UNIVERSITÀ DEGLI STUDI DI PADOVA

FACOLTÀ DI SCIENZE STATISTICHE

CORSO DI LAUREA

IN STATISTICA E GESTIONE DELLE IMPRESE

TESI DI LAUREA TRIENNALE

LA TEORIA DEI VALORI ESTREMI PER L'IDENTIFICAZIONE DI GENI DIFFERENZIALMENTE ESPRESSI

RELATORE: PROF.SSA LAURA VENTURA

CORRELATORE: DOTT.SSA CHIARA ROMUALDI

LAUREANDA: PATRIZIA DETIMO

ANNO ACCADEMICO 2004-2005

INDICE

INTRODUZIONE

Ormai da circa 5 anni il completamento della mappatura del DNA è avvenuto! Infatti, nell'Aprile del 2000 la Celera Genomics, società privata di biotecnologie, in corsa con gli scienziati del Progetto Genoma Umano per la decifrazione del nostro DNA ha annunciato di aver completato l'opera. I ricercatori della Celera sono riusciti a mettere nel giusto ordine tutte le "lettere chimiche" con cui è scritto il codice della vita di un essere umano.

L'analisi del DNA rappresenta, senza dubbio, un valido metodo d'indagine sia in campo clinico sia nella diagnostica di un numero sempre maggiore di malattie genetiche. È ormai dimostrato che le diverse manifestazioni patologiche, che insieme vanno sotto il nome di cancro, sono caratterizzate dalla crescita incontrollata e invasiva di gruppi di cellule geneticamente alterate. Il cancro è, dunque, una malattia genetica.

Un utile strumento, al fine di individuare geni con un coinvolgimento diretto in specifici eventi cellulari ed in particolari condizioni fisiopatologiche è la tecnologia del *DNA microarray*, che consente di valutare il livello d'espressione di decine di migliaia di geni (al limite anche l'intero genoma umano).

Proprio l'espressione genica, infatti, è l'obiettivo degli studi che vengono condotti oggi sulla genetica, specialmente in campo medico: capire quando, dove e in che quantità ogni gene è espresso, ossia attivo.

Nel presente elaborato ci si propone lo studio delle leucemie con lo scopo di effettuare un'accurata estrapolazione dei valori estremi. A tal fine verrà utilizzata la Teoria dei Valori Estremi, un'insieme di procedure scientificamente e statisticamente razionali utili a studiare il comportamento estremo di variabili aleatorie. Applicazioni pratiche di modelli per valori estremi si possono trovare in diverse discipline scientifiche: nelle scienze ambientali, finanziarie, biologiche e così via.

Da un punto di vista statistico il problema della Teoria dei Valori Estremi è un problema di estrapolazione. Infatti, considerati i dati prodotti dal processo cui siamo interessati, vogliamo predire il comportamento del processo per livelli più grandi di quelli osservati. L'estrapolazione è quindi ottenuta analizzando il comportamento del modello oltre il *range* dei dati osservati.

Fondamentalmente esistono buone approssimazioni per la coda di una distribuzione, ma è difficile capire dove iniziare ad approssimare la coda. Questo significa che non sappiamo il livello oltre il quale poter considerare le osservazioni come osservazioni estreme. Ed è proprio in questo contesto che entra in gioco la cosiddetta Teoria dei Valori estremi. Dall'utilizzo dei dati *microarray*, infatti, scopo di questa tesi è l'approfondimento di un modello (*Threshold Model*) che possa, in qualche modo, aiutarci a risolvere il problema della selezione della soglia.

Da ricerche bibliografiche è emerso che questo tipo di approccio non è mai stato adottato su dati derivanti da *microarray* e per verificare l'attendibilità dei risultati si è reso necessario effettuare un confronto con uno dei test più usati in ambito biologico per l'identificazione di geni differenzialmente espressi: il *Significance analisys of microarray* (SAM).

Il confronto conduce a risultati simili in termini di numerosità, ma i geni estrapolati non sono identici: solo il 45% dei geni risulta comune ad entrambi i metodi; è evidente che sarebbero necessarie delle ricerche biologiche più dettagliate per ottenere risultati più concreti, ma questo esula dagli interessi del presente elaborato che si propone come una ricerca tutt'altro che definitiva.

La struttura è la seguente: dopo il primo capitolo, in cui vengono richiamati alcuni concetti base di biologia cellulare ed effettuata una rassegna sulle scoperta riguardanti il Genoma e sulla *Microarray Tecnology*, nel capitolo due viene descritta la Teoria dei Valori Estremi, soffermandosi, in particolare, sul *Modello della Soglia*. Nel capitolo tre viene presentato il dataset, si passa all'applicazione e, con l'ausilio di grafici e tabelle, si mostrano alcuni interessanti risultati, ottenuti anche grazie al confronto effettuato con il test Sam. Infine, nel capitolo quattro, dedicato alle conclusioni, vengono messi in luce vantaggi e svantaggi della Teoria dei Valori Estremi e i problemi non ancora risolti.

1

Il Genoma Umano e la Microarray Tecnology

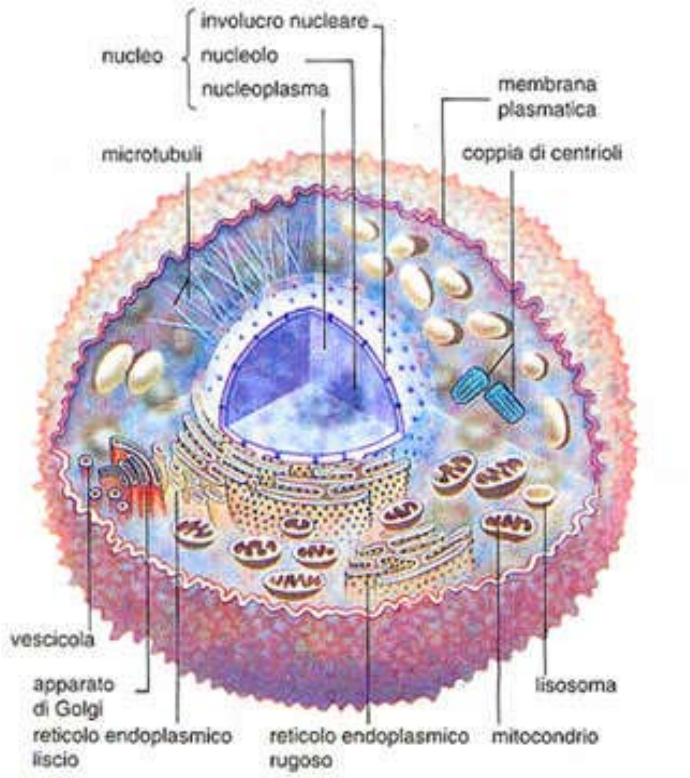
1.1 Introduzione

Al fine di agevolare la trattazione della fase sperimentale, in questo capitolo verranno richiamati alcuni concetti base di biologia cellulare (cellula, DNA, sintesi proteica, ...). Successivamente verrà effettuata una breve trattazione riguardo agli sviluppi più recenti sul genoma e alla nuova tecnologia utilizzata: i microarray.

Poiché i concetti qui trattati sono esaustivi ai fini della presente tesi, per una trattazione più dettagliata degli argomenti si rinvia a testi più specializzati (Alberts, 1999).

1.2 Alcune nozioni di biologia

La *cellula* è l'unità fondamentale di ogni organismo vivente. Al microscopio si presenta come una delimitazione nettissima, indice di una membrana che la racchiude, denominata *membrana plasmatica* la quale, oltre a garantire l'integrità funzionale della cellula, regola il passaggio delle sostanze dall'interno verso l'esterno e viceversa.

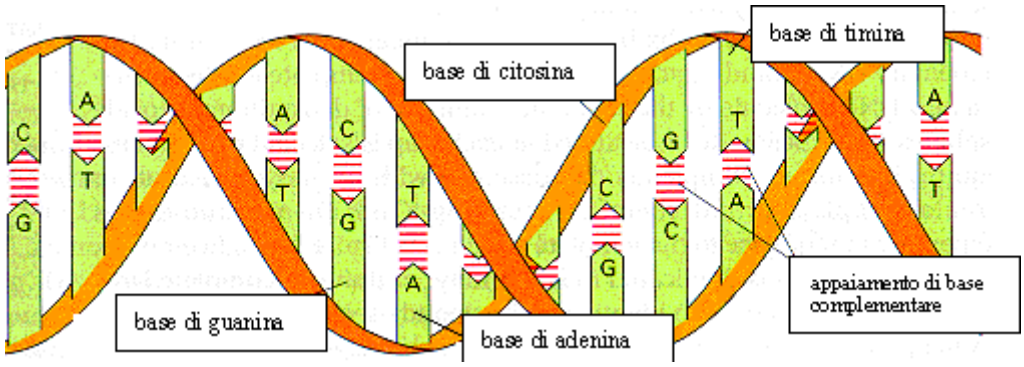


All'interno della cellula si trova il *citoplasma* e il *reticolo endoplasmatico*, una soluzione acquosa concentrata, contenente enzimi, ioni e molecole disciolte, oltre ad un certo numero di organuli con funzioni specifiche.

Tra questi organuli rivestono particolare interesse i *ribosomi*, ossia i siti in cui ha luogo l'assemblaggio e la sintesi proteica. Essi possono ricoprire il reticolo endoplasmatico oppure trovarsi liberi nel citoplasma. Altri organelli presenti nel citoplasma sono: i *mitocondri*, in cui avvengono le reazioni chimiche che forniscono energia per le attività cellulari, *l'apparato di Golgi*, dove sono immagazzinate le molecole sintetizzate nella cellula, i *lisosomi* e i *perossisomi*, che sono delle vescicole in cui le molecole di rifiuto vengono scomposte in elementi più semplici che possono essere riutilizzati dalla cellula oppure eliminati definitivamente. Il citoplasma è inoltre fornito di un *citoscheletro*, che determina la forma della cellula, le consente di muoversi e fissa i suoi organuli (*Alberts, 2002*).

Al centro della cellula, immerso nel citoplasma, domina un grosso corpo tondeggiante, il nucleo: esso contiene l''informazione genetica dell''organismo, immagazzinata nella struttura molecolare del DNA.

Il DNA (acido desossiribonucleico) è una lunga molecola elicoidale costituita da due filamenti superavvolti l'uno sull'altro e uniti da ponti infinitesimali detti *ponti idrogeno*. Lo scheletro dei due filamenti è costituito da una serie ripetuta di una molecola di fosfato alternata con una molecola di zucchero desossiribosio. I ponti idrogeno sono formati tra le coppie di basi azotate. Uno zucchero desossiribosio, un gruppo fosfato e una base azotata costituiscono un *nucleotide*.



Esistono quattro tipi di basi nel DNA: adenina (A), timina (T), citosina (C), guanina (G) che in base alla loro complementarietà di struttura appaiano sempre nello stesso modo, adenina con timina e citosina con guanina. Esse sono una sorta di alfabeto col quale viene scandito il messaggio genetico: la successione delle triplette nel DNA individua la sequenza di un particolare gene, che è per l'appunto un tratto di DNA che possiede regioni regolative ed altre codificanti per la sintesi delle proteine (*Lewin, 1994*).

Quando una molecola di DNA si duplica, i due filamenti si separano grazie alla rottura dei legami idrogeno e ciascuno, con le proprie basi azotate, funge da stampo per la formazione di un nuovo filamento complementare. E' così che l'informazione ereditaria si trasmette fedelmente da una cellula madre alla cellula figlia in quella che viene detta *duplicazione semiconservativa*.

La successione di nucleotidi nella molecola di DNA, in base al codice genetico determina una sequenza di amminoacidi, ossia delle unità elementari che insieme costituiscono la struttura primaria di una proteina: in generale, una serie di tre

nucleotidi (detta *codone*) codifica per un amminoacido.

Il processo attraverso il quale il DNA viene tradotto in proteine avviene in due fasi fondamentali: *trascrizione* e *traduzione*. Durante la prima fase, l'informazione viene *trascritta* da un filamento singolo di DNA in un filamento singolo di RNA detto messaggero (mRNA). L'RNA messaggero è del tutto simile al DNA. La sola differenza è che al posto della timina si trova un'altra base azotata: *l'uracile*. Una volta trascritto, l'mRNA lascia il nucleo e si sposta nel *citosol* dove, a livello dei ribosomi, ha luogo la sintesi proteica o *traduzione.* I ribosomi sono costituiti da subunità formate da RNA ribosomiale (rRNA) e proteine. Nel processo di traduzione interviene anche un'altra molecola di RNA che è detta RNA di trasporto (tRNA) e provvede al trasporto degli amminoacidi in sede ribosomale. Il tRNA è munito di una tripletta di basi, detta *anticodone*, specifica per l'amminoacido che trasporta. Durante la sintesi proteica, il tRNA mette in corrispondenza ciascuna tripletta di basi (codone) del mRNA che si trova all'interno del ribosoma con il suo anticodone e rilascia il suo amminoacido alla catena proteica in accrescimento, solo se l'anticodone appaia perfettamente con il codone nel mRNA.

In questo modo, in base alla sequenza dettata inizialmente dal DNA, le unità amminoacidiche vengono allineate una dopo l'altra andando ad assemblare la catena polipeptidica, ossia la *proteina nucleotidica*.

Molte malattie genetiche sono il risultato della mancanza o dell'inattività di enzimi o altre proteine non enzimatiche, a causa di mutazioni nei geni che ne codificano la loro struttura.

Le mutazioni sono cambiamenti nella sequenza del DNA, dovute ad aggiunta, delezione o sostituzione di uno o più nucleotidi (*Lewin, 1992*).

1.3 Gli sviluppi più recenti sul genoma

Recenti studi hanno permesso di leggere tutto il patrimonio genetico umano; nell'ottobre del 1990, infatti, è ufficialmente partito il Progetto Genoma Umano, avente l'obiettivo di mappare l'esatta sequenza di tutto il corredo cromosomico contenuto nelle cellule umane e con il fine ultimo di prevenire e curare tutte le malattie che hanno causa genetica (*Di Giorgio, 2000*). Sono già stati pubblicati molti risultati da parte di scienziati facenti parte dell'Human Genome Project, ma naturalmente il cammino è tutt'altro che banale. Gli ostacoli da affrontare sono parecchi: il DNA è costituito in gran parte da porzioni di materiale non-codificante (INTRONI), e poi le malattie più diffuse sono causate dall'azione di più geni contemporaneamente, e quindi è difficile seguirne l'alterazione di generazione in generazione. Una malattia che lo studio del genoma può sicuramente aiutare a sconfiggere è il cancro, che è noto derivi da malfunzionamenti cellulari. Al giorno d'oggi non si conosce quale sia la natura specifica del problema: essi si identificano solo per il fatto di colpire un certo organo (*Boncinelli, 2004*).

Il futuro? Sarà definire il profilo genetico di un individuo per calcolare il suo rischio di sviluppare un certo tumore, effettuare diagnosi sempre più precise e precoci e disegnare farmaci mirati e adatti a ogni individuo. In questo senso, i ricercatori del Cancer Genome Project di Cambridge stanno già impegnando i risultati del Progetto Genoma Umano per confrontare il DNA di persone sane con quello di persone malate, attraverso i DNA microarray, vetrini delle dimensioni di pochi centimetri sui quali viene depositato un grande numero di geni (*Sorvillo, 2000*). I microarray consentono di verificare quali geni sono attivi, qual è il loro livello di espressione e quali variazioni occorrono in condizioni patologiche.

Per comprendere la tecnica dei *microarray chip* è fondamentale notare che, nella fase di trascrizione, ciascuna cellula produce RNA solamente per quei geni (ossia quei segmenti di DNA) che sono attivi in quel momento; pertanto un modo per indagare quali sono i geni attivi e quali quelli inattivi in un determinato istante sarà quello di analizzare l'RNA prodotto dalla cellula, ed è da questo punto che parte l'intuizione della *DNA microarray tecnology* (*Tilstone, 2003*)

1.4 Le nuove tecnologie: i *microarray*

L'innovazione si chiama appunto *DNA microarray*, una mistura genetica, chimica, elettronica ed informatica (*Allison e David, 2002*). Si osservi anche come i *DNA chip* o *chip genetici*, sono uno strumento importante delle cosiddette "nanotecnologie". Essi sono utili per lo studio dell'espressione genica e di grande interesse per i ricercatori che studiano le basi molecolari del cancro e di altre malattie complesse oltre che, in ambito farmacologico, per l'individuazione di nuovi farmaci.

Messi sul mercato nel 1996 consentono di analizzare contemporaneamente l'attività di decine di migliaia di geni (fino a poco tempo fa, i ricercatori potevano analizzare solo un gene alla volta, tanto che si diceva: "un gene, una vita"). I *chip* sono formati da moltissime molecole di DNA (detti *sonde*) depositate in una posizione nota su un supporto a formare una microgriglia (da cui il nome *microarray*) che consente di identificarle in modo univoco. Il supporto di solito è un vetrino da microscopio che ha le dimensioni, più o meno, di un pollice della mano. Ogni sonda è costituita da un segmento di DNA a singola elica di un gene e, nel loro insieme, tutte le sonde di un DNA chip rappresentano tutti, o la maggior parte, dei geni di un organismo. I concetti chiave per la misura del livello d'espressione genica sono due: la *retro-trascrizione* e l'*ibridazione*.

La *retro-trascrizione* è il processo inverso rispetto alla trascrizione. Nella trascrizione, un filamento singolo di DNA viene usato come stampo per la costruzione di una molecola di RNA. Durante la retro-trascrizione, avviene il contrario: una molecola di mRNA viene in questo caso retro-trascritta in DNA complementare o cDNA. Si noti che il mRNA di una cellula può essere isolato sperimentalmente e retrotrascritto andando così a costituire la libreria di cDNA della cellula. Inoltre il cDNA rappresenta una molecola molto più stabile del mRNA e quindi più adatta ad essere studiata e manipolata.

L'*ibridazione*, invece, è un processo di appaiamento delle due eliche singole di DNA o RNA. Esse si separano ad una temperatura caratteristica di 95°C. Se la temperatura viene fatta scendere e mantenuta al di sotto di quella di separazione, le

due eliche tornano ad unirsi con la loro controparte. Tale processo è basato sul principio di appaiamento delle basi che consente l'unione, ossia l'*ibridazione*, solo di segmenti di DNA o RNA tra loro complementari (*Pierotti e Garibaldi, 2004*).
Come esempio di applicazione dei *microarray* si considera l'identificazione dei geni peculiarmente espressi o non espressi in un tessuto tumorale rispetto al relativo tessuto normale. Quando i geni sono attivamente espressi, vale a dire sono attivamente "trascritti", nelle cellule di questo tessuto sarà presente un numero elevato di molecole di RNA messaggero corrispondente ai geni espressi rispetto al tessuto sano. Si estrae pertanto l'RNA dai due tipi di tessuti (sano e tumorale), si converte l'mRNA nella coppia più stabile a DNA (cDNA) e vi si lega un marcatore fluorescente: ad esempio verde per il cDNA ottenuto da cellule tumorali e rosso per quello ottenuto da cellule sane. Si applicano poi i cDNA marcati al *chip*.
Quando il cDNA trova la sua sequenza di basi complementari tra le decine di migliaia di sonde depositate sul chip, vi si appaia. In quel punto del *microarray* si ha emissione di fluorescenza, indice dell'espressione di quel determinato gene. I *chip* vengono quindi analizzati con uno *scanner*, strumento che valuta il quadro di fluorescenza e i risultati sono elaborati da un computer. Si ottiene come risposta una mappa a colori: segnale rosso se un gene è espresso solo nel tessuto sano e verde, se un gene è espresso solo nel tessuto tumorale, e infine diverse gradazioni di giallo (rosso + verde) se un gene è espresso in entrambi i tessuti a livelli diversi. In altre parole si ottiene quello che viene definito un *profilo d'espressione*, che consente di confrontare i quadri di espressione genica in tessuti diversi, o nello stesso tessuto in differenti condizioni, oppure in cellule a stadi diversi di sviluppo.
La tecnologia, che è alla base dei chip, in realtà è molto complessa e l'utilizzo nella ricerca biomedica di questi utili "cacciatori di geni" è solo agli inizi (*Techini, 2004*).

Il dataset cui si dispone è stato costruito per mezzo di questa tecnologia e verrà presentato nell'ultimo capitolo, insieme con la relativa applicazione che consiste nella selezione della soglia, effettuata tramite il "*Thresholding Model"*, oggetto del prossimo capitolo.

2

Teoria dei Valori Estremi: Modello della Soglia

2.1 Introduzione

Questo capitolo presenta una breve rassegna sulla Teoria dei Valori Estremi, partendo dalla descrizione del modello classico di tale teoria per poi giungere ad illustrare, in modo più approfondito, il modello della soglia (Threshold Model).

Il tradizionale approccio (Metodo Classico, Gumbel 1958) è basato sulla distribuzione limite del valore estremo già identificato da Fisher e Tippet (1928). Una tipica applicazione consiste nella distribuzione generalizzata dei valori estremi (GEV), stimata per massimi annuali.

I metodi soglia, invece, sono basati sulla stima di un modello stocastico per valutare le eccedenze o i picchi oltre una certa soglia.

Metodi simili in qualche forma hanno girato per lungo tempo, ma il primo sviluppo sistematico fu nelle opere di Todorovic e Zelenhasic (1970) e in quelle di Todorovic e Rousselle (1971). Successivamente le versioni hanno assunto un processo di Poisson non omogeneo per modellare le eccedenze oltre una certa soglia, combinato con variabili casuali esponenziali per rappresentare le quantità dalle quali la soglia è superata.

North (1980) ha proposto una versione stagionale nella quale entrambe le intensità del processo di eccedenze e il livello medio del processo dipendono da componenti sinusoidali nel tempo, ma sono state proposte numerose varianti più recenti.

Il nostro approcio è basato sulla distribuzione generalizzata di Pareto (GDP), idea suggerita inizialmente da Pickands (1975), ma sviluppata molto dopo da Dumauchel (1983), Davison (1984 a,b), Smith (1984, 1987), Von Montfort e Witter (1985, 1986), Hosking Willis (1987) e Joe (1987), (Davison e Smith, 1990).

2.2 La Teoria dei Valori Estremi

La teoria dei valori estremi ha avuto sviluppi molto rapidi negli ultimi cinquant'anni. Può essere definita come un insieme di procedure scientificamente e statisticamente razionali utili a stimare il comportamento estremo di variabili o processi casuali (*Coles, 2001*).

Date *n* variabili casuali indipendenti e identicamente distribuite (i.i.d.) X_1, X_2, ..., X_n da una distribuzione $F(\cdot)$, il punto di partenza per l'analisi dei valori estremi è lo studio del comportamento di:

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

chiamato Ordine statistico massimo. La sua funzione di ripartizione è:

$$\begin{aligned} P\{M_n \leq x\} &= P\{X_1 \leq x, \ldots, X_n \leq x\} \\ &= P\{X_1 \leq x\} \times \cdots \times P\{X_n \leq x\} \\ &= \{F(x)\}^n. \end{aligned} \tag{2.1}$$

Tuttavia, non conoscendo $F(\cdot)$, tale probabilità non risulta semplice da stimare.

Una possibilità sarebbe quella di stimare $F(\cdot)$ con tecniche statistiche e sostituire la stima in $F(x)^n$. Però piccole discrepanze nella stima di $F(\cdot)$ ci porterebbero a sostanziali discrepanze per $F(\cdot)^n$.

L'alternativa è quella di adottare un approccio asintotico, cioè studiare i limiti della distribuzione di M_n, con *n* tendente ad infinito e usare questa famiglia come un'approssimazione di M_n con *n* finito. Sappiamo che necessariamente, con probabilità pari a 1, la distribuzione di M_n per $n \to \infty$ converge all'estremo superiore di F, quindi $F^n(x) \to 0$. A questo punto adottiamo lo stesso approccio per la stima della distribuzione della media campionaria, giustificato dal Teorema del Limite Centrale. In quel caso, con probabilità pari a 1 la media di X_n converge a μ, quindi si adotta una normalizzazione dei dati e si trova che X_n si distribuisce come una Normale standardizzata. Nel caso di M_n usiamo invece un'altra modifica che,

vedremo, si distribuirà come una GEV(μ,σ,ξ):

$$M_n* = \frac{M_n - b_n}{a_n} \sim G(\mu,\sigma,\xi),$$

con a_n e b_n sequenze di coefficienti normalizzati.

2.2.1 Distribuzione Generalizzata dei Valori estremi

L'insieme delle possibili distribuzioni per $M_n* = (M_n - b_n)/a_n$ è dato dal Teorema 1:

Teorema 1, Extremal types theorem: se esistono sequenze di costanti $a_n > 0$ e b_n, con n tendente ad infinito, tale che la probabilità

$$P\left\{\frac{M_n - b_n}{a_n} \le x\right\} \to G(x),$$

dove G è una distribuzione non degenerata, allora G segue una delle seguenti funzioni di ripartizione:

I: G(x) = $\exp\left\{-\exp\left(-\frac{x-b}{a}\right)\right\} \quad -\infty < x > \infty;$ distrib. Gumbel

II: G(x) = $\begin{cases} 0 & \text{se } x \le b \\ \exp\left\{-\left(\frac{x-b}{a}\right)^{-\alpha}\right\} & \text{se } x > b, \alpha > 0; \end{cases}$ distrib. Fréchet

III: G(x) = $\begin{cases} \exp\left\{-\left[-\left(\frac{x-b}{a}\right)\right]^{\alpha}\right\} & \text{se } x < b, \alpha > 0 \\ 1 & \text{se } x \ge b; \end{cases}$ distrib. Weibul

Ogni famiglia ha un parametro di posizione b e uno di scala a, in più la Fréchet e la Weibull hanno un ulteriore parametro di forma α.

Le tre distribuzioni hanno una forma di comportamento distinta, in corrispondenza al differente comportamento delle code della distribuzione F. Per fini statistici non è conveniente lavorare con tre classi di distribuzioni diverse, si usa perciò utilizzare un modello che li contenga tutti e tre.

Quest'ultimo viene definito Distribuzione generalizzata dei Valori estremi G(μ,σ,ξ), dall'inglese GEV (Generalized Extreme Value distribution), con funzione di ripartizione:

$$G(x) = \exp\left\{-\left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}\right\}, \tag{2.2}$$

definita per $1+\xi\left(\frac{x-\mu}{\sigma}\right) > 0; -\infty < \mu > +\infty; \sigma > 0; -\infty < \xi < +\infty.$

La distribuzione ha tre parametri: μ, σ e ξ, rispettivamente parametri di posizione, di scala e di forma.

2.3 Modello della soglia

Il modello classico precedentemente analizzato è un approccio dispendioso dell'analisi dei valori estremi in quanto considera solo i dati nelle estremità, anche se altri dati sono accessibili. Nel nostro caso risulta più naturale considerare come eventi estremi non solo i massimi ma tutti quei valori che eccedono una determinata soglia u.

Siano $x_1, \ldots, x_n$ realizzazioni i.i.d. di una ignota distribuzione F. Consideriamo quindi quelle x_i per cui

$$x_i > u \qquad \text{con } u = \text{soglia fissata,}$$

che vengono rinominate $x_{(1)}, \ldots, x_{(k)}$. L'analisi si basa su,

$$y_j = x_{(j)} - u, \qquad \text{per } j = 1, \ldots, k$$

Indicando con x, un termine arbitrario della sequenza degli $x_{(i)}$, segue che una descrizione del comportamento stocastico di questi punti estremi è data dalla probabilità condizionata,

$$P(x > u + y \mid x > u) = \frac{1 - F(u + y)}{1 - F(u)} \qquad \text{con } y > 0.$$

2.3.1 Distribuzione generalizzata di Pareto

Non conoscendo F, a partire dal Teorema 2, troviamo un'approssimazione utile ripercorrendo la logica utilizzata per la distribuzione GEV.

Teorema 2: Data una sequenza di variabili indipendenti X_1 ... X_n, con distribuzione comune F, se vale il Teorema 1, secondo cui:

$$P(M_n \le z) \approx G(z) \qquad \text{dove } G(x) = \exp\left\{-\left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}\right\},$$

allora la distribuzione di Y, definita come $X-u$ con $x > u$ e u molto grande, è approssimativamente

$$H(y) = 1-\left(1+\frac{\xi y}{\tilde{\sigma}}\right)^{-\frac{1}{\xi}} \tag{2.3}$$

$$\text{con } \tilde{\sigma} = \sigma + \xi(u-\mu), \tag{2.4}$$

dove $y > 0$, $\quad 1+\frac{\xi y}{\tilde{\sigma}} > 0$.

Questa famiglia di distribuzioni prende il nome di Distribuzione Generalizzata di Pareto (GPD). Il teorema implica che le due distribuzioni, la GEV e la GPD, sono in correlazione: i parametri della famiglia di Pareto sono infatti determinati unicamente da quelli della distribuzione GEV e, in particolare ξ è uguale tra le due. Questo significa che il parametro di forma ξ è dominante nel determinare il comportamento qualitativo della distribuzione GPD, come lo era per GEV.
In particolare:

- Se $\xi < 0$ la distribuzione degli eccessi ha un limite superiore,
- Se $\xi > 0$ la distribuzione non ha limiti,
- Se $\xi = 0$ si prende in considerazione il limite di $\xi \to 0$ che porta a

$$H(y) = 1-\exp\left(-\frac{y}{\tilde{\sigma}}\right) \qquad y > 0,$$

corrispondente a un esponenziale con parametro $1/\tilde{\sigma}$.

Possiamo quindi concludere che le y_j sono visti come realizzazioni indipendenti di una variabile aleatoria, la cui distribuzione può essere approssimata da un membro della famiglia generalizzata di Pareto.

Giustificazioni del teorema: Partiamo dal risultato del Teorema 1

$$F^n(z) \approx \exp\left\{-\left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}\right\},$$

per parametri μ, $\sigma > 0$ e ξ,

$$n\log(F) \approx -\left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}. \qquad (2.5)$$

A questo punto ci avvaliamo dell'espressione di Taylor: $\log F(z) \approx -1-F(z)$, valida per valori grandi di z, che sostituita in (2.5) ci dà, proprio come stavamo cercando, un'approssimazione di F

$$1-F(u) \approx \frac{1}{n}\left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}.$$

Possiamo dunque stimare la probabilità:

$$P(X > u+y \mid X > u) \approx \frac{n^{-1}\left[1+\xi(u+y-\mu)/\sigma\right]^{-\frac{1}{\xi}}}{n^{-1}\left[1+\xi(u-\mu)/\sigma\right]^{-\frac{1}{\xi}}},$$

$$= \left[1+\frac{\xi y}{\tilde{\sigma}}\right]^{-\frac{1}{\xi}},$$

dove $\tilde{\sigma} = \sigma + \xi(u-\mu)$.

2.3.2 Il problema della selezione della soglia

La soglia *u* deve essere fissata in modo da assicurare un equilibrio tra distorsione e varianza: una soglia troppo bassa va contro le ipotesi asintotiche del modello e fa aumentare la distorsione, mentre una soglia troppo alta genera pochi valori estremi su cui stimare il modello, e quindi la varianza sarebbe elevata. Esistono due metodi per individuare *u*:

1. Tecnica esplorativa, basata sulla media della distribuzione generalizzata di Pareto.
2. Ricerca della stabilità dei parametri, basata sulla stima di vari modelli in un rango di differenti soglie.

1. Quando $\xi < 1$, la media della GPD è $E(Y) = \sigma/(1 - \xi)$, quando $\xi \geq 1$ è, invece, infinita. Se u_0 è la soglia prescelta, la media è

$$E(X - u_0 \mid X > u_0) = \frac{\sigma_{u0}}{1 - \xi} \quad \text{con } \xi < 1,$$

dove indichiamo con σ_{uo} il parametro di scala riferito a u_o. Ma se la distribuzione GPD è valida per u_o, allora è valida ugualmente anche per tutte le soglie $u > u_o$. Quindi

$$E(X - u \mid X > u) = \frac{\sigma_u}{1 - \xi} = \frac{\sigma_{u0} + \xi u}{1 - \xi}, \tag{2.6}$$

in virtù della (2.4). Per $u > u_o$ l'espressione della media è una funzione lineare in *u*, dove $u < x_{max}$. Rappresentando, in un grafico, la stima della media al variare della soglia, otteniamo un grafico di punti:

$$\left\{ u, \frac{1}{n_u} \sum_{i=1}^{nu} (x_{(i)} - u) \right\},$$

dove n_u rappresenta il numero di valori che eccedono la soglia u.
Il grafico che si ottiene si chiama *Mean residual life plot* e permette di individuare u_o come quel punto dell'asse della ascisse per cui, al di sopra di esso, la curva delle medie appare approssimativamente lineare. Non sempre però il grafico è facile da interpretare. Per maggiori conferme utilizziamo la seconda procedura.

2. Il secondo metodo per la scelta di u_o è quello di stimare il modello sotto un rango di soglie, fino a raggiungere la stabilità dei parametri.
Al variare degli u, maggiori di u_o, per i quali l'ipotesi asintotica della distribuzione generalizzata di Pareto rimane valida, si verifica che:

- la stima di ξ deve rimanere *costante* perchè i parametri di forma sono identici al variare di u, $\xi_{uo} = \xi_u$.
- la stima di σ deve essere *lineare in u* perchè, per la (2.4)

$$\sigma_u = \sigma_{uo} + \xi(u - u_o),$$

il parametro di scala cambia con u. Per renderlo costante lo riparametrizziamo:

$$\sigma^* = \sigma_u - \xi_u.$$

Verifichiamo queste due ipotesi rappresentando in un grafico ξ e σ^* contro u, e selezioniamo la soglia u_0 come il valore più basso di u per il quale le stime rimangono quasi-costanti.

L'applicazione pratica dei metodi visti sopra sarà oggetto del prossimo capitolo, in cui verrà presentato anche un confronto con il Test SAM.

3

Dataset con Applicazione

3.1 Introduzione

Nel capitolo precedente, dedicato alla rassegna dei modelli per i valori estremi e, in particolare al modello della soglia, abbiamo posto le basi, da un punto di vista teorico, per poter ora analizzare e confrontare in modo puramente applicativo i risultati cui giungiamo attraverso l'implementazione del "Threshold Model".

I risultati ottenuti verranno successivamente confrontati con il "test Sam" che, tra i metodi conosciuti per l'identificazione dei geni differenzialmente espressi, risulta essere il più adatto per i dati cui si dispone.

Le analisi che verranno proposte nel seguito sono state condotte utilizzando il software statistico R disponibile gratuitamente sul sito http://www.r-project.org.

Le funzioni utilizzate per il "Threshold Model" sono state realizzate da Alec Sthephenson, dell'Università di Lancaster, e sono raccolte nella libreria "ismev", che verrà presentata nei paragrafi successivi.

3.2 Il dataset

Il dataset è stato fornito dal centro di ricerca Interdipartimentale per le Biotecnologie Innovative (CRIBI) dell'Università degli studi di Padova (*Romualdi et al, 2003*). Si dispone dell'espressione di 4992 geni che si riferiscono a 10 soggetti affetti da leucemia linfoblastica acuta-B (ALL-B), (*Smith, 2002*).

Le leucemie sono malattie neoplastiche caratterizzate da una proliferazione tumorale dei tessuti linfoidi (malattie linfoproliferative) o mieloidi (malattie mieloproliferative), che presentano un aumento della proliferazione cellulare, una diminuzione della differenziazione e un'alterazione importante dei meccanismi che controllano la morte cellulare programmata. Possono essere acute se la proliferazione interessa cellule incapaci di differenziarsi e maturare completamente, o croniche, se la proliferazione interessa cellule in grado di differenziarsi e maturare. Nelle leucemie linfoidi la cellula neoplastica è già determinata in senso linfoide (linfociti T e B). Nelle leucemie mieloidi la trasformazione tumorale interessa una cellula altamente indifferenziata o una cellula in senso mieloide. La sede di insorgenza è il midollo.

Le leucemie rappresentano il tumore che si osserva con maggiore frequenza nella popolazione infantile (30-35% di tutte le neoplasie). Le forme acute linfoblastiche (acute lymphoblastic leukemia, ALL) sono più frequenti (75% dei casi), mentre le mieloidi (acute lyeloid leucemia, AML) riguardano circa il 15% dei pazienti. Il restante 10% dei casi è costituito dalle leucemie mieloidi croniche e dalle mielodisplasie.

3.3 Imputazione dei valori mancanti

Il primo problema che si incontra nell'analisi di dati derivanti da *microarray* è l'esistenza di valori mancanti nella matrice dei dati.

Le ragioni per cui spesso vengono a mancare le misure relative ad un certo gene sono diverse e possono essere connesse agli strumenti utilizzati (presenza di polvere o graffi nei vetrini…), oppure al trattamento computazionale dell'immagine e al processo di trasformazione del segnale luminoso in dato numerico (insufficiente risoluzione, alterazione dell'immagine…).

Dato che, sfortunatamente, i pacchetti utilizzati per l'analisi non prevedono la presenza di dati mancanti, sorge il problema dell'imputazione degli stessi (*Troyaskaya et al, 2001*). Il metodo più indicato per il dataset di cui si dispone è il *Metodo d'imputazione k-nearest neighbors (knn)*. Tale metodo prevede l'utilizzo, nella stima dei dati mancanti, dei geni che più somigliano a quello che presenta il valore mancante (*Laursen, 1999 – Hastie et al, 1999*).

Questa tecnica prevede la partizione del *dataset* in due matrici: X^c e X^m; la prima contiene tutti e soli i geni che non presentano valori mancanti, la seconda raccoglie invece i geni che possiedono almeno un valore mancante. Indicato con x^* un qualsiasi vettore appartenente alla matrice X^m, l'algoritmo consiste nel calcolo della distanza tra x^* e tutti i geni contenuti nella matrice X^c usando solo le coordinate dei valori di x^* che non risultano mancanti. Successivamente il dato da imputare viene calcolato con una media pesata dei k geni che risultano più prossimi (o "vicini") ad x^*. La procedura d'imputazione sarà presentata, insieme a tutti gli altri comandi, nel paragrafo seguente.

3.4 Presentazione dei comandi in R

Il primo passo da compiere è caricare il dataset, composto da 4992 righe e 10 colonne. Le righe si riferiscono al tipo di gene, mentre le colonne al singolo soggetto. Quindi si ha:

```
> dati<- read.table("H:\\NormalizedData.txt", header=T,  sep="\t",
row.names=1)
> dati<-data.matrix(dati)
> dim(dati)
[1] 4992 10
```

Dato che il pacchetto *ismev* non prevede la presenza di valori mancanti, adottiamo l'imputazione *knn* (presentata nel paragrafo precedente) contenuta nella libreria *impute*:

```
> library(impute)
> dati.imputed<-impute.knn(dati,k=5)
```

Ponendo K=5, vengono presi i 5 punti più vicini, viene effettuata la media pesata delle loro coordinate e si ottengono, così, le nuove coordinate del punto mancante.

Il passo successivo consiste nel caricare la libreria ***ismev*** e procedere utilizzando le funzioni implementate in essa. Come già precedentemente accennato (vedi par. 2.2.3), si ricorda che esistono due procedure per la selezione della soglia:

- Tecnica esplorativa, basata sulla media della distribuzione generalizzata di Pareto.
- Ricerca della stabilità dei parametri, basata sulla stima di vari modelli in un rango di differenti soglie.

Per la prima, faremo riferimento alla funzione "mrl.plot", mentre per la seconda procedura utilizzeremo la funzione "gpd.fitrange".

mrl.plot

La sintassi completa della routine è:

```
> mrl.plot(data, umin = min(data), umax = max(data) - 0.1,
  conf = 0.95, nint = 100)
```

La funzione restituisce, in output, un grafico: Mean Residual Life plot; la riesaminazione degli argomenti sarà fornita di seguito, ma per ulteriori particolari si rimanda all'help fornito dal software R.

L'argomento `data` fa riferimento al vettore numerico dei dati, sul quale selezionare la soglia.
Gli argomenti `umin` e `umax` definiscono le soglie minime e massime su cui la funzione è calcolata.
L'argomento `conf` specifica il coefficiente per gli intervallo di confidenza rappresentati nel grafico.
L'argomento `nint` indica il numero di punti a cui è calcolato il grafico Mean residual Life Plot.

gpd.fitrange

La sintassi completa della routine è:

```
> gpd.fitrange(data, umin, umax, nint = 10, show = FALSE)
```

L'output della funzione è rappresentato da due grafici che mostrano stime di massima verosimiglianza e intervalli di confidenza dei parametri di scala e forma che sono prodotti sopra un range di soglie.
Gli argomenti `data, umin, umax, nint` hanno le stesse caratteristiche della funzione precedente. L'argomento `show`, se accompagnato da `True`, stampa i dettagli di ciascuna stima.

Considerando che il dataset è una matrice 4992 x 10, è possibile applicare il metodo della soglia in due diversi modi: si crea un vettore delle medie (effettuando la media per ogni riga), oppure si lavora separatamente su ogni singola colonna/soggetto.
Ulteriore considerazione deve essere fatta circa la natura dei dati: le osservazioni si basano sul rapporto "sogg. Sano/sogg. Malato" e assumono valori sia positivi che negativi, a seconda che il gene sia sottoespresso o sovraespresso nel sano; pertanto le applicazioni, per l'estrapolazione dei valori estremi, verranno effettuate su entrambe le code del processo, con l'obiettivo di identificare geni sregolati.

- *Vettore delle medie*

Tramite la funzione *apply* creiamo il vettore delle medie, xm:

```
> xm<-apply(dati.imputed, 1, mean)
```

Separiamo ora i positivi dai negativi e applichiamo le funzioni per il calcolo della

soglia (si ricorda che in questo paragrafo ci si limita alla sola presentazione dei comandi, mentre i grafici con i relativi commenti saranno oggetto del prossimo paragrafo):

- *threshold per i minimi*

```
> xmmin<- xm[xm<0]
> length(xmmin)
[1] 1956
> xm_min<- -xmmin
> library(ismev)
> mrl.plot(xm_min)
> gpd.fitrange(xm_min,0,4,nint=20)
```

- *threshold per i massimi*

```
> xm_max<- xm[xm>=0]
> length(xm_max)
[1] 3036
> mrl.plot(xm_max)
>gpd.fitrange(xm_max,0,2,nint=20)
```

- *Vettori dei singoli soggetti*

Come per il vettore delle medie, lavoriamo separatamente con i minimi (cambiati di segno) e con i massimi. I comandi per il primo soggetto sono:

- *threshold per i minimi*

```
sogg1<- dati.imputed[,1][dati.imputed[,1]<0]
soggnew1<- -sogg1
mrl.plot(soggnew1)
gpd.fitrange(soggnew1, 0, 3.5, nint=20)
```

- *threshold per i massimi*

```
sog1<- dati.imputed[,1][dati.imputed[,1]>=0]
mrl.plot(sog1)
gpd.fitrange(sog1,0,3.5,nint=20)
```

La procedura è analoga per i restanti soggetti.

3.5 Applicazione al dataset

Dopo aver presentato, nel paragrafo precedente, i comandi utili allo scopo, nel seguito verranno analizzati e commentati i relativi grafici, facendo riferimento al solo vettore delle medie, mentre per i vettori dei singoli soggetti, verranno presentati solo i grafici e una tabella con le relative soglie.

1. **Applicazione al vettore delle medie**

Applichiamo il metodo della soglia *u* al vettore delle medie, considerando i soli valori negativi (cambiati di segno): scelta che servirà a definire quante, tra tutte le osservazioni medie, saranno i valori estremi.

Il primo metodo, basato sull'analisi del Mean Residual Life Plot, rappresentato in figura 3.1, ci porta a scegliere $u = 2.2$, perché il grafico risulta curvo fino a lì e poi scende abbastanza linearmente.

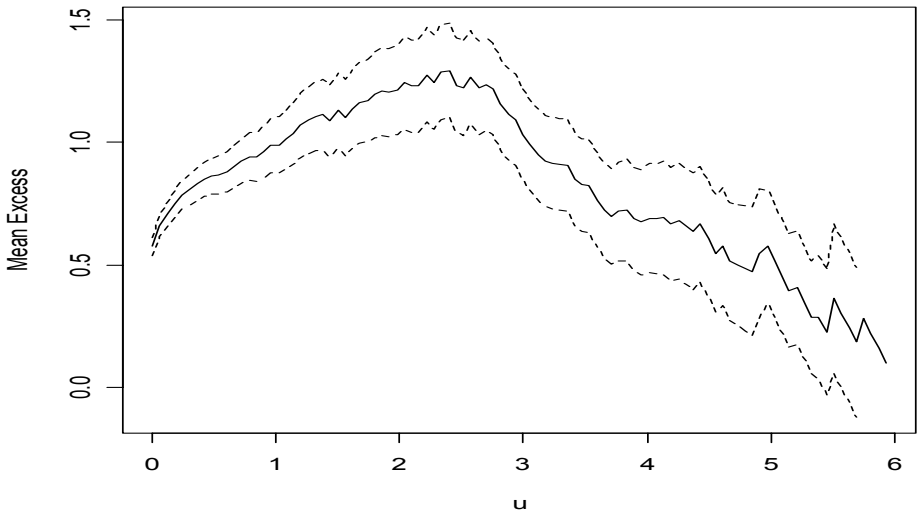


Figura 3.1: Mean Residual Life Plot per il vettore delle medie sui minimi.

Anche il secondo metodo, basato sui grafici di figura 3.2, conferma la scelta: dopo la soglia u=2.2 infatti, l'intervallo di confidenza tende ad aumentare. Questa scelta porta ad avere "94" valori eccedenti (sottoespressi) su un totale di 4992 osservazioni.

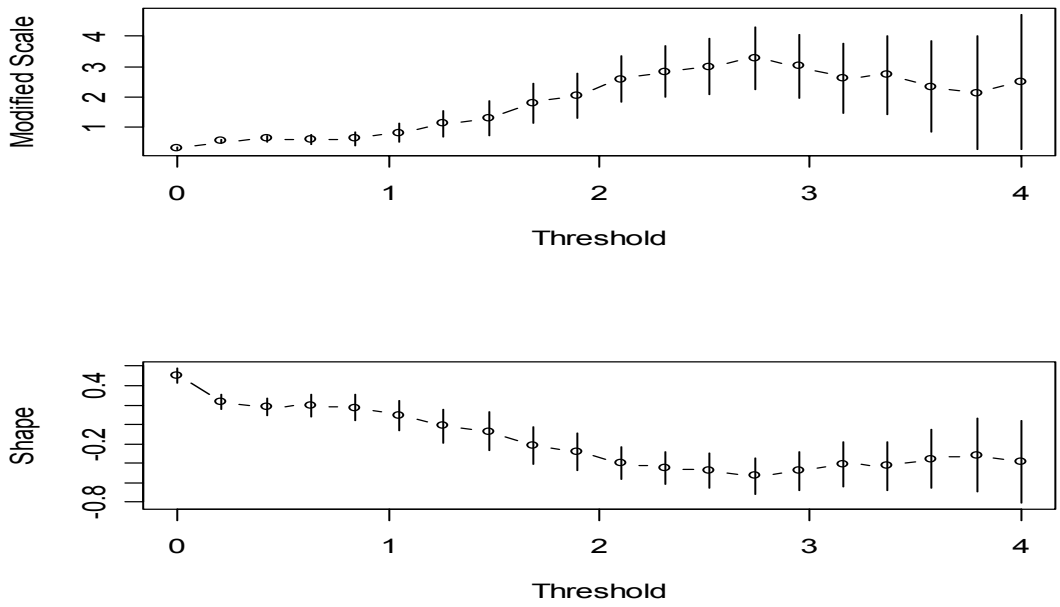


Fig. 3.2: Stime dei parametri contro la soglia per il vettore delle medie sui minimi

Per quanto concerne, invece, l'identificazione dei geni sovraespressi, in figura 3.3 è riportato il Mean Residual Life Plot: il grafico non risulta molto facile da interpretare.

La seconda procedura, illustrata in figura 3.4, ci suggerisce però di scegliere u = 1.4 visto che, da questa soglia in poi, l'intervallo di confidenza tende ad aumentare notevolmente.

Questa scelta porta ad avere "125" valori eccedenti su un totale di 4992.

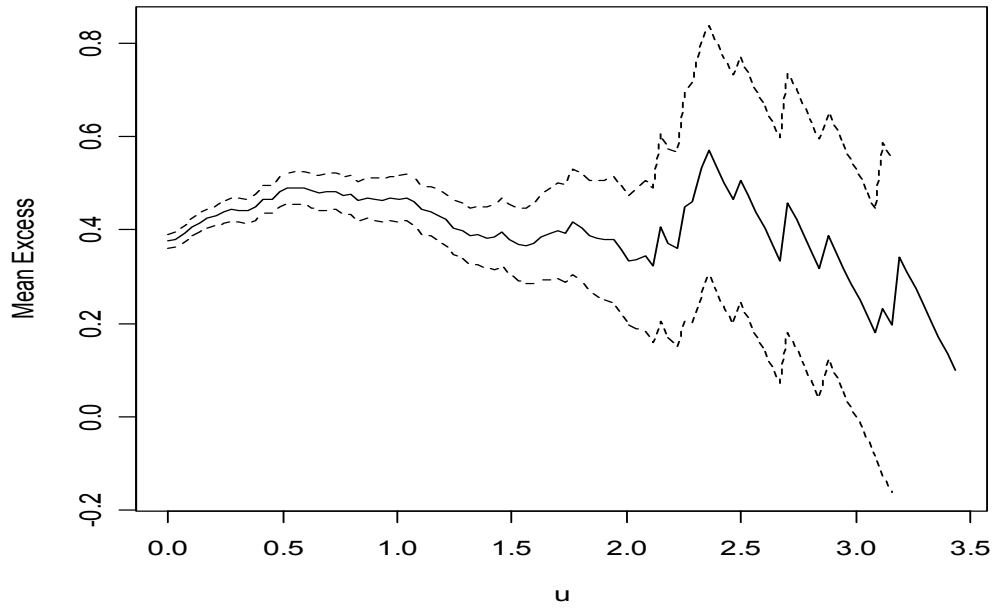


Figura 3.3: Mean Residual Life Plot per il vettore delle medie sui massimi.

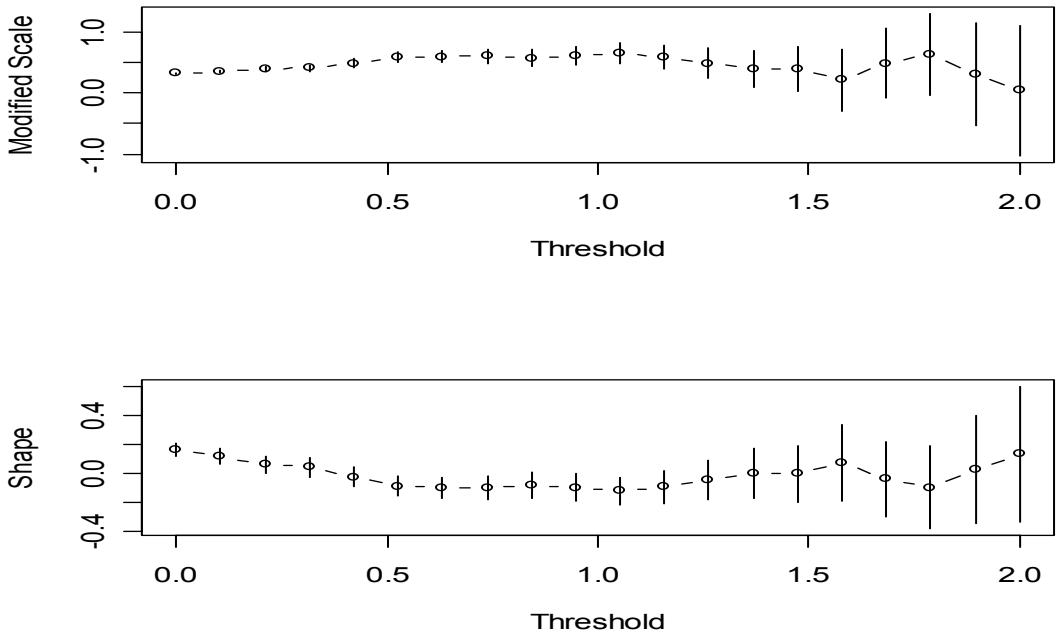


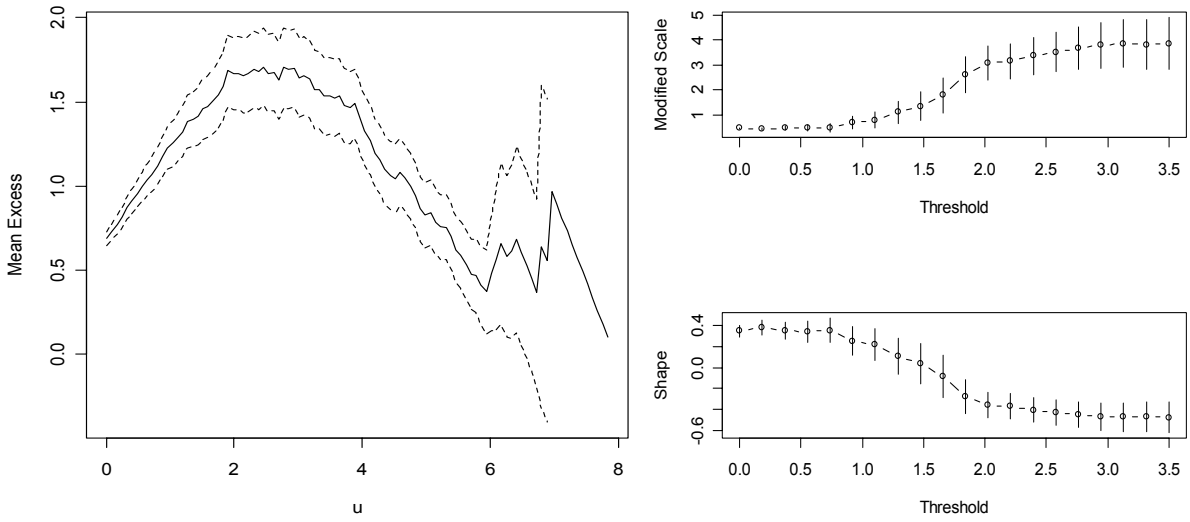
Figura 3.4:Stime dei parametri contro la soglia per il vettore delle medie sui massimi

2. Applicazione ai vettori dei singoli soggetti

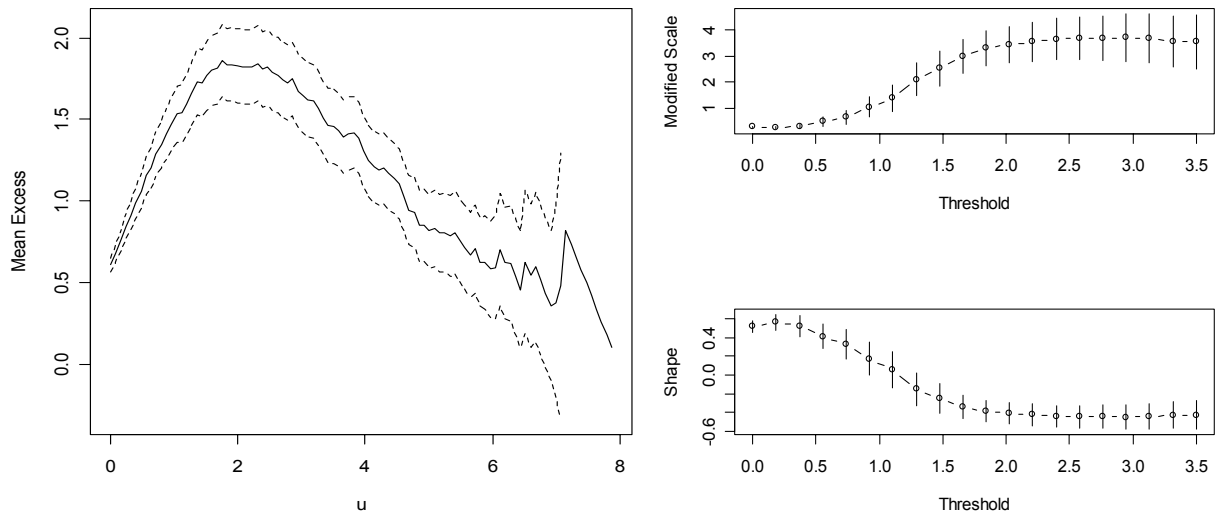
A scopo illustrativo, applicando ora il metodo della soglia *u* per ogni singolo soggetto, su entrambe le code del processo, si ottengono i seguenti grafici che saranno poi riassunti da una successiva tabella con le relative soglie:

→ *Threshold model per i minimi*

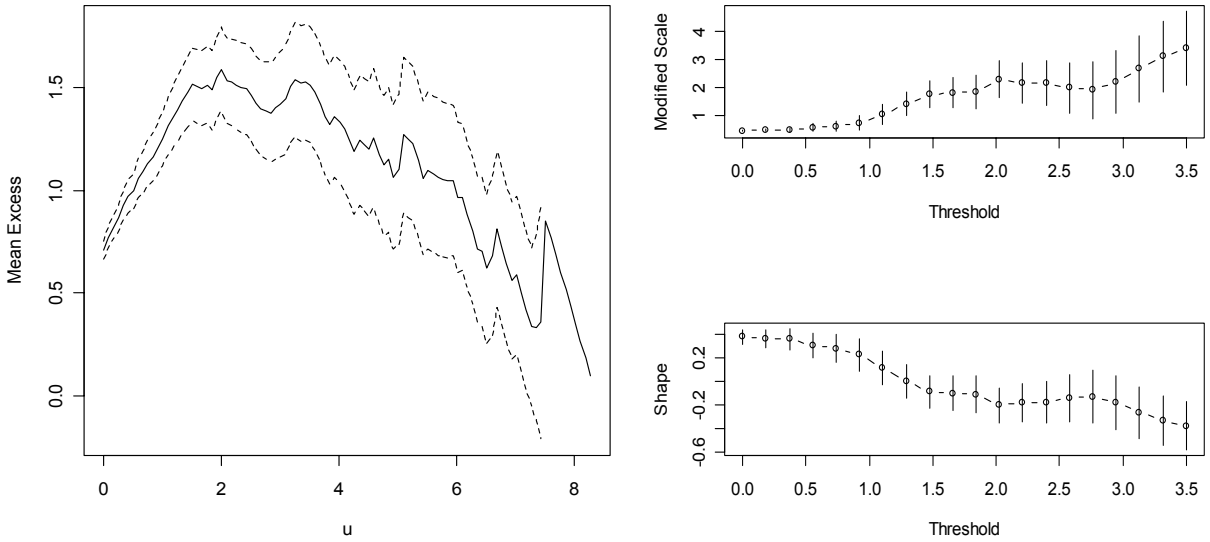
- Soggetto 1



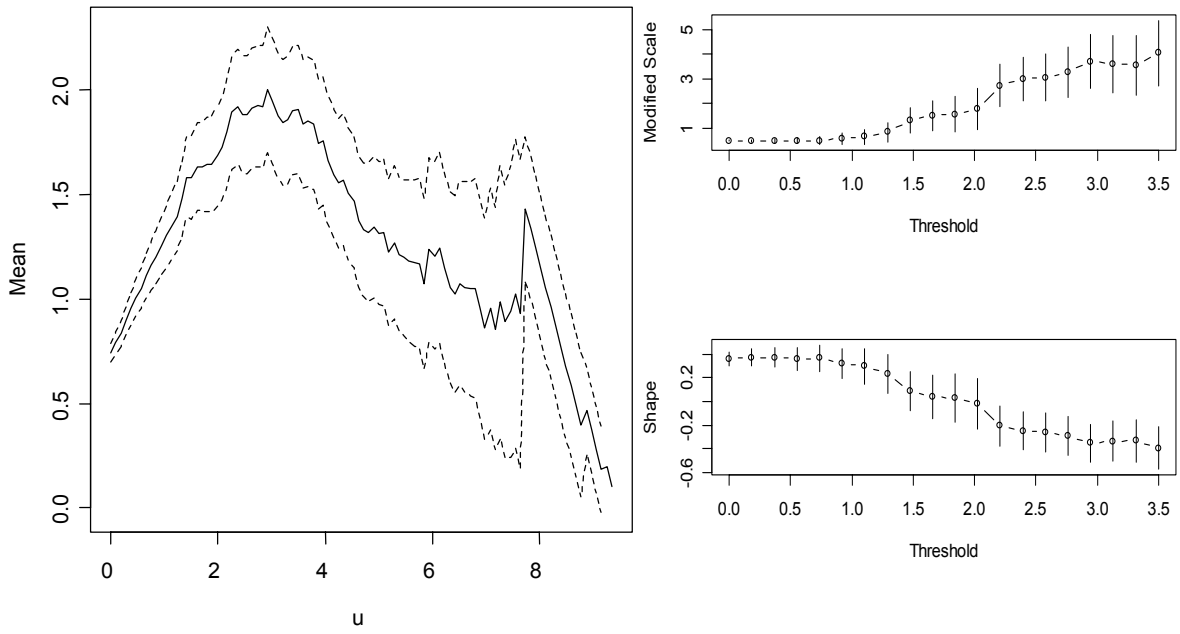
- Soggetto2



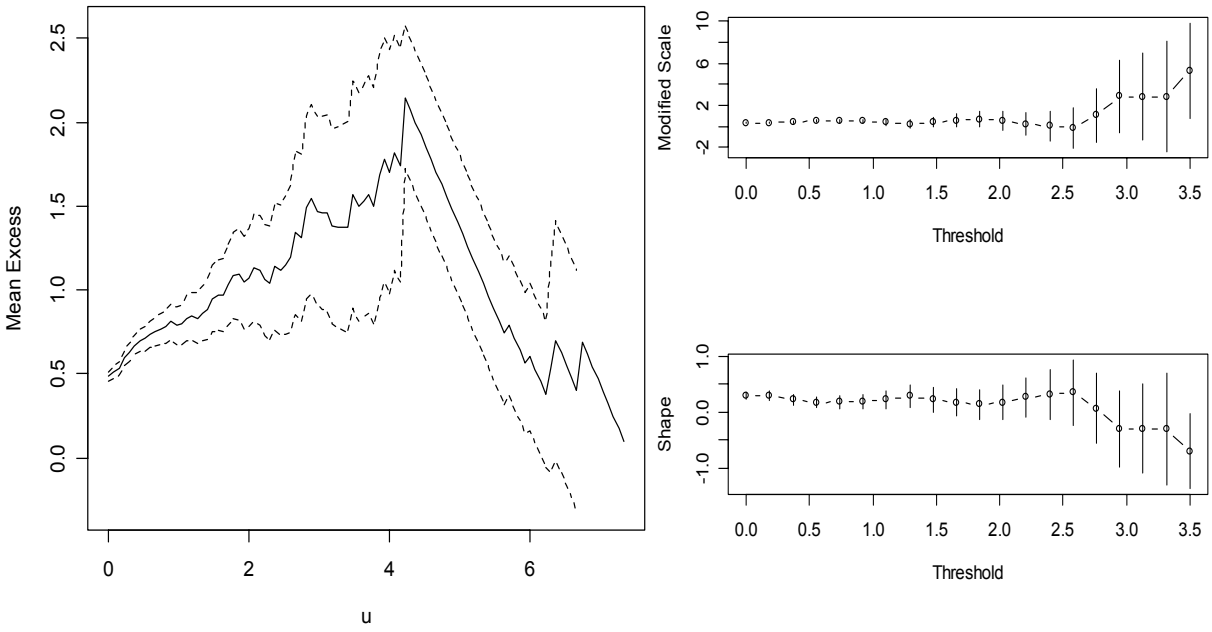
- Soggetto 3



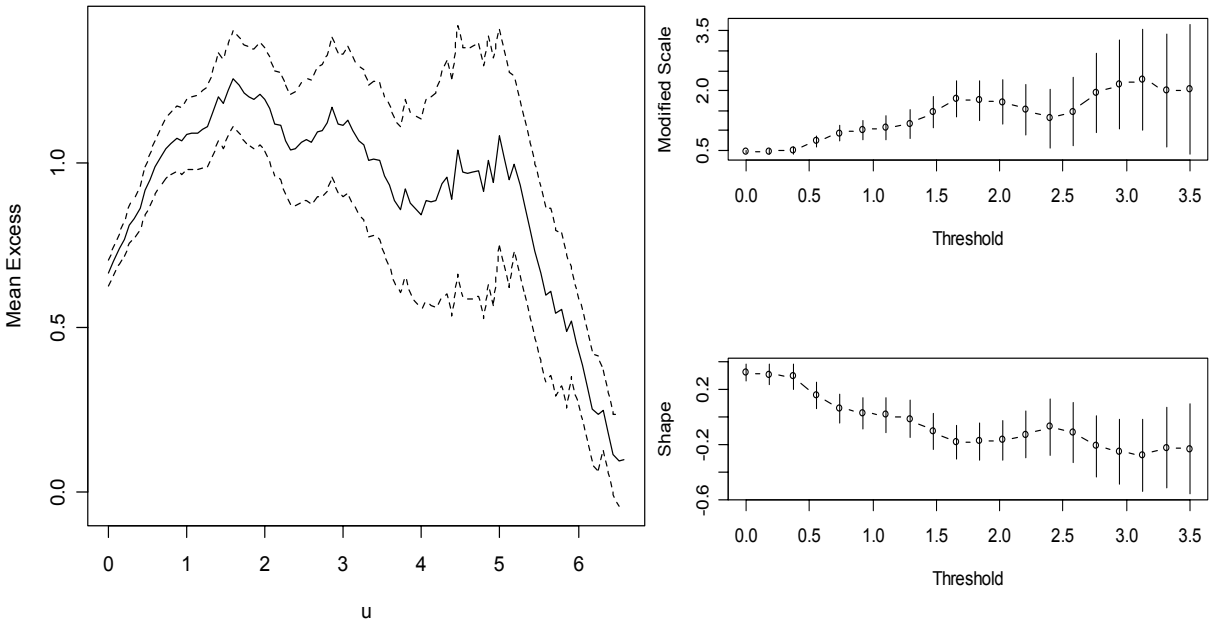
- Soggetto 4



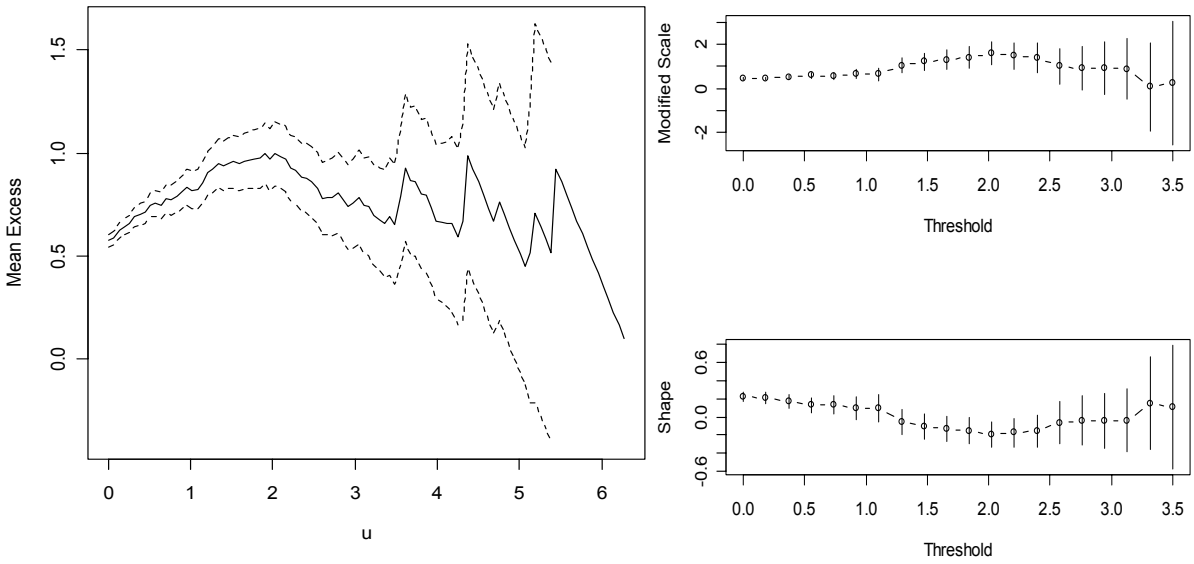
- Soggetto 5



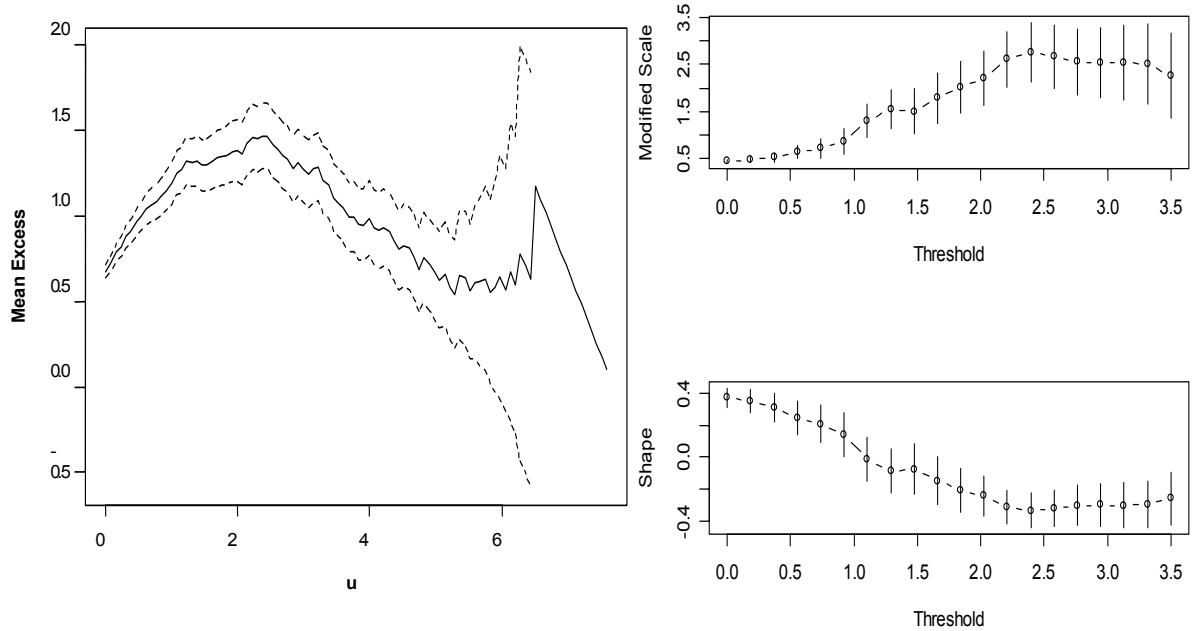
- Soggetto 6



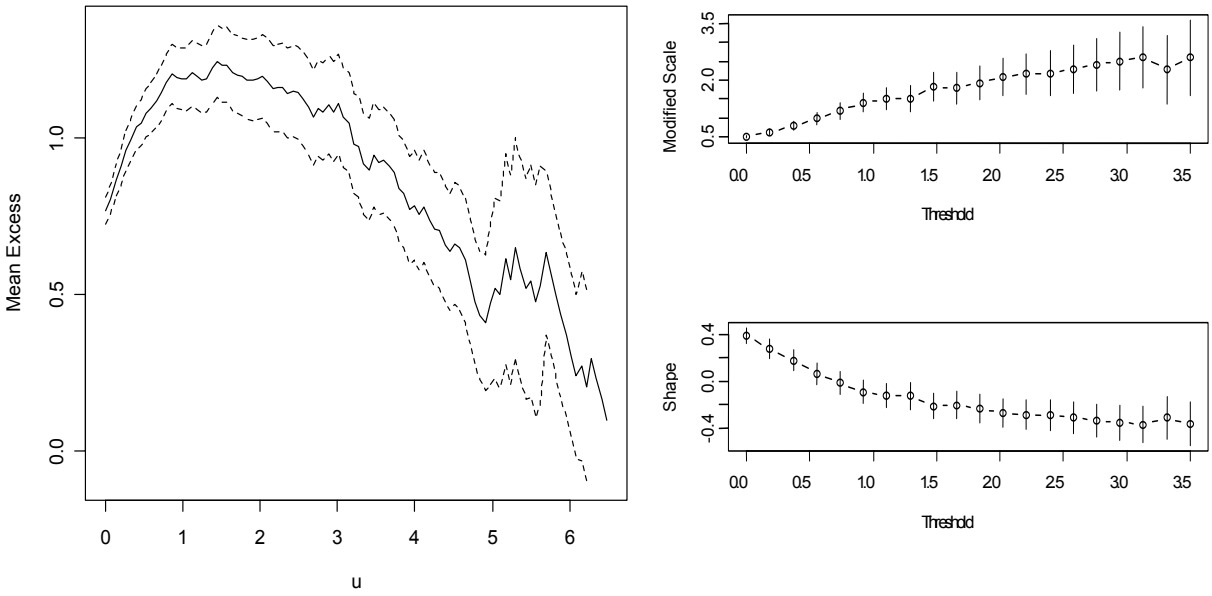
- Soggetto 7



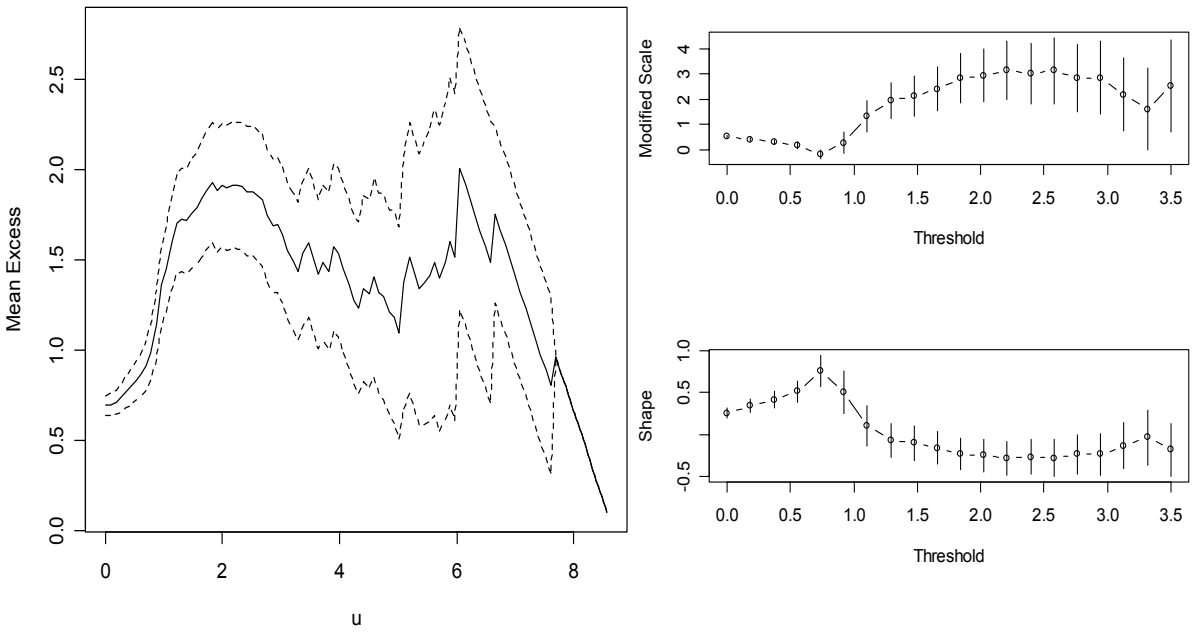
- Soggetto 8



- Soggetto 9

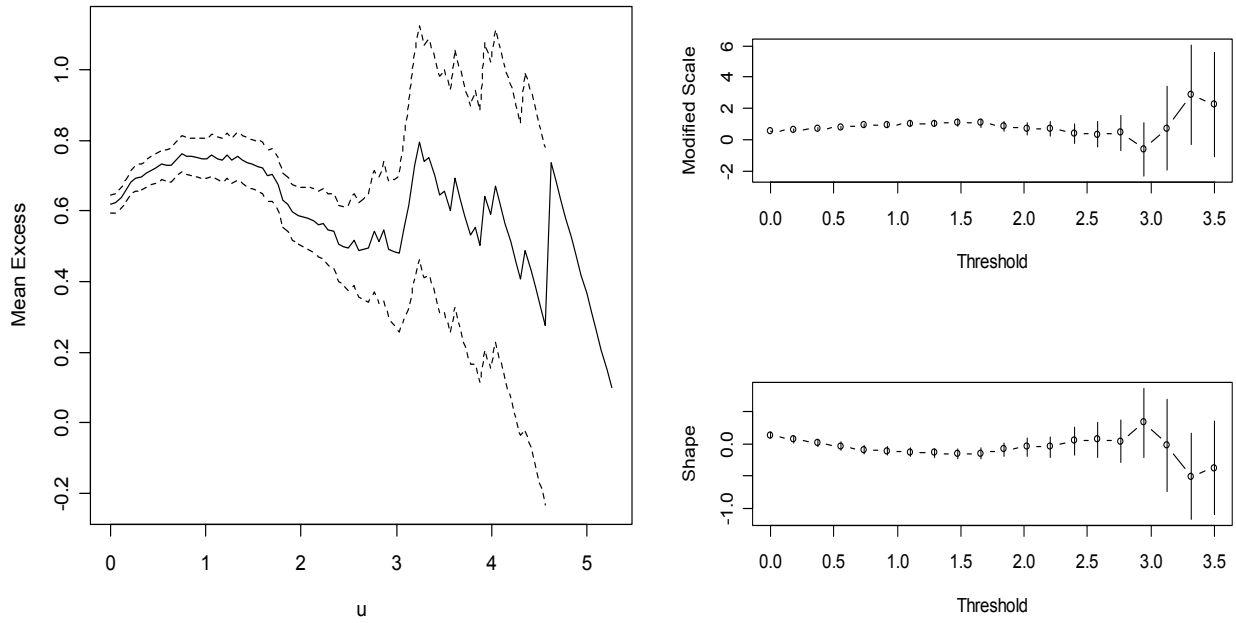


- Soggetto 10

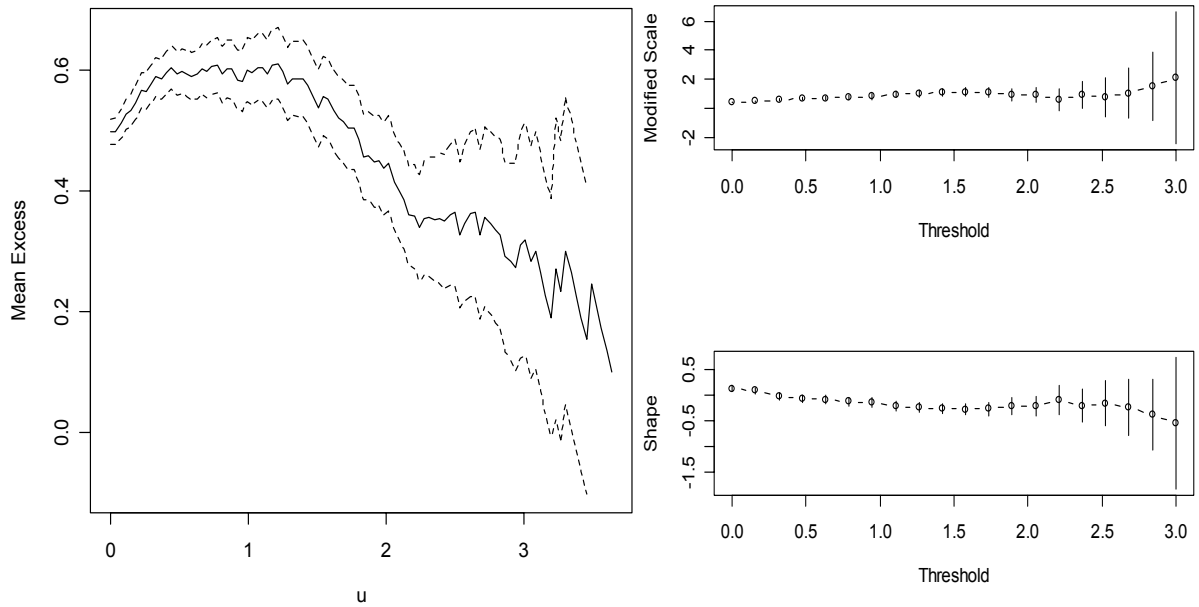


→ *Threshold model per i massimi*

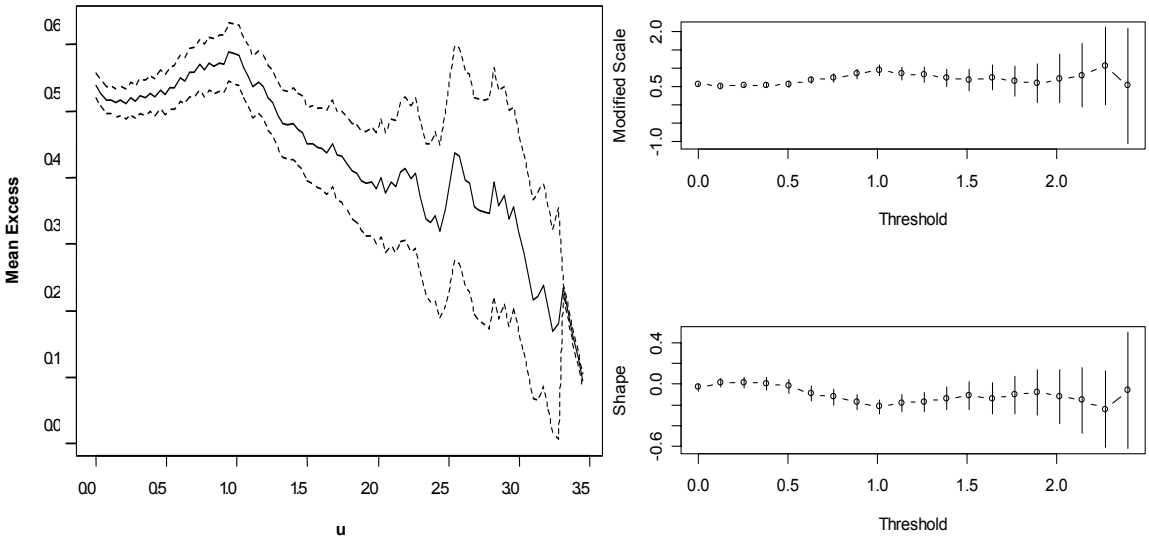
- Soggetto 1



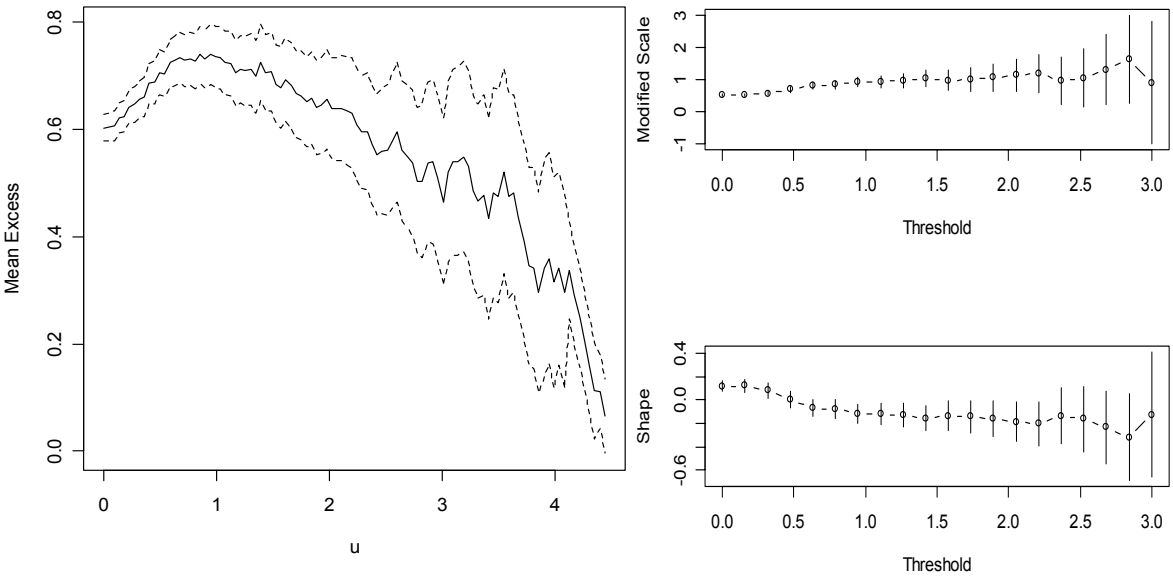
- Soggetto 2



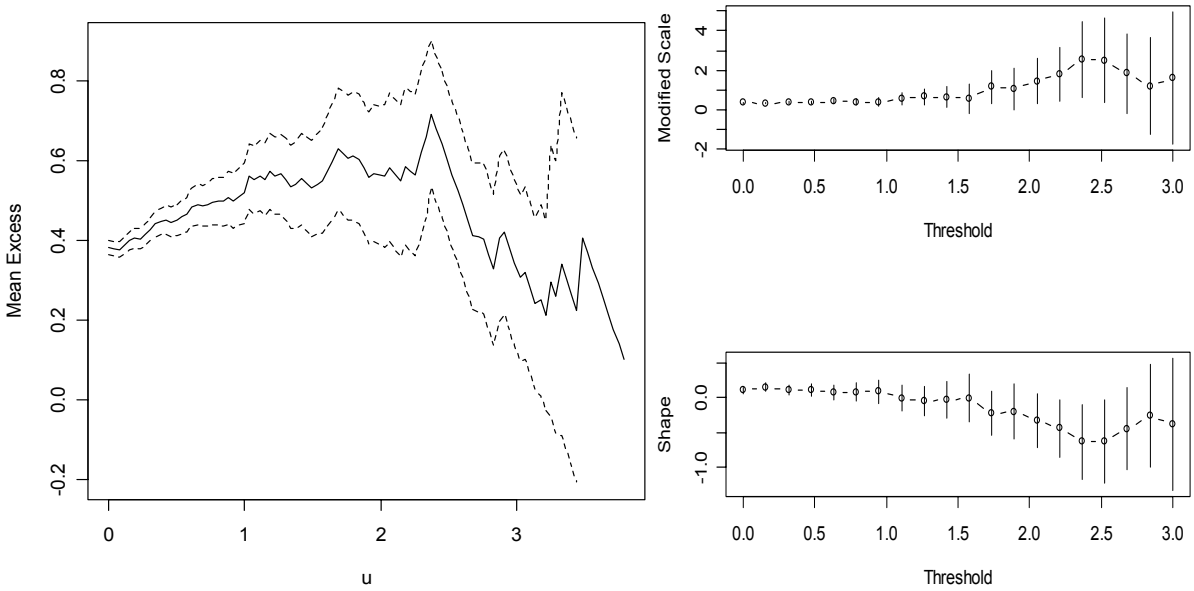
- Soggetto 3



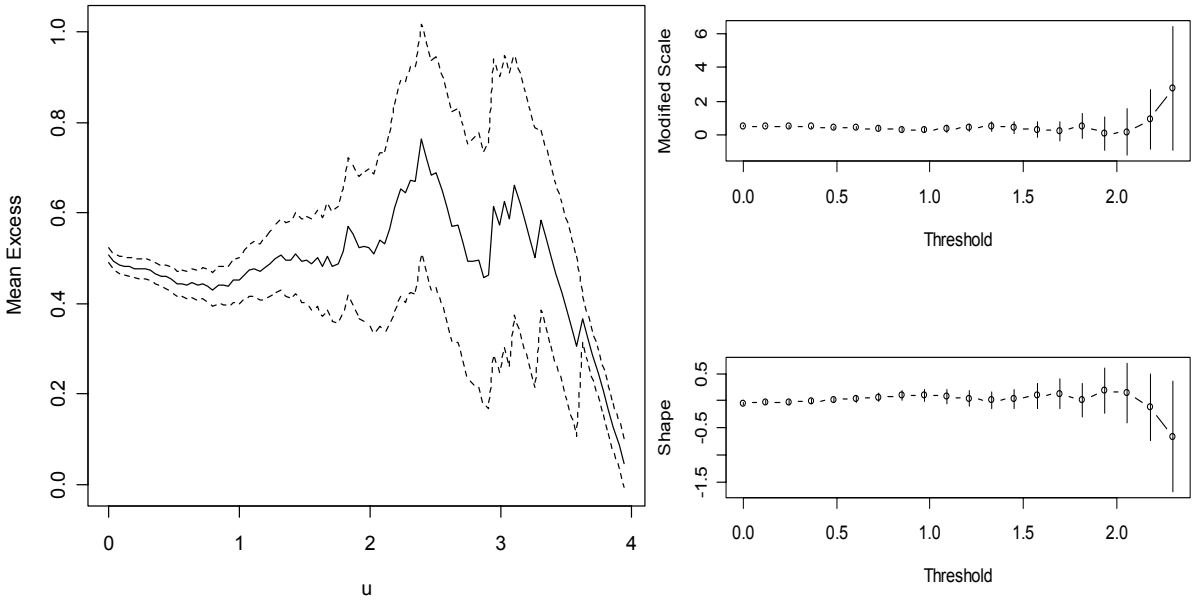
- Soggetto 4



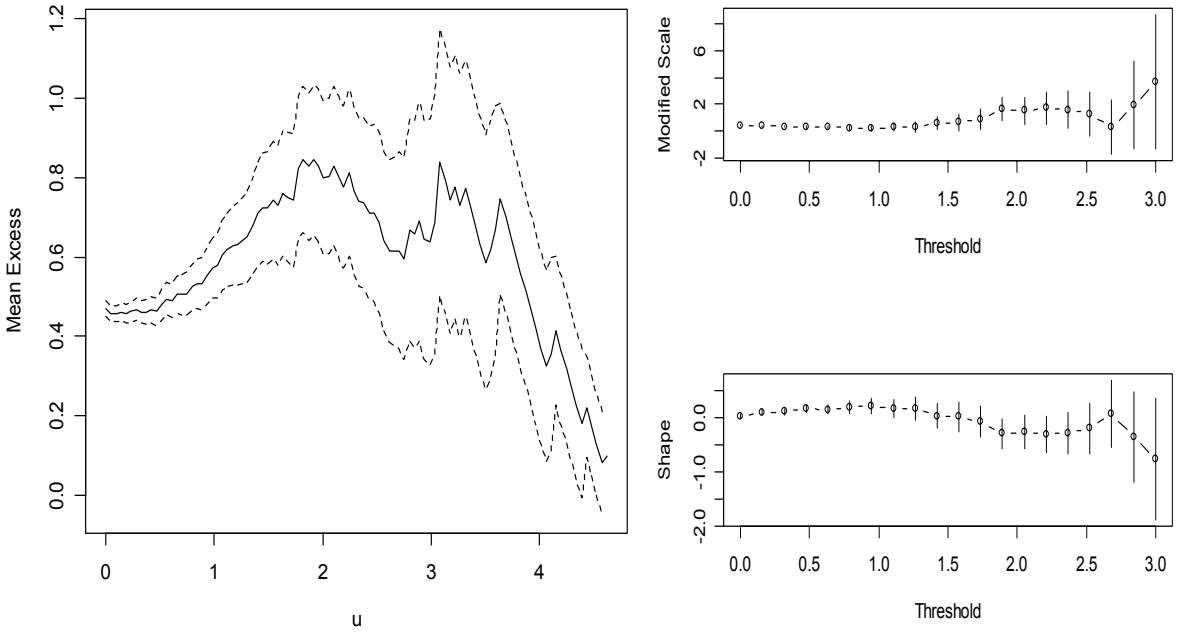
- Soggetto 5



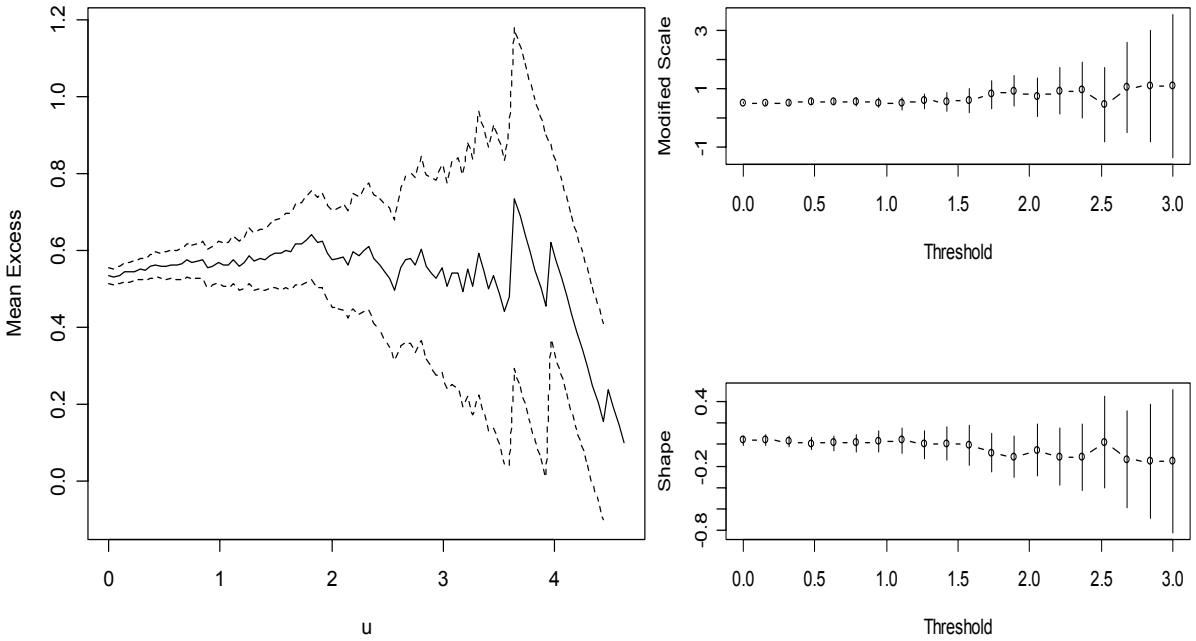
- Soggetto 6



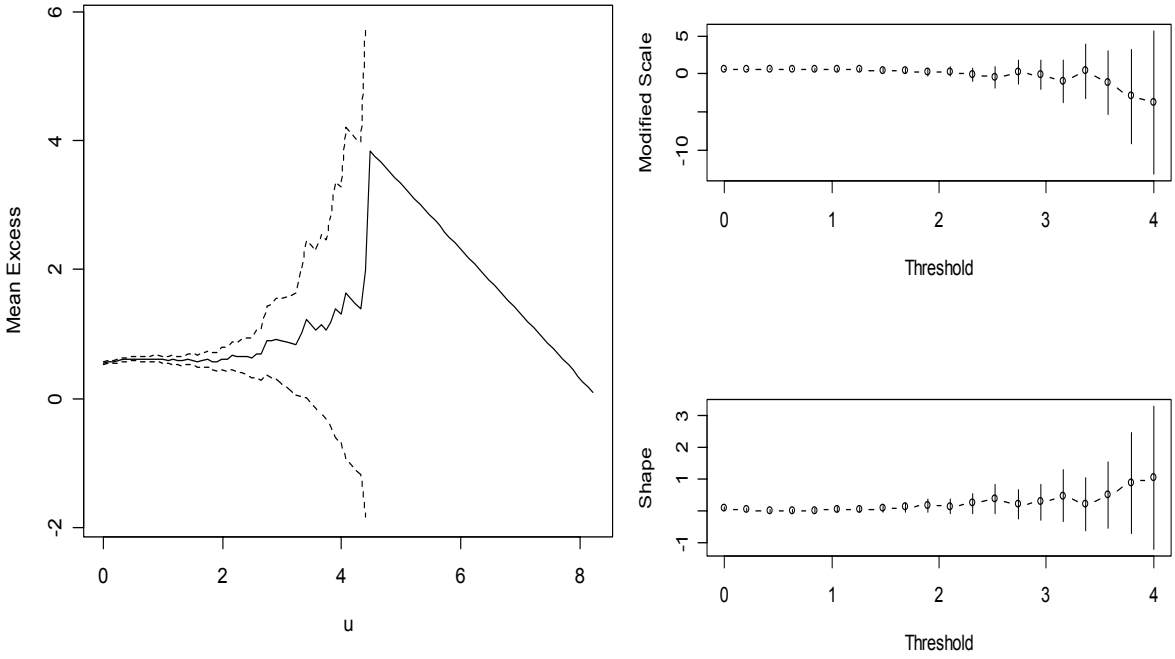
- Soggetto 7



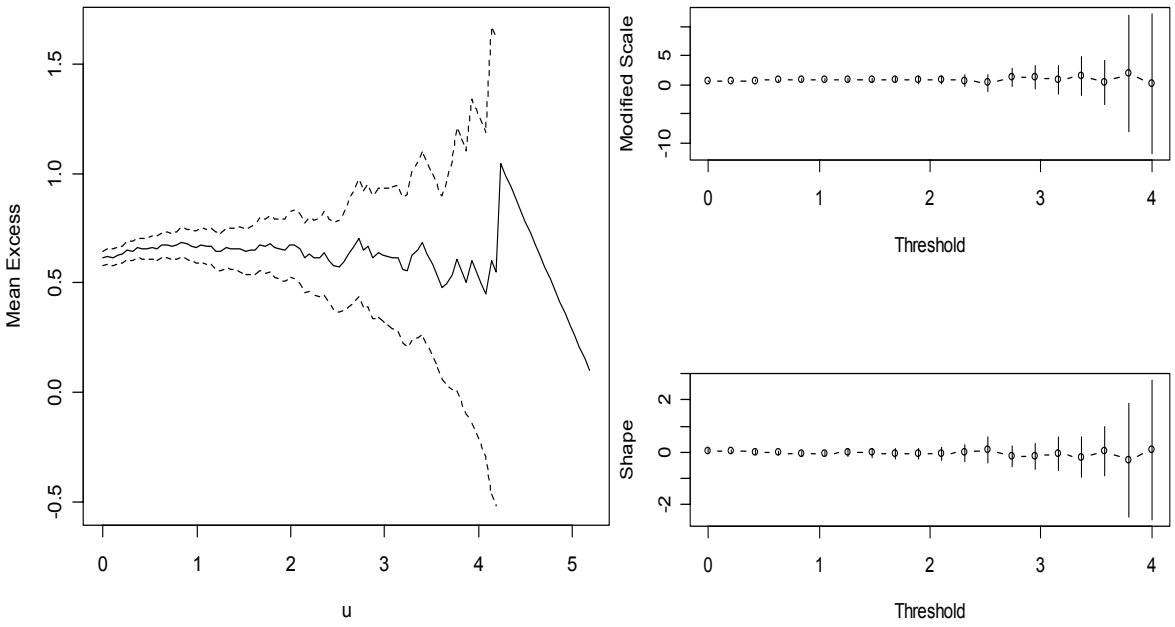
- Soggetto 8



- Soggetto 9



- Soggetto 10



Nella tabella seguente vengono riassunti i risultati derivanti dall'analisi dei grafici mostrati precedentemente; per ogni soggetto sono state scelte le relative soglie e contate le osservazioni estreme, identificando così sia i geni sottoespressi che sovraespressi.

In linea di massima, si nota una certa costanza sia nelle soglie che nelle relative osservazioni estreme: per i minimi le prime si aggirano intorno al valore –2 e le relative osservazioni estreme su 130-140, mentre per i massimi la soglia media è 1.6, con 148 osservazioni estreme medie.

Soggetto	**M I N I M I**		**M A S S I M I**	
	Soglia *u*	**Oss. Estreme**	**Soglia *u***	**Oss. Estreme**
1	*- 2,2*	147	2,1	143
2	*- 2,2*	137	1,7	135
3	*- 2*	174	1,3	283
4	*- 2,5*	126	1,8	188
5	*- 1,5*	129	1,2	128
6	*- 2*	162	1,5	119
7	*- 1,8*	134	1,4	126
8	*- 2,2*	138	1,6	143
9	*- 2,3*	173	1,8	131
10	*- 1,3*	130	1,8	89

Tab. 1: Soglie u e relative osservazioni estreme per massimi e minimi

Nel paragrafo successivo verrà presentato il Test SAM e i relativi risultati saranno, poi, confrontati con quelli ottenuti dal *Threshold Model* applicato al vettore delle medie.

3.6 Confronto con SAM

Si possono utilizzare diverse metodologia per identificare i geni differenzialmente espressi (*Dudoit et al, 2002*): è tuttavia necessario porre attenzione alle caratteristiche dei dati, nonchè alla loro distribuzione. Ad esempio, un approccio di tipo *Bayesiano*, basato sulle differenze e che ipotizza la normalità dei dati, non andrebbe bene in quanto il dataset a nostra disposizione si basa sul rapporto "sogg. sano/sogg. malato" e i singoli vettori della matrice non hanno distribuzione Normale (di tipo Gaussiano).

In questo paragrafo verrà analizzata la *Significance Analysis of microarray* (SAM) e i risultati saranno poi confrontati con quelli ottenuti precedentemente.

L'analisi si basa sulle differenze relative nell'espressione dei geni

$$d_j = \frac{\bar{y}_{j(1)} - \bar{y}_{j(2)}}{s_j + s_0}, \tag{3.1}$$

dove $\bar{y}_{j(1)}$ e $\bar{y}_{j(2)}$ indicano i livelli medi di espressione per il gene *j* nelle 2 condizioni rispettivamente, s_j lo scarto quadratico medio delle misure d'espressione ripetute - ossia la dispersione gene-specifica. È funzione degli scarti quadratici medi delle due popolazioni $s_{j(1)}$, $s_{j(2)}$

$$s_j = \sqrt{a\left\{\sum_{k1=1}^{K1}[y_{jk1} - \bar{y}_{j(1)}]^2 + \sum[y_{jk2} - \bar{y}_{j(2)}]^2\right\}},$$

dove $\sum_{k1=1}^{K1}$ e $\sum_{k2=1}^{K2}$ sono le sommatorie delle misure d'espressione nei due stati, rispettivamente, e $a=(1/K_1 + 1/K_2)/(K_1 + K_2 - 2)$.

Le differenze ottenute si confrontano successivamente con quelle attese, e, in base ad una soglia predefinita, si decide se il gene è da considerarsi differenzialmente espresso oppure no.

Per assumere che le d_j siano realizzazioni di variabili casuali identicamente distribuite, è necessario che la distribuzione da cui hanno origine sia indipendente dal livello di espressione dei geni. Questo non è sempre vero. Infatti, a bassi livelli d'espressione (quando la quantità di mRNA prodotto è esigua) la dispersione è ridotta; s_j assume, di conseguenza, valori molto piccoli, e la variabilità delle differenze diviene elevata. Per assicurare questa indipendenza (della varianza delle d_j dal livello di espressione) a denominatore è stata aggiunta una costante $s_0 > 0$, calcolata in modo da minimizzare il coefficiente di variazione delle d_j.

A questo punto, si pone il problema di associare un livello di significatività alla statistica utilizzata: Tusher *et al* (2001) propongono di eseguire i controlli tramite permutazioni dei dati. In pratica, si genera un numero elevato di controlli calcolando le differenze relative $d_{p(j)}$ da permutazioni dei dati. Le differenze vengono ordinate in base al loro valore ($d_{(1)}$ è la più grande) e per ogni gene si definiscono le differenze attese $d_{E(j)}$ come media delle permutazioni, $d_{E(j)} = \sum_p d_{p(j)} / N$, dove N = numero di permutazioni. Riportando queste quantità in un grafico, è immediato identificare i geni espressi.

In un diagramma di dispersione delle $d_{E(J)}$ contro le d_j, l'andamento dell'espressione genica sotto H_0 è lineare, lungo la bisettrice; le differenze osservate si disporranno più o meno regolarmente lungo tale retta. Fissato un intervallo entro cui, per ogni gene, ci si aspetta che sia compreso il valore della d_j sotto H_0, (ad esempio - Δ, +Δ) l'insieme di geni significativi è costituito da quelli la cui *dj* si colloca al di fuori degli estremi.

Infine, non è da escludere che questo metodo, come tutti gli altri, commetterà degli errori nell'identificazione dei geni, in una certa percentuale.

Nella tabella sottostante, sono stati riportati i risultati derivanti dall'applicazione di questo metodo: 120 geni positivi e 93 negativi, ordinati in base alla maggior espressione d_j .

120 Positive Significant Genes				***93 Negative Significant Genes***			
Gene Name	**Score (d)**	**Numerator (r)**	**Denomin. (s+s0)**	**Gene Name**	**Score (d)**	**Numerator (r)**	**Denomin. (s+s0)**
BL-003H09	16,224428	1,6309	0,1005213	2-030G10	-14,270978	-4,2246	0,2960274
2-014A04	14,780519	2,067	0,1398462	BL-003D02	-12,668475	-1,507	0,1189567
2-017H05	12,694758	1,857	0,1462808	BL-004A12	-12,535604	-1,671	0,1333003
BL-008H04	12,523941	1,2985	0,1036814	BL-006C05	-11,995382	-3,005	0,2505131
2-015D05	12,186908	2,1278	0,1745972	2-022A03	-11,261806	-1,306	0,1159672
2-002F08	11,728728	1,698	0,1447727	BL-009A09	-11,194422	-4,3377	0,3874876
2-018C10	11,474157	2,1116	0,1840309	2-026H05	-11,055478	-1,68	0,1519609
2-015F09	11,370759	1,401	0,1232108	BL-003C12	-10,941937	-1,948	0,1780306
BL-009D05	11,179183	1,399	0,1251433	BL-001G01	-10,675541	-3,4979	0,3276555
2-015B09	10,721964	2,973	0,2772813	2-036G05	-10,486547	-3,465	0,3304233
2-016C09	10,534365	2,102	0,1995374	2-011H04	-10,341964	-3,2038	0,3097864
2-029E03	10,507135	2,1108	0,2008921	2-010C12	-10,244296	-2,918	0,2848415
BL-010A12	10,446326	1,0774	0,1031367	2-002C08	-10,089914	-1,952	0,1934605
BL-004B05	10,2355	1,752	0,171169	BL-003F02	-10,080581	-3,253	0,3226997
2-036H05	10,028404	1,3769	0,1373	BL-008H06	-10,067031	-1,6972	0,1685899
2-010C10	9,8908233	2,187	0,221114	BL-004B09	-10,024573	-4,309	0,4298438
2-029D03	9,7480898	0,708	0,0726296	2-031D11	-9,8655333	-4,157	0,421366
2-029G09	9,719662	1,869	0,1922906	BL-002B05	-9,8605498	-3,3104	0,3357216
2-006E11	9,6905535	2,643	0,2727398	2-032A08	-9,7147753	-2,29	0,2357234
2-010F01	9,6512127	1,151	0,1192596	2-020A07	-9,5734519	-1,1184	0,1168231
2-029E07	9,513003	3,3181	0,3487963	2-006G11	-9,5145949	-2,0058	0,210813
2-007A04	9,4124431	1,227	0,1303594	2-028G09	-9,4694058	-1,067	0,1126787
BL-005D01	9,3708487	1,18	0,1259224	2-035F10	-9,3321991	-2,5869	0,2772015
2-018A04	9,3627836	1,291	0,1378863	2-013B05	-9,3217896	-1,195	0,1281943
2-018E06	9,3270896	2,059	0,2207548	BL-004C05	-9,3133137	-0,9712	0,1042808
2-017A05	9,2655233	0,799	0,0862337	2-032F09	-9,2129101	-5,3949	0,5855804
2-011D01	9,1985198	1,3891	0,1510134	2-022A06	-9,0056933	-0,823	0,0913866
2-024A01	8,90143	1,456	0,1635692	2-004B08	-8,987434	-0,891	0,0991384
2-010A12	8,872338	3,025	0,3409473	BL-002G03	-8,9259554	-3,9929	0,4473359
2-002F09	8,824585	1,4818	0,1679172	2-026A09	-8,8882924	-0,747	0,0840431
2-001C12	8,8048711	2,102	0,2387315	BL-001H09	-8,7919483	-4,188	0,4763449
2-027C04	8,7868938	1,365	0,155345	2-005F12	-8,6688066	-1,106	0,1275839
2-035D12	8,7479694	1,738	0,1986747	2-021B05	-8,6086127	-1,293	0,1501984
2-010C07	8,6968835	0,964	0,1108443	2-005F11	-8,5479117	-3,429	0,4011506
2-020G11	8,640985	1,3608	0,157482	2-014C04	-8,5049936	-4,073	0,4788951
BL-010B08	8,6250807	1,827	0,2118241	BL-005B08	-8,4119382	-2,2233	0,2643029
2-002A12	8,4920873	2,0373	0,2399057	2-017C01	-8,2856973	-1,5684	0,18929
2-018E11	8,4858471	1,753	0,2065793	2-012F01	-8,2834402	-0,69	0,0832987
2-040F06	8,4396813	2,1325	0,2526754	2-014H03	-8,2792239	-0,8	0,0966274
2-017H06	8,4323993	0,987	0,1170485	2-024C05	-8,2267379	-1,461	0,1775917
2-002B05	8,3465214	1,3568	0,1625587	2-005B01	-8,1913976	-2,1117	0,2577948
2-026G09	8,3114841	2,2235	0,2675214	2-038F12	-8,0889678	-3,9611	0,4896916
2-018D04	8,2642081	1,542	0,1865877	BL-004D11	-8,0882262	-3,837	0,4743933
2-002F05	8,242752	2,527	0,3065724	2-003F04	-8,0813811	-1,2153	0,1503827
2-002H11	8,2013418	1,6561	0,2019304	BL-006A12	-8,0661937	-3,5595	0,4412862
2-011B09	8,18422	1,312	0,1603085	2-006C09	-8,0434766	-1,098	0,1365081
BL-008D01	8,1502009	2,406	0,2952074	2-003A12	-8,0408103	-1,8865	0,2346157
2-029C03	8,1305305	1,2937	0,1591163	2-005E07	-8,0262049	-1,9379	0,2414466
2-010G06	8,0912248	1,02	0,1260625	BL-003D04	-8,0206294	-3,7749	0,4706488

BL-004B08	8,0851926	2,393	0,2959732	2-012E09	-7,9952661	-0,648	0,081048
2-001H09	8,0782886	1,1721	0,1450926	2-006D09	-7,9945358	-2,501	0,3128387
2-018H05	8,0594756	1,736	0,2153986	BL-010E10	-7,9701506	-1,336	0,1676254
2-030D01	8,0306886	1,715	0,2135558	BL-001B05	-7,9092988	-1,982	0,2505911
2-002C10	8,0233348	2,131	0,2656003	BL-008B07	-7,8845465	-1,5441	0,1958388
2-029D01	8,007323	1,563	0,1951963	2-026F06	-7,8844326	-1,7487	0,2217915
2-018E08	7,9973702	1,6323	0,2041046	2-012G06	-7,8824566	-1,2437	0,1577808
2-030C04	7,9287976	1,234	0,1556352	BL-003E05	-7,8774315	-0,604	0,0766747
2-010D02	7,9252377	1,221	0,1540648	BL-008A04	-7,8752331	-1,323	0,167995
2-023C01	7,8650928	0,975	0,1239655	BL-001G07	-7,8538683	-1,03	0,1311456
2-017D08	7,8476235	0,906	0,115449	BL-005H12	-7,8264275	-5,153	0,6584102
2-002D08	7,794879	0,915	0,1173848	2-038A04	-7,8096168	-2,115	0,2708199
2-030C06	7,7937776	1,072	0,1375456	BL-003H11	-7,7861041	-2,839	0,3646239
BL-008C11	7,738469	1,1695	0,1511281	2-037G07	-7,5325396	-2,7549	0,3657332
BL-001C06	7,6997098	0,9752	0,1266541	BL-002E09	-7,5288015	-0,8372	0,1111996
2-023B05	7,6858374	1,315	0,1710939	2-006G06	-7,4793707	-2,5183	0,3366995
2-002A01	7,6335692	1,473	0,1929635	2-010H04	-7,4341418	-1,019	0,1370703
2-035D08	7,6296896	1,4868	0,1948703	2-039D02	-7,410758	-1,5252	0,2058089
BL-001E11	7,6114246	2,0056	0,2634986	2-018A11	-7,4070119	-1,9571	0,2642226
2-020E05	7,6064134	0,634	0,0833507	2-005F09	-7,3390008	-2,3715	0,3231366
BL-002C04	7,6045331	1,503	0,1976453	2-011E04	-7,2880852	-2,0768	0,2849582
2-013C04	7,5622477	1,648	0,2179246	BL-001G05	-7,270845	-1,333	0,1833349
2-018C08	7,5520187	0,9481	0,1255426	2-019A08	-7,2584485	-2,246	0,3094325
BL-003F10	7,5272339	2,896	0,3847363	BL-003G12	-7,2460853	-0,9636	0,1329821
BL-010E03	7,4999037	0,962	0,1282683	2-033C08	-7,2384683	-1,205	0,1664717
BL-007A09	7,4882836	1,4638	0,1954787	BL-004B11	-7,217325	-4,2065	0,5828337
2-002B03	7,453758	1,4464	0,1940498	2-003C09	-7,1994279	-3,7964	0,5273197
2-020A01	7,4466262	0,806	0,1082369	BL-008G08	-7,1596247	-1,553	0,2169108
2-020C01	7,4369101	1,437	0,1932254	BL-008G07	-7,1434275	-1,9277	0,2698565
2-025E02	7,3958237	1,014	0,1371044	BL-005A06	-7,0946912	-1,2977	0,1829114
2-018G06	7,3908873	0,9218	0,1247212	BL-002D08	-7,0799677	-3,5752	0,5049741
2-024G01	7,3755112	1,7657	0,2394004	2-014E04	-7,0716113	-2,7983	0,395709
2-002E10	7,3629602	0,777	0,1055282	BL-008H08	-7,0346909	-3,894	0,5535424
2-010G02	7,3285343	1,6147	0,2203306	BL-001H08	-7,0345286	-2,492	0,3542526
2-018G05	7,2967315	0,9731	0,1333611	BL-003A01	-7,0158213	-0,95	0,1354082
2-015E08	7,2717526	1,2564	0,1727782	BL-004D06	-7,0092142	-2,175	0,3103058
2-011F05	7,2700082	1,525	0,2097659	2-005B07	-6,9670447	-1,0715	0,1537955
2-035E04	7,2676292	1,0075	0,1386284	BL-002C11	-6,9534053	-1,584	0,2278021
2-002A02	7,2397656	0,955	0,1319103	2-023A12	-6,9007832	-0,628	0,0910042
2-026E06	7,2353253	1,441	0,1991617	BL-009E01	-6,8822075	-1,2501	0,1816423
2-018A12	7,2338623	0,872	0,1205442	BL-003H06	-6,8720214	-1,322	0,1923743
2-002C07	7,2321133	1,447	0,2000798	BL-001B09	-6,8552068	-1,512	0,2205623
2-029G05	7,2311782	1,547	0,2139347	2-012G01	-6,8304749	-0,355	0,051973
2-029B04	7,2288298	2,1348	0,2953175	2-012E03	-6,8243824	-1,2103	0,1773494
2-002F12	7,2140425	0,965	0,1337669				
2-018B03	7,1550041	1,289	0,1801536				
2-035F02	7,1095932	1,678	0,2360191				
2-034D06	7,0675487	1,059	0,1498398				
2-001D01	7,0673446	0,7652	0,1082726				
2-010B09	7,0364822	1,1707	0,1663757				
BL-006D11	6,9788291	1,8322	0,2625369				
2-001F10	6,962066	1,6293	0,2340254				
2-035E05	6,9520496	0,825	0,11867				
2-027D01	6,9485194	0,508	0,0731091				

2-023H04	6,9404052	0,393	0,0566249				
BL-003B04	6,9311162	1,064	0,1535106				
2-018A06	6,896996	0,755	0,1094679				
2-010F08	6,8951474	1,2785	0,1854203				
2-001D07	6,8327659	1,9176	0,2806477				
2-001C03	6,8035006	0,8744	0,1285221				
2-022B11	6,7933549	1,137	0,1673694				
2-017E10	6,7608133	1,847	0,273192				
2-032A04	6,7582964	1,261	0,1865855				
2-017E08	6,7577908	1,422	0,2104238				
2-019A02	6,7524599	0,4081	0,0604372				
2-032C01	6,7424219	1,382	0,2049709				
2-007G10	6,7360711	2,156	0,3200679				
2-016H06	6,7311546	0,642	0,0953774				
2-029C08	6,7308092	1,415	0,2102273				
BL-006E09	6,7282257	2,403	0,3571521				
2-018H06	6,7253713	0,916	0,1362007				

Tab. 2: Risultati ottenuti con il test SAM

Interpretazione:

La prima colonna si riferisce al nome dei geni identificati, la seconda fa riferimento al valore dell'espressione d_j, ottenuta dal rapporto tra la terza e quarta colonna, come risulta dalla (3.1):

$$d_j = \frac{\bar{y}_{j(1)} - \bar{y}_{j(2)}}{s_j + s_0}.$$

In riguardo al Modello della Soglia, invece, si ottengono i seguenti risultati, rappresentati in Tabella 3:

125 Positive Significant Genes: u=1.4				94 Negative Significant Genes: u=2.2			
Cumuni al Sam	media	Non comuni	media	Comuni al Sam	media	Non comuni	media
2-001C12	2,111111	2-001C06	1,521528	2-003C09	4,027222	2-001A10	3,520407
2-001D07	2,029556	2-001C09	1,400611	2-005F11	3,645556	2-003C10	2,662037
2-001F10	1,661333	2-001H05	1,546167	2-006D09	2,572917	2-005C12	2,205435
2-002A01	1,616111	2-002A04	1,415556	2-006G06	2,663889	2-005E01	3,678333
2-002A12	1,738407	2-002B06	1,662083	2-006G11	2,215741	2-005F05	2,334648
2-002B05	1,582500	2-002B08	2,680722	2-010C12	2,367556	2-005F07	3,138639
2-002C07	1,590556	2-002C01	1,624213	2-011E04	2,321000	2-005F08	4,051389
2-002C10	2,047000	2-002D04	1,757667	2-014C04	3,996981	2-005F10	4,388667
2-002F05	2,239444	2-002E05	1,560463	2-014E04	2,363000	2-006G07	2,538889
2-002F08	1,776111	2-002E09	2,115639	2-030G10	3,810556	2-006H02	3,025370
2-002F09	1,508444	2-002G04	2,124889	.2-031D11	3,731667	2-010B03	3,322389
2-002H11	1,561944	2-005D11	1,715639	2-032A08	2,533611	2-010D05	3,057500
2-007G10	2,223889	2-007B10	1,904778	2-032F09	6,031322	2-011C06	3,244852
2-010A12	2,845556	2-009E03	1,417222	2-035F10	2,663889	2-013F01	4,917778
2-010C10	1,752361	2-012A12	1,666167	2-036G05	3,176583	2-014B11	4,135889
2-010F08	1,504528	2-014H10	1,410000	2-038F12	4,119213	2-014E07	4,215417
2-010G02	1,594222	2-015D05	2,127917	BL-001H08	2,688417	2-014F06	5,465380
2-011F05	1,552778	2-015D07	1,426778	BL-001H09	3,816111	2-017B12	2,328843
2-014A04	2,011667	2-015E10	1,944333	BL-002B05	3,183500	2-018B01	3,236852
2-015B09	3,095833	2-016C02	1,436111	BL-002D08	3,267222	2-018D05	2,471250
2-016C09	2,062778	2-016F05	1,667639	BL-003D04	3,302731	2-021A10	2,258917
2-017E08	1,431111	2-016G01	1,657778	BL-003F02	3,347296	2-029A01	2,204556
2-017E10	1,848889	2-016H09	2,138972	BL-003H11	2,332778	2-029A02	2,527885
2-018C10	2,271407	2-017A09	2,057667	BL-004B09	4,214444	2-031F09	2,343979
2-018D04	1,775000	2-017G09	1,477065	BL-004B11	4,349444	2-031H01	2,495907
2-018E06	2,035556	2-017H05	1,689444	BL-004D11	3,716204	2-033E05	5,719054
2-018E08	1,849111	2-018C01	1,465556	BL-006A12	3,698241	2-033E08	2,563935
2-018E11	1,833889	2-018D09	2,345722	BL-006C05	2,891111	2-033E09	2,344414
2-018H05	1,725000	2-020C09	1,982546	BL-008H08	5,377361	2-033F07	3,501074
2-020C01	1,571000	2-020H07	1,488778	BL-009A09	4,755926	2-034E08	2,853889
2-020G11	1,897444	2-021E09	1,907556			2-035B07	2,300000
2-023B05	1,605111	2-021H11	2,208222			2-036D08	2,489685
2-026E06	1,731667	2-023C07	1,869583			2-037A10	3,917361
2-027C04	1,422593	2-023E11	1,439833			2-037C09	2,877556
2-029B04	2,238444	2-024F09	1,444944			2-038C05	2,702361
2-029C08	1,501111	2-024G01	1,647778			2-038F07	2,741843
2-029D01	1,608472	2-027E09	1,550333			2-040C07	4,010648
2-029E03	2,321389	2-027F01	1,481667			2-040E10	4,376093
2-029E07	3,531583	2-030G12	1,431889			BL-001A01	4,847500
2-029G05	1,755000	2-031F07	1,431852			BL-001C07	2,226333
2-029G09	1,751889	2-033E07	2,309333			BL-001E12	5,202815
2-030D01	1,888472	2-033F10	1,664222			BL-001G10	3,763056
2-032C01	1,425556	2-034D10	1,649444			BL-001H10	4,572222
2-035D08	1,623565	2-035C10	1,629111			BL-002A06	3,880556
2-035D12	1,667778	2-035D11	1,415556			BL-002F05	3,969519
2-035F02	1,530694	2-040E09	1,586222			BL-002F08	3,917815
2-036H05	1,475417	2-041A12	1,721667			BL-002F10	4,566074
BL-001E11	2,133000	BL-002F09	1,454852			BL-002G03	4,722185
BL-006D11	1,748056	BL-002H12	1,440278			BL-002G04	3,801130
BL-006E09	2,474444	BL-003A02	1,412222			BL-002G06	3,762130

BL-007A09	1,626528	BL-003D12	1,598444			BL-003A05	4,282204
BL-008D01	2,229444	BL-003F10	3,167222			BL-003B09	4,858185
BL-010B08	1,584889	BL-003H09	1,757944			BL-003D03	3,222361
		BL-003H12	1,409056			BL-003E01	2,292454
		BL-004B08	2,694778			BL-003H03	3,429722
		BL-005E12	1,692046			BL-004A02	2,322537
		BL-006B04	1,524167			BL-004D02	3,116222
		BL-006C03	1,451361			BL-005D04	4,084352
		BL-006H09	1,531944			BL-005H12	5,493546
		BL-007B11	1,546306			BL-006H08	2,621111
		BL-007H05	1,615722			BL-007B01	3,510509
		BL-008D02	1,629222			BL-007G05	4,864120
		BL-008D09	1,433852			BL-008C01	3,419722
		BL-008F05	1,629444			BL-010F09	2,561722
		BL-008H05	1,531333				
		BL-010A01	1,545417				
		BL-010B06	1,832111				
		BL-010C07	1,935111				
		BL-010E05	1,494259				
		BL-010E06	1,920611				
		BL-010E11	2,126222				
		BL-010H02	1,593241				

Tab. 3: Risultati ottenuti con il Modello della Soglia.

Con il *Threshold Model* si ottengono, in base alle relative soglie, 125 geni positivi e 94 negativi, contro i 120 positivi e 93 negativi del test Sam.

Si nota una differenza in termini di estrapolazione: alcuni geni sono comuni ad entrambi i metodi, altri no.

Ciò è dovuto a diversi fattori che differenziano un metodo dall'altro: la costante s_0, presente nella (3.1) influenza sicuramente la scelta dei geni, nel metodo Sam; inoltre, il metodo della Soglia è stato applicato al vettore delle medie e di conseguenza non considera la variabilità tra i soggetti nel selezionare la soglia, diversamente da quanto accade, invece, con il metodo Sam.

Nel capitolo successivo, dedicato alle considerazioni conclusive, verranno messi in risalto i relativi vantaggi e svantaggi che si hanno nell'applicare la Teoria dei Valori estremi a dati derivanti da *microarray*.

4

Considerazioni Conclusive

In questa tesi sono stati analizzati dati riguardanti l'espressione genica di 10 soggetti affetti da leucemia linfoblastica acuta-B (ALL-B). L'analisi era finalizzata alla selezione di una soglia *u* in grado di individuare geni con una elevata espressione, sia in senso positivo che negativo.

Per tale scopo si è utilizzata la Teoria dei Valori Estremi ed è stato applicato il modello della Soglia (*Threshold Model*) basato sulla distribuzione generalizzata di Pareto (GPD), confrontato poi, con il test SAM. Inoltre, le analisi sono state condotte facendo riferimento al "soggetto medio", mentre per i singoli soggetti sono stati presentati solamente i plot illustrativi delle tecniche di selezione della soglia.

Dall'analisi condotta emerge che, le tecniche esplorative per la selezione della soglia si basano su valutazioni alquanto soggettive, dedotte dall'andamento del Mean residual Life Plot e dall'ampiezza dell'intervallo di confidenza.

Per l'identificazione dei geni sottoespressi, si è scelta la soglia u=2.2 che ci ha portato ad avere 94 valori eccedenti su un totale di 4992.

In riguardo all'identificazione dei geni sovraespressi, la soglia identificata è stata scelta pari a 1.4, con l'estrapolazione di 125 geni.

Considerata la soggettività della tecnica, si è ritenuto opportuno confrontarla con il Test SAM, molto noto in letteratura, nonché il più usato per analisi di questo tipo. Dal confronto è emerso che:

	VALORI ECCEDENTI	
	Thershold Model	Test SAM
Geni sottoespressi	94	96
Geni sovraespressi	125	120

I massimi/minimi estrapolati hanno circa la stessa numerosità in entrambi i metodi, ma non conducono agli stessi geni: solo il 45% dei geni estrapolati sono risultati comuni (vedi tabella 3, capitolo precedente).

È difficile capire quale metodo sia il migliore, si possono solo fare considerazioni in merito: di certo, la Teoria dei Valori Estremi non è mai stata applicata a questa tipologia di dati, quindi come primo impatto sicuramente il test Sam sarà il preferito. Inoltre, per quanto concerne la dipendenza tra dati, è possibile affermare che tra i geni esiste una correlazione, ma è sconosciuta (teoricamente un gene può rimanere "alto" all'infinito), diversamente rispetto ai dati di una serie storica (i più usati nella teoria dei valori estremi): anche questo potrebbe essere oggetto di preferenza. Ci sono, però, valide ragioni per preferire un'analisi basata sulla teoria dei Valori Estremi.

- Innanzitutto, il modello della Soglia lavora su ogni colonna della matrice o su un singolo vettore; il test Sam, invece, considera anche le repliche, ossia l'intera matrice.
- I geni estrapolati con il Modello della Soglia, godono di una distribuzione nota, che si svincola dai test, la cosiddetta Distribuzione Generalizzata di Pareto (GPD); tale risultato è importantissimo a livello biologico, nonché statistico, in quanto da la possibilità di poter stimare i relativi parametri e così giungere a risultati più precisi.

Complessivamente si può affermare che la tecnica analizzata ha portato a risultati abbastanza sensati anche dal punto di vista biologico. Dal confronto con il test Sam si nota che sarebbero necessarie delle ricerche biologiche più dettagliate per avere evidenze più concrete. Questo comunque esula dagli interessi del presente elaborato che si propone come una ricerca tutt'altro che definitiva.

Molte questioni restano comunque aperte, dando spazio alla ricerca e alla collaborazione tra diverse discipline. Molto probabilmente il continuo sviluppo tecnologico oltre che della biologia molecolare darà accesso, nei prossimi anni, a misure più attendibili, cosa che garantirà anche alla ricerca statistica delle basi su cui fondare il proprio lavoro.

BIBLIOGRAFIA

- Alberts, (2002), *Elementi di biologia cellulare*, Zanichelli.

- Allison, David B., Gadbury, Gary L., Heo Moonseoong, Fernandez, Josè R., Lee, Cheal-koo, Prolla, Tomas A. and Weindruch Richard, (2002), *A Mixture model approch for the analysis of microarray gene expression data*, Computazional statistics and data Analysis.

- Boncinelli, E. (2004), *Genoma: il grande libro dell'uomo.*

- Davison e Smith, (1990), Models for excedances over high thresholds, *Journal of the Royal Statistical Society B*, **Vol. 52**, pagg.: 393- 442.

- Di Giorgio, C. (2000), Completata la sequenza del Genoma Umano, *Repubblica*.

- Gordon K. Smith, Yee Hwa Yang abd terry Speed, (2002), *Statistical Issues in cDna Microarray Data Analysis*.

- Hastie, T, Tibshirani, R. and Friedman J., (1991), *The elements of statistical Learning. Data Mining, Inference, and Prediction*, Springer-Verlag, New York.

- Laursen, B. (1999), *Statistical Analisys of genetics distance data*.

- Lewin, B. (1994), *Genes V*, Oxford University Press.

- Lewin, B. (1992), *Il Gene IV*, Zanichelli.

- Pierotti Marco A. e Garibaldi Manuela, (2004), Dip. di oncologia sperimentale, *Istituto Nazionale Tumori Milano*.

- Romualdi Chiara, Campanaro Stefano, Campagna Davide, Celegato Barbara, Cannata Nicola, Toppo Stefano, Valle Giorgio, Lanfranchi Gerolamo, (2003), Pattern recognition ingene expression profile using DNA array: a comparative study of different statistical method applied to cancer classification, *Human molecolar genetics*, **Vol. 12**, pag.: 8.

- Sorvillo, S. (2000), Abbiamo decifrato il Genoma Umano, *Saninforma*.

- Stuart, Coles (2001), *An Introduction to Statistical modelling of Extreme Value*, Springer Series in statistics.

- Techini, M. L. (2004), Dip. Di Biologia e genetica per le Scienze Mediche, *Università Milano*.

- Tilstone, C. (2003), DNA microarrays: Vital Statistics, *Nature*.

- Troyaskaya O., Cantor M., Sherlock G., Brown P., Hastie T., Tibshirani R., Botstein D. and Altman R.B., (2001), *Missing value estimation methods for Dna microarra*, Bioinformatics.