UNIVERSITÀ DEGLI STUDI DI PADOVA

FACOLTÀ DI SCIENZE STATISTICHE

CORSO DI LAUREA IN STATISTICA E INFORMATICA

Tesi di Laurea

ANALISI DEI DATI DI ESPRESSIONE GENICA: STUDIO COMPARATIVO PER LA VALUTAZIONE DELL'IMPATTO DELLA NORMALIZZAZIONE SULL'INFERENZA STATISTICA

Relatore: Dott. Chiara Romualdi

Laureando: Davide Risso

Anno Accademico 2007/2008

Indice

Capitolo 1

Introduzione

Nel 1975 Edwin Southern dimostrò come fosse possibile fissare il DNA ad un supporto solido ed estrarre una catena complementare di DNA. Tale processo, noto come "*Southern blotting*" si può considerare come l'origine della tecnologia dei *microarray*: fu poi Fodor che nel 1991 fabbricò i primi *microarray*, combinando il metodo fotolitografico, usato per i semiconduttori, per realizzarne i primi fissando degli oligonucleotidi su superfici di vetro.

La tecnologia dei *DNA microarray* è oggi uno strumento ampiamente usato in molte ricerche in campo medico e biologico. La sua forza è permettere di misurare l'espressione di migliaia di geni simultaneamente, in un unico esperimento.

Gli studi sull'espressione genica cercano di determinare la quantità di RNA messaggero (mRNA) trascritto nel sistema biologico di riferimento, in quanto maggiore quantità di RNA si produce, maggiori saranno le proteine create. Risulta evidente quanto lo studio dell'espressione genica possa essere importante per studiare quelle malattie, ad esempio i tumori, caratterizzate da particolari alterazioni geniche, dalle quali dipendono il decorso della malattia e l'efficacia di farmaci specifici. Solo confrontando l'espressione genica di migliaia di geni in individui affetti da patologie diverse, è possibile individuare geni differenzialmente espressi, e quindi potenzialmente responsabili delle diversità nei gruppi.

La statistica, nell'analisi dei dati ricavati dagli esperimenti di *microarray*, è strumento indispensabile, non solo per l'identificazione di geni differenzialmente espressi, ma anche per altri obiettivi biologici non meno importanti, quali l'individuazione di geni co-regolati. Nel DNA sono infatti presenti geni detti "regolatori", che portano alla formazione di proteine che hanno il ruolo di interruttori per l'attivazione di un gene. I *microarray* possono essere inoltre usati come strumento di classificazione tra soggetti sani e malati, o tra sottoclassi di una data malattia. Spesso infatti pazienti affetti da una stessa patologia possono avere diverse risposte a diversi trattamenti farmacologici. Attraverso il loro profilo molecolare, e quindi attraverso studi di espressione, si è visto che in realtà quella stessa patologia poteva essere divisa in più sottoclassi patologiche. Questo aspetto quindi rende i microarray estremamente potenti sia sotto l'aspetto diagnostico che sotto quello prognostico.

1.1 Dal DNA alle proteine

Per capire meglio l'utilità dello studio dell'espressione genica è necessario introdurre alcune nozioni di base di biologia molecolare. La cellula è l'unità biologica fondamentale, alla quale si riconducono tutte le funzioni vitali. Essa è composta dal nucleo e dal citoplasma, nel quale si trovano i ribosomi, in cui avviene l'assemblaggio degli amminoacidi per la sintesi delle proteine. Nel nucleo, invece, si trova il DNA, struttura a doppia elica che ricorda quella di una scala a pioli disposta a spirale. I "lati" della scala sono costituiti da uno zucchero e da un gruppo fosfato, mentre i "pioli" consistono in due nucleotidi collegati debolmente da legami a idrogeno. Ad ogni base corrisponde la sua complementare: cosí l'Adenina (A) è appaiata alla Timina (T) e la Guanina (G) alla Citosina (C).

La sintesi delle proteine avviene principalmente in due fasi:

1. fase di trascrizione: nella quale un piccolo frammento di DNA, nel quale sono contenute le informazioni per creare una proteina, viene copiato su un filamento di RNA messaggero;

2. fase di traduzione: nella quale il filamento di mRNA esce dal nucleo cellulare per unirsi ai ribosomi e creare le catene di amminoacidi che andranno a formare le proteine;

1.2 La tecnologia dei *microarray*

Si è già detto di come sia importante la possibilità di misurare l'espressione genica in differenti condizioni, e di come i *microarray* siano una tecnologia in grado di farlo su migliaia di geni simultaneamente. Esistono diversi tipi di *microarray* in commercio, ma nel presente studio ci si è concentrati in modo particolare sui *microarrays a cDNA*.

1.2.1 *cDNA microarrays*

I *microarrays a cDNA* sono uno strumento di misurazione dell'espressione genica a doppio canale che agisce in modo indiretto: essi misurano infatti l'intensità di colore di filamenti di cDNA (i.e. sano vs malato) etichettati con due sostanze fluorescenti e ibridati sullo stesso *array* (ibridazione competitiva). Il cDNA (DNA complementare) è la copia negativa dell'RNA, creata dagli enzimi attraverso un'operazione detta di "retro trascrizione". Le molecole di cDNA sono più stabili di quelle di RNA e permettono di avere informazione sui geni attivi in una cellula in un dato momento.

Durante un esperimento di *microarray* i segmenti di cDNA sintetico, creati in laboratorio, che riguardano i geni che si pensa possano essere importanti per il caso in studio, detti sonde (*probes*), vengono collocati su superfici piatte, come vetrini o *microchip*. Questa operazione è eseguita attraverso un braccio meccanico che dispone in griglia sul vetrino minuscole quantità di cDNA, dell'ordine dei nanolitri, su ogni punto (detto *spot*) attraverso delle particolari punte, dette *print-tip*. Il cDNA proveniente dalla retro trascrizione dell'mRNA dei campioni biologici studiati viene detto "*target*": per rendere il *cDNA target* "visibile" si aggiunge una sostanza fluorescente in modo che la quantità che si lega alla superficie del vetrino possa essere mi-

surata attraverso uno scanner ottico. Le sostanze solitamente utilizzate per marcare i geni sono il *Cy3*(*Cyanine 3*), di colore verde, e il *Cy5*(*Cyanine 5*), di colore rosso. A questo punto viene eseguito il processo di ibridazione competitiva, in cui i *cDNA target* dei campioni biologici vengono depositati sui *probe*: le sequenze del *cDNA probe* si legano alle sequenze del cDNA *target*, grazie allo stesso tipo di legame a idrogeno che lega le due eliche del DNA. In questo modo, se in una posizione del vetrino sono presenti le sequenze di *probe* relative ad un gene k, a tali sequenze si legheranno le corrispondenti sequenze del *target* in quantità proporzionale all'espressione dello stesso gene k. Il procedimento è schematizzato in Figura 1.1. Il risultato finale dell'esperimento è una griglia di *spot* fluorescenti, analizzati poi attraverso uno scanner ottico, che permette di quantificare le intensità della fluorescenza dei due campioni studiati (Figura 1.2).

1.3 La matrice dei dati

Si è già accennato all'importante ruolo che la statistica riveste nell'analisi dei dati provenienti da esperimenti di *microarray*. In questo paragrafo verrà presentata la struttura della matrice dei dati per chiarire meglio le operazioni preliminari e le successiva inferenza.

Dal momento che l'unità statistica viene usualmente definita come una replicazione indipendente soggetta a condizioni o trattamenti di interesse, nel caso dei *microarray* si dovrebbe considerare come unità statistica ogni esperimento di *array*. Tuttavia nel caso degli studi di espressione l'unità di interesse è il gene, quindi è più comune considerare come unità statistica il gene (Wit e McClure (2004)).

Il *dataset* si presenta quindi come una matrice di p righe (numero di geni) per $2 \cdot n$ colonne (n numero di esperimenti): n per il canale marcato con Cy3, che verrà in seguito denominato "canale verde", e n per il canale marcato con Cy5, denominato "canale rosso". Un esempio di come si presenta la matrice dei dati è illustrato in Tabella 1.1, dove con $x_{g,i}$ e $y_{g,i}$ si fa riferimento

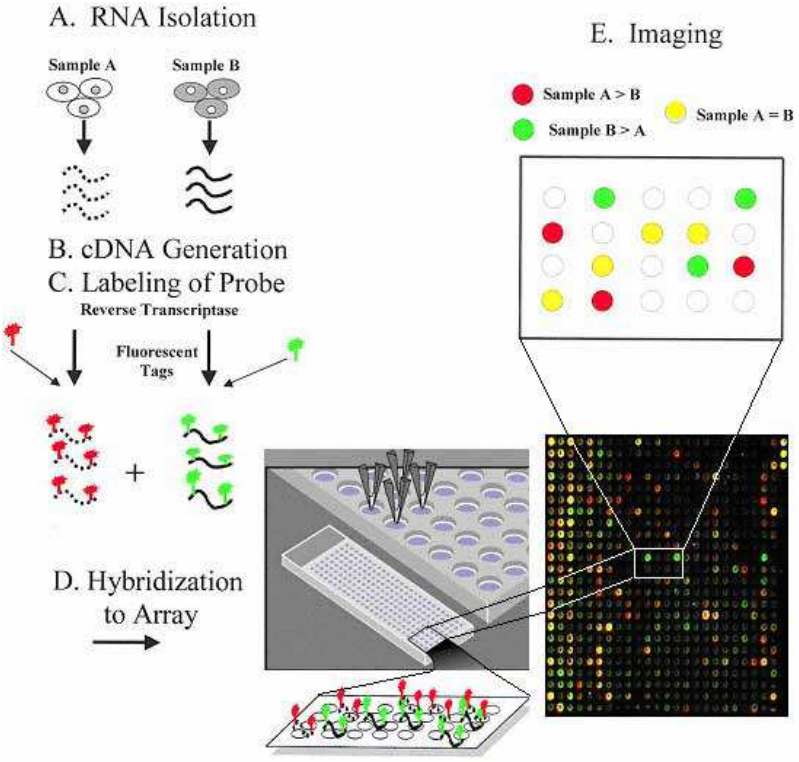

Figura 1.1: Struttura di un esperimento di *microarray*.

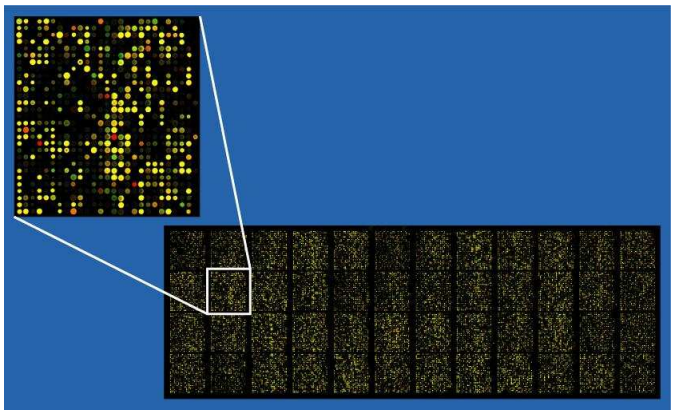

Figura 1.2: Immagine finale dell'esperimento.

all'espressione del gene g, rispettivamente nel canale rosso e nel canale verde, per l'i-esimo *array*.

	Cy5			*Cy3*		
gene 1	$x_{1,1}$	...	$x_{1,n}$	$y_{1,1}$	...	$y_{1,n}$
...						
gene g	$x_{g,1}$	...	$x_{g,n}$	$y_{g,1}$	...	$y_{g,n}$
...						
gene p	$x_{p,1}$	...	$x_{p,n}$	$y_{p,1}$	...	$y_{p,n}$

Tabella 1.1: Esempio di matrice dei dati.

1.4 Normalizzazione

La struttura dei dati appena descritta mette in evidenza la peculiarità dell'analisi di dati provenienti da esperimenti di *microarray*: siamo davanti ad un *dataset* con un numero molto elevato di geni, a fronte di un numero decisamente più basso di replicazioni a disposizione per ogni gene (paradigma del "*large p, small n*"). Proprio per il fatto che le replicazioni sono molto poche se confrontate col numero di geni da analizzare, anche un piccolo errore sistematico può provocare distorsioni significative nella fase di inferenza.

La normalizzazione è l'operazione preliminare all'inferenza necessaria per l'identificazione e la rimozione degli errori sistematici all'interno dello stesso *array* e tra *arrays* diversi. Tali distorsioni vengono introdotte in fase di preparazione del *microarray*, di esecuzione dell'esperimento e di scansione del risultato.

La maggior parte degli algoritmi di normalizzazione recentemente proposti ed efficaci nella riduzione o eliminazione degli errori sistematici hanno però delle assunzioni sull'espressione e sulla simmetria dei dati:

1. Espressione: solo un numero relativamente piccolo di geni varia significativamente nell'espressione tra i due canali;

2. Simmetria: il numero di geni sovra-espressi è equivalente (almeno approssimativamente) al numero di geni sotto-espressi.

Visto oramai che le recenti piattaforme di *microarray* contengono più di 50,000 probe, queste assunzioni sono rispettate dalla maggioranza degli esperimenti di *microarray*. Nel nostro studio in particolare, utlizzando dati simulati, le assunzioni sono soddisfatte. Si è deciso infatti di simulare dati in cui solo il 5-6% dei geni è generato come differenzialmente espresso, inoltre il piano di simulazione dei geni differenzialmente espressi rende la loro distribuzione simmetrica (per i dettagli si veda il Capitolo 2).

In letteratura recentemente sono stati proposti alcuni metodi più o meno efficaci per la normalizzazione dei dati di espressione che verranno descritti esaustivamente nel Capitolo 3.

Possiamo comunque classificare tali metodi in tre grandi categorie:

- metodi di normalizzazione globale: assumono che le intensità del canale rosso e del canale verde siano in relazione attraverso una costante. La normalizzazione, quindi consiste semplicemente nello spostamento della media (o mediana) della distribuzione dei logaritmi dei rapporti sullo zero:
$$\log_2 \frac{x_g}{y_g} \rightarrow \log_2 \frac{x_g}{y_g} - m \tag{1.1}$$
- metodi di normalizzazione dipendenti dall'intensità (*intensity-dependent normalization*): la relazione tra i due canali non è più considerata costante, ma è una funzione che dipende dalla media dei logaritmi delle intensità nei due canali:
$$\log_2 \frac{x_g}{y_g} \rightarrow \log_2 \frac{x_g}{y_g} - c\,(A) \tag{1.2}$$
dove:
$$A = \frac{1}{2}\left(\log_2 x_g + \log_2 x_g\right) \tag{1.3}$$
La funzione $c\,(A)$ può essere ad esempio stimata tramite una regressione locale (Yang *et al.* (2002)).

- metodi di normalizzazione spaziale (*spatial-dependent normalization*): la relazione tra i due canali è considerata dipendente dalle coordinate spaziali delo *spot* nell'*array*:

$$\log_2 \frac{x_g}{y_g} \rightarrow \log_2 \frac{x_g}{y_g} - c\,(X, Y) \tag{1.4}$$

La necessità di considerare la dipendenza dall'intensità media per normalizzare le espressioni geniche è giustificata principalmente dal fatto che per bassi valori di intensità nei due canali anche piccole variazioni sono più influenti. Inoltre gli scanner ottici faticano a catturare le intensità troppo basse (*spot* troppo poco luminosi) e quelle troppo alte, perché l'intensità può raggiungere il livello di massima intensità che lo scanner può registrare ($2^{16} - 1$ unità) creando così un effetto di saturazione.

La ragione della dipendenza spaziale è meno chiara: Yang *et al.* (2002) suggeriscono che la distorsione spaziale sia dovuta all'utilizzo di differenti *print-tip* in zone diverse dell'*array*. Futschik e Crompton (2004) osservano che non essendoci distorsioni a blocchi, ma un trend spaziale continuo su tutto l'*array*, è più verosimile che tale distorsione dipenda da una ibridazione non omogenea. Per tale motivo il metodo di normalizzazione proposto da Yang *et al.* (2002) utilizza le *print-tip* per eliminare la distorsione spaziale, mentre Futschik e Crompton (2004) utilizzano una funzione delle coordinate dello *spot*. Rimandiamo comunque al Capitolo 3 per i dettagli.

1.5 Identificazione di geni differenzialmente espressi

Uno dei risultati più importanti che si possono ottenere da un esperimento di *microarray* è l'identificazione di geni detti "differenzialmente espressi" nei due o più gruppi di campioni biologici presi in considerazione nello studio. I geni differenzialmente espressi sono quei geni che mostrano una differenza significativa nell'espressione in un gruppo rispetto ad un altro.

Esistono in letteratura molti test che possono essere impiegati per l'individuazione di geni con differente livello di espressione. Nel presente studio si è deciso di utilizzare il test SAM (*Significance Analysis of Microarray*), proposto da Tusher *et al.* (2001). SAM è una variante del test t per il confronto dell medie, in cui però una costante positiva viene introdotta al denominatore. Questa modifica, detta *moderazione* rende la statistica test meno sensibile alla variabilità dei livelli di espressione genica.

Il test viene effettuato gene per gene sull'intera matrice di dati (per ciascuna riga della matrice) ottenendo quindi per ciascun gene un livello di significatività.

Per identificare i geni è poi necessario definire una soglia oltre la quale si rifiuta l'ipotesi di uguaglianza dell'espressione dei geni considerati. In questo studio la soglia è stata fatta variare per osservare, attraverso una curva ROC, l'impatto delle normalizzazione sula frazione di falsi positivi e negativi del test (si vedano i dettagli nel Capitolo 3). Negli esperimenti reali, invece, la soglia è scelta in modo da controllare il *False Discovery Rate* (FDR), definito come il numero atteso di geni equivalentemente espressi tra quelli dichiarati differenzialmente espressi:

$$FDR = E\left[\frac{F_p}{S}\right] \tag{1.5}$$

dove F_p è il numero di geni equivalentemente espressi classificati come differenzialmente espressi, e S è il numero di geni dichiarati differenzialmente espressi. È chiaro quindi che avere un FDR basso equivale a tenere bassa la probabilità di falsi positivi.

Se in un esperimento si fosse interessati all'espressione di un singolo gene sarebbe sufficiente considerare il *p-value* per accettare o rifiutare l'ipotesi nulla di equivalente espressione del gene nei due gruppi. Dal momento però che la forza dei *microarray* consiste proprio nel testare contemporaneamente migliaia di geni, è necessaria una procedura che combini un numero finito di test dipendenti. Molto spesso nel caso di test multipli si utilizzano procedure che controllano il *Familywise Error Rate* (FWER), come la correzione di

Bonferroni o la procedura di Hochberg. Il FWER è definito come la probabilità che almeno un gene equivalentemente espresso sia classificato dal test come differenzialmente espresso:

$$FWER = Pr\left(F_p > 0\right) \tag{1.6}$$

È evidente come il FWER sia un criterio molto conservativo: specialmente con un grande numero di test parziali è impensabile che sia bassa la probabilità di sbagliare anche solo una volta.

Per questo motivo nel campo dei *microarray* si preferiscono procedure che controllano il FDR. Le procedure che si basano sul controllo del FDR permettono di decidere un margine (i.e. 5%) di geni equivalentemente espressi tra quelli dichiarati differenzialmente espressi. In Tusher *et al.* (2001) il FDR è stimato tramite permutazioni per ogni valore della soglia.

1.6 Scopo della Tesi

In questo elaborato l'interesse è rivolto all'influenza che vari metodi di normalizzazione presentati recentemente in letteratura hanno sulle successive analisi inferenziali per l'identificazione di geni differenzialmente espressi.

L'impatto delle normalizzazioni sulle procedure inferenziali è stato valutato attraverso le stime di sensibilità e specificità del test SAM su dati simulati e attraverso l'ordine della statistica test su un *dataset* reale.

La complessa struttura di correlazione tra geni e la presenza di variabilità spuria causata da errori sistematici in alcune fasi dell'esperimento rendono la simulazione dei dati un passo delicato. A questo scopo si sono presi in considerazione più modelli di simulazione.

Per ognuno dei modelli di simulazione presi in considerazione (si veda il Capitolo 2) si sono generate 10 matrici di dati, quindi 10 esperimenti da 15 *array*, ad ognuna delle quali si sono applicati i metodi di normalizzazione considerati. La scelta di 10 simulazioni per ogni modello, che può sembrare un numero limitato di ripetizioni, è stato un compromesso tra numero di

normalizzazioni da confrontare e il costo computazionale per le simulazioni. Il numero totale di matrici su cui è stato effettuato il test SAM è di 400. E' chiaro come aumentare il numero di ripetizioni avrebbe innalzato molto il tempo macchina.

Gli strumenti più adatti per il confronto tra sensibilità e specificità del test SAM sono le curve ROC, grafici con in ascissa la proporzione di falsi positivi sul totale dei geni equivalentemente espressi (1-specificità) e in ordinata la proporzione di veri positivi sul totale dei differenzialmente espressi (sensibilità).

Per meglio comprendere la definizione di sensibilità e specificità è utile fare riferimento alla Tabella 1.2: le colonne corrispondono all'esito del test, mentre le righe alla reale espressione dei geni.

	Espressione prevista dal test		
Espressione reale	Equivalente	Differente	Totale
Equivalente	a	b	a + b
Differente	c	d	c + d
Totale	a + c	b + d	

Tabella 1.2: Tabella per l'identificazione dei geni correttamente classificati dal test.

La sensibilità corrisponde alla frazione di geni differenzialmente espressi correttamente identificati:

$$Sensibilità = \frac{d}{c+d}$$

La specificità alla frazione di geni equivalentemente espressi correttamente identificati:

$$Specificità = \frac{a}{a+b}$$

Oltre all'impatto su sensibilità e specificità si è utilizzato come ulteriore indice per il confronto l'ordinamento della statistica test: si sono ordinati i geni identificati dal test come differenzialmente espressi in base al valore assoluto assegnato loro dalla statistica test e si sono confrontate le liste dei

geni con rango più alto, utilizzando poi la percentuale di geni presenti in entrambe le liste come indice per il confronto tra le due normalizzazioni.

La stessa procedura è stata utilizzata anche per analizzare un set di dati proveniente da un esperimento reale, in cui, non conoscendo a priori i geni differenzialmente espressi, non si sono potuti utilizzare come indici di bontà del modello sensibilità e specificità.

Uno schema generale di come si è svolto il lavoro si può osservare in Figura 1.3.

Il secondo capitolo di questo elaborato si propone di approfondire i tre diversi modelli di simulazione utilizzati nello studio, mentre il terzo capitolo si concentrerà su alcune delle diverse tecniche di normalizzazione proposte in letteratura, soffermandosi sulle caratteristiche e sui metodi statistici sottostanti. Sarà poi presentato il test SAM, utilizzato per il confronto tra le normalizzazioni e il metodo di imputazione KNN. Infine saranno presentati e discussi i risultati ottenuti con i dati generati. Le simulazioni e le analisi proposte sono state condotte utilizzando i *software* R[1] e *BioConductor*[2].

[1] http://www.r-project.org

[2] http://www.bioconductor.org

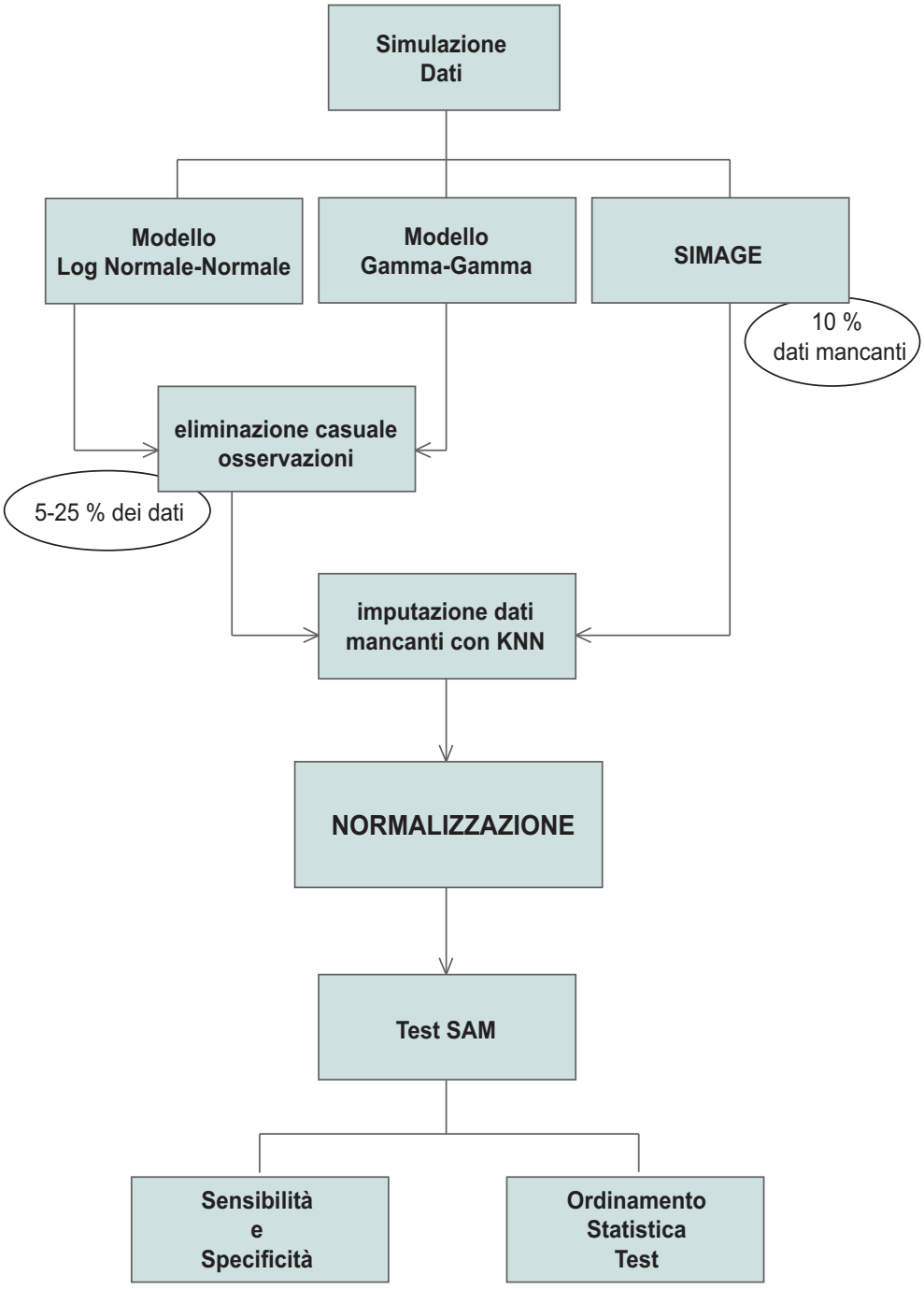

Figura 1.3: Struttura dell'analisi svolta.

Capitolo 2

Modelli di Simulazione

2.1 Introduzione

Come accennato nel Capitolo 1, la necessità di lavorare con dati simulati è duplice: da una parte essi hanno il grande vantaggio di avere caratteristiche note e dall'altro di essere disponibili in quantità pressoché illimitate a costo zero. L' unica limitazione vincolante è data dal tempo computazionale delle analisi. Conoscendo a priori quali geni sono stati generati come egualmente espressi e quali come differenzialmente espressi, si ha la possibilità di utilizzare come indice di bontà di un test statistico la sua sensibilità e specificità.

Nel presente studio si sono presi in considerazione tre diversi modelli di simulazione:

- due modelli bayesiani parametrici gerarchici: il modello LogNormale-Normale e il modello Gamma-Gamma (Newton *et al.* (2003));
- un metodo recentemente proposto chiamato SIMAGE (Albers *et al.* (2006)).

I modelli parametrici gerarchici sono modelli bayesiani in cui ogni gene è stocasticamente classificato come egualmente o differenzialmente espresso, e in base a tale classificazione la sua espressione nei due canali viene generata

da una specifica densità di probabilità congiunta. Ovviamente è possibile scegliere la percentuale di geni differenzialmente espressi, specificando la probabilità a priori δ di un gene di essere differenzialmente espresso nei due canali. Entrambi questi modelli generano dati che non presentano alcun tipo di distorsione sistematica. I dati simulati con SIMAGE, invece, sono generati da un modello più complicato che prende in considerazione molti parametri, per generare dati con distorsioni sistematiche e non, il più simili possibile a quelle riscontrate nei dati reali; per i dettagli si rimanda al Paragrafo 4.3.

2.2 I modelli bayesiani gerarchici

Newton *et al.* (2003) propongono un modello mistura di tipo gerarchico per l'identificazione di geni differenzialmente espressi. Questo modello permette l'utilizzo di dati di espressione genica provenienti da cellule sotto differenti condizioni e con la possibilità di più replicazioni per ogni condizione. Il numero di condizioni da cui sono ottenute le misure di espressione porta ad avere un sistema di ipotesi con un numero diverso di possibili *pattern* di espressione: ci limitiamo a trattare il caso di due diverse condizioni, in cui sono possibili solo due *pattern* di espressione, ovvero espressione equivalente oppure espressione differenziale, dal momento che siamo interessati esclusivamente a generare dati simulando esperimenti di *microarray* a due canali. In ogni caso l'estensione a situazioni con più di due *pattern* di espressione è immediata.

Indichiamo dunque con $x_g = (x_{g,1}, \ldots, x_{g,n})$ l'espressione del gene g nella prima condizione e con $y_g = (y_{g,1}, \ldots, y_{g,n})$ l'espressione dello stesso gene nella seconda condizione. A questo punto si può considerare una distribuzione mistura, considerando il fatto che ogni gene può essere differenzialmente espresso (DE), con una probabilità δ, oppure equivalentemente espresso (EE), con una probabilità $1 - \delta$. Se il gene g è EE allora la sua espressione proviene da una densità di probabilità congiunta $f_0(x_g, y_g)$, se al contrario è DE allora la densità di probabilità congiunta è $f_1(x_g, y_g)$. Allora la distribuzione

marginale dei dati sarà:

$$f(x_g, y_g) = \delta\, f_1\,(x_g, y_g) + (1 - \delta)\; f_0\,(x_g, y_g) \tag{2.1}$$

Nell'approccio bayesiano empirico, si suppone che x_g e y_g siano campioni casuali semplici rispettivamente dalla distribuzione $f_{obs}\,(x_g|\mu_g)$ e $f_{obs}\,(y_g|\mu_g)$. Se si considera il caso EE, allora supponiamo che le $2\,n$ misurazioni siano indipendenti e identicamente distribuite secondo la distribuzione $f_{obs}\,(\cdot|\mu_g)$, in altre parole si suppongono x_g e y_g provenienti da una distribuzione comune, supponendo che non ci siano variazioni sistematiche tra i due canali. Le forme parametriche che consideriamo per $f_{obs}\,(\cdot|\mu_g)$ sono la Gamma e la Log-Normale; entrambe sembrano avere caratteristiche simili alla distribuzione delle espressioni geniche derivanti da esperimenti reali.

Si considera ora la distribuzione $\pi\,(\mu_g)$, che rappresenta le variazioni nel livello di espressione medio fra tutti i geni considerati nell'esperimento. Quindi sotto l'ipotesi EE si può esprimere la distribuzione marginale come:

$$f_0\,(x_g, y_g) = \int \left(\prod_{i=1}^{n} f_{obs}\,(x_{g,i}|\mu_g)\right) \left(\prod_{i=1}^{n} f_{obs}\,(y_{g,j}|\mu_g)\right) \pi\,(\mu_g)\, d\mu_g \tag{2.2}$$

Sotto l'ipotesi DE, la media latente μ_{g1} relativa al campione $x_{g,i}$ (con $i = 1, \ldots, n$) è diversa dalla media μ_{g2} relativa al campione $y_{g,i}$ (con $i = 1, \ldots, n$). I due valori della media sono generati in modo indipendente dalla distribuzione $\pi\,(\mu_g)$. Questo porta alla seguente rappresentazione:

$$f_1\,(x_g, y_g) = f_0\,(x_g)\; f_0\,(y_g) \tag{2.3}$$

dove:

$$f_0\,(x_g) = \int \left(\prod_{i=1}^{n} f_{obs}\,(x_{g,i}|\mu_g)\right) \pi\,(\mu_g)\, d\mu_g \tag{2.4}$$

$$f_0\,(y_g) = \int \left(\prod_{i=1}^{n} f_{obs}\,(y_{g,i}|\mu_g)\right) \pi\,(\mu_g)\, d\mu_g \tag{2.5}$$

Per la simulazione dei livelli di espressione genica è quindi sufficiente estrarre casualmente l'espressione di ogni gene dalla distribuzione marginale

data dalla (2.1), dopo avere specificato opportunamente la probabilità δ di differente espressione e aver parametrizzato in modo opportuno la densità condizionata $f_{obs}(\cdot|\mu_g)$.

Come visto in precedenza, il modello mistura è specificato dalla distribuzione sulla singola osservazione $f_{obs}(\cdot|\mu_g)$, che caratterizza la variabilità relativa alle misure ripetute di un gene avente la stessa media di espressione μ_g, e da una seconda componente $\pi(\mu_g)$ che descrive la variabilità in queste medie tra geni.

Le due distribuzioni che saranno specificate sono di tipo parametrico e ipotizzano entrambe il coefficiente di variazione costante per tutti i geni analizzati. Questa è un'assunzione ragionevole in quanto in molti *dataset* reali il coefficiente di variazione dei dati risulta pressochè costante.

2.2.1 Il modello Gamma-Gamma

Nel modello Gamma-Gamma (GG) la distribuzione sulla singola osservazione è una Gamma con parametro di forma $\alpha > 0$ e media μ_g. Il parametro di scala λ è dunque pari a α/μ_g. La distribuzione assume quindi la seguente forma (per $z > 0$):

$$f_{obs}(z|\mu_g) = \frac{\lambda^\alpha z^{\alpha-1} \exp\{-\lambda z\}}{\Gamma(\alpha)} \tag{2.6}$$

Il coefficiente di variazione della variabile $Z|\mu_g \sim Ga(\alpha,\lambda)$ è pari a:

$$\frac{SE(Z|\mu_g)}{\mathrm{E}(Z|\mu_g)} = \frac{\sqrt{\alpha/\lambda^2}}{\alpha/\lambda} = \frac{1}{\sqrt{\alpha}} \tag{2.7}$$

ed è quindi costante per tutti i geni dal momento che il parametro α non dipende dal singolo gene. Appare a questo punto evidente come la distribuzione Gamma sia una buona scelta, in quanto garantisce, oltre alla facilità di calcolo, supporto positivo e coefficiente di variazione costante. Fissato α è necessario scegliere la distribuzione marginale $\pi(\mu_g)$, che viene scelta come inversa di una Gamma. In questo modo, fissando α, la quantità α/μ_g ha distribuzione Gamma con parametro di forma α_0 e parametro di scala ν: $\lambda \sim Ga(\alpha_0, \nu)$. In totale sono quindi coinvolti tre parametri: $\theta = (\alpha, \alpha_0, \nu)$.

2.2.2 Il modello LogNormale-Normale

Nel modello LogNormale-Normale (LNN) si ipotizza che la distribuzione della trasformata logaritmica della singola osservazione sia normale. Indichiamo con $z_{g,i}$ la misura di espressione e con $\tilde{z}_{g,i} = \log(z_{g,i})$ il suo logaritmo naturale. La variabile $\tilde{Z}_{g,i}|\mu_g$ si distribuisce come una Normale, con media dipendente dal singolo gene e varianza comune per tutti i geni: $\tilde{Z}_{g,i}|\mu_g \sim N(\mu_g, \sigma^2)$. La distribuzione a priori di μ_g è anch'essa Normale: $N(\mu_0, \tau_0^2)$.

Si noti che la media latente μ_g è ora una media per i logaritmi delle misure di espressione. Il coefficiente di variazione per le misure nella scala originale è costante anche in questo modello:

$$\frac{SE(Z|\mu_g)}{\mathrm{E}(Z|\mu_g)} = \sqrt{\exp(\sigma^2) - 1} \tag{2.8}$$

I parametri coinvolti sono anche in questo caso tre: $\theta = (\mu, \sigma^2, \tau_0)$.

2.2.3 Simulazione

Nonostante i modelli bayesiani gerarchici siano stati proposti in Newton *et al.* (2003) come metodo di inferenza, per l'identificazione dei geni differenzialmente espressi, si prestano particolarmente bene a simulare espressioni geniche simili a quelle reali. L'assunzione che sta alla base dei metodi inferenziali basati su questi modelli, infatti, è che i dati di espressione su cui fare inferenza siano provenienti da tali distribuzioni.

Nel presente elaborato si sono simulate 10 matrici utilizzando il modello GG e 10 utilizzando il modello LNN, ognuna con 2 condizioni, 10000 geni e 15 ripetizioni per ogni condizione. La probabilità a priori che un gene sia differenzialmente espresso è stata posta al 5% in entrambi i modelli ($\delta = 0.05$). Per quanto riguarda gli altri parametri, si sono considerati i seguenti valori numerici:

- modello GG: $\theta = (\alpha, \alpha_0, \nu) = (1, 1.1, 45.4)$;
- modello LNN: $\theta = (\mu, \sigma, \tau_0) = (6.58, 0.9, 1.13)$.

La scelta di tali parametri assicura un coefficiente di variazione pari circa a 1 in entrambi i modelli, come si può facilmente calcolare dalla (2.7) e dalla (2.8):

- modello GG: $CV = \frac{1}{\sqrt{\alpha}} = 1$
- modello LNN: $CV = \sqrt{\exp(\sigma^2) - 1} \cong 1.2$

In Figura 2.1 si può osservare il coefficiente di variazione stimato dai dati simulati con il modello GG in funzione della media dell'espressione genica. Analogamente si può osservare il coefficiente di variazione stimato dai dati simulati con il modello LNN in Figura 2.2.

2.2.4 Distorsione

Dal momento che l'obiettivo del presente elaborato è testare differenti metodi di normalizzazione, si è deciso di introdurre una distorsione arbitraria in entrambi i modelli: i dati generati con i modelli GG e LNN infatti assumono la forma di dati di espressione "puliti", nei quali cioè non è presente alcun tipo di distorsione sistematica.

Prima di concentrarsi sulla trasformazione dei dati e sulla filosofia che sta alla base della scelta della distorsione, è opportuno introdurre brevemente lo strumento degli MA-Plot. Un MA-Plot è un grafico con in ascissa la media dei logaritmi in base 2 dell'espressione dei geni nei due canali (A) e in ordinata il logaritmo in base 2 del rapporto tra le due espressioni geniche (M). Ricordando che con x_g indichiamo l'espressione del gene g nella prima condizione e con y_g l'espressione del gene g nella seconda condizione, si ha:

$$A = \frac{1}{2}(\log_2 x_g + \log_2 y_g) = \frac{1}{2}\log_2(x_g y_g) \tag{2.9}$$

$$M = \log_2\left(\frac{x_g}{y_g}\right) \tag{2.10}$$

Gli MA-Plot, che non sono altro che una rotazione di 45° in senso antiorario del grafico dei logaritmi delle intensità nei due canali, con un'opportuna trasformazione di scala, sono strumenti essenziali per rappresentare graficamente i dati di espressione di un singolo *array*.

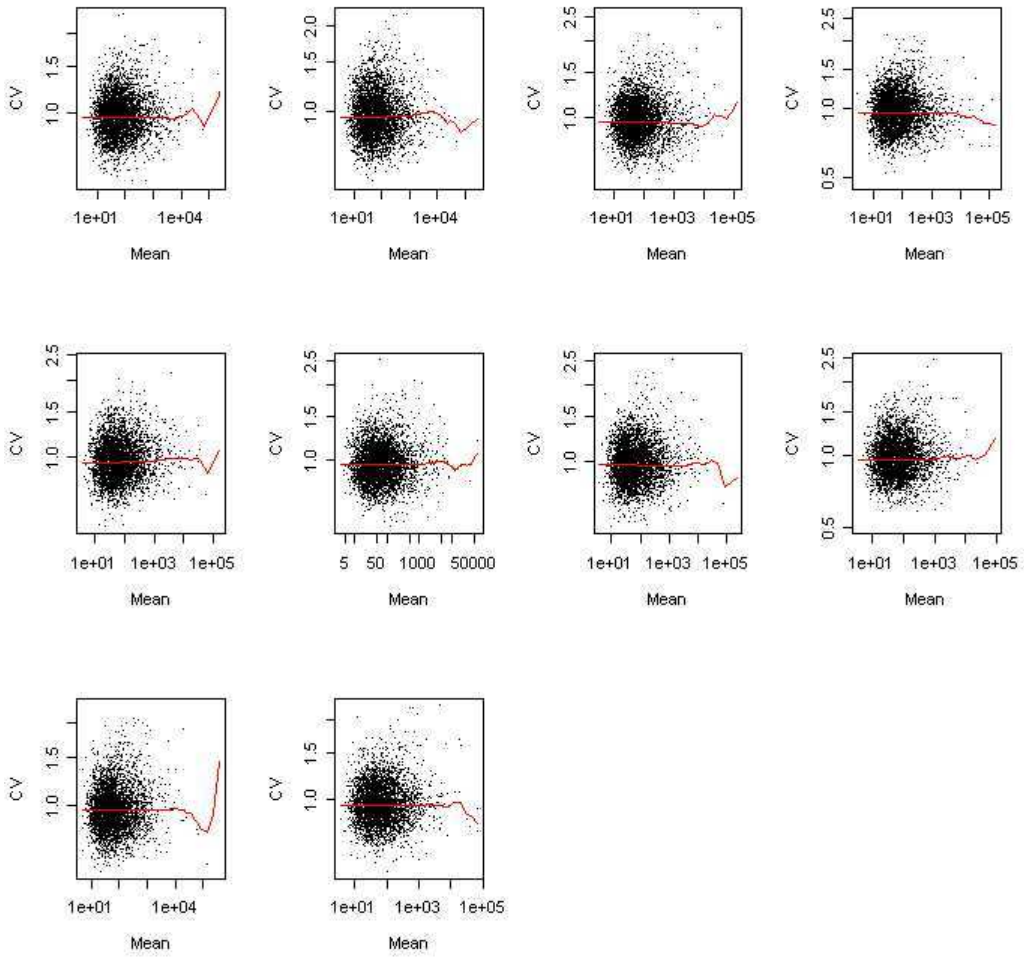

Figura 2.1: Coefficiente di variazione nelle misure di espressione simulate con il modello GG.

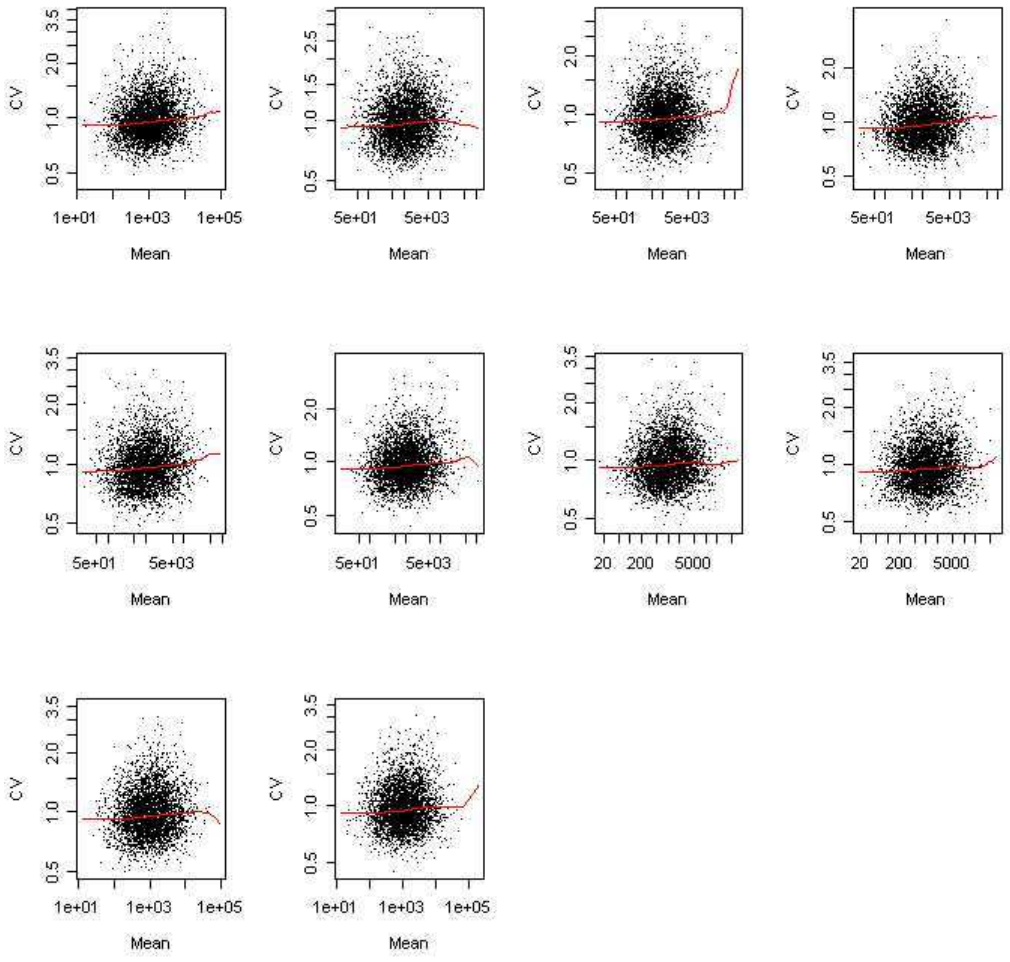

Figura 2.2: Coefficiente di variazione nelle misure di espressione simulate con il modello LNN.

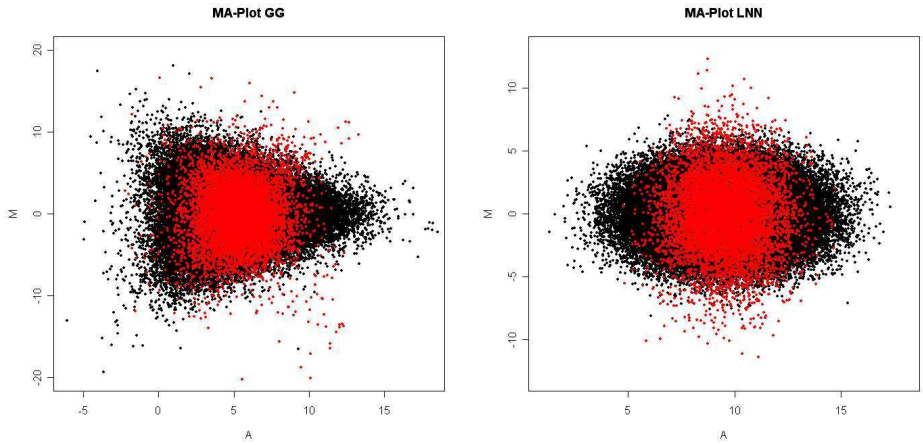

Figura 2.3: MA-Plot dei dati di espressione della prima matrice generata dal modello GG (sinistra) e dal modello LNN (destra). Sono evidenziati i geni differenzialmente espressi.

Dudoit *et al.* (2002) infatti dimostrano come l'utilizzo degli MA-Plot faciliti l'identificazione di distorsioni dovute alle varie fasi sperimentali della tecnologia. In un esperimento di *microarray* non distorto l' MA-Plot si presenta come una nuvola ellittica di punti simmetrica rispetto alla retta $M = 0$, negli esperimenti reali invece, i dati non normalizzati hanno spesso un andamento non lineare, con una coda per valori bassi di espressione media verso il basso o verso l'alto, indice di errori sistematici nei processi di marcatura e ibridazione.

Come si può osservare negli MA-Plot di Figura 2.3 i modelli GG e LNN simulano dati di espressione senza alcun tipo di distorsione; in altre parole le espressioni geniche dei due canali sono confrontabili, e le differenze nell'espressione tra il canale rosso e il verde non sono imputabili ad errori sistematici ma solo a differenze biologiche nei due campioni. Appare evidente come in una situazione di questo tipo qualsiasi normalizzazione diventa superflua.

Partendo dalla rappresentazione grafica dei livelli di espressione, si è quindi deciso di introdurre nei dati simulati una distorsione, dipendente dalla intensità e che facesse assumere all'MA-Plot un andamento simile a quello

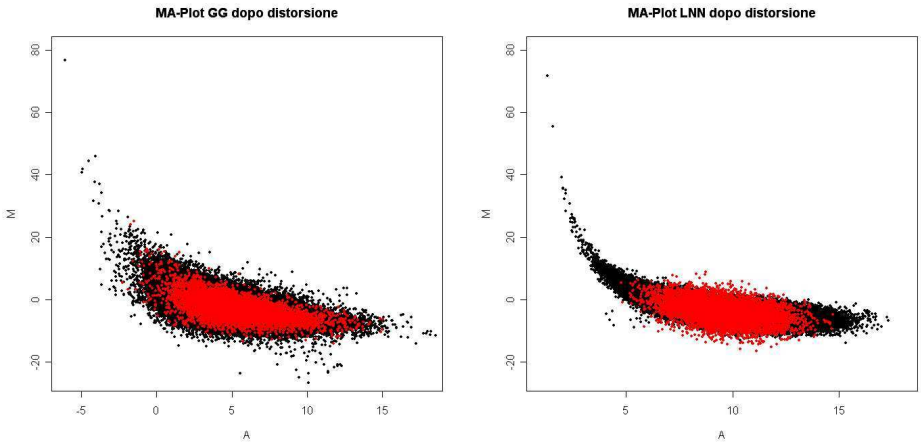

Figura 2.4: MA-Plot dei dati di espressione della prima matrice generata dal modello GG (sinistra) e dal modello LNN (destra), dopo l'introduzione della distorsione sistematica. Sono evidenziati i geni differenzialmente espressi.

reale. Si è deciso di sommare al logaritmo del rapporto delle intensità una funzione del tipo $\frac{1}{A}$ opportunamente scalata. Dopo vari tentativi si è deciso di utilizzare la seguente trasformazione per il modello GG:

$$\log_2\left(\frac{\tilde{x}_g}{\tilde{y}_g}\right) = \log_2\left(\frac{x_g}{y_g}\right) + \frac{300}{\log_2\left(x_g y_g\right) + 15} - 15 \tag{2.11}$$

Per il modello LNN, invece, i dati sono stati trasformati nel modo seguente:

$$\log_2\left(\frac{\tilde{x}_g}{\tilde{y}_g}\right) = \log_2\left(\frac{x_g}{y_g}\right) + \frac{300}{\log_2\left(x_g y_g\right)^{4/3}} - 10 \tag{2.12}$$

In Figura 2.4 si possono osservare i dati di espressione dei due modelli dopo la trasformazione appena descritta.

Mettendo a sistema l'equazione di trasformazione del logaritmo del rapporto M, con la condizione che l'intensità media A debba rimanere inalterata, si possono ottenere le espressioni algebriche dell'intensità del singolo gene g dopo la distorsione. Dalla (2.11) si possono ottenere i valori di espressione per il modello GG:

$$\begin{cases} \log_2(\tilde{x}_g\tilde{y}_g) = \log_2(x_g y_g) \\ \log_2\left(\frac{\tilde{x}_g}{\tilde{y}_g}\right) = \log_2\left(\frac{x_g}{y_g}\right) + \frac{300}{\log_2(x_g y_g)+15} - 15 \end{cases} \tag{2.13}$$

da cui:

$$\begin{cases} \tilde{x}_g = x_g \cdot 2^{\frac{300-15\log(x_g y_g)-225}{2\log(x_g y_g)+30}} \\ \tilde{y}_g = y_g \cdot 2^{-\frac{300-15\log(x_g y_g)-225}{2\log(x_g y_g)+30}} \end{cases} \tag{2.14}$$

Dalla (2.12), invece, si possono ricavare i valori di espressione distorti per il modello LNN:

$$\begin{cases} \log_2(\tilde{x}_g\tilde{y}_g) = \log_2(x_g y_g) \\ \log_2\left(\frac{\tilde{x}_g}{\tilde{y}_g}\right) = \log_2\left(\frac{x_g}{y_g}\right) + \frac{300}{\log_2(x_g y_g)^{4/3}} - 10 \end{cases} \tag{2.15}$$

da cui:

$$\begin{cases} \tilde{x}_g = x_g \cdot 2^{\frac{300-10\log(x_g y_g)^{4/3}}{2\log(x_g y_g)^{4/3}}} \\ \tilde{y}_g = y_g \cdot 2^{-\frac{300-10\log(x_g y_g)^{4/3}}{2\log(x_g y_g)^{4/3}}} \end{cases} \tag{2.16}$$

2.3 SIMAGE

Oltre ai dati provenienti dai modelli di simulazione bayesiani gerarchici, si è deciso di utilizzare dati di espressione generati con il *software* SIMAGE[1]. SIMAGE, acronimo di *SImulation of MicroArray Gene Expression data*, è l'implementazione di un modello di simulazione proposto da Albers *et al.* (2006). Il modello di simulazione proposto tiene in considerazione diversi fattori (imputabili alle varie fasi dei protocolli sperimentali e alla tecnologia) che influenzano l'esperimento e che possono essere a loro volta stimati a partire da un *dataset* reale. Questo garantisce una maggiore similarità tra i dati reali e i dati simulati. La notazione utilizzata in questo paragrafo è

[1] liberamente fruibile all'indirizzo web http://bioinformatics.biol.rug.nl/websoftware/simage

leggermente diversa rispetto al paragrafo precedente, la scelta è stata fatta per uniformità al lavoro di Albers *et al.* (2006).

2.3.1 Il modello

Un esperimento di DNA *microarray* consiste di un numero di *spot*, su cui si va ad ibridare il cDNA *target*, disposti in una griglia, divisa in $n_{row} \times n_{col}$ quadranti, ognuno, a sua volta, con un numero di *spot* pari a n_{spot}^2. In questo modo il numero totale di *spot* per ogni *array* è dato da $n_{row} \times n_{col} \times n_{spot}^2$. In totale si ha un numero di replicazioni pari a n_{slide} per ogni esperimento, in cui ogni *spot* fornisce una misura di intensità per il canale rosso e una per il canale verde.

Indichiamo con y_{ijkl} il logaritmo in base due dell'espressione dell'intensità del canale k, nello spot in posizione (i, j) dell'*array* l.

Consideriamo ora il seguente modello complessivo, che consiste di 29 parametri, di cui 6 sono costanti note (si pensi ad esempio al numero di *spot* nell'*array*) per poi analizzarlo in dettaglio elemento per elemento:

$$y_{ijkl} = g_{nl}\left(t_\delta\left(f_{nl}\left(bg_{ijl} + z_{ijkl}\right)\right)\right) \cdot m_{ijl} \tag{2.17}$$

con:

$$z_{ijkl} = G_{g,k} + D_{g,k} + C_k + S_{pin(ij)} + X_{g,k} + \varepsilon_{ijkl} \tag{2.18}$$

dove:

- g è il gene in posizione (i, j) nell'*array* l;
- g_{nl} è la trasformatione dovuta alla quantizzazione dell'intensità e alla saturazione del segnale;
- *pin* (i, j) è la *spot pin* usata per depositare il DNA nello *spot* in posizione (i, j);
- f_{nl} è la trasformazione dovuta alla non linearità nella dipendenza tra il log-rapporto M e l'intensità media A;

- t_δ è una trasformazione che mima il cosiddetto effetto a "coda di pesce" (*fish tail*) spesso osservabile negli MA-Plot di dati reali;
- m_{ijl} è una funzione pari a 0 se lo *spot* è da considerarsi dato mancante, 1 altrimenti;
- bg_{ijl} è il livello di *background*;
- $G_{g,k}$ è il livello di espressione del gene g nel canale k;
- $D_{g,k}$ è il cambiamento di espressione dovuto alla sovra o sotto espressione;
- C_k è il cambiamento di espressione dovuto all'effetto del canale Cy3 o Cy5;
- $S_{pin(ij)}$ simula l'effetto delle diverse *spot pin* utilizzate per depositare il DNA sulla superficie dell'*array*;
- $X_{g,k}$ è l'interazione gene-marcatore, dovuta al fatto che geni diversi interagiscono in modo diverso con le molecole dei marcatori;
- ε_{ijkl} è l'errore casuale legato alle ripetizioni della misurazione.

Nel Paragrafo 2.3.2 si definiranno in modo più formale il livello di espressione del gene, considerando assenza di alcun tipo di distorsione. Mentre nel Paragrafo 2.3.3 si prenderanno in considerazione tutte le variabili che rappresentano distorsioni casuali e sistematiche.

2.3.2 Espressione genica

Assumendo assenza di ogni tipo di errore sistematico, il log-rapporto tra i valori di espressione nei due canali sarà zero se il gene è egualmente espresso, mentre sarà maggiore di zero se il gene è espresso maggiormente nel canale rosso piuttosto che nel canale verde (sovra-espresso), minore di zero se il gene è espresso maggiormente nel canale verde (sotto-espresso):

$$\log_2 \frac{R_g}{G_g} = \log_2 R_g - \log_2 G_g \begin{cases} < 0 & g \text{ sotto espresso} \\ = 0 & g \text{ egualmente espresso} \\ > 0 & g \text{ sovra espresso} \end{cases} \tag{2.19}$$

Indichiamo con y^*_{ijkl} la parte di y_{ijkl} che non dipende da alcun tipo di distorsione sistematica, cioè la parte che identifica propriamente il livello di espressione. Appare chiaro come y^*_{ijkl} sarà uguale in ogni *spot* che descrive l'espressione del gene g; possiamo quindi utilizzare la notazione y_g. Dividiamo y_g in due parti: $G_{g,k}$ che rappresenta l'espressione base del gene g nel canale k e $D_{g,k}$ che rappresenta la possibile variazione dovuta alla sovra o sotto espressione:

$$y^*_{g,k} = G_{g,k} + D_{g,k} \tag{2.20}$$

Consideriamo $G_{g,k}$ come una variabile casuale latente distribuita come una Normale: $G_{g,k} \sim N(\mu, \sigma^2_G)$. In questo modo i geni egualmente espressi si distribuiscono simmetricamente rispetto a μ. Nella maggior parte degli esperimenti di *microarray* i due canali non si possono considerare indipendenti, è necessario quindi introdurre un fattore di correlazione:

$$COV(G_{g,1}, G_{g,2}) = \rho\, \sigma^2_G \tag{2.21}$$

dove ρ è il coefficiente di correlazione tra i livelli di intensità dei due canali.

Per ogni gene infine è necessario specificare la probabilità π_0 di essere egualmente espresso, la probabilità π_+ di essere sovra espresso e la probabilità $\pi_- = 1 - \pi_0 - \pi_+$ di essere sotto espresso.

Se definiamo $D_{g,k}$ come:

$$\begin{pmatrix} D_{g,1} \\ D_{g,2} \end{pmatrix} = \begin{pmatrix} k_1 \\ k_2 \end{pmatrix} \mu_D \tag{2.22}$$

dove:

$$(k_1, k_2) = \begin{cases} (-1, 1) & g \text{ sotto espresso} \\ (0, 0) & g \text{ egualmente espresso} \\ (1, -1) & g \text{ sovra espresso} \end{cases} \tag{2.23}$$

allora il log-rapporto dei geni sovra espressi e sotto espressi è incrementato o diminuito di una quanità pari a $2\mu_D$.

Modellare y_g^* come una combinazione di densità normali, facilita la stima dei numerosi parametri del modello; cosa che una modellazione più fine dal punto di vista biologico non avrebbe permesso.

2.3.3 Variazioni casuali e sistematiche

Oltre alla variabilità biologica nell'espressione dei geni nei due canali, i livelli di espressione vengono alterati da distorsioni sperimentali. SIMAGE prende in considerazione diversi tipi di alterazioni, casuali e sistematiche, che possono più o meno influire sul risultato finale di un esperimento di *microarray*.

Replicazioni tecniche

Spesso per avere maggiore precisione nella misurazione dell'espressione dei geni si preferisce riservare ad ogni gene più di uno spot. Questo procedimento viene detto *multiple spotting* ed è modellato in SIMAGE specificando il numero di replicazioni tecniche da effettuare per ogni gene nell'*array*. Le variazioni nelle replicazioni possono essere dovute alla procedura di ibridazione, all'effetto delle *spot-pin*, o ad altri fattori. Queste variazioni vengono modellate nel seguente modo: si estraggono in modo casuale due errori ε_{ij1l} e ε_{ij2l} da una variabile casuale $N(0, \sigma_\varepsilon^2)$ e si sommano alle due misure ripetute y_{ij1l}^* e y_{ij2l}^*. Ogni gene sarà replicato n_{rep} volte.

Superficie di *background*

Il livello di espressione di un gene è determinato dalla differenza tra l'intensità dello *spot* specifico per quel gene e l'intensità di tutto quello che non è *spot*. Tutta la parte del vetrino che non è *spot* è chiamato *background*. In particolare si intende per *background* il livello di intensità locale intorno allo *spot*. In un *microarray* il valore di intensità del *background* dovrebbe

essere vicino allo zero (non dovrebbe dare fluorescenza). Può capitare però che il processo di ibridazione abbia subito delle alterazioni e che quindi abbia generato della fluorescenza aspecifica su tutto o solo su una parte del vetrino.

Il modo più semplice per modellare il gradiente della superficie del *background* è quello di utilizzare un piano inclinato: questo modello descrive la situazione in cui un lato dell'*array* ha valori di espressione sistematicamente più alti. Un modo più raffinato è quello di utilizzare anzichè un piano una funzione quadratica, ma anche in questo caso il modello risulta troppo semplice, in quanto consente al massimo un solo estremo relativo. Albers *et al.* (2006) propongono un metodo che permette l'esistenza di più estremi relativi: vengono introdotte n_{bg} densità normali bivariate, con posizione nel vetrino generata in modo casuale, così come casuale è la matrice di varianza/covarianza. Tali densità sono poi moltiplicate per un'ampiezza I, estratta da una uniforme in $[-1, 1]$. Si sommano poi tutte le densità e tale somma costituisce il *background*.

Più nello specifico ogni densità bivariata è una Normale con vettore delle medie $(\mu_{bg,X}, \mu_{bg,Y})$, varianze rispettivamente $\sigma^2_{bg,X}$ e $\sigma^2_{bg,Y}$ e correlazione ρ_{bg}. Le varianze derivano dal valore σ_{bg} specificato dall'utente, che rappresenta la deviazione standard media, nel seguente modo:

$$\sigma^2_{bg,X}, \sigma^2_{bg,Y} \sim \left(\sigma_{bg} \min\left(n_{row}, n_{col}\right) n_{spot}\right) U\left(\frac{1}{2}, \frac{3}{2}\right) \tag{2.24}$$

La correlazione ρ_{bg} è invece estratta da una distribuzione $U\left(-\frac{1}{2}, \frac{1}{2}\right)$.

Il *background* così definito è poi sommato ad un gradiente lineare con inclinazione minore di s. La massima inclinazione s è specificata dall'utente. Una pendenza orizzontale s_H e una pendenza verticale s_V sono estratte casualmente da una distribuzione $U(-s, s)$. Per ogni *spot* (i, j) viene quindi definita la seguente quantità, relativa alla superficie lineare di *background*:

$$bg_{lin}(i, j) = \left(i - \frac{n_{col}}{2}\right) s_H + \left(j - \frac{n_{row}}{2}\right) s_V \tag{2.25}$$

Come detto le n_{bg} densità di *background* vengono moltiplicate per un'intensità I e poi sommate insieme; durante questa operazione si somma anche un effetto di rumore casuale estratto da una Normale centrata in zero:

$noise\,(i,j) \sim N\left(0, \sigma^2_{\varepsilon,bg}\right)$:

$$bg\,(i,j) = noise\,(i,j) + \sum_{m=1}^{n_{bg}} I_m f_m\,(i,j) \tag{2.26}$$

Il modello risulta un modello additivo, in cui cioè la parte relativa al *background* viene sommata alla parte relativa all'espressione dei geni. Tuttavia può capitare che il *background* risulti negativo, e in ogni caso è difficile regolare il suo valore in funzione del valore di espressione dei geni. Per questo si applica una trasformazione lineare:

$$bg_{ij} = b\,\mu\,\left(bg\,(i,j) - \min_{i,j} bg\,(i,j)\right) + bg_{lin}\,(i,j) \tag{2.27}$$

dove b (specificato dall'utente) indica la percentuale massima di segnale di *background* relativamente al segnale di espressione genica μ. In questo modo il minimo valore del *background* sarà sempre zero e il massimo sarà il $b\%$ del segnale di espressione.

In Figura 2.5 si può osservare lo schema completo di come viene simulato dal modello il livello di *background* che viene poi sommato al livello di espressione genica.

Canali, interazione gene-marcatore e *spot-pin*

Ognuno di questi tre effetti viene introdotto nel modello in modo simile, andando cioè ad aggiungere al livello di espressione un valore estratto da una distribuzione normale.

Come già esposto nell'introduzione del presente lavoro, in molti esperimenti di *microarray* la media del segnale del canale marcato con Cy3 è significativamente diversa dalla media del segnale del canale marcato con Cy5: le ragioni, oltre ad una differente efficienza globale dei due marcatori, sono da ricercare nelle differenze di quantità di marcatori nella preparazione del DNA e nelle variazioni da gene a gene della capacità di incorporare tali marcatori. Questo effetto viene inserito nel modello estraendo in modo casuale e indipendente una quantità C_1 per il canale Cy3 e una quantità C_2 per il canale Cy5 da una distribuzione $N\,(0, \sigma^2_{channel})$.

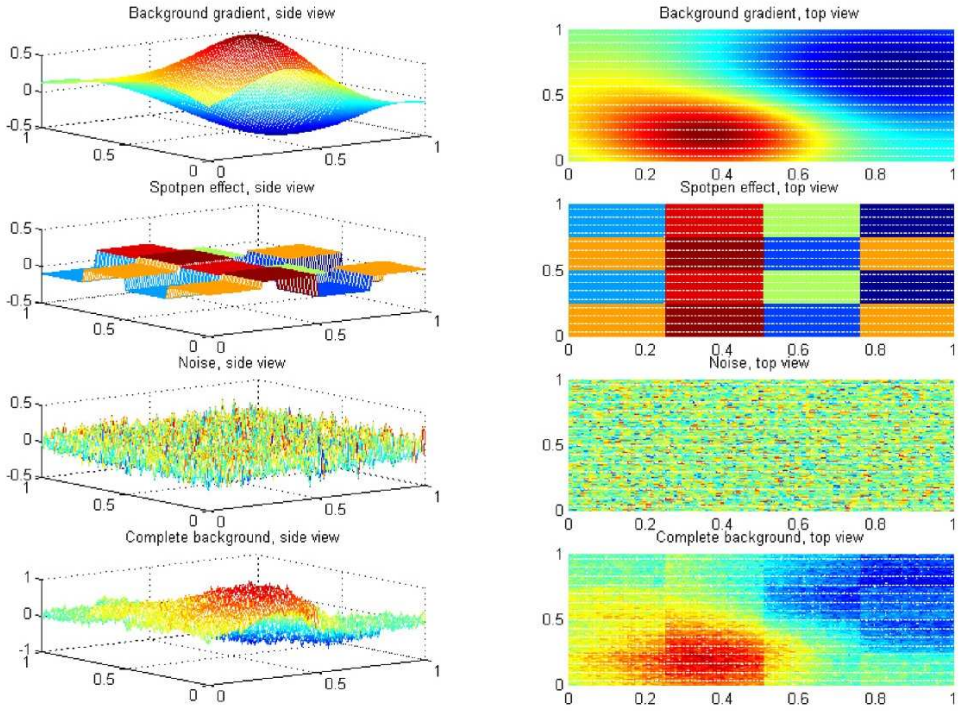

Figura 2.5: Struttura del livello di *background*. A sinistra si può osservare la vista laterale e a destra la vista dall'alto; il colore blu indica un basso livello di *background*, mentre il colore rosso indica un livello elevato di *background*.

La differente capacità da parte dei geni di incorporare le molecole dei marcatori porta alla necessità di inserire nel modello anche un effetto di interazione tra gene e marcatore. A tutte le misure di espressione relative al gene g e al marcatore k viene aggiunta una distorsione casuale X_{gk} estratta dalla distribuzione $N\left(0, \sigma^2_{g\times k}\right)$.

L'uso di differenti *spot-pin* mostra distorsioni sistematiche nella quantità di DNA *probe* depositato sulla superficie del vetrino. Per modellare questo effetto vengono usate n_{pin} estrazioni da una distribuzione $N\left(0, \sigma^2_{pin}\right)$. L'effetto viene considerato uguale per entrambi i canali.

Non linearità

In molti casi gli MA-Plot di esperimenti reali presentano un aspetto non lineare: si osserva una certa curvatura e inclinazione, molto spesso con maggiori deviazioni dalla linearità per valori bassi di espressione media. Per introdurre nel modello la non linearità si utilizza una funzione f_{nl} con due parametri di interesse: α_1 specifica la quantità massima di curvatura, mentre α_2 specifica la massima inclinazione. La trasformazione agisce direttamente sull'MA-Plot, andando a trasformare in funzione di A (media dei logaritmi dell'espressione nei due canali) il valore M che si ricorda essere il log-rapporto tra i segnali dei due canali. La funzione di trasformazione è basata su una funzione polinomiale di grado 2, ma per evitare di applicare una distorsione troppo grande per i valori di A lontani dalla media di espressione μ si introducono due rette tangenti nei punti di ascissa L e R, estratti in modo casuale da:

$$L \sim U\left(\mu - 2.5\sigma_A, \mu - \sigma_A\right) \tag{2.28}$$

$$R \sim U\left(\mu + \sigma_A, \mu + 1.5\sigma_A\right) \tag{2.29}$$

dove σ_A è la deviazione standard delle intensità A. In questo modo la funzione f_{nl} avrà andamento parabolico tra i due punti L e R e lineare all'esterno di essi. L'ascissa del punto di estremo della parabola (C) è anch'essa estratta in modo casuale dalla distribuzione $U\left(\mu - \sigma_A, \mu + \sigma_A\right)$. Si noti che il valore L è più distante di R, questo garantisce che l'effetto di non linearità sia più accentuato per valori bassi di intensità A. I valori dei coefficienti del polinomio sono specificati nel modo seguente:

$$a_1 \sim U\left(-\alpha_1, \alpha_1\right) \tag{2.30}$$

$$a_2 \sim U\left(-\alpha_2, \alpha_2\right) \tag{2.31}$$

Indichiamo infine l'espressione matematica della trasformazione:

$$\tilde{M}_i = M_i + \begin{cases} (a_2 + 2a_1(L - C))A_i + (a_1(C^2 - L^2) - a_2C) + \gamma & se\, A_i < L \\ a_1(A_i - C)^2 + a_2(A_i - C) + \gamma & se\, L < A_i < R \\ (a_2 + 2a_1(R - C))A_i + (a_1(C^2 - R^2) - a_2C) + \gamma & se\, A_i > R \end{cases} \tag{2.32}$$

Code di pesce

Un altro effetto comune visibile negli MA-Plot è noto come effetto "coda di pesce" (*fish tail*). Tale effetto indica che per valori bassi di intensità media A i livelli di espressione tendono ad essere più alterati, questo è dovuto essenzialmente al fatto che gli *spot* con intensità più bassa sono più soggetti agli errori. Un'altra possibile ragione potrebbe essere la sovra-trasformazione dovuta alla trasformazione logaritmica delle espressioni.

L'effetto è inserito nel modello attraverso il parametro δ, che aumenta il log-rapporto degli *spot* con espressione media A_{ijl} inferiore a μ di un fattore pari a:

$$\left(1 + \delta\sigma_M^{-2}(A_{ijl} - \mu)^2\right) \tag{2.33}$$

dove σ_M è la deviazione standard della distribuzione dei log-rapporti M. Valori alti di δ porteranno ad un effetto maggiore, mentre $\delta = 0$ indica assenza di effetto *fish tail*. L'effetto è inserito nel modello in modo casuale; ogni gene egualmente espresso ha probabilità 1/2 di essere influenzato da δ.

In Figura 2.6 si può osservare un esempio di non linearità combinato ad un effetto "coda di pesce". L'MA-Plot è stato generato con SIMAGE.

Quantizzazione e saturazione

Un ulteriore elemento di distorsione proviene dallo scanner ottico che ha il compito di quantificare le intensità luminose dei marcatori. Gli scanner attualmente in commercio usano 16 bit per la descrizione del segnale: i segnali saranno quindi in un *range* compreso tra $\log_2(2^0) = 0$ e $\log_2(2^{16}) = 16$. Inoltre spesso gli scanner tendono a sovrastimare i valori bassi di intensità e

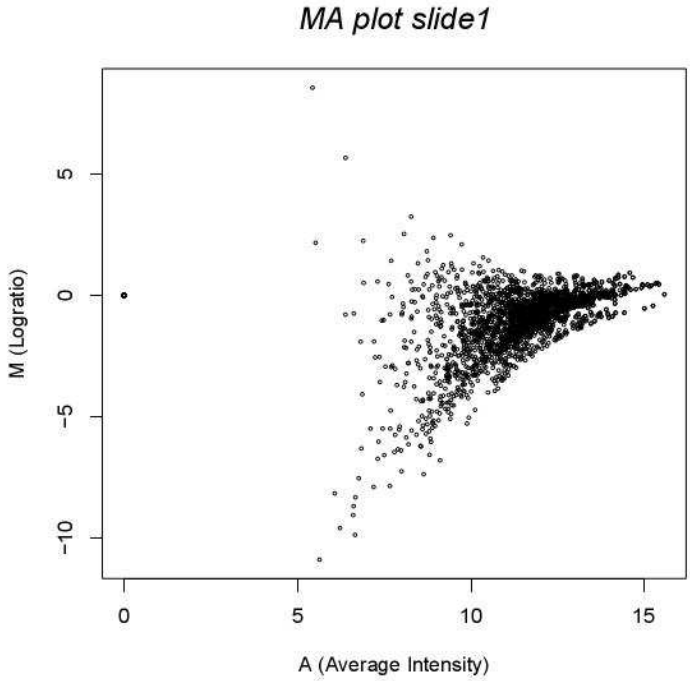

Figura 2.6: Esempio dell'effetto di non linearità e *fish tail* sull'MA-Plot. Si noti come sia presente una inclinazione e una curvatura (non linearità) e come per bassi valori di intensità i livelli di espressione abbiano maggiore rumore casuale (*fish tail*).

a sottostimare i valori alti di intensità. SIMAGE, attraverso un parametro w compreso tra 0 e 1, introduce questa distorsione, modellando una situazione intermedia tra il caso $l_0(y)$ dove si considerano solo i limiti fisici dello scanner e il caso $l_1(y)$ in cui i valori bassi vengono aumentati e i valori alti diminuiti. Tutto questo è inserito nel modello attraverso la funzione:

$$g_{nl} = (1 - w)\, l_0(y) + w l_1(y) \tag{2.34}$$

dove:

$$l_0(y) = \min(\max(0, y), 16) \tag{2.35}$$

$$l_1(y) = \frac{16}{1 + \exp\left(2 - \frac{y}{4}\right)} \tag{2.36}$$

Dati mancanti

Quando si analizzano dati provenienti da esperimenti di *microarray* reali, ci si trova spesso a lavorare con dati mancanti. Le ragioni sono molteplici: imperfezioni nella superficie del vetrino, problemi di ibridazione, mancato deposito del DNA sulla superficie durante la procedura di *spotting*, presenza di polvere o impurità oppure effetto saturazione.

Il presente modello considera tre tipi di dati mancanti:

- segmenti, che simulano la presenza di un pelo sul vetrino;
- corone circolari, che comprendono diversi *spot* e che simulano la presenza di un'impurità;
- singoli *spot* disposti casualmente.

Il numero di occorenze di ogni tipo di dato mancante, così come la sua massima estensione, sono specificabili dall'utente. La posizione sul vetrino è invece casuale, in modo uniforme su tutto l'*array*.

In Figura 2.7 si può osservare una rappresentazione delle diverse occorrenze di valori mancanti simulate dal modello.

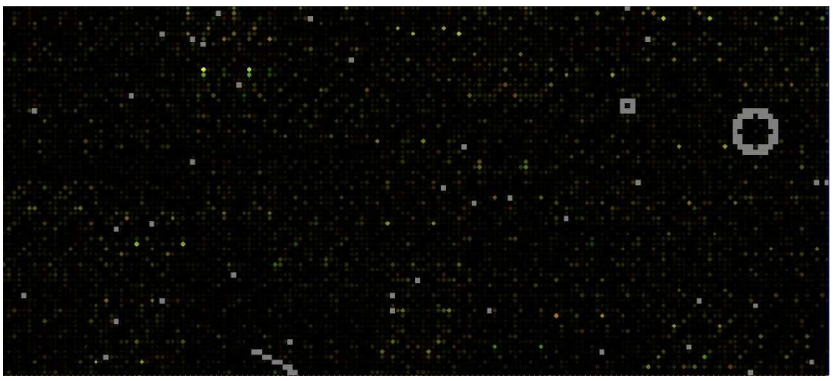

Figura 2.7: Esempio di dati mancanti in un *array*, indicati in grigio. Si noti la presenza di corone circolari con raggio diverso, di segmenti e di singoli *spot*.

Capitolo 3

Normalizzazione e metodi statistici

3.1 Normalizzazione globale

Come già accennato nell'introduzione dell' elaborato, un'operazione fondamentale per la buona riuscita di un esperimento di *microarray* è la normalizzazione dei dati. La normalizzazione, infatti, permette il confronto tra valori di espressione derivanti da esperimenti diversi, riducendo o eliminando errori e distorsioni che potrebbero alterare la reale regolazione dei geni nelle due condizioni in esame.

La più semplice normalizzazione dei dati è costituita dalla "normalizzazione globale", che consiste nello spostamento della media (o mediana) della distribuzione dei log-rapporti sullo zero. Se indichiamo con x_g e y_g (con $g = 1, \ldots, p$) i vettori dei livelli di espressione del gene g rispettivamente nella prima e nella seconda condizione sperimentale e con m la media della

distribuzione dei log-rapporti, si ottiene:

$$\log_2 \frac{\tilde{x}_g}{\tilde{y}_g} = \log_2 \frac{x_g}{y_g} - m \tag{3.1}$$

$$= \log_2 \frac{x_g}{y_g} - \log_2 2^m \tag{3.2}$$

$$= \log_2 \frac{x_g}{y_g \cdot 2^m} \tag{3.3}$$

dove $\tilde{x}_g$ e $\tilde{y}_g$ sono i valori di espressione normalizzati del gene g rispettivamente nella prima e nella seconda condizione sperimentale. La normalizzazione si può quindi ridurre ad una trasformazione del livello di espressione del secondo canale:

$$\tilde{y}_g = y_g \cdot 2^m \tag{3.4}$$

3.2 Regressione locale

Yang *et al.* (2002) propongono un'ulteriore trasformazione dei dati basata su una regressione locale di tipo *lowess* sui valori dell'MA-Plot da farsi successivamente alla normalizzazione globale. Una regressione locale permette infatti di normalizzare i dati considerando anche le distorsioni dipendenti dall'intensità degli *spot* (*intensity-dependent normalization*). La normalizzazione avviene in due passaggi: si stima l'andamento dei punti nel grafico attraverso la regressione locale e poi si sottrae tale curva ai dati per eliminare distorsioni sistematiche dipendenti dall'intensità. Ricordando che $M_g = \log_2 (x_g/y_g)$ (con $g = 1, \ldots, p$) indica il log-rapporto delle intensità nei due canali, allora si può scrivere:

$$M_g^{'} = M_g - c\,(A_g) \tag{3.5}$$

dove $c\,(A_g)$ è la funzione *lowess* interpolata sull'MA-Plot, considerando M_g come risposta e $A_g = \frac{1}{2} \log_2 (x_g\, y_g)$ come esplicativa.

La funzione utilizzata per la regressione locale è la *robust local weighted regression* introdotta da Cleveland (1979). *Lowess* è un metodo di lisciamento che permette di riassumere dati multivariati, attraverso curve (o superfici)

regolari. Il lisciamento avviene interpolando localmente una funzione polinomiale di grado d; questa operazione quindi equivale al calcolo di una media mobile. Nella procedura, la curva di regressione è calcolata usando il metodo dei minimi quadrati pesati in cui i pesi sono scelti in modo inversamente proporzionale alla distanza dei punti stessi dalla curva di regressione. In questo modo i punti più vicini influenzano maggiormente la regressione, rispetto a quelli più lontani. Per ogni punto (A_g, M_g) si considerano i valori $\hat{\beta}_j(A_g)$ (con $j = 0, \ldots, d$) che minimizzano la quantità:

$$\sum_{k=1}^{p} \left(M_k - \beta_0 - \beta_1 A_k - \ldots - \beta_d A_k^d\right)^2 w_k(A_g) \tag{3.6}$$

dove A e M sono rispettivamente ascissa e ordinata del grafico da "lisciare" e d è il grado del polinomio usato per la regressione locale, $w_k(A_g)$ è un sistema di pesi così definito:

$$w_k(A_g) = W\left(\frac{1}{h_i}(A_k - A_g)\right) \tag{3.7}$$

in cui $W(\cdot)$, è una funzione simmetrica (tipicamente biquadratica o tricubica) e centrata in A_g, che assegna pesi positivi in $(-1, 1)$, e nulli al di fuori dell'intervallo; h è detto "parametro di lisciamento" e influisce su quanto la curva seguirà l'andamento dei punti: valori maggiori o minori di h rendono la curva rispettivamente più o meno liscia. Nella funzione *lowess* il parametro h è in funzione della frazione di punti, α, considerati per la regressione. Yang *et al.* (2002) suggeriscono di utilizzare un valore di α pari a 0.4. Il parametro h_g è calcolato come la distanza tra A_g e l'r-esimo punto più vicino, dove r è la parte intera di $n \cdot \alpha$.

Cleveland (1979) propone inoltre un sistema di pesi robusti, basati su una funzione biquadratica da applicare sui residui della regressione. I pesi sono definiti nel modo seguente:

1. Siano $e_k = M_k - \hat{M}_k$ i residui della regressione sopra descritta, sia s la mediana dei $|e_k|$ e sia B la funzione biquadratica definita da:

$$B(x) = \begin{cases} (1 - x^2)^2 & |x| < 1 \\ 0 & |x| \geq 1 \end{cases} \tag{3.8}$$

2. Si calcoli:
$$\delta_k = B\left(\frac{e_k}{6\,s}\right) \tag{3.9}$$

3. Si calcolino le nuove stime di $\hat{M}_k$ utilizzando come sistema di pesi la funzione $\delta_k\, w_k\,(A_g)$

4. Si ripeta la procedura dal punto 1) al punto 3) N volte.

Utilizzare una regressione robusta, mette al riparo da deviazioni dovute alla presenza di *outlier*. Nel caso dei *microarray* questo aspetto è molto importante perchè permette di normalizzare l'andamento basandosi principalmente sui dati di espressione equivalente, senza che la funzione di normalizzazione sia influenzata dai pochi geni differenzialmente espressi (caratterizzati da valori di log-rapporto molto grandi).

3.2.1 *Print-tip Lowess*

Oltre alle distorsioni dipendenti dall'intensità media A, spesso si osservano nei dati di *microarray* distorsioni dipendenti dalle diverse punte che si utilizzano per depositare il DNA sul vetrino (*print-tip group*). Per questo motivo Yang *et al.* (2002) propongono un metodo di normalizzazione dipendente contemporaneamente dall'intensità A e dal gruppo di punte utilizzato per depositare il DNA (*(print-tip + A)-dependent normalization*). Questa procedura è nota come *Print-tip Lowess*, e consiste semplicemente nel normalizzare i dati attraverso un numero I di curve *lowess*, utilizzando per l'i-esima regressione gli *spot* creati con l'i-esimo gruppo di punte:
$$M^{'} = M - c_i\,(A) \tag{3.10}$$
dove $c_i\,(A)$ (con $i = 1, \ldots, I$) è la funzione *lowess* interpolata sull'MA-Plot dei punti depositati dall'i-esimo *print-tip*.

Si suppone infine che i log-rapporti dell'i-esimo *print-tip group* si distribuiscano come una Normale di media nulla e varianza dipendente dal gruppo di *print-tip*: $N\,(0, a_i^2\sigma^2)$, dove σ^2 è la varianza dei log-rapporti e a_i^2 è un fattore di scala per l'i-esimo gruppo di punte. Per ottenere il fattore di scala per

la normalizzazione spaziale si stima e si elimina a_i^2. Considerando il vincolo naturale dato dalla somma $\sum_{i=1}^{I} \log a_i^2 = 0$, una stima robusta per a_i è:

$$\hat{a}_i = \frac{MAD_i}{\sqrt[I]{\prod_{i=1}^{I} MAD_i}} \tag{3.11}$$

dove la deviazione mediana assoluta (*median absolute deviation*, MAD) è definita da:

$$MAD_i = median_j \left\{ |M_{ij} - median_j \left(M_{ij} \right)| \right\} \tag{3.12}$$

dove M_{ij} indica il j-esimo log-rapporto dell'i-esimo *print-tip group*.

3.3 Regressione locale ottima

Un metodo di normalizzazione simile a quello di Yang *et al.* (2002), almeno nella filosofia iniziale, è proposto in Futschik e Crompton (2004). In quest'ultimo lavoro infatti si discutono due modelli di normalizzazione che tengono conto di una dipendenza non lineare dall'intensità e dalla posizione degli *spot*. Tali modelli sono anch'essi basati su una regressione locale; la novità introdotta riguarda i parametri usati per il lisciamento. Futschik e Crompton (2004) osservano infatti che spesso la scelta di tali parametri viene lasciata al singolo ricercatore, e al più vengono forniti dei valori "consigliati". I modelli qui proposti, invece, prevedono una procedura per ottenere i parametri ottimi, attraverso una *cross-validation* generalizzata (GCV).

Lo schema proposto per la normalizzazione è basato su una procedura iterativa di regressione locale. La regressione locale è calcolata attraverso la funzione *locfit* (Loader (1999)), basata sullo stesso modello del metodo *lowess*. Esattamente come *lowess*, *locfit* richiede la scelta di un parametro α, che indica la frazione di punti da considerare nella regressione e che controlla quindi il parametro di lisciamento h.

La scelta dei parametri è cruciale per il "lisciamento" della curva di regressione e quindi per l'efficienza della normalizzazione. Per questo Futschik e

Crompton (2004) introducono una procedura di ottimizzazione dei parametri attraverso la GCV.

La procedura standard di *cross-validation* divide i dati in k parti, di cui $k-1$ vengono utilizzate per la costruzione del modello e la restante per la validazione. Questa operazione è ripetuta k volte in modo che ogni parte sia utilizzata per la costruzione e per la validazione. Si stima poi l'errore di previsione mediando l'errore quadratico medio nei k passaggi. Se la *cross-validation* è utilizzata per il confronto tra più modelli, viene selezionato il modello che produce l'errore di previsione più basso. Generalmente la scelta di k equivale al numero di punti (*leave-one-out cross-validation*).

Dal momento che la *cross validation* ha un costo computazionale proibitivo per il numero di dati provenienti da un esperimento di *microarray*, si introduce la GCV che è meno pesante dal punto di vista computazionale. La GCV è una procedura che approssima i risultati ottenuti attraverso la *leave-one-out cross-validation* senza la necessità di costruire più modelli di regressione. Se indichiamo con $\hat{\mu}(\cdot)$ il modello di regressione locale, possiamo così definire il criterio di GCV:

$$GCV(\hat{\mu}) = n \frac{\sum_{i=1}^{n} (Y_i - \hat{\mu}(x_i))^2}{(n-\nu)^2} \tag{3.13}$$

dove n è il numero di punti considerati nella regressione e ν indica i gradi di libertà equivalenti della curva. Per i dettagli su come calcolare i gradi di libertà si veda Craven e Wahba (1979).

Indichiamo con α_A il parametro di lisciamento per la regressione locale di M rispetto ad A, con α_{XY} il parametro di lisciamento per la regressione locale di M rispetto alle coordinate spaziali X e Y e infine con s_Y il parametro di scala che permette di avere un diverso lisciamento nella direzione di Y rispetto a quella di X.

3.3.1 *Optimized Local Intensity-dependent Normalization*

Il primo modello descritto in Futschik e Crompton (2004) è chiamato *Optimized Local Intensity-dependent Normalization* (OLIN); in esso vengono effettuate diverse regressioni locali di M rispetto ad A, con un insieme di valori per il parametro α_A. Si confrontano i modelli sulla base della GCV e si indica con α_A^* il parametro ottimo. Si sottrae quindi la curva di regressione calcolata con tale parametro:

$$M^{'} = M - c_{\alpha_A^*}(A) \tag{3.14}$$

Analogamente si considerano poi diverse curve di regressione di $M^{'}$ rispetto alle coordinate spaziali, con un insieme di valori per il parametro α_{XY} e un insieme di valori per il parametro s_Y. Si confrontano i modelli sulla base della GCV e si indica la curva ottimale con $c_{\alpha_{XY}^* s_Y^*}$:

$$M^{''} = M^{'} - c_{\alpha_{XY}^* s_Y^*}(A) \tag{3.15}$$

La procedura viene ripetuta finché il numero massimo di iterazioni N non è raggiunto. Dopo le N iterazioni $M^{''}$ indica il log-rapporto delle intensità normalizzato.

Per le normalizzazioni effettuate nel presente elaborato si sono utilizzati i seguenti valori dei parametri, in base alle indicazioni presenti in Futschik e Crompton (2004):

- per i parametri α_A e α_{XY} si è scelto il valore ottimale nell'intervallo $[0.2, 1]$ con un passo di 0.2;
- per il parametro s_Y si è scelto tra i valori $\{0.1, 1, 10, 20\}$;
- il numero di iterazioni è stato fissato a 3.

3.3.2 *Optimized Scaled Local Intensity-dependent Normalization*

Il secondo dei modelli proposti, detto *Optimized Scaled Local Intensity-dependent Normalization* (OSLIN), è un'ulteriore normalizzazione sui dati già precedentemente normalizzati con OLIN. Infatti sebbene OLIN normalizzi i dati anche rispetto a variabili spaziali, è ancora possibile che vi siano distorsioni spaziali nell'*array*. E' possibile, infatti, che le zone dell'*array* con più alto valore assoluto di M mostrino meno variabilità delle regioni con valore assoluto più basso. Per questo motivo in OSLIN è stato introdotto un fattore di scala che dipende dal valore assoluto di M.

Come prima cosa bisogna normalzzare i log-rapporti delle intensità attraverso il modello OLIN; sui valori così ottenuti si considera un insieme di parametri di lisciamento α e un insieme di parametri di scala s e si calcolano le corrispondenti regressioni locali sui punti considerando come risposta i valori assoluti di M e come esplicative le coordinate X e Y. Si confrontano i modelli attraverso la GCV e si scelgono i valori ottimi α^* e s^*: si considera la curva ottimale $c^{abs}_{\alpha^* s^*}$ come fattore di scala:

$$M''' = \frac{M''}{c^{abs}_{\alpha^* s^*}} \tag{3.16}$$

M''' è poi aggiustato in modo che la variabilità di M'' rimanga inalterata:

$$M^{IV} = M''' \sqrt{\frac{\text{var}\,(M'')}{\text{var}\,(M''')}} \tag{3.17}$$

M^{IV} è il log-rapporto normalizzato.

Si sono considerati per le normalizzazioni i seguenti valori dei parametri:

- per il parametro α si è scelto il valore ottimale nell'intervallo $[0.2, 1]$ con un passo di 0.2;
- per il parametro s si è scelto tra i valori $\{0.1, 1, 10, 20\}$.

Normalizzare utilizzando come fattore di scala il valore assoluto di M sottointende la seguente assunzione: la variabilità dei log-rapporti all'interno

dell'*array* è costante. In generale tale assunzione è rispettata purchè i geni siano disposti casualmente nell'*array*.

3.4 Reti Neurali

Un altro possibile approccio per la normalizzazione delle distorsioni spaziali e dipendenti dall'intensità è descritto nel lavoro di Tarca e Cooke (2005) che propone un modello basato su reti neurali robuste. Il log-rapporto M viene corretto attraverso una funzione che dipende dall'intensità A e dalla posizione dello *spot* nell'*array* attraverso una rete neurale. La robustezza nei confronti degli *outlier* è garantita assegnando ad ogni *spot* un peso che dipende dalla distanza tra il log-rapporto calcolato in quello *spot* e la mediana della distribuzione dei log-rapporti di geni con valore simile di intensità A, oltre che utilizzando delle pseudo-coordinate, anzichè il numero di riga e di colonna dello *spot*.

Il modello basato su rete neurale utilizza come regressori l'intensità media A e le coordinate spaziali dello *spot*. Invece di considerare direttamente come coordinate spaziali il numero di riga e di colonna dello *spot*, si dividono i *sub-array* costituiti dagli *spot* depositati dalla stessa *print-tip* in blocchi quadrati di 9 *spot*. Ogni *spot* all'interno di ogni blocco avrà la stessa coordinata spaziale. Usando questa procedura si mantiene inalterata la differenza tra l'intensità degli *spot*, che può dipendere dalle variazioni biologiche nei due gruppi, eliminando la variazione spaziale con un livello di "risoluzione" più basso di quello che si avrebbe lavorando sulle coordinate del singolo *spot*.

Nel modello proposto si sottrae alla distribuzione dei log-rapporti una funzione stimata attraverso la rete neurale, ottenendo così i valori normalizzati:

$$M^{'} = M - c_i\left(A, X, Y\right) \tag{3.18}$$

dove per ogni *print-tip* i (con $i = 1, \ldots, I$) si considera una diversa curva c_i.

Senza il bisogno di supporre alcuna forma parametrica particolare, le reti neurali sono in grado di stimare complesse relazioni non lineari tra più

variabili esplicative (dette di *input*) e più variabili risposta (dette di *output*), utilizzando un insieme di dati per il *training* della rete e un insieme per la validazione del modello. In questo caso si considera come unica variabile risposta M e come esplicative A, X e Y.

Possiamo scrivere la forma parametrica della rete neurale nel seguente modo:

$$f\left(\underline{z}, \underline{w}\right) = \sigma^{(2)}\left(\sum_{j=1}^{J+1}\left(w_j \sigma_j^{(1)}\left(\sum_{i=1}^{L+1}\left(z_i w_{i,j}\right)\right)\right)\right) \tag{3.19}$$

dove:

- $\underline{z}$ è il vettore che ha come componenti gli $L = 3$ fattori X, Y e A più un valore costante pari a 1;
- $w_{i,j}$ sono i parametri della rete, detti "pesi";
- $\sigma_j^{(1)}$ sono le funzioni di attivazione che legano le variabili di *input* ai nodi latenti;
- $\sigma^{(2)}$ è la funzione di attivazione che lega la risposta ai nodi latenti;
- J è il numero di nodi nello strato latente.

Entrambe le funzioni $\sigma_j^{(1)}$ e $\sigma^{(2)}$ sono funzioni sigmoidee. La funzione sigmoidea è definita come segue:

$$\sigma\left(x\right) = \frac{1}{1 + e^{-x}} \tag{3.20}$$

Una rete neurale con funzione di attivazione sigmoidea è sufficiente ad approssimare bene qualsiasi tipo di funzione, purchè i nodi latenti siano in numero sufficiente. Dopo aver scelto un valore ragionevole per J, si crea la rete neurale utilizzando un insieme di dati detto di *training*. I classici algoritmi per la creazione delle reti neurali minimizzano la somma degli scarti al quadrato tra i veri valori e i valori predetti dalla rete, per identificare i valori ottimi per i pesi:

$$\min_{w} \sum_{k=1}^{p}\left(f\left(z_k, \underline{w}\right) - M_k\right)^2 \tag{3.21}$$

dove p è il numero totale di *spot*.

Il principale problema in cui si può incorrere approssimando una funzione attraverso una rete neurale, è quello di provocare un sovra-adattamento ai dati; in questo caso la curva segue troppo le fluttuazioni casuali dei punti senza descriverne l'andamento generale. Per evitare questo tipo di problema Tarca e Cooke (2005) consigliano di inserire nello strato latente un numero basso di nodi (i.e. $J = 3$).

Le stime dei parametri ottenute con il classico approccio di minimizzare la somma degli scarti al quadrato, descritto dalla (3.21), non sono molto robuste in presenza di valori estremi. Considerare coordinate uguali per blocchi di *spot* permette ai risultati di non essere influenzati dai geni il cui valore di M è atipico rispetto agli *spot* vicini. Per ottenere un metodo più robusto agli *outlier* in funzione di A gli autori considerano un sistema di pesi, descritto dalla seguente procedura:

1. si considerano i punti $(M, A)_k$ con $k = 1, \ldots, p$, ordinati in modo crescente rispetto ad A, in modo che $A_{(1)} \leq A_{(2)} \leq \ldots \leq A_{(n)}$;

2. si dividono i punti in B gruppi (i.e. $B = 20$) in modo che ogni gruppo abbia circa lo stesso numero di punti e preservando l'ordine su A. Si indicano con M_k^l i valori dei punti con $l = 1, \ldots, B$, $k = 1, \ldots, N_l$ e $\sum_{l=1}^{B} N_l = p$;

3. per ogni valore M_k^l si sottrae la mediana dei valori nel gruppo l e si ottengono i nuovi valori $\breve{M}_k^l$, in modo che per ogni gruppo la mediana risulti spostata sullo 0;

4. si trasformano linearmente i $\breve{M}_k^l$ ottenendo $\tilde{M}_k^l$ in modo che:

$$\max\left(\tilde{M}^l\right) = 1 \quad se \quad \left|\max\left(\breve{M}^l\right)\right| > \left|\min\left(\breve{M}^l\right)\right| \tag{3.22}$$

   oppure:

$$\min\left(\tilde{M}^l\right) = -1 \quad altrimenti \tag{3.23}$$

   dopo questa trasformazione i valori ottenuti sono centrati in 0 e appartengono all'intervallo $[-1, 1]$;

5. si ottengono i pesi per ogni *spot* k nel gruppo l, attraverso una tricubica:

$$W_k^l = \left(1 - \left|\tilde{M}_k^l\right|^3\right)^3 \tag{3.24}$$

La procedura appena descritta assegna a tutti gli *spot* pesi compresi in $[0, 1]$, basati su quanto è distante il valore M rispetto alla mediana dei valori degli *spot* con valore simile di A.

Nel presente elaborato si sono considerate reti neurali con $J = 3$ nodi latenti, che corrispondono a 16 parametri da stimare. Idealmente per ottenere il fattore di normalizzazione sarebbe opportuno utilizzare una rete neurale costruita su tutti i punti tranne il k-esimo (usato per la validazione) al variare di k, ma tale operazione risulta troppo pesante dal punto di vista computazionale, perciò si è scelto di utilizzare una *cross-validation* basata sulla divisione dei punti in 4 gruppi in cui ogni gruppo è per 3 volte nell'insieme di *training* e per una volta nell'insieme di validazione.

3.5 *Q-spline*

Il metodo di normalizzazione proposto in Workman *et al.* (2002), al contrario di quelli precedentemente descritti, è stato proposto per i *microarray* a singolo canale, ma trova applicazione anche nei *microarray* a cDNA. Come per i precedenti anche questo modello considera una normalizzazione spaziale e dipendente dall'intensità. Il modello, che prende il nome di *q-spline*, utilizza i quantili della distribuzione dell'intensità di ogni canale separatamente e quelli di una distribuzione "*target*" (v) per calcolare *spline* di lisciamento. Le *spline* sono quindi usate come funzioni di normalizzazione dipendente dall'intensità.

Nel caso specifico dei *microarray* a cDNA si normalizza uno dei due canali utilizzando come *target* l'altro canale. Nel presente elaborato si è deciso di effettuare prima della normalizzazione attraverso il modello *q-spline* una normalizzazione globale, in modo che le intensità dei due canali fossero

confrontabili. Successivamente si sono considerate e confrontate due normalizzazioni, facendo svolgere il ruolo di *target* prima al canale rosso e poi al canale verde. Questo approccio è utilizzato anche in Wu *et al.* (2005) dove il modello è applicato a esperimenti di *microarray* a due canali, proponendo quindi due modelli di normalizzazione:

- *q-splineR* in cui i valori di espressione di entrambi i canali di ogni *array* sono normalizzati utilizzando come *target* la media geometrica del canale rosso;
- *q-splineG* in cui i valori di espressione di entrambi i canali di ogni *array* sono normalizzati utilizzando come *target* la media geometrica del canale verde.

Per prima cosa si definisce il *target* calcolato come la media geometrica del canale rosso (caso *qsplineR*):

$$v_i = \left(\prod_{j=1}^{n} x_{ij} \right)^{\frac{1}{n}} \tag{3.25}$$

o del canale verde (caso *qsplineG*):

$$v_i = \left(\prod_{j=1}^{n} y_{ij} \right)^{\frac{1}{n}} \tag{3.26}$$

dove n è il numero di *array* e $i = 1, \ldots, p$ indica lo *spot*. Per ogni canale x_j e y_j e per il vettore v si calcolano i percentili, che indichiamo rispettivamente con $q_{x,j}$, $q_{y,j}$ e q_v. Per ogni coppia $(q_{\cdot,j}, q_v)$ si costruisce una *spline* cubica $s_j = f\left(q_{\cdot,j}, q_v\right)$, dove f è un generatore di *spline* che adatta i parametri di una *spline* cubica naturale (*B-spline*). Per la creazione di tale funzione Workman *et al.* (2002) utilizzano la funzione *splinefun* presente nel pacchetto base di R. Per i dettagli si rimanda a Hastie (1992).

La funzione di interpolazione definita sul k-esimo intervallo, con i parametri a_{jk} e z_{jk} è data da:

$$s_j(x) = a_{jk1} + a_{jk2}(x - z_{jk}) + a_{jk3}(x - z_{jk})^2 + a_{jk4}(x - z_{jk})^3 \tag{3.27}$$

In un'ottica iterativa, i quantili utilizzati sono estratti in modo casuale dai percentili, corretti per un fattore di compensazione $p \cdot p_0/N$, dove p_0 è il primo percentile e N è il numero di iterazioni. Questo assicura un differente insieme di quantili equispaziati per ogni curva da adattare ai dati. Si è considerato un numero di iterazioni pari a 5 e si sono poi mediati i risultati.

Per tenere conto della distorsione spaziale si utilizza una finestra mobile di 10×10 *spot* centrata su ogni *spot*. Questa finestra definisce gli spot tra loro "vicini". Viene poi definito un sistema di pesi basato sulla distanza euclidea degli *spot* dal centro della finestra usando una distribuzione Normale con deviazione standard pari a 3.

3.6 Trasformazione per la stabilizzazione della varianza

Nel presente paragrafo viene presentata una trasformazione introdotta in Durbin *et al.* (2002) per stabilizzare la varianza nei dati da *microarray*. Questa trasformazione dei dati consente di ottenere una distribuzione simmetrica e la cui varianza non dipende dal valore della media, ma è costante lungo tutto il *range* dei valori di espressione. Consideriamo questa trasformazione al pari delle normalizzazioni in quanto Durbin *et al.* (2002) suggeriscono che dopo aver trasformato in questo modo i dati di espressione le assunzioni su cui sono basate molte delle più comuni tecniche statistiche (come l'ANOVA o la regressione) vengono rispettate. I dati così trasformati sono quindi pronti per essere testati attraverso il test SAM senza alcun bisogno di ulteriori trasformazioni.

Durbin *et al.* (2002) assumono che il livello di espressione dei dati proveniente da *microarray* possa essere modellato come segue:

$$Y_g = \alpha_g + \mu_g \, e_g^{\eta} + \varepsilon_g \tag{3.28}$$

dove Y_g è la variabile che misura il livello di espressione del gene g per un singolo canale, α_g è il rumore medio di *background* nel relativo *spot*, μ_g è il vero

valore di espressione del gene, mentre η_g e ε_g sono termini di errore distribuiti secondo una distribuzione Normale: $\eta_g \sim N\left(0, \sigma_\eta^2\right)$ e $\varepsilon_g \sim N\left(0, \sigma_\varepsilon^2\right)$.

A livelli di espressione bassi (i.e. $\mu_g \to 0$) la misura dell'espressione può essere approssimata da:

$$Y_g \approx \alpha_g + \varepsilon_g \tag{3.29}$$

Questo implica che la misura di espressione si distribuisce approssimativamente come una Normale con varianza costante: $Y_g \sim N\left(\alpha_g, \sigma_\varepsilon^2\right)$.

Analogamente a livelli di espressione alti (i.e. $\mu_g \to +\infty$) il secondo termine della (3.28) domina sugli altri:

$$Y_g \approx \mu_g \, e^{\eta_g} \tag{3.30}$$

In questo caso la misura di espressione si distribuisce approssimativamente come una Log-Normale: $Y_g \sim logN\left(\log \mu_g, \sigma_\eta^2\right)$. Si noti che in questo caso la deviazione standard di Y_g varia linearmente con la media dell'espressione μ_g, dal momento che:

$$Var\left(Y_g\right) \approx \mu_g^2 S_\eta^2 \tag{3.31}$$

dove:

$$S_\eta^2 = e^{\sigma_\eta^2}\left(e^{\sigma_\eta^2} - 1\right) \tag{3.32}$$

Quando il livello di espressione μ_g ha un valore intermedio tra i due casi estremi appena considerati, la misura dell'espressione Y_g si distribuisce come una combinazione lineare di una Normale e di una Log-Normale. La varianza di Y_g si può quindi scrivere come:

$$Var\left(Y_g\right) = \mu_g^2 \, S_\eta^2 + \sigma_\varepsilon^2 \tag{3.33}$$

Anche in questo caso la varianza dipende dalla media di espressione μ_g, ma in un modo più complicato, rendendo la struttura di errore dei dati da *microarray* piuttosto complessa. Si cerca quindi una trasformazione dei dati che stabilizzi la varianza asintotica su tutto il *range* dei dati.

Si vuole trovare una funzione $f\left(\cdot\right)$ sufficientemente liscia, tale che la varianza asintotica della trasformata sia costante. Indicando con $AV\left(Y_g\right)$ la

varianza asintotica di Y_g, si ha:

$$AV\left(f\left(Y_g\right)\right) = f'\left(Y_g\right)^2 Var\left(Y_g\right) = k^2 \tag{3.34}$$

dove f' indica la derivata prima di f e k^2 è costante.

Dopo alcuni passaggi si ottiene il seguente risultato:

$$f'\left(Y_g\right)^2 = \frac{k^2}{Var\left(Y_g\right)} \tag{3.35}$$

$$f'\left(Y_g\right) = \frac{k}{\sqrt{\mu_g^2 S_\eta^2 + \sigma_\varepsilon^2}} \tag{3.36}$$

da cui:

$$\int f'\left(Y_g\right) dy = \int \frac{k}{\sqrt{\mu_g^2 S_\eta^2 + \sigma_\varepsilon^2}} dy \tag{3.37}$$

Una delle soluzioni dell'integrale (3.37) è data dalla funzione:

$$f\left(Y_g\right) = \ln\left(Y_g - \alpha_g + \sqrt{\left(Y_g - \alpha_g\right)^2 + c}\right) \tag{3.38}$$

con $c = \sigma_\varepsilon^2 / S_\eta^2$.

La trasformazione $f(\cdot)$ prende il nome di *Generalized Logarithm Transformation* (GLOG), introdotta per la prima volta nel contesto dei *microarray* da Munson (2001). La funzione $f(\cdot)$ è monotona crescente per tutti i valori di Y_g, positivi o negativi. Per valori di μ_g tendenti a zero la funzione è approssimativamente lineare, mentre per valori elevati di μ_g assume una distribuzione logaritmica. La varianza asintotica dei dati trasformati è costante e uguale a S_η^2.

3.7 Imputazione dei dati mancanti

Molto spesso può capitare che in un esperimento reale di *microarray* vi siano valori mancanti. Questo può accadere per ragioni diverse, quali la presenza di impurità sul vetrino, la risoluzione insufficiente dello scanner o un'ibridazione aspecifica che aumenta il valore del *background*. Quando si lavora con la trasformazione logaritmica dei dati, spesso agli *spot* mancanti viene sostituito

il valore zero, o meno frequentemente il valore medio della riga (i.e. delle altre misure di espressione dello stesso gene). Questi metodi di imputazione, però, non tengono conto della struttura di correlazione dei dati. Troyanskaya *et al.* (2001) propongono un metodo basato sull'algoritmo dei *k-nearest neighbours* (KNN): tale algoritmo considera i geni con profilo di espressione simile a quello del gene che presenta il dato mancante per la sua imputazione.

Supponiamo che il gene g abbia un valore mancante nell'esperimento l. Il metodo proposto seleziona k geni con misura di espressione non mancante nell'esperimento l e con profilo di espressione il più simile possibile a quello del gene g, negli $n-1$ restanti esperimenti. Una media pesata dei valori per l'esperimento l dei k geni selezionati è quindi utilizzata per imputare il dato mancante. Come indice di similarità utilizzato per pesare il contributo alla media di ogni gene viene utilizzata la distanza euclidea.

3.8 Test SAM

Come già accennato nel Capitolo 1, si è deciso di utilizzare come test per individuare i geni differenzialmente espressi il test SAM (*Significance Analysis of Microarray*) proposto da Tusher *et al.* (2001). Come detto il test SAM viene effettuato gene per gene, fornendo un livello di significatività per ciascun gene considerato nell'esperimento.

Ricordiamo che $x_g = (x_{g,1}, \ldots, x_{g,n})$ indica l'espressione del gene g nella prima condizione e che $y_g = (y_{g,1}, \ldots, y_{g,n})$ indica l'espressione del gene g nella seconda condizione (con $g = 1, \ldots, p$). Possiamo a questo punto considerare un test che abbia come ipotesi nulla l'uguaglianza nell'espressione del gene nelle due condizioni e come ipotesi alternativa l'espressione differenziale nei due canali:

$$H_{0,g} : \mu_x(g) - \mu_y(g) = 0$$
$$H_{1,g} : \mu_x(g) - \mu_y(g) \neq 0$$

con $g = 1, \ldots, p$, dove $\mu_x(g)$ è il valore atteso dell'espressione del gene g nella prima condizione e $\mu_y(g)$ è il valore atteso dell'espressione del gene g

nella seconda condizione. Per ogni gene consideriamo quindi la differenza nell'espressione nelle due condizioni, utilizzando la seguente statistica test:

$$d_g = \frac{\bar{x}_g - \bar{y}_g}{s_g - s_0} \tag{3.39}$$

dove $\bar{x}_g$ e $\bar{y}_g$ sono rispettivamente il livello medio di espressione nella prima e nella seconda condizione. La deviazione standard s_g è definita nel seguente modo per il gene g:

$$s_g = \sqrt{\alpha \left\{ \sum_{j=1}^{n} (x_{g,j} - \bar{x}_g)^2 + \sum_{j=1}^{n} (y_{g,j} - \bar{y}_g)^2 \right\}} \tag{3.40}$$

dove n è il numero di replicazioni, mentre α è dato da:

$$\alpha = \frac{1}{n(n-1)} \tag{3.41}$$

Per quanto riguarda la scelta del parametro s_0, essa è condizionata al fatto che a bassi livelli di espressione la varianza di d_g può essere alta a causa del basso valore di s_g. Per poter confrontare i valori di d_g per tutti i geni è necessario che la distribuzione di d_g sia indipendente dal livello di espressione dei geni. Per questo motivo il parametro s_0 è scelto in modo che il coefficiente di variazione di d_g sia costante indipendentemente dal valore della deviazione standard s_g. Si è dimostrato che la costante s_0 può essere calcolata come il novantesimo percentile della distribuzione degli s_g.

Il valore della statistica test e la sua significatività possono essere calcolati tramite un approccio permutale attraverso i seguenti punti:

1. Si calcolano i valori osservati:

$$d_g = \frac{\bar{x}_g - \bar{y}_g}{s_g + s_0} \tag{3.42}$$

   per ogni $g = 1, \ldots, p$.

2. Si ordinano i valori del test in modo che

$$d_{(1)} \geq d_{(2)} \geq \ldots \geq d_{(p)} \tag{3.43}$$

3. Per ciascuna permutazione k (con $k = 1, \ldots, K$) dei valori di espressione si calcola:
$$d_g(k) = \frac{\bar{x}_g - \bar{y}_g}{s_g + s_0} \tag{3.44}$$

4. si ordinano in modo crescente i valori $d_g(k)$:
$$d_{(1)}(k) \geq d_{(2)}(k) \geq \ldots \geq d_{(p)}(k) \tag{3.45}$$

5. infine si definisce la quantità media:
$$d_{(g)}(E) = \frac{1}{K} \sum_{k=1}^{K} d_{(g)}(k) \tag{3.46}$$

Per poter identificare i geni differenzialmente espressi è necessario confrontare le statistiche $d_{(g)}$ originali con i valori $d_{(g)}(E)$ appena calcolati. Si definisce poi una soglia oltre la quale si rifiuta l'ipotesi di uguaglianza di espressione. Le soglie t_1 e t_2 sono definite rispettivamente come il più piccolo valore positivo e il più grande valore negativo tali che $|d_g - d_g(E)| > \Delta$. Il gene g è considerato differenzialmente sovra-espresso se $d_g \geq t_1$ oppure sotto-espresso se $d_g \leq t_2$. Il valore di Δ è scelto in modo da controllare il FDR: mantenere basso tale errore controlla la presenza di falsi positivi, come descritto nel Paragrafo 1.5.

Capitolo 4

Risultati e discussione

4.1 Introduzione

Scopo della Tesi è di analizzare l'impatto che le normalizzazioni usate per l'analisi dei dati di *microarray* hanno sia sulla sensibilità e specificità del test SAM che sull'ordinamento di tale statistica. In particolare, si vuole valutare *(i)* se alcuni modelli sono più efficaci di altri nel ridurre possibili errori sitematici, *(ii)* se l'uso di alcune normalizzazioni quando non necessarie possa alterare i valori di espressione riducendo poi la sensibilità e specificità del test SAM e *(iii)* se le liste di geni differenzialmente espressi identificati dal test SAM su matrici normalizzate con modelli diversi presentano differenze trascurabili.

L'utilizzo delle liste di geni sregolati permette di confrontare i risultati ottenuti da dati simulati con quelli ottenuti da dati reali, per i quali non è possibile utilizzare sensibilità e specificità, non conoscendo a priori i geni differenzialmente espressi. Nel Paragrafo 4.5 si utilizzerà tale strumento per analizzare dati di espressione nel Sarcoma di Ewing.

4.2 Modelli bayesiani gerarchici

Nella prima parte del lavoro si sono confrontate le diverse normalizzazioni utilizzando le matrici generate con i modelli GG e LNN. I dati di *microarray* sono spesso caratterizzati da un elevato numero di valori mancanti causati spesso da probemi di ibridazione aspecifica, questi valori mancati sono poi imputati attraverso il metodo KNN che si è dimostrato essere il più efficace (Troyanskaya *et al.* (2001)). Per valutare l'infleunza che i valori mancanti imputati possono avere sui modelli e sulle loro prestazioni, una volta generate le matrici, si è eliminata in modo casuale una percentuale crescente (5%, 10%, 15% e 25%) di valori di espressione reimputati poi attraverso il metodo KNN. Osservando le curve ROC per il confronto delle normalizzazioni al variare della percentuale di dati mancanti (Figure 4.1 e 4.2), si osserva una diminuzione delle prestazioni del test, diminuzione che però è indipendentemente dal modello che si sia usa per normalizzare i dati. Per brevità riportiamo solo le curve relative al 5% e al 25% di dati mancanti. Da questo momento in avanti quindi riporteremo esclusivamente i risultati relativi alle normalizzazioni effettuate sulle matrici con il 5% di dati mancanti.

Nelle Figure 4.1 e 4.2 sono riportate per chiarezza solo le ROC per la normalizzazione globale, per la *lowess* e per la GLOG, tutte e otto le normalizzazioni infatti hanno un andamento indistinguibile e includerle tutte in uno stesso grafico darebbe un'idea molto confusa dell'andamento generale. I modelli bayesiani gerarchici simulano dati privi di distorsioni sistematiche, quindi già i dati grezzi (i.e. non normalizzati) portano ad una buona prestazione in termini di sensibilità e specificità del test.

Per quanto riguarda invece il confronto tra i due modelli di simulazione, si può osservare come il test sui dati simulati da LNN classifica in generale più correttamente i geni. Questo è sottolineato dal fatto che la curva ROC relativa ai dati grezzi nel modello GG è uniformemente più bassa di quella relativa ai dati grezzi nel modello LNN. Ad esempio, ad una percentuale di falsi positivi pari a 12% si ha per LNN un livello di sensibilità pari al 74%, mentre per GG la sensibilità è 65%. Il confronto tra i due modelli si può

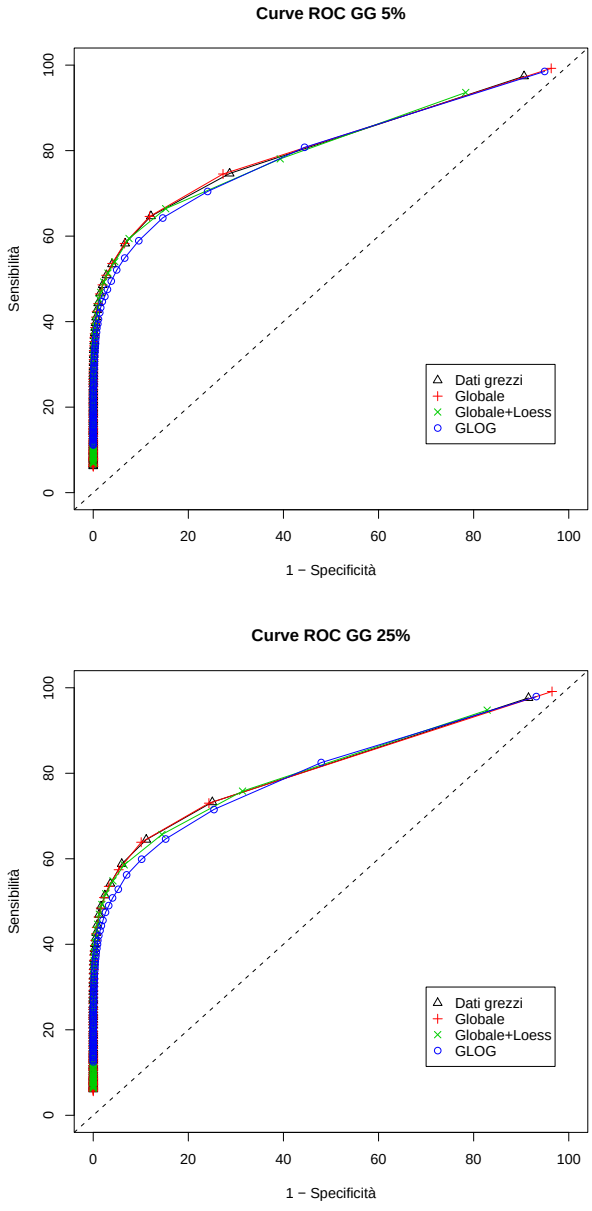

Figura 4.1: Confronto tra le normalizzazioni per il modello GG al variare della percentuale di dati mancanti.

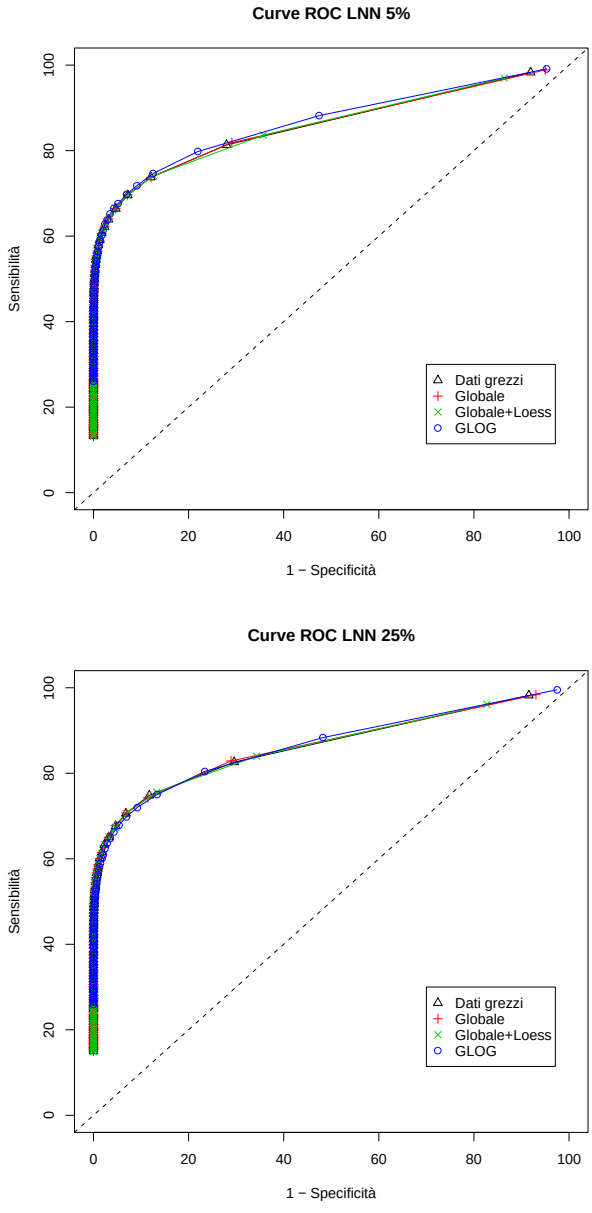

Figura 4.2: Confronto tra le normalizzazioni per il modello LNN al variare della percentuale di dati mancanti.

osservare in Figura 4.3.

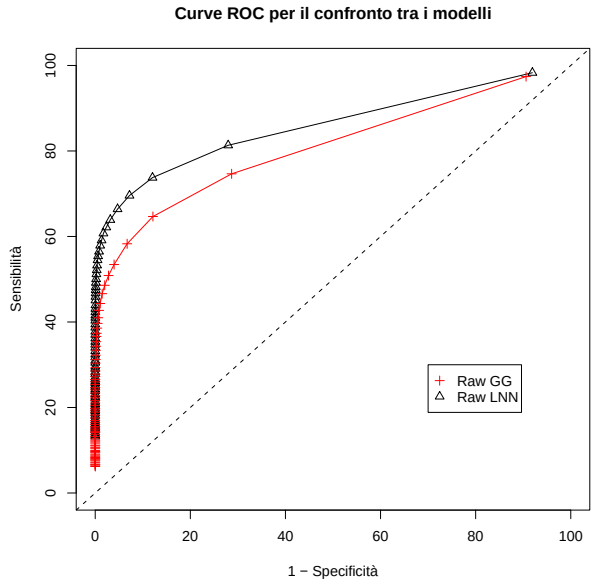

Figura 4.3: Confronto tra le curve ROC relative ai dati grezzi simulati con il modello GG e con il modello LNN.

Parallelamente alle curve ROC presentiamo anche le liste dei geni riconosciuti come differenzialmente espressi dal test dopo ogni normalizzazione. Per brevità riportiamo solo le liste relative al modello GG, dal momento che i risultati sono del tutto analoghi per il modello LNN. Ad ogni gene si è assegnato un rango basato sul valore assoluto del risultato della statistica test e si sono presi in considerazione i primi 20, 50, 100 e 500 geni ordinati in base a tale rango. Indicando con d_g il valore della statistica test associato al gene g e con $|d_g|$ il suo valore assoluto, il rango r_g è definito come la posizione di $|d_g|$ nella lista ordinata in modo decrescente:

$$\left|d_{(1)}\right| \geq \left|d_{(2)}\right| \geq \ldots \geq \left|d_{(p)}\right| \tag{4.1}$$

A valori assoluti alti di SAM corrisponderanno quindi valori bassi nel rango. Quando in seguito si parlerà di geni più significativi si intenderà quelli con rango più basso. Nel grafico in Figura 4.4, i valori in ordinata rappresentano i

geni in comune tra la lista di geni ottenuta dai dati normalizzati attraverso il modello preso in considerazione e quella ottenuta dai dati grezzi. Ad esempio il punto di coordinate $(20, 0.78)$ significa che il 78% dei dati nella lista dei 20 più significativi partendo dai dati grezzi è presente anche nella lista dei 20 più significativi partendo dai dati normalizzati con la trasformazione GLOG.

Dalla Figura 4.4 si osserva come i dati grezzi, la normalizzazione globale e la *lowess* individuano gli stessi geni, mentre esiste circa un 20% di geni che sono individuati dai dati grezzi, ma non dalla trasformazione GLOG. Questo è sottolineato anche dal grafico in Figura 4.5 che confronta i primi 20, 50, 100, 500 geni individuati da GLOG con la totalità dei geni individuati come differenzialmente espressi dagli altri modelli. In questo grafico si sottolinea come i geni individuati solo da un modello abbiano tutti rango alto, questo sta a sottolineare che i geni classificati in modo diverso dai due modelli sono quelli "più in dubbio" (i.e. con valore della statistica più vicino al valore critico di accettazione dell'ipotesi nulla). I primi 100 geni per ordine della statistica sono individuati come differenzialmente espressi da tutti i modelli di normalizzazione considerati.

Questo risultato è in accordo con le curve ROC di Figura 4.1, in quanto se le prestazioni del test a seguito delle diverse normalizzazioni sono quasi identiche, ci si aspetta che la maggior parte dei geni siano riconosciuti allo stesso modo dai diversi modelli. I geni che non sono in comune saranno in ogni caso classificati correttamente o meno in modo simmetrico per i modelli. Essendo tali geni al limite della zona di accettazione, è ragionevole pensare che siano classificati DE o meno a causa di fluttuazioni casuali. Il fatto invece che la normalizzazione globale e la *lowess* riconoscano come DE esattamente gli stessi geni riconosciuti a partire dai dati grezzi è dovuto al fatto che tali normalizzazioni alterano i dati in modo trascurabile. Già i dati grezzi infatti hanno la distribuzione dei log-rapporti centrata in zero senza andamenti sistematici dipendenti dall'intensità.

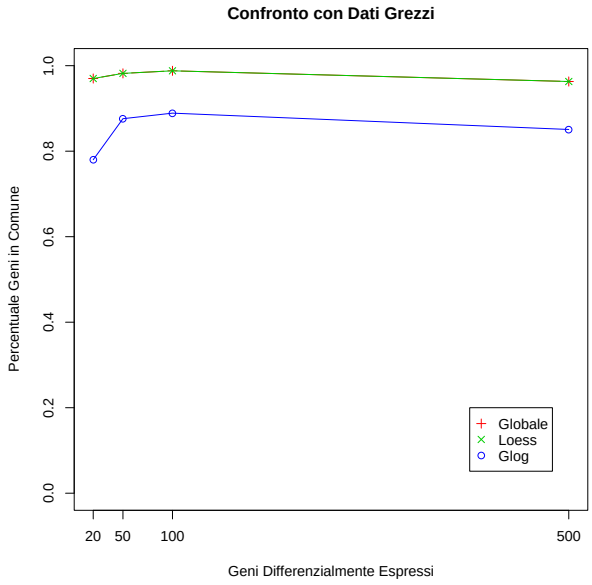

Figura 4.4: Confronto tra i dati grezzi e le altre normalizzazioni relativamente ai primi geni per rango della statistica SAM.

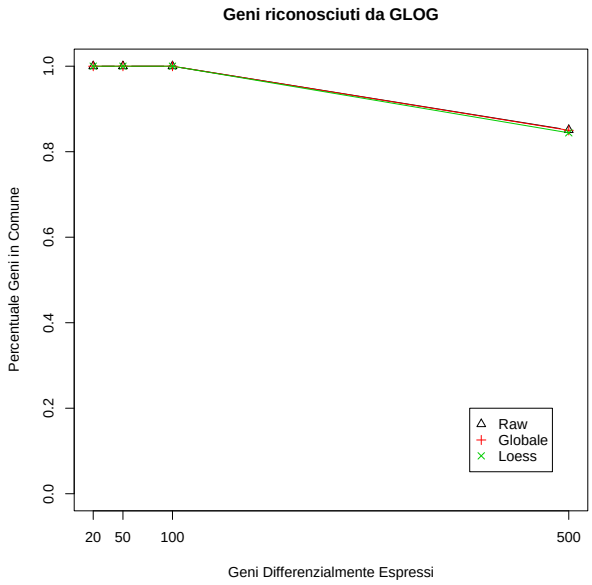

Figura 4.5: Confronto tra i primi geni per rango della statistica SAM dopo la GLOG e i 500 geni classificati come DE dal test dopo le altre normalizzazioni.

4.2.1 Distorsione

Introduciamo ora nei dati generati dai modelli GG e LNN la distorsione presentata nel Paragrafo 2.2.4. Dopo l'introduzione di una distorsione sistematica nei dati, ci aspettiamo che il test SAM calcolato a partire dai dati grezzi non sia più in grado di classificare correttamente i geni, e che quindi normalizzare i dati apporti un notevole miglioramento in termini di sensibilità e specificità del test.

In Figura 4.6 si possono osservare le curve ROC relative al modello GG. Si osservi come il test effettuato sui dati grezzi non sia più in grado di discriminare correttamente tra geni egualmente e differenzialmente espressi. Il fatto che la curva ROC relativa ai dati non normalizzati sia molto vicina alla diagonale indica che la percentuale di geni correttamente classificati è equivalente a quella che si avrebbe classificando i geni in modo casuale. Dopo la normalizzazione globale le cose migliorano ma la curva rimane molto schiacciata sulla diagonale, ad indicare che la percentuale di falsi positivi e negativi è ancora molto alta. Anche la normalizzazione data dalla trasformazione GLOG, sebbene meglio della globale, non sembra soddisfacente. Le altre normalizzazioni permettono al test di discriminare decisamente meglio i geni. Soprattutto i modelli OLIN e OSLIN , così come la normalizzazione *lowess* sembrano avere prestazioni migliori rispetto agli altri. In una situazione intermedia si collocano i due modelli che utilizzano le *spline* e la rete neurale che non sembra rispondere quanto le regressioni locali.

Per quanto riguarda invece il modello LNN il confronto tra le normalizzazioni è riportato in Figura 4.7. In questo caso il test sui dati grezzi fornisce risultati ancora peggiori, a significare che la distorsione applicata ai dati generati dal modello LNN è più marcata di quella applicata al modello GG. Come si può osservare anche dall'MA-Plot di Figura 2.4, la non linearità per valori bassi di A è maggiore per il modello LNN. Come nel modello GG, anche la normalizzazione globale non è sufficiente; la trasformazione GLOG normalizza meglio i dati, portando però ad una prestazione del test in termini di specificità e sensibilità ancora inferiore rispetto alle normalizzazioni

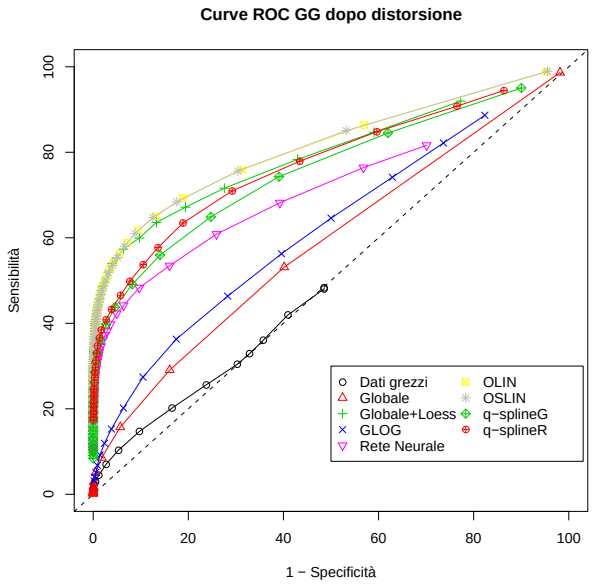

Figura 4.6: Confronto tra le normalizzazioni per il modello GG dopo la distorsione.

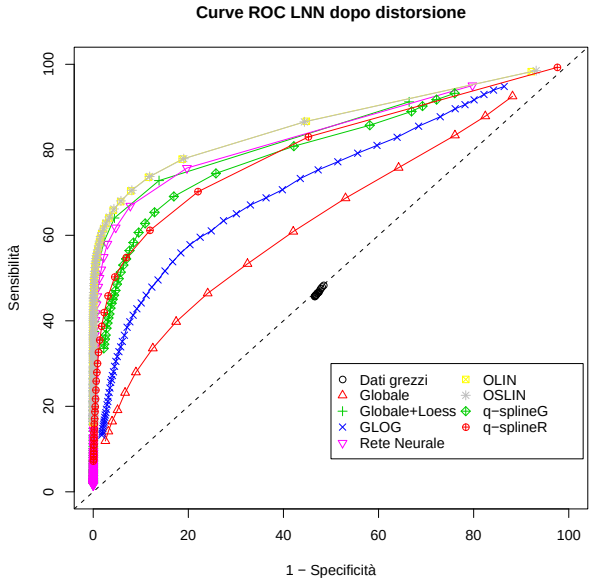

Figura 4.7: Confronto tra le normalizzazioni per il modello LNN dopo la distorsione.

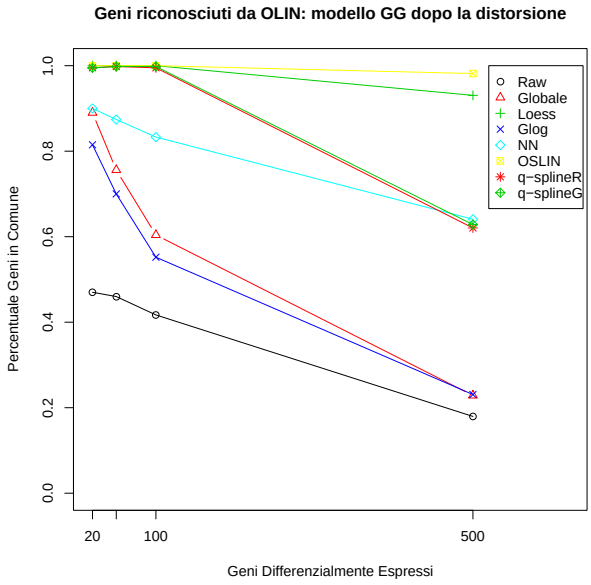

Figura 4.8: Confronto tra i primi geni per rango di SAM dopo OLIN e i 500 geni classificati come DE dal test dopo le altre normalizzazioni per GG.

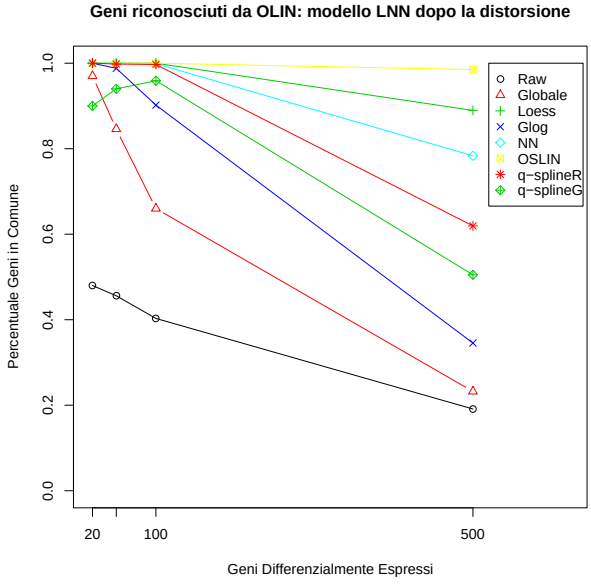

Figura 4.9: Confronto tra i primi geni per rango di SAM dopo OLIN e i 500 geni classificati come DE dal test dopo le altre normalizzazioni per LNN.

basate sulle regressioni locali (*lowess*, OLIN, OSLIN). In questo caso la rete neurale si comporta decisamente meglio e porta a risultati paragonabili a quelli dati dalle regressioni locali. Le *spline* sono anche in questo caso in una situazione intermedia. Quello che si osserva è che per i dati simulati con LNN i due modelli basati sulle *spline* non si comportano in modo analogo: *qsplineG* sembra avere un andamento più irregolare e anomalo rispetto alle altre normalizzazioni, anche se come valori di sensibilità e specificità è paragonabile a *qsplineR*.

Per quanto riguarda invece le liste di geni, a partire da questo Paragrafo si è deciso per brevità di riportare solamente i grafici relativi al confronto dei primi geni di una normalizzazione di riferimento con la totalità dei geni classificati DE da SAM dopo le altre normalizzazioni. La normalizzazione di riferimento è scelta sempre come quella che sembra portare a risultati migliori in termini di sensibilità e specificità. In appendice A sono riportati tutti i grafici relativi alle liste di geni delle normalizzazioni di riferimento confrontate con la percentuale crescente di geni con rango più basso per le altre normalizzazioni, che sono stati omessi in questo Capitolo.

Osserviamo quali geni individuati nei primi posti dal test SAM a seguito della normalizzazione OLIN sono presenti nei 500 geni classificati come differenzialmente espressi dagli altri modelli. Si è scelto come modello di riferimento OLIN in quanto sembra essere quello che normalizza meglio i dati in entrambi i modelli GG e LNN in termini di sensibilità e specificità. In Figura 4.8 è riportato il grafico relativo al modello GG, mentre in Figura 4.9 il grafico relativo al modello LNN.

Nel modello GG si può osservare come i primi 100 geni ordinati da SAM a seguito della normalizzazione OLIN siano tutti presenti nei geni classificati come DE dai modelli OSLIN, *lowess* e *q-spline*, che sono anche i modelli che meglio normalizzano i dati. Vi è circa un 30% di geni considerato DE se si normalizza con OLIN, ma non se si normalizza per mezzo delle *spline*. Tuttavia tali geni hanno tutti rango superiore a 100, nell'ordinamento decrescente della statistica test. Le scarse prestazioni della rete neurale e ancora

di più della GLOG sono evidenziate anche dal fatto che già considerando i 100 geni con *score* più alto per OLIN, rispettivamente il 17% e il 55% di essi sono classificati come EE dalla rete neurale e da GLOG. Verosimilmente tali geni sono falsi negativi per questi modelli, in quanto la rete neurale e la trasformazione GLOG non sembrano normalizzare bene i dati, come si vede in Figura 4.6.

Anche nel modello LNN si può notare che i primi 100 geni ordinati da SAM a seguito della normalizzazione OLIN sono tutti presenti nei geni classificati come DE dai modelli OSLIN, *lowess* e *q-splineR*, mentre come già osservato in precedenza *qsplineG* presenta un andamento anomalo, anche se comunque il 96% di tali geni è classificato DE da tale modello. Come detto in precedenza la rete neurale sembra funzionare meglio con i dati simulati dal modello LNN: anch'essa infatti classifica come DE tutti i geni ordinati nei primi 100 posti da SAM dopo OLIN. Consideriamo infine il confronto con GLOG: al contrario di ciò che avviene nel modello GG, la normalizzazione GLOG fornisce risultati più simili a quelli ottenuti dopo le altre normalizzazioni: nonostante solo il 35% dei geni considerati DE dopo OLIN siano considerati DE anche dopo GLOG, se consideriamo solo i primi 100 geni di OLIN, il 90% di essi è considerato DE anche normalizzando con GLOG. Questo sottolinea ancora una volta il fatto che la differenza nelle prestazioni del test dopo le diverse normalizzazioni è influenzata dai geni con rango alto, e quindi con valore della statistica vicino alla soglia di accettazione.

In entrambi i modelli la normalizzazione più accurata sembra essere il modello OLIN. Prendendo in considerazione tutti i 500 geni classificati come DE da OLIN, si osserva che:

- OSLIN individua esattamente gli stessi geni come DE rispetto a OLIN (98% sia in GG sia in LNN): questo potrebbe essere indice di come l'ulteriore trasformazione dei dati presente in OSLIN sia superflua;
- *lowess* ha una percentuale di geni in comune con OLIN molto alta: 93% in GG e 89% in LNN. Questo indica che il valore scelto per il parametro di lisciamento della *lowess* ($\alpha = 0.4$) non si allontana dal valore ottimo;

- la rete neurale e la GLOG sembrano fare maggiore difficoltà a normalizzare i dati simulati con GG piuttosto che quelli simulati da LNN (Figure 4.6 e 4.7).

Ultima considerazione si può esprimere in relazione alle scarse prestazioni di GLOG rispetto alle altre normalizzazioni: la natura della distorsione sistematica dei dati, potrebbe favorire le normalizzazioni che vanno a stimare l'andamento di una funzione di distorsione presente nell'MA-Plot e la vanno a sottrarre ai dati. La distorsione qui introdotta è infatti proprio una funzione dell'intensità A sommata al log-rapporto M: è facile quindi ad esempio per *lowess* stimare l'andamento di tale funzione, sottrarla alle ordinate dei punti dell'MA-Plot e riportare i dati ad una situazione priva di errori sistematici. Inoltre il fatto che la GLOG faccia più fatica a normalizzare i dati provenienti dal modello GG può essere dovuto all'assunzione sulla distribuzione dei dati sottostante alla scelta della trasformazione: Durbin *et al.* (2002) infatti assumono che i dati di espressione provengano da una variabile casuale combinazione di una Normale e di una LogNormale (Paragrafo 3.6).

4.3 SIMAGE

In questo Paragrafo andremo a considerare 10 matrici simulate con il *software* SIMAGE. Per quanto riguarda la scelta dei parametri si è deciso di cambiare il meno possibile quelli presenti di *default*. I parametri di *default* infatti sono stati stimati utilizzando un esperimento reale, realizzato al *Groningen Biomolecular Sciences and Biotechnology Institute* (Albers *et al.* (2006)). garantendo quindi che i dati simulati siano molto più realistici. Premesso questo, si è comunque deciso di cambiare alcuni parametri per rendere i dati simulati con SIMAGE in qualche modo confrontabili con quelli simulati dai modelli bayesiani. Per questo si è deciso di avere per ogni simulazione 15 *array* da 10000 geni. Inoltre si sono impostate al 3% la percentuale di geni sotto-espressi e quella di geni sovra-espressi, per avere una percentuale di differenzialmente espressi simile a quella impostata nei modelli bayesiani,

che era al 5%. Infine si è posta al 50% la percentuale massima di intensità di *background* rispetto all'intensità dello *spot* (si veda il Paragrafo 2.3.3). In Appendice B si trova la lista completa dei parametri usati per la simulazione.

In Figura 4.10 si trovano le curve ROC relative al test SAM effettuato sui dati così simulati. Per prima cosa si può notare come i dati grezzi e la normalizzazione globale portino a valori di sensibilità e specificità già accettabili. In secondo luogo si può osservare come i modelli *q-spline* e il modello GLOG sembrano normalizzare meglio i dati almeno da un certo punto in poi: per valori di specificità minori di 90 le *q-spline* portano a risultati migliori rispetto agli altri modelli; per valori di specificità minori di 80 anche la GLOG sembra normalizzare meglio degli altri, pur rimanendo ad un livello di sensibilità inferiore a quello delle *q-spline*.

Questo risultato potrebbe dipendere dalla presenza di molti dati negativi, che sono considerati come mancanti dai modelli basati sulla normalizzazione globale così come per le normalizzazioni ottenute attraverso regressione locale: tali modelli infatti ricorrono alla trasformazione logaritmica che utilizza solo i dati di espressione positivi, ignorando (e trattando come fossero mancanti) i dati negativi. La trasformazione GLOG invece, proprio per la sua natura, è una trasformazione che può prendere in argomento anche valori di espressione negativi, come si evince dalla (3.38). Anche i modelli *q-spline* ammettono la presenza di valori negativi, in quanto essi normalizzano i canali separatamente (senza calcolare i log-rapporti), utilizzando i quantili della distribuzione.

La presenza di valori di espressione negativi si ottiene quando il valore di *background* di uno *spot* è più grande del valore di intensità registrato per quello stesso *spot*. Nell'ipotesi che la presenza di valori negativi sia responsabile della maggiore bontà dei modelli che li ammettono, abbiamo provato a ripetere i test sostituendo, preliminarmente alla normalizzazione, i dati negativi con un valore positivo arbitrario, scelto in modo tale da non dare problemi alla trasformazione logaritmica e da risultare vicino a zero, per modellare il fatto che il valore di espressione in quello *spot* sia basso. Tale

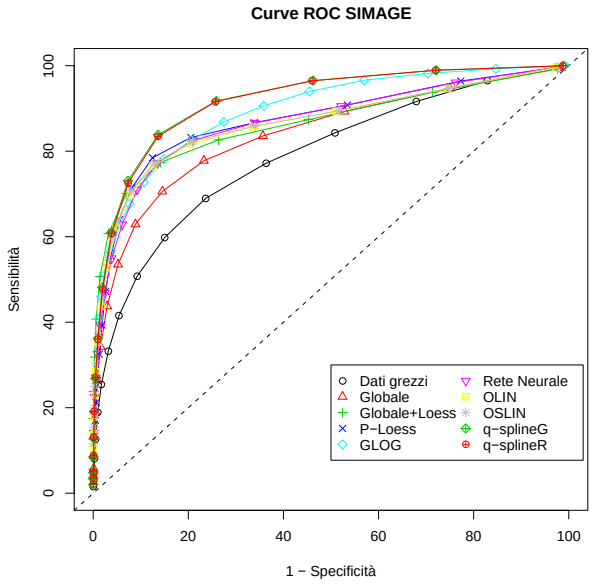

Figura 4.10: Curve ROC relative al modello SIMAGE con livello di *background* al 50% rispetto all'intensità dello *spot*.

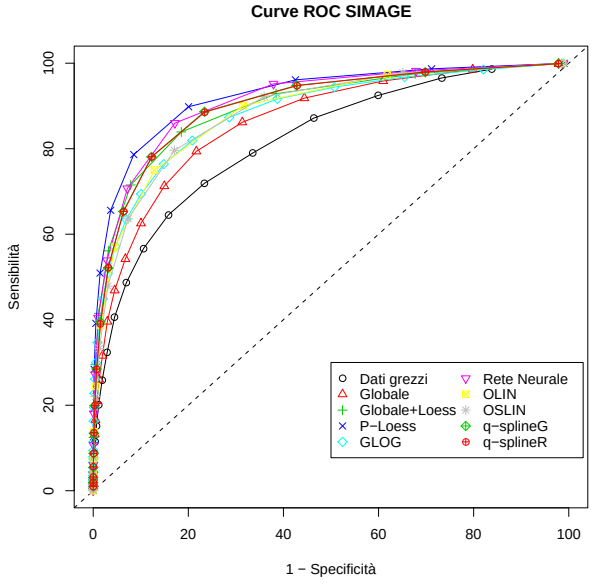

Figura 4.11: Curve ROC relative al modello SIMAGE con livello di *background* al 50%, dopo aver sostituito i valori di intensità negativi.

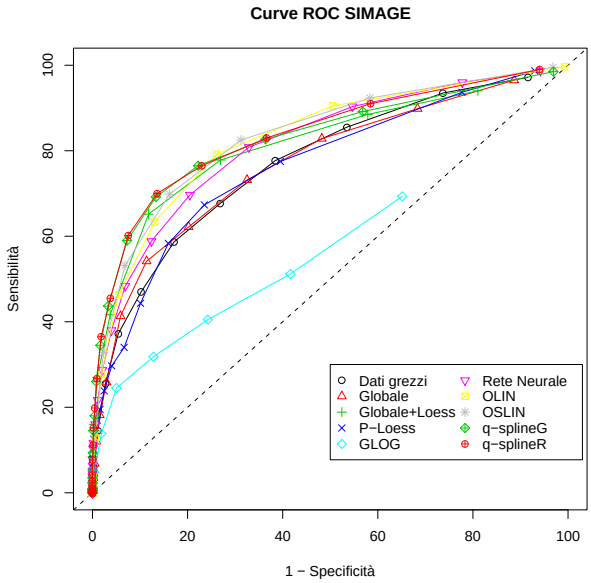

Figura 4.12: Curve ROC relative al modello SIMAGE con livello di *background* al 150%, dopo aver sostituito i valori di intensità negativi.

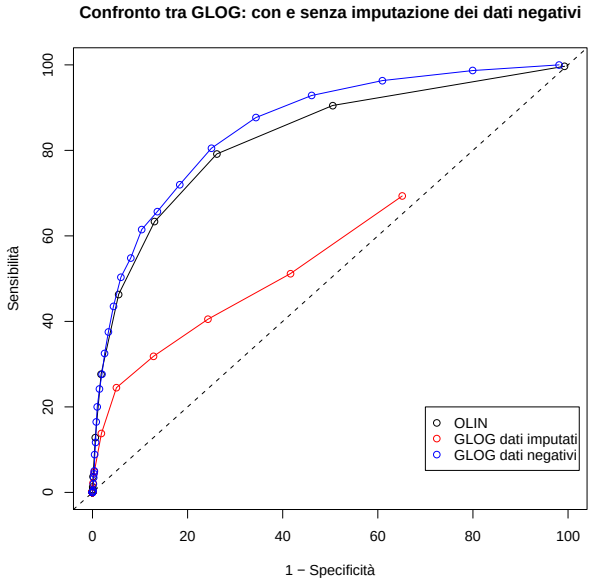

Figura 4.13: Confronto tra la GLOG applicata ai dati originali (in blu) e quella applicata dopo aver sostituito i valori negativi (in rosso).

operazione è peraltro frequente nell'analisi di dati provenienti da esperimenti reali; nel presente elaborato si sono sostituiti i dati negativi con il valore 4.

Le curve ROC relative al test ottenuto dai dati così modificati sono presentate in Figura 4.11. Si può osservare come l'impatto su sensibilità e specificità dei modelli basati su regressione locale cambi notevolmente: il modello che sembra normalizzare al meglio i dati è ora la *P-lowess*, ma anche gli altri modelli hanno adesso prestazioni paragonabili a quelle delle *q-spline* e della GLOG. Questi due modelli presentano un andamento leggermente peggiore, anche se in modo non significativo, rispetto alla situazione in cui i dati negativi erano mantenuti nell'analisi.

Quello che si può ipotizzare e che le prestazioni di tali modelli peggiorino all'aumentare dei dati sostituiti con un valore positivo arbitrario. Per verificare tale ipotesi si può considerare la situazione in cui la matrice dei dati presenta molti più dati negativi, che si andranno poi a sostituire: per ottenere tali dati simuliamo con SIMAGE un esperimento in cui il livello di *background* è stato posto al 150% del livello di espressione e gli altri parametri sono mantenuti inalterati.In Figura 4.12 si possono osservare le curve ROC relative a tale simulazione. Dal confronto tra le diverse normalizzazioni si osserva come i modelli basati su regressione locale non risentano della sostituzione, se si esclude un peggioramento delle prestazioni dovuto all'aumento di dati non informativi. Anche le *q-spline* sembrano mantenere buoni livelli di specificità e sensibilità, mentre la GLOG, con i dati così simulati, porta a risultati insoddisfacenti.

In Figura 4.13 sono riportate la curva ROC del modello OLIN che sembra normalizzare bene i dati così simulati, la curva relativa ai dati trasformati con GLOG dopo la sostituzione, e quella relativa ai dati trasformati con GLOG mantenendo nell'analisi i dati negativi: si osserva che in questo caso le prestazioni della normalizzazione GLOG sono addirittura superiori a quelle di OLIN, ad indicare come in presenza di valori alti di *background* essa normalizzi al meglio i dati; nel caso però che non si disponga dei dati negativi, perchè già eliminati in una fase precedente all'analisi, la GLOG non permette

più di normalizzare correttamente i dati.

Consideriamo ora un ulteriore modello in cui il livello di *background* è stato posto al 10% relativamente al segnale di *non-background*. Questo indica un esperimento in cui l'ibridazione è riuscita meglio e l'intensità registrata negli *spot* è più "pulita". Gli altri parametri di SIMAGE sono stati lasciati inalterati, e si sono simulate ancora 10 matrici, ognuna con 15 *array* da 10000 geni. Le curve ROC relative al test SAM effettuato sui dati così simulati sono riportate in Figura 4.14.

Appare evidente come in questo caso le normalizzazioni riescano a trasformare i dati in modo che quasi la totalità dei geni sia classificata correttamente. Infatti ad una percentuale di falsi positivi del 20% si ha una sensibilità pari quasi al 100%. I modelli si comportano tutti in modo molto simile se si eccettua la normalizzazione globale e la GLOG che come sempre assume un andamento diverso rispetto agli altri modelli, essendo meno performante, anche se in modo non significativo, nell'intervallo di specificità compreso tra 60 e 90.

Per quanto riguarda le liste di geni, prendiamo come normalizzazione di riferimento per i dati simulati con il 50% di *background* il modello *qsplineR*, in quanto è quello insieme a *qsplineG* che sembra normalizzare al meglio i dati (Figura 4.10).

In Figura 4.15 osserviamo la percentuale di geni con rango più basso dopo la normalizzazione *qsplineR* riconosciuti come DE anche dopo le altre normalizzazioni. In questo caso la totalità dei geni classificati DE è 600 in quanto abbiamo simulato dati con il 3% di geni sotto-espressi e con il 3% di geni sovra-espressi. Dal grafico si osserva come i due modelli basati sulle *spline* individuino esattamente gli stessi geni come DE: questo indica che non è cruciale la scelta del canale da considerare come *target* per la normalizzazione. Se si eccettuano la normalizzazione globale e i dati grezzi, anche le altre normalizzazioni individuano la maggior parte dei geni individuati da *qsplineR*: se si considerano i 100 geni con rango più basso, essi sono presenti in media al 95% nelle liste dei geni DE degli altri modelli, mentre se si considera la

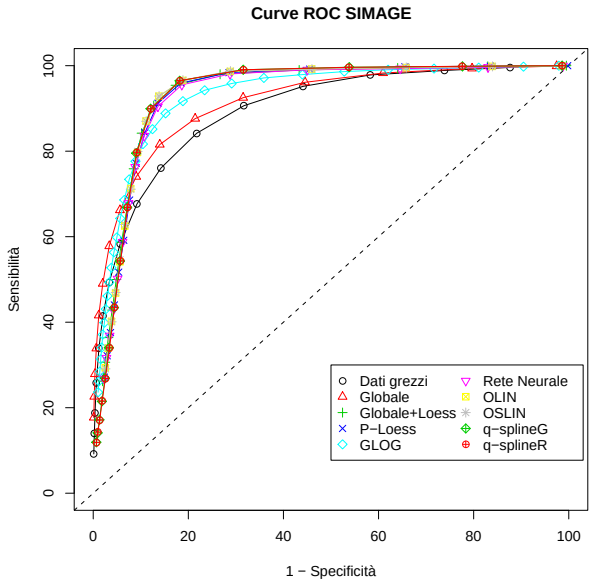

Figura 4.14: Curve ROC relative al modello SIMAGE con livello di *background* al 10% rispetto all'intensità dello *spot*.

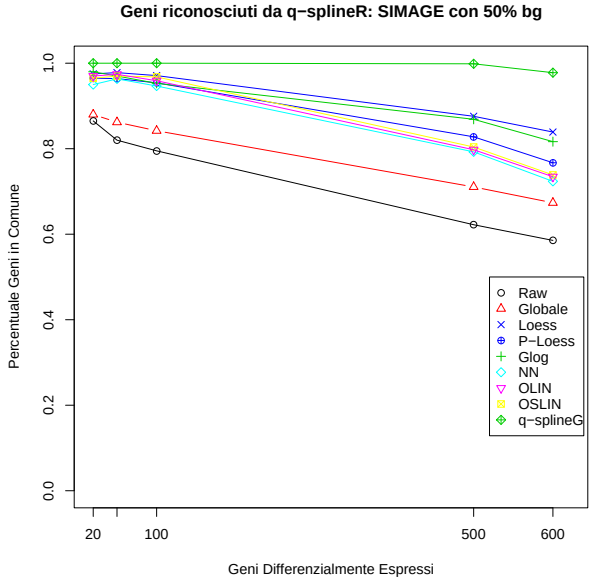

Figura 4.15: Confronto tra i primi geni per rango di SAM dopo *qsplineR* e i 600 geni classificati come DE dal test dopo le altre normalizzazioni.

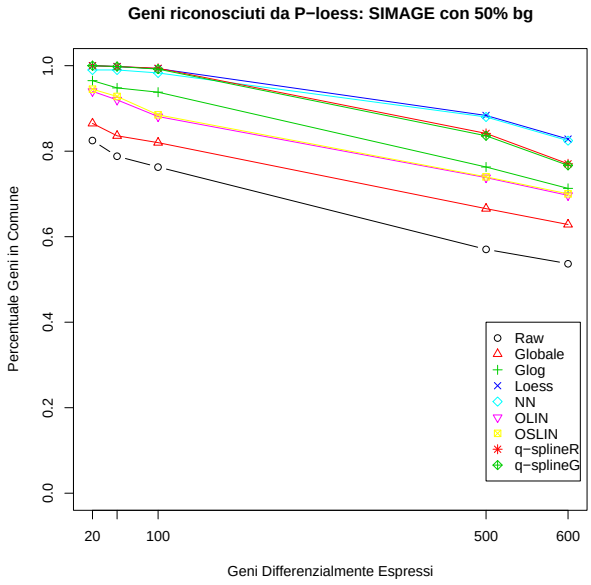

Figura 4.16: Confronto tra i primi geni per rango di SAM dopo *P-lowess* e i 600 geni classificati come DE dal test dopo le altre normalizzazioni.

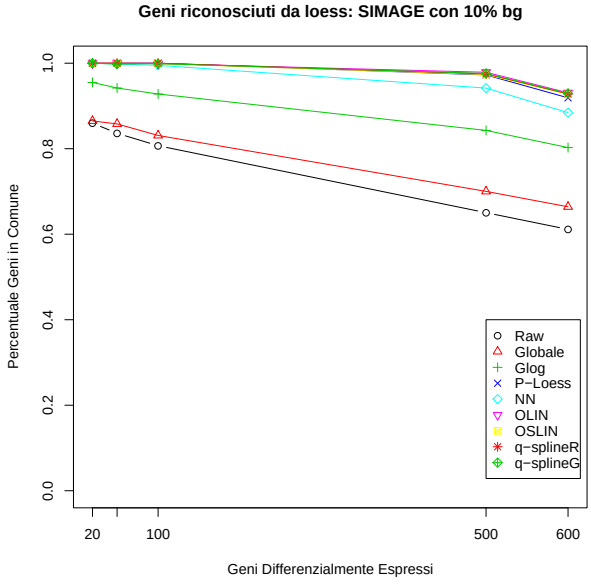

Figura 4.17: Confronto tra i primi geni per rango di SAM dopo *lowess* e i 600 geni classificati come DE dal test dopo le altre normalizzazioni.

totalità dei geni DE vi è una corrispondenza media tra il 72% e l'83%, a seconda del modello considerato.

Come detto, se consideriamo la situazione in cui, prima di normalizzare i dati, si sono sostituiti i valori negativi con zero, i modelli basati sulle regressioni locali hanno prestazioni migliori. Prendiamo quindi come modello di riferimento la *P-lowess*, che sembra comportarsi leggermente meglio delle altre normalizzazioni.

In Figura 4.16 è riportata la percentuale di geni con rango più basso per la statistica SAM dopo la *P-lowess* riconosciuti come DE anche dagli altri modelli. Il fatto che tutte le normalizzazioni portino a risultati soddisfacenti si riscontra anche osservando le liste di geni: considerando i primi 100 geni con rango più basso si osserva come essi siano presenti al 100% nella lista dei geni classificati DE dalla *lowess*, dalla rete neurale e dalle *spline*; al 94% dei geni classificati DE dalla GLOG e all'88% di quelli classificati DE da OLIN e OSLIN. Per quanto riguarda il confronto con la totalità dei geni si va da una corrispondenza del 70% di OLIN e OSLIN ad una dell'83% di *lowess*.

In ultima analisi consideriamo le liste di geni provenienti dai dati simulati con SIMAGE con un livello di *background* del 10%. In Figura 4.17 si può osservare la corrispondenza tra i geni individuati da *lowess* e quelli individuati dagli altri modelli.

Come si può notare anche dalla curva ROC di Figura 4.14 le normalizzazioni sembrano funzionare tutte molto bene per i dati così simulati. In particolare non si osservano differenze nelle normalizzazioni ad eccezione della globale e dei dati grezzi, e in parte della GLOG. Lo stesso risultato si ottiene osservando il grafico in Figura 4.17: vi è una corrispondenza di più del 90% tra i 600 geni individuati come DE da *lowess* e quelli individuati dagli altri modelli, ad eccezione della rete neurale, per cui la percentuale di corrispondenza è leggermente più bassa (88%). Vi è invece un 20% di geni classificati DE da *lowess* che non sono individuati come tali dalla GLOG: è proprio questo 20% che determina l'andamento leggermente peggiore della curva ROC di tale modello.

4.4 Confronto tra i modelli di simulazione

Si è visto nei Paragrafi precedenti come il confronto tra le normalizzazioni dipende, in modo significativo, dal modello utilizzato per la simulazione dei dati. In particolare la trasformazione GLOG risulta meno performante per i dati provenienti dal modello GG. Per i dati così simulati le prestazioni della GLOG sono infatti insoddisfacenti (si veda la curva ROC di Figura 4.6). I risultati migliorano se si utilizzano dati provenienti dal modello LNN, ma anche in questo caso la GLOG non raggiunge livelli di sensibilità e specificità paragonabili alle altre normalizzazioni (Figura 4.7). Le differenze di prestazioni della GLOG tra il modello GG e LNN, potrebbero in parte essere dovute all'assunzione di normalità dei dati da trasformare: come detto nel Paragrafo 3.6, infatti, Durbin *et al.* (2002) assumono che i dati di espressione genica si distribuiscano come una combinazione di una variabile Normale e una LogNormale. La GLOG ha quindi più facilità a trasformare i dati provenienti da LNN, piuttosto che quelli da GG che si distribuiscono come una Gamma.

Se prendiamo in considerazione il modello SIMAGE, invece, le prestazioni della GLOG migliorano notevolmente, risultando addirittura migliori delle normalizzazioni basate su regressione locale nel caso di valori di *background* molto elevati (Figure 4.10 e 4.11). Questo è dovuto alla natura stessa della trasformazione (Paragrafo 3.6), che ammette, al contrario della trasformazione logaritmica utilizzata per le altre normalizzazioni, anche valori di espressione negativi, presenti quando il valore di *background* è più alto dell'intensità dello *spot*.

Le scarse prestazioni della GLOG nei modelli bayesiani, in confronto a quelle osservate sui dati simulati da SIMAGE, potrebbero inoltre dipendere dalla scarsa capacità di tali modelli (soprattutto del modello GG) di generare dati di espressione genica simili a quelli reali, almeno con i parametri utilizzati per le simulazioni: i parametri infatti svolgono un ruolo cruciale nella simulazione, in quanto variando di poco il loro valore si ottengono dati di espressione genica più o meno verosimili (si pensi a valore medio e

variabilità).

Per verificare la correttezza di questa ipotesi, abbiamo generato 10 matrici attraverso il *software* SIMAGE, senza alcun tipo di distorsione, ovvero generando solo il livello di espressione genica. Partendo da tali matrici stimiamo attraverso la funzione *emfit* del pacchetto *EBarrays* di R i parametri dei modelli GG e LNN, e mediamo i risultati rispetto alle 10 stime ottenute per ogni parametro. Mentre le stime dei parametri del modello LNN non sono molto diverse dai valori utilizzati per le simulazioni, le stime dei parametri del modello GG si discostano notevolmente da quelle considerate in precedenza. In Tabella 4.1 sono riportati i nuovi valori dei parametri, stimati dalle matrici generate con SIMAGE, e i valori utilizzati nelle simulazioni precedenti per il modello GG, mentre in Tabella 4.2 sono riportati quelli per il modello LNN.

	Parametri per il modello GG		
	α	α_0	ν
Valori precedenti	1	1.1	45.4
Nuove stime	3.64	2.37	1761.19
Deviazione Standard	0.35	0.18	428.57

Tabella 4.1: Valori utilizzati per la simulazione e nuove stime dei parametri del modello GG.

	Parametri per il modello LNN		
	μ	σ	τ
Valori precedenti	6.58	0.9	1.13
Nuove stime	7.96	0.164	0.895
Deviazione Standard	0.006	0.001	0.006

Tabella 4.2: Valori utilizzati per la simulazione e nuove stime dei parametri del modello LNN.

Abbiamo utlizzato i valori così stimati per generare nuovamente 10 matri-

ci per il modello GG e 10 matrici per il modello LNN sempre considerando 15 esperimenti e 10000 geni e questa volta il 6% di dati DE, alle quali andremo ad introdurre la stessa distorsione sistematica considerata in precedenza. In Figura 4.18 sono riportate le curve ROC relative alla sensibilità e specificità del test SAM dopo le normalizzazioni per il modello GG.

Si osserva come la trasformazione GLOG applicata ai dati generati da GG assume valori di sensibilità e specificità confrontabili con quelli delle altre normalizzazioni. Inoltre, per ogni modello considerato, si ha un aumento nelle prestazioni della normalizzazione rispetto ai valori dei parametri usati in precedenza. Si osservi infatti che nelle curve ROC di Figura 4.6 anche il modello che normalizzava meglio i dati (OLIN) raggiungeva un valore di sensibilità pari al 60% in corrispondenza del 10% di falsi positivi. Con i nuovi valori dei parametri, invece, in corrispondenza del 10% di falsi positivi si raggiunge un valore di sensibilità di oltre il 70%. Per quanto riguarda la GLOG, se ad un valore di falsi positivi del 10% corrispondeva una sensibilità del 25%, con i nuovi parametri tale valore è aumentato fino a raggiungere il 65% circa.

Per quanto riguarda invece il modello LNN, si possono osservare le curve ROC in Figura 4.19: appare evidente come in questo caso le normalizzazioni permettano al test SAM di individuare molti più geni correttamente che con i valori usati precedentemente, raggiungendo il 90% di sensibilità già ad una percentuale di falsi positivi pari a 10%. Anche in questo caso la GLOG risulta leggermente meno performante rispetto alle altre normalizzazioni, anche se permette anch'essa al test ottimi risultati in termini di sensibilità e specificità.

I risultati appena descritti indicano come la scelta dei valori dei parametri sia fondamentale per ottenere valori di espressione realistici, in quanto utilizzando la stessa distorsione applicata ai modelli con i nuovi valori dei parametri si ottengono risultati, in termini di sensibilità e specificità, molto diversi. La differenza residua nelle prestazioni della GLOG rispetto alle altre normalizzazioni, è verosimilmente da ricercare nella natura della distorsione introdotta nei dati (Paragrafo 2.2.4), che proprio perchè funzione di M

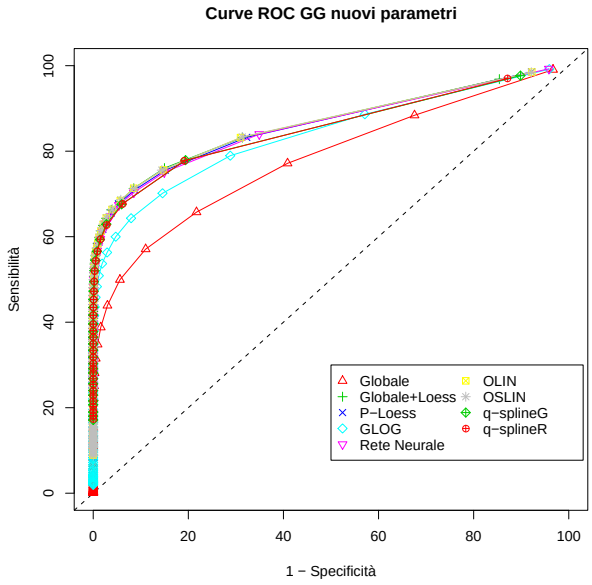

Figura 4.18: Curve ROC relative alle diverse normalizzazioni dei dati provenienti da GG con i nuovi valori dei parametri.

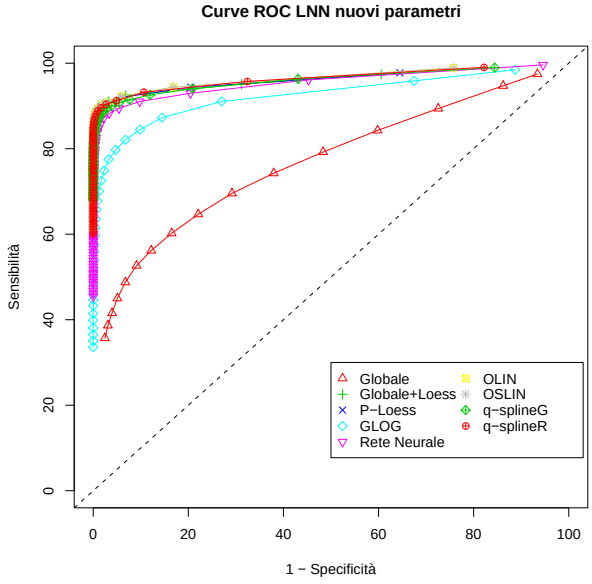

Figura 4.19: Curve ROC relative alle diverse normalizzazioni dei dati provenienti da LNN con i nuovi valori dei parametri.

e di A potrebbe favorire quelle normalizzazioni che agiscono stimando un andamento dei punti sistematico nell'MA-Plot.

Come detto i parametri della GG stimati dai dati simulati con SIMAGE sono molto diversi da quelli utilizzati per le simulazioni iniziali. In particolare il valore molto alto del parametro ν indica che il parametro λ della Gamma, da cui provengono i dati simulati, ha una distribuzione quasi degenere: ricordando che nel modello GG si ha che la distribuzione condizionata dell'espressione genica è $Y|\mu_g \sim Ga(\alpha, \lambda)$ e che $\lambda \sim Ga(\alpha_0, \nu)$, per i valori considerati dei parametri vale:

$$\hat{E}(\lambda) = \frac{\hat{\alpha}_0}{\hat{\nu}} = \frac{2.37}{1761.19} \approx 0.001 \tag{4.2}$$

$$\widehat{Var}(\lambda) = \frac{\hat{\alpha}_0}{\hat{\nu}^2} = \frac{2.37}{1761.19^2} \approx 7 \cdot 10^{-7} \tag{4.3}$$

Il parametro λ assumerà quindi valori molto concentrati intorno al suo valore atteso. Sostituendo tale valore atteso nel calcolo della media della variabile Y si ottiene che i valori di espressione genica con questi valori dei parametri del modello GG saranno intorno ai 1500, vale a dire intorno a 10.5 in scala logaritmica (su base due). Si possono fare due considerazioni su tali risultati:

- il valore così basso per la varianza di λ potrebbe derivare dal fatto che SIMAGE simula dati da un modello normale in un'ottica frequentista: andando a stimare la distribuzione del parametro in un'ottica bayesiana, esso assume una distribuzione quasi degenere proprio per modellare un fattore deterministico nella scelta del parametro, ovvero una forte conoscenza a priori del suo valore;

- con i valori precedenti dei parametri, il modello GG stimava valori di espressione genica con valor medio intorno a 200, che in scala logaritmica corrisponde a circa 7.5. Con i nuovi parametri si hanno invece valori più alti, che più spesso occorrono negli esperimenti di *microarray*, e che sono ora in accordo con i valori generati dal modello LNN.

In ultima analisi osserviamo le liste di geni: per entrambi i modelli abbiamo ora una situazione in cui le normalizzazioni individuano per la gran

parte gli stessi geni come DE, dal momento che anche la GLOG ora porta a risultati soddisfacenti.

In Figura 4.20 si possono osservare le liste dei geni per il modello GG, mentre in Figura 4.21 sono riportate quelle per il modello LNN. In entrambi i casi utilizziamo come normalizzazione di riferimento la *lowess* che sembra normalizzare al meglio i dati, al pari delle altre normalizzazioni basate su regressione locale.

Per quanto riguarda il modello GG si osserva come la totalità dei 100 geni con rango più basso per *lowess* sia classificata DE da tutte le altre normalizzazioni. Considerando invece tutti i 600 geni si ha una corrispondenza del 97% con OLIN e OSLIN, dell'88% con la *P-lowess*, le *spline* e la rete neurale. Vi è ancora un 28% di geni individuati da *lowess* e non da GLOG. Ad ulteriore conferma del fatto che i geni classificati in modo differente tra le diverse normalizzazioni hanno generalmente rango alto, si può osservare come la percentuale di geni in comune cali notevolmente se si considerano i primi 500 piuttosto che tutti i 600 geni individuati da *lowess*: ad esempio la *P-lowess* ha una percentuale di geni in comune pari a 96% se si considerano i primi 500 geni: tale percentuale scende a 88% se si prendono in considerazione gli ulteriori 100 geni. Quindi se considerando i primi 500 geni erano circa 20 i geni classificati in modo differente, considerando tutti i 600 geni i geni classificati in modo diverso dai due modelli sono 72, ad indicare che dei 100 geni con rango più alto di 500, 52 sono classificati in modo diverso dai due modelli.

Per quanto riguarda i dati generati da LNN, tutti i modelli di normalizzazione individuano in gran parte gli stessi geni come DE. Se si esclude la GLOG, la globale e i dati grezzi, infatti, la percentuale di geni in comune con quelli individuati da *lowess* varia tra il 90% della rete neurale al 95% della *P-lowess*. La trasformazione GLOG individua circa un 23% di geni in modo differente rispetto alle altre normalizzazioni, ma in ogni caso i 100 geni con rango più basso dopo *lowess* sono tutti classificati DE anche dopo la GLOG, e se il confronto si ferma ai primi 500 geni si osserva una percentuale di geni

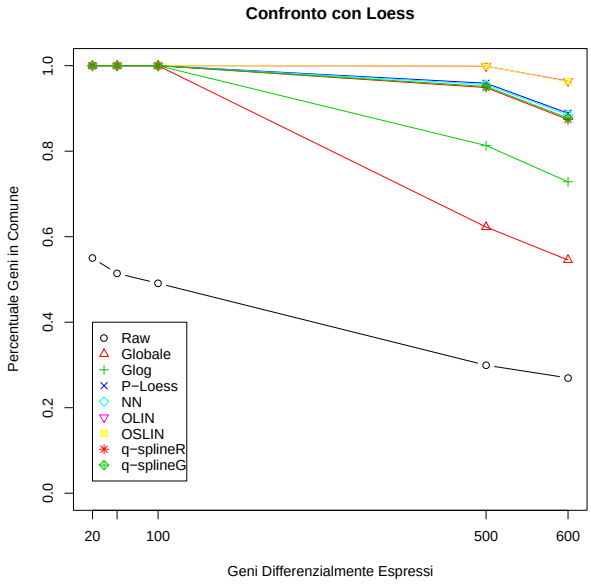

Figura 4.20: Confronto tra i primi geni per rango di SAM dopo *lowess* e i 600 geni DE dopo le altre normalizzazioni per GG con i nuovi parametri.

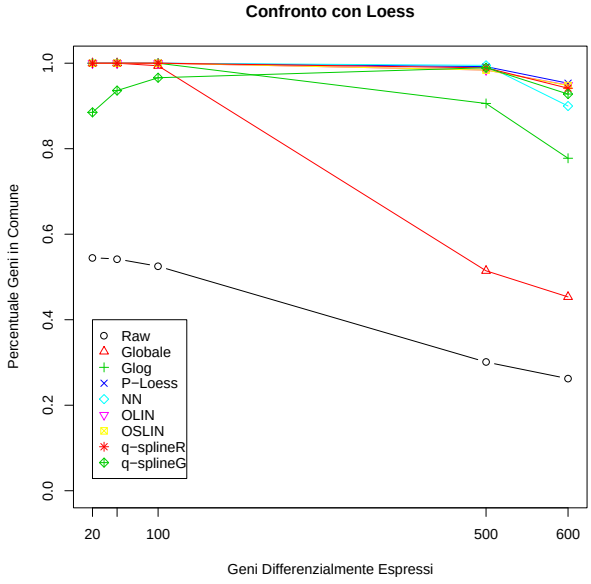

Figura 4.21: Confronto tra i primi geni per rango di SAM dopo *lowess* e i 600 geni DE dopo le altre normalizzazioni per LNN con i nuovi parametri.

in comune pari al 90%.

4.5 Applicazione a dati reali

Come detto nell'introduzione del presente Capitolo si è deciso di utilizzare lo strumento delle liste di geni ordinati per rango della statistica SAM anche per l'analisi di un *dataset* proveniente da un esperimento reale.

I dati presi in considerazione provengono da un esperimento descritto in Baird *et al.* (2005), in cui attraverso l'analisi di dati di espressione si cercano indizi sulla differenziazione cellulare dei sarcomi, una particolare classe di tumori di origine mesenchimale. Lo studio originale era su un campione di 181 tumori, in rappresentanza di 16 classi di sarcomi delle ossa e dei tessuti molli. Le analisi sono state effettuate a partire dai dati di *microarray* a cDNA con 12,601 *probe*. Nel presente elaborato si è utilizzata solo una parte dei dati a disposizione degli autori: quella relativa al sarcoma di Ewing, che corrisponde a 18 esperimenti in cui con il Cy5 è marcato l'RNA del campione in studio (canale rosso) e con il Cy3 è marcato il controllo (canale verde), costituito da un gruppo di cinque linee di cellule tumorali (*pooled group*). La matrice dei dati presenta quindi 12,601 righe e 36 colonne (18 per il canale rosso e 18 per il canale verde). A causa della natura del controllo la matrice dei dati presenta molti zeri, che verosimilmente non sono zeri reali, ma introdotti in fase post-sperimentale per sostituire valori negativi (i.e. con valore di *background* maggiore del valore di espressione) oppure per sostituire i dati mancanti. Prima di procedere alla normalizzazione e alla successiva inferenza, si è quindi deciso di eliminare le righe con più dell'80% di valori nulli sul totale delle replicazioni, in modo indipendente per il canale rosso e per il canale verde. In altre parole si sono eliminate le misure di espressione relative ai geni che presentavano più di 14 zeri sulle 18 replicazioni relative all'intensità di uno dei due canali. Al termine di questa operazione il *dataset* a disposizione è costituito da 6,154 righe corrispondenti ad altrettanti geni. Le misure di espressione nulle ancora presenti sono sostituite da un valore

non nullo basso per evitare problemi nell'utilizzo del log-rapporto.

4.5.1 Il sarcoma di Ewing[1]

Il sarcoma di Ewing deve il suo nome al Dott. James Ewing, che ha descritto questa neoplasia per la prima volta nel 1920. Esso è un tumore che può svilupparsi in qualsiasi distretto del corpo, sebbene origini più frequentemente dalle ossa. Qualsiasi osso può essere sede di malattia, ma la pelvi, il femore e la tibia sono le sedi più comuni. In Italia il numero di nuovi casi/anno con interessamento scheletrico stimato è di circa 60, di cui 2/3 di età inferiore ai 20 anni. Negli Stati Uniti l'incidenza è di 3.4 nuovi casi/anno per milione di abitanti. Colpisce prevalentemente adolescenti e più comunemente i maschi delle femmine.

Sebbene il sarcoma di Ewing sia un tumore prevalentemente osseo, può anche originare dai tessuti molli. In questo caso prende il nome di sarcoma di Ewing extaosseo. Sarcoma di Ewing, *primitive neuroectodermal tumours* (PNET) e tumori di Askin, vengono inquadrati nel gruppo dei "tumori della famiglia di Ewing", un gruppo di neoplasie di origine neuroectodermica con caratteristiche istologiche, immunoistochimiche e citogenetiche comuni. Nel 20-25% dei pazienti sono presenti clinicamente evidenti localizzazioni secondarie di malattia all'esordio, che possono interessare i polmoni o l'osso. I tumori della famiglia degli Ewing non sono comunemente associati ad altre malattie congenite e non vi sono evidenze convicenti che questi tumori siano ereditari.

4.5.2 Risultati

In Figura 4.22 si possono osservare le percentuali di geni riconosciuti DE dai modelli di normalizzazione, tra quelli con più basso rango della statistica SAM dopo la normalizzazione OLIN. Utilizziamo OLIN come normalizzazione di riferimento in quanto si è visto dai risultati sui dati simulati come i

[1]Dott. Stefano Ferrari - Ewing's sarcoma in children (http://www.ior.it/)

modelli basati su regressione locale normalizzino in generale bene i dati. La procedura di ottimizzazione del parametro presente in OLIN, inoltre, permette a tale modello di ottenere nella normalizzazione risultati migliori o almeno uguali a quelli ottenuti a partire da *lowess*.

Se si considerano i 100 geni con rango più basso dato da SAM dopo OLIN, essi sono presenti quasi totalmente nella lista dei geni individuati come DE da SAM dopo le altre normalizzazioni, se si esclude la GLOG. Considerando i primi 200 geni la percentuale di geni in comune resta molto alta, e anche con 300 geni tale percentuale rimane sopra l'80%. Considerando la totalità dei geni, infine, si osserva come OSLIN abbia una percentuale di geni in comune con OLIN pari al 94%, la rete neurale pari all'80%, i due modelli basati sulle *spline* circa al 70% e i modelli *lowess* e *P-lowess* circa al 66%. Il fatto che vi sia più di un 30% di geni che sono classificati come DE da OLIN e non da *lowess* può indicare che il valore del parametro di lisciamento ($\alpha = 0.4$) scelto per la regressione è lontano dal suo valore ottimo, selezionato dal modello OLIN. Considerando i primi 300 geni le percentuali sono tutte molto alte, ad indicare, anche nei dati reali, che i geni classificati in modo diverso dalle diverse normalizzazioni hanno per la maggior parte rango alto, e sono quindi i più vicini alla regione di accettazione dell'ipotesi nulla di espressione equivalente.

Discorso a parte va affrontato per quanto riguarda la GLOG: già considerando i 20 geni con rango più basso per OLIN il 30% di essi (6 geni) è classificato come EE da GLOG. Considerando tutti i 500 geni, la percentuale di geni classificati come DE sia dal modello OLIN che dalla GLOG è pari al 55%.

Osserviamo in ultima analisi la percentuale di geni con rango della statistica SAM più basso dopo la GLOG riconosciuti DE anche dopo le altre normalizzazioni (Figura 4.23): i primi 20 geni sono classificati come DE dagli altri modelli per più del 90%. Considerando i primi 100 geni la percentuale scende tra il 60% e l'80%. Considerando i 500 geni classificati DE da GLOG, essi sono classificati DE dagli altri modelli in una percentuale che varia dal

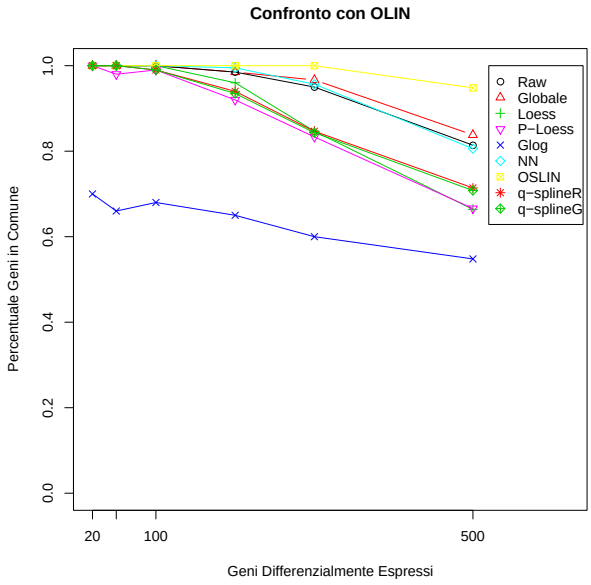

Figura 4.22: Confronto tra i primi geni per rango di SAM dopo OLIN e i 500 geni DE dopo le altre normalizzazioni sui dati del Sarcoma di Ewing.

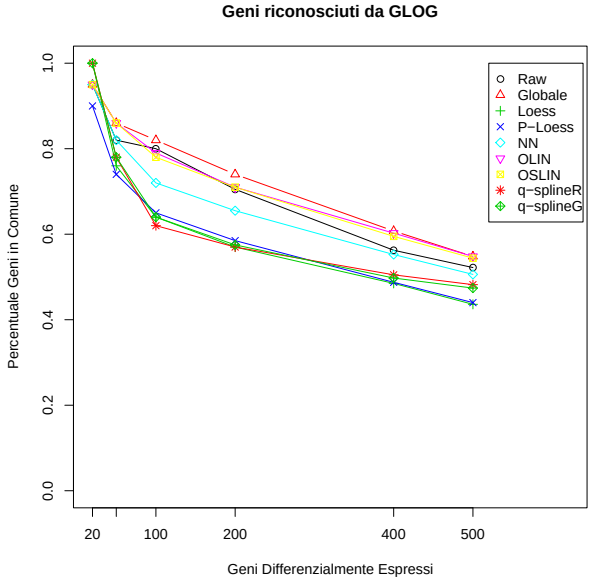

Figura 4.23: Confronto tra i primi geni per rango di SAM dopo GLOG e i 500 geni DE dopo le altre normalizzazioni sui dati del Sarcoma di Ewing.

43% della *lowess* al 55% della OLIN.

Il test SAM dopo la trasformazione GLOG classifica i geni in modo molto diverso dalla classificazione che si ottiene dopo le altre normalizzazioni. Considerando 500 geni, infatti, solo la metà di quelli individuati come DE da GLOG sono individuati come DE anche dagli altri modelli e viceversa. Questo potrebbe dipendere dall'eliminazione dei valori di espressione negativi effettuata in fase post-sperimentale: tale operazione può aver danneggiato la GLOG che lavora meglio se usata come unica trasformazione dei dati, quindi applicata anche ai dati negativi.

Le differenze così marcate tra la GLOG e le altre normalizzazioni portano a pensare che la trasformazione GLOG normalizzi i dati in modo non soddisfacente. Questa considerazione è supportata anche dalle curve ROC presentate in Figura 4.12 relative ai dati simulati con il 150% di *background* dopo aver sostituito i valori negativi con un valore positivo arbitrario: si osservava che la GLOG portava a risultati di SAM insoddisfacenti dal punto di vista della sensibilità e della specificità.

Nel caso reale qui descritto, i dati imputati con valore positivo sono in percentuale confrontabile rispetto a quella dei dati relativi ai dati simulati con 150% di *background*. Nel caso dei dati simulati si erano imputati il 32% dei dati, mentre nel caso reale in analisi la percentuale di dati nulli e quindi imputati è di 25%. È quindi ragionevole pensare che le prestazioni del test SAM dopo la GLOG siano insoddisfacenti anche in questo caso.

Capitolo 5

Conclusioni

La normalizzazione è una fase molto delicata dell'analisi statistica dei dati di *microarray* a causa della estrema sensibilità della tecnologia che porta spesso ad avere valori distorti in modo sistematico. Lo scopo della Tesi è quello di confrontare queste normalizzazioni sulla base della loro influenza nella successiva fase inferenziale. In particolare, il confronto è stato fatto in termini di sensibilità e specificità del test SAM (uno dei test statistici più utilizzati per l'identificazione di geni differenzialmente espressi) e ordinamento della statistica test.

La scelta del modello di simulazione dei dati di espressione è risultato fondamentale: il confronto tra le varie normalizzazioni, infatti, dipende fortemente dai dati di partenza. Non esiste quindi una normalizzazione sempre preferibile alle altre ma a seconda delle caratteristiche dei dati è opportuno utilizzare un modello rispetto ad un altro.

In particolare, per quanto riguarda i modelli bayesiani, la scelta dei parametri di simulazione influenza fortemente l'impatto delle normalizzazioni sull'inferenza: soprattutto per il modello GG il *set* di parametri stimato a partire da SIMAGE sembra portare a dati di espressione più realistici di quelli osservati utilizzando i parametri scelti all'inizio dell'elaborato.

Per quanto riguarda invece i modelli di normalizzazione, alcune considerazioni possono essere fatte a partire dai risultati delle curve ROC e dal

confronto tra le liste di geni. Tali considerazioni sono riportate punto per punto nella restante parte del Paragrafo.

- Quando i dati sono simulati senza alcuna distorsione sistematica, tutte le normalizzazioni hanno lo stesso impatto sulla sensibilità e specificità del test, in questi casi anche le liste dei geni maggiormente sregolati presentano differenze trascurabili.

- Generalmente i geni che nelle liste dei differenzialmente espressi non sono confermati presentano tutti rango elevato, ovvero sono quelli più vicini alla zona di accettazione dell'ipotesi nulla di espressione equivalente.

- Nel caso di errori dipendenti dall'intensità, i modelli di normalizzazione che agiscono cercando di stimare un andamento sistematico dell'MA-Plot per poi sottrarlo ai dati, sembrano normalizzare meglio rispetto a *q-spline* e GLOG.

- In presenza di dati di espressione negativi, dovuti a valori alti di *background*, le *q-spline* e la GLOG invece normalizzano meglio degli altri modelli.

- L'uso della trasformazione GLOG in presenza di molti dati negativi o nulli sostituiti con un valore basso positivo arbitrario (procedura non corretta ma comune nell'ambito dei dati di microarray) porta a valori di sensibilità e specificità molto insoddisfacenti.

- L'ulteriore normalizzazione presente in OSLIN (rispetto ad OLIN) non apporta nessun vantaggio in termini di sensibilità e specificità del test SAM e nessuna differenza di ordinamento della statistica rispetto alla normalizzazione OLIN.

- Il valore del parametro di lisciamento utilizzato da *lowess* non è lontano da quello ottimo stimato in OLIN dato che non vi è molta differenza tra i risultati dei due modelli; tuttavia dato che il costo computazionale di

OLIN è equivalente a quello di *lowess* tale modello riamane preferibile dato che garantisce l'ottimizzazione del parametro.

- Non si sono osservate differenze significative tra le prestazioni della *lowess* generale e della *lowess print-tip* (*P-lowess*).

- Per i dati simulati con i modelli e i parametri utilizzati nel presente elaborato, il modello basato su reti neurali non fornisce risultati preferibili agli altri modelli: il suo costo computazionale elevato rispetto agli altri porta a preferire altri tipi di normalizzazione.

- Il costo computazionale alto del modello NN è verosimilmente legato alla procedura di *cross-validation* utilizzata: sostituire tale procedura con la *cross-validation* generalizzata, utilizzata con buoni risultati in OLIN, potrebbe abbattere i costi elevati della procedura.

- Al contrario i modelli *q-spline* hanno un costo computazionale inferiore agli altri modelli e le loro prestazioni in tutte le situazioni considerate sono almeno paragonabili a quelle degli altri modelli; inoltre si è visto come non sia cruciale la scelta del canale da utilizzare come *target* nella normalizzazione, purchè si sia prima effetuata sui dati una normalizzazione globale.

- La trasformazione GLOG si comporta sempre in modo particolare rispetto alle altre normalizzazioni: sembra fare più fatica a normalizzare gli andamenti sistematici dipendenti dall'intensità, ma è preferibile nel caso si disponga di molti dati negativi (non utilizzabili nei modelli che si basano sul log-rapporto) oppure nel caso in cui le assunzioni di espressione e simmetria su cui si basano gli altri modelli non siano rispettate.

Infine, dall'analisi sui dati reali si è evidenziato come le liste di geni siano uno strumento utile per il confronto delle normalizzazioni quando non sono disponibili indici di bontà che presuppongono la conoscenza a priori dei geni

differenzialmente espressi. In questo modo è possibile evidenziare quali normalizzazioni si comportano in maniera differente e quale sia il rango dei geni riconosciuti da tutti i modelli o soltanto da alcuni.

Appendice A

Grafici relativi al confronto tra le liste di geni

Riportiamo in Appendice i grafici relativi al confronto tra le liste dei geni. In questi grafici, al contrario di quelli presentati nel Capitolo 4, il confronto è effettuato tra i primi geni per rango del test SAM dopo la normalizzazione di riferimento contro i primi geni per rango del test SAM dopo le altre normalizzazioni: i primi 20 contro i primi 20, i primi 100 contro i primi 100 e così via.

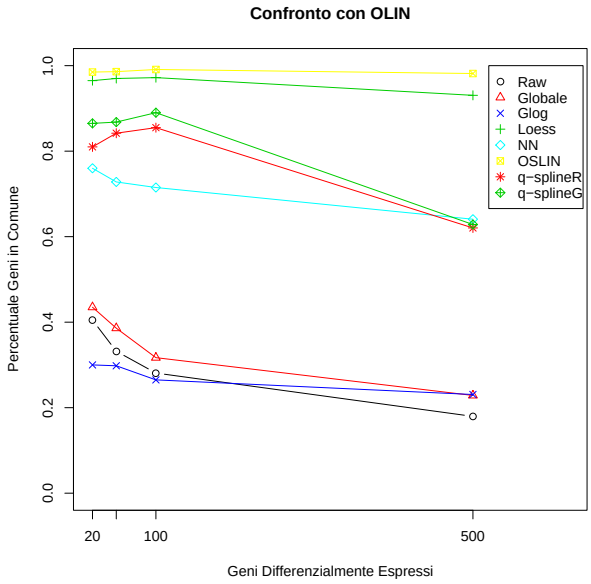

Figura A.1: Confronto con OLIN: dati GG con distorsione.

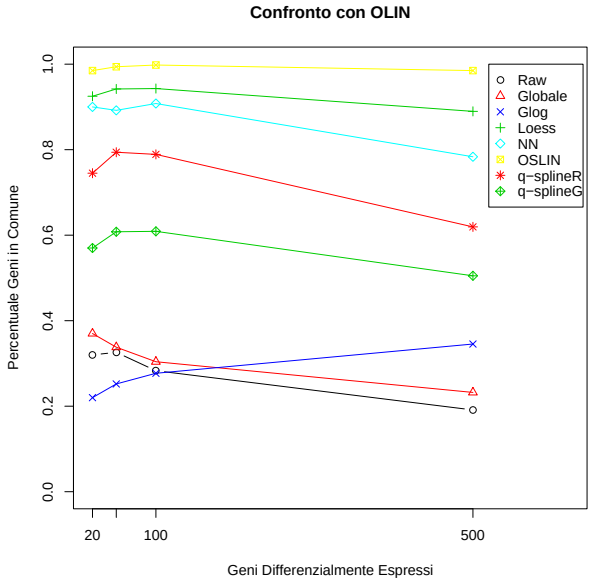

Figura A.2: Confronto con OLIN: dati LNN con distorsione.

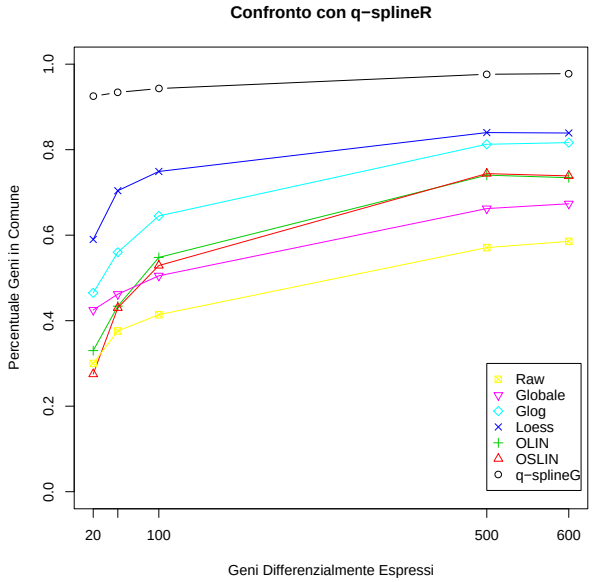

Figura A.3: Confronto con *qsplineR*: SIMAGE con 50% di *background*.

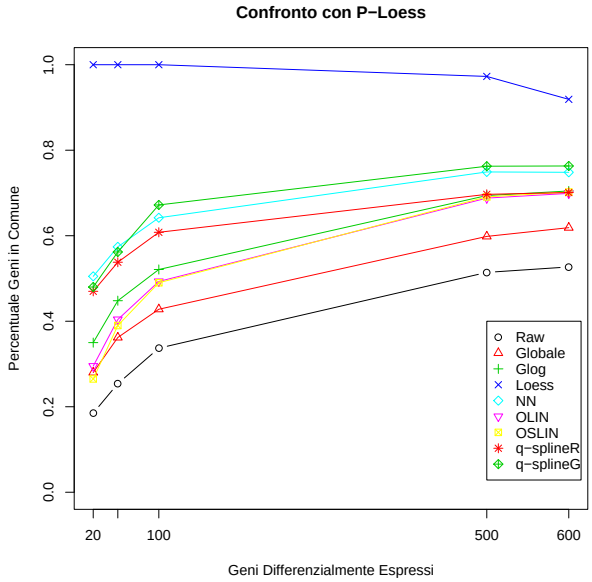

Figura A.4: Confronto con *P-lowess*: SIMAGE con 50% di *background* con dati negativi sostituiti.

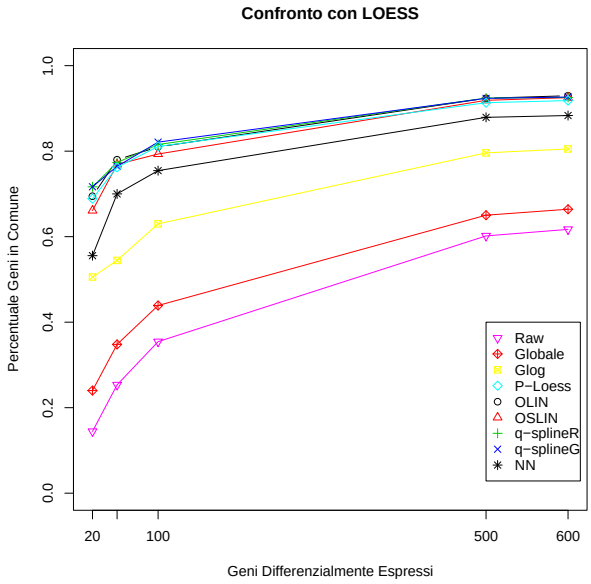

Figura A.5: Confronto con *lowess*: SIMAGE con 10% di *background*.

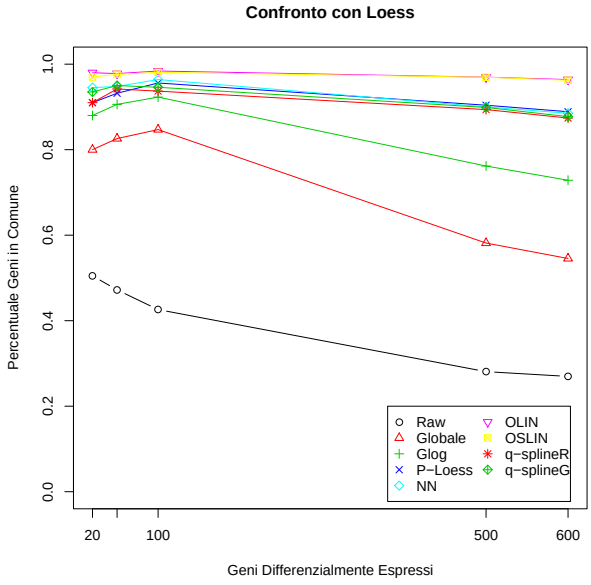

Figura A.6: Confronto con *lowess*: dati GG con nuovi parametri.

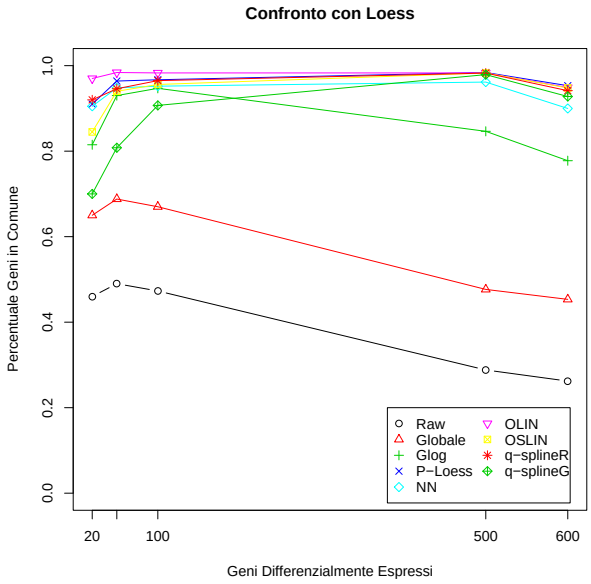

Figura A.7: Confronto con *lowess*: dati LNN con nuovi parametri.

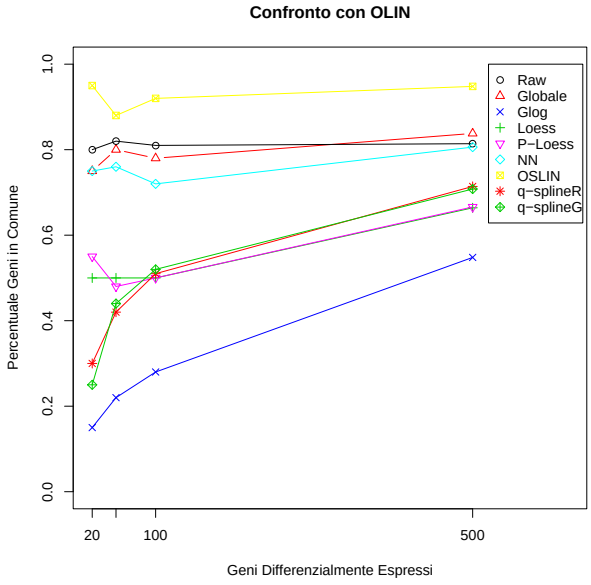

Figura A.8: Confronto con OLIN: dati del Sarcoma di Ewing.

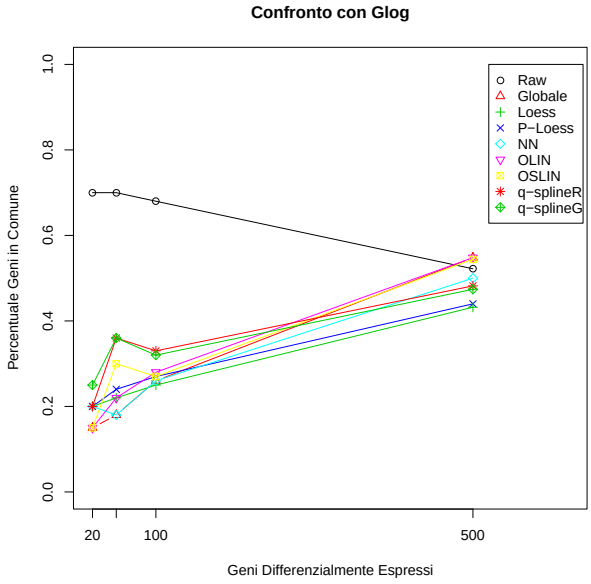

Figura A.9: Confronto con GLOG: dati del Sarcoma di Ewing.

Appendice B

Parametri usati per le simulazioni con SIMAGE

```
Array number of grid rows  9
Array number of grid columns  4
Number of spots in a grid row  18
Number of spots in a grid column  18
Number of spot pins  16
Number of technical replicates  1
Number of genes (0 = max) 10000
Number of slides  15
Perform dye swaps no
Gene expression filter yes
Reset gene filter for each slide no
Mean signal  11.492
Change in log2ratio due to upregulation  0.832
Change in log2ratio due to downregulation  -0.605
Variance of gene expression  1.775
% of upregulated genes  3
% of downregulated genes  3
Correlation between channels  0.89
```

```
Dye filter no
Reset dye filter for each slide yes
Channel (dye) variation  0.51
Gene x Dye  0
Error filter yes
Reset error filter for each slide yes
Random noise standard deviation  0.219
Tail behaviour in the MA plot  0.11
Non-linearity filter yes
Reset non-linearity filter for each slide yes
Non-linearity parameter curvature  0.025
Non-linearity parameter tilt  0.777
Non-linearity from scanner filter yes
Reset non-linearity scanner filter for each slide yes
Scanning device bias (0 = clipped; 1 = fully non-linear) 0.295
spotpin deviation filter yes
Reset spotpin filter for each slide no
spotpin variation  0.36
Background filter yes
Reset background filter for each slide yes
Number of background densities  2
Mean standard deviation per background density  0.3
Maximum of the background signal (%)
 relative to the non-background signals  50
Standard deviation  of the random noise for the background signals
0.1
Background gradient filter yes
Reset gradient filter for each slide yes
Maximum slope of the linear tilt  700
Missing values filter yes
Reset missing spots filter for each slide yes
```

```
Number of hairs  10
Maximum length of hair  20
Number of discs  6
Average radius disc  10
Number of missing spots  1000
```

Bibliografia

Albers C.; Jansen R.; Kok J.; Kuipers O.; van Hijum S. (2006). Simage: simulation of dna-microarray gene expression data. *BMC Bioinformatics*, **7**(1), 205.

Baird K.; Davis S.; Antonescu C. R.; Harper U. L.; Walker R. L.; Chen Y.; Glatfelter A. A.; Duray P. H.; Meltzer P. S. (2005). Gene Expression Profiling of Human Sarcomas: Insights into Sarcoma Biology. *Cancer Res*, **65**(20), 9226–9235.

Cleveland W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, **74**(368), 829–836.

Craven P.; Wahba G. (1979). Smoothing noisy data with spline functions. *Numerische Mathematik*, **31**(4), 377–403.

Dudoit S.; Yang Y.; Callow M.; Speed T. (2002). Statistical methods for identifying differentially expressed genes in replicated cdna microarray experiments. *Statistica Sinica*, **12**, 111–139.

Durbin B.; Hardin J.; Hawkins D.; Rocke D. (2002). A variance-stabilizing transformation for gene-expression microarray data. *Bioinformatics*, **18**(suppl_1), S105–110.

Futschik M.; Crompton T. (2004). Model selection and efficiency testing for normalization of cdna microarray data. *Genome Biology*, **5**(8), R60.

Hastie T. (1992). Generalized additive models In *Statistical Models in S*. A cura di Chambers J., Hastie T., capitolo 7. Wadsworth & Brooks/Cole.

Loader C. (1999). *Local regression and likelihood*. Springer.

Munson P. (2001). A "consistency" test for determining the significance of gene expression changes on replicate samples and two convenient variance-stabilizing transformations. In *Genelogic Workshop on Low Level Analysis of Affymetrix Genechip® data*. http://www.stat.berkeley.edu/∼terry/zarray/Affy/GL_Workshop/genelogic2001.html.

Newton M. A.; Kendziorski C. M.; Lan H.; Gould M. N. (2003). On parametric empirical bayes methods for comparing multiple groups using replicated gene expression profiles. *Statistics in Medicine*, **22**(24), 3899–3914.

Tarca A. L.; Cooke J. E. K. (2005). A robust neural networks approach for spatial and intensity-dependent normalization of cDNA microarray data. *Bioinformatics*, **21**(11), 2674–2683.

Troyanskaya O.; Cantor M.; Sherlock G.; Brown P.; Hastie T.; Tibshirani R.; Botstein D.; Altman R. (2001). Missing value estimation methods for dna microarrays. *Bioinformatics*, **17**(6), 520–525.

Tusher V. G.; Tibshirani R.; Chu G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences*, **98**(9), 5116–5121.

Wit E.; McClure J. (2004). *Statistics for microarrays : design, analysis and inference*. Wiley.

Workman C.; Jensen L.; Jarmer H.; Berka R.; Gautier L.; Nielser H.; Saxild H.-H.; Nielsen C.; Brunak S.; Knudsen S. (2002). A new non-linear normalization method for reducing variability in dna microarray experiments. *Genome Biology*, **3**(9), research0048.1–research0048.16.

Wu W.; Xing E.; Myers C.; Mian I. S.; Bissell M. (2005). Evaluation of normalization methods for cdna microarray data by k-nn classification. *BMC Bioinformatics*, **6**(1), 191.

Yang Y. H.; Dudoit S.; Luu P.; Lin D. M.; Peng V.; Ngai J.; Speed T. P. (2002). Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucl. Acids Res.*, **30**(4), e15–.