la Gamma, e, in minor misura, la Log-Normale e la Weibull.

Purtroppo, i metodi numerici per verificare la bontà dell'adattamento non possono essere applicati in quanto la non validità dell'ipotesi di indipendenza tra i campioni farebbe ottenere dei risultati poco significativi. Per esempio, il test di Pearson χ^2 fornisce un p-value uguale a zero per tutte le distribuzioni, utilizzando sia il metodo di Moore sia quello di Sturges §3.2.3, tranne che per la distribuzione Pareto, per la quale si ottiene invece un p-value uguale a uno. Questi risultati sono in netta contraddizione con quanto affermato

Distribuzione	**Grafico**	**Pearson** r	**Pearson** r CI_5	ΔA_r
Cauchy				
	P-P	0.98398	(0.98385, 0.98413)	0.06151
	Q-Q	0.08901	(0.08453, 0.09350)	7.17948
Esponenziale				
	P-P	0.99393	(0.99388, 0.99399)	$2.66 \cdot 10^{-6}$
	Q-Q	0.99151	(0.99143, 0.99158)	0.75437
Gamma				
	P-P	0.98956	(0.98947, 0.98966)	$2.65 \cdot 10^{-6}$
	Q-Q	0.99320	(0.99314, 0.99327)	0.46385
GEV				
	P-P	0.98997	(0.98988, 0.99006)	0.00599
	Q-Q	0.45350	(0.44990, 0.45708)	0.98958
GPD				
	P-P	0.99152	(0.99144, 0.99159)	$7.08 \cdot 10^{-5}$
	Q-Q	0.67899	(0.67655, 0.68142)	12.3193
Logistica				
	P-P	0.97081	(0.97055, 0.97109)	0.03818
	Q-Q	0.87883	(0.87779, 0.87985)	7.55616
Log-Normale				
	P-P	0.98755	(0.98749, 0.98766)	0.00012
	Q-Q	0.96458	(0.96427, 0.96490)	0.60896
Normale				
	P-P	0.94819	(0.94773, 0.94864)	0.04042
	Q-Q	0.87223	(0.87115, 0.87331)	8.29692
Pareto				
	P-P	0.91446	(0.91371, 0.91520)	0.99998
	Q-Q	0.06683	(0.06233, 0.07133)	1.00000
Phase-Type continua				
	P-P	0.99064	(0.99056, 0.99072)	$2.65 \cdot 10^{-6}$
	Q-Q	0.99291	(0.99284, 0.99297)	0.38803
Weibull				
	P-P	0.98841	(0.98831, 0.98851)	$2.66 \cdot 10^{-6}$
	Q-Q	0.98637	(0.98620, 0.98645)	1.13644

Tabella 7.3: Coefficienti di Correlazione e Differenze di Area dei grafici P-P e Q-Q per i Tempi di Interarrivo (livello Grid).

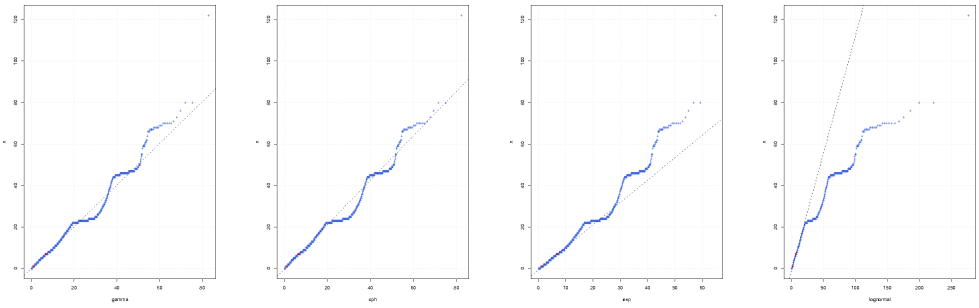


Figura 7.10: Q-Q plot di Gamma, Phase-Type, Esponenziale, Log-Normale rispetto ai Tempi di Interarrivo (livello Grid).

in precedenza: il grafico delle densità delle distribuzioni, sovrapposto all'istogramma della densità empirica (Fig. 7.8), mostra, chiaramente, come la distribuzione Pareto sia ben lontana dall'andamento caratteristico dei dati; ciò dimostra, quindi, come l'utilizzo del test di Pearson χ^2 in presenza di campioni dipendenti porti a conclusioni decisamente errate. Un discorso simile vale per i test di Kolmogorov-Smirnov e Anderson-Darling, usati in congiunzione con il metodo del "bootstrap" parametrico; per poter utilizzare la tecnica del "bootstrap" è necessario che il campione sia indipendente.

Basandosi, quindi, sui soli test visuali, grafico P-P e Q-Q, e sul calcolo del relativo coefficiente di correlazione lineare e della differenza di area relativa, si conclude che le distribuzioni Phase-Type continua e Gamma sembrano quelle che meglio si adattano sia al corpo sia alla coda (destra) della distribuzione empirica; anche la distribuzione Esponenziale descrive entrambe le parti della distribuzione in modo discreto, sebbene l'adattamento tenda a peggiorare lievemente mano a mano che ci si sposta verso i valori più estremi della coda. Altre possibili alternative includono la Weibull e la Log-Normale. Di seguito si riportano i valori dei parametri di queste distribuzioni, stimati dall'insieme delle osservazioni:

- *Esponenziale*.

Metodo	MLE
Rate	1.1978563

- *Gamma*.

Metodo	MLE
Shape	0.7258343
Rate	0.1436109

- *Log-Normale*.

Metodo	MLE
Log-Mean	1.370986
Log-StdDev	0.934164

- *Phase-Type continua*.

Metodo	MoM
Initial Vector	$[1, 0]$
Generator	$\begin{bmatrix} -0.496289 & 0.227110 \\ 0.000000 & -0.150569 \end{bmatrix}$

- *Weibull*.

Metodo	MLE
Shape	1.119601
Scale	6.289857

Riepilogo

A livello Grid, i tempi di interarrivo dei job sono caratterizzati da una discreta dispersione e asimmetria destra e da una curtosi piuttosto alta; dal grafico EDA sulla forma della distribuzione si nota un certo allontanamento dalla normalità e una possibile presenza di coda lunga.

La coda, molto probabilmente, non è di tipo "heavy", in quanto nessun test grafico (Curva di Lorenz, grafico Mass-Count Disparity e grafico dello stimatore di Hill) ne mostra la presenza.

L'analisi sull'autocorrelazione mostra una forte presenza di dipendenza a breve e lungo termine; questo ha come conseguenza la non validità dell'assunzione di indipendenza tra le osservazioni; in base alle precedenti considerazioni, non si pensa che la presenza di dipendenza a lungo termine sia causata da una coda "heavy".

Per quanto concerne l'adattamento di una distribuzione teorica ai dati, la maggior parte delle distribuzioni considerate si adatta molto bene al corpo e discretamente alla coda della distribuzione; fra tutte, quelle che possono considerarsi le migliori, per l'adattamento dell'intera distribuzione, sono, principalmente, la Phase-Type continua e la Gamma, e, in minor misura, la Log-Normale e la Weibull.

7.2.3 Tempi di Interarrivo – Livello VO

L'analisi dei tempi di interarrivo a livello di VO, è stata limitata alle VO dalle quali proviene il maggior numero di job (si veda §7.2.1); in particolare, nel presente documento, verrà descritta quella relativa alla VO *alice*.

Organizzazione Virtuale ALICE

Bonifica dei Dati

Come si è fatto notare in §7.2.1, nel giorno 2005-11-24 ci sono stati solo 4 job sottomessi da questa VO; dal seguente estratto di traccia:

```
...
1132838484 cms 43 lxgate13.cern.ch 111873
...
1132838639 lhcb 26 gw39.hep.ph.ic.ac.uk 143
1132838641 lhcb 26 gridgate.cs.tcd.ie 133
1132838651 alice 53 lxgate13.cern.ch 436
1132838662 lhcb 26 lcg-ce.usc.cesga.es 134
1132838664 cms 53 t2-ce-02.lnl.infn.it 1547
1132838672 lhcb 26 lcgce01.gridpp.rl.ac.uk 133
```

```
1132838689 cms 53 t2-ce-02.lnl.infn.it 1689
1132838703 lhcb 26 ce.keldysh.ru 78
1132838709 lhcb 26 ce101.grid.ucy.ac.cy 203
1132838718 cms 53 t2-ce-02.lnl.infn.it 1664
1132838721 alice 53 lxgate13.cern.ch 436
1132838724 dteam 33 dgce0.icepp.jp 7226
1132838728 alice 53 lxgate13.cern.ch 256
1132838734 lhcb 26 lcg-ce.ecm.ub.es 78
1132838734 lhcb 26 lcg-ce.ecm.ub.es 77
1132838746 cms 53 t2-ce-02.lnl.infn.it 1724
1132838780 lhcb 26 hudson.datagrid.jussieu.fr 81
1132838788 cms 53 t2-ce-02.lnl.infn.it 1729
1132838797 alice 53 lxgate13.cern.ch 256
1132838802 lhcb 26 lcgce01.triumf.ca 258
1132838814 cms 53 t2-ce-02.lnl.infn.it 1670
...
```

non sembrano esserci dati anomali:

- il "resource broker" *lxgate13.cern.ch* non sembra aver avuto problemi in quanto ha schedulato, nella stessa giornata, job provenienti anche da altre VO; ad es.:

  ```
  1132838484 cms 43 lxgate13.cern.ch 111873
  ```

- l'utente 53 ha sottomesso dei job anche in altre VO; ad es.:

  ```
  1132838664 cms 53 t2-ce-02.lnl.infn.it 1547
  ```

- i tempi di esecuzione non hanno valori anomali (in ordine temporale, 436, 436, 256 e 256).

Non ci sono quindi valide ragioni per escludere queste osservazioni dall'analisi statistica.

In Fig. 7.11 è possibile notare la grande dispersione che caratterizza questo insieme di dati; il tipo di asimmetria della densità è decisamente destra e il

picco è estremamente alto e stretto; ciò può essere verificato osservando anche alcune delle statistiche per la centralità e la dispersione dei dati:

- Media: 54.981 ($\mathrm{CI}_5\,(32.11910, 77.84319)$);
- Mediana: 8 ($\mathrm{CI}_5\,(7, 75)$);
- Scarto Quadratico Medio: 1398.381 ($\mathrm{CI}_5\,(1382.401, 1414.738)$);
- Asimmetria: 53.802
- Curtosi: 3112.875

La distribuzione empirica mostra parecchi valori estremi; a causa di ciò, non è possibile capire se, tra questi valori, quelli più estremi rappresentino o meno dei possibili "outlier"; per esempio, il valore massimo è dato dall'osservazione avente un tempo di interarrivo pari a 94823, equivalente a poco più di un giorno (nel caso l'unità di misura sia il secondo); il fatto che questo valore sia anomalo o regolare può essere solo verificato attraverso una precisa conoscenza della realtà da cui è stata ricavata la traccia; in generale, tale valore, benchè estremo può essere considerato ragionevale e quindi non eliminabile dall'analisi dei dati. In base a queste considerazioni, l'analisi statistica verrà effettuata tenendo conto di tutte le osservazioni.

Analisi delle Proprietà Statistiche

Nella traccia, relativamente alla VO *alice*, sono presenti 14372 osservazioni (circa il 7.6% del totale); dalle principali misure di dispersione e centralità (Tab. 7.4), si può notare come il centro della distribuzione (mediana) e il punto di concentrazione dei dati (media) siano molto distanti fra loro (circa, di un fattore 10); in effetti, la distribuzione dei dati è completamente sbilanciata, ben lontana dalla normalità (ad es., la curtosi è tre ordini di grandezza superiore a quella di una Normale, mentre l'asimmetria ne è cinquanta volte più grande). Dalla tabella si constata come la stima intervallare al 95% di confidenza della media e della mediana fornisca una stima grossolana a causa della dispersione dei dati; per esempio, la mediana, pari a 8, ha associato un intervallo di

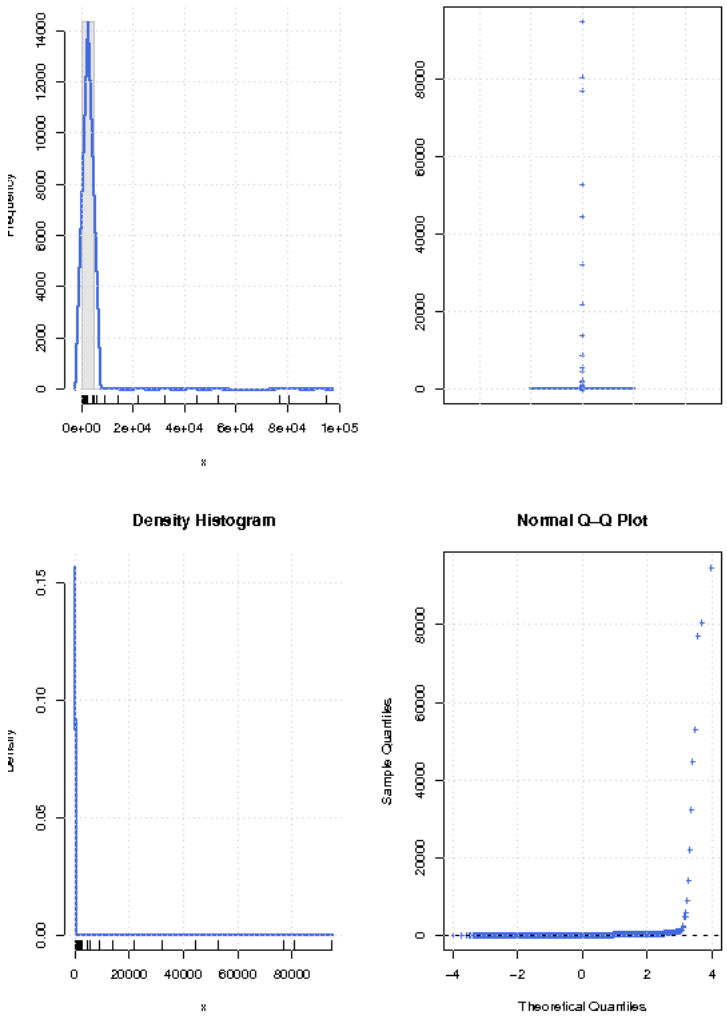


Figura 7.11: Grafico EDA per la forma della distribuzione dei Tempi di Interarrivo (livello VO: ALICE).

Statistica	**Valore**	CI_5
Min	0	–
Primo Quartile	7	–
Mediana	8	$(7, 75)$
Terzo Quartile	11	–
Max	94823	–
IQR	4	–
MAD	1.4826	–
Asimmetria	53.802	–
Quartile-Asimmetria	0.5	–
Curtosi	3112.875	–
Curtosi Eccesso	3109.875	–
Media	54.981	$(32.11910, 77.84319)$
Deviazione Standard	1398.381	$(1382.401, 1414.738)$
CV	25.433	–

Tabella 7.4: Riepilogo delle misure di centralità e dispersione dei Tempi di Interarrivo (livello VO: ALICE).

confidenza la cui larghezza supera di circa 18 volte la larghezza dell'intervallo interquartile; il valore medio, invece, lontano dal centro della distribuzione per una distanza pari a circa 7 volte la mediana, ha associato uno scarto quadratico medio che lo supera di circa 25 volte. Tutto questo suggerisce che i risultati ottenuti dai metodi statistici basati sui momenti vanno interpretati con una certa criticità.

Prima di passare alla scelta del modello, è utile analizzare il tipo di dipendenza tra le osservazioni e le caratteristica della coda della distribuzione empirica.

Per quanto riguarda l'autocorrelazione, a livello Grid (§7.2.2) si era notato che i tempi di interarrivo mostravano un'autocorrelazione sia a breve sia a lungo termine. In Fig. 7.12, si osserva che in questa VO i tempi di interarrivo, nel breve termine, sono lascamente correlati. Anche la dipendenza a lungo termine, determinata attraverso la stima dell'esponente di Hurst (Tab. 7.5) ne rivela una debole presenza, molto meno accentuata di quella misurata a livello Grid.

Per quanto concerne il tipo di coda della distribuzione, dalla Fig. 7.14 si

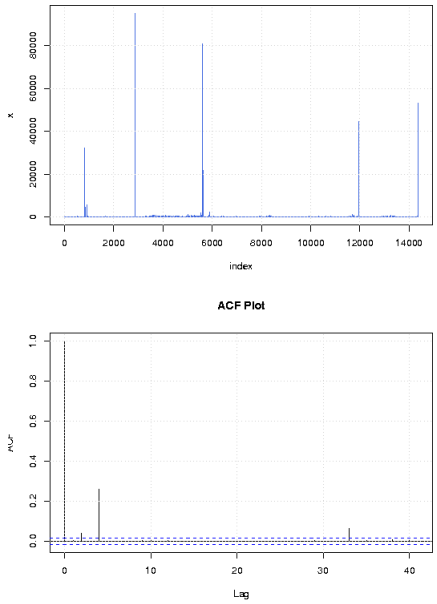


Figura 7.12: Autocorrelazione nella distribuzione dei Tempi di Interarrivo (livello VO: *alice*).

Metodo	H
Variance	0.522
R/S	0.455
Periodogram	0.626

Tabella 7.5: Esponente di Hurst per i Tempi di Interarrivo (livello VO: *alice*).

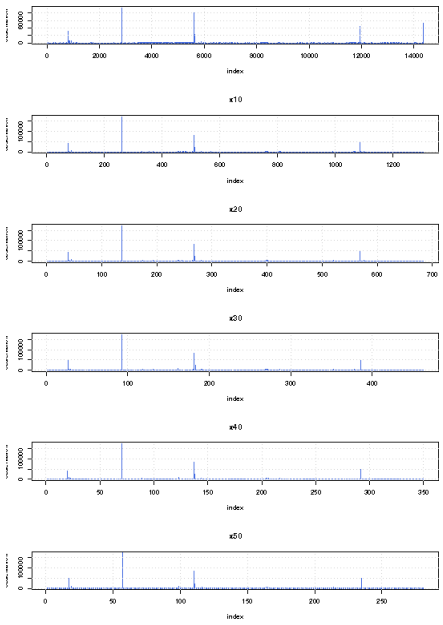


Figura 7.13: Invarianza di Scala nei Tempi di Interarrivo (livello VO: *alice*).

nota una presenza di una coda lunga; per esempio, dal grafico della curva di Lorenz e da quello della Mass-Count Disparity si può osservare la tendeza dei valori estremi a predominare su quelli più numerosi. Il grafico di Hill, non mostra un andamento costante; ciò potrebbe significare che la coda, pur essendo lunga, non sia "heavy".

Scelta e Verifica del Modello

La Fig. 7.15 è difficile da interpretare; dal grafico log-log CCDF si nota che la maggior parte delle distribuzioni teoriche mostra un discreto adattamento fino all'osservazione avente valore 190, circa; dopodiché il comportamento della maggioranza delle distribuzioni tende ad allontanarsi da quello della distribuzione dei dati; le distribuzioni Phase-Type continua e Gamma sembrano essere le uniche a riuscire ad approssimare la parte terminale della distribuzione; tuttavia, la distribuzione Gamma si comporta abbastanza male nel corpo della distribuzione. Occorre comunque far notare che anche il comportamento della Phase-Type si differenzia da quello della distribuzione dei dati nella parte

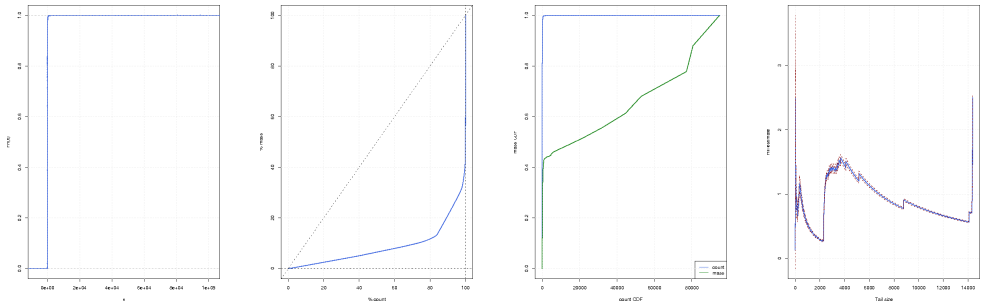


Figura 7.14: ECDF, Curva di Lorenz, Mass-Count Disparity e Hill plot dei Tempi di Interarrivo (livello VO: *alice*).

terminale della coda: a partire dall'osservazione avente valore 80604, il decadimento della coda nella distribuzione dei dati diventa più veloce di quello della Phase-Type. Questo è dovuto al fatto che, per quella parte della coda, sono a disposizione solo due osservazioni; inoltre, quando ci si accinge a effettuare il "fitting" di distribuzioni, è abbastanza frequente che verso la parte terminale della distribuzione dei dati, la distribuzione teorica si comporti in modo differente da quella empirica. Dal grafico log-log CDF è possibile analizzare l'adattamento per il corpo della distribuzione; in questo caso si nota che la maggior parte delle distribuzioni si discosta dalla distribuzione dei dati; solo la Valori Estremi Generalizzata mostra un buon adattamento. La distribuzione Phase-Type continua, che esibiva un buon adattamento per la coda della distribuzione, in questa parte della distribuzione non si comporta altrettanto bene.

Per verificare la validità delle suddette affermazioni, si procede all'analisi dei grafici P-P e Q-Q. La valutazione grafica è piuttosto complicata da interpretare; per questo motivo si passa direttamente alla valutazione dei coefficienti di correlazione r di Pearson e delle differenze di area relativa ΔA_r, calcolati sui due grafici (Tab. 7.6):

- per il corpo della distribuzione empirica, si nota che tutte le distribuzioni, a parte la Normale e la Pareto, mostrano una differenza di area relativa molto piccola, ossia sono molto vicine alla "reference line"; fra queste,

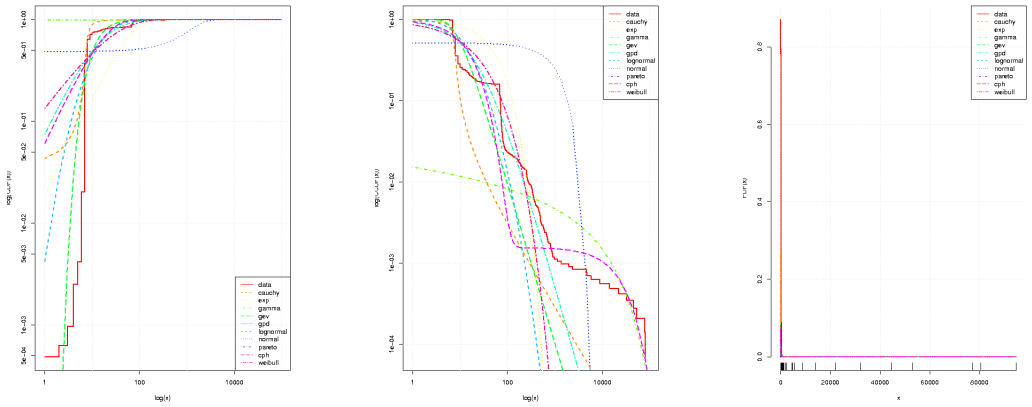


Figura 7.15: Fit per la distribuzione dei Tempi di Interarrivo (livello VO: *alice*).

le distribuzioni Cauchy e Valori Estremi Generalizzata sembrano esibire un buon adattamento (in termini di linearità e distanza), seguite dalla Phase-Type continua, dalla Gamma e dalla Pareto Generalizzata (si veda Fig. 7.16);

- per quanto concerne la coda della distribuzione empirica, si osserva che, mentre per alcune distribuzioni il coefficiente di correlazione lineare è molto vicino a 1, in generale tutte le distribuzioni sono lontane della "reference line"; il miglior compromesso si ottiene dalle distribuzioni Phase-Type continua e Gamma, seguite dalla Pareto e, in minor misura, dalla Log-Normale (si veda Fig. 7.17);

- infine, la distribuzione che mostra, complessivamente, il miglior adattamento pare essere la Phase-Type continua, seguita dalla Valori Estremi Generalizzata, dalla Pareto e dalla Log-Normale.

Di seguito si riportano i valori dei parametri, stimati dall'insieme delle osservazioni, delle distribuzioni che sono risultate le migliori dal punto di vista dell'adattamento:

- *Cauchy*.

Distribuzione	**Grafico**	**Pearson** r	**Pearson** r CI_5	ΔA_r
Cauchy				
	P-P	0.95910	(0.95777, 0.96039)	0.00137
	Q-Q	0.59782	(0.58721, 0.60823)	1829.43
Esponenziale				
	P-P	0.72778	(0.71999, 0.73538)	$3.48 \cdot 10^{-5}$
	Q-Q	0.20170	(0.18596, 0.21732)	2187.18
Gamma				
	P-P	0.09513	(0.07890, 0.11131)	$3.72 \cdot 10^{-6}$
	Q-Q	0.91595	(0.91328, 0.91854)	1.00000
GEV				
	P-P	0.86279	(0.85856, 0.86691)	$3.48 \cdot 10^{-5}$
	Q-Q	0.73049	(0.72277, 0.73802)	133.3966
GPD				
	P-P	0.80837	(0.80263, 0.81396)	$3.47 \cdot 10^{-5}$
	Q-Q	0.04618	(0.02985, 0.06248)	7591.85
Log-Normale				
	P-P	0.81844	(0.81297, 0.82376)	$3.48 \cdot 10^{-5}$
	Q-Q	0.38000	(0.36593, 0.39391)	715.093
Normale				
	P-P	0.36656	(0.35232, 0.38063)	0.3063774
	Q-Q	0.10112	(0.08491, 0.11727)	5250.98
Pareto				
	P-P	0.47909	(0.46639, 0.49158)	0.99990
	Q-Q	0.60764	(0.59723, 0.61785)	1.00000
Phase-Type continua				
	P-P	0.80462	(0.79878, 0.81031)	$7.58 \cdot 10^{-7}$
	Q-Q	0.95732	(0.95593, 0.95866)	3.38714
Weibull				
	P-P	0.78448	(0.77811, 0.79069)	$3.48 \cdot 10^{-5}$
	Q-Q	0.31160	(0.29677, 0.32629)	895.337

Tabella 7.6: Coefficienti di Correlazione e Differenze di Area dei grafici P-P e Q-Q per i Tempi di Interarrivo (livello VO: *alice*).

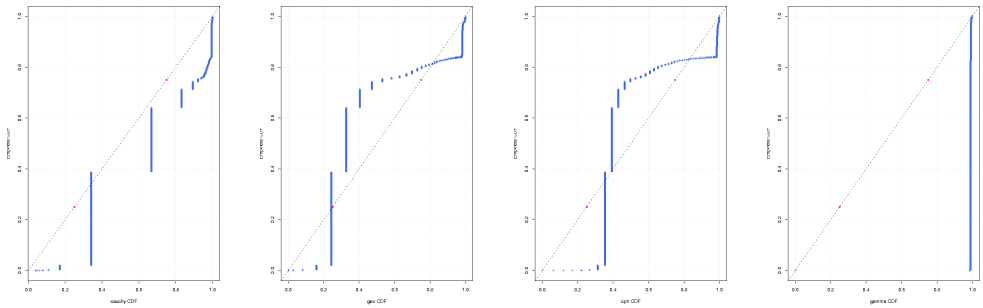


Figura 7.16: P-P plot di Cauchy, GEV, Phase-Type, Gamma rispetto ai Tempi di Interarrivo (livello VO: *alice*).

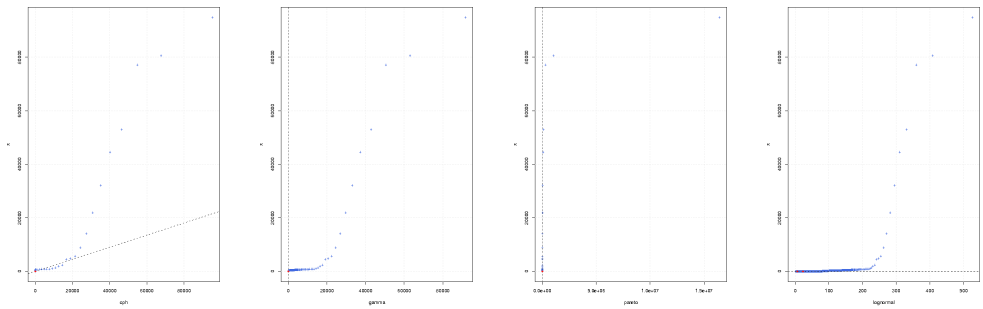


Figura 7.17: Q-Q plot di Phase-Type, Gamma, Pareto, Log-Normale rispetto ai Tempi di Interarrivo (livello VO: *alice*).

Metodo	MLE
Location	7.486256
Scale	0.884692

- *Gamma*.

Metodo	MLE
Shape	0.001546
Rate	$2.81 \cdot 10^{-5}$

- *Log-Normale*.

Metodo	MLE
Log-Mean	2.496144
Log-StdDev	0.947936

- *Pareto*.

Metodo	MLE
Location	0.000122
Shape	0.400618

- *Phase-Type continua*.

Metodo	MoM
Initial Vector	$[1, 0]$
Generator	$\begin{bmatrix} -0.062462 & 9.70 \cdot 10^{-5} \\ 0.000000 & -3.98 \cdot 10^{-5} \end{bmatrix}$

- *Valori Estremi Generalizzata*.

Metodo	MLE
Location	8.530473
Shape	0.506378
Scale	4.807556

Riepilogo

Per la VO *alice*, i tempi di interarrivo dei job sono caratterizzati da un'elevata dispersione e asimmetria destra, e da una curtosi estrema, di tre ordini di grandezza maggiore di quella di una Normale; questo significa, chiaramente, che la distribuzione dei dati è molto lontanta dalla normalità.

Rispetto ai tempi di interarrivo a livello Grid, non sembra esserci presenza di autocorrelazione a breve o a lungo termine, se non qualche leggera fluttuazione, che può essere dovuta alla casualità.

I test sul tipo di coda, mostrano la presenza di una coda lunga; ciò nonostante si pensa che non si tratti di una coda "heavy" in quanto in tal caso si manifesterebbe anche una forma di dipendenza a lungo termine.

Per quanto concerne l'adattamento di una distribuzione teorica ai dati, si è notato che la maggior parte delle distribuzioni prese in considerazione si adatta molto bene al corpo, ma approssima in modo pessimo la coda. Fra queste distribuzioni, quella che si potrebbe utilizzare per descrivere l'intera distribuzione dei dati è la Phase-Type continua; in minor misura, è possibile utilizzare anche la Valori Estremi Generalizzata.

7.2.4 Tempi di Esecuzione – Livello Grid

Bonifica dei Dati

L'ispezione della traccia, relativamente ai tempi di esecuzione, non ha rilevato alcun errore di forma dei dati (ad es., valore negativi o mancanti); tuttavia, si è notata la presenza di tempi di esecuzione molto piccoli o addirittura nulli. In particolare, su 188041 osservazioni, ne sono presenti 521 con valore pari a 0, 367 con valore uguale a 1, 151 con valore pari a 2, ...Occorre innanzitutto precisare che non è nota l'unità di misura dei tempi di esecuzione; nel caso fosse il secondo, valori inferiori al minuto potrebbero rappresentare qualche condizione anomala, come un errore di esecuzione; d'altro canto, per unità di misura multiple del minuto, i valori piccoli potrebbero semplicemente rappresentare dei job molto brevi. Inoltre, il fatto che siano presenti questo tipo di osservazioni , può ritenersi una caratteristica rappresentativa del carico, in

Statistica	**Valore**	CI_5
Min	0	–
Primo Quartile	136	–
Mediana	255	$(69, 40307)$
Terzo Quartile	4490	–
Max	586702	–
IQR	4354	–
MAD	263.903	–
Asimmetria	7.154	–
Quartile-Asimmetria	0.945	–
Curtosi	65.940	–
Curtosi Eccesso	62.940	–
Media	8970.918	$(8822.518, 9119.318)$
Deviazione Standard	32833.073	$(32728.47, 32938.35)$
CV	3.660	–

Tabella 7.7: Riepilogo delle misure di centralità e dispersione dei Tempi di Esecuzione (livello Grid).

quanto, nella pratica, è abbastanza ragionevole che alcuni job falliscano la propria esecuzione e, in più, uno "scheduler" spenderebbe comunque del tempo per effettuare la loro schedulazione. Per tali motivi, e per il fatto che non si hanno sufficienti conoscenze sui dati per implementare un criterio di selezione ragionevole, si è deciso di mantenere queste osservazioni nella traccia e di utilizzarle nell'analisi statistica.

Dai quattro grafici sulla forma della distribuzione (Fig. 7.18) si nota una grande dispersione fra i dati e una netta asimmetria destra. Con una così forte dispersione è difficile capire se siano presenti degli "outlier".

Visto che non vi sono forti motivazioni che spingano a escludere particolari osservazioni, l'analisi statistica verrà effettuata sull'intero insieme dei dati.

Analisi delle Proprietà Statistiche

In Tab. 7.7 sono illustrate i valori delle principali misure di centralità e di dispersione. Un'analisi della tabella permette di constatare la grande dispersione dei dati rispetto al centro della distribuzione; la mediana, pari a 3 ha associato

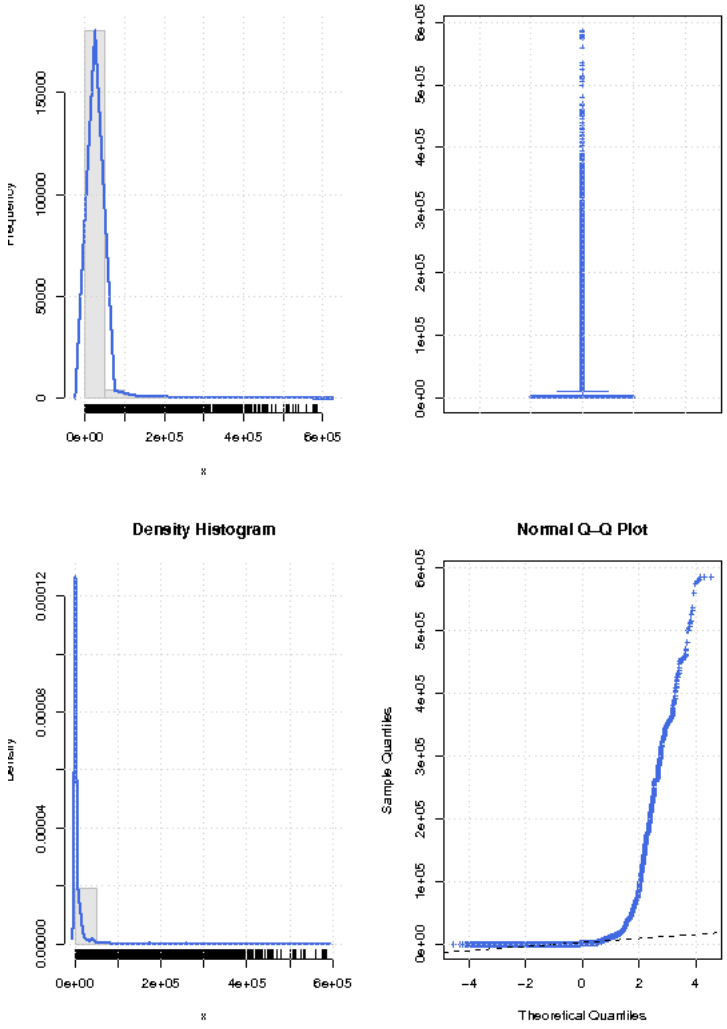


Figura 7.18: Grafico EDA per la forma della distribuzione dei Tempi di Esecuzione (livello Grid).

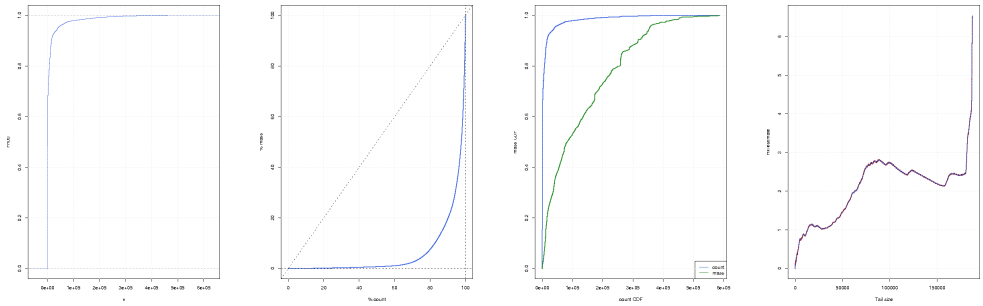


Figura 7.19: ECDF, Curva di Lorenz, Mass-Count Disparity e Hill plot dei Tempi di Esecuzione (livello Grid).

un intervallo di confidenza la cui larghezza supera di nove volte, circa, l'intervallo interquartile; la media, cioè dove tendono a concentrarsi le osservazioni, è molto lontana dal centro della distribuzione e il relativo scarto quadratico medio è pari a quattro volte il valore della media.

Dalla Fig. 7.18 risulta chiara l'asimmetria destra, il picco allungato e la coda destra molto pronunciata della densità della distribuzione dei dati. Da tutti e quattro i grafici, e in particolare dal grafico Normal Q-Q, si può osservare l'allontanamento dalla condizione di normalità; in effetti, il valore dell'asimmetria è 7 volte quello di una Normale, mentre il valore della curtosi è addirittura 63 volte, circa, quello di una Normale.

Per quanto concerne il tipo di coda della distribuzione, dalla Fig. 7.19 si sospetta la presenza di una coda lunga; per esempio, dal grafico della curva di Lorenz e da quello della Mass-Count Disparity si può esservare la tendeza dei valori estremi a dominare su quelli più probabili; tuttavia, la presenza di coda "heavy" non è certa in quanto il grafico di Hill non mostra un andamento costante; ciò potrebbe significare che la coda, pur essendo lunga, non sia "heavy".

Passando all'analisi per l'indipendenza delle osservazioni, dalla Fig. 7.20 si nota la presenza di autocorrelazione positiva per tutti i *lag* considerati. Al crescere del "lag", l'autocorrelazione sembra decrescere; è interessante quindi verificare se nel lungo termine tale dipendenza sparisca o venga mantenuta. Dalla Tab. 7.8 risulta emergere una chiara presenza di dipendenza a lungo

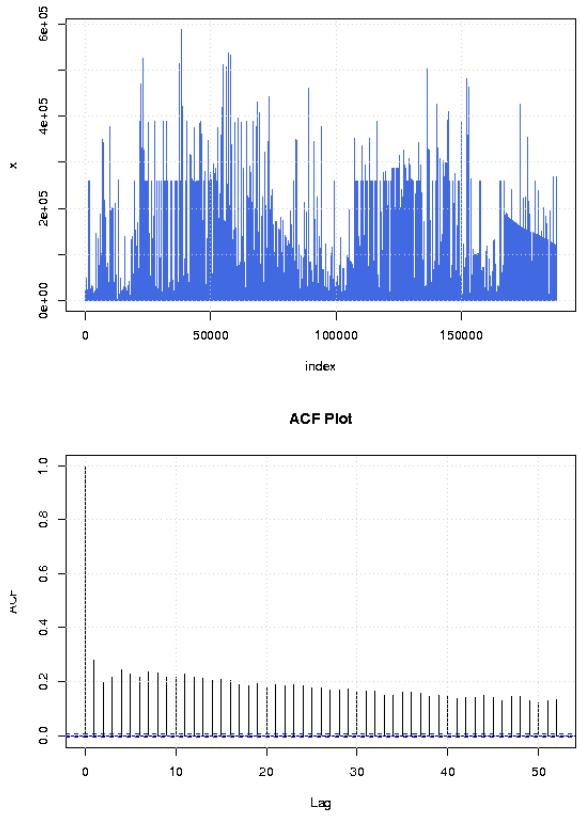


Figura 7.20: Autocorrelazione nella distribuzione dei Tempi di Esecuzione (livello Grid).

termine; la Fig. 7.21 permette di verificare la proprietà di invarianza di scala nel lungo termine.

La presenza di dipendenza tra le osservazioni, nel breve e nel lungo termine, implica la non validità di tutti quei metodi statistici che fanno uso dell'assunzione di indipendenza dei campioni (ad es., come molti dei test numerici sulla bontà dell'adattamento o il metodo del "bootstrap"). Inoltre, potrebbe essere anche un sintomo di presenza di code "heavy".

Scelta e Verifica del Modello

Dalla Fig. 7.22 si possono effettuare le seguenti considerazioni:

- per il corpo della distribuzione (grafico log-log CDF), sembra che le di-

Metodo	*H*
Variance	0.882
R/S	0.421
Periodogram	0.828

Tabella 7.8: Esponente di Hurst per i Tempi di Esecuzione (livello Grid).

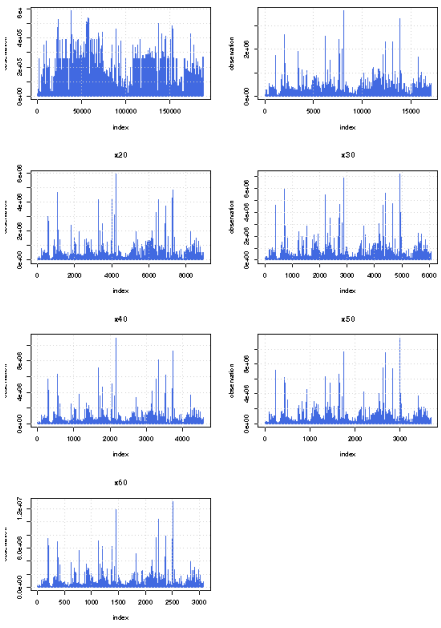


Figura 7.21: Invarianza di Scala nei Tempi di Esecuzione (livello Grid).

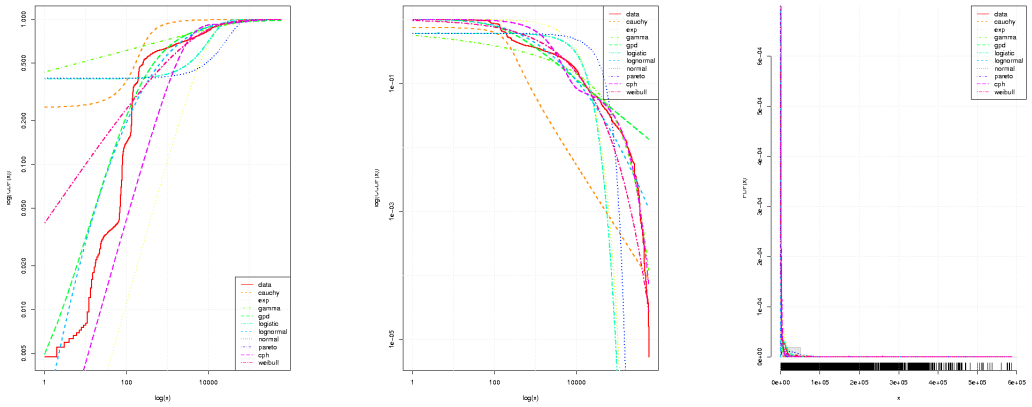


Figura 7.22: Fit per la distribuzione dei Tempi di Esecuzione (livello Grid).

stribuzioni, che meglio ne descrivono il comportamento, sono la Phase-Type continua, la Log-Normale, la Pareto Generalizzata; anche la distribuzione Esponenziale pare avere lo stesso andamento della distribuzione dei dati, benchè sembra esserci uno spostamento nella locazione;

- la coda della distribuzione (grafico log-log CCDF) sembra essere ben descritta dalla Phase-Type continua e dalla Gamma, sebbene per quest'ultima il buon adattamento pare iniziare dopo l'osservazione 10000; anche la Weibull sembra seguire l'andamento della coda, benché il comportamento nella parte terminale, cioè nelle osservazioni più estreme, si differenzia: la coda della distribuzione empirica ha un decadimento più rapido di quello della Weibull.

Per verificare la validità delle suddette affermazioni, si prendono in considerazione i grafici P-P e Q-Q. La valutazione dei coefficienti di correlazione r di Pearson e delle differenze di area relativa ΔA_r, calcolati sui due grafici (Tab. 7.9) porta alle seguenti considerazioni:

- per il corpo della distribuzione, risulta che la distribuzione Weibull ed Esponenziale siano quelle che lo descrivono nel miglior modo; in particolare, la Weibull, avendo un coefficiente di correlazione più elevato, tende a descrivere meglio il comportamento della distribuzione dei dati, mentre l'Esponenziale, con una differenza di area inferiore, imita in

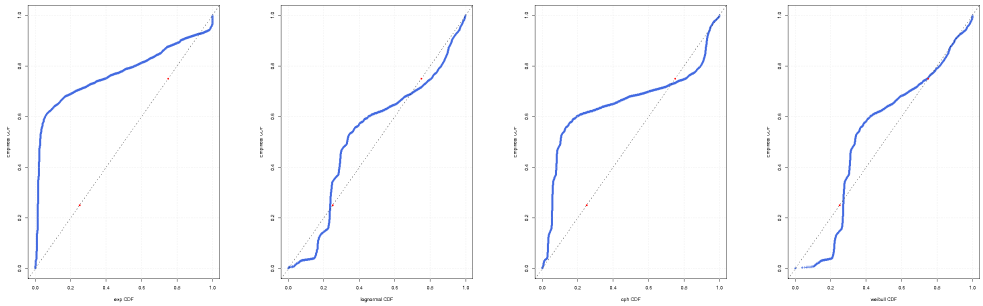


Figura 7.23: P-P plot di Esponenziale, Log-Normale, Phase-Type, Weibull rispetto ai Tempi di Esecuzione (livello Grid).

modo più fedele i valori di probabilità; oltre a queste due distribuzioni, si potrebbe anche utilizzare la Phase-Type continua, la Log-Normale, la Gamma, la Pareto Generalizzata e la Cauchy (si veda Fig. 7.24);

- riguardo alla coda della distribuzione, nessuna distribuzione sembra approssimare in modo soddisfacente la distribuzione empirica; le distribuzioni migliori risultano essere la Weibull e la Phase-Type continua, seguite dalla Gamma e, in minor misura, dalla Log-Normale (si veda Fig. 7.24).

- complessivamente, le distribuzioni che sembrano esibire il miglior adattamento sono, principalmente, la Weibull e la Phase-Type continua, e, in minor misura, la Gamma.

Riepilogo

A livello Grid, i tempi di esecuzione dei job sono caratterizzati da una dispersione molto elevata e da un'altrettanto accentuata asimmetria destra e curtosi; da queste statistiche si sospetta la presenza di code lunghe e quindi di distribuzioni sub-esponenziali.

Alcuni test grafici sulla normalità (come il grafico Normal Q-Q) mostrano, in effetti, un forte allontanamento dalla normalità; la presenza di code lunghe

Distribuzione	**Grafico**	**Pearson** r	**Pearson** r CI_5	ΔA_r
Cauchy				
	P-P	0.97017	(0.96991, 0.97044)	0.06559
	Q-Q	0.10361	(0.09914, 0.10808)	63.9201
Esponenziale				
	P-P	0.83258	(0.83118, 0.83396)	$2.66 \cdot 10^{-6}$
	Q-Q	0.77824	(0.77645, 0.78001)	10.7637
Gamma				
	P-P	0.92079	(0.92010, 0.92148)	0.00011
	Q-Q	0.98835	(0.98824, 0.98845)	0.79080
GPD				
	P-P	0.96989	(0.96962, 0.97016)	0.01362
	Q-Q	0.50119	(0.49780, 0.50457)	71.6596
Logistica				
	P-P	0.78262	(0.78086, 0.78436)	0.17866
	Q-Q	0.56897	(0.56590, 0.57202)	93.7396
Log-Normale				
	P-P	0.95787	(0.95749, 0.95824)	0.00111
	Q-Q	0.59215	(0.58921, 0.59508)	0.97617
Normale				
	P-P	0.65487	(0.65228, 0.65745)	0.18193
	Q-Q	0.52651	(0.52323, 0.52976)	96.3305
Pareto				
	P-P	0.89836	(0.89748, 0.89922)	0.98805
	Q-Q	0.04061	(0.03610, 0.04512)	1.00000
Phase-Type continua				
	P-P	0.89998	(0.89912, 0.90083)	$6.41 \cdot 10^{-5}$
	Q-Q	0.99115	(0.99107, 0.99123)	0.53356
Weibull				
	P-P	0.94264	(0.94213, 0.94314)	$2.48 \cdot 10^{-5}$
	Q-Q	0.96527	(0.96496, 0.96558)	0.35233

Tabella 7.9: Coefficienti di Correlazione e Differenze di Area dei grafici P-P e Q-Q per i Tempi di Esecuzione (livello Grid).

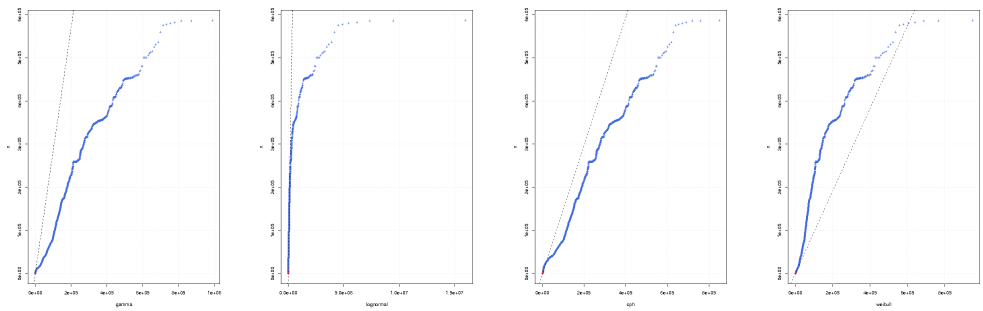


Figura 7.24: Q-Q plot di Gamma, Log-Normale, Phase-Type, Weibull rispetto ai Tempi di Esecuzione (livello Grid).

è confermata dai grafici Mass-Count Disparity e dalla Curva di Lorenz; tuttavia, non si può essere certi della presenza di una coda "heavy" in quanto il grafico dello stimatore di Hill non converge a un valore costante e non sembra nemmeno esserci una regione di stabilità.

L'analisi sull'autocorrelazione mostra una forte presenza di dipendenza a breve e lungo termine; questo ha due implicazioni principali: l'assenza di indipendenza tra le osservazioni e la possibile presenza di code "heavy".

Per quanto concerne l'adattamento di una distribuzione teorica ai dati, si è notato che la maggior parte delle distribuzioni prese in considerazione si adatta molto bene al corpo, ma non altrettanto bene alla coda; fra queste, quelle che potrebbe essere utilizzate per descrivere l'intera distribuzione dei dati sono, principalmente, la Weibull e la Phase-Type continua, e, in minor misura, la Gamma.

7.2.5 Tempi di Esecuzione – Livello VO

L'analisi dei tempi di esecuzione a livello di VO, è limitata alle VO dalle quali proviene il maggior numero di job (si veda §7.2.1); in particolare, nel presente documento, verrà descritta quella relativa alla VO *alice*.

Statistica	**Valore**	CI_5
Min	15	–
Primo Quartile	2448.75	–
Mediana	6099	(136, 21513)
Terzo Quartile	10081	–
Max	95250	–
IQR	7632.25	–
MAD	5708.01	–
Asimmetria	3.52589	–
Quartile-Asimmetria	0.04347	–
Curtosi	21.0677	–
Curtosi Eccesso	18.0677	–
Media	7772.230	(7632.395, 7912.065
Deviazione Standard	8553.145	(8455.402, 8653.190)
CV	1.10047	–

Tabella 7.10: Riepilogo delle misure di centralità e dispersione dei Tempi di Esecuzione (livello VO: ALICE).

Organizzazione Virtuale ALICE

Bonifica dei Dati

Non si sono riscontrate particolari anomalie nei dati; la Fig. 7.25 mostra molti valori al di fuori di ciò che può essere interpretato come il corpo della distribuzione; il numero di questi valori è così grande che non possono essere considerati degli "outlier". Ne segue che nessuna osservazione verrà esclusa dall'analisi statistica.

Analisi delle Proprietà Statistiche

In Tab. 7.10 sono riassunte le principali misure di centralità e dispersione ricavate dai 14372 tempi di esecuzione relativi alla VO *alice*. Dalla tabella si nota come i dati siano altamenti dispersi rispetto al valor medio e come tendano a concentrarsi in un punto molto distante dal centro della distribuzione (mediana). L'asimmetria, non troppo accentuata, è destra e il valore della curtosi indica un picco abbastanza alto e stretto.

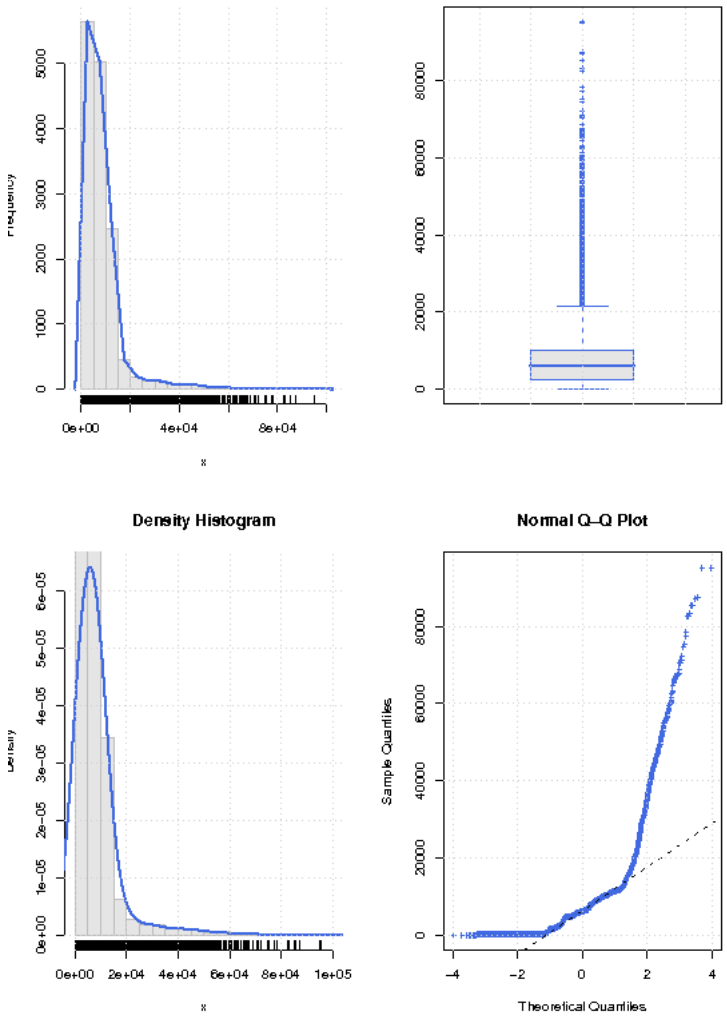


Figura 7.25: Grafico EDA per la forma della distribuzione dei Tempi di Esecuzione (livello VO: ALICE).

Metodo	H
Variance	0.937
R/S	0.941
Periodogram	1.154

Tabella 7.11: Esponente di Hurst per i Tempi di Esecuzione (livello VO: *alice*).

Per quanto riguarda l'autocorrelazione, dalla Fig. 7.26, si nota un'elevata dipendenza a breve termine; questo fatto è interessante, in quanto a livello Grid, la dipendenza a breve termine era presente in modo meno accentuato; questo potrebbe significare che a livello Grid la dipendenza fra le osservazioni potrebbe essere attenuata a causa della sovrapposizione dei tempi di esecuzione di varie VO. Anche la dipendenza a lungo termine sembra essere presente in modo marcato; ciò puó essere osservato dalla Tab. 7.11, la quale mostra la stima dell'esponente di Hurst, ottenuta applicando diversi metodi; il fatto che la stima dell'esponente di Hurst fornita del metodo Periodogram sia maggiore di 1 può essere dovuto alla presenza di particolari "pattern" nei dati, come dei "trend' [63]. Dalla Fig. 7.27 è possibile notare un effetto della dipendenza a lungo termine: la proprietà di invarianza di scala. La presenza di dipendenza nel breve e nel lungo termine, causa la inapplicabilità di tutti quei metodi statistici che effettuano un'assunzione di indipendenza sui campioni, come molti dei metodi numerici per la verifica della bontà dell'adattamento di una distribuzione a un insieme di dati. Inoltre, la dipendenza a lungo termine può essere un segnale di presenza di code "heavy".

Dato che è stata osservata una probabile presenza di dipendenza a lungo termine, è interessante capire che tipo di coda possiede la distribuzione dei dati. Dalla Fig. 7.28 non si nota nessun particolare segno di code "heavy": la Curva di Lorenz non è eccessivamente sbilanciata, il grafico Mass-Count Disparity non mostra nessuna dominanza della "count distribution" sulla "mass distribution", il grafico dello stimatore di Hill diverge al crescere della dimensione della coda.

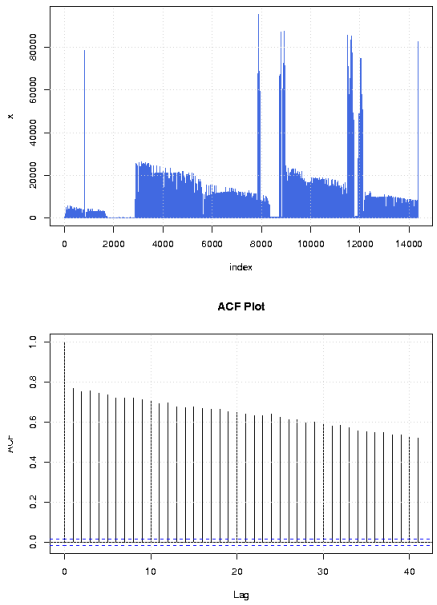


Figura 7.26: Autocorrelazione nella distribuzione dei Tempi di Esecuzione (livello VO: *alice*).

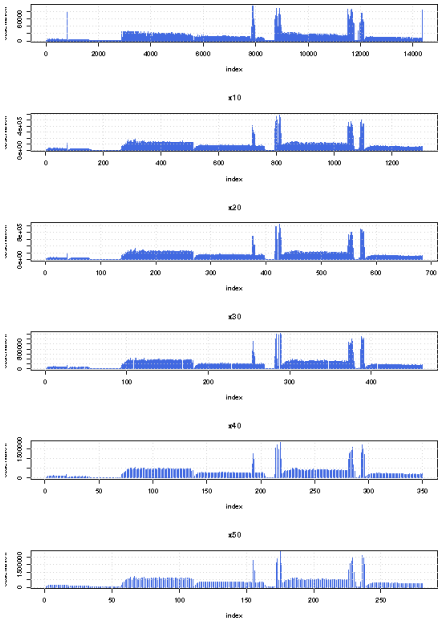


Figura 7.27: Invarianza di Scala nei Tempi di Esecuzione (livello VO: *alice*).

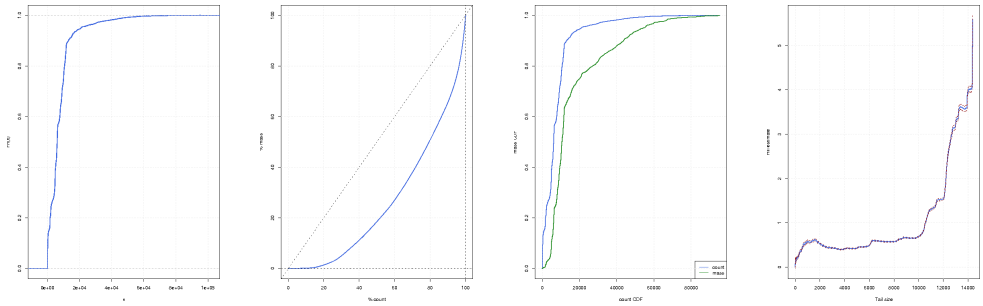


Figura 7.28: ECDF, Curva di Lorenz, Mass-Count Disparity e Hill plot dei Tempi di Esecuzione (livello VO: *alice*).

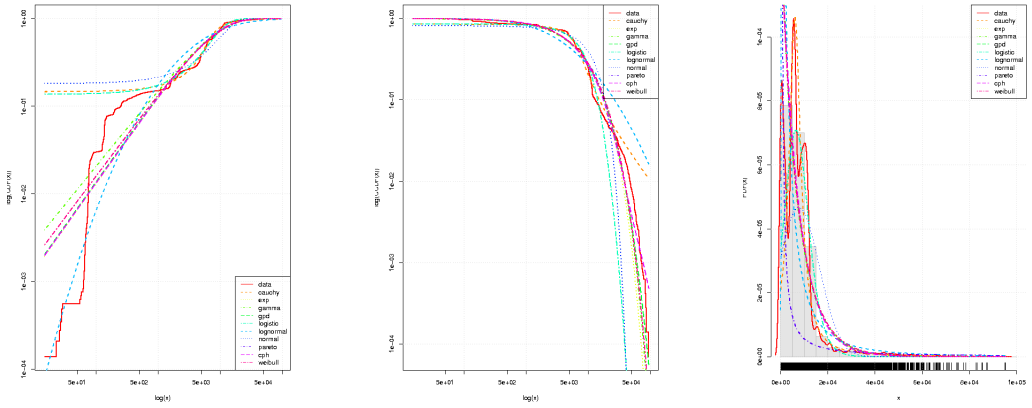


Figura 7.29: Fit per la distribuzione dei Tempi di Esecuzione (livello VO: *alice*).

Scelta e Verifica del Modello

La Fig. 7.29 mostra un buon adattamento alla coda della distribuzione empirica per la maggior parte delle distribuzioni; invece, al contrario di quanto di solito succede, il corpo della distribuzione sembra essere ben approsimato solo dalla distribuzione Log-Normale, dalla Gamma, dalla Weibull e dalla Phase-Type.

Per verificare le precedenti affermazioni, si procede all'analisi dei grafici P-P e Q-Q. La valutazione dei coefficienti di correlazione r di Pearson e delle differenze di area relativa ΔA_r (Tab. 7.12) porta alle seguenti considerazioni:

- per il corpo della distribuzione empirica, tutte le distribuzioni, al di fuo-

ri della Pareto, si comportano molto bene; quelle in assoluto migliori sono la Weibull e la Gamma, seguite dalla Esponenziale e dalla Pareto Generalizzata (Fig. 7.30);

- per la coda della distribuzione empirica, le migliori distribuzioni sono la Weibull, la Gamma, l'Esponenziale e la Phase-Type continua (Fig. 7.31);
- per l'adattamento dell'intera distribuzione empirica, le migliori distribuzioni sono, principalmente, la Weibull e la Gamma, e, in minor misura, l'Esponenziale e la Phase-Type continua.

Come si può notare la Log-Normale non rientra fra le distribuzioni migliori; eppure dalla Fig. 7.29 sembrava descrivere molto bene il corpo della distribuzione; questa esclusione può essere motivata dal fatto che nel grafico P-P, usato per individuare le distribuzioni che meglio si avvicinano al corpo della distribuzione empirica, hanno anche una certa influenza i valori della coda; dato che la coda della Log-Normale si discosta molto da quella della distribuzione empirica, ciò contribuisce alla sua esclusione dalle distribuzioni migliori. In effetti per poter effettuare una reale distinzione tra la migliore distribuzione per il corpo e quella per la coda della distribuzione empirica occorrerebbe trattare le due parti separatamente; questo però implica la scelta di un opportuno valore fra le osservazioni che funga da punto di divisione tra corpo e coda; purtroppo questa scelta non è banale.

Dalla Tab. 7.12 si può osservare che la distribuzione Esponenziale sembra comportarsi meglio della distribuzione Phase-Type continua; questo sembrerebbe un controsenso in quanto la Phase-Type continua è, in un certo senso, una generalizzazione della distribuzione Esponenziale. In effetti questo comportamento è quasi sicuramente causato dall'utilizzo di differenti tecniche di adattamento ai dati: per l'Esponenziale è stato utilizzato il metodo MLE, mentre per la Phase-Type continua si è usato il metodo dei momenti (tecnica che in generale fornisce stime meno precise di quelle ottenute tramite il metodo MLE).

Le Fig. 7.30 e Fig. 7.31 mostrano, rispettivamente, i grafici P-P e Q-Q delle distribuzioni risultate migliori dal punto di vista dell'adattamento ai dati.

Distribuzione	**Grafico**	**Pearson** r	**Pearson** r CI_5	ΔA_r
Cauchy				
	P-P	0.99212	(0.99186, 0.99237)	0.03353
	Q-Q	0.19709	(0.18133, 0.21276)	2.41170
Esponenziale				
	P-P	0.97836	(0.97765, 0.97905)	$2.64 \cdot 10^{-5}$
	Q-Q	0.95597	(0.95454, 0.95735)	0.32703
Gamma				
	P-P	0.97371	(0.97285, 0.97454)	$5.09 \cdot 10^{-6}$
	Q-Q	0.95929	(0.95797, 0.96057)	0.24430
GPD				
	P-P	0.97557	(0.97477, 0.97635)	$2.51 \cdot 10^{-5}$
	Q-Q	0.59087	(0.58013, 0.60141)	1.53075
Logistica				
	P-P	0.99271	(0.99247, 0.99295)	0.01908
	Q-Q	0.84206	(0.83724, 0.84676)	3.43374
Log-Normale				
	P-P	0.93848	(0.93650, 0.94040)	0.01647
	Q-Q	0.84805	(0.84339, 0.85258)	0.91917
Normale				
	P-P	0.96452	(0.96336, 0.96564)	0.03430
	Q-Q	0.82245	(0.81709, 0.82767)	4.12953
Pareto				
	P-P	0.92710	(0.92477, 0.92936)	0.91972
	Q-Q	0.08551	(0.06926, 0.10172)	1.00000
Phase-Type continua				
	P-P	0.97686	(0.97610, 0.97759)	0.00045
	Q-Q	0.96727	(0.96620, 0.96830)	0.36052
Weibull				
	P-P	0.97585	(0.97505, 0.97661)	$6.90 \cdot 10^{-6}$
	Q-Q	0.96052	(0.95924, 0.96177)	0.21649

Tabella 7.12: Coefficienti di Correlazione e Differenze di Area dei grafici P-P e Q-Q per i Tempi di Esecuzione (livello VO: *alice*).

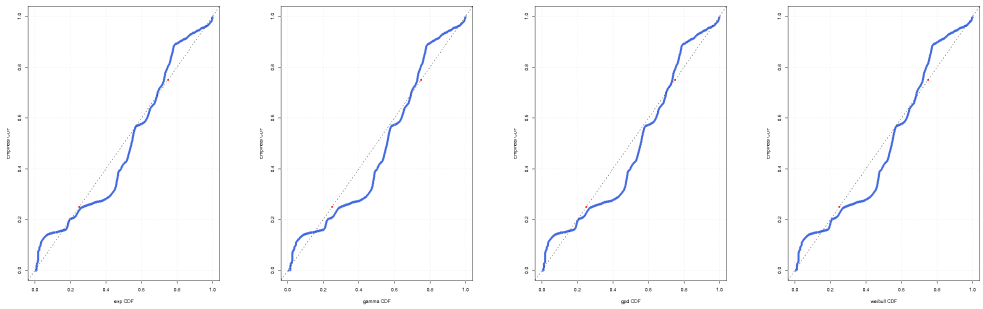


Figura 7.30: P-P plot di Esponenziale, Gamma, GPD, Weibull rispetto ai Tempi di Esecuzione (livello VO: *alice*).

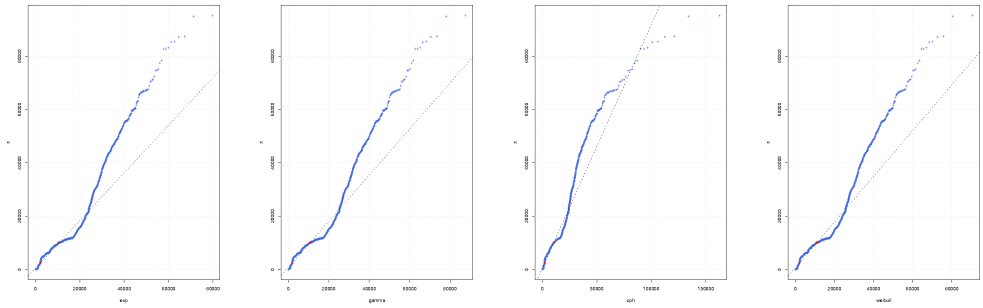


Figura 7.31: Q-Q plot di Esponenziale, Gamma, Phase-Type, Weibull rispetto ai Tempi di Esecuzione (livello VO: *alice*).

Riepilogo

I tempi di esecuzione relativi alla sola VO *alice* sono caratterizzati da un'alta dispersione e asimmetria destra; da ciò, risulta evidente un netto allontamento dalla normalità da parte della distribuzione dei dati.

Questo fatto è ulteriormente confermato dalla presenza molto accentuata di autocorrelazione a breve e a lungo termine; tale caratteristica rende quindi inutilizzabile tutte le tecniche statistiche che prevedono indipendenza nei campioni (ad es., test numerici per l'adattamento, tecnica del "bootstrap", ...).

La presenza di dipendenza a lungo termine potrebbe essere un segno di code "heavy"; tuttavia, dalla valutazione della Curva di Lorenz, del grafico Mass-Count Disparity e del grafico dello stimatore di Hill, non si osserva nessuna presenza di una coda con questa caratteristica.

Le distribuzioni che meglio descrivono l'intera distribuzione empirica sono risultate essere, principalmente, la Weibull e la Gamma, e, in minor misura, l'Esponenziale e la Phase-Type continua.

7.3 Riepilogo e Considerazioni Finali

Dall'analisi della traccia LCG è emerso che vi è una forte dispersione e asimmetria fra le osservazioni di ogni caratteristica del carico presa in considerazione, sia a livello Grid, sia a livello VO. Le osservazioni sono inoltre caratterizzate dalla presenza di autocorrelazione; in particolare l'autocorrelazione a breve termine sembra che sia il tipo di dipendenza più forte e quella più presente nella maggior parte dei casi analizzati. La presenza di dipendenza a breve termine può essere l'effetto di due cause:

- i job analizzati potrebbero far parte di Bag-of-Task; ciò implica che la sottomissione di job avvenga sottoforma di gruppi di task alla volta; in tal modo sia i tempi di interarrivo sia quelli di esecuzione dei job ne sono influenzati: gruppi di job arrivano a uno scheduler a istanti di tempo ravvicinati e la durata della loro esecuzione risulta generalmente simile;

- la maggioranza dei job (il 90% circa) proviene solo da 5 VO; questo può

determinare un sovraccarico degli scheduler locali, contribuendo quindi a introdurre dipendenza tra le caratteristiche del carico a livello VO.

La presenza di dipendenza a lungo termine non sembra essere causata da distribuzioni "heavy tailed"; difatti le distribuzioni che sono risultate più adeguate per descrivere il comportamento dell'intera distribuzione dei dati sono la Gamma, la Phase-Type continua e la Weibull; fra le altre distribuzioni che potrebbero essere usate per descrivere il corpo o la coda della distribuzione empirica, vi è la Log-Normale, l'Esponenziale e la Pareto Generalizzata. In seguito alla presenza di dipendenza a lungo termine, l'unico modo effettivo per verificare la bontà dell'adattamento è stata la valutazione dei grafici P-P e Q-Q, insieme ai relativi coefficienti di correlazione lineare e alle differenze di area relativa.

A causa dell'elevata dispersione e asimmetria nei dati, un possibile futuro approccio all'analisi da adottare potrebbe essere quello di dividere la fase di adattamento in due parti: l'adattamento del corpo della distribuzione e quello della coda. Per applicare questo tipo di analisi, occorre però sviluppare delle tecniche che permettano di trovare il punto di "taglio" che divide il corpo di una distribuzione dalla relativa coda. Si tratta di un aspetto della statistica ancora sotto studio, per il quale sembrano non esistere, al momento, delle tecniche generali. Inoltre, vista la presenza di autocorrelazione sia a breve sia a lungo termine, potrebbe essere necessario valutare l'utilizzo di particolari processi stocastici, come i processi *Markov Modulated Poisson Process* (*MMPP*) o i più generici *Markov Arrival Process* (*MAP*), per cercare di includere nel modello teorico alcune delle dipendenze scoperte empiricamente.

Infine, potrebbe essere interessante valutare l'eventuale dipendenza tra caratteristiche del carico differenti tramite un'analisi multivariata; ciò è motivato dalle considerazioni effettuate in §7.2.1; per esempio, si è notato che la distribuzione del numero dei job risulta essere sbilanciata verso poche entità (ad es., utenti, VO, ...); potrebbe, quindi, essere interessante capire se ci sia qualche tipo di dipendenza incrociata fra queste entità e i tempi di interarrivo o di esecuzione, sia a livello Grid sia a livello VO.

Capitolo 8

Analisi della traccia TeraGrid

La traccia analizzata in questo capitolo è stata generata attraverso il sistema *TeraGrid* [3], un'infrastruttura collaborativa fondata dal *National Science Foundation* (*NSF*) e coordinata dal *Grid Infrastructure Group* (*GIG*) dell'Università di Chicago; è costituita da diversi *resource provider* (*RP*) e, al momento, include più di 250 Teraflops di potenza computazionale e più di 30 Petabyte di capacità di memorizzazione.

Il capitolo è organizzato nel seguente modo: la sezione §8.1 descrive il formato interno della traccia e il significato dei vari campi che compongono ogni sua riga; la sezione §8.2 è dedicata all'analisi statistica della traccia e, in particolare, allo studio della distribuzione dei tempi di interarrivo dei job (§8.2.2) e dei relativi tempi di esecuzione (§8.2.3); infine, la sezione §8.3 fornisce un riepilogo dell'analisi effettuata sulla traccia e le relative considerazioni.

8.1 Formato

La traccia contiene la registrazione dell'esecuzione di 162362 job nel periodo compreso tra il 2004-01-05 e il 2005-08-10.

Ogni riga della traccia, relativa all'esecuzione di un particolare job, ha il seguente formato:

```
<timestamp> <nodes> <starttime> <endtime>
```

dove:

`timestamp`: data di sottomissione del job;

`nodes`: numero di nodi del cluster utilizzati dal job per l'esecuzione;

`starttime`: istante di inizio dell'esecuzione del job;

`endtime`: istante di fine esecuzione del job.

Sebbene non specificato il campo `nodes` potrebbe essere utilizzato come informazione sulla dimensione di una Bag-of-Task; tuttavia, per poter effettuare un'accurata analisi basata su BoT occorrerebbe conoscere le informazioni sui singoli task che compongono una certa BoT; in particolare, ciò che manca è l'effettiva durata dell'esecuzione di ogni task appartenente a una BoT; con i dati che si hanno a disposizione, l'unica ipotesi che potrebbe essere effettuata è quella di uguale durata di esecuzione per ogni task, cioè:

$$\langle \text{BoT task duration} \rangle = \frac{\langle \text{endtime} \rangle - \langle \text{starttime} \rangle}{\langle \text{nodes} \rangle}$$

Tuttavia tale assunzione, oltre a non essere in grado di provarne la correttezza, non aggiungerebbe nessun aspetto interessante rispetto all'analisi individuale dei tempi di interarrivo e di esecuzione; per tali motivi, nel resto del capitolo, non verrà presa in considerazione.

8.2 Analisi Statistica

L'analisi dei tempi di interarrivo e di esecuzione può essere effettuata solo a livello Grid, in quanto nella traccia mancano le informazioni relative alle Organizzazioni Virtuali.

Il resto della sezione prevede la descrizione delle proprietà statistiche generali della traccia (§8.2.1) e quindi di quelle dei tempi di interarrivo dei job (§8.2.2) e dei relativi tempi di esecuzione (§8.2.3).

8.2.1 Caratteristiche Generali

La traccia è composta da 162362 job osservati nel periodo compreso tra il 2004-01-05 e il 2005-08-10. Attraverso una procedura di "pre-processing" del file di log è stata rilevata la presenza di alcuni dati anomali; i tipi di anomalie riscontrate sono:

1. `<timestamp> > <starttime>`: l'arrivo del job è avvenuto dopo l'inizio dell'esecuzione dello stesso job. Le righe della traccia affette da questa anomalia devono essere rimosse o corrette in modo opportuno (ad es., impostando il campo `timestamp` uguale al campo `starttime`).

2. `<starttime> > <endtime>`: la terminazione dell'esecuzione di un job è avvenuta prima che il job venisse eseguito. Una riga della traccia con un'anomalia di questo tipo è sicuramente da rimuovere in quanto rappresenta una condizione di errore.

3. `<starttime> = <endtime>`: la terminazione dell'esecuzione di un job è avvenuta nello stesso momento in cui l'esecuzione è iniziata. La presenza di una riga nella traccia con un'anomalia di questo tipo può essere dovuta:

   - alla risoluzione temporale non abbastanza fine (cioè l'esecuzione di un job è durata meno di un secondo);
   - al fallimento dell'inizio dell'esecuzione del job;
   - a una generica condizione di errore.

   Nel primo caso, la riga della traccia non deve essere rimossa in quanto contribuisce a caratterizzare il carico (presenza di job molto corti). Negli altri casi, invece, l'osservazione associata alla riga rappresenta un "outlier" e deve essere rimossa. Non avendo sufficienti informazioni sull'origine di queste osservazioni, e quindi per evitare di influenzare erroneamente le statistiche (introducendo un peso sulla numerosità ma non sulla massa), si è deciso di escluderle dall'analisi statistica.

	Interarrivo		**Esecuzione**	
	Full	*All Fixed*	*Full*	*All Fixed*
Numerosità	162362.0	162310.0	162362.0	162310.0
Min	0.0	0.0	0.0	1.0
Primo Quartile	4.0	4.0	3.0	3.0
Mediana	49.0	49.0	3.0	3.0
Media	310.2	310.3	3627.6	3629.0
Terzo Quartile	217.0	217.0	6.0	6.0
Max	314338.0	318245.0	359030.0	35900.0
Std. Dev.	1741.0	1796.8	21612.6	21616.9
CV	5.6	5.8	6.0	6.0

Tabella 8.1: Statistiche sui Tempi di Interarrivo e di Esecuzione con e senza anomalie.

4. `nodes` ≤ 0: il job è stato assegnato a un numero di nodi nullo o negativo. Una riga della traccia avente questo tipo di anomalia deve essere rimossa.

L'analisi della traccia per la rilevazione delle suddette anomalie ha riscontrato: 15 anomalie del tipo 1 e 52 anomalie di tipo 3 (i due tipi di anomalie colpiscono differenti righe della traccia). La correzione delle anomalie (tramite la correzione di quelle del tipo 1 e la rimozione di quelle del tipo 3) ha ridotto l'insieme dei dati da 162362 a 162310 osservazioni; in Tab. 8.1 e in Fig. 8.1 è mostrata la distribuzione dei tempi di interarrivo e della durata dell'esecuzione, prima (etichetta "full") e dopo (etichetta "all fixed") la correzione di tutte le anomalie; come si può notare, la distribuzione della durata dell'esecuzione ha subito solamente una piccola variazione; i tempi di interarrivo, invece, sono stati maggiormente influenzati dall'eliminazione delle anomalie. L'analisi statistica presentata nelle prossime sezioni è basata sulla versione della traccia privata delle anomalie.

Prima di passare all'analisi dei tempi di interarrivo e della durata delle esecuzioni, è utile capire come sono distribuiti i job nell'arco temporale considerato e la relativa dimensione caratteristica (in termini di numero di nodi computazionali). In Fig. 8.2 è mostrata la distribuzione del numero di job suddivisi per data (figura a sinistra) e per numero di nodi (figura a destra). Dalla

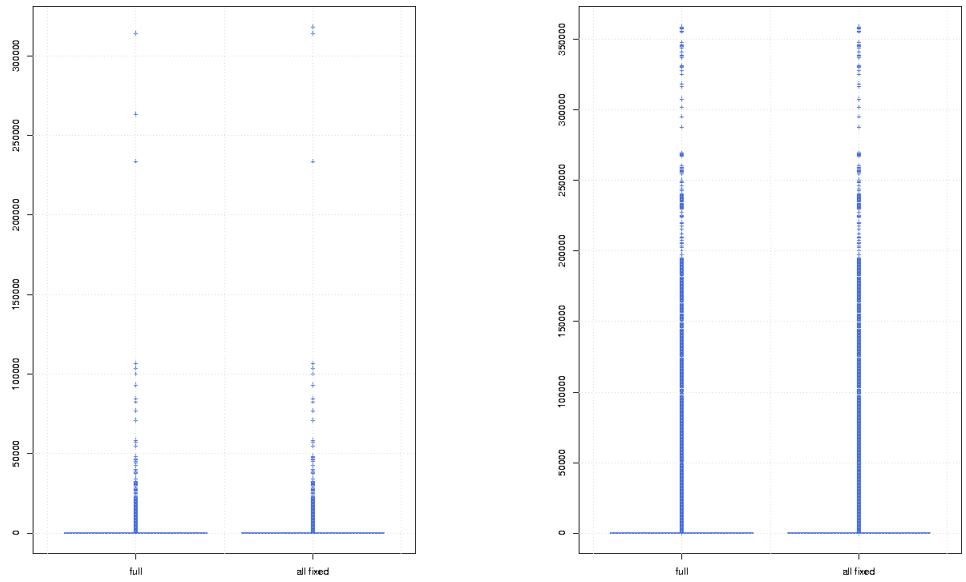


Figura 8.1: Tempi di Interarrivo (a sinistra) e di Esecuzione (a destra) con e senza anomalie.

distribuzione del numero di job per data, si nota un picco elevato che ha il suo massimo nella giornata del 2004-10-27, in cui sono stati sottomessi 3155 job; benché si tratta di un caso fuori dal comune non è possibile escluderlo dalla traccia in quanto non si potrebbe fornire una spiegazione plausibile. Per quanto riguarda la distribuzione del numero di job per numero di nodi di computazione, si nota che la maggior parte dei job (per la precisione, 140984 job, che corrispondono al 86.9% del numero totale di job) ha dimensione 1; ciò significa che un Bag-of-Task non è la principale tipologia di job che viene eseguita nel sistema Grid considerato.

8.2.2 Tempi di Interarrivo

Bonifica dei Dati

Dalla Fig. 8.3 si nota una grande concentrazione di dati nella zona comprendente le osservazioni tra 0 e 50000 circa e una presenza di pochi valori estremi che superano il valore 300000 (cioè 6 volte il valore massimo della zona ad alta

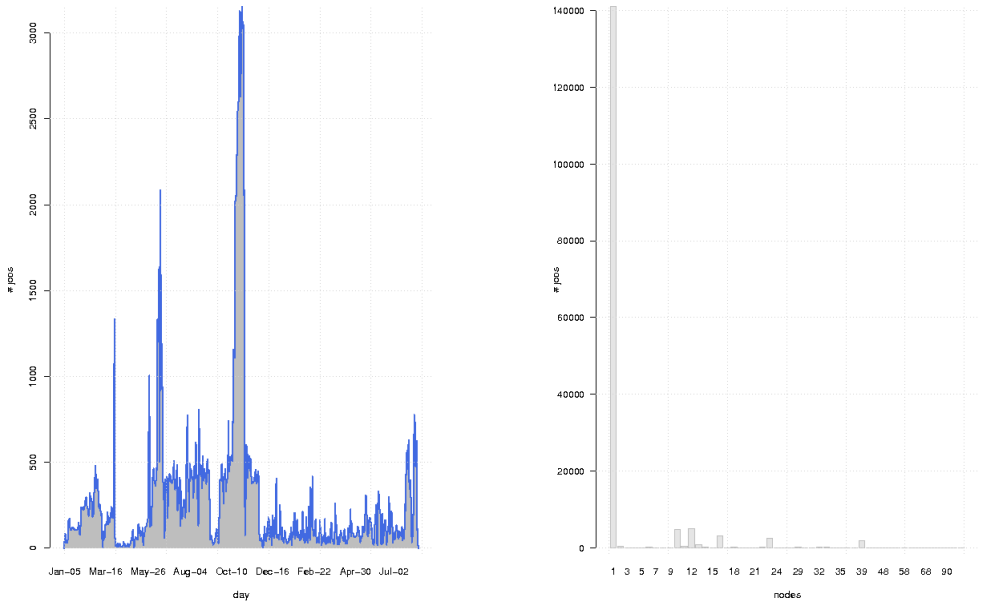


Figura 8.2: Numero di Job per Data (a sinistra) e per Nodi (a destra).

densità). Dato che la presenza di questi valori potrebbe essere un segnale di coda lunga, non sono stati esclusi dall'analisi statistica.

Analisi delle Proprietà Statistiche

Il numero di osservazioni presenti nella versione della traccia senza anomalie è pari a 162310, delle quali 162295, ossia il 99.991% risulta minore o uguale 50000; la distribuzione mostra una grande dispersione in quanto il valore massimo, cioè 318245, risulta molto maggiore del valore più caratteristico. Ciò può essere osservato dalle statistiche sulla centralità e dispersione riassunte in Tab. 8.2; il centro della distribuzione e il valore medio sono molto distanti, l'intervallo interquartile e deviazione standard hanno un valore elevato e molto differente. Anche l'asimmetria (destra) è molto grande (90 volte più accentuata di una Normale) e la curtosi supera di addirittua 4 ordini di grandezza quello di una Normale.

Dal grafico dell'autocorrelazione (Fig. 8.4) si osserva la presenza di correlazione a breve termine per tutti i valori del lag considerati; inoltre, dalla stima

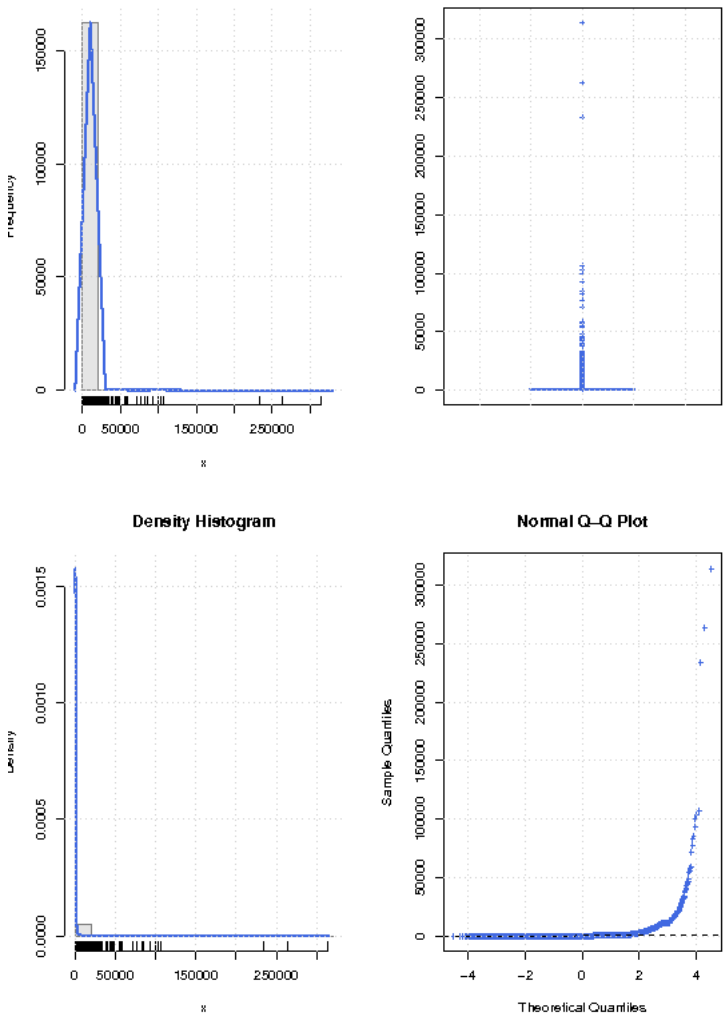


Figura 8.3: Grafico EDA per la forma della distribuzione dei Tempi di Interarrivo.

Statistica	**Valore**	CI_5
Min	0	–
Primo Quartile	4	–
Mediana	49	$(0, 1000)$
Terzo Quartile	217	–
Max	318245	–
IQR	213	–
MAD	71.165	–
Asimmetria	90.696	–
Quartile-Asimmetria	0.577	–
Curtosi	13948.337	–
Curtosi Eccesso	13945.337	–
Media	310.318	$(301.577, 319.060)$
Deviazione Standard	1796.820	$(1790.66, 1803.022)$
CV	5.790	–

Tabella 8.2: Riepilogo delle misure di centralità e dispersione dei Tempi di Interarrivo.

Metodo	H
Variance	0.788
R/S	0.602
Periodogram	1.1.183

Tabella 8.3: Esponente di Hurst per i Tempi di Interarrivo.

dell'esponente di Hurst (Tab. 8.3) risulta essere presente anche dipendenza a lungo termine, fatto confermato anche dall'invarianza di scala mostrata in Fig. 8.5 (ad es., si confrontino gli aggregati da $\times 20$ a $\times 60$). Il fatto che il metodo Periodogram fornisca un valore dell'esponente di Hurst superiore a 1 può essere dovuto alla presenza di eventuali "pattern" nei dati [63]. La presenza di autocorrelazione implica la inapplicabilità dei metodi statistici che assumono l'indipendenza dei dati; inoltre la presenza di dipendenza a lungo termine può significare che il tipo di coda della distribuzione sia di tipo "heavy".

Per analizzare il tipo di coda della distribuzione, si ricorre ai grafici della Curva di Lorenz, al Mass-Count Disparity e allo stimatore di Hill (Fig. 8.6); dal grafico della Curva di Lorenz, benchè si osservi uno sbilanciamento verso la curva di perfetta iniquità, non si nota una particolare predominanza di pochi

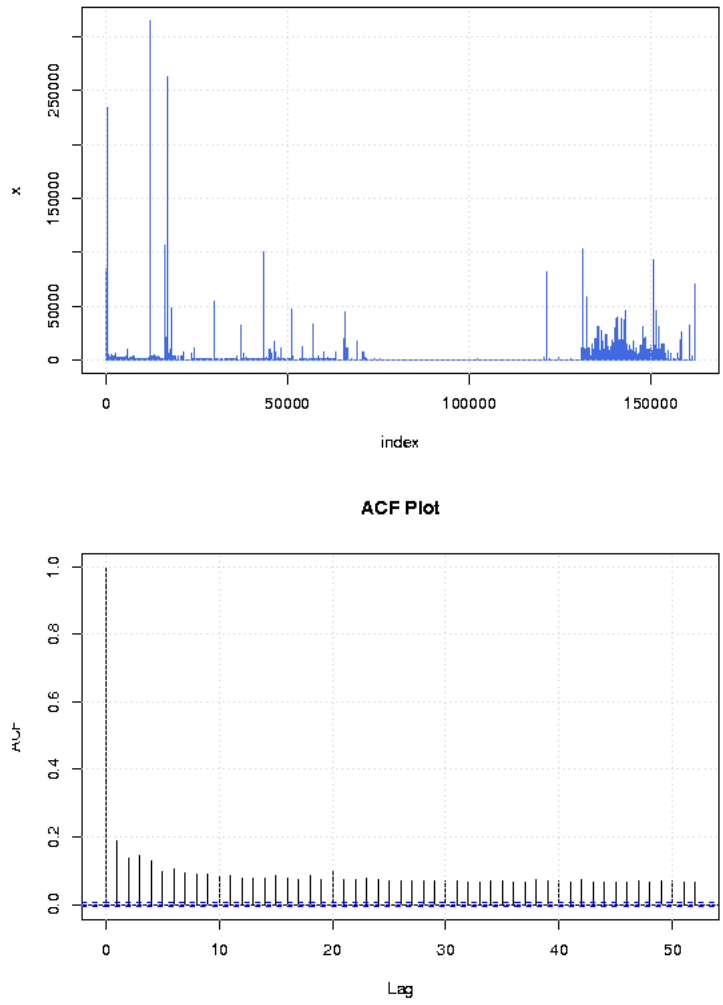


Figura 8.4: Autocorrelazione nella distribuzione dei Tempi di Interarrivo.

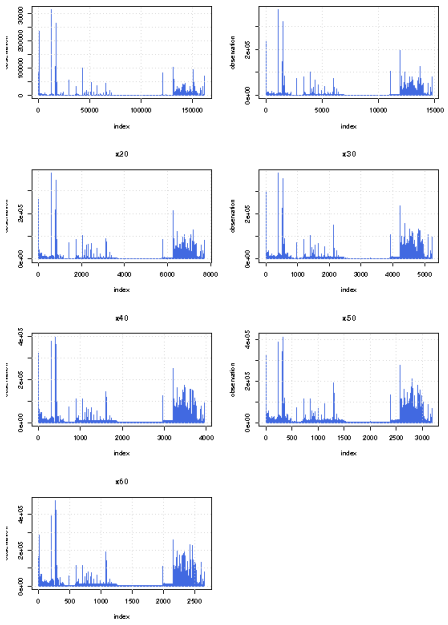


Figura 8.5: Invarianza di Scala nei Tempi di Interarrivo.

valori estremi; il grafico della Mass-Count Disparity mostra chiaramente l'assenza di predominanza dei valori estremi: ai primi 50000 valori (cioè quelli più frequenti) risulta associata la quasi totalità della massa; infine, il grafico dello stimatore di Hill mostra una divergenza al crescere della dimensione della coda.

Scelta e Verifica del Modello

In Fig. 8.7 sono mostrati i grafici log-log CDF, log-log CCDF e PDF relativi all'adattamento di alcune distribuzioni teoriche ai tempi di interarrivo; il corpo della distribuzione sembra ben descritto dalla Pareto Generalizzata, dalla Log-Normale, dalla Weibull e dalla Phase-Type continua (grafico log-log CDF); per quanto riguarda la coda, si reputa che le distribuzioni migliori siano la Log-Normale e la Phase-Type continua.

La verifica dell'adattamento non può essere effettuata con i test numerici a causa della dipendenza tra le osservazioni; perciò la bontà dell'adattamento viene valutata in base ai grafici P-P e Q-Q (Fig. 8.8 e Fig. 8.9, rispettivamente), e

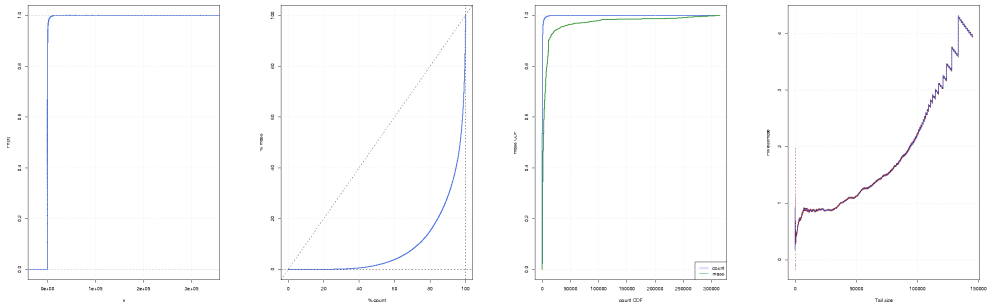


Figura 8.6: ECDF, Curva di Lorenz, Mass-Count Disparity e Hill plot dei Tempi di Interarrivo.

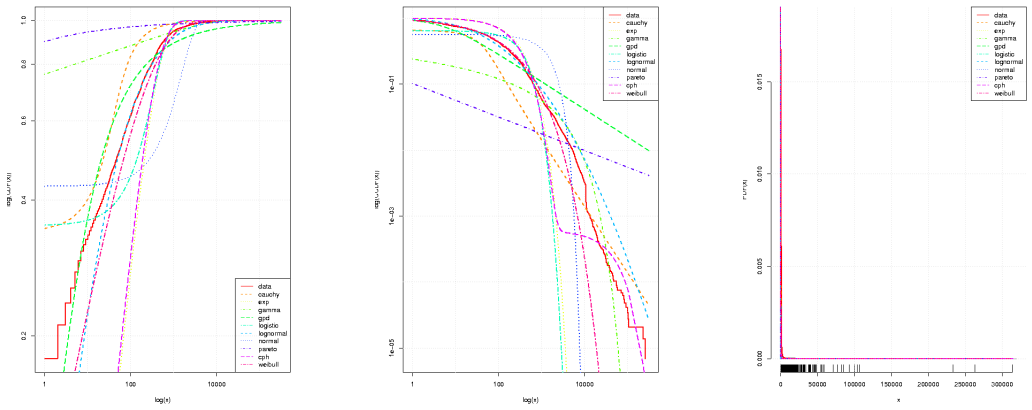


Figura 8.7: Fit per la distribuzione dei Tempi di Interarrivo.

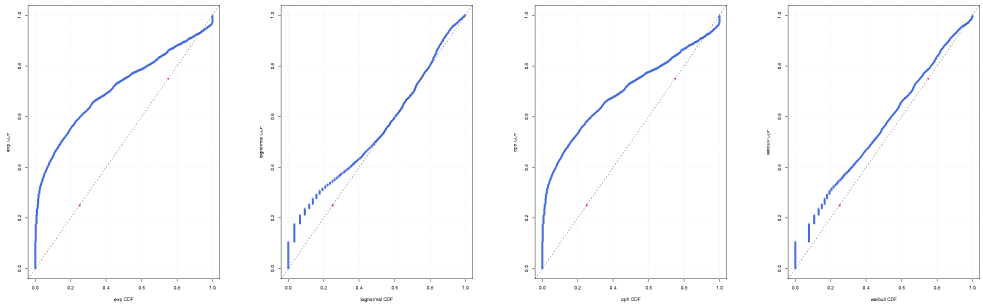


Figura 8.8: P-P plot di Esponenziale, Log-Normale, Phase-Type, Weibull rispetto ai Tempi di Interarrivo.

all'analisi dei relativi coefficienti di correlazione r di Pearson e delle differenze di area relativa ΔA_r (Tab. 8.4):

- per il corpo della distribuzione, le migliori distribuzioni risultano essere la Weibull, la Log-Normale, la Phase-Type continua e l'Esponenziale;
- per la coda della distribuzione, le uniche distribuzioni che mostrano un discreto adattamento sono la Log-Normale e la Phase-Type continua; fra le rimanenti, la Weibull e la Pareto sembrano quelle migliori;
- complessivamente, l'intera distribuzione empirica è descritta bene dalla Log-Normale e dalla Phase-Type, seguite dalla Weibull e, in minor misura, dalla Pareto.

Di seguito si riportano i valori dei parametri, stimati dall'insieme delle osservazioni, delle distribuzioni che sono risultate le migliori dal punto di vista dell'adattamento:

- *Esponenziale*.

Metodo	MLE
Rate	0.003222499

- *Log-Normale*.

Distribuzione	**Grafico**	**Pearson** r	**Pearson** r CI_5	ΔA_r
Cauchy				
	P-P	0.96901	(0.96871, 0.96931)	0.13115
	Q-Q	0.51395	(0.51036, 0.51752)	941.348
Esponenziale				
	P-P	0.92524	(0.92454, 0.92594)	$3.08 \cdot 10^{-6}$
	Q-Q	0.49593	(0.49226, 0.49959)	131.125
Gamma				
	P-P	0.68178	(0.67917, 0.68437)	$3.08 \cdot 10^{-6}$
	Q-Q	0.76486	(0.76284, 0.76688)	0.99224
GPD				
	P-P	0.97254	(0.97227, 0.97280)	0.00453
	Q-Q	0.29744	(0.29300, 0.30186)	797.121
Logistica				
	P-P	0.86320	(0.86196, 0.86444)	0.14080
	Q-Q	0.35866	(0.35441, 0.36289)	980.523
Log-Normale				
	P-P	0.99351	(0.99345, 0.99357)	$2.26 \cdot 10^{-5}$
	Q-Q	0.96371	(0.96337, 0.96406)	0.64425
Normale				
	P-P	0.65227	(0.64947, 0.65506)	0.22871
	Q-Q	0.32218	(0.31781, 0.32653)	1198.11
Pareto				
	P-P	0.95260	(0.95215, 0.95305)	0.99849
	Q-Q	0.44552	(0.44161, 0.44941)	1.00000
Phase-Type continua				
	P-P	0.93233	(0.93169, 0.93296)	$1.40 \cdot 10^{-6}$
	Q-Q	0.85566	(0.85435, 0.85696)	0.62901
Weibull				
	P-P	0.99727	(0.99724, 0.99730)	$3.08 \cdot 10^{-6}$
	Q-Q	0.70740	(0.70496, 0.70982)	16.6636

Tabella 8.4: Coefficienti di Correlazione e Differenze di Area dei grafici P-P e Q-Q per i Tempi di Interarrivo.

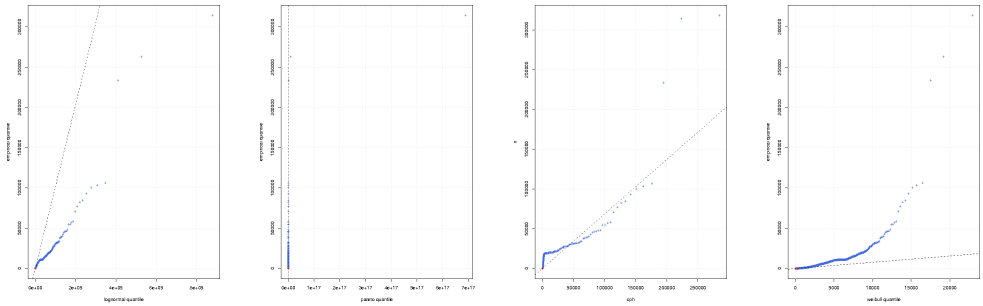


Figura 8.9: Q-Q plot di Log-Normale, Pareto, Phase-Type, Weibull rispetto ai Tempi di Interarrivo.

Metodo	MLE
Log-Mean	3.946299
Log-StdDev	2.154433

- *Pareto*.

Metodo	MLE
Location	0.0001220703
Shape	0.253402

- *Phase-Type continua*.

Metodo	MoM
Initial Vector	$[1, 0]$
Generator	$\begin{bmatrix} -0.00354553 & 1.800 \cdot 10^{-5} \\ 0.000000 & -1.795 \cdot 10^{-5} \end{bmatrix}$

- *Weibull*.

Metodo	MLE
Shape	0.5040407
Scale	148.2247

Riepilogo

I tempi di interarrivo sono caratterizzati da una dispersione molto elevata: il 99.99% delle osservazioni non supera il valore 50000, mentre l'osservazione con il valore massima supera di 6 volte, circa, tale valore. Le osservazioni sono caratterizzate da una modesta autocorrelazione a breve e a lungo termine, e la distribuzione ha una coda destra piuttosta lunga; non si tratta di una coda "heavy" in quanto nessun test grafico ne ha rivelato la presenza.

Per quanto riguarda l'adattamento di distribuzioni teoriche ai dati, fra le distribuzioni considerate, quelle risultate complessivamente migliori nel descrivere l'intera distribuzione sono la Log-Normale, la Phase-Type continua e, in minor misura, la Weibull.

8.2.3 Tempi di Esecuzione

Bonifica dei Dati

Dalla Fig. 8.10 si nota una grande concentrazione di dati nella zona comprendente le osservazioni tra 0 e 20000 (circa il 96% delle osservazioni), una meno intensa concentrazione tra i valori 20000 e 40000 e tra 80000 e 100000 e una presenza di pochi valori estremi oltre il valore 300000. Dato che la presenza di questi valori potrebbe essere un segnale di coda lunga, essi non sono stati esclusi dall'analisi statistica.

Analisi delle Proprietà Statistiche

Il numero di osservazioni presenti nella versione della traccia senza anomalie è pari a 162310, delle quali 97.2% risulta minore o uguale a 50000, mentre il 99.8% è minore o uguale a 200000; la distribuzione mostra una grande dispersione in quanto il valore massimo, cioè 318245, risulta molto maggiore del valore più caratteristico. Ciò può essere osservato dalle statistiche sulla centralità e dispersione riassunte in Tab. 8.5; il centro della distribuzione e il valore medio sono molto distanti, l'intervallo interquartile e la deviazione standard hanno un valore elevato e molto differente. Anche l'asimmetria (destra) è molto gran-

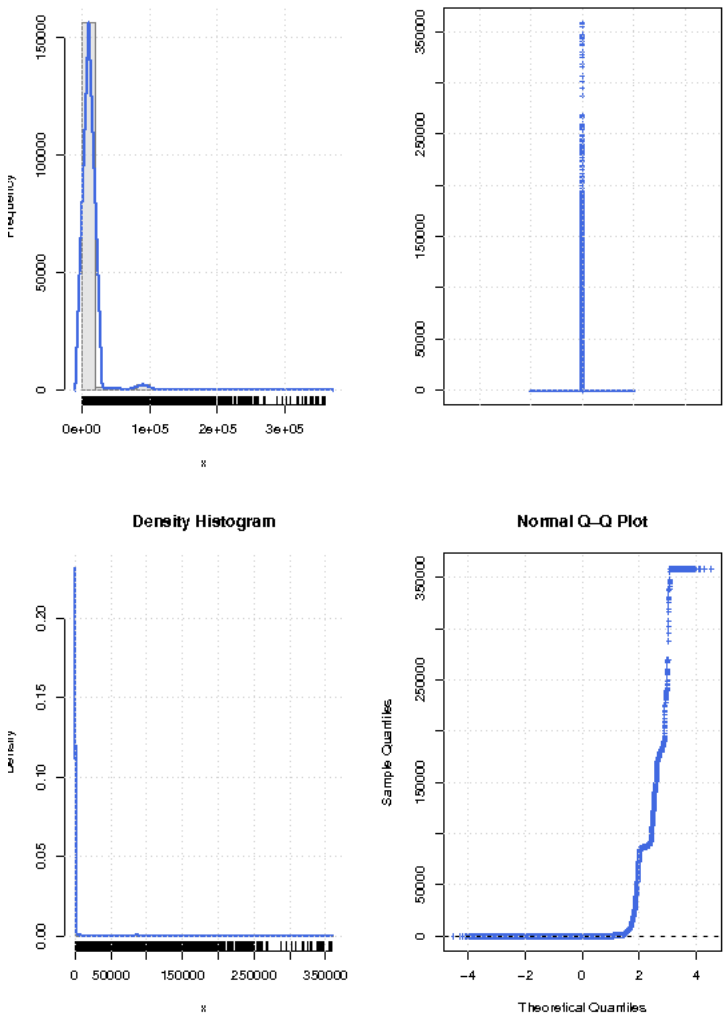


Figura 8.10: Grafico EDA per la forma della distribuzione dei Tempi di Esecuzione.

Statistica	**Valore**	CI_5
Min	1	–
Primo Quartile	3	–
Mediana	3	$(2, 6991)$
Terzo Quartile	6	–
Max	359030	–
IQR	3	–
MAD	1.4826	–
Asimmetria	9.390	–
Quartile-Asimmetria	1.000	–
Curtosi	118.154	–
Curtosi Eccesso	15.154	–
Media	3628.735	$(3523.575, 3733.895)$
Deviazione Standard	21615.96	$(21541.86, 21690.58)$
CV	5.957	–

Tabella 8.5: Riepilogo delle misure di centralità e dispersione dei Tempi di Esecuzione.

Metodo	H
Variance	0.849
R/S	0.736
Periodogram	1.1.379

Tabella 8.6: Esponente di Hurst per i Tempi di Esecuzione.

de (9 volte più accentuata di una Normale) e la curtosi supera di 2 ordini di grandezza quello di una Normale.

Dal grafico dell'autocorrelazione (Fig. 8.11) si nota una discreta presenza di correlazione a breve termine; inoltre, sembra essere presente anche della dipendenza a lungo termine, come mostrato dalla stima dell'esponente di Hurst (Tab. 8.6) e anche, empiricamente, dalla presenza della proprietà di invarianza di scala (Fig. 8.12). Il fatto che il metodo Periodogram fornisca un valore dell'esponente di Hurst superiore a 1 può essere dovuto alla presenza di eventuali "pattern" nei dati [63]. La presenza di autocorrelazione implica l'impossibilità di utilizzare tutti quei metodi statistici che assumono la proprietà di indipendenza nelle osservazioni.

Per quanto riguarda il tipo di coda della distribuzione, dalla Fig. 8.13 e in

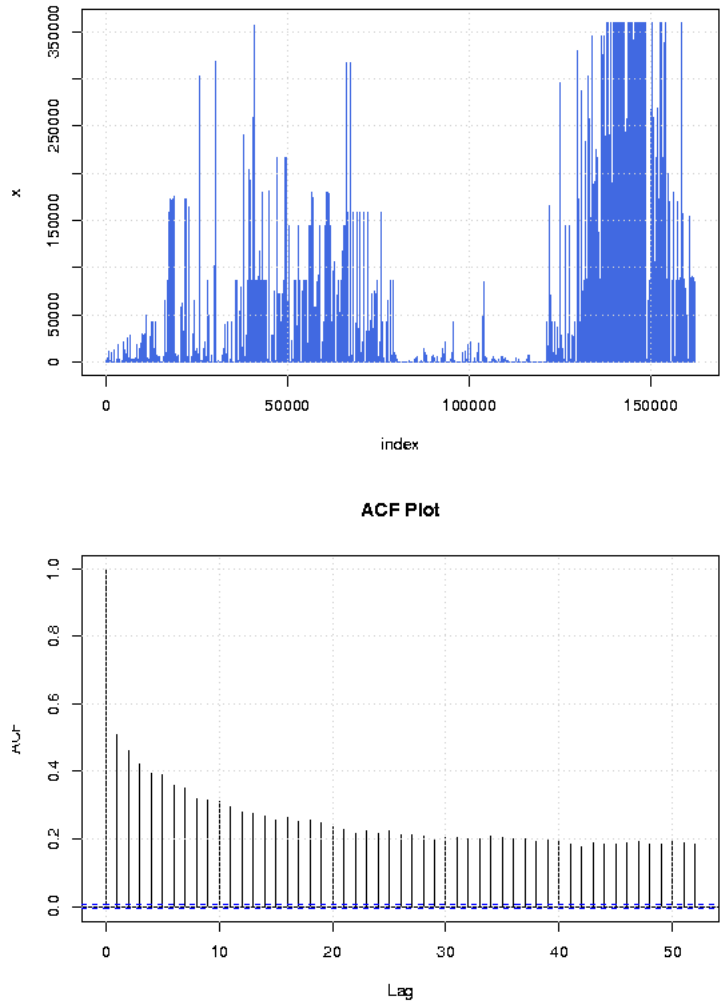


Figura 8.11: Autocorrelazione nella distribuzione dei Tempi di Esecuzione.

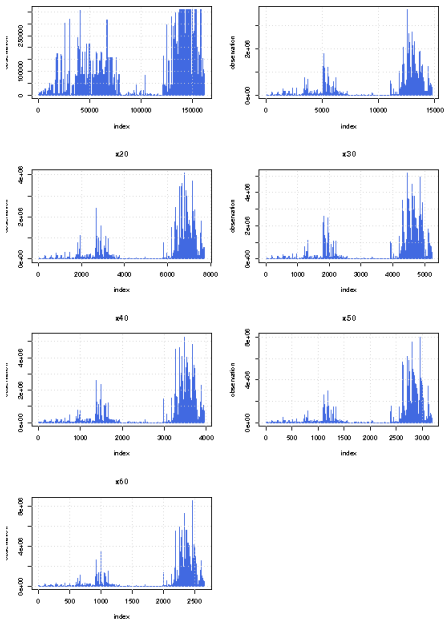


Figura 8.12: Invarianza di Scala nei Tempi di Esecuzione.

particolare dal grafico della Curva di Lorenz e da quello della Mass-Count Disparity, si nota una netta dominanza dei pochi valori estremi rispetto ai valori più frequenti (e più piccoli); dal grafico di Hill non è chiaro se si è in presenza di una coda "heavy", anche se nella parte finale destra sembra esservi una zona di stabilità compresa tra 1 e 2.

Scelta e Verifica del Modello

Dalla Fig. 8.14 si osserva che il corpo della distribuzione sembra ben approssimato dalla Pareto Generalizzata (grafico log-log CDF), mentre nella coda risulta più difficile trovare una distribuzione che ne descriva il comportamento: la Phase-Type continua descrive abbastanza bene la parte finale, mentre per la parte iniziale non è chiaro quale sia la migliore distribuzione.

Per valutare l'adattamento, data l'impossibilità di utilizzo dei test numerici a causa delle presenza di dipendenza fra le osservazioni, si ricorre ai grafici P-P e Q-Q (Fig. 8.15 e Fig. 8.16, rispettivamente), e all'analisi dei relativi coefficienti di correlazione r di Pearson e delle differenze di area relativa ΔA_r (Tab. 8.7):

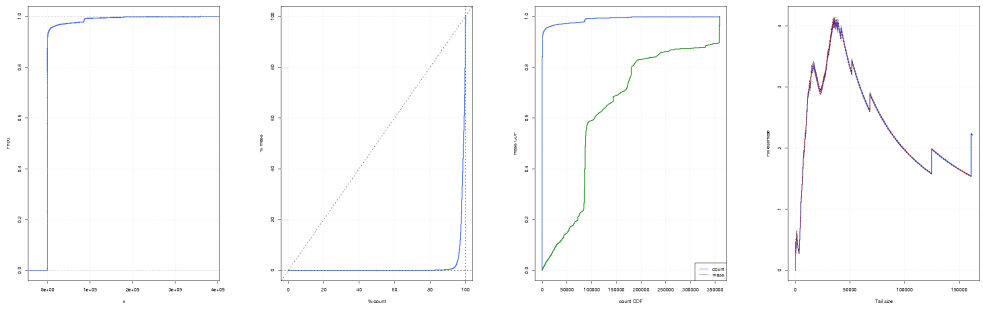


Figura 8.13: ECDF, Curva di Lorenz, Mass-Count Disparity e Hill plot dei Tempi di Esecuzione.

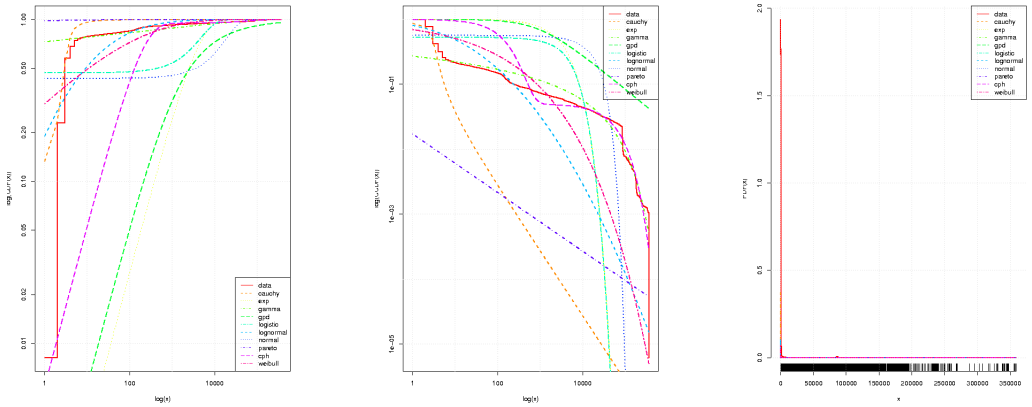


Figura 8.14: Fit per la distribuzione dei Tempi di Esecuzione.

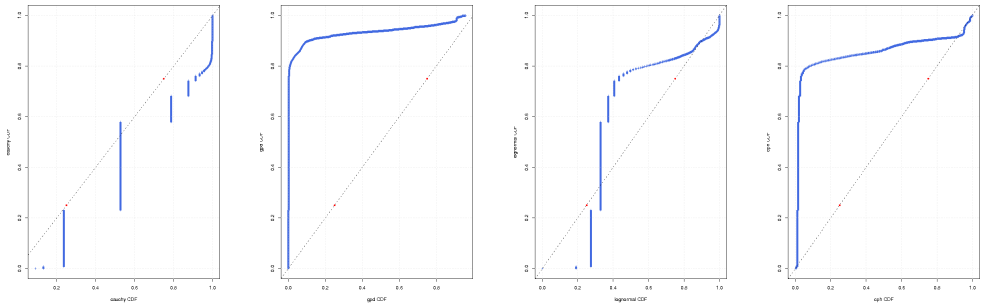


Figura 8.15: P-P plot di Cauchy, GPD, Log-Normale, Phase-Type rispetto ai Tempi di Esecuzione.

- per il corpo della distribuzione, le migliori distribuzioni risultano essere la Pareto Generalizzata, la Cauchy, la Log-Normale e, in minor misura, la Phase-Type continua, la Weibull e la Gamma;
- per la coda della distribuzione, nessuna distribuzione sembra adattarsi all'insieme dei dati; fra queste, quelle che esibiscono un migliore adattamento sono la Log-Normale, la Phase-Type continua, la Weibull e, in minor misura, la Gamma;
- complessivamente, l'intera distribuzione empirica potrebbe essere descritta dalla Log-Normale, dalla Phase-Type continua o dalla Weibull, anche se per tutte e tre le distribuzioni il comportamento nella coda risulta differente da quello della distribuzione dei dati.

Di seguito si riportano i valori dei parametri, stimati dall'insieme delle osservazioni, delle distribuzioni che sono risultate le migliori dal punto di vista dell'adattamento:

- *Cauchy*.

Metodo	MLE
Location	2.923599
Scale	0.8533599

- *Gamma*.

Distribuzione	**Grafico**	**Pearson** r	**Pearson** r CI_5	ΔA_r
Cauchy				
	P-P	0.96566	(0.96533, 0.96599)	0.01798
	Q-Q	0.11185	(0.10705, 0.11665)	18623.5
Esponenziale				
	P-P	0.47584	(0.47207, 0.47960)	$3.01 \cdot 10^{-6}$
	Q-Q	0.64873	(0.64590, 0.65154)	9737.10
Gamma				
	P-P	0.68246	(0.67986, 0.68505)	0.08500
	Q-Q	0.89430	(0.89332, 0.89527)	990.021
GPD				
	P-P	0.89877	(0.89783, 0.89970)	0.00208
	Q-Q	0.36042	(0.35618, 0.36465)	41436.0
Logistica				
	P-P	0.44565	(0.44174, 0.44954)	0.28243
	Q-Q	0.44876	(0.44486, 0.45263)	38826.7
Log-Normale				
	P-P	0.81923	(0.81762, 0.82082)	0.03789
	Q-Q	0.49154	(0.48784, 0.49522)	5.72228
Normale				
	P-P	0.38233	(0.37817, 0.38647)	0.23122
	Q-Q	0.39970	(0.39561, 0.40378)	38203.1
Pareto				
	P-P	0.42760	(0.42361, 0.43156)	0.99885
	Q-Q	0.04720	(0.04234, 0.05205)	1.00000
Phase-Type continua				
	P-P	0.66541	(0.66269, 0.66811)	0.00031
	Q-Q	0.98279	(0.98262, 0.98295)	38.0883
Weibull				
	P-P	0.77417	(0.77221, 0.77611)	0.10037
	Q-Q	0.84683	(0.84545, 0.84820)	37.5680

Tabella 8.7: Coefficienti di Correlazione e Differenze di Area dei grafici P-P e Q-Q per i Tempi di Esecuzione.

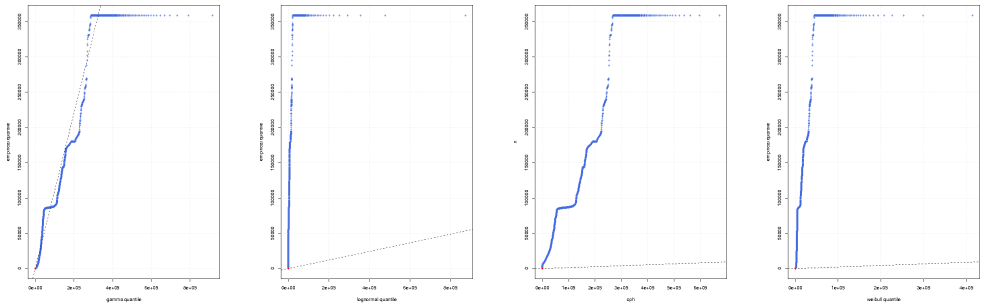


Figura 8.16: Q-Q plot di Gamma, Log-Normale, Phase-Type, Weibull rispetto ai Tempi di Esecuzione.

Metodo	MLE
Shape	0.1305049
Scale	27805.35

- *Log-Normale*.

Metodo	MLE
Log-Mean	2.216622
Log-StdDev	2.535900

- *Pareto Generalizzata*.

Metodo	MLE
Location	0.9987656
Scale	2.664534
Shape	2.026969

- *Phase-Type continua*.

Metodo	MoM
Initial Vector	$[1, 0]$
Generator	$\begin{bmatrix} -0.00569733 & 2.836 \cdot 10^{-4} \\ 0.000000 & -1.441 \cdot 10^{-5} \end{bmatrix}$

- *Weibull*.

Metodo	MLE
Shape	0.27552
Scale	40.96114

Riepilogo

I tempi di esecuzione dei job sono caratterizzati da una dispersione molto elevata, da una discreta autocorrelazione a breve e a lungo termine, e da una coda destra lunga; il tipo di coda potrebbe essere, in base ai grafici della Curva di Lorenz e della Mass-Count Disparity, di tipo "heavy"; tuttavia né il grafico dello stimatore di Hill né il grafico log-log CCDF ne mostra una presenza.

Per quanto riguarda l'adattamento ai dati, la coda della distribuzione si è rivelata molto più difficile da descrivere rispetto al corpo, tanto che nessuna, fra le distribuzioni prese in considerazione, si può ritenere adatta a descriverne il comportamento; volendo scegliere le distribuzioni che, fra quelle considerate, hanno mostrato un comportamento più simile a quello delll'intera distribuzione empirica, si sceglierebbe la Log-Normale, la Phase-Type continua e la Weibull.

8.3 Riepilogo e Considerazioni Finali

L'analisi della traccia TeraGrid ha messo in luce una forte dispersione sia nei i tempi di interarrivo sia in quelli di esecuzione; per ognuna delle due caratteristiche considerate, la presenza di valori molto elevati (e poco frequenti) ha influito sulle proprietà statistiche della distribuzione sottostante, causando valori di asimmetria e di curtosi molto elevati e un allontanamento del valor medio dal centro della distribuzione (cioè dalla mediana), ciò malgrado che la maggior parte delle osservazioni fosse concentrata in una zona relativamente ristretta. Questo è sicuramente un segnale di presenza di una coda lunga.

Mentre si è certi che il tipo di coda (destra) della distribuzione dei dati sia una coda lunga, si hanno dei dubbi sul fatto che sia una coda "heavy"; tali dubbi derivano dal fatto che dal grafico dello stimatore di Hill non è stata

osservata nessuna zona di stabilità; occorre comunque ricordare che vi sono forti ragioni nel credere che l'effettiva utilità dello stimatore di Hill si abbia quando la distribuzione sottostante sia una Pareto [34]; per esempio, nei tempi di esecuzione, sia il grafico della Curva di Lorenz, sia quello della Mass-Count Disparity hanno mostrato una probabile presenza di code "heavy".

A causa della presenza di dipendenza a lungo termine, potrebbe essere interessante valutare la costruzione di un modello basato su processi stocastici che siano in grado di modellare, almeno in parte, le dipendenza fra i dati, come, ad esempio, i processi *Markov Modulated Poisson Process* (*MMPP*) o i più generici *Markov Arrival Process* (*MAP*). Il vantaggio di questi processi è la loro flessibilità nel descrivere le dipendenze fra i dati; tuttavia hanno lo svantaggio di essere onerosi dal punto di vista computazionale. Per quest'ultimo motivo sarebbe preferibile utilizzare una distribuzione computazionalmente più semplice, come una Weibull o una Log-Normale, le quali si sono rivelate abbastanza adatte per descrivere l'intera distribuzione dei dati o almeno una parte. Anche con la Phase-Type continua si sono ottenuti dei buoni risultati, tuttavia si è constatato sperimentalmente che dal punto di vista della complessità computazionale risulta molto più onerosa delle altre distribuzioni.

Capitolo 9

Considerazioni Finali

9.1 Risultati

Dall'analisi statistica delle tracce è emerso che la distribuzione dei tempi di interarrivo sembra essere descritta in maniera soddisfacente dalla distribuzione Phase-Type continua e dalla Log-Normale; in alcuni casi è anche possibile utilizzare una Weibull (ad es., in TeraGrid). Invece, per quanto riguarda i tempi di esecuzione, pare che la distribuzione Weibull sia quella che si adatti meglio ai dati osservati, insieme alla distribuzione Phase-Type continua; in alcuni casi è possibile utilizzare anche la Gamma (ad es., in LCG) e la Log-Normale (ad es., in TeraGrid). Occorre, tuttavia, far notare che la stima dei parametri di una distribuzionie e quindi la relativa valutazione della bontà dell'adattamento ai dati, non è indipendente dal metodo utilizzato per l'adattamento; la bontà dell'adattamento di una particolare distribuzione a un insieme di dati può variare a seconda della tecnica di adattamento utilizzata. Un esempio che mostra questo tipo di dipendenza è quello relativo all'adattamento dei tempi di esecuzione per la VO *alice* nella traccia LCG (§7.2.5); in quel caso, la distribuzione Esponenziale esibisce un miglior adattamento rispetto alla Phase-Type continua (che può essere considerata una sua generalizzazione) grazie al fatto di utilizzare un metodo di "fitting" (tecnica MLE) che fornisce, in genere, stime più precise rispetto a quello utilizzato per la distribuzione Phase-Type continua (metodo dei momenti).

Come ci si aspettava, si è notata una maggiore difficoltà nell'adattamento della coda della distribuzione empirica, cioè di quella zona in cui si trovano dei valori estremi aventi una probabilità di occorrenza molto bassa. Il problema principale nell'adattamento della coda di una distribuzione è la presenza di poche informazioni. Anche l'utilizzo di distribuzioni di probabilità più flessibili, come la Phase-Type, non semplifica l'operazione di adattamento della coda: il comportamento della coda, oltre l'ultima osservazione ricavata dai dati, non è, in genere, conosciuto; se, da un lato, questo potrebbe essere visto come un problema secondario per il fatto che le osservazioni in questa zona sono rare, resta il problema che l'occorrenza di un'osservazione rara, ed estrema, potrebbe avere un effetto molto pesante sulle prestazioni del sistema (ad es., l'arrivo nel sistema di job caratterizzati da un tempo di esecuzione estremamente elevato potrebbe causare, se non adeguatamente modellato, la congestione del sistema stesso). Mentre questo è un problema difficilmente risolvibile, se non attraverso un'evidenza sperimentale del comportamento di un certo tipo di caratteristica del carico, occorre far notare che l'utilizzo di una distribuzione Phase-Type porta, in ogni caso, a un decadimento esponenziale della coda oltre l'ultima osservazione che si ha a disposizione.

L'analisi statistica ha rivelato, con una certa sorpresa, la presenza di autocorrelazione nella maggior parte dei casi analizzati; si tratta di autocorrelazione sia a breve sia a lungo termine. La presenza di autocorrelazione, specialmente a breve termine, può essere dovuta alla eventuale presenza di Bag-of-Task; in tal caso, infatti, l'arrivo di una Bag-of-Task corrisponde a un arrivo *batch* di task che compongono la Bag-of-Task. Anche i tempi di esecuzione ne possono essere influenzati in quanto, di solito, una Bag-of-Task è costituita da applicazioni identiche che possono essere eseguite in modo indipendente le une dalle altre. La presenza di correlazione provoca qualche problema nella conduzione di un'analisi statistica; infatti l'utilizzo dei classici test statistici che assumono l'indipendenza o la normalità fra i campioni (ad es., i test numerici sulla bontà di adattamento) è praticamente inutile. La presenza di autocorrelazione ha un altra importante conseguenza: la classica assunzione di utilizzo di una distribuzione Esponenziale per modellare la distribuzione

dei tempi di interarrivo sembra non essere applicabile nel contesto dei sistemi Grid. Ciò significa che le simulazioni di sistemi Grid, utilizzate, ad esempio, per la verifica delle prestazioni di euristiche di scheduling, come in [71], non dovrebbero utilizzare una distribuzione di Poisson per modellare gli arrivi di job nel sistema. Malgrado la presenza di dipendenza fra le osservazioni, una distribuzione Esponenziale potrebbe comunque risultare una buona approssimazione dei dati (come di fatti succede in §7.2.4 o in §8.2); tuttavia, la distribuzione Esponenziale non riuscirebbe a catturare l'autocorrelazione, a causa della proprietà di assenza di memoria, e a modellare code sub-esponenziali, a causa del decadimento esponenziale della propria coda.

Per quanto riguarda la tipologia della coda delle distribuzioni relative alle caratteristiche del carico analizzate, secondo i risultati ottenuti si può concludere che non vi è particolare evidenza di presenza di code "heavy"; tuttavia in alcuni casi (ad es., in §8.2.3) la dominanza dei valori estremi su quelli più probabili risulta piuttosto palese dai grafici della Curva di Lorenz e della Mass-Count Disparity; purtroppo il grafico dello stimatore di Hill non ha mai fornito una valida regione di stabilità, indice di presenza di code "heavy".

9.2 Lavori Correlati

L'analisi di tracce per la caratterizzazione del carico è uno studio piuttosto frequente in quanto rappresenta uno dei modi più diretti per la costruzione di modelli realistici di un particolare sistema. Nel contesto dei sistemi paralleli, come i cluster, la caratterizzazione del carico è un campo di ricerca ormai esplorato da parecchi anni; invece, per i sistemi Grid, l'interesse è emerso solo recentemente. In generale, le proprietà e le caratteristiche del carico di sistemi paralleli sono diverse da quelle relative al carico di sistemi Grid, a causa della natura differente dei due sistemi (ad es., nei sistemi paralleli le macchine sono generalmente omogenee, mentre nei sistemi Grid le macchine sono eterogenee).

In [67] viene effettuata l'analisi della medesima traccia analizzata nel Cap. 7 del presente documento, ossia la traccia LCG; nell'articolo, si presenta l'analisi

per la costruzione di un modello generativo per la distribuzione degli arrivi dei job; il modello viene costruito effettuando un adattamento sia sui dati sia sulla funzione di autocorrelazione (ACF). La scelta del tipo di processo, per la generazione degli arrivi dei job, viene effettuata tra il classico processo di *Poisson*, il processo *Interrupted Poisson Process* (*IPP*), e tre differenti processi *Markov Modulated Poisson Process* (*MMPP*) con 2, 3 e 4 stati, rispettivamente. Il processo di Poisson è caratterizzato da tempi di interarrivo Esponenziali e indipendenti. Un processo IPP è un processo stocastico *ON–OFF* caratterizzato dall'alternarsi di periodi *ON*, in cui gli arrivi sono distribuiti secondo un processo di Poisson con tasso costante, e di periodi *OFF*, in cui non avviene nessun arrivo nel sistema. Un process MMPP è un processo stocastico in cui ogni stato rappresenta un processo di Poisson, il cui tasso dipende dallo stato stesso; le transizioni tra gli stati, governate da specifici tassi di transizione, permettono di modellare alcune delle possibili forme di autocorrelazione tra gli arrivi. I risultati ottenuti mostrano che i processi MMPP sembrano offrire migliori risultati, dal punto di vista dell'adattamento, rispetto agli altri processi considerati; tuttavia il processo IPP sembra migliore nel modellare situazioni di autocorrelazione a breve termine. Inoltre viene fatto notare come l'utilizzo di processi MMPP con un numero di stati superiore a 2 comporti un aggravio dal punto di vista della complessità computazionale e della convergenza del metodo utilizzato per effettuarne la stima dei parametri; sulla base di queste considerazioni, gli autori dell'articolo sostengono che un processo MMPP a due stati dovrebbe rappresentare un buon compromesso per modellare in modo soddisfacente i tempi di interarrivo, comprese le eventuali forme di autocorrelazione.

Oltre all'articolo citato non ne sono stati trovati altri inerenti alla caratterizzazione del carico per sistemi Grid. Un'ottima fonte di informazioni è il libro sulla caratterizzazione del carico scritto da D. Feitelson [41], scaricabile gratuitamente dal sito del *Parallel Workloads Archive* [39]; tale libro, anche se rivolto in modo particolare alla caratterizzazione del carico per sistemi paralleli, si è rivelato un ottimo punto di inizio per la comprensione delle problematiche inerenti la caratterizzazione del carico e un riferimento costante per lo

svolgimento del presente studio.

9.3 Sviluppi Futuri

I risultati ottenuti da questa analisi aprono la strada a vari lavori futuri:

- *Analisi multivariata delle caratteristiche del carico*. Nel presente lavoro sono state analizzate, in modo individuale, alcune caratteristiche del carico; ci si aspetta, tuttavia, di trovare una presenza di correlazione (*cross-correlation*) tra due o più caratteristiche; per esempio, i tempi di interarrivo e quelli di esecuzione potrebbero mostrare una forma di dipendenza, in particolare se i job rappresentano dei task di una Bag-of-Task.

- *Verifica sperimentale dei risultati ottenuti*. I risultati ottenuti dalla presente analisi, hanno permesso di individuare delle possibili distribuzioni utilizzabili per descrivere i tempi di interarrivo e di esecuzione dei job. Tuttavia non è stato possibile effettuare nessun tipo di generalizzazione a causa del numero limitato di tracce analizzate e della totale assenza di informazioni sul grado di rappresentatività dei dati in esse contenuti; per la traccia TerGrid, si può supporre che le informazioni in essa contenute siano una buona rappresentazione del carico, grazie alla dimensione piuttosto grande dell'arco temporale coperto; invece, per la traccia LCG, il periodo di tempo considerato copre solamente dieci giorni, un arco temporale troppo piccolo per ritenere rappresentative le informazioni sul carico in essa contenute. È quindi assolutamente necessario recuperare altre tracce, in modo da poter estendere, confermare o confutare i risultati ottenuti nella presente analisi.

- *Valutazione di nuove Distribuzioni o Processi*. Una famiglia di distribuzioni che si intende prendere in considerazione in successive analisi è quella delle distribuzioni α-Stable, una classe di distribuzioni in grado di descrivere molte delle distribuzioni "heavy-tailed" (e non) e per cui vale la proprietà di invarianza di scala; per il momento, si è deciso di escluderle

dall'analisi a causa di alcuni problemi di instabilità numerica. Vi sarebbe, inoltre, un certo interesse nel valutare i risultati e le prestazioni che si otterrebbero utilizzando dei processi stocastici come gli MMPP, in modo simile a quanto effettato in [67], al fine di effettuare dei confronti con i risultati ottenuti in quest'analisi; si vorrebbero valutare anche le prestazioni di altri tipi di processi stocastici come i MAP o i BMAP (Batch MAP), ...Uno dei motivi per cui interessa esplorare questi tipi di processi, è legato alla valutazione della complessità computazionale; infatti, sebbene questi proccessi siano molto flessibili e permettano di modellare alcune forme di correlazione, hanno lo svantaggio di essere molto esigenti dal punto computazionale e di soffrire, a volte, di problemi di instabilità numerica. Per esempio, si è tentato di utilizzare i processi MMPP per modellare la distribuzione degli arrivi; si è cercato di effettuare la stima dei parametri del processo MMPP attraverso l'uso di uno degli algoritmi più diffusi (algoritmo di Ryden), basato sulla tecnica statistica EM; tuttavia, a causa di problemi di instabilità numerica (come matrici singolari, tolleranza non raggiunta,...) non si è riusciti a ottenere alcun risultato.

- *Applicabilità pratica e Ottimizzazione*. Nel punto precedente è stata esposta la difficoltà di utilizzo di processi stocastici, come i MAP o gli MMPP, causata da problemi di instabilità numerica. Un altro problema associato a tali processi è la possibile elevata complessità computazionale. Anche per la distribuzione Phase-Type continua esiste un problema simile; in particolare, si è notato sperimentalmente che il tempo necessario per la generazione di quantili distribuiti secondo una Phase-Type continua passa da un valore dell'ordine di ore a uno dell'ordine di giorni, aumentando, semplicemente, di una unità la dimensione del generatore della distribuzione (da una matrice di ordine 2 a una di ordine 3). Il metodo attuale di generazione dei quantili consiste nella simulazione della Catena di Markov sottostante; occorre capire se esistono metodi più efficienti di questo. Inoltre, l'utilizzo di distribuzioni o processi caratterizzati da un alto numero di parametri rende meno intuitiva l'utilizzabilitá dal punto di vista sperimentale; per esempio, nel caso si scoprisse, che

la distribuzione Phase-Type continua o un processo MMPP riesca a rappresentare in modo soddisfancente la distribuzione empirica, vi sarebbe una certa difficoltà nello scegliere il numero e il valore dei parametri da utilizzare in tali modelli; decisione sicuramente più semplice nel caso si utilizzasse una distribuzione Esponenziale o Log-Normale; oltre a questo aspetto, occorre anche tenere in considerazione che più complesso è il modello e più è facile che variando di poco il numero o i valori dei parametri del modello si ottengano dei risultati molto differenti (problema del *overfitting*).

- *Stima separata del Corpo e della Coda*. Nella maggior parte delle analisi statistiche descritte nei precedenti capitoli, si è osservato che l'adattamento di una distribuzione empirica sarebbe stato più semplice e avrebbe fornito risultati migliori nel caso in cui l'insieme dei dati fosse stato diviso in due parti: *corpo* e *coda* della distribuzione; infatti, è noto che queste due parti tendono a comportarsi in maniera differente. Il problema nell'applicare questo approccio è la stima del cosidetto valore di *threshold*, cioè di quel valore a partire dal quale finisce il corpo della distribuzione e inizia la coda. Non sembra esistere un metodo diretto per la stima di tale valore; una prima indagine su tale argomento ha fatto emergere l'esistenza di alcuni approcci empirici, tra cui quello di far terminare il corpo della distribuzione in corrispondenza del k-esimo percentile (ad es., 95-esimo percentile), oppure di ipotizzare un particolare valore massimo, dipendente dalla caratteristica del carico studiata, a partire dal quale considerare tutte le osservazioni più grandi come delle osservazioni estreme della coda della distribuzione (ad es., supporre che il tempo di esecuzione massimo per un job sia di un mese). Anche se non vi è stato modo per effettuare una verifica sperimentale, si ritiene che questi metodi siano utilizzabili solo quando si ha una conoscenza specifica dello scenario che si sta modellando; negli altri casi, l'utilizzo di questi metodi potrebbe introdurre un bias difficilmente trascurabile; ad es., l'utilizzo di un valore massimo errato potrebbe far considerare come estreme delle osservazioni che in realtà non lo sono.

Parte III

Appendici

Appendice A

Architettura del Codice Sorgente

In questo capitolo viene illustrata la struttura del codice sorgente utilizzato per effettuare gli esperimenti sulla caratterizzazione del carico.

La maggior parte del codice sorgente è stato implementato nel linguaggio *R* [45], un linguaggio di programmazione ad alto livello rivolto allo sviluppo di applicazioni statistiche [1]. Alcune funzionalità critiche (come l'esponenziale di una matrice) sono state sviluppate nel linguaggio *C* e integrate in *R* tramite l'utilizzo di librerie dinamiche. Sono stati, infine, utilizzati alcuni script scritti in linguaggio *Matlab* [77] (ad es., il fitting di distribuzioni Phase-Type) e in linguaggio *Perl* [33] (per la ricerca di anomalie nelle tracce).

L'intero codice utilizzato per la conduzione degli esperimenti descritti nel presente progetto è disponibile nel CD allegato al documento.

Nelle successive sezioni viene presentata la struttura dei file contenenti le funzioni sviluppate nel linguaggio *R*.

A.1 Funzioni di Libreria

mg_consts.R. Alcune costanti utilizzate dalle funzioni di libreria e dalle applicazioni, come l'identificare di distribuzioni di probabilità, dei test di adattamento, del tipo di grafico, ...

[1] *R* è la versione *open-source* del linguaggio statistico *Splus*, distribuito da *Insightful*.

mg_debug.R. Funzioni di utilità per effettuare il *debug* del codice.

mg_dists_cph.R. Funzioni per ottenere i valori della densità, della funzione di distribuzione, dei quantili e per la generazione di numeri casuali relativi alla distribuzione Phase-Type continua.

mg_dists_frechet.R. Funzioni per ottenere i valori della densità, della funzione di distribuzione, dei quantili e per la generazione di numeri casuali relativi alla distribuzione Frechét.

mg_dists_gev.R. Funzioni per ottenere i valori della densità, della funzione di distribuzione, dei quantili e per la generazione di numeri casuali relativi alla distribuzione Valori Estremi Generalizzata.

mg_dists_gpd.R. Funzioni per ottenere i valori della densità, della funzione di distribuzione, dei quantili e per la generazione di numeri casuali relativi alla distribuzione Pareto Generalizzata.

mg_dists_gumbel.R. Funzioni per ottenere i valori della densità, della funzione di distribuzione, dei quantili e per la generazione di numeri casuali relativi alla distribuzione Gumbel.

mg_dists_pareto.R. Funzioni per ottenere i valori della densità, della funzione di distribuzione, dei quantili e per la generazione di numeri casuali relativi alla distribuzione Pareto.

mg_dists_stable.R. Funzioni per ottenere i valori della densità, della funzione di distribuzione, dei quantili e per la generazione di numeri casuali relativi alla distribuzione Stable.

mg_dists_weibull3.R. Funzioni per ottenere i valori della densità, della funzione di distribuzione, dei quantili e per la generazione di numeri casuali relativi alla distribuzione Weibull a 3 parametri.

mg_eda.R. Funzioni di utilità per effettuare l'analisi dei dati secondo l'approccio EDA.

mg_fit.R. Funzioni per l'utilizzo di vari stimatori di parametri nel contesto dell'adattamento di una distribuzione teorica a un insieme di dati (come funzioni per lo stimatore MLE, per il Metodo dei Momenti, per lo stimatore di Hill, ...).

mg_fit_utils.R. Funzioni di utilità per l'adattamento di una distribuzione teorica a un insieme di dati.

mg_gof_ad.R. Test di adattamento di Anderson-Darling a uno e a due campioni, e in versione semplice o con bootstrap.

mg_gof_chisq.R. Test di adattamento di Pearson χ_2 per distribuzioni di probabilità discrete e continue, e in versione semplice o parametrica.

mg_gof_ks.R. Test di adattamento di Kolmogorov-Smirnov a uno e a due campioni, e in versione semplice o con bootstrap.

mg_gof_pp.R. Test di adattamento basato sul grafico P-P a uno e a due campioni; calcola anche il coefficiente di correlazione lineare di Pearson e la differenza di area assoluta e relativa.

mg_gof_qq.R. Test di adattamento basato sul grafico Q-Q a uno e a due campioni; calcola anche il coefficiente di correlazione lineare di Pearson e la differ enza di area assoluta e relativa.

mg_gof.R. Funzioni di utilità per la verifica dell'adattamento di una distribuzione teorica ai dati o per la verifica di uguale distribuzione fra due insiemi di dati.

mg_lcg.R. Classe per la gestione della traccia LCG.

mg_lrd.R. Funzioni per la verifica di dipendenza a lungo termine (ad es., metodo della Verianza, del R/S, Periodogram, Cumulative Periodogram, aggregazione, ...).

mg_math_funcs.R. Alcune funzioni matematiche.

mg_math_integral.R. Alcune funzioni per l'integrazione numerica.

mg_math_poly.R. Alcune funzioni per la valutazione numerica di polinomi.

mg_matrix.R. Alcune funzioni per la manipolazione di matrici.

mg_moments.R. Funzioni per il calcolo di momenti campionari (centrati o non) di qualsiasi ordine, della asimmetria e della curtosi.

mg_plot_lrd.R. Funzioni per la visualizzazione grafica dei metodi di verifica di dipendenza a lungo termine.

mg_plot.R. Funzioni per la visualizzazione di vari grafici (grafico P-P, grafico Q-Q, istogrammi, run sequence plot, box-plot, autocorrelazione, ...).

mg_rvg_alias.R. Implementazione del metodo del *alias* per la generazione di numeri casuali.

mg_statsutils.R. Funzione per il calcolo di generiche statistiche (intervalli di confidenza, coefficiente di variazione, ...).

mg_teragrid.R. Classe per la gestione della traccia TeraGrid.

mg_utils.R. Funzioni generiche di utilitá (look-up di tabelle, ordinamento di *data.frame*, ricerca di file, ...).

mg_vector.R. Alcune funzioni per la manipolazione di vettori.

A.2 Applicazioni

- **mg_app_conf.R**. Parametri di Configurazione comuni.
- **mg_app_lcg-overall-iat-all.R**. Analisi dei Tempi di Interarrivo dei job, a livello Grid, per la traccia LCG.
- **mg_app_lcg-overall-rt-all.R**. Analisi dei Tempi di Esecuzione dei job, a livello Grid, per la traccia LCG.
- **mg_app_lcg-vo-iat-all.R**. Analisi dei Tempi di Interarrivo dei job, a livello Organizzazione Virtuale, per la traccia LCG.

- **mg_app_lcg-vo-rt-all.R**. Analisi dei Tempi di Esecuzione dei job, a livello Organizzazione Virtuale, per la traccia LCG.
- **mg_app_teragrid-anomalies.R**. Analisi delle anomalie presenti nella traccia TeraGrid.
- **mg_app_teragrid-overall-iat-all.R**. Analisi dei Tempi di Interarrivo dei job, a livello Grid, della traccia TeraGrid.
- **mg_app_teragrid-overall-rt-all.R**. Analisi dei Tempi di Esecuzione dei job per la traccia TeraGrid.

Appendice B

Elementi di Probabilità e Statistica

In questo capitolo vengono presentati, in maniera schematica, alcuni concetti di base della teoria delle Probabilità e della Statistica; per maggiori chiarimenti e approfondimenti, si consulti un qualsiasi libro di testo specifico per questo argomento come [80, 101].

B.1 Elementi di Probabilità

Definizione B.1.1 (Spazio di Probabilità)**.** Uno *Spazio di Probabilità* è una tripla $(\Omega, \mathcal{A}, \Pr)$ dove Ω è un insieme non vuoto, $\mathcal{A}$ è una σ-algebra, e $\Pr$ è una misura positiva che soddisfa gli *Assiomi di Probabilità* secondo Kolmogorov. Inoltre:

- Ω è detto *Spazio dei Campioni* e i suoi elementi sono chiamati *risultati* o *stati di natura*;
- $\mathcal{A}$ è detto *Spazio degli Eventi* e i suoi elementi sono chiamati *eventi* (insieme di risultati per i quali si può richiedere il calcolo della probabilità);
- $\Pr$ è una *Misura di Probabilità* e rappresenta una funzione da $\mathcal{A}$ all'insieme $[0, 1]$; il valore ottenuto da questa funzione, rispetto a un certo evento $A \in \mathcal{A}$, è detto *probabilità* di A.

Definizione B.1.2 (Variabile Casuale)**.** Una *Variabile Casuale* (o *Aleatoria*) è una funzione che mette in relazione un insieme di eventi (ad es., i risultati di un

esperimento) con un insieme di valori. Più formalmente, se $(\Omega, \mathcal{A}, \mathcal{P})$ è uno *spazio di probabilità*, una *Variabile Casuale* è una funzione $X : \Omega \to \mathcal{S}$ tale che:

$$\{\omega \in \Omega : X(\omega) \leq s\} \in \mathcal{A}, \quad \forall s \in \mathcal{S}$$

dove $\mathcal{S} \subseteq \mathbb{R}$ è detto *spazio degli stati*. Se il numero di valori che una variabile casuale può assumere è finito, la variabile è detta *discreta*; viceversa, la variabile è detta *continua*.

Definizione B.1.3 (Funzione di Distribuzione Cumulativa (CDF))**.** Data una variabile aleatoria X, si definisce *Funzione di Distribuzione Cumulativa* (*CDF*) la funzione

$$F : \mathbb{R} \to [0, 1] \text{ t.c. } F(x) = \Pr\{X \leq x\} \tag{B.1.1}$$

La CDF è anche chiamata *Funzione di Ripartizione* o, semplicemente, *Funzione di Distribuzione*.

Valgono le seguenti proprietà:

1. Ogni CDF è monotona (non-strettamente) crescente e continua a destra.
2. $\lim_{x \to -\infty} F(x) = 0$.
3. $\lim_{x \to \infty} F(x) = 1$.
4. $\Pr\{a \leq X \leq b\} = F(b) - F(a)$.
5. Se X è una variabile casuale continua e $F(\cdot)$ è derivabile, allora $F(x) = \int_{-\infty}^{x} f(t)\, dt$, dove $f(\cdot)$ prende il nome di *funzione di densità di probabilità* (*PDF*).
6. Una CDF descrive completamente la distribuzione di probabilità associata.

Definizione B.1.4 (Momento k-esimo)**.** Data un variabile aleatoria X con funzione di probabilità $f(\cdot)$, si definisce *Momento k-esimo (centrato intorno all'origi-*

ne) di X la quantità:

$$\mu'_k = \begin{cases} \sum_{x \in X} x^k f(x) & \text{se } X \text{ è discreta} \\ \int_{-\infty}^{\infty} x^k f(x)\, dx & \text{se } X \text{ è continua} \end{cases} \tag{B.1.2}$$

con $k = 1, 2, \ldots$.

Definizione B.1.5 (Momento k-esimo centrale)**.** Data un variabile aleatoria X con funzione di probabilità $f(\cdot)$, si definisce *Momento k-esimo centrale* (o *Momento k-esimo centrato intorno alla media*) di X la quantità:

$$\mu_k = \begin{cases} \sum_{x \in X} (x - \mu'_1)^k f(x) & \text{se } X \text{ è discreta} \\ \int_{-\infty}^{\infty} (x - \mu'_1)^k f(x)\, dx & \text{se } X \text{ è continua} \end{cases} \tag{B.1.3}$$

con $k = 1, 2, \ldots$.

Il momento del primo ordine μ'_1 di una variabile aleatoria X corrisponde al suo valore atteso $\mathrm{E}[X]$ e spesso si denota con μ; il momento del secondo ordine centrale μ_2 corrisponde, invece, alla varianza $\mathrm{Var}(X)$ e spesso si indica con σ^2.

Teorema B.1.1 (Teorema del Limite Centrale (CLT) classico)**.** *Sia $X_1, \ldots, X_n$ una sequenza di variabili aleatorie i.i.d., tutte com media μ e varianza σ^2, entrambe finite. Si può dimostrare che per n tendente all'infinito, la distribuzione della somma:*

$$S_n = X_1 + \cdots + X_n$$

tende a una distribuzione Normale con media $n\mu$ e varianza $n\sigma^2$. Detto in altri termini, sia

$$Z = \frac{S_n - n\mu}{\sigma\sqrt{n}}$$

la variabile standardizzata *di S_n; allora Z tende a una distribuzione Normale con media 0 e varianza 1.*

Definizione B.1.6 (Funzione Quantile)**.** Data una variabile aleatoria X, si defi-

nisce *Funzione Quantile* la funzione [1]

$$Q(y) = \inf_{x \in \mathbb{R}} \{F(x) \geq y\}, \quad y \in (0,1) \tag{B.1.4}$$

In tal caso $Q(y)$ viene detto *quantile di ordine* y (o $(y \times 100)$-esimo percentile). Se $F(\cdot)$ è invertibile (cioè, monotona strettamente crescente e continua), allora

$$Q(y) = F^{-1}(y), \quad y \in [0,1] \tag{B.1.5}$$

dove $F^{-1}(\cdot)$ è la funzione inversa di $F(\cdot)$.

Valgono le seguenti proprietà:

1. I valori $Q(y^-) = \lim_{t \to y^-} Q(t)$ e $Q(y^+) = \lim_{t \to y^+} Q(t)$ sono determinati in modo univoco, mentre la quantità $Q(y)$ in generale non è unica, cioè $Q(y) \in \{Q(y^-), Q(y^+)\}$; $Q(y)$ è unica quando $Q(y) = F^{-1}(y)$.
2. $Q(\cdot)$ è una funzione monotona crescente in $(0,1)$.
3. $Q(\cdot)$ è continua $\Leftrightarrow$ $F(\cdot)$ è strettamente crescente.
4. $Q(\cdot)$ è strettamente crescente $\Leftrightarrow$ $F(\cdot)$ è continua.
5. $F(x^-) \leq y \Leftrightarrow Q(y^+) \geq x$.
6. $F(x^+) \geq y \Leftrightarrow Q(y^-) \leq x$.
7. $F(x^-) \leq y \leq F(x^+) \Leftrightarrow Q(y^-) \leq x \leq Q(y^+)$.

Definizione B.1.7 (Mediana)**.** La *mediana* di una distribuzione è quel valore che divide a metà la distribuzione di probabilità; detto in altri termini, rappresenta il 50-esimo quantile. Nel caso di distribuzioni di probabilità con funzione di distribuzione $F(\cdot)$ invertibile, si può affermare che la mediana m è il valore assunto dalla funzione di distribuzione cumulativa inversa $F^{-1}(\cdot)$ nel punto 0.5:

$$m = F^{-1}(0.5)$$

[1] Si noti che di solito gli estremi $y = 0$ e $y = 1$ sono esclusi dalla definizione della funzione quantile; per includerli, si dovrebbe specificare che $F^{-1}(0) = -\infty$, in quanto $-\infty$ è il minimo valore per cui la funzione quantile vale 0.

B.2 Elementi di Statistica

B.2.1 Teoria dei Campioni

Definizione B.2.1 (Momento campionario k-esimo)**.** Dato un campione $X_1, \ldots, X_n$, si definisce *Momento campionario k-esimo (centrato intorno all'origine)* la variabile aleatoria:

$$M'_k = \frac{1}{n} \sum_{i=1}^{n} X_i^k \tag{B.2.1}$$

con $k = 1, 2, \ldots$. Una volta selezionato il campione $X_1 = x_1, \ldots, X_n = x_n$, il relativo valore del Momento campionario k-esimo è dato da:

$$m'_k = \frac{1}{n} \sum_{i=1}^{n} x_i^k \tag{B.2.2}$$

Definizione B.2.2 (Momento campionario k-esimo centrale)**.** Dato un campione di osservazioni $X_1 = x_1, \ldots, X_n = x_n$, si definisce *Momento campionario k-esimo centrale* (o *Momento campionario k-esimo centrato intorno alla media campionaria*) la variabile aleatoria:

$$M_k = \frac{1}{n} \sum_{i=1}^{n} (X_i - M_1)^k \tag{B.2.3}$$

con $k = 1, 2, \ldots$. Una volta selezionato il campione $X_1 = x_1, \ldots, X_n = x_n$, il relativo valore del Momento campionario k-esimo centrale è dato da:

$$m_k = \frac{1}{n} \sum_{i=1}^{n} (x_i - m'_1)^k \tag{B.2.4}$$

Talvolta il momento campionario del primo ordine M'_1 (m'_1) è indicato con il simbolo $\bar{X}$ ($\bar{x}$) e viene chiamato *media campionaria*, mentre per il momento campionario del secondo ordine centrale M_2 (m_2) si utilizza il simbolo $\hat{\Sigma}^2$ ($\hat{\sigma}^2$) e viene detto *varianza campionaria*.

Definizione B.2.3 (Statistica)**.** Dato un campione di osservazioni $X_1, \ldots, X_n$, si

definisce *statistica* $\hat{\Theta} = h(X_1, \ldots, X_n)$ una variabile aleatoria che è una funzione (osservabile) del campione $X_1, \ldots, X_n$.

Di seguito si indicherà con $\hat{\Theta}_\theta = h(X_1, \ldots, X_n)$ una statistica calcolata rispetto al campione $X_1, \ldots, X_n$ e relativa a un parametro θ; inoltre si denoterà con Θ lo spazio dei parametri θ e con $\mathcal{T}$ l'insieme dei valori della statistica $\hat{\Theta}_\theta$. Le proprietà più importanti di una statistica sono:

- *Sufficienza*: una statistica $\hat{\Theta}_\theta$ è *sufficiente* per un parametro $\theta \in \Theta$ se rappresenta in modo sintetico tutta l'informazione relativa a θ contenuta nel campione $X_1, \ldots, X_n$, cioè se il campione $X_1, \ldots, X_n$ è condizionalmente indipendente da θ, dato $\hat{\Theta}_\theta$:

$$\begin{gathered} \Pr\left\{X_1 = x_1, \ldots, X_n = x_n | \hat{\Theta}_\theta = \hat{\theta}, \theta\right\} \\ = \\ \Pr\left\{X_1 = x_1, \ldots, X_n = x_n | \hat{\Theta}_\theta = \hat{\theta}\right\} \end{gathered} \tag{B.2.5}$$

- *Completezza*: una statistica $\hat{\Theta}_\theta$, a valori in $\mathcal{T}$, è *completa* per un parametro θ se per ogni funzione reale $g(\cdot)$ su $\mathcal{T}$:

$$\begin{gathered} \mathrm{E}\left[g\left(\hat{\Theta}_\theta\right)|\theta\right] = 0 \text{ per ogni } \theta \in \Theta \\ \Rightarrow \\ \Pr\left\{g\left(\hat{\Theta}_\theta\right) = 0 | \theta\right\} = 1 \text{ per ogni } \theta \in \Theta \end{gathered} \tag{B.2.6}$$

- *Anciliarità*: una statistica $\hat{\Theta}_\theta$ è *anciliare* se la sua distribuzione di probabilità non dipende da θ.

Si noti che una statistica è una variabile casuale *osservabile*, a differenza di una parametro di una distribuzione, che rappresenta una quantità in generale non osservabile che descrive una proprietà di una popolazione statistica. La distribuzione di probabilità di una statistica è detta *distribuzione campionaria*.

Alcuni esempi di statistiche comprendono: la media campionaria (o media aritmetica), la varianza campionaria, le statistiche d'ordine, le statistiche dei test d'ipotesi, ...

B.2.2 Statistiche d'Ordine

Definizione B.2.4 (Statistiche d'ordine)**.** Dato un campione di osservazioni $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$ lo si riordini in senso crescente in $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ in modo tale che $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$; allora il k-esimo elemento più piccolo $x_{(k)}$, con $1 \leq k \leq n$, è detto *statistica di ordine* k e k è chiamato *rango* della k-esima osservazione ordinata. In particolare:

$$x_{(1)} = \min\{x_1, \ldots, x_n\}, \quad x_{(n)} = \max\{x_1, \ldots, x_n\}$$

si definiscono, rispettivamente, *statistica d'ordine estrema del minimo* e *del massimo* (*minima* e *maxima extreme order statistic*).

Definizione B.2.5 (Coppia Concordante)**.** Una *Coppia Concordante* è una coppia $\langle (x_i, y_i), (x_j, y_j) \rangle$ di un campione bivariato, tale per cui:

$$(x_i < x_j \wedge y_i < y_j) \vee (x_i > x_j \wedge y_i > y_j)$$

cioè se:

$$\operatorname{sign}(x_j - x_i) = \operatorname{sign}(y_j - y_i)$$

In maniera analoga, una *Coppia Discordante* è una coppia $\langle (x_i, y_i), (x_j, y_j) \rangle$ di una campione bivariato, tale che:

$$(x_i < x_j \wedge y_i > y_j) \vee (x_i > x_j \wedge y_i < y_j)$$

cioè se:

$$\operatorname{sign}(x_j - x_i) = -\operatorname{sign}(y_j - y_i)$$

dove sign è la funzione *segno*, definita come:

$$\mathrm{sign}\,(x) = \begin{cases} -1 & : \quad x < 0 \\ 0 & : \quad x = 0 \\ 1 & : \quad x > 0 \end{cases}$$

Definizione B.2.6 (Tie)**.** Dato un insieme di dati $x_1, \ldots, x_n$, si definisce *tie* un valore che nell'insieme appare più di una volta; più formalmente, t è un *tie* per l'insieme $x_1, \ldots, x_n$ se:

$$\exists i, j : i \neq j \wedge x_i = t \wedge x_j = t, \quad 1 \leq i, j \leq n$$

Definizione B.2.7 (Funzione di Distribuzione Empirica (EDF))**.** Si definisce *Funzione di Distribuzione Empirica* (*EDF* o *ECDF*) di un campione casuale di osservazioni $X_1 = x_1, \ldots, X_n = x_n$ la funzione:

$$\begin{aligned} F_n\,(x) &= \frac{\#\,\{x_i | x_i \leq x, i = 1, 2, \ldots, n\}}{n} \\ &= \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\,(\{x_i \leq x\}) \end{aligned} \tag{B.2.7}$$

cioè la frazione di osservazioni minori o uguali a x.

Definizione B.2.8 (Funzione Quantile Campionaria)**.** Dato $X_1, \ldots, X_n$ un campione casuale estratto da una popolazione con distribuzione X, la *Funzione Quantile Campionaria*

$$\hat{Q}_n\,(p) = \min\,\{q | \text{le frazioni } p \text{ delle osservazioni} \leq q\} \tag{B.2.8}$$

rappresenta una stima della distribuzione dei quantili di X, cioè della funzione quantile $Q\,(\cdot)$ (Def. B.1.6) della distribuzione X. Sono state fornite varie espressioni per $\hat{Q}_n\,(\cdot)$; quelle maggiormente utilizzate si basano su una o due statistiche d'ordine del campione di osservazioni $X_1 = x_1, \ldots, X_n = x_n$. In [59] sono presentati 9 tipi di funzione quantile campionaria, tutti basati sulla

seguente definizione:

$$Q_i(p) = (1-\gamma)\, x_{(j)} + \gamma x_{(j+1)} \quad \text{t.c. } \frac{j-m}{n} \leq p < \frac{j-m+1}{n} \tag{B.2.9}$$

dove:

- $1 \leq i \leq 9$ è il tipo di definizione;
- $x_{(j)}$, con $1 \leq j \leq n$ è la j-esima statistica d'ordine;
- n è la numerosità del campione;
- $m \in \mathbb{R}$ è una costante che è funzione del tipo di quantile campionario (continuo o non continuo);
- $0 \leq \gamma \leq 1$ è una quantità che dipende dalla parte frazionaria di $np+m-j$.

Da ciò emerge anche che, in generale, la relazione $x = Q_n(F_n(x))$, non è sempre soddisfatta. Fra queste varianti di funzione quantile campionaria, non sembra essercene una migliore in assoluto; ad esempio, se la densità della popolazione, da cui proviene il campione, fosse continua a tratti, una buona scelta sarebbe quella di utilizzare la variante proposta da Weibull [113]:

$$F_n(x) = \frac{i}{n+1} \tag{B.2.10}$$

dove i è il rango di x. Tale stimatore ha il vantaggio di non escludere eventuali valori più estremi di quelli osservati in un certo campione [2]:

$$\Pr\left\{X < X_{(1)}\right\} = \Pr\left\{X > X_{(n)}\right\} = \frac{1}{n+1}$$

La variante utilizzata nel presente progetto è la definizione numero 7 in [59] ed è quella fornita di "default" dal linguaggio *R*:

$$F_n(x) = \frac{i-1}{n-1} \tag{B.2.11}$$

[2]Questa proprietà risulta particolarmente interessante per distribuzioni caratterizzate da code a legge polinomiale, come le distribuzioni "heavy tailed".

dove i è il rango di x; anche per questa variante, gli eventuali valori più estremi di quelli osservati hanno associata una probabilità non nulla.

Definizione B.2.9 (Grafici e Punti di Probabilità)**.** Un problema strettamente correlato alla stima della funzione quantile è la scelta dei *punti di probabilità* $p(k)$ (*probability points* o *probability plotting positions*) per un grafico di probabilità che coinvolge le statistiche d'ordine $X_{(k)}$. Dato un insieme di osservazioni ordinate $X_{(1)} = x_{(1)}, \ldots, X_{(n)} = x_{(n)}$ di un campione casuale i.i.d. $X_1, \ldots, X_n$, con funzione di distribuzione $F(\cdot)$, e una distribuzione di probabilità teorica con funzione di distribuzione $G(\cdot)$ (o eventualmente un altro campione di statistiche d'ordine), un *grafico di probabilità* è un grafico che utilizza, direttamente o indirettamente, i valori $x_{(i)}$ e $G\left(x_{(i)}\right)$. Per esempio, per disegnare un grafico Q-Q §3.2.1, occorre scegliere i valori di probabilità da utilizzare con la funzione di distribuzione teorica,; in maniera analoga, per tracciare un grafico P-P §3.2.2, è necessario ricavare i valori di probabilità della distribuzione delle statistiche d'ordine campionarie.

Non esiste una soluzione universale a questo problema (ad es., si veda [70, 69, 59]); fra le diverse varianti in circolazione, la forma più utilizzata sembra essere la seguente:

$$p(k) = \frac{k - \alpha}{n + 1 - \alpha - \beta} \tag{B.2.12}$$

dove $1 \leq k \leq n$ è il rango associato alla statistica d'ordine del campione sotto osservazione, mentre α e β sono due numeri reali (di solito, $0 < \alpha, \beta < 1$). La forma più semplice sarebbe:

$$p(k) = \frac{k}{n}$$

ma, in generale, produce una sovrastima, dà luogo a una situazione asimmetrica (in quanto l'ultima osservazione è associata al punto di probabilità 1, ma nessuna è associata al punto 0) e non è adatta a distribuzioni dotate di una coda media o lunga (valori più estremi di quelli osservati avrebbero probabilità nulla). La versione fornita dal linguaggio *R* e utilizzata nel presente progetto

è:

$$p(k) = \begin{cases} \frac{k-\frac{3}{8}}{n-\frac{2}{8}}, & n \leq 10 \\ \frac{k-0.5}{n}, & n > 10 \end{cases} \tag{B.2.13}$$

Definizione B.2.10 (Mediana Campionaria)**.** Dato un campione $X_1, \ldots, X_n$, la *mediana campionaria* rappresenta una stima della *mediana* B.1.7 della popolazione da cui il campione è stato estratto. Si ottiene considerando le statistiche d'ordine $X_{(1)}, \ldots, X_{(n)}$ del campione $X_{(1)}, \ldots, X_{(n)}$, ottenute ordinando in modo crescente i valori del campione, e scegliendo come valore per la mediana:

- il valore centrale:

  $$M = X_{(\lfloor n/2 \rfloor + 1)}$$

  se n è dispari;

- la media aritmetrica tra i due valori centrali:

  $$M = \frac{X_{(n/2)} + X_{(n/2+1)}}{2}$$

  se n è pari;

Definizione B.2.11 (Test di Ipotesi)**.** Il *Test d'Ipotesi* consiste nel verificare la validità di una certa inferenza statistica fatta su una popolazione, dipendente da un parametro incognito θ, rispetto a un certo campione $X_1, \ldots, X_n$; per *ipotesi statistica* generalmente si intende un'asserzione su uno o più parametri della distribuzione della popolazione da cui provengono i campioni. L'ipotesi fatta sul parametro θ che deve essere verificata dal test, prende il nome di *ipotesi nulla* e si indica con $\mathcal{H}_0$; l'ipotesi contraria, invece, è chiamata *ipotesi alternativa*, e si indica con $\mathcal{H}_1$. Il risultato del test è un valore di probabilità, chiamato p-value, dal quale si decide se *rifiutare* o *non rifiutare* l'ipotesi nulla; tale probabilità rappresenta la probabilità minima di commettere un *Errore di Prima Specie* (*Type I Error*), indicato spesso con α, cioè di rifiutare l'ipotesi nulla quando questa in realtà è vera. È possibile commettere un altro tipo di errore, detto *Errore di Seconda Specie* (*Type II Error*), indicato spesso con β, rappresentante la probabilità di non rifiutare l'ipotesi nulla quando questa in realtà è falsa; l'errore di

seconda specie dipende dal paramertro θ e quindi, in generale, si avrà una funzione $\beta(\theta)$, detta *Curva Operativa Caratteristica* (*Operating Characteristics Curve*, in breve *OC*). La *Potenza* del test rappresenta la probabilità di rifiutare l'ipotesi nulla quando questa è effettivamente falsa (o, in maniera analoga, di accettare l'ipotesi alternativa quando questa è effettivamente vera); dato che la potenza varia in funzione del parametro θ considerato, essa sarà data da $1 - \beta(\theta)$, per un certo θ fissato.

Ad esempio, per verificare se un certo parametro θ (ad es., la media di una popolazione) è uguale a un certo valore θ_0, si può effettuare un test d'ipotesi, specificando, come ipotesi nulla, l'uguaglianza tra il parametro θ e il valore θ_0, mentre, come ipotesi alternativa, la diversità tra θ e θ_0. Il test può essere rappresentato in maniera compatta nel seguente modo:

$$\mathcal{H}_0 : \theta = \theta_0$$
$$\mathcal{H}_1 : \theta \neq \theta_0$$

Indicando con T la statistica del test, cioè la variabile casuale distribuita secondo una particolare distribuzione individuata dal test, si può calcolare il cosidetto *valore critico* c, ossia il valore della statistica del test calcolato in funzione di θ_0 e del campione $X_1 = x_1, \ldots, X_n = x_n$, attualmente sotto osservazione. Il p-value del test è dato dalla probabilità che la statistica T superi il valore critico c, sotto l'ipotesi di validità di $\mathcal{H}_0$, cioè:

$$p\text{-value} = \Pr\{T > c | \mathcal{H}_0\}$$

L'ipotesi nulla va quindi rifiutata per tutti i valori $\alpha \geq p$-value; la quantità α è detta *Livello di Significatività* e valori tipici sono 0.05 (*test significativo*), 0.01 (*test molto significativo*), 0.001 (*test altamente significativo*). Si noti che, nel caso di non rifiuto dell'ipotesi nulla, non è corretto affermare che "l'ipotesi nulla viene accettata" o che "l'ipotesi nulla è vera", in quanto il test viene condotto sulla base di un campione e non sull'intera popolazione, da cui il campione viene estratto; in effetti, l'unica cosa che si può dire in caso di non rifiuto, è che l'ipotesi nulla è compatibile con il campione analizzato o, in modo analogo,

che non vi è evidenza sperimentale sufficiente che induca a rifiutare l'ipotesi nulla.

Definizione B.2.12 (Coefficienti di Correlazione Campionaria)**.** I *Coefficienti di Correlazione Campionaria* forniscono una misura del grado di correlazione tra due insiemi di osserevazioni; il tipo di correlazione può, in generale, essere anche non lineare, sebbene non tutti i coefficienti sono in grado di catturarla. Fra i vari coefficienti di correlazione esistenti, quelli presi in considerazione nel progetto sono [3]:

- *Coefficiente di Correlazione secondo Pearson*

$$r_{xy} = \frac{\sum_{i=1}^{N} (x_i - \bar{x})\,(y_i - \bar{y})}{(N-1)\, s_x s_y}$$

  dove N è la numerosità dei due campioni $x_1, \ldots, x_N$ e $y_1, \ldots, y_N$, mentre s_x e s_y ne sono le rispettivte varianze campionarie.

- *Coefficiente di Correlazione secondo Kendall*

$$\tau = \frac{S}{\binom{N}{2}} = \frac{2S}{N(N-1)}$$

  dove N è il numero di elementi e S è la differenza tra il numero di *coppie concordanti* e il numero di *coppie discordanti* (Def. B.2.5).

- *Coefficiente di Correlazione secondo Spearman*

$$\rho = 1 - \frac{6\sum_{i=1}^{N} d_i^2}{N(N^2-1)} = 1 - \frac{6\sum_{i=1}^{N} \left(\mathrm{rank}(x_i) - \mathrm{rank}(y_i)\right)^2}{N(N^2-1)}$$

  dove $d_i = \mathrm{rank}(x_i) - \mathrm{rank}(y_i)$ è la differenza tra i ranghi dei corrispondenti valori di $x_1, \ldots, x_N$ e $y_1, \ldots, y_N$, e N è la numerosità dei due campioni.

Tutti e tre i coefficienti forniscono un valore compreso tra -1 e 1 Si noti che il coefficiente di Pearson è utilizzabile solo per determinare l'esistenza di una

[3] Nel caso in cui in uno dei due insiemi di dati vi siano dei "tie" (Def. B.2.6), le formule devono essere opportunamente riadattate.

relazione di tipo lineare; invece, i coefficienti di Kendall e Spearman sono dei coefficienti non parametrici più generici che riescono a catturare relazioni anche non lineari. Il coefficiente di Pearson ha, inoltre, lo svantaggio di essere sensibile alla presenza di "outlier" e di assumere la presenza di una relazione di normalità tra i due campioni (difatti, si tratta di una misura parametrica). Invece i coefficienti di Kendall e Spearman sono considerati più robusti in quanto si basano sulle statistiche d'ordine, cioè misurano l'associazione solo in funzione dei ranghi (per questo sono detti coefficienti *rank correlation*); di conseguenza, trasformazioni monotone preservano la correlazione basata sul rango. Tuttavia, nel caso in cui l'ipotesi di normalità tra i due campioni sia valida, la potenza dei coefficienti basati sul rango è di solito inferiore a quella del coefficiente di Pearson.

Quando l'assunzione di normalità è valida, esiste una relazione che lega il coefficiente di Pearson a quello di Kendall:

$$\tau = \frac{2}{\pi} \arcsin(r)$$

Appendice C

Trasformazione Integrale di Probabilità

In questo capitolo viene presentato un particolare tipo di trasformazione probabilistica che trova spesso utilizzo nei metodi di campionamento e in particolare nella generazione di numeri casuali. Questo metodo di trasformazione utilizza il concetto di funzione di distribuzione cumulativa e di funzione quantile; nella sezione §C.1, prima di presentare la definizione formale della trasformazione, verranno quindi richiamati alcuni dei concetti di base relativi alle distribuzioni di probabilità; la sezione §C.2 mostra, invece, alcuni esempi di utilizzo della trasformazione in questione.

C.1 Definizioni e Proprietà

Il metodo della *Trasformazione Integrale di Probabilità* utilizza i concetti di funzione di distribuzione cumulativa (CDF) e di funzione quantile; per maggiori informazioni su questi argomenti si vedano le Def. B.1.3 e B.1.6.

Teorema C.1.1 (Trasformazione Integrale di Probabilità (PIT))**.** *Sia X una variabile casuale la cui distribuzione di probabilità è rappresentata dalla CDF $F(\cdot)$. Si supponga che $F(\cdot)$ sia continua e che esista la sua inversa $F^{-1}(\cdot)$ (Funzione Quantile):*

$$F^{-1}(u) = \inf \{x | F(x) = u, 0 < u < 1\}$$

Ne segue che:

1. *se U è una variabile casuale Uniforme in $[0, 1]$, $F^{-1}(U)$ ha funzione di distribuzione $F(\cdot)$;*

2. *se X è una variabile casuale con funzione di distribuzione $F(\cdot)$, $F(X)$ è uniformemente distribuita in $[0, 1]$.*

Dimostrazione. Per dimostrare la prima affermazione è sufficiente notare che, data una variabile casuale U Uniforme in $[0, 1]$, si ha che per ogni $x \in \mathbb{R}$:

$$\begin{aligned}
\Pr\left\{F^{-1}(U) \leq x\right\} &= \Pr\left\{\inf_{t\in\mathbb{R}}\{t|F(t) = U\} \leq x\right\} \quad \text{(per la definizione di } F^{-1}(\cdot)\text{)} \\
&= \Pr\{U \leq F(x)\} \quad \text{(per la monotonicità di } F(\cdot)\text{)} \\
&= F(x) \quad \text{(per la definizione di CDF di una Uniforme in } [0,1]\text{: } F_U(u) = u\text{)}
\end{aligned}$$

Dall'ultima uguaglianza, segue che la variabile $X = F^{-1}(U)$ ha funzione di distribuzione $F(\cdot)$.

Per dimostrare la seconda affermazione, si parte dal fatto che per ogni $u \in [0, 1]$:

$$\begin{aligned}
\Pr\{F(X) \leq u\} &= \Pr\left\{X \leq F^{-1}(u)\right\} \quad \text{(per la monotonicità di } F\text{)} \\
&= F\left(F^{-1}(u)\right) \quad \text{(per la definizione di } F\text{)} \\
&= u \quad \text{(per la definizione di funzione inversa)}
\end{aligned}$$

Confrontando il primo e l'ultimo membro dell'uguaglianza, si ottiene la definizione di funzione di distribuzione di una distribuzione Uniforme in $[0, 1]$; ne segue che la variabile $U = F(X)$ è uniformemente distribuita in $[0, 1]$. □

C.2 Esempi di Utilizzo

In questa sezione verranno presentati alcuni esempi di applicazione del PIT.

C.2.1 Generazione di Numeri Casuali

Uno dei campi di applicazione del PIT più noti è sicuramente quello riguardante la generazione di numeri casuali; il metodo che ne deriva prende il nome di *Metodo dell'Inversione* (si veda [32]). Questo metodo permette di generare un numero casuale r distribuito secondo una certa distribuzione continua X, a partire da un numero casuale u, generato secondo una distribuzione Uniforme in $[0, 1]$, e dalla funzione $F^{-1}(\cdot)$, rappresentante l'inversa della CDF $F(\cdot)$ di X. Il metodo si compone di due passi:

1. Si genera un numero casuale u secondo una distribuzione Uniforme in $[0, 1]$.

2. Si calcola la quantità $x = F^{-1}(u)$.

Il valore x rappresenta un numero casuale generato secondo la distribuzione X.

C.2.2 Test di Adattamento a Distribuzioni

In questa sezione verrà mostrato come l'utilizzo del PIT permetta di rendere indipendenti dalla distribuzione di probabilità alcuni tipi di test di adattamento a distribuzioni teoriche, basati sulla funzione di distribuzione empirica (EDF). In particolare si evidenzierà come l'impiego del PIT permetta di considerare questi test come *Test di Uniformità* (cioè di adattamento a una distribuzione Uniforme – in particolare, Uniforme in $[0, 1]$); ciò può essere ottenuto osservando che i valori $F(x_1), \ldots, F(x_n)$, ricavati da un campione casuale $x_1, \ldots, x_n$ i.i.d. applicando la CDF di una distribuzione teorica, possono essere visti come un campione $u_1, \ldots, u_n$ i.i.d. Uniforme in $[0, 1]$. Questo fatto permette, ad esempio, di:

- stimare il *p-value* del test attraverso una simulazione *Monte Carlo*, utilizzando una qualsiasi distribuzione teorica e quindi anche la distribuzione Uniforme in $[0, 1]$;

- utilizzare questi test senza la necessità di conoscere i valori critici e i relativi *p-value* specifici per una particolare distribuzione teorica.

Si ricordi che un test di adattamento non è altro che un test di ipotesi in cui l'ipotesi nulla $\mathcal{H}_0$ afferma che il campione $x_1, \ldots, x_n$, sottoposto al test, sia distribuito secondo una certa distribuzione teorica con funzione di probabilità $p_0(\cdot)$:

$$\begin{aligned} \mathcal{H}_0 &: \Pr\{X = x_i\} = p_0(x_i), \quad \text{per ogni } i = 1, \ldots, n \\ \mathcal{H}_1 &: \Pr\{X = x_i\} \neq p_0(x_i), \quad \text{per qualche } i = 1, \ldots, n \end{aligned}$$

L'idea di base, per l'applicazione del PIT, è la seguente: dato un campione $x_1, \ldots, x_n$ i.i.d.[1], per verificare l'ipotesi $\mathcal{H}_0$ che il campione provenga da una certa distribuzione avente funzione di distribuzione $F(x|\mathcal{H}_0)$, è possibile considerare l'ipotesi $\mathcal{H}_0'$ che il corrispondente campione $u_1 = F(x_1|\mathcal{H}_0), \ldots, u_n = F(x_n|\mathcal{H}_0)$, ottenuto per mezzo del PIT, rappresenti un campione casuale proveniente dalla distribuzione Uniforme $U = F(x|\mathcal{H}_0)$, ossia che la funzione di distribuzione $F_U(u|\mathcal{H}_0')$ rappresenta la CDF di una distribuzione Uniforme in $[0, 1]$ (si veda [94]).

Si deve far notare che questo metodo non è, in generale, applicabile nei seguenti casi:

- la distribuzione di probabilità è multivariata; in questo caso si può comunque ricorrere alla cosidetta *Trasformazione Integrale di Probabilità Condizionata* (*CPIT*), in cui l'utilizzo delle distribuzioni condizionate (ottenute dalla scomposizione della distribuzione congiunta tramite la *chain-rule*) permette di ottenere campioni multivariati i.i.d. Uniformi in $[0, 1]^d$, dove d è la dimensione della distribuzione multivariata (si veda, ad es., [100, 72]);

- i parametri della distribuzione sono stati stimati dallo stesso campione per cui si vuole effettuare il test di adattamento (si veda, ad es., [30]);

[1] L'assunzione di "campione identicamente distribuito" non è strettamente necessaria (si veda [94]).

- le variabili casuali non sono continue (si veda, ad es., [31]).

Gli esempi seguenti rappresentano due test di adattamento basati sulla EDF: il test di Kolmogorov-Smirnov e il test di Anderson-Darling; per ognuno di essi, il PIT verrà applicato in due modi:

- alla CDF teorica $F\left(x|\mathcal{H}_0\right)$ dell'ipotesi nulla $\mathcal{H}_0$:

$$U = F\left(X|\mathcal{H}_0\right) \sim U(0,1) \tag{C.2.1}$$

  cioè la CDF $F(\cdot)$ può essere vista come una variabile Uniforme in $[0,1]$;

- alla definizione di EDF $F_n\left(\cdot\right)$, supponendo la validità dell'ipotesi nulla $\mathcal{H}_0$:

$$\begin{aligned} F_n\left(x|\mathcal{H}_0\right) &= \frac{\#\left\{x_i | x_i \leq x, i = 1,2,\ldots,n\right\}}{n} \\ &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{I}\left(\left\{x_i \leq x\right\}\right) \\ &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{I}\left(\left\{x_i \leq F^{-1}\left(u|\mathcal{H}_0\right)\right\}\right) \quad \text{(per la def. di } F^{-1}(\cdot)\text{, con } u = F(x)\text{)} \\ &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{I}\left(\left\{F\left(x_i|\mathcal{H}_0\right) \leq u\right\}\right) \quad \text{(per la monotonicità di } F(\cdot)\text{)} \\ &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{I}\left(\left\{U_i \leq u\right\}\right) \quad \text{(per il PIT)} \end{aligned} \tag{C.2.2}$$

  dove U_i, con $i = 1,\ldots,n$, sono variabili aleatorie i.i.d. con distribuzione Uniforme in $[0,1]$.

Di seguito, per comodità, si eviterà di specificare nella CDF e nella EDF l'assunzione di validità dell'ipotesi nulla $\mathcal{H}_0$.

C.2.3 Test di Kolmogorov-Smirnov

Sia X_i, con $i = 1,2,\ldots,n$, un campione i.i.d. di numerosità n, e sia $F_n\left(\cdot\right)$ la relativa funzione di distribuzione empirica. Si supponga, inoltre, che la fun-

zione $F(\cdot)$ sia una CDF invertibile associata alla distribuzione di probabilità teorica dell'ipotesi nulla $\mathcal{H}_0$ del test di adattamento. Il test di Kolmogorov-Smirnov (a un campione) §3.2.4 è definito come la massima distanza tra la EDF e la CDF della distribuzione dell'ipotesi nulla $\mathcal{H}_0$, calcolata sul campione sotto l'assunzione di validità dell'ipotesi nulla $\mathcal{H}_0$:

$$D = \sup_{x \in \mathbb{R}} |F_n(x) - F(x)|$$

Indicando con d il valore critico della statistica del test D, e applicando il PIT (Eq. C.2.1 e C.2.2), si può scrivere che:

$$\begin{aligned}
\Pr\{D \leq d\} &= \Pr\left\{\sup_{x \in \mathbb{R}} |F_n(x) - F(x)| \leq d\right\} \\
&= \Pr\left\{\sup_{u \in [0,1]} \left|\frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(\{U_i \leq u\}) - u\right| \leq d\right\} \\
&= \Pr\left\{\sup_{i=1,\ldots,n} \left\{\frac{i}{n} - U_{(i)}, U_{(i)} - \frac{i-1}{n}\right\} \leq d\right\}
\end{aligned}$$

dove $U_{(i)}$, con $i = 1, \ldots, n$ è il campione U_i, con $i = 1, \ldots, n$ riordinato in modo crescente.

Tale uguaglianza afferma che il test di adattamento alla distribuzione $F(\cdot|\mathcal{H}_0)$, secondo Kolmogorov-Smirnov, può essere visto come un test di uniformità effettuato sul campione $U_i = F(x_i|\mathcal{H}_0)$, con $i = 1, \ldots, n$ i.i.d., rispetto alla distribuzione Uniforme in $[0, 1]$; ciò dimostra che il test di Kolmogorov-Smirnov è indipendente dalla distribuzione $F(\cdot|\mathcal{H}_0)$ (*distribution-free*). Per maggiori informazioni si veda [101].

C.2.4 Test di Anderson-Darling

Si supponga che $F(\cdot)$ sia una CDF invertibile associata alla distribuzione teorica dell'ipotesi nulla $\mathcal{H}_0$ del test di adattamento. Sia, inoltre, X_i, con $i = 1, 2, \ldots, n$, un campione i.i.d. di numerosità n, e sia $F_n(\cdot)$ la relativa funzione di distribuzione empirica. Il test di Anderson-Darling (a un campione) §3.2.5 è definito come la somma pesata delle distanze quadratiche tra la EDF

e la CDF della distribuzione dell'ipotesi nulla $\mathcal{H}_0$, calcolata sul campione sotto l'assunzione di validità dell'ipotesi nulla $\mathcal{H}_0$:

$$A_n^2 = n \int_{-\infty}^{\infty} \frac{\left[F_n(x) - F(x)\right]^2}{F(x)\left(1 - F(x)\right)} dF(x)$$

Indicando con a il valore critico della statistica del test A_n^2, e applicando il PIT (Eq. C.2.1 e C.2.2), si può scrivere che:

$$\begin{aligned}
\Pr\left\{A_n^2 \leq a\right\} &= \Pr\left\{ n \int_{-\infty}^{\infty} \frac{\left[F_n(x) - F(x)\right]^2}{F(x)\left(1 - F(x)\right)} dF(x) \leq a \right\} \\
&= \Pr\left\{ n \int_0^1 \frac{\left[\frac{1}{n}\sum_{i=1}^n \mathbb{I}\left(\{U_i \leq u\}\right) - u\right]^2}{u\left(1-u\right)} du \leq a \right\} \\
&= \Pr\left\{ -n - \frac{1}{n}\sum_{i=1}^{n} \left(2i-1\right)\left[\ln U_{(i)} + \ln\left(1 - U_{(n-i+1)}\right)\right] \leq a \right\}
\end{aligned}$$

dove $U_{(i)}$, con $i = 1, \ldots, n$ è il campione U_i, con $i = 1, \ldots, n$ riordinato in modo crescente.

Per gli stessi ragionamenti effettuati nella sezione precedente, per il test di Kolmogorov-Smirnov, da questa equazione si evince che ili test di Anderson-Darling è un test *distribution-free*.

Appendice D

Metodi delle Trasformate

I metodi delle trasformate sono tecniche matematiche che si rivelano particolarmente utili, ad esempio, nella risoluzione di alcune tipologie di equazioni differenziali e di problemi che coinvolgono la somma di variabili casuali indipendenti.

Prima di illustrare i vari metodi di trasformazione, nella sezione §D.1 viene introdotta la cosidetta *Funzione Generatrice dei Momenti*, dalla quale si derivanno il metodo della *Trasformata-z*, nella sezione §D.2, e il metodo della *Trasformata di Laplace-Stieltjes*, nella sezione §D.3.

D.1 Funzione Generatrice dei Momenti (MGF)

Definizione D.1.1 (Funzione Generatrice dei Momenti (MGF))**.** Data una variabile casuale X, si definisce *Funzione Generatrice dei Momenti* (*MGF*) $M_X(\theta)$ di X la funzione:

$$M_X(\theta) = E\left[e^{X\theta}\right], \quad \theta \in \mathbb{R} \tag{D.1.1}$$

ammesso che il valor medio esista.

Dalla definizione, segue immediatamente che:

$$M_X(\theta) = \begin{cases} \sum_i e^{x_i\theta} p_X(x_i), & \text{se } X \text{ discreta} \\ \int_{-\infty}^{\infty} e^{x\theta} f_X(x)\, dx, & \text{se } X \text{ continua} \end{cases}$$

La relazione tra la MGF e i momenti della variabile X è data dalla seguente:

Proposizione D.1.1. *Data una variabile aleatoria X, supponendo che la MGF $M_X(\theta)$ esista, il k-esimo momento di X è dato da:*

$$E\left[X^k\right] = M_X^{(k)}(0) = \frac{d}{d\theta^k} M_X(\theta)\Big|_{\theta=0}, \quad k = 1, 2, \ldots \tag{D.1.2}$$

La MGF gode di tre interessanti proprietà, di seguito citate.

Proposizione D.1.2 (Traslazione Lineare)**.** *Siano X e Y due variabili aleatorie tale che $Y = aX + n$, con $a, b \in \mathbb{R}$. allora*

$$M_Y(\theta) = e^{b\theta} M_X(a\theta)$$

Proposizione D.1.3 (Convoluzione)**.** *Sia $X = \sum_{i=1}^{n} X_i$ la somma di n variabili aleatorie $X_1, \ldots, X_n$ i.i.d. e mutuamente esclusive; supponendo che $M_{X_i}(\theta)$ esista per ogni i, allora $M_X(\theta)$ esiste ed è pari a:*

$$M_X(\theta) = \sum_{i=1}^{n} M_{X_i}(\theta)$$

Proposizione D.1.4 (Corrispondenza o Univocità)**.** *Date due variabili aleatorie X_1 e X_2, se $M_{X_1}(\theta) = M_{X_2}(\theta)$, per ogni θ, allora $F_1(x) = F_2(x)$, per ogni x.*

D.2 Trasformata-z (PGF)

Il concetto di *Trasformata-z* [109], o *Funzione Generatrice delle Probabilità*, è un utile strumento per la semplificazione dei calcoli che coinvolgono variabili causali discrete a valori non negativi.

Definizione D.2.1 (Trasformata z (PGF))**.** Data una variabile casuale X discreta e con supporto (intero) non negativo, la sua *Trasformata-z*, o *Funzione Generatrice di Probabilità* (*PGF*), è definita come:

$$G_X(z) = E\left[z^X\right] = M_X(\ln z) = \sum_{x=0}^{\infty} p_X(x)\, z^x \tag{D.2.1}$$

Si può dimostrare che la Trasformata-z $G_X(z)$ di X converge per ogni $z \in \mathbb{C}$ tale che $|z| \leq 1$ [1]. Inoltre, dato che la PGF è un caso particolare della MGF, ne condivide le proprietà; ad es. per la proprietà D.1.4, se due variabili discrete X e Y hanno la stessa PGF, allora hanno anche la stessa distribuzione di probabilità.

D.3 Trasformata di Laplace-Stieltjes (LST)

Il concetto di *Trasformata di Laplace-Stieltjes* [109], insieme a quello della *Trasformata di Laplace*, è un'utile tecnica per l'analisi di sistemi lineari invarianti nel tempo, che serve per semplificare la descrizione funzionale del comportamento nel sistema (a volte si dice che permette di passare dal "dominio del tempo" al "dominio delle frequenze").

Definizione D.3.1 (Trasformata di Laplace (LT) Unilaterale)**.** Data una generica funzione $\varphi : \mathbb{R} \to \mathbb{R}$, la sua *Trasformata di Laplace* (*LT*) *Unilaterale*, è definita come:

$$\{\mathcal{L}\varphi\}(s) = \varphi^*(s) = \int_0^{\infty} e^{-st}\varphi(x)\, dt, \quad s \in \mathbb{C} \tag{D.3.1}$$

Definizione D.3.2 (Trasformata di Laplace-Stieltjes (LST) Unilaterale)**.** Data una generica funzione $\varphi : \mathbb{R} \to \mathbb{R}$, la sua *Trasformata di Laplace-Stieltjes* (*LST*) *Unilaterale*, è definita come:

$$\{\mathcal{L}^*\varphi\}(s) = \int_0^{\infty} e^{-st}\, d\varphi(t), \quad s \in \mathbb{C} \tag{D.3.2}$$

[1] Sia $z \in \mathbb{C}$, il suo valore assoluto è $|z| = \sqrt{\Re(z)^2 + \Im(z)^2}$.

Se X è una variabile casuale continua con CDF $F_X(x)$ e e PDF $f_X(x)$, si ha che:

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\, dt = \int_{-\infty}^{x} dF_X(t)$$

perciò è possibile affermare che:

Proposizione D.3.1. *Data una variabile casuale X continua e con supporto non negativo, la LST Unilaterale di $F_X(\cdot)$, è definita come la LT Unilaterale di $f_X(\cdot)$, ossia:*

$$\{\mathcal{L}^* F_X\}(s) = \int_0^{\infty} e^{-sx}\, dF_X(x) = \int_0^{\infty} e^{-sx} f_X(x)\, dx = \{\mathcal{L} f_X\}(s) \qquad \text{(D.3.3)}$$

Inoltre, dalla definizione di MGF D.1.1, risulta immediatamente che:

$$\{\mathcal{L} f_X\}(s) = \{\mathcal{L}^* F_X\}(s) = M_X(-s)$$

Appendice E

Catene di Markov

Questo capitolo fornisce un'introduzione sulle *Catene di Markov*, uno degli strumenti analitici più utilizzati non solo nella valutazione delle prestazioni dei sistemi, ma anche in molte altre aree scientifiche, come la biologia, la fisica, la meteorologia, ...Nel contesto di questo documento, le Catene di Markov sono fondamentali per la comprensione di alcuni degli argomenti illustrati nei capitoli successivi (ad es. il capitolo sulle distribuzioni *Phase-Type* Cap. 5).

La prima sezione §E.1 fornisce una visione generica sulle Catene di Markov, motivando la loro importanza attraverso alcuni casi di utilizzo teorici e pratici, e introduce le definizioni di base. La sezione §E.2 descrive una particolare classe di Catene di Markov, chiamata Catene di Markov a Tempo Discreto, o semplicemente Catene di Markov Discrete; alcune delle proprietà esposte in questa sezione sono necessarie per la comprensione anche di altre classi di Catene di Markov, descritte nelle sezioni successive. La sezione §E.3 illustra un'altra importante classe di Catene di Markov, chiamata Catene di Markov a Tempo Continuo, o semplicementente Catene di Markov Continue. Infine, nella sezione §E.4 viene fornito un accenno a una famiglia di processi, detti processi Semi-Markoviani, per i quali le proprietà Markoviane valgono solo in parte. Per ulteriori approfondimenti si veda, ad esempio, [76, 16, 109].

E.1 Panoramica sulle Catene di Markov

Le *Catene di Markov* (*MC*) rappresentano una particolare classe di *Processi Stocastici*; la loro definizione si deve agli studi del matematico russo *A. A. Markov* (1856-1922), da cui hanno ereditato il nome. Il matematico russo *A. N. Kolmogorov* (1903-1987) introdusse, successivamente, una generalizzazione per spazi degli stati infinitamente enumerabili.

Le MC trovano applicazione nei più svariati campi della scienza: nella fisica, nella biologia, in bio-informatica, nella teoria delle code, in intelligenza artificiale, ... Prima di procedere a una definizione formale, vengono presentati alcuni esempi di utilizzo sia pratico sia teorico.

Esempio E.1.1 (Meteorologia)**.** Nella meteorologia le MC vengono spesso utilizzate come modello probabilistico per la quantità giornaliera di pioggia, per lo studio della dinamica del vento, ... In quest'ultimo caso, ad esempio, la serie temporale delle velocità del vento può essere divisa in vari stati a seconda delle media e varianza campionaria; a questo punto la MC può essere utilizzata per generare, attraverso simulazione stocastica, delle serie sintetiche di velocità eoliche.

Esempio E.1.2 (Google "PageRank")**.** L'algortimo *PageRank* [91, 90, 92], usato da *Google* per assegnare un punteggio alle pagine "Web" in base alla loro popolarità, utilizza le MC: il "Web" viene visto come un grafo diretto in cui i nodi sono le pagine e gli archi sono i collegamenti ipertestuali tra le pagine. Il "PageRank" di un sito i rappresenta la probabilità stazionaria di un nodo associato a i in una MC costruita sul grafo del "Web"; le probabilità di transizione p_{ij} della MC modellano le probabilità che un utente casuale (*random surfer*) passi dalla pagina i alla pagina j.

Esempio E.1.3 (Inferenza Bayesiana)**.** L'inferenza esatta su reti Bayesiane "multiply-connected" (cioè contenenti dei nodi connessi da più cammini indiretti) è un problema *NP-Hard*; nasce quindi la necessità di utilizzare algoritmi approssimati. Uno dei migliori algoritmi approssimati per inferenza su reti Bayesiane è un algoritmo "Markov Chain Monte Carlo" (MCMC) chiamato

Gibbs sampler [51, 102], in cui l'inferenza viene effettuata attraverso la simulazione di una MC: ogni stato corrisponde a una particolare assegnazione delle variabili (nodi) della rete, e le transizioni rappresentano il passaggio da una particolare configurazione delle variabili a un'altra.

Esempio E.1.4 (Generazione di Numeri Casuali)**.** Le MC hanno acquistato anche grande importanza nella generazione di numeri casuali relativi a una certa distribuzione di probabilità (generalmente complicata); la simulazione stocastica tramite algoritmi "Markov Chain Monte Carlo" (MCMC) [51] è una delle tecniche più utilizzate.

Esempio E.1.5 (Processi Branching o Jump)**.** Si consideri un particolare tipo di particelle che possa generarne delle altre; il numero di particelle create ξ_i, con $i = 1, \ldots, \rho$, da ogni particella ρ nella generazione n è indipendente e identicamente distribuito secondo una distribuzione $\mathcal{D}$. Ogni particella, dopo aver generato i suoi discendenti, muore. Il numero $X(n)$ di particelle presenti nella generazione n rappresenta una MC. Lo stato in cui la popolazione di particelle è "esaurita" (ossia nessuna nuova particella può essere generata) è chiamato "stato assorbente"; una volta che la MC vi entra non potrà più uscirne.

Esempio E.1.6 (Processi Nascita-Morte)**.** Si consideri una MC che da un certo stato x possa solo spostarsi, nel passo successivo, in uno degli stati "vicini" a quello corrente: lo stato $x - 1$, rappresentante una "morte", lo stato x, oppure lo stato $x + 1$, rappresentante una "nascita". La nascita in uno stato x avviene secondo un tasso $\lambda_x > 0$ ("tasso di nascita"), mentre la relativa morte, avviene secondo un tasso $\mu_x > 0$ ("tasso di morte"). Un processo stocastico siffatto prende il nome di "Processo Nascita-Morte". Processi di questo tipo si incontrano spesso, per esempio, nella teoria delle code in cui una nascita può essere interpretata come l'arrivo di un job nel sistema, mentre una morte, come la partenza di un job dal sistema.

Dagli esempi sopra esposti, dovrebbe risultare abbastanza chiara l'importanza che le MC hanno nella costruzione di modelli probabilistici. Prima di procedere alla distinzione tra le varie classi di MC, vengono fornite alcune definizioni di base.

Definizione E.1.1 (Processo Stocastico)**.** Un processo stocastico $\mathcal{X} = \{X(t)|t \in T\}$ è una famiglia di variabili casuali $X(t)$, definite su un particolare spazio di probabilità, indicizzata da un parametro t appartenente a un insieme di indici T.

I valori assunti dalla variabile casuale $X(t)$ sono detti *stati*; l'insieme di tutti i possibili stati costituisce lo *spazio degli stati* $\mathcal{S}$ del processo. L'insieme dei parametri T viene anche chiamato "tempo".

Definizione E.1.2 (Processo di Markov)**.** Un *Processo di Markov* è un processo stocastico $\{X(t)|t \in T\}$ in cui per ogni $t_0 < t_1 < \cdots < t_n < t_{n+1}$ vale la *Proprietà di Markov*, cioè:

$$\begin{gathered}\Pr\left\{X\left(t_{n+1}\right) \leq x_{k_{n+1}} | X\left(t_n\right) = x_{k_n}, \ldots, X\left(t_0\right) = x_{k_0}\right\} \\ = \\ \Pr\left\{X\left(t_{n+1}\right) \leq x_{k_{n+1}} | X\left(t_n\right) = x_{k_n}\right\}\end{gathered}$$

La proprietà di Markov, anche detta *di assenza di memoria*, afferma che, dato lo stato corrente del processo ($X\left(t_n\right) = x_{k_n}$), lo stato futuro ($X\left(t_{n+1}\right) \leq x_{k_{n+1}}$) è indipendente da quello passato ($X\left(t_k\right) = x_{k_h}$, con $0 \leq h < n$), cioè lo stato futuro dipende esclusivamente dallo stato corrente e non da come si è giunti in tale stato. Detto in altri termini, ogni stato futuro è *condizionalmente indipendente* da ogni stato precedente, dato lo stato corrente.

Definizione E.1.3 (Catena di Markov)**.** Una *Catena di Markov* (*MC*) è un processo di Markov $\{X(t)|t \in T\}$ a stati discreti (cioè con spazio degli stati finito o infinitamente enumerabile), ovvero un processo stocastico discreto in cui ad ogni istante di tempo t si estrae una variabile casuale discreta e per il quale vale la *Proprietà di Markov*:

$$\begin{gathered}\Pr\left\{X\left(t_{n+1}\right) = x_{k_{n+1}} | X\left(t_n\right) = x_{k_n}, \ldots, X\left(t_0\right) = x_{k_0}\right\} \\ = \\ \Pr\left\{X\left(t_{n+1}\right) = x_{k_{n+1}} | X\left(t_n\right) = x_{k_n}\right\}\end{gathered}$$

E.2 Catene di Markov a Tempo Discreto (DTMC)

In questa sezione viene descritta una particolare classe di Catene di Markov chiamata *Catene di Markov a Tempo Discreto*, in cui l'insieme degli indici T è discreto; dopo aver introdotto le definizioni di base, verranno presentate le equazioni di *Chapman-Kolmogorov*, le quali permettono di descrivere il comportamento di una Catena di Markov; seguirà quindi una caratterizzazione degli stati di una Catena di Markov, lo studio della Catena *a regime* e, infine, il caso di Catene di Markov aventi stati *assorbenti*, il quale ritornerà utile per l'analisi delle distribuzioni *Phase-Type* Cap. 5.

Definizione E.2.1 (Catena di Markov a Tempo Discreto (DTMC))**.** Una *Catena di Markov a Tempo Discreto* (*DTMC*), o semplicemente *Catena di Markov Discreta*, è una Catena di Markov $\{X_n | n \in T\}$ in cui l'insieme dei parametri T é discreto, cioè $T = \{0, 1, 2, \dots\}$.

La Fig. E.1 mostra una possibile rappresentazione grafica dell'evoluzione di una DTMC: sull'asse delle ascisse è indicato il tempo n (in passi), mentre su quello delle ordinate vi sono i possibili stati che il processo X_n può assumere; ogni punto rappresenta lo stato in cui il processo X_n si trova al passo n.

Si indica con $p_i\,(n)$ la *probabilitá di stato* al tempo n, cioè la probabilità incondizionata $\Pr\{X_n = i\}$ che il processo si trovi nello stato i al tempo n, mentre con $p_{ij}\,(m, n)$, la *probabilità di transizione* nello stato j al tempo n sapendo che il processo si trovava nello stato i al tempo m, ossia la probabilità condizionata $\Pr\{X_n = j | X_m = i\}$. Spesso è conveniente utilizzare una notazione matriciale:

- *vettore delle probabilitá di stato* al tempo n: il vettore $\vec{\mathbf{p}}(n) = [p_i\,(n)]$, con $i \in \mathcal{S}$ e $\mathcal{S}$ spazio degli stati;
- *vettore delle probabilitá iniziali*: il vettore $\vec{\mathbf{p}}_0 = \vec{\mathbf{p}}(0)$;
- *matrice delle probabiità di transizione* dal tempo m al tempo n: la matrice $\mathbf{P}(m, n) = [p_{ij}(m, n)]$, con $i, j \in \mathcal{S}$ e $\mathcal{S}$ spazio degli stati.

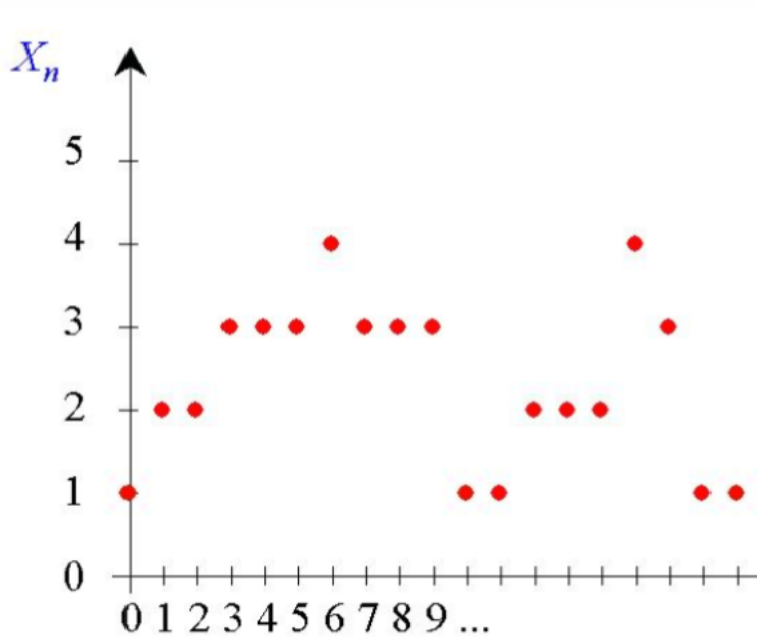


Figura E.1: Rappresentazione grafica dell'evoluzione di una DTMC.

Per il vettore $\vec{\mathbf{p}}$ e la matrice $\mathbf{P}$ devono valere le seguenti *condizioni di normalizzazione*:

$$\begin{aligned} \vec{\mathbf{p}}\,(n)\,\vec{\mathbf{1}}^T &= 1 \\ \mathbf{P}(m,n)\vec{\mathbf{1}}^T &= \vec{\mathbf{1}}^T \end{aligned} \tag{E.2.1}$$

Una matrice $\mathbf{P}$ che soddisfa tali condizioni è detta *matrice stocastica* [1].

E.2.1 Omogeneità

Definizione E.2.2 (Omogeneità)**.** Una DTMC è detta *omogenea (nel tempo)* se vale la proprietà di invarianza rispetto all'origine temporale n, cioè se per ogni n:

$$\Pr\left\{X_{n+1} = x_{k_{n+1}} | X_n = x_{k_n}\right\} = \Pr\left\{X_n = x_{k_n} | X_{n-1} = x_{k_{n-1}}\right\}$$

La proprietà di *omogeneità*, grazie alla quale le transizioni non dipendono dal tempo ma solo dalla differenza temporale, permette di scrivere le probabilità di transizione come $p_{ij}\,(m,n) = p_{ij}\,(n-m)$, con $m \leq n$. Il vettore $\vec{\mathbf{p}}\,(n)$ è anche chiamato *vettore delle probabilità di stato a n passi*, mentre la matrice $\mathbf{P}\,(n)$ è detta *matrice delle probabilità di transizione a n passi*.

E.2.2 Equazioni di Chapman-Kolmogorov

Proposizione E.2.1 (Equazioni di Chapman-Kolmogorov)**.** *L'evoluzione di una DTMC è descritta dalle seguenti* Equazioni di Chapman-Kolmogorov*:*

$$\begin{aligned} p_j\,(n) &= \sum_{k\in\mathcal{S}} p_k\,(m)\,P_{kj}\,(m,n) \\ P_{ij}\,(m,n) &= \sum_{k\in\mathcal{S}} P_{ik}\,(m,u)\,P_{kj}\,(u,n) \end{aligned} \tag{E.2.2}$$

per ogni $i, j \in \mathcal{S}$ e $m < n$. In forma matriciale si ha:

$$\begin{aligned} \vec{\mathbf{p}}\,(n) &= \vec{\mathbf{p}}\,(m)\,\mathbf{P}\,(m,n) \\ \mathbf{P}\,(m,n) &= \mathbf{P}\,(m,u)\,\mathbf{P}\,(u,n) \end{aligned} \tag{E.2.3}$$

[1] Una matrice $\mathbf{A}$ è *stocastica* se è quadrata, se per ogni elemento a_{ij} vale $0 \leq a_{ij} \leq 1$ e se ogni riga somma a 1, cioè $\forall i : \sum_j a_{ij} = 1$.

con $0 \leq m < u < n$. *Le condizioni di normalizzazione sono:*

$$\begin{aligned} \vec{\mathbf{p}}\,(n)\,\vec{\mathbf{1}}^T &= 1 \\ \mathbf{P}(m,n)\vec{\mathbf{1}}^T &= \vec{\mathbf{1}}^T \end{aligned} \tag{E.2.4}$$

Se la DTMC è omogenea, *posto* $\mathbf{P} = \mathbf{P}\,(1)$, *le equazioni diventano:*

$$\begin{aligned} \vec{\mathbf{p}}\,(n) &= \vec{\mathbf{p}}\,(n-1)\,\mathbf{P} = \vec{\mathbf{p}}_0\mathbf{P}^n \\ \mathbf{P}\,(n) &= \mathbf{P}\,(n-1)\,\mathbf{P} = \mathbf{P}^n \end{aligned} \tag{E.2.5}$$

mentre le condizioni di normalizzazione sono:

$$\begin{aligned} \vec{\mathbf{p}}\,(n)\,\vec{\mathbf{1}}^T &= 1 \\ \mathbf{P}(n)\vec{\mathbf{1}}^T &= \vec{\mathbf{1}}^T \end{aligned} \tag{E.2.6}$$

Alle equazioni di Chapman-Kolmogorov è possibile dare un'interpretazione associata al *principio di conservazione del flusso*:

$$\langle \text{rate of build-up} \rangle = \langle \text{rate of flow-in} \rangle - \langle \text{rate of flow-out} \rangle$$

che, nel caso omogeneo, si esprime come:

$$p_j\,(n) - p_j\,(n-1) = \sum_{i \neq j} p_i\,(n-1)\,P_{ij} - p_j\,(n-1) \sum_{i \neq j} P_{ji} \tag{E.2.7}$$

La suddetta equazione si ricava dalle equazioni di Chapman-Kolmogorov E.2.5 e dalle condizioni di normalizzazione E.2.6:

$$\begin{aligned} p_j\,(n) &= \sum_i p_i\,(n-1)\,P_{ij} \\ &= \sum_{i \neq j} p_i\,(n-1)\,P_{ij} + p_j\,(n-1)\,P_{jj} \\ &= \sum_{i \neq j} p_i\,(n-1)\,P_{ij} + p_j\,(n-1)\left(1 - \sum_{i \neq j} P_{ji}\right) \\ &= \dots \end{aligned}$$

E.2.3 Riducibilità

Definizione E.2.3 (Riducibilità)**.** Uno stato j è *raggiungibile* da uno stato i ($i \rightarrow j$) se vi è una probabilità non nulla di trovarsi, nel futuro, nello stato j, sapendo che al momento ci si trovi nello stato i, cioè se $\exists m, n : m < n \wedge p_{ij}(m, n) > 0$. Due stati i e j sono *comunicanti* ($i \leftrightarrow j$) se sono mutuamente raggiungibili. Un sottoinsieme C dello spazio degli stati $\mathcal{S}$ rappresenta una *classe comunicante* se ogni coppia di stati in C è comunicante, cioè se $\forall i, j \in C : p_{ij} > 0$. Una classe comunicante C si dice *chiusa* se la probabilità di lasciare la classe è nulla, cioè se $\forall i, j : i \in C \wedge j \notin C \wedge p_{ij} = 0$. Quando un insieme chiuso contiene un solo stato i, si dice che i è uno stato *assorbente*; in tal caso $p_{ii} = 1$. Una MC si dice *irriducibile* se tutti gli stati appartengono a un singolo insieme chiuso, cioè se lo spazio degli stati $\mathcal{S}$ è chiuso. Ciò implica che in una MC irriducibile tutti gli stati sono fra loro comunicanti.

E.2.4 Periodicità

Definizione E.2.4 (Periodicità)**.** Uno stato i ha *periodo* d_i se il ritorno a tale stato avviene in un numero di passi multiplo di d_i, dove d_i è il più grande numero intero positivo che soddisfa tale proprietà; cioè, il periodo d_i di uno stato i si definisce come:

$$d_i = \gcd\{n : p_{ii}(n) > 0\}$$

Se $d_i = 1$, lo stato è detto *aperiodico*; se $d_i > 1$, lo stato è detto *periodico* di periodo d_i. Si può dimostrare che ogni stato appartenente alla stessa classe comunincante ha lo stesso periodo. Una MC è detta *ergodica* se è irriducibile e tutti i suoi stati sono aperiodici.

E.2.5 Ricorrenza

Definizione E.2.5 (Ricorrenza)**.** Uno stato i si dice *transiente* se esiste una probabilità non nulla di non farvi mai ritorno, supponendo che si parta da esso; uno stato è *ricorrente* se la probabilità di ritorno è l'evento certo; formalmente, sia $f_{ii}(n) = \Pr\{X_n = i, X_{n-1} \neq i, \ldots, X_1 \neq i | X_0 = i\}$ la probabilità che, dato

che la DTMC si trovi nello stato i, il prossimo ritorno allo stato i avvenga in n passi, cioè che il *tempo di ricorrenza* $T_i = \min_k \{X_k = i | X_0 = i\}$ (*hitting time*) nello stato i valga n; la probabilità che la DTMC ritorni prima o poi nello stato i vale:

$$f_{ii} = \sum_{k=1}^{\infty} f_{ii}(k)$$

Uno stato è *ricorrente* se $f_{ii} = 1$ mentre è *transiente* se $f_{ii} < 1$. Se il *tempo medio di ricorrenza* $E[T_i] = \mu_i$ nello stato i è finito, lo stato i è detto *positivo ricorrente*; se $\mu_i = \infty$, lo stato i è detto *ricorrente nullo*. Il *tempo medio di ritorno* μ_i nello stato i può essere calcolato come:

$$E[T_i] = \mu_i = \sum_{k=1}^{\infty} k f_{ii}(k)$$

Uno stato è *assorbente* se, una volta che il processo vi è entrato, non riesce più a uscirne, cioè:

$$\forall i, j \in \mathcal{S} : i \neq j \wedge p_{ii} = 1 \wedge p_{ij} = 0$$

Lo spazio degli stati $\mathcal{S}$ di una DTMC può essere partizionato in un insieme di stati transienti e in un insieme chiuso di stati ricorrenti (eventualmente assorbenti).

E.2.6 Ergodicità

Definizione E.2.6 (Ergodicità)**.** Uno stato è *ergodico* se è ricorrente positivo e aperiodico.

Una DTMC è *ergodica* se tutti i suoi stati sono ergodici, cioè se è irriducibile, aperiodica, e tutti i suoi stati sono ricorrenti positivi.

E.2.7 Analisi a Regime di una DTMC omogenea

Risulta particolarmente interessante conoscere quale sia la probabilità π_i, con $i \in \mathcal{S}$, che una DTMC omogenea si trovi, a regime (cioè nel lungo termine), in un certo stato i; in generale non è detto che per una MC esista una *distri-*

buzione stazionaria; se però esiste si possono ottenere importanti risultati, come l'indipendenza dalla distribuzione iniziale una volta che la MC raggiunge la condizione di equilibrio. Occorre quindi conoscere quale siano le condizioni di esistenza delle probabilità π_i, se i valori π_i costituiscano una distribuzione di probabilità, e in che modo è possibile calcolarli.

Definizione E.2.7 (Probabilità di Stato Limite)**.** Si definisce *vettore delle probabilità di stato limite* o *distribuzione limite* il vettore:

$$\vec{\pi} = \lim_{n\to\infty} \vec{\mathbf{p}}\,(n) \tag{E.2.8}$$

Se tale limite esiste, cioè se $\forall i \in \mathcal{S} : \pi_i \in \mathbb{R}$, si può dimostrare che $\sum_{i\in\mathcal{S}} \pi_i \leq 1$.

Definizione E.2.8 (Probabilità di Stato Stazionario)**.** Se la distribuzione limite $\vec{\pi}$ esiste e se risulta che:

$$\vec{\pi}\vec{\mathbf{1}}^T = 1 \tag{E.2.9}$$

il vettore $\vec{\pi}$ è detto *vettore delle probabilità di stato stazionario* o *distribuzione stazionaria*.

L'esistenza di tale vettore indica che nel lungo termine, l'influenza dello stato iniziale e gli effetti degli stati transienti sulla MC sono cessati e la MC ha raggiunto uno stato stazionario; le probabilità π_i possono essere viste come la porzione di tempo che la MC spende nello stato i nel lungo termine. Per la definizione di distribuzione stazionaria e per le considerazioni fatte in precedenza §E.2.2, si ottiene il sistema:

$$\begin{cases} \vec{\pi} = \vec{\pi}\mathbf{P} \\ \vec{\pi}\vec{\mathbf{1}}^T = 1 \\ \vec{\pi} \geq \vec{\mathbf{0}} \end{cases} \tag{E.2.10}$$

Una possibile interpretazione di questo sistema è che per ogni stato i il "flusso entrante" $\sum_{k\in\mathcal{S},k\neq i} \pi_k p_{ki}$ deve essere uguale al "flusso uscente" $\pi_i\,(1 - p_{ii})$. Un'altra possibile interpretazione è che l vettore $\vec{\pi}$ può essere visto come un *autovettore* della matrice $\mathbf{P}$ relativo all'*autovalore* 1.

Quando un tale vettore $\vec{\pi}$ esiste, risulta che:

1. La MC si "dimentica" da dove ha iniziato (cioè la distribuzione iniziale) e converge a una distribuzione stazionaria:

$$\lim_{n\to\infty} \mathbf{P}^n = \begin{bmatrix} \vec{\pi} \\ \vdots \\ \vec{\pi} \end{bmatrix} = \vec{\mathbf{1}}^T \vec{\pi}$$

2. Per ogni $\vec{\mathbf{p}}_0$:

$$\lim_{n\to\infty} \vec{\mathbf{p}}\,(n) = \lim_{n\to\infty} \vec{\mathbf{p}}_0 \mathbf{P}^n = \vec{\mathbf{p}}_0 \vec{\mathbf{1}}^T \vec{\pi} = \vec{\pi}$$

3. Se $\vec{\mathbf{p}}_0 = \vec{\pi}$, allora $\vec{\mathbf{p}}\,(n) = \vec{\pi}$ per ogni n.

Teorema E.2.2. *In una DTMC ergodica (irriducibile, aperiodica con tutti gli stati ricorrenti positivi), il vettore limite $\vec{\pi}$ esiste, è unico e rappresenta il vettore delle probabilità di stato stazionario.*

Inoltre risulta che:

$$\pi_i = \lim_{n\to\infty} p_i\,(n) = \frac{1}{\mu_i} \tag{E.2.11}$$

essendo μ_i il tempo medio di ricorrenza nello stato i §E.2.5.

E.2.8 DTMC con stati assorbenti

In una DTMC composta da un insieme di stati transienti e un insieme chiuso di stati ricorrenti, la matrice delle probabilità di transizione (a un passo) può essere così partizionata:

$$\mathbf{P} = \begin{bmatrix} \mathbf{T} & | & \mathbf{B} \\ -- & -- & -- \\ \mathbf{0} & | & \mathbf{E} \end{bmatrix}$$

dove:

- $\mathbf{T}$: matrice quadrata sub-stocastica (con almeno una somma di riga inferiore a 1) che descrive le connessioni tra gli stati transienti;

- $\mathbf{B}$: matrice rettangolare [2] che contiene le connessioni dagli stati transienti a quelli ricorrenti;
- $\mathbf{E}$: matrice quadrata stocastica che descrive le connessioni tra gli stati ricorrenti;
- $\mathbf{0}$: matrice rettangolare [3] di zeri.

Per gli stati s transienti la corrispondente probabilità stazionaria π_s è uguale a zero.

È utile definire in forma matriciale alcune misure di interesse:

- *Probabilità di assorbimento da un qualsiasi stato transiente a un qualsiasi stato ricorrente in n passi*:

$$\mathbf{f}(n) = \mathbf{T}^{n-1}\mathbf{B}$$

- *Probabilità di assorbimento da un qualsiasi stato transiente a un qualsiasi stato in ricorrente*:

$$\mathbf{f}^{(a)} = \sum_{k=1}^{\infty} \mathbf{f}(k) = \left(\mathbf{I} + \mathbf{T} + \mathbf{T}^2 + \dots\right)\mathbf{B} = (\mathbf{I} - \mathbf{T})^{-1}\mathbf{B} = \mathbf{M}\mathbf{B}$$

  dove $\mathbf{M} = \sum_{k=0}^{\infty} \mathbf{T}^k = (\mathbf{I} - \mathbf{T})^{-1}$ è chiamata *matrice fondamentale*.

- *Numero medio di visite allo stato transiente i partendo dallo stato transiente j prima di entrare nell'insieme ricorrente*:

$$E[X_{ij}] = m_{ij}$$

  dove:

  - X_{ij} è la variabile aleatoria rappresentante il numero di visite allo stato transiente j partendo dallo stato transiente i prima di raggiungere uno degli stati ricorrenti;

[2] Eventualmente un vettore colonna se l'insieme di stati ricorrenti è composto da un unico stato assorbente.

[3] Eventualmente un vettore riga se l'insieme di stati ricorrenti è composto da un unico stato assorbente.

  - m_{ij} è l'elemento (i, j)-esimo della matrice $\mathbf{M}$.

- Se vi è un unico stato assorbente, la matrice $\mathbf{B}$ è un vettore colonna $\vec{\mathbf{B}}$; in questo caso si ottiene che:

  - l'assorbimento da ogni stato transiente è un evento certo:

$$\mathbf{f}^{(a)} = \vec{\mathbf{f}}^{(a)\,T} = \mathbf{MB} = \mathbf{M}\vec{\mathbf{B}}^{T} = \vec{\mathbf{1}}^{T}$$

  - il *tempo medio all'assorbimento* partendo da un qualsiasi stato transiente è:

$$\begin{aligned}
\mathbf{m}^{(a)} &= \vec{\mathbf{m}}^{(a)\,T} \\
&= \sum_{k=1}^{\infty} k\vec{\mathbf{f}}\,(k) \\
&= \sum_{k=1}^{\infty} k\mathbf{T}^{k-1}\vec{\mathbf{B}}^{T} \\
&= (\mathbf{I} - \mathbf{T})^{-2}\,\vec{\mathbf{B}}^{T} \\
&= (\mathbf{I} - \mathbf{T})^{-1}\,\vec{\mathbf{f}}^{(a)\,T} \\
&= (\mathbf{I} - \mathbf{T})^{-1}\,\vec{\mathbf{1}}^{T}
\end{aligned}$$

E.3 Catene di Markov a Tempo Continuo (CTMC)

In questa sezione viene descritta una seconda, non meno importante, classe di Catene di Markov chiamata *Catene di Markov a Tempo Continuo*, in cui l'insieme degli indici T è continuo; parte delle definizioni di base e della caratterizzazione degli stati verranno omesse in quanto identiche a quelle introdotte per le DTMC; saranno, quindi, presentate le equazioni di *Chapman-Kolmogorov*, che in questo caso devono essere espresse in forma differenziale; seguirà lo studio della Catena *a regime* e, infine, si descriveranno le cosidette *Catene di Markov Embedded*, un particolare tipo di Catene immerse nella struttura di una Catena di Markov a Tempo Continuo.

Definizione E.3.1 (Catena di Markov a Tempo Continuo (CTMC))**.** Una *Catena di Markov a Tempo Continuo* (*CTMC*), o semplicemente *Catena di Markov Continua*, è una Catena di Markov $\{X(t) | t \in T\}$ in cui l'insieme dei parametri T è continuo, cioè $T = [0, \infty)$.

La Fig. E.2 mostra una possibile rappresentazione grafica dell'evoluzione di una CTMC: sull'asse delle ascisse è indicato il tempo t, mentre su quello delle ordinate vi sono i possibili stati che il processo $X(t)$ può assumere; ogni punto rappresenta lo stato in cui il processo $X(t)$ si trova al tempo t.

L'analisi di una CTMC è simile a quella di una DTMC, fatta eccezione che ora le transizioni da uno stato a un altro possono avvenire in un qualunque istante di tempo. Anche in questo caso si definisce il vettore delle *probabilità di stato* $\vec{\mathbf{p}}(t)$, con $p_i(t)$ probabilità che la MC si trovi nello stato i al tempo t:

$$p_i(t) = \Pr\{X(t) = i\}$$

e la matrice delle *probabilità di transizione* $\mathbf{P}(u, v)$, con $p_{ij}(u, v)$ probabilità che vi sia una transizione nello stato j al tempo v condizionata al fatto di trovarsi nello stato i al tempo u:

$$p_{ij}(u, v) = \Pr\{X(v) = j | X(u) = i\}$$

con $u < v$.

E.3.1 Equazioni di Chapman-Kolmogorov

Proposizione E.3.1 (Equazioni di Chapman-Kolmogorov)**.** *L'evoluzione di una CTMC è descritta dalle seguenti* Equazioni di Chapman-Kolmogorov *(in forma matriciale):*

$$\begin{cases} \vec{\mathbf{p}}(v) = \vec{\mathbf{p}}(u)\,\mathbf{P}(u, v) \\ \mathbf{P}(t, v) = \mathbf{P}(t, u)\,\mathbf{P}(u, v) \\ \mathbf{P}(v, v) = \mathbf{I} \end{cases} \tag{E.3.1}$$

con $0 \leq t < u < v$*. Sotto ipotesi di* sufficiente regolarità*, le equazioni possono essere riscritte come un sistema di equazioni differenziali:*

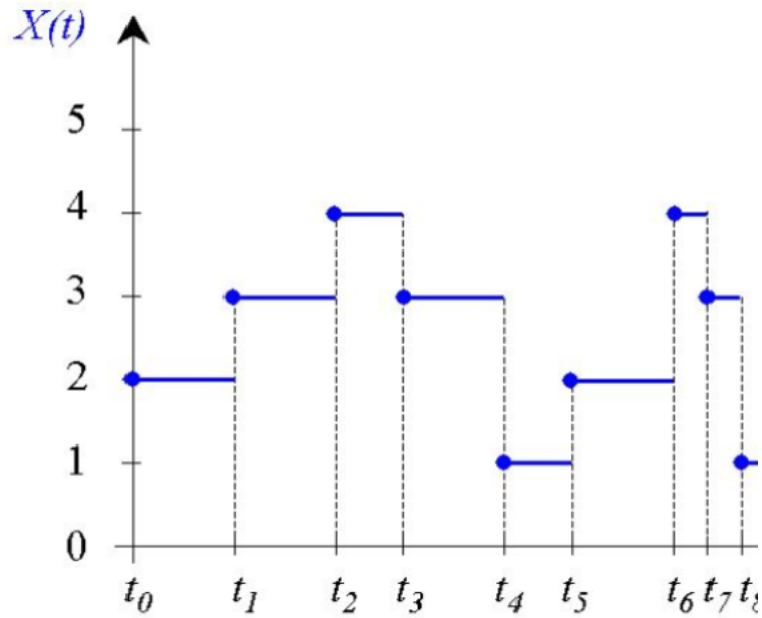


Figura E.2: Rappresentazione grafica dell'evoluzione di una CTMC.

i. Equazioni di Kolmogorov in avanti *(*Forward Kolmogorov Equations*):*

$$\begin{cases} \dfrac{\partial}{\partial v}\mathbf{P}\left(t,v\right)=\mathbf{P}\left(t,v\right)\mathbf{Q}\left(v\right) \\ \dfrac{d}{dv}\vec{\mathbf{p}}\left(v\right)=\vec{\mathbf{p}}\left(v\right)\mathbf{Q}\left(v\right) \end{cases} \tag{E.3.2}$$

ii. Equazioni di Kolmogorov all'indietro *(*Backward Kolmogorov Equations*):*

$$\begin{cases} \dfrac{\partial}{\partial t}\mathbf{P}\left(t,v\right)=\mathbf{Q}\left(t\right)\mathbf{P}\left(t,v\right) \\ \dfrac{d}{dt}\vec{\mathbf{p}}\left(t\right)=\vec{\mathbf{p}}\left(t\right)\mathbf{Q}\left(t\right) \end{cases} \tag{E.3.3}$$

dove $\mathbf{Q}\left(t\right)$ *è detta* matrice dei tassi di transizione *o* generatore infinitesimale*.*

La matrice $\mathbf{Q}\left(t\right)$ ha le seguenti caratteristiche:

- è una matrice quadrata;
- gli elementi diagonali $q_{ii}\left(v\right)$:

  $$q_{ii}\left(t\right)==-\frac{\partial}{\partial t}p_{ii}\left(u,t\right)\Big|_{u=t}=\lim_{h\to 0}\frac{1-p_{ii}\left(t,t+h\right)}{h}$$

  sono negativi e possono essere interpretati come il *tasso* con cui si verifica la transizione "uscita dallo stato i";
- gli elementi q_{ij}, con $i\neq j$, non appartenenti alla diagonale:

  $$q_{ij}\left(t\right)=\frac{\partial}{\partial t}p_{ij}\left(u,t\right)\Big|_{u=t}=\lim_{h\to 0}\frac{p_{ij}\left(t,t+h\right)}{h}$$

  sono positivi e possono essere interpretati come il *tasso* con cui si verifica la transizione "passaggio dallo stato i allo stato j";
- ogni riga di $\mathbf{Q}$ somma a zero:

  $$\mathbf{Q}\vec{\mathbf{1}}^{T}=\vec{\mathbf{0}}^{T}\quad\equiv\quad\forall i:q_{ii}=-\sum_{j,j\neq i}q_{ij}$$

- la relazione tra i tassi di transizione e le probabilità di transizione è la seguente [4]:

$$\begin{aligned} p_{ij}(t, t+h) &= q_{ij}(t)\, h + o(h), & i \neq j \\ p_{ii}(t, t+h) &= 1 - q_{ii}(t)\, h + o(h), & i = j \end{aligned}$$

Proposizione E.3.2 (Equazioni di Chapman-Kolmogorov in CTMC omogenee)**.** *Se la CTMC è* omogenea *(cioè se* $\mathbf{P}(u, v)$ *dipende solo da* $v - u$*), le* Equazioni di Chapman-Kolmogorov *diventano:*

$$\begin{cases} \dfrac{\partial}{\partial t}\mathbf{P}(t) = \mathbf{P}(t)\,\mathbf{Q} \\ \dfrac{d}{dt}\vec{\mathbf{p}}(t) = \vec{\mathbf{p}}(t)\,\mathbf{Q} \end{cases} \tag{E.3.4}$$

con condizioni iniziali:

$$\begin{cases} \mathbf{P}(0) = \mathbf{I} \\ \vec{\mathbf{p}}(0) \end{cases} \tag{E.3.5}$$

Le soluzioni del sistema di equazioni differenziali per una CTMC omogenea è:

$$\begin{aligned} \mathbf{P}(t) &= \exp(\mathbf{Q}t) \\ \vec{\mathbf{p}}(t) &= \vec{\mathbf{p}}(0)\exp(\mathbf{Q}t) \end{aligned} \tag{E.3.6}$$

dove $\exp(\mathbf{Q}t) = \mathbf{I} + \sum_{n=1}^{\infty} \mathbf{Q}^n \frac{t^n}{n!}$ *rappresenta l'operazione di esponenziale matriciale Cap. F.*

Si possono ricavare alcune misure di interesse:

- Il *tempo di permanenza* (*sojourn time*) in uno stato i è esponenzialmente distribuito:

$$p_i(t) = 1 - \exp\left(-\sum_{j, j \neq i} q_{ij} t\right) = 1 - \exp(q_{ii} t) = 1 - \exp(-q_i t) \tag{E.3.7}$$

[4] La relazione $f(x) \in o(g(x))$ si legge "$f(x)$ è o-piccolo di $g(x)$" e, intuitivamente, significa che $g(x)$ cresce più rapidamente di $f(x)$. Formalmente: $f(x) \in o(g(x)) \equiv \lim_{x \to \infty} \left| \frac{f(x)}{g(x)} \right| = 0$.

con $q_i = -q_{ii}$. Ne segue che il *tempo medio di permanenza in uno stato* i è la media di una distribuzione esponenziale di parametro q_i (*inverse scale* o *rate*).

- La *probabilità che la permanenza nello stato* i *termini con una transizione nello stato* j è data da:

$$p_{ij}(t) = \frac{q_{ij}}{q_i} \exp(-q_i t) \tag{E.3.8}$$

con $q_i = -q_{ii}$.

- Il *tempo di permanenza medio nello stato* i *nell'intervallo* $[0, t]$ è calcolabile come:

$$E[\theta_i(t)] = E\left[\int_0^t \Pr\{X(u) = i\}\, du\right] = \int_0^t p_i(u)\, du \tag{E.3.9}$$

dove $\theta(t)$ è la variabile casuale rappresentante il tempo speso dalla CTMC $X(t)$ nello stato i nell'intervallo $[0, t]$. Se si indica con $\vec{\theta}(t)$ il vettore $[E[\theta_i(t)]]$ si ha:

$$\vec{\theta}(t) = \vec{\theta}(0) + \vec{\mathbf{p}}_0 \sum_{k=0}^{\infty} \mathbf{Q}^k \frac{t^{k+1}}{(k+1)!} \tag{E.3.10}$$

Come effettuato per le DTMC, anche per le CTMC è possibile dare un'interpretazione delle equazioni di Chapman-Kolmogorov secondo il *principio di conservazione del flusso*:

$$\frac{\partial}{\partial t} p_j(t) = \sum_{i \neq j} p_i(t)\, q_{ij} - p_j(t) \sum_{i \neq j} q_{ji} \tag{E.3.11}$$

E.3.2 Classificazione degli Stati

La classificazione degli stati per una CTMC è simile a quella di una DTMC, fatta eccezione per l'assenza di stati periodici/aperiodici. Uno stato i è detto *assorbente* se $q_{ij} = 0$ per ogni $i \neq j$, cioè se, una volta che il processo vi entra, non riesce più a uscirne. Uno stato j è *raggiungibile* dallo stato i se $\exists t > 0 : p_{ij}(t) > 0$. Una CTMC è *irriducibile* se ogni suo stato è raggiungibile da un

qualsiasi altro stato. Lo spazio degli stati di una CTMC può essere partizionato in un insieme di stati transienti e in un insieme chiuso di stati ricorrenti.

E.3.3 Analisi a Regime

Teorema E.3.3. *In unna CTMC* omogenea *e* irriducibile *le condizioni di equilibrio esistono sempre e sono indipendenti dalle condizioni iniziali:*

$$\lim_{t\to\infty} p_i(t) = \lim_{t\to\infty} p_{ki}(t) = \pi_i, k, i \in \mathcal{S} \tag{E.3.12}$$

Se i limiti π_i esistono, allora:

$$\lim_{t\to\infty} \frac{d}{dt} p_i(t) = 0 \tag{E.3.13}$$

Usando questo risultato nelle equazioni di Chapman-Kolmogorov si ottiene:

Proposizione E.3.4 (Equazioni di Chapman-Kolmogorov per CTNC omogenee in Condizioni di Equilibrio)**.** *Sia $\vec{\pi}$ il vettore di stato stazionario per una CTMC omogenea, le* Equazioni di Chapman-Kolmogorov *sono date dal sistem:*

$$\begin{cases} \vec{\pi}\mathbf{Q} = 0 \\ \vec{\pi}\vec{\mathbf{1}}^T = 1 \end{cases} \tag{E.3.14}$$

Si noti che senza le condizioni iniziali $\vec{\pi}\vec{\mathbf{1}}^T = 1$, una soluzione sempre valida sarebbe $\vec{\pi} = \vec{\mathbf{0}}$ e l'esistenza di almeno un'altra soluzione $\vec{\pi}^*$, diversa da quella nulla, porterebbe ad avere infinite soluzioni del tipo $\alpha\vec{\pi}^*$, con $\alpha \in \mathbb{R}$.

La distribuzione stazionaria gode di alcune importanti proprietà:

- Per ogni probabilità iniziale $\vec{\mathbf{p}}_0$, le probabilità di stato $p_i(t)$ tendono a un valore costante π_i per $t \to \infty$ e il vettore $\vec{\pi} = [\pi_i]$ forma una distribuzione di probabilità.

- Se la probabilità iniziale è π_i, allora $p_i(t) = \pi_i$ per ogni t, cioè la catena non evolve più.

- La porzione di tempo $E[\theta_i]$ speso, in media, nello stato i nell'intervallo $[0, t]$ tende a π_i per $t \to \infty$:

$$\lim_{t\to\infty} \frac{E[\theta_i]}{t} = \pi_i$$

- Le equazioni di stato stazionario possono essere interpretate come equazioni di bilanciamento: per ogni stato i, il "flusso uscente" $\pi_i q_{ii}$ eguaglia il "flusso entrante" $\sum_{k\neq i} \pi_k q_{ki}$.

E.3.4 CTMC con stati assorbenti

In una CTMC composta da un unico stato assorbente a, il generatore infinitesimale $\mathbf{Q}(t)$ può essere così partizionato:

$$\mathbf{Q}(t) = \begin{bmatrix} \mathbf{A} & | & \vec{\mathbf{B}}^T \\ -- & -- & -- \\ \vec{\mathbf{0}} & | & 0 \end{bmatrix}$$

dove:

- $\mathbf{A}$: matrice quadrata che contiene i tassi di transizione tra gli stati transienti;
- $\vec{\mathbf{B}}^T$: vettore colonna, tale che $\vec{\mathbf{B}}^T = -\mathbf{A}\vec{\mathbf{1}}^T$, contenente i tassi di transizione dagli stati transienti a quello assorbente;
- $\vec{\mathbf{0}}$: vettore di zeri.

L'equazione di Chapman-Kolmogorov per il vettore delle probabilità di stato può essere riscritta come:

$$\frac{\partial}{\partial t}\left[\vec{\mathbf{p}}(t), p_a(t)\right]\mathbf{Q}(t) \tag{E.3.15}$$

che esplicitando diventa:

$$\begin{cases} \dfrac{\partial}{\partial t}\vec{\mathbf{p}}\left(t\right) = \vec{\mathbf{p}}\left(t\right)\mathbf{A} \\ \dfrac{\partial}{\partial t}p_a\left(t\right) = \vec{\mathbf{p}}\left(t\right)\vec{\mathbf{B}}^T \end{cases} \tag{E.3.16}$$

La soluzione di questo sistema è:

$$\begin{cases} \vec{\mathbf{p}}\left(t\right) = \vec{\mathbf{p}}_0 \exp\left(\mathbf{A}t\right) \\ p_a\left(t\right) = \vec{\mathbf{p}}_0 \exp\left(\mathbf{A}t\right)\vec{\mathbf{B}}^T \end{cases} \tag{E.3.17}$$

In molti casi risulta particolarmente utile ricavare delle informazioni riguardo Il *tempo all'assorbimento* τ_a.

Proposizione E.3.5 (Tempo all'assorbimento)**.** *Sia τ_a la variabile casuale rappresentante il* tempo all'assorbimento*, cioè il tempo necessario alla CTMC $X\left(t\right)$ omogenea a raggiungere lo stato di assorbimento a a partire da $t = 0$; la distribuzione di τ_a è data da:*

$$F_{\tau_a}\left(t\right) = \Pr\left\{\tau_a \leq t\right\} = \Pr\left\{X\left(t\right) = a\right\} = p_a\left(t\right) = 1 - \vec{\mathbf{p}}\left(t\right)\vec{\mathbf{1}}^T = 1 - \vec{\mathbf{p}}_0 \exp\left(\mathbf{A}t\right)\vec{\mathbf{1}}^T \tag{E.3.18}$$

e il valore medio $E\left[\tau_a\right]$ vale:

$$E\left[\tau_a\right] = \vec{\mathbf{p}}_0\left(-\mathbf{A}\right)^{-1}\vec{\mathbf{1}}^T \tag{E.3.19}$$

E.3.5 Catene di Markov Embedded (EMC)

Le *Catene di Markov Embedded* giocano un ruolo molto importante per l'analisi di processi stocastici sia di tipo Markoviano sia di tipo non-Markoviano. Prima di darne una definizione formale ne viene fornito un semplice esempio.

Esempio E.3.1 (Catena di Markov embedded in una Coda M/M/1)**.** Si consideri un coda *M/M/1* [109], cioè un sistema a server singolo con tempi di interarrivo e di servizio distribuiti secondo una distribuzione esponenziale negativa. Sia $N(t)$ il numero di job nel sistema (sia in coda sia in servizio) al tempo t; il

processo $\{N(t)|t \geq 0\}$ è un Processo Nascita-Morte (si veda l'esempio E.1.6). Il sistema può essere analizzato nel seguente modo: si considerino gli istanti di tempo in cui i job lasciano il sistema; questi punti temporali, anche chiamati *punti di rigenerazione*, possono essere utilizzati come insieme degli indici di un nuovo processo stocastico definito nella seguente maniera: sia t_n, con $n \in= 1, 2, \ldots$, l'istante di partenza dell'n-esimo job dal sistema (immediatamente dopo la fine del servizio), e sia X_n il numero di job nel sistema al tempo t_n. Ne segue che:

$$X_n = N(t_n), \quad n = 1, 2, \ldots$$

Il processo stocastico $\{X_n|n = 1, 2, \ldots\}$ è una DTMC omogenea chiamata *Catena di Markov Embedded della CTMC* $\{N(t)|t \geq 0\}$.

In generale si può dimostrare che ad ogni CTMC è possibile associare una Catena di Markov Embedded. Un metodo per trovare la distribuzione stazionaria di una CTMC è quello di risolvere prima la Catena di Markov Embedded ad essa associata e, a partire dalle probabilità di transizione di quest'ultima, trovare le probabilità di stato stazionario.

Definizione E.3.2 (Catena di Markov Embedded (EMC))**.** Data una CTMC $\{X(t)|t \geq 0\}$ con generatore infinitesimale $\mathbf{Q} = [q_{ij}]$, si considerino solo gli istanti di tempo θ_n, con $n = 0, 1, 2, \ldots$, in cui avviene un cambiamento di stato nel sistema, e le variabili Y_n, con $n = 0, 1, 2, \ldots$, che assumono valori in $[\theta_n, \theta_{n+1})$; il processo stocastico $\{Y_n|n = 0, 1, 2, \ldots\}$ è una DTMC chiamata *Catena di Markov Embedded* (*EMC*) della CTMC $\{X(t)|t \geq 0\}$.

La Fig. E.3 mostra una possibile CTMC $\{X(t)|t \geq 0\}$ con associata la relativa EMC $\{Y_n|n = 0, 1, \ldots\}$. Le probabilità di transizione v_{ij} della EMC sono date da

$$\begin{aligned} v_{ij} &= \Pr\{Y_{n+1} = j|Y_n = i\} \\ &= \Pr\{X(\theta_{n+1}) = j|X(\theta_n) = i\}, \quad \forall i, j \in \mathcal{S}, i \neq j \\ v_{ii} &= 0, \quad \forall i \in \mathcal{S} \end{aligned}$$

con θ_n rappresentante l'n-esimo punto di rigenerazione; per le proprietà di una

CTMC è possibile scrivere:

$$\Pr\{Y_{n+1} = j, (\theta_{n+1} - \theta_n) > \tau | Y_n = i\} = v_{ij} \exp(-q_i \tau), \quad \forall i, j \in \mathcal{S}, \tau \geq 0$$

dove $q_i = -q_{ii}$ e $\theta_{n+1} - \theta_n$ è il tempo di permanenza nello stato j dopo che la catena è stata nello stato i. È chiaro quindi che le probabilità di transizione v_{ij} della EMC possono essere ottenute a partire dai tassi di transizione q_{ij} della CTMC. Inoltre, se la CTMC è ergodica, si può utilizzare la distribuzione limite e scrivere:

$$v_{ij} = \begin{cases} -\frac{q_{ij}}{q_i} & i \neq j \\ 0 & i = j \end{cases}$$

Per come è stata definita la EMC, i relativi tempi di permanenza nei suoi stati sono pari alla costante 1.

È possibile derivare le probabilità a regime $\pi^{(X)}$ di una CTMC $\{X(t) | t \geq 0\}$ ergodica a partire dalla distribuzione stazionaria $\pi^{(Y)}$ della relativa EMC $\{Y_n | n = 0, 1, 2, \dots\}$ [5]; la distribuzione a regime $\pi^{(X)}$ del processo $X(t)$ pàri a:

$$\pi_i^{(X)} = \lim_{t \to \infty} \Pr\{X(t) = i\}, \quad \forall i \in \mathcal{S}$$

mentre la distribuzione stazionaria $\pi^{(Y)}$ del processo Y_n si ottiene risolvendo il seguente sistema:

$$\begin{cases} \pi_j^{(Y)} = \sum_{i \in \mathcal{S}} \pi_i^{(Y)} v_{ij}, \quad \forall j \in \mathcal{S} \\ \sum_{i \in \mathcal{S}} \pi_i^{(Y)} = 1 \end{cases} \tag{E.3.20}$$

[5] Si noti che se una CTMC $\{X(t) | t \geq 0\}$ è ergodica, la corrispondente EMC $\{Y_n | n = 0, 1, 2, \dots\}$ è irriducibile e ricorrente; non è detto però che sia aperiodica e quindi ergodica.

Tra le due distribuzioni vi è la seguente relazione:

$$\begin{aligned} \pi_i^{(X)} &= \frac{\pi_i^{(Y)}/q_i}{\sum_{j\in\mathcal{S}} \pi_j^{(Y)}/q_j} \\ \pi_i^{(Y)} &= \frac{\pi_i^{(X)} q_i}{\sum_{j\in\mathcal{S}} \pi_j^{(X)} q_j} \end{aligned}, \quad \forall i \in \mathcal{S} \tag{E.3.21}$$

Una EMC rappresenta solo le probabilità di transizione da uno stato all'altro e trascura l'informazione relativa alla frequenza di transizioni, ai tempi medi di permanenza negli stati, ...Quindi si può affermare che una CTMC è caratterizzata dalla relativa EMC (che definisce il "dove muoversi") e dai tempi di permanenza esponenziali (che descrivono il "quando muoversi").

E.4 Processi Semi-Markoviani (SMP)

I processi Markoviani trovano applicazione nello studio delle code *M/M/**; queste code hanno il vantaggio di avere tempi di interarrivo e di servizio in cui vale la proprietà di assenza di memoria. Nei sistemi reali, tuttavia, si incontrano spesso situazioni in cui non è possibile assumere la validità di tale proprietà (ad es. i tempi di servizio non sono distribuiti secondo una distribuzione esponenziale). In questi casi si può descrivere la grandezza non distribuita in modo esponenziale come una composizione di un certo numero di stadi esponenziali in serie o in parallelo (ad esempio, si veda Cap. 5) oppure descrivere il sistema in specifici punti temporali (*punti di rigenerazione*). Quest'ultimo caso è l'oggetto di questa sezione.

Un *Processo Semi-Markoviano* (*SMP*) è una generalizzazione del Processo di Markov e del Processo di Rinnovo [109]. Come per le CTMC §E.3, a un SMP è possibile associare una EMC §E.3.5.

Definizione E.4.1 (Processo Semi-Markoviano (SMP))**.** In riferimento alla Fig. E.3, si consideri un processo stocastico $\{X(t) | t \geq 0\}$ omogeneo a tempo continuo; si indichino gli istanti di cambiamento di stato con θ_n, dove $n = 0, 1, 2, \ldots$, e si supponga che le variabili Y_n assumano valori nell'intervallo $[\theta_n, \theta_{n+1})$; si

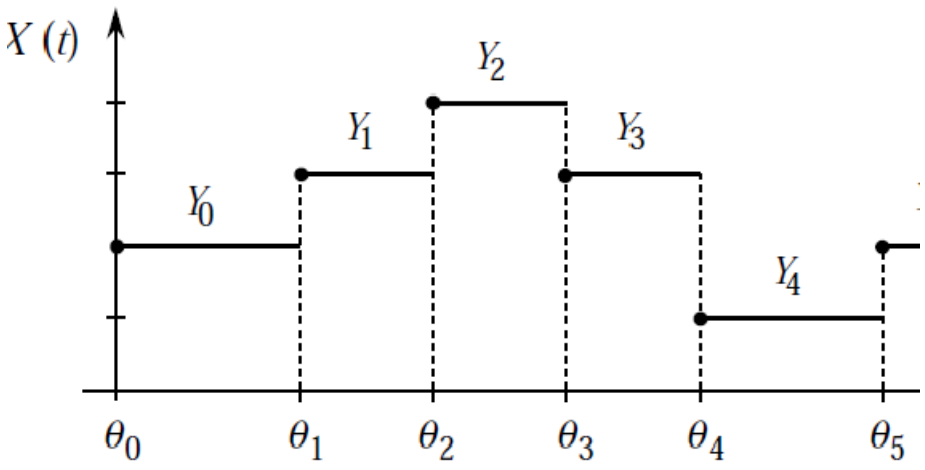


Figura E.3: Esempio di CTMC con la relativa EMC.

supponga inoltre che, $\forall i, j \in \mathcal{S}$:

$$\begin{aligned} k_{ij}(\tau) &= \Pr\{Y_{n+1} = j, (\theta_{n+1} - \theta_n) \leq \tau | Y_n = i, \ldots, Y_0 = k, \theta_n = t_n, \ldots, \theta_0 = t_0\} \\ &= \Pr\{Y_{n+1} = j, (\theta_{n+1} - \theta_n) \leq \tau | Y_n = i\} \\ p_{ij} &= \Pr\{Y_{n+1} = j | Y_n = i\} \\ &= \lim_{\tau \to \infty} k_{ij}(\tau) \end{aligned}$$

Si noti che i tempi di soggiorno possono avere una distribuzione arbitraria (quindi non necessariamente esponenziale) e le probabilità possono dipendere sia dallo stato corrente sia da quello futuro. Un processo stocastico aventi le suddette caratteristiche è un *Processo Semi-Markoviano* (*SMP*).

Il processo stocastico $\{Y_n | n = 0, 1, 2, \ldots\}$ rappresenta una DTMC §E.2 e viene detto *Catena di Markov Embedded* (*EMC*) del processo stocastico $\{X(t) | t \geq 0\}$ §E.3.5.

In modo analogo a quanto è stato fatto per le MC, è possibile definire le Catene Semi-Markoviane a partire dai Processi Semi-Markoviani.

Definizione E.4.2 (Catena Semi-Markoviana (SMC))**.** Una *Catena Semi-Markoviana* (*SMC*) è un Processo Semi-Markoviano $\{X(t) | t \in T\}$ a stati discreti (cioè con spazio degli stati finito o infinitamente enumerabile).

Nelle SMC, il tasso di transizione da uno stato i a uno stato j può dipendere dal tempo già speso nello stato i (dall'ultima visita) ma non dagli stati visitati prima di entrare nello stato i e nemmeno dai precedenti tempi di soggiorno. La matrice di transizione $\mathbf{K}(t)$ diventa quindi dipendente dal tempo e prende il nome di *kernel di una SMC*; ogni elemento $k_{ij}(t)$ rappresenta la probabilità che la transizione dallo stato i allo j avvenga in al più t unità di tempo. Durante i cambianmenti di stato la SMC si comporta come una DTMC; perciò le probabilità di stato stazionario possono essere ricavate a partire dalle probabilità di stato relative alla corrispondente EMC.

Di seguito viene proposto un esempio in cui si utilizza una SMC per modellare il comportamento di un sistema.

Esempio E.4.1 (Catena Semi-Markoviana in una coda M/G/1)**.** Si consideri una coda *M/G/1* [109], cioè un sistema a server singolo con tempi di interarrivo distribuiti secondo una distribuzione esponenziale negativa e tempi di servizio distribuiti secondo una generica distribuzione di probabilità. Sia $N(t)$ il numero di job nel sistema (sia in coda sia in servizio) al tempo t; se $N(t) \geq 1$, significa che c'è un job in servizio e, dato che in generale la distribuzione dei tempi di servizio non gode della proprietà di assenza di memoria, il comportamento futuro del sistema dipende sia da $N(t)$ sia dal tempo speso da un job in servizio. Ne consegue che il processo stocastico $\{N(t) | t \geq 0\}$ non è un processo Markoviano. Tuttavia, se si considerano gli istanti di tempo t_n in cui un job esce dal sistema (*punti di rigenerazione*), si individua un nuovo processo stocastico $\{X_n | n = 1, 2, \ldots\}$, dove X_n rappresenta il numero di job nel sistema al tempo t_n, cioè il numero di job dopo che l'n-esimo job ha lasciato il sistema (immediatamente dopo il servizio); ne risulta quindi che $X_n = N(t_n)$ con $n = 1, 2, \ldots$ Il processo stocastico $\{N(t) | t \geq 0\}$ è una SMC e il processo $\{X_n | n = 1, 2, \ldots\}$ rappresenta la relativa EMC.

Appendice F

Esponenziale di una Matrice

L'operazione di esponenziale di una matrice compare in molte teorie scientifiche; ad esempio, la soluzione di una CTMC §E.3 con generatore infinitesimale $\mathbf{Q}$ prevede il calcolo dell'esponenziale di $\mathbf{Q}$; oppure, la distribuzione di probabilità di una PH$(\vec{\alpha}, \mathbf{T})$ Cap. 5 implica il calcolo dell'esponenziale del generatore $\mathbf{T}$.

In questo capitolo viene fornita una panoramica sull'operazione di esponenziale di una matrice e vengono proposti alcuni metodi numerici per effettuarne il calcolo.

F.1 Definizione e Alcune Proprietà

Definizione F.1.1 (Esponenziale di una Matrice)**.** Data una matrice quadrata $\mathbf{A} \in \mathbb{C}^{n \times n}$ di ordine n, l'esponenziale di $\mathbf{A}$ è una matrice quadrata di ordine n data dalla serie di potenze:

$$\exp(A) = e^A = \sum_{k=0}^{\infty} \frac{A^k}{k!} \tag{F.1.1}$$

La serie F.1.1 è ricavata direttamente dallo sviluppo in serie di Taylor della funzione $\exp(x)$ e converge sempre per qualunque matrice $\mathbf{A}$ quadrata di ordine n.

Di seguito vengono mostrate alcune delle proprietà che l'esponenziale di una matrice gode:

- $\exp(\mathbf{0}) = \mathbf{I}$;
- $\exp(c\mathbf{A})\exp(d\mathbf{A}) = \exp[(c+d)\,\mathbf{A}]$, con $c, d \in \mathbb{R}$;
- $\exp(\mathbf{A})\exp(-\mathbf{A}) = \mathbf{I}$ (l'esponenziale di una matrice è sempre invertibile e l'inversa è data da $\exp(-\mathbf{A})$);
- Se $\mathbf{AB} = \mathbf{BA}$, allora $\exp(\mathbf{A})\exp(\mathbf{B}) = \exp(\mathbf{A}+\mathbf{B})$;
- Se $\mathbf{B}$ è invertibile allora $\exp(\mathbf{BAB}^{-1}) = \mathbf{B}\exp(\mathbf{A})\,\mathbf{B}^{-1}$.

F.2 Metodi Numerici

In questa sezione vengono esposti alcuni metodi numerici utilizzabili per il calcolo dell'esponenziale di una matrice, dando maggiore enfasi a quelli utilizzati per effettuare gli esperimenti descritti nel presente documento; per un elenco più completo si veda [79].

F.2.1 Matrici Diagonali

Se la matrice $\mathbf{A}$ è diagonale, il relativo esponenziale può essere ottenuto applicando l'esponenziale a ogni elemento della diagonale, cioè se

$$\mathbf{A} = \operatorname{diag}(a_{11}, a_{22}, \ldots, a_{nn})$$

allora:

$$\exp[\mathbf{A}] = \exp[\operatorname{diag}(a_{11}, \ldots, a_{nn})] = \operatorname{diag}[\exp(a_{11}), \ldots, \exp(a_{nn})] \qquad \text{(F.2.1)}$$

F.2.2 Matrici Nilpotenti

Definizione F.2.1 (Matrice Nilpotente)**.** Una matrice $\mathbf{A}$ è *nilpotente* se esiste un intero positivo k tale per cui $\mathbf{A}^k = \mathbf{0}$. Si può dimostrare che una matrice $\mathbf{A}$ è *nilpotente* se e solo se tutti i suoi autovalori sono nulli.

Nel caso di matrici nilpotenti, l'esponenziale può essere calcolato espandendo direttamente la serie F.1.1, troncandola al termine k-esimo (in quanto il termine $(k+1)$-esimo implica il calcolo di $\mathbf{A}^k$ che si sa essere uguale a $\mathbf{0}$):

$$\exp(\mathbf{A}) = \mathbf{I} + \mathbf{A} + \frac{1}{2}\mathbf{A}^2 + \frac{1}{6}\mathbf{A}^3 + \cdots + \frac{1}{(k-1)!}\mathbf{A}^{k-1}. \tag{F.2.2}$$

F.2.3 Caso Generale

I metodi proposti di seguito sono quelli che sono stati utilizzati durante lo svolgimento degli esperimenti. La tendenza attuale (si veda ad es. [79, 55]) sembra dare la preferenza, sia in termini di efficienza sia in fatto di affidabilità, verso la cosidetta tecnica dello *Scaling e Squaring*, in particolare usata in congiunzione con l'*approssimazione di Padé*; anche nel lavoro descritto nel presente documento si è scelto di utilizzare l'unione di questi due metodi.

Serie di Taylor

La tecnica diretta consiste nell'espandere la serie di Taylor F.1.1 troncandola a un determinato termine n-esimo:

$$\exp(\mathbf{A}) \approx \mathbf{I} + \mathbf{A} + \frac{1}{2}\mathbf{A}^2 + \frac{1}{6}\mathbf{A}^3 + \cdots + \frac{1}{(n-1)!}\mathbf{A}^{n-1}. \tag{F.2.3}$$

Questo metodo oltre a essere poco efficiente potrebbe soffrire di problemi di stabilità numerica (ad es. errori di cancellazione).

Approssimazione di Padé

Definizione F.2.2 (Approssimazione di Padé)**.** L'*approssimazione di Padé* $[p/q]$ per una serie di potenze $A(x)$:

$$A(x) = \sum_{j=0}^{\infty} a_j x^j$$

è data dalla funzione razionale:

$$[p/q]_A(x) = R_{p/q}(x) = \frac{N_{pq}(x)}{D_{pq}(x)} = \frac{n_0 + n_1 x + n_2 x^2 + \cdots + n_p x^p}{1 + d_0 + d_1 x + d_2 x^2 + \cdots + d_q x^q} \tag{F.2.4}$$

dove:

$$N_{pq}(x) = \sum_{j=0}^{p} \frac{(p+q-j)!p!}{(p+q)!\,(p-j)!j!} x^j$$
$$D_{pq}(x) = \sum_{j=0}^{q} \frac{(p+q-j)!q!}{(p+q)!\,(q-j)!j!} (-x)^j$$

Applicando la definizione di approssimazione di Padé $[p/q]$ alla funzione esponenziale di matrici $\exp(\mathbf{A})$ si ottiene:

$$R_{p/q}(\mathbf{A}) = [D_{pq}(\mathbf{A})]^{-1} N_{pq}(\mathbf{A}) \tag{F.2.5}$$

dove:

$$N_{pq}(\mathbf{A}) = \sum_{j=0}^{p} \frac{(p+q-j)!p!}{(p+q)!j!\,(p-j)!} \mathbf{A}^j$$
$$D_{pq}(\mathbf{A}) = \sum_{j=0}^{q} \frac{(p+q-j)!q!}{(p+q)!j!\,(q-j)!} (-\mathbf{A})^j$$

Così definito, questo metodo, oltre a soffrire di errori di cancellazione, potrebbe generare errori di arrotondamento. Tuttavia, utilizzando la tecnica descritta nel prossimo paragrafo si possono ottenere dei buoni risultati.

Scaling e Squaring

La cosidetta tecnica della *Scalatura ed Elevamento al quadrato* (*Scaling e Squaring*) [112, 79], quando applicata a uno dei precedenti metodi (Eq. F.2.3 e F.2.5), può far aumentare l'affidabilità e l'efficienza del metodo in questione; ad es., se utilizzata insieme all'approssimazione di Padé, si ottiene un metodo particolarmente efficiente, stabile e, in generale, migliore del metodo di Taylor . Tale tecnica consiste nello sfruttare la proprietà dell'esponenziale di una matrice:

$$\exp(\mathbf{A}) = \exp(\mathbf{A}/m)^m, \quad m \in \mathbb{R} \tag{F.2.6}$$

scegliendo come m una potenza di due in modo che $\exp(\mathbf{A}/m)$ (l'esponenziale della matrice *scaled*) sia affidabile ed efficiente da calcolare [1]; il calcolo dell'esponenziale della matrice *scaled* può essere effettuato, ad esempio, con uno dei precedenti metodi (Eq. F.2.3 e F.2.5). La matrice $\exp(\mathbf{A}/m)^m$ può essere quindi calcolate attraverso degli elevamenti al quadrato ripetuti.

Autovettori

Un'approccio alternativo a quelli precedenti è quello di applicare il metodo di *trasformazione per similitudine* [98] e considerare la scomposizione della matrice $\mathbf{A}$ in:

$$\mathbf{A} = \mathbf{SBS}^{-1}$$

anche chiamata, scherzosamente, *rappresentazione a sandwitch di* $\mathbf{A}$, e quindi quella della matrice $\exp(\mathbf{A})$ in:

$$\exp(\mathbf{A}) = \mathbf{S}\exp(\mathbf{B})\mathbf{S}^{-1}$$

in modo tale che le matrici $\mathbf{S}$ e $\mathbf{B}$ (e $\exp(\mathbf{B})$) siano più semplici da calcolare. Un modo per ottenere questo tipo di scomposizione è il *metodo QR* [98].

Nel metodo degli *autovettori*, l'idea è quella di scegliere come matrice $\mathbf{S}$, la

[1] Un buon criterio è che $\|\mathbf{A}\|/m \leq 1$.

matrice le cui colonne sono autovettori di $\mathbf{A}$:

$$\mathbf{S} = \left[\vec{\mathbf{v}}_1^T \cdots \vec{\mathbf{v}}_n^T\right]$$

tale che:

$$\mathbf{A}\vec{\mathbf{v}}_j^T = \lambda_j \vec{\mathbf{v}}_j^T, \quad j = 1, 2, \ldots, n$$

con λ_j autovalore di $\mathbf{A}$ e $\vec{\mathbf{v}}_j$ un corrispondente autovettore. Indicando con $D = \operatorname{diag}\{\lambda_1, \ldots, \lambda_n\}$, il sistema di n equazioni può essere riscritto come:

$$\mathbf{AS} = \mathbf{SD}$$

Inoltre per le proprietà dell'esponenziale di una matrice diagonale:

$$\exp\left(\operatorname{diag}\{d_{11}, \ldots, d_{nn}\}\right) = \operatorname{diag}\{\exp(d_{11}), \ldots, \exp(d_{nn})\}$$

e per il fatto che $\mathbf{S}$ sia invertibile, risulta infine che:

$$\exp(\mathbf{A}) = \exp\left(\mathbf{SDS}^{-1}\right) = \mathbf{S}\exp(\mathbf{D})\,\mathbf{S}^{-1} = \mathbf{S}\operatorname{diag}\{\exp(\lambda_1), \ldots, \exp(\lambda_n)\}\,\mathbf{S}^{-1} \tag{F.2.7}$$

Questo metodo risulta molto accurato ed efficiente per matrici simmetriche, ortogonali [2] e altri tipi di matrici normali [3]; tuttavia, al crescere del numero di condizionamento [4] l'accuratezza diminuisce fino a dare risultati pessimi in caso di matrici difettive [5].

[2] Una matrice quadrata $\mathbf{A}$ è *ortogonale* se $\mathbf{AA}^T = \mathbf{A}^T\mathbf{A} = \mathbf{I}$.

[3] Una matrice quadrata $\mathbf{A} \in \mathbb{C}^{n\times n}$ è *normale* se $\mathbf{AA}^H = \mathbf{A}^H\mathbf{A}$, dove $\mathbf{A}^H$ è la coniugata transposta di $\mathbf{A}$.

[4] Il *numero di condizionamento* di un problema dà un'indicazione su quanto delle piccole perturbazioni sui dati incidano sulle soluzioni del problema; maggiore è il numero di condizionamento e più il problema è *mal condizionato*.

[5] Una matrice quadrata di ordine n è *difettiva* se e solo se non possiede una base completa di autovettori, cioè se non ha n autovettori linearmente indipendenti.

Appendice G

Composizioni Matriciali di Kronecker

G.1 Prodotto di Kronecker

G.1.1 Definizione

Definizione G.1.1 (Prodotto di Kronecker)**.** Siano $\mathbf{A} = [a_{ij}]$ e $\mathbf{B} = [b_{ij}]$ due matrici di dimensione $n_1 \times m_1$ e $n_2 \times m_2$, rispettivamente. Si definisce *prodotto di Kronecker* (o *prodotto matriciale diretto*) tra $\mathbf{A}$ e $\mathbf{B}$:

$$\mathbf{K} = \mathbf{A} \otimes \mathbf{B}$$

una matrice a blocchi [1] di dimensione $(n_1 n_2) \times (m_1 m_2)$ definita nel seguente modo:

$$\mathbf{K} = \mathbf{A} \otimes \mathbf{B} = \begin{bmatrix} a_{11}\mathbf{B} & a_{12}\mathbf{B} & \cdots & a_{1m_1}\mathbf{B} \\ \vdots & \vdots & & \vdots \\ a_{n_1 1}\mathbf{B} & a_{n_1 2}\mathbf{B} & \cdots & a_{n_1 m_1}\mathbf{B} \end{bmatrix} \tag{G.1.1}$$

[1] Una matrice a *blocchi* è una matrice definita a partire da matrici più piccole chiamate *blocchi*.

cioè tale per cui:

$$k_{\alpha\beta} = a_{ij} b_{hl}$$

dove:

$$\alpha = n_1 (i-1) + h$$
$$\beta = m_1 (j-1) + l$$

G.1.2 Proprietà

Di seguito vengono presentante alcune delle proprietà che il prodotto di Kronecker possiede:

$$\mathbf{A} \otimes \mathbf{B} \otimes \mathbf{C} = \mathbf{A} \otimes (\mathbf{B} \otimes \mathbf{C}) = (\mathbf{A} \otimes \mathbf{B}) \otimes \mathbf{C} \quad \text{(Associativa)} \tag{G.1.2}$$
$$\mathbf{A} \otimes (\mathbf{B} + \mathbf{C}) = \mathbf{A} \otimes \mathbf{B} + \mathbf{A} \otimes \mathbf{C} \quad \text{(Distributiva)} \tag{G.1.3}$$
$$\mathbf{A} \otimes \mathbf{B} \neq \mathbf{B} \otimes \mathbf{A} \quad \text{(Non Commutativa)} \tag{G.1.4}$$
$$(c_1 \mathbf{A}) \otimes (c_2 \mathbf{B}) = c_1 c_2 (\mathbf{A} \otimes \mathbf{B}), \quad c_1, c_2 \in \mathbb{R} \tag{G.1.5}$$
$$(\mathbf{A} \otimes \mathbf{B})(\mathbf{C} \otimes \mathbf{D}) = (\mathbf{AC}) \otimes (\mathbf{BD}) \tag{G.1.6}$$
$$(\mathbf{A} \otimes \mathbf{B})^T = \mathbf{A}^T \otimes \mathbf{B}^T \tag{G.1.7}$$
$$(\mathbf{A} \otimes \mathbf{B})^{-1} = \mathbf{A}^{-1} \otimes \mathbf{B}^{-1} \tag{G.1.8}$$

G.2 Somma di Kronecker

G.2.1 Definizione

Definizione G.2.1 (Somma di Kronecker)**.** Siano $\mathbf{A} = [a_{ij}]$ e $\mathbf{B} = [b_{ij}]$ due matrici quadrate di ordine n e m, rispettivamente. Si definisce *somma di Kronecker* (o *somma matriciale diretta*) tra $\mathbf{A}$ e $\mathbf{B}$:

$$\mathbf{K} = \mathbf{A} \oplus \mathbf{B}$$

una matrice quadrata a blocchi di ordine nm definita nel seguente modo:

$$\mathbf{K} = \mathbf{A} \oplus \mathbf{B} = \mathbf{A} \otimes \mathbf{I}_m + \mathbf{I}_n \otimes \mathbf{B} \tag{G.2.1}$$

G.3 Esempi

Esempio G.3.1. Si considerino le due matrici:

$$\mathbf{A} = \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}$$

$$\mathbf{B} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

I prodotti di Kronecker $\mathbf{A} \otimes \mathbf{B}$ e $\mathbf{B} \otimes \mathbf{A}$ sono dati da:

$$\mathbf{A} \otimes \mathbf{B} = \left[\begin{array}{ccc|ccc} 2 & 0 & 0 & -1 & -0 & -0 \\ 0 & 2 & 0 & -0 & -1 & -0 \\ 0 & 0 & 2 & -0 & -0 & -1 \\ \hline -1 & -0 & -0 & 2 & 0 & 0 \\ -0 & -1 & -0 & 0 & 2 & 0 \\ -0 & -0 & -1 & 0 & 0 & 2 \end{array}\right]$$

$$\mathbf{B} \otimes \mathbf{A} = \left[\begin{array}{cc|cc|cc} 2 & -1 & 0 & 0 & 0 & 0 \\ -1 & 2 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 2 & -1 & 0 & 0 \\ 0 & 0 & -1 & 2 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 2 & -1 \\ 0 & 0 & 0 & 0 & -1 & 2 \end{array}\right]$$

Esempio G.3.2. Si considerino le due matrici:

$$\mathbf{A} = \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}$$

$$\mathbf{B} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Le somme di Kronecker $\mathbf{A} \oplus \mathbf{B}$ e $\mathbf{B} \oplus \mathbf{A}$ sono date da:

$$\mathbf{A} \oplus \mathbf{B} = \left[\begin{array}{ccc|ccc} 3 & 0 & 0 & -1 & 0 & 0 \\ 0 & 3 & 0 & 0 & -1 & 0 \\ 0 & 0 & 3 & 0 & 0 & -1 \\ \hline -1 & 0 & 0 & 3 & 0 & 0 \\ 0 & -1 & 0 & 0 & 3 & 0 \\ 0 & 0 & -1 & 0 & 0 & 3 \end{array}\right]$$

$$\mathbf{B} \oplus \mathbf{A} = \left[\begin{array}{cc|cc|cc} 3 & -1 & 0 & 0 & 0 & 0 \\ -1 & 3 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 3 & -1 & 0 & 0 \\ 0 & 0 & -1 & 3 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 3 & -1 \\ 0 & 0 & 0 & 0 & -1 & 3 \end{array}\right]$$

Bibliografia

[1] Large Hadron Collider. http://public.web.cern.ch/.

[2] The LCG Real Time Monitor. http://gridportal.hep.ph.ic.ac.uk/rtm/.

[3] TeraGrid. http://www.teragrid.org/.

[4] The worldwide LHC Computing Grid project. http://lcg.web.cern.ch/LCG/.

[5] Robert J. Adler, Raisa E. Feldman, and Murad S. Taqqu. *A Practical Guide to Heavy Tails. Statistical Techniques and Applications*. Birkhāuser, 1998.

[6] J. H. Ahrens and U. Dieter. Computer methods for sampling from the exponential and normal distributions. *Communications of the ACM*, 15(10):873–882, 1972.

[7] J. H. Ahrens and U. Dieter. Generating gamma variates by a modified rejection technique. *Communications of the ACM*, 25(1):47–54, 1982.

[8] T. W. Anderson. On the distribution of the two-sample cramer-von mises criterion. *The Annals of Mathematical Statistics*, 33(3):1148–1159, 1962.

[9] T. W. Anderson and D. A. Darling. Asymptotic theory of certain "goodness of fit" criteria based on stochastic processes. *The Annals of Mathematical Statistics*, 23(2):193–212, 1952.

[10] T. W. Anderson and D. A. Darling. A test of goodness of fit. *Journal of the American Statistical Association*, 49(268):765–769, 1954.

[11] Martin Arlitt and Tai Jin. Workload characterization of the 1998 world cup web site. 1999.

[12] Søren Asmussen, Olle Nerman, and Marita Olsson. Fitting phase-type distributions via the em algorithm. *Scand. J. Stat.*, 23(4):419–441, 1996.

[13] Gutti J. Babu and Radhakrishna C. Rao. Bootstrap methodology. In *Handbook of Statistics*, volume 9, pages 627–660. Elsevier Science Publishers, The Netherlands, Amsterdam, 1993.

[14] Gutti J. Babu and Radhakrishna C. Rao. Goodness-of-fit tests when parameters are estimated. *Sankhya*, 66(1):63–74, 2004.

[15] D. J. Best and D. E. Roberts. Algorithm as 91: The percentage points of the χ^2 distribution. *Applied Statistics*, 24(3):385–388, 1975.

[16] Andrea Bobbio. Appunti delle lezioni di "Modelli Quantitativi", 1997.

[17] Andrea Bobbio, András Horváth, and Miklós Telek. Matching three moments with minimal acyclic phase type distributions. *Stochastic Models*, 21:303–326, 2005.

[18] Peter Buchholz, Gianfranco Ciardo, Susanna Donatelli, and Peter Kemper. Complexity of memory-efficient kronecker operations with applications to the solution of markov models. *INFORMS Journal of Computing*, 12(3), 2000.

[19] Maria Calzarossa, Luisa Massari, and Daniele Tessera. Workload characterization issues and methodologies. In *Performance Evaluation: Origins and Directions*, pages 459–481, London, UK, 2000. Springer-Verlag.

[20] Henri Casanova, James Hayes, and Yang Yang. Algorithms and software to schedule and deploy independent tasks in grid environments. In *Workshop on Distributed Computing, Metacomputing and Resource Globalization*, Aussois, France, Dec 2002.

[21] Henri Casanova, Arnaud Legrand, Dmitrii Zagorodnov, and Francine Berman. Heuristics for scheduling parameter sweep applications in grid environments. In *Heterogeneous Computing Workshop*, pages 349–363, Cancun, Mexico, May 2000.

[22] Enrique Castillo. *Extreme Value Theory in Engineering*. Academic Press Inc., San Diego, 1988.

[23] Herman Chernoff and E. L. Lehmann. The use of maximum likelihood estimates in χ^2 tests for goodness of fit. *The Annals of Mathematical Statistics*, 25(3):579–586, 1954.

[24] William G. Cochran. Some methods for strengthening the common χ^2 tests. *Biometrics*, 10(4):417–451, 1954.

[25] David R. Cox. A use of complex probabilities in the theory of stochastic processes. volume 51, pages 313–319, 1955.

[26] Mark E. Crovella and Lester Lipsky. Long-lasting transient conditions in simulations with heavy-tailed workloads. In *Winter Simulation Conference*, pages 1005–1012, 1997.

[27] Mark E. Crovella and Murad S. Taqqu. Estimating the heavy tail index from scaling properties. *Methodology and Computing in Applied Probability*, 1(1), 1999.

[28] Mark E. Crovella, Murad S. Taqqu, and Azer Bestavros. Heavy-tailed probability distributions in the world wide web. In Robert J. Adler, Raisa E. Feldman, and Murad S. Taqqu, editors, *A Practical Guide to Heavy Tails. Statistical Techniques and Applications*. Birkhāuser, 1998.

[29] Harold L. Crutcher. A note on the possible misuse of the kolmogorov-smirnov test. *Journal of Applied Meteorology*, 14(8):1600–1603, 1975.

[30] F. N. David and N. L. Johnson. The probability integral transformation when parameters are estimated from the sample. *Biometrika*, 35(1/2):182–190, 1948.

[31] F. N. David and N. L. Johnson. The probability integral transformation when the variable is discontinuous. *Biometrika*, 37(1/2):42–49, 1950.

[32] Luc Devroye. *Non-Uniform Random Variate Generation*. Springer-Verlag, New York, 1986.

[33] The Perl Directory. Perl. http://www.perl.org/.

[34] Holger Drees, Sidney Resnick, and Laurens de Haan. How to make a hill plot. *The Annals of Statistics*, 28(1):254–274, 2000.

[35] B. Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):1–26, 1979.

[36] Agner K. Erlang. Solution of some problems in the theory of probabilities of significance in automatic telephone exchanges. *Elektroteknikerem*, 13:5–13, 1917.

[37] Leonhard Euler. De Progressionibus Harmonicus Observationes. *Commentarii Academiæ Scientarum Imperialis Petropolitanæ*, 1734(7):150–161, 1735.

[38] Julian J. Faraway. *Practical Regression and Anova using R*. 2002. http://citeseer.ist.psu.edu/642417.html.

[39] Dror Feitelson. Parallel workloads archive. http://www.cs.huji.ac.il/labs/parallel/workload/.

[40] Dror G. Feitelson. Workload modeling for performance evaluation. In M. Calzarossa and S. Tucci, editors, *Performance Evaluation of Complex Systems: Techniques and Tool*, volume 2459, pages 114–141. Springer-Verlag, Aug 2002. Lecture Notes in Computer Science.

[41] Dror G. Feitelson. *Workload Modeling for Computer System Performance Evaluation*. 2007. http://www.cs.huji.ac.il/ feit/wlmod/.

[42] Dror G. Feitelson and Dan Tsafrir. Metrics for mass-count disparity. In *Modeling, Analysis, and Simulation of Computer and Telecommunication Systems*, pages 61–68, Sep 2006.

[43] Dror G. Feitelson and Dan Tsafrir. Workload sanitation for performance evaluation. In *IEEE International Symposium on Performance Analysis of Systems and Software*, pages 221–230, Mar 2006.

[44] Anja Feldmann and Ward Whitt. Fitting mixtures of exponentials to long-tail distributions to analyze network performance models. In *INFOCOM '97: Proceedings of the INFOCOM '97. Sixteenth Annual Joint Conference of the IEEE Computer and Communications Societies. Driving the Information Revolution*, page 1096, Washington, DC, USA, 1997. IEEE Computer Society.

[45] The R Foundation for Statistical Computing. The r project for statistical computing. http://www.r-project.org.

[46] I. Foster and C. Kesselman. *The Grid: Blueprint for e New Computing Infrastructure*. Morgan Kaufmann, 1999.

[47] Ian Foster, Carl Kesselman, and Steven Tuecke. The anatomy of the grid: Enabling scalable virtual organizations. *International Journal of High Performance Computer Application*, 15(3):200–222, 2001.

[48] David Freedman and Persi Diaconis. On the histogram as a density estimator: L_2 theory. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 57(4):453–476, 1981.

[49] Michael Frigge, David C. Hoaglin, and Boris Iglewicz. Some implementations of the boxplot. *The American Statistician*, 43(1):50–54, Feb 1989.

[50] Emil J. Gambel. *Statistics of Extremes*. Columbia University Press, New York, 1958.

[51] Dani Gamerman. *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*. Chapman & Hall, 1997.

[52] F. F. Gan and Kenneth J. Koehler. Goodness-of-fit tests based on P-P probability plots. *Technometrics*, 32(3):289–303, 1990.

[53] F. F. Gan, Kenneth J. Koehler, and John C. Thompson. Probability plots and distribution curves for assessing the fit of probability models. *The American Statistician*, 45(1):14–21, 1991.

[54] Frank E. Grubbs. Sample criteria for testing outlying observations. *The Annals of Mathematical Statistics*, 21(1):27–58, Mar 1950.

[55] Nicholas J. Higham. The scaling and squaring method for the matrix exponential revisited. *SIAM Journal on Matrix Analysis and Applications*, 26(4):1179–1193, 2005.

[56] Bruce M. Hill. A simple general approach to inference about the tail of a distribution. *The Annals of Statistics*, 3(5):1163–1174, 1975.

[57] Eric. B. Holmgren. The P-P plot as a method for comparing treatment effects. *Journal of the Amirican Statistical Association*, 90(429):360–365, 1995.

[58] András Horváth and Miklós Telek. PhFit: a general phase-type fitting tool. In *12th Performance TOOLS*, volume 2324, London, UK, Apr 2002. Imperial College.

[59] Rob J. Hyndman and Yanan Fan. Sample quantiles in statistical packages. *The American Statistician*, 50(4):361–365, November 1996.

[60] C. Arthur Williams Jr. On the choice of the number and width of classes for the chi-square test of goodness of fit. *Journal of the American Statistical Association*, 45(249):77–86, 1950.

[61] Frank J. Massey Jr. The kolmogorov-smirnov test for goodness of fit. *Journal of the American Statistical Association*, 46(253):68–78, 1951.

[62] M. Kac, J. Kiefer, and J. Wolfowitz. On tests of normality and other tests of goodness of fit based on distance methods. *The Annals of Mathematical Statistics*, 26(2):189–211, 1955.

[63] T. Karagiannis, M. Faloutsos, and R. Riedi. Long-range dependence: Now you see it, now you don't! In *Global Internet Symposium*, Taipei, Taiwan, Nov 2002.

[64] Richard A. Kronmal and Jr. Arthur V. Peterson. On the alias method for generating random variables from a discrete distribution. *The American Statistician*, 33(4):214–218, Nov 1979.

[65] Guy Latouche and Vaidyanathan Ramaswami. *Introduction to Matrix Analytic Methods in Stochastic Modeling*. ASA-SIAM, Philadelphia, PA, 1999.

[66] Will E. Leland, Murad S. Taqqu, Walter Willinger, and Daniel V. Wilson. On the self-similar nature of ethernet traffic (extended version). *IEEE/ACM Trans. Netw.*, 2(1):1–15, 1994.

[67] Hui Li, Michael Muskulus, and Lex Wolters. Modeling job arrivals in a data-intensive grid. In *12th International Workshop on Job Scheduling Strategies for Parallel Processing (JSSPP)*, Saint Malo, France, Jun 2006. Springer-Verlag.

[68] Virginia Lo, Jens Mache, and Kurt Windisch. A comparative study of real workload traces and synthetic workload models for parallel job scheduling. *Lecture Notes in Computer Science*, 1459:25–46, 1998.

[69] Robert A. Lodder and Gary M. Hieftje. Quantile analysis: A method for characterizing data distributions. *Applied Spectroscopy*, 42(8):1512–1519, nov 1988.

[70] Stephen W. Looney and Jr. Thomas R. Gulledge. Probability plotting positions and goodness of fit for the normal distribution. *The Statistician*, 34(3):297–303, 1985.

[71] Muthucumaru Maheswaran, Shoukat Ali, Howard Jay Siegel, Debra A. Hensgen, and Richard F. Freund. Dynamic matching and scheduling of a class of independent tasks onto heterogeneous computing systems. In *Heterogeneous Computing Workshop*, pages 30–44, 1999.

[72] C. L. Mallows. On the probability integral transformation. *Biometrika*, 46(3/4):481–483, 1959.

[73] H. B. Mann and A. Wald. On the choice of the number of class intervals in the application of the chi square test. *The Annals of Mathematical Statistics*, 13(3):306–317, 1942.

[74] G. Marsaglia. Generating discrete random variables in a computer. *Communications of ACM*, 6(1):37–38, 1963.

[75] George Marsaglia, Wai Wan Tsang, and Jingbo Wang. Evaluating kolmogorov's distribution. *Journal of Statistical Software*, 8(18), 2003.

[76] M. Ajmone Marsan and F. Neri. Appunti delle lezioni di "Modelli di Sistemi TLC", 2002.

[77] The MathWorks. Matlab and simulink for technical computing. http://www.mathworks.com/.

[78] Stefan Mittnik, Svetlozar T. Rachev, and Marc S. Paolella. Stable paretian modeling in finance: Some empirical and theoretical aspects. In Robert J. Adler, Raisa E. Feldman, and Murad S. Taqqu, editors, *A Practical Guide to Heavy Tails. Statistical Techniques and Applications*. Birkhāuser, 1998.

[79] Cleve Moler and Charles Van Loan. Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later. *SIAM Review*, 45(1):3–49, 2003.

[80] Douglas C. Montgomery and George C. Runger. *Applied Statistics and Probability for Engineers*. John Wiley & Sons, Inc., 3rd edition, 2002.

[81] D. S. Moore. Tests of chi-squared type. In R. B. D'Agostino and M. A. Stephens, editors, *Goodness-of-Fit Techniques*, pages 97–193. Marcel Dekker, New York, 1986.

[82] Randolph Nelson. *Probability, stochastic processes, and queueing theory: the mathematics of computer performance modeling*. Springer-Verlag New York, Inc., New York, NY, USA, 1995.

[83] Marcel F. Neuts. Probability Distributions of Phase Type. In *Liber Amicorum Prof. Emeritus H. Florin, University of Louvain*, pages 173–206. Dept. of Mathematics, University of Louvain, Louvain, Belgium, 1975.

[84] Marcel F. Neuts. *Matrix-Geometric Solutions in Stochastic Models – An Algoritmic Approach*. The John Hopkins University Press, Baltimore, Maryland, 1981.

[85] Marcel F. Neuts and Miriam E. Pagano. Generating random variates from a distribution of phase type. In *Winter Simulation Conference*, 1981.

[86] M. E. J. Newman. Power laws, pareto distributions and zipf's law. *Contemporary Physics*, 46:323, 2005.

[87] John P. Nolan. *Stable Distributions - Models for Heavy Tailed Data*. Birkhäuser, Boston, 2007. In progress, Chapter 1 online at academic2.american.edu/∼jpnolan.

[88] Takayuki Osogami and Mor Harchol-Balter. Closed form solutions for mapping general distributions to quasi-minimal ph distributions. *Performance Evaluation*, 63(6):524–552, 2006.

[89] Marita Osslon. The EMpht programme. Technical report, Department of Mathematics, Chalmers University of Technology, Jun 1998. http://www.math.1th.se/matstat/staff/asmus/pspapers.htm.

[90] Lawrence Page and Sergey Brin. Method for node ranking in a linked database - us patent 7058628.

[91] Lawrence Page and Sergey Brin. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30(1–7):107–117, 1998.

[92] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford Digital Library Technologies Project, 1998.

[93] Vern Paxson and Sally Floyd. Wide area traffic: the failure of Poisson modeling. *IEEE/ACM Transactions on Networking*, 3(3):226–244, 1995.

[94] E. S. Pearson. The probability integral transformation for testing goodness of fit and combining independent tests of significance. *Biometrika*, 30(1/2):134–148, 1938.

[95] Karl Pearson. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to hove arisen from random sampling. *The Philosophical Magazine*, 50(5):157–175, 1900.

[96] Karl Pearson. On the probability that two independent distributions of frequency are really samples from the same population. *Biometrika*, 8(1/2):250–254, 1911.

[97] A. N. Pettitt. A two-sample anderson-darling rank statistic. *Biometrika*, 63(1):161–168, 1976.

[98] Alfio Quarteroni, Riccardo Sacco, and Fausto Salieri. *Matematica Numerica*. Springer-Verlag, Italia, Milano, 2nd edition, 2000.

[99] Alfio Quarteroni and Fausto Salieri. *Introduzione al Calcolo Scientifico. Esercizi e Problemi risolti con MATLAB*. Springer-Verlag, Italia, Milano, 2nd edition, 2004.

[100] Murray Rosenblatt. Remarks on a multivariate transformation. *The Annals of Mathematical Statistics*, 23(3):470–472, 1952.

[101] Sheldom M. Ross. *Introduction to Probability and Statistics for Engineers and Scientists*. Academic Press, 2nd edition, 2000.

[102] Stuart Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach*. Prentice-Hall, Englewood Cliffs, NJ, 2nd edition, 2003.

[103] David W. Scott. On optimal and data-based histograms. *Biometrika*, 66(3):605–610, 1979.

[104] George W. Snedecor and William G. Cochran. *Statistical Methods*. Iowa State University Press, Iowa, 7th edition, 1980.

[105] M. A. Stephens. EDF statistics for goodness of fit and some comparisons. *Journal of the American Statistical Association*, 69(347):730–737, 1974.

[106] M. A. Stephens. Tests based on EDF statistics. In R. B. D'Agostino and M. A. Stephens, editors, *Goodness-of-Fit Techniques*, pages 97–193. Marcel Dekker, New York, 1986.

[107] Murad S. Taqqu and Vadim Teverovsky. On estimating the intensity of long-range dependence in finite and infinite variance time series. In Robert J. Adler, Raisa E. Feldman, and Murad S. Taqqu, editors, *A Practical Guide to Heavy Tails. Statistical Techniques and Applications*. Birkhāuser, 1998.

[108] Murad S. Taqqu, Vadim Teverovsky, and Walter Willinger. Estimators for long-range dependence: an empirical study. *Fractals*, 3(4):785–798, 1995.

[109] Kishor S. Trivedi. *Probability and Statistics with Reliability, Queuing, and Computer Science Applications*. John Wiley and Sons, New York, 2nd edition, 2001.

[110] John Tukey. *Exploratory Data Anlysis*. Addison-Wesley, 1977.

[111] Alastair J. Walker. An efficient method for generating discrete random variables with general distribution. *ACM Transactions on Mathematical Software*, 3(3):253–256, 1977.

[112] Robert C. Ward. Numerical computation of the matrix exponential with accuracy estimate. *SIAM*, 14(4):600–610, Sep 1977.

[113] W. Weibull. The phenomenon of rupture in solids. *Ingeniörsvetenskapsakademiens Handlingar*, 153(17), 1939.

[114] W. Weibull. A statistical distribution function of wide applicability. *Journal of Applied Mechanics*, 18(3):293–297, 1951.

[115] M. B. Wilk and R. Gnanadesikan. Probability plotting methods for the analysis of data. *Biometrika*, 55(1):1–17, 1968.

[116] Ian H. Witten and Eibe Frank. *Data Mining. Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan-Kaufmann, 1999.