UNIVERSITÀ DEGLI STUDI DI PADOVA

FACOLTÀ DI SCIENZE STATISTICHE

CORSO DI LAUREA IN STATISTICA, ECONOMIA E FINANZA

RELAZIONE FINALE

Analisi esplorativa di dati di espressione genica di tipo time course

Relatore: Ch.ma Prof.ssa MONICA CHIOGNA

Laureando: MARCO BANTERLE

Matricola n. 580485/SEF

ANNO ACCADEMICO 2009-2010

Indice

Introduzione

Gli esperimenti di tipo time course riguardano geni la cui espressione viene misurata ripetutamente, al fine di studiarne una eventuale dinamica dipendente dal tempo. Nello studio in esame vengono trattati geni regolatori di un ciclo biologico giornaliero, che vengono di conseguenza osservati, cioè ne viene rilevata l'espressione, per un'intera giornata ad intervalli regolari di 4 ore. Queste misurazioni vengono inoltre effettuate sotto quattro condizioni differenti, in modo da evidenziare possibli differenze in realzione agli stimoli ricevuti.

In questo contesto vengono quindi trattati metodi grafici per l'analisi esplorativa di questo tipo di esperimenti, dove le misurazioni ripetute portano ad avere dati multivariati con alta dimensionalità, e tecniche di segmentazione statistica per individuare in modo formale gruppi di geni affini o comportamenti particolarmente diversi dalla norma.

Verranno esposte quindi le principali caratteristiche dei grafici in coordinate parallele, per giungere poi ai diagrammi di Andrews, loro generalizzazione, che saranno utilizzati anche nel seguito per la visualizzazione dei risultati ottenuti tramite le classificazioni. Si cercherà di verificare infatti se l'ora e la condizione in cui viene misurata l'espressione abbiano un effetto significativo e soprattutto se questo effetto è diverso per qualche gruppo di geni rispetto agli altri.

Per lo studio in analisi è inoltre presentata inizialmente una partizione di geni che vengono definiti da chi ha reso disponibili i dati come 'oscillanti'. Si cercherà quindi di verificare l'effettiva presenza di questa classificazione e di proporne delle nuove con lo scopodi facilitare il lavoro di analisi, restringendo l'altrimenti cospicuo numero di geni in esame.

Capitolo 1: I dati

I dati che verranno analizzati riguardano un esperimento cosiddetto time course nel quale vengono misurati ad intervalli regolari i livelli di espressione di un insieme di geni che regolano un ciclo biologico non meglio specificato.
Più precisamente viene presentata una lista di circa diciassettemila probe set, cioè delle porzioni di DNA progettate specificatamente per rilevare l'espressione di un singolo gene dell'array in esame, per ognuno dei quali vengono riportate 24 rilevazioni: sei misure consecutive, una ogni quattro ore dalle 00.00 alle 20.00, in quattro esperimenti diversi, nei quali i tessuti sono sottoposti a diverse codizioni.
Il livello di espressione viene misurato nel dataset in questione come (una trasformata de) la quantità di mRNA, trascritto dal DNA della cellula per la sintesi di macromolecole funzionali (generalmente si tratta di proteine), rilevato nel campione sotto analisi ed è solitamente codificato in quantità diverse a seconda della condizione in cui si trova la cellula stessa.
Nel caso in esame vengono effettuate le misurazioni sottoponendo i tessuti a stimoli elettrici repentini o meno (rispettivamente condizione veloce o lenta) in modo da indurre condizioni innervate oppure viceversa questi stimoli verranno fatti mancare per misurare la risposta in condizioni denervate.
Trattandosi, inoltre, di geni che regolano un ciclo biologico giornaliero (detto circadiano, dal latino *circa* attorno *diem* giorno) ci si aspetta che durante la giornata essi producano quantità diverse di mRNA e di conseguenza vengano rilevati diversi livelli di espressione.
Ad ogni momento d'interesse, per ogni condizione, vengono quindi effettuate numerose misurazioni 'grezze' per ogni probe set che vengono quindi sintetizzate, attraverso un algoritmo denominato PLIER, Probe Logaritmic Intensity Error (per dettagli sull'algoritmo si veda ad esempio la nota tecnica *"Affymetrix, Guide to Probe Logarithmic Intensity Error (PLIER) Estimation"*), in un unico valore di espressione che viene infine riportato nei dati.

Sono stati resi disponibili due dataset, uno che verrà chiamato dataset completo, contenente tutti i ventiquattro livelli di espressione sui 17313 probe set ed uno meno numeroso in cui vengono selezionati solo 577 geni che 'oscillano' in tutte e quattro le condizioni, che verrà denominato dataset ridotto. La definizione di oscillazione in questo caso è però prettamente soggettiva e stabilita dai biologi che hanno fornito i dati, o almeno così pare, che etichettano come 'oscillanti' tutti quei geni per i quali il rapporto tra il massimo ed il minimo livello di espressione supera il valore 1.5 e che fanno registrare almeno una misurazione superiore al valore 100. Non sono state però fornite indicazioni sulla scelta di questi valori soglia e quindi si cercherà innanzitutto di capire se sia statisticamente possibile distinguere tra i due gruppi o se le soglie scelte non permettano una separazione netta dei due gruppi.

Capitolo 2: Metodi

2.1 Visualizzazione di dati multidimensionali

2.1.1 Coordinate Parallele

La visualizzazione dei dati è sempre una parte importante, perlomeno delle analisi preliminari, in ogni studio statistico.
L'elevato numero di situazioni in cui l'espressione viene misurata in questo esperimento rende però inaccessibili le classiche visualizzazioni basate sui diagrammi a dispersione o simili.
Anche considerando una sola condizione, infatti, la caratteristica in esame viene misurata sei volte, rendendo impossibile, o quantomeno dispersiva e poco intuitiva, la visualizzazione in coordinate cartesiane.
Proprio per lo studio di dati multivariati sono state progettate quindi le cosiddette Coordinate Parallele che propongono di disegnare un punto di uno spazio, supponiamo k-dimensionale, come una spezzata in uno spazio bidimensionale i cui vertici si trovano su k assi verticali (generalmente paralleli ed equispaziati).
La posizione del vertice della linea sull' i-esimo asse rappresenta quindi la i-esima coordinata del punto. (cfr Figura1)

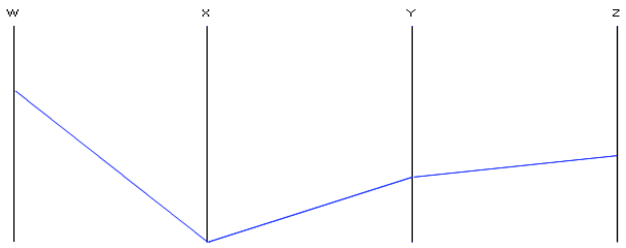

Figura 1: Semplice esempio di coordinate parallele per un solo punto in quattro dimensioni

Anche avendo scelto come rappresentare tutte le caratteristiche per ogni unità statistica, lo studio in analisi pone però un ulteriore problema, oltre al fatto

ovvio che parte dell'informazione verrà persa per via della proiezione da uno spazio k-dimensionale ad uno bidimensionale.

Su questo tipo di coordinate è stata sviluppata un'ampia letteratura, sia puramente matematica (sono per esempio note le traiettorie tracciate dalle più comuni figure geometriche e gli effetti di trasformazioni lineari e non), che più prettamente statistica (criteri per l'ordinamento delle variabili ad esempio, cruciali per ricavare dai grafici informazioni utili), ma se per un numero limitato di punti questa rappresentazione permette di avere un'idea chiara di come 'si muovono' i dati, e quindi di trarre importanti conclusioni, per un numero elevato di unità si rischia di disegnare un grafico molto confuso. Questo è, ad esempio, quello che si ottiene disegnando senza criterio i dati in questione (espressione dei geni del dataset ridotto nella prima condizione, in questo caso) in coordinate parallele (cfr Fig.2).

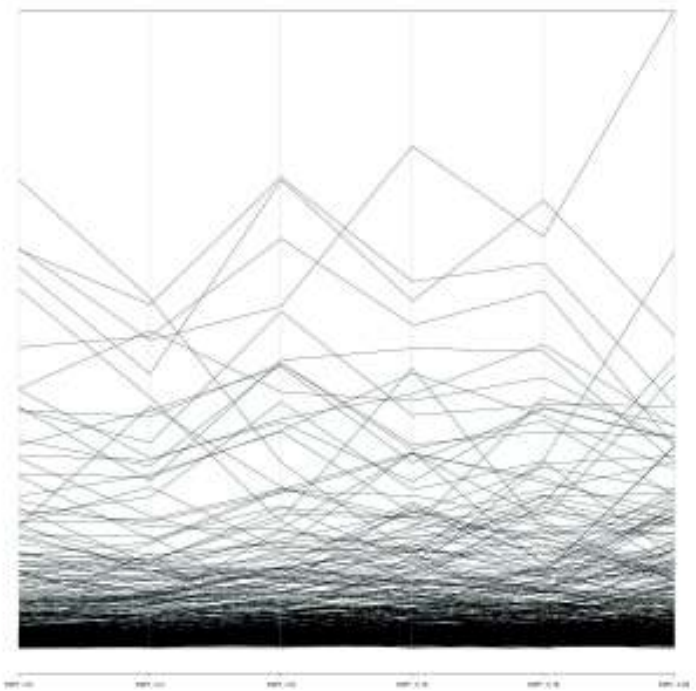

Figura 2: Grafico in coordinate parallele dei livelli di espressione dei geni del dataset ridotto

Si guadagna quindi in completezza, poiché per ogni probe-set è ora possibile osservare tutti i livelli di espressione rilevati, anche tutti e ventiquattro contemporaneamente, ma il risultato è probabilmente troppo caotico per essere chiaramente interpretato.

Il passo successivo dunque è cercare una generalizzazione di questi sistemi di coordinate che permetta di osservare più chiaramente le traiettorie seguite dai probe-set in esame, per individuare pattern comuni, outliers e così via.
Fortunatamente anche in questo senso si ha a disposizione molto lavoro già svolto, ad esempio *Moustafa e Wegman (2002 e 2006),* che suggeriscono alcune trasformazioni, come rotazioni o trasformazioni lineari degli assi coordinati, o generalizzazioni degli assi stessi in piani, ottenendo la possibilità di rappresentare coppie variabili assieme e di visualizzare traiettorie in tre dimensioni (cfr Fig.3).

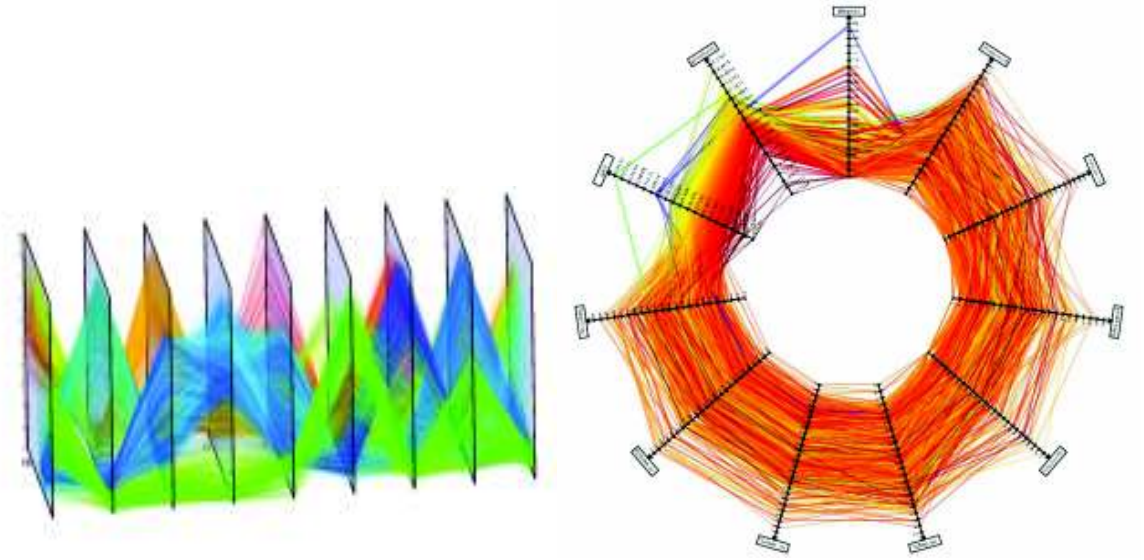

Figura 3: Alcune generalizzazioni dei grafici in coordinate parallele

2.1.2 Diagrammi di Andrews

Il tipo di grafico che verrà usato più di frequente in questo lavoro è infatti ottenibile attraverso una trasformata delle coordinate parallele, in particolare tramite un'interpolazione in serie di Fourier dello stesso grado della dimensione dei dati (si veda *Moustafa e Wegman, 2002* per maggiori dettagli), seppure sia stato sviluppato indipendentemente (ed antecedentemente) al lavoro di questi autori, da D. F. Andrews nel 1972 (cfr *Andrews, 1972).*
In questo caso l'idea per rappresentare un punto k-dimensionale in due dimensioni è quella di definire una famiglia di dimensione k di funzioni

periodiche di una variabile in una variabile, i cui coefficienti (in numero $\geq k$) dipenderanno dai k valori osservati per ogni unità. Il risultato è quindi, similmente (ma solo a livello grafico) a quello che prima era una spezzata, una curva per ogni unità statistica, disegnata per convenzione (visto che le funzioni risultano periodiche in questo intervallo, ma qualsiasi intervallo di lunghezza maggiore od uguale fornisce le stesse informazioni) in $(-\pi, \pi)$.

In particolare, nei cosiddetti diagrammi di Andrews, la famiglia di funzioni è definita per $x = (x_1, x_2, x_3, ...x_k)$ come

$$f_x(t) = x_1 / \sqrt{2} + x_2 \sin(t) + x_3 \cos(t) + x_4 \sin(2t) + x_5 \cos(2t) + \ldots .$$

Il risultato è un diagramma simile a quello che segue (cfr Fig. 4).

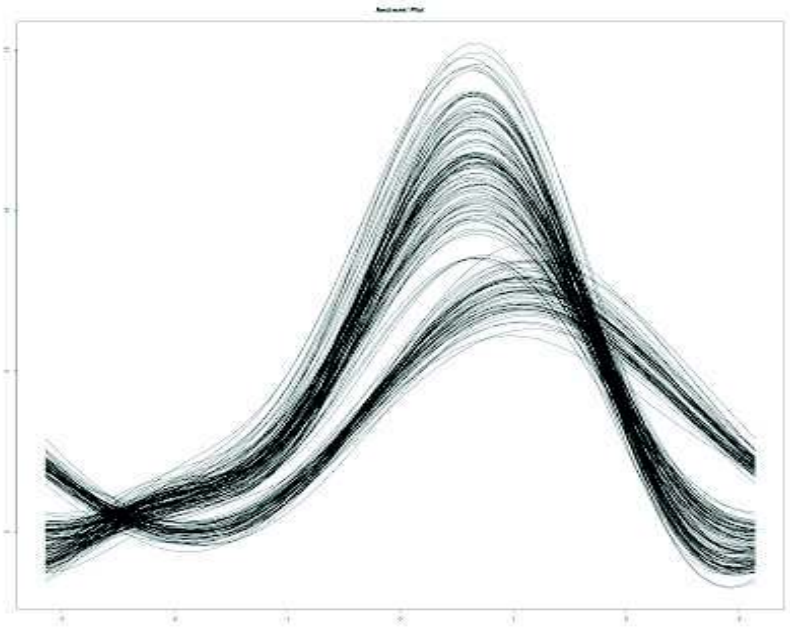

Figura 4: Andrews Plot sui dati degli Iris di Fisher

A prima vista però potrebbe sembrare che l'utilizzo di una complicata serie di Fourier, il cui sforzo computazionale per essere calcolata non è certo insignificante (seppure non troppo elevato), soprattutto all'aumentare della dimensione dei punti, non valga il risultato ottenuto.

Ora si hanno delle curve al posto delle spezzate ma non sempre si sarà così fortunati nell'individuazione dei gruppi come nell'esempio appena riportato. Altre volte il grafico risulterà confuso e difficilmente interpretabile proprio come accadeva per le coordinate parallele. Il vantaggio dei diagrammi di Andrews, però, stà in alcune interessanti proprietà di cui queste rappresentazioni godono.

- **Mantengono la media.**

Se $\bar{x}$ è la media di n osservazioni multivariate x_i, allora la funzione rappresentante $\bar{x}$ è la media delle n funzioni corrispondenti alle osservazioni.

$$f_{\bar{x}}(t) = 1/n \sum_{i=1}^{n} f_x(t)$$

Di conseguenza la media delle funzioni visivamente individuata nel grafico risulta essere realmente la funzione corrispondente alla media delle osservazioni.

- **Preservano** una proporzionalità con **la varianza.**

Se le componenti dei dati sono incorrelate e con varianza comune σ^2 allora per ogni x la funzione in un generico t ha varianza

$$var[f_x(t)] = \sigma^2 \left(1/2 + \sin^2(t) + \cos^2(t) + \sin^2(2t) + ...\right)$$

che si riduce quindi alla costante $(1/2)\sigma^2 k$ se k è dispari e sta tra $(1/2)\sigma^2(k-1)$ e $(1/2)\sigma^2(k+1)$ se k è pari. (In quest'ultimo caso la varianza dipende leggermente da t, dipendenza che diminuisce all'aumentare di k).

- **Mantengono le distanze.**

Punti vicini nello spazio k-dimensionale risulteranno essere funzioni vicine nel grafico; in particolare l'integrale del quadrato dell'area compresa tra due funzioni (quantità facilmente interpretabile come distanza 'visiva' tra due curve) è proporzionale alla distanza euclidea tra due punti nello spazio k-dimensioanle, e ciò favorisce notevolmente il loro utilizzo con scopi di raggruppamento.

- **Inducono proiezioni su uno spazio mono-dimensionale.**

Per un dato valore t_0 , il valore della funzione $f_x(t_0)$ è proporzionale alla proiezione del punto $x=(x_1, x_2, ..., x_k)$ su

$$f_1(t_0) = (1/\sqrt{2}, \sin(t_0), \cos(t_0), \sin(2t_0), \cos(2t_0), ...)$$

Queste proiezioni sono facilmente osservabili e potrebbero rivelare ad esempio gruppi nei dati o marcarli come outlier rispetto agli altri.

Il vantaggio di questo tipo di rappresentazione sta proprio nel riuscire a mostrare un elevato numero (in teoria infinito) di queste proiezioni su un grafico solo e ciò tornerà utile in seguito.

Queste proprietà permettono inoltre di fare alcune notevoli osservazioni:

– se un fascio di funzioni si mantiene compatto in tutto il grafico, i punti da cui le funzioni sono ricavate si troveranno vicini anche in termini euclidei e quindi formeranno con ogni probabilità un cluster in una possibile analisi dei gruppi;

– per un fissato valore di t è possibile formulare un test per verificare che $f_x(t) = f_y(t)$ per due punti x e y generici.

Data la varianza della funzione $f_x(t)$, se viene assunto che le componenti di x siano indipendenti e normalmente distribuite si ottiene che

$$z = \left[f_x(t) - f_y(t)\right] / var\left[f_x(t)\right]^{1/2}$$

si distribuisce secondo una normale standard sotto l'ipotesi nulla.

È possibile quindi testare per un valore di t , scelto a priori, se due punti siano uguali a livello statistico oppure per un singolo punto costruire un intervallo di livello α per cui le funzioni che vi cadono all'interno non sono distinguibili da quella di riferimento;

– se invece si volesse creare un test “complessivo” sull’intera funzione è possibile, ricordando che è stato assunto che le x_i siano indipendenti e normalmente distribuite, con varianza comune σ^2 e media μ_i, allora $w = \| x-y \|^2 / \sigma^2$ si distribuisce come una χ^2_k.

Sappiamo inoltre che la distanza al quadrato della proiezione $p(x-y)$ di $x-y$ su un qualsiasi spazio monodimensionale non è più grande della sua distanza euclidea al quadrato, ovvero

$$p(x-y) \quad \leq \quad \| x-y \|^2$$

e quindi, considerato che la proiezione sul vettore $v = f_1(t)/[f_1'(t)f_1(t)]^{1/2}$ non è altro che $|(x-y)'v|^2 = |f_x(t)-f_y(t)|^2/[f_1'(t)f_1(t)]$, con probabilità $1-\alpha$, per tutti i valori di t,

$$|f_x(t)-f_y(t)|^2 \quad \leq \quad \sigma^2 |f_1(t)| \chi^2_k(\alpha) \quad \leq \quad (\frac{k+1}{2})\sigma^2\chi^2_k(\alpha)$$

dove $\chi^2_k(\alpha)$ è il quantile α della distribuzione χ^2_k.

La funzione $f_x(t)$ descrive quindi attorno a sé una banda di confidenza di livello $1-\alpha$ di ampiezza fissata e se $f_y(t)$ cade al di fuori di questa banda per qualche t c’è evidenza contro l’ipotesi nulla $H_0 : x = y$.

2.2 Tecniche statistiche di classificazione

2.2.1 Introduzione alla classificazione

Classificare significa assegnare un’unità statistica ad un gruppo con cui condivide alcune caratteristiche in esame. Si distinguono generalmente due tipi di classificazione, denominate supervisionata (o analisi discriminante) e non supervisionata (o analisi dei gruppi).

Nella classificazione supervisionata è nota, per un campione di dati, la popolazione di appartenenza per ogni unità statistica. Si cerca quindi di individuare un modello che permetta di stimare per unità delle quali si conosce la provenienza, la probabilità di appartenenza ad ogni gruppo condizionatamente alle caratteristiche osservate e quindi di assegnare ogni nuova unità ad una popolazione, minimizzando gli errori.

Al contrario, l'analisi dei gruppi (o classificazione non supervisionata) cerca di dividere i dati in gruppi omogenei di unità tra loro simili, non sapendo se realmente delle strutture di gruppo esistano nell'insieme dei dati. Va da sé che un'analisi di quest'ultimo tipo è più informativa più i gruppi appaiono con naturalezza dai dati. Esistono inoltre due principali tipologie di analisi dei gruppi: quella gerarchica e quella non gerarchica; la prima ci permette di formare una famiglia di partizioni di dimensione da 1 a n, partendo da quella banale in cui tutte le unità formano un gruppo distinto per arrivare a quella che prevede n gruppi, uno per ogni unità; i metodi non gerarchici invece forniscono una sola partizione in un numero g fissato a priori di gruppi.

In questo lavoro si tratterà, ovviamente, visto che non è noto alcun raggruppamento possibile dei geni, principalmente di analisi dei gruppi ma in particolare verranno sfruttate le tecniche di questo ambito per effettuare quella che in gergo si chiama dissezione. Spesso verrà infatti forzata una struttura di gruppo sui dati, verosimile o no, semplicemente per poter scartare la possibilità che esistano davvero sottogruppi all'interno del dataset o per identificare eventuali outlier.

2.2.2 Classificazione supervisionata

In presenza di un vettore, diciamo y, di lunghezza n, che descriva per ogni unità il gruppo a cui essa appartiene, pare naturale cercare di costruire la regola per prevedere la popolazione d'appartenenza attraverso l'apprendimento di un modello che utilizzi y come risposta e la matrice X delle variabili come covariate.

In un' unica situazione però si avrà la possibilità di conoscere y, per verificare un'ipotesi di interesse ma non strettamente inerente ai dati, ed esso assumerà due sole modalità; descriveremo solo una di queste tecniche quindi, quella strettamente utilizzata nel seguito, e cioè la regressione logistica (per una trattazione approfondita riguardante le tecniche di classificazione si rimanda ad esempio ad *Azzalini e Scarpa, 2004*).

Codificate le due modalità di y come 0 ed 1 e definita la funzione logistica come $f(z)=\frac{e^z}{1+e^z}$, la sua immagine rappresenta in questo caso la probabilità attesa di appartenenza alla categoria '1' dell' i-esima unità dato il vettore delle caratteristiche x_i. Quindi

$E[y_i \mid x_i]= f(z)$ con $z = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_p x_{pi}$.

Verranno quindi classificate come facenti parte del gruppo '1' tutte quelle unità per le quali tale probabilità attesa sarà più alta della soglia di 0.5 e viceversa per le altre.

2.2.3 Classificazione non supervisionata : introduzione

Parlando invece di metodi di analisi dei gruppi, in genere è sensato innanzitutto verificare l'esistenza di strutture di gruppo all'interno dei dati per non rendere inutile se non addirittura fuorviante l'uso di tali tecniche.
Esistono diversi metodi per saggiare quest'ipotesi, dai più formali test d'ipotesi ad altri di tipo più esplorativo come ad esempio l'osservazione di un istogramma k-dimensionale. Come già detto, però, in quest'analisi si è particolarmente interessati a 'dissezionare' i dati, segmentandoli in gruppi non necessariamente sensati per individuare eventuali outlier e quindi questa premessa non verrà presa in considerazione.

Sia per essere in grado di formare dei gruppi di elementi simili sensati che per poterli segmentare con una qualche regola semplicemente matematica, è necessario in ogni caso definire cosa intendiamo per *simili*.
Deve essere quindi scelto un indice di similarità e, nel caso dei metodi gerarchici, una definizione univoca di distanza tra due gruppi.
Se le unità statistiche sono descritte esclusivamente tramite variabili continue, l'indice di prossimità utilizzato è generalmente un qualche tipo di distanza (quasi sempre euclidea); altrimenti, vengono utilizzati dei più semplici indici di uguaglianza tra caratteristiche delle unità statistiche (ad esempio indici funzione delle co-presenze o simili, si veda *Zani, 2000* per maggiori informazioni a riguardo).
Quest'ultimi hanno però un problema all'aumentare della dimensionalità dei vettori delle caratteristiche, in quanto, ovviamente, il confronto tra k caratteristiche, con $k >> 1$, porta ad utilizzare indici molto imprecisi che spesso ottengono valori molto bassi.
Fortunatamente i dati qui analizzati sono tutti di natura continua e quindi verrà utilizzata in tutti i casi la distanza euclidea (che ricordiamo essere preservata dalla rappresentazione di Andrews) e per i metodi gerarchici generalmente verrà definita la distanza tra due gruppi come la distanza tra gli elementi più vicini che ne fanno parte (nearest neighbour):

$$d(C_1, C_2) = min(d_{rs}), perr \in C_1 es \in C_2$$

Questo tipo di distanza tra cluster viene anche chiamato col nome di legame singolo, per differenziarsi da altri possibili tipi che possono essere definiti, tra cui il legame completo (o furthest neighbour) od il legame medio.
Per tutti i metodi è uso comune verificare la sensatezza ed in particolare la stabilità delle partizioni ottenute, ad esempio al variare delle variabili considerate tra tutte le possibili o delle partizioni iniziali, tramite l'analisi grafica o attraverso degli indici adatti.

2.2.4 Analisi dei gruppi : Metodi gerarchici

Come già specificato, un metodo di analisi dei gruppi di tipo gerarchico fornisce una famiglia di partizioni e non una sola possibilità di suddivisione, unendo tutti gli elementi che presentano distanze inferiori o uguali a tutti i livelli di distanza tra 0 (partizione banale con tutti gli elementi separati) e $+\infty$ (partizione banale formata dall'intero insieme dei dati).

Esistono diversi tipi di algoritmi che effettuano queste suddivisioni, in particolare alcuni partono dalla massima aggregazione e via via dividono da essa i singoli elementi con distanza massima, chiamati disgregativi, o che partono dai singoli disgiunti elementi e via via aggregano i più vicini fino a formare la partizione che li contiene tutti; quale che sia 'la direzione' scelta, la partizione corrispondente ad un livello di distanza contiene sempre tutte le partizioni ottenute con i livelli più bassi. Per i metodi aggregativi ad esempio ciò significa che due elementi che sono stati uniti dall'algoritmo non verranno più divisi nei passi successivi.

Queste partizioni vengono generalmente rappresentate attraverso un dendrogramma, cioè un albero n-dimensionale i cui nodi rappresentano le partizioni successive che dalla radice (partizione contenente tutti gli elementi) arrivano alle singole foglie/elementi. Ad esso viene generalmente affiancata un'indicazione del livello di distanza, in modo da poter disegnare i rami proporzionali in lunghezza al livello di distanza necessario per quell'aggregazione. (cfr Figura 5)

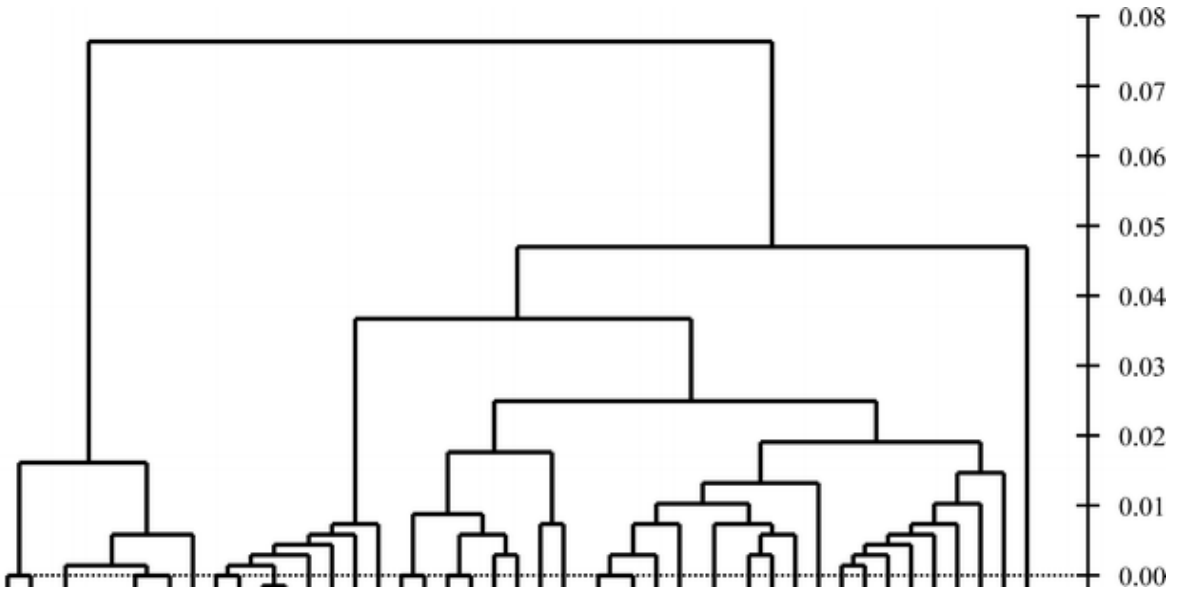

Figura 5: dendrogramma e relativa scala dei livelli di distanza

I problemi principali che limitano l'uso di tali metodi in questa sede sono entrambi legati all'alta numerosità dei dati.

Un dendrogramma con circa seicento foglie non sempre risulta di facile interpretazione, diventando praticamente impossibile da leggere se disegnato per il dataset completo; inoltre ad ogni passo sulla scala delle distanze l'algoritmo deve ri-calcolare l'intera matrice di distanza utilizzate, comportando un dispendio di tempo non indifferente.

Si deve disporre inoltre di un criterio che permetta di decidere quale partizione utilizzare per i dati, una volta ottenuta l'intera famiglia, e spesso l'osservazione del dendrogramma è necessaria ma non sufficientemente rigorosa per fornirci la risposta.

Sono però molto utili nell'individuazione dei cosiddetti outliers, che per definizione sono dati in un certo senso 'distanti' dalla norma e quindi vengono spesso aggregati (o rispettivamente disgregati) molto tardi (presto) e solo al 'costo' di un livello di distanza molto elevato.

2.2.5 Analisi dei gruppi : Metodi non gerarchici

Contrariamente alle tecniche esposte fin'ora, i metodi non gerarchici, o di partizione, portano alla formazione di una sola suddivisione con un numero di gruppi g fissato a priori dal ricercatore.

È possibile specificare una funzione obiettivo, tipicamente funzione della scomposizione della devianza totale in devianza tra i gruppi e devianza nei gruppi, ed arrivare al risultato tramite un semplice problema di ottimizzazione, risolto solitamente in modo iterativo ed eventualmente fermato al raggiungimento di alcune condizioni limite.

Il punto di forza dei metodi non gerarchici è proprio quello di sopperire alle problematiche evidenziate invece per i metodi gerarchici; essi sono cioè spesso rapidi nell'esecuzione e non richiedono un uso intensivo del calcolatore. Sono inoltre facilmente rappresentabili e non è necessario produrre complessi dendrogrammi per visualizzare l'intera famiglia di partizioni.

Anche questi metodi non sono però esenti da problemi, ad esempio non è banale scegliere in modo corretto il numero dei gruppi, a meno che non sia in qualche modo noto, né la configurazione di partenza dei cluster (necessaria per il primo passo degli algoritmi iterativi); è inoltre praticamente impossibile essere sicuri di aver minimizzato la funzione obiettivo e di non trovarsi invece in, od in prossimità di, un minimo locale, vista l'altissima numerosità di configurazioni possibili anche con n (numero di unità) e g (numero di gruppi) non troppo grandi.

Generalmente si ovvia a questi inconvenienti usando dei semplici accorgimenti; per determinare il numero ottimo di gruppi, ad esempio, è prassi comune ripetere l'analisi per alcuni valori ritenuti verosimili, che tengano conto del principio di parsimonia, e confrontare le partizioni ottenute eventualmente anche tramite degli indici di coesione o stabilità.

Non è raro, inoltre, decidere il range di valori utilizzati per g tramite l'osservazione di un dendrogramma fornito da una segmentazione gerarchica sugli stessi dati.

2.2.6 L' algoritmo delle k - medie

Senza il rischio di perdere di generalità, trattandosi dell'algoritmo più comunemente usato, verrà trattato come esempio di analisi dei gruppi non gerarchico, il metodo delle k-medie.

Esso si articola in un semplice algoritmo iterativo in tre passi:

- Si sceglie la configurazione iniziale, ovvero i k semi (punti nello spazio p-dimensionale, con p dimensione del vettore delle caratteristiche in esame), e si costruiscono i k gruppi iniziali associando ogni unità al seme più vicino.
- Si calcolano i centroidi dei k cluster e la distanza da essi di ogni unità.

Se la distanza minima non è ottenuta con il centroide del gruppo di appartenenza l'unità cambia cluster in favore del più vicino.

- Si ripete il secondo punto fino a raggiungere la stabilità dei cluster.

Nella realtà, spesso, raggiungere la stabilità completa comporta tempi di calcolo molto lunghi e quindi si decide di interrompere la procedura quando i centroidi non si spostano al di sopra di una certa soglia o, in alcuni casi limite, al raggiungere di un numero massimo di iterazioni.

Si può dimostrare che per l'algoritmo delle k-medie l'impiego della norma eucliedea come indice di prossimità garantisce la convergenza e che la partizione così ottenuta ha come funzione da ottimizzare un criterio di coesione fondato sulla devianza nei gruppi.

L'unico punto in sospeso resta quindi il metodo di scelta dei semi per la partenza dell'algoritmo; seppure spesso si proponga di eleggere a seme k unità scelte a caso (o le prime k), pare più sensato scegliere k punti in modo che siano rappresentativi dello spazio p-dimensionale dove l'analisi si svolge.

Tale accorgimento riduce spesso sensibilmente il tempo di convergenza del metodo ed inoltre sembra più corretto a livello intuitivo. È spesso perseguito attraverso la scelta di punti sufficientemente spaziati e/o che siano tra loro più distanti di quanto non lo siano con le unità che li circondano.

Nota: in quest'ultimo paragrafo, per coerenza con la letteratura di riferimento, si è deciso di utilizzare p come dimensione dello spazio delle caratteristiche, in favore di k (utlizzato in precedenza) che ora rappresenta il numero di gruppi scelto, da cui il nome dell'algoritmo "k-medie".

Capitolo 3: Analisi dei dati

3.1 I due insiemi di dati

Riepilogate in maniera formale le tecniche che verranno utilizzate, verrà ora descritta più agevolmente la loro applicazione all'analisi ed i risultati ottenuti.
Su suggerimento di chi ha fornito i dati, le analisi cominciano considerando solo i probe set di maggior interesse, contenuti nel dataset ridotto di circa seicento unità. Vengono quindi prodotte le curve di Andrews relative ad esempio ad una delle quattro condizioni per avere un idea visiva di come i geni si comportano durante la giornata di osservazione (cfr Figura 6).
Quello che maggiormente si nota è la sincronicità delle traiettorie delle funzioni di Andrews calcolate sui dati; nessuno dei probe set, pur differenziandosi anche di molto per valori assunti, sembra seguire un andamento diverso dagli altri.

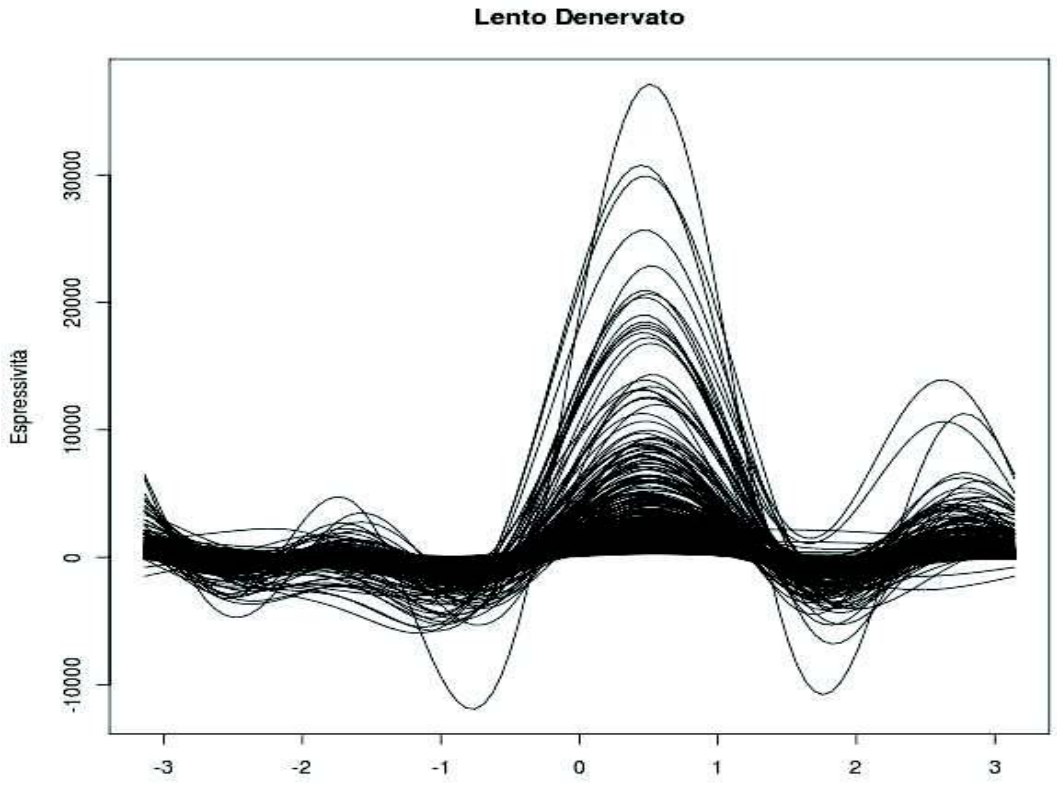

Figura 6: Curve di Andrews per il dataset ridotto, condizione lenta denervata

Questo risultato sembra plausibile visto che si stanno analizzando i livelli di espressione di geni coinvolti nello stesso ciclo biologico. Ma, a questo punto, sorge istintivo chiedersi se la suddivisione effettuata dai biologi in geni che “oscillano” e “non oscillano”, sia statisticamente sensata. Per controllarlo inizialmente a livello intuitivo, si è disegnato lo stesso tipo di grafico appena presentato sul dataset completo, cercando sostanziali differenze tra i due (cfr Figura 7).

Il risultato è un diagramma ovviamente più denso di linee, visto il consistente aumento del numero di geni considerati, soprattutto nella parte bassa (dove compaiono un gran numero di probe set che non superano almeno in una osservazione il livello di soglia di espressione di 100) e che estende il campo di variazione fino a tre volte tanto il precedente; non vengono evidenziate però sostanziali differenze con il precedente per quanto riguarda le traiettorie seguite dalle curve, né è immediato individuare dei gruppi distinti nello stesso.

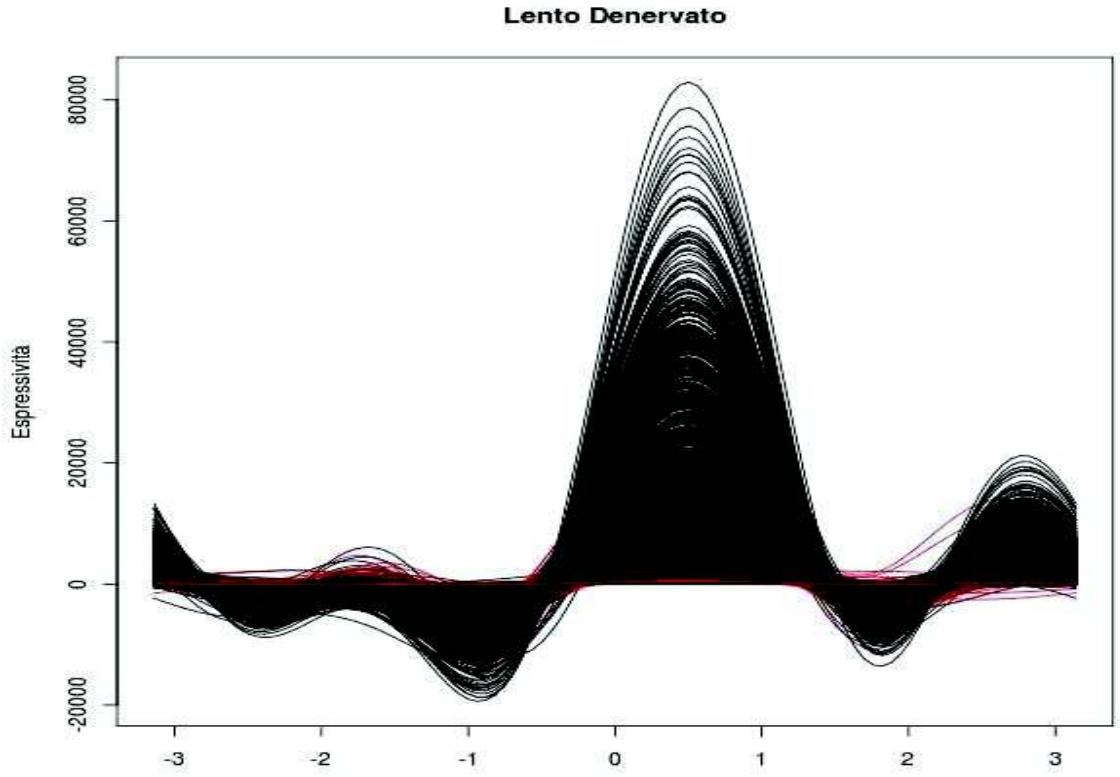

Figura 7: Curve di Andrews, condizione lenta denervata. In rosso i probe-set presenti solo nel dataset ridotto, in nero i restanti del dataset completo

Si decide comunque di esplorare questa non distinguibilità tra gli inclusi e gli esclusi dal set di dati ridotto attraverso alcuni approcci statistici più rigorosi, dapprima tramite una analisi dei gruppi di tipo gerarchico aggregativo e, successivamente, implementando un semplice modello logistico per la probabilità di appartenenza ad uno dei due gruppi e valutandone le capacità predittive.

Il dendrogramma che descrive le aggregazioni successive, nel quale vengono evidenziati con colori diversi i due gruppi, mostra come i raggruppamenti non lascino intendere una maggiore vicinanza tra elementi definiti “oscillanti” rispetto agli altri, qualunque sia il tipo di distanza considerata.

Anche il modello logistico che considera come variabile risposta il gruppo di appartenenza e come covariate sia le 24 misurazioni complessive che le sole 6 misurazioni di un esperimento scelto a caso porta a considerazioni analoghe.

Tramite convalida incrociata, il modello stimato porta a stimare la probabilità di errata classificazione attorno al 4%; sembra adattarsi quasi esclusivamente al gruppo più numeroso, tendendo a classificare come tali tutte le unità statistiche (e di conseguenza sbagliando poco in termini assoluti vista l'alta numerosità di tali elementi nei dati) ad esclusione di quelle che si assestano in una fascia di valori di espressone dove troviamo, pur in minoranza rispetto agli altri, la maggior parte dei geni “oscillanti”; qui il modello assegna, erroneamente, una più alta probabilità di appartenenza al dataset ridotto e ne risulta quindi una classificazione addirittura peggiore di quella ottenibile costruendo una regola che assegna tutte le unità al gruppo degli esclusi (il più numeroso) che otterrebbe una analoga misura d'errore pari circa al 3%, cioè la frequenza relativa di geni di tale gruppo.

Si può pensare però di esplorare meglio creando ad hoc un campione casuale di probe set bilanciato (cioè contenente un numero di “non oscillanti”, scelti a caso, pari a quello degli inclusi nel dataset ridotto) ed, usando la stessa procedura, arrivare ad una frequenza di errata classificazione tramite convalida incrociata più facilmente interpretabile.

La scelta del campione viene ripetuta più volte, per essere certi che il caso non abbia portato ad avere un campione dal nostro punto di vista sfortunato (ad esempio che presenta gruppi perfettamente separati) ed il valore della funzione di perdita considerata viene quindi ottenuto come media.
La stima della probabilità di errata classificazione risulta questa volta essere attorno al 50%, come quella limite di una regola che sceglie a caso, e viene deciso di conseguenza che i due gruppi non possono essere separati.

3.2 Analisi dei gruppi : le variazioni temporali

Pur riprendendo le analisi con l'insieme completo, può essere sensato chiedersi se esistano dei gruppi di geni, diversi da quello precedentemente menzionato, che evidenzino caratteristiche particolari.
Non avendo informazioni a priori possiamo usare in questa fase solo tecniche di apprendimento non supervisionate e quindi tecniche di analisi dei gruppi.
Su un numero di elementi così elevato, i metodi gerarchici potrebbero consegnarci informazioni non facili da gestire e quindi, pur dovendo fare una forte assunzione iniziale sull'esistenza dei gruppi e sul loro numero, si sceglie di utilizzare il metodo delle k-medie per dividere i probe set in un numero esiguo (da quattro a sei) di cluster.
I cluster ottenuti utilizzando cinque centroidi sembrano particolarmente sensati; non si ottengono evidenti sovrapposizioni e le curve di Andrews risultano ben stratificate. Viene qui riportato come esempio il medesimo grafico mostrato in precedenza, i geni nella condizione lenta denervata, colorando in modo diverso le funzioni corrispondenti a geni classificati in gruppi differenti (cfr Figura 8).

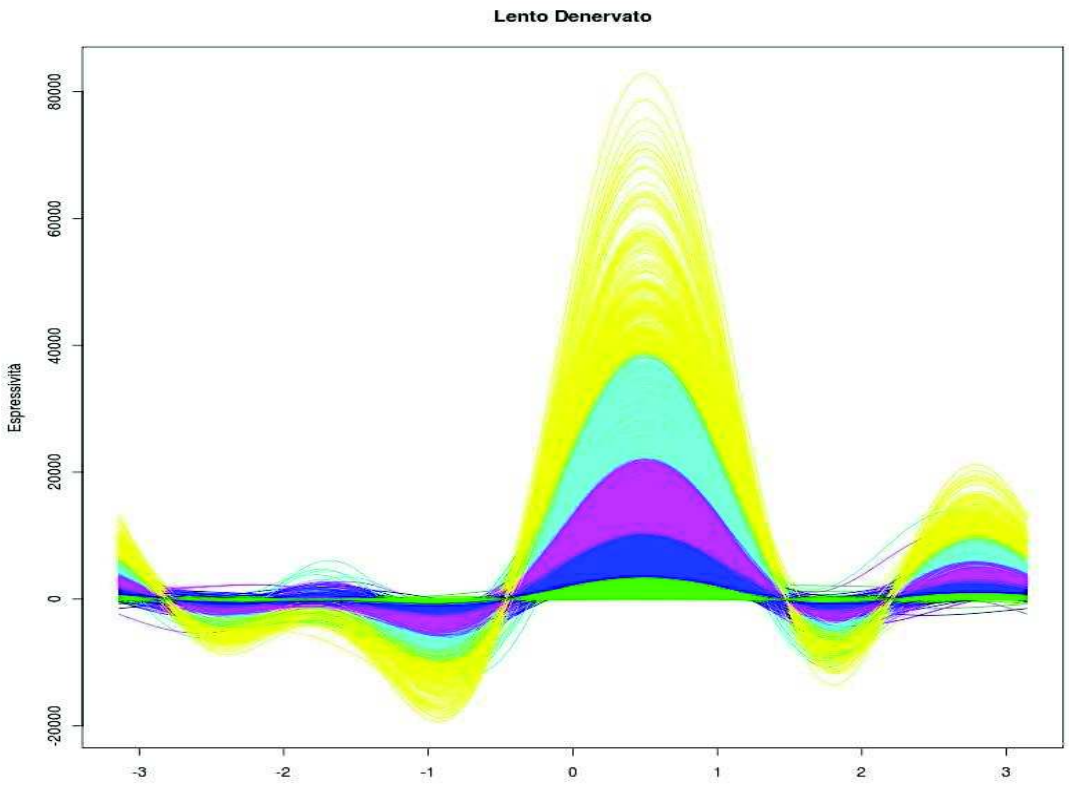

Figura 8: Condizione Lenta Denervata, classificazione con il metodo delle k-medie

Un risultato del genere però non è molto significativo in sé, perchè trattandosi il metodo delle k-medie di un algoritmo di ottimizzazione, arrivare a gruppi apparentemente ben definiti è piuttosto comune. Potremmo quindi interessarci piuttosto al fatto che questi cluster si conservino nelle varie condizioni, ovvero potrebbe avere un significato scoprire se le divisioni effettuate sulla prima condizione sono simili a quelle ottenute con una procedura analoga in una delle altre.

Una prima idea per esplorare la stabilità dei gruppi potrebbe essere, ad esempio, ricavare i cluster analizzando la seconda condizione e, mantenendoli fissi, colorare poi le funzioni di Andrews ottenute nelle quattro condizioni in base a questi gruppi (cfr Figura 9).

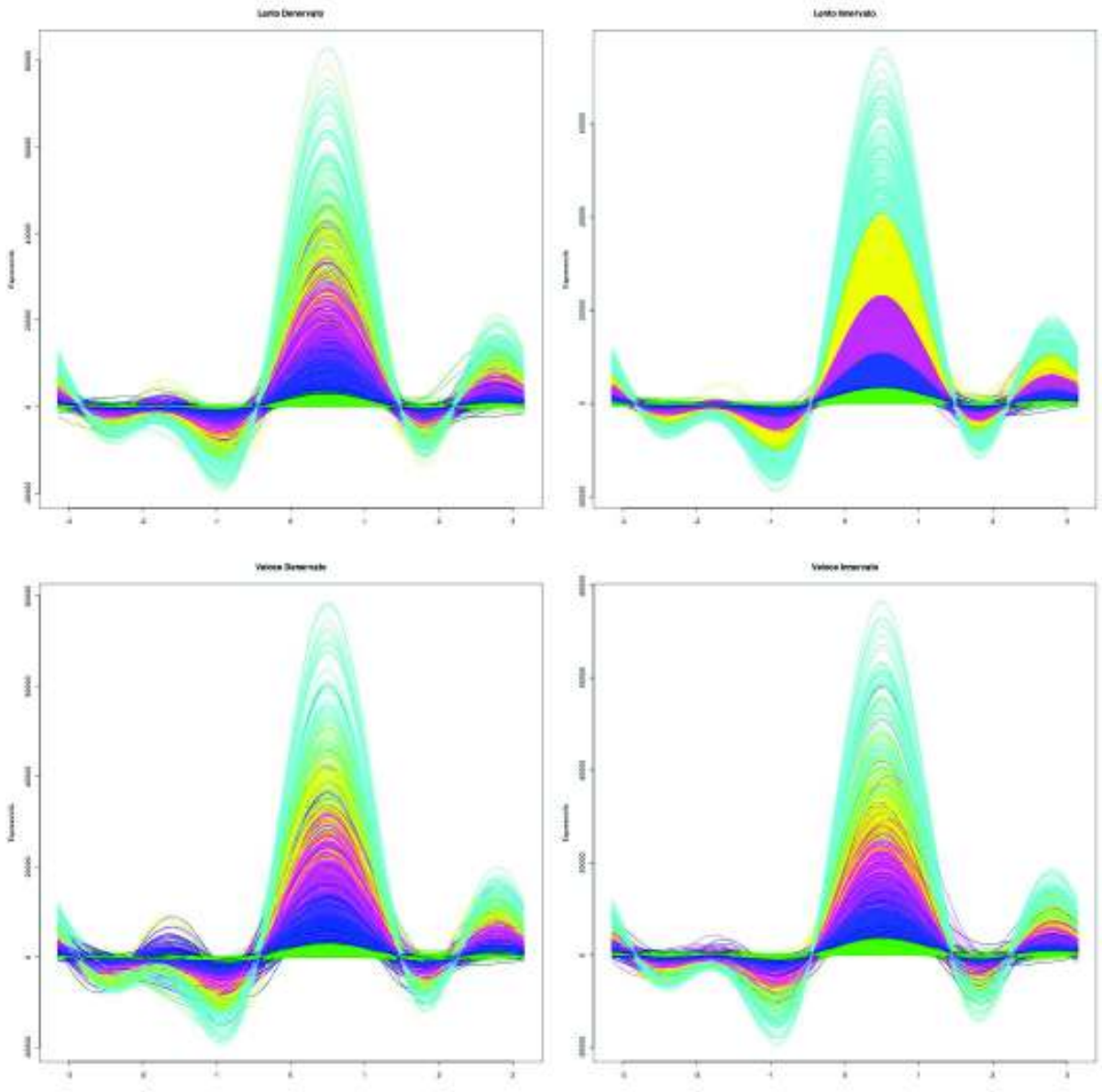

Figura 9: procedura visiva per l'analisi visiva della stabilità dei gruppi; diagrammi di Andrews dei quattro esperimenti colorati in base ai cluster ottenuti sulla seconda condizione

I grafici relativi alle condizioni diverse da quella su cui abbiamo effettuato l'analisi cluster sembrano essere confusi e sono numerose le curve che si sovrappongono; sembra quindi che i gruppi non si conservino. Dobbiamo ricordare però che le funzioni mostrate in questo grafico sono oltre 17000, molte delle quali sovrapposte almeno in parte per via della bassa distanza tra geni simili, ed è quindi lecito chiedersi se non sia solo una piccola parte di queste curve che, cambiando gruppo e risultando fuori dalla zona del loro colore, creano questa sensazione di confusione all'intero grafico.

Si prova quindi a verificare l'ipotesi di stabilità calcolando la segmentazione separatamente per ogni condizione e confrontando i risultati attraverso la costruzione di una matrice di confusione da cui stimare la probabilità di diversa classificazione come la somma degli elementi fuori dalla diagonale diviso per il totale dei probe set.

Il problema sta nel fatto che non conoscendo a priori un elemento rappresentativo per ogni classe e non potendo controllare lo spostamento dei centroidi durante l'esecuzione dell'algoritmo delle k-medie, ad ogni esecuzione l'algoritmo assegnerà delle etichette praticamente casuali ad ognuno dei gruppi, dipendentemente dalla posizione iniziale nello spazio p-dimensionale dei cinque punti iniziali e dagli elementi a loro vicini, il che porta a non essere in grado di individuare direttamente lo stesso gruppo in due classificazioni diverse e quindi a non poter stimare la probabilità di diversa classificazione.

Osservando il grafico in Figura 9 si nota però che, almeno approssimativamente, le stratificazioni successive restano individuate dallo stesso colore predominante; vengono rinominati quindi i cluster in modo che nella matrice di confusione risultino sulla diagonale gli incroci con la maggiore frequenza assoluta, assumendo cioè che la maggior parte dei geni rimanga classificata allo stesso modo.

A conferma di ciò si arriva inoltre ad avere gli stessi abbinamenti tra etichette e gruppi ordinando i gruppi in base all'osservazione dei grafici per ogni condizione, nominandoli cioè in ordine crescente partendo dal gruppo le cui funzioni risultano essere più schiacciate sull'asse orizzontale e via via a salire.

Un esempio di matrice ottenuta in questo modo è quella relativa al confronto tra l'analisi cluster ottenuta sulla prima condizione e quella sulla terza condizione (cfr Tabella 1).

Cluster sulla Prima Condizione, gruppo \ Cluster sulla Terza Condizione, gruppo	1	2	3	4	5
5	98	0	10	43	3
4	0	485	2	65	109
3	0	1	13827	0	319
2	16	45	1	168	2
1	1	92	304	1	1721

Tabella 1: Matrice di confusione per il confronto delle classificazioni ottenute rispettivamente sulla prima e sulla terza condizione. (Condizione rapida denervata ed innervata)

Vengono calcolate quindi le percentuali di diversa classificazione come la somma degli elementi fuori dalla diagonale principale in rapporto al numero totale di geni e si scopre che sorprendentemente sono molto basse; in particolare mediando per tutti gli abbinamenti possibili tra le quattro classificazioni, risultano sotto al 10%. Si modifica quindi l'idea che era naturale farsi dalla sola visione del grafico presentato in precedenza, dove l'ipotesi di stabilità appariva quantomeno forzata.

L'aver manualmente permutato le etichette potrebbe però, in caso non fosse stato lecito farlo, aver indotto una sottostima di questa probabilità e quindi, per avere un'ulteriore conferma, si è pensato di controllare al posto dei singoli geni, le combinazioni di due elementi.

Intuitivamente l'idea è di contare il numero di volte in cui due geni sono classificati nello stesso gruppo (o in gruppi diversi) in due cluster ottenuti su condizioni diverse e, dividendo per il totale di possibili combinazioni di probe-

set, ottenere quindi una percentuale di diversa classificazione indipendentemente da come le si assegnano le etichette ai gruppi. Viene calcolato, cioè, l'indice di Rand che, fortunatamente, possiede una formulazione equivalente calcolabile partendo dalle stesse matrici di confusione illustrate poc'anzi, senza dover controllare iterativamente tutte le coppie di geni (si veda *S.Zani, 2000*, per ua descrizione approfondita dell'indice e per una dimostrazione formale di entrambe le formulazioni).
Comunque si effettuino i calcoli viene ottenuto per tutte le coppie considerate un valore di tale indice attorno allo 0.9 e, ricordando che quest'indice assume valori tra 0 (perfetto disaccordo tra i cluster) ed 1 (cluster perfettamente coincidenti), si conclude quindi che i gruppi ottenuti sono decisamente stabili al variare degli esperimenti.

3.3 Analisi dei gruppi : Variazioni tra gli esperimenti

È dunque solo un numero relativamente ristretto (ma comunque presente) di geni che si sposta di gruppo se analizziamo per ogni condizione il comportamento durante la giornata, cioè solo alcuni sembrano rispondere in modo diverso alle diverse condizioni in cui vengono effettuate le misurazioni di espressione.
Il passo successivo dell'analisi è, di conseguenza, osservare il comportamento dei dati non più in una sola condizione facendo variare le ore, ma fissando il tempo e confrontando i valori di espressione osservati nei quattro esperimenti, cercando di individuare pattern particolari.
Ancora una volta ci vengono in aiuto i diagrammi di Andrews e, per iniziare, viene presentato a titolo esemplificativo il grafico relativo alle funzioni calcolate sull'ora dodici (cfr Figura 10) .

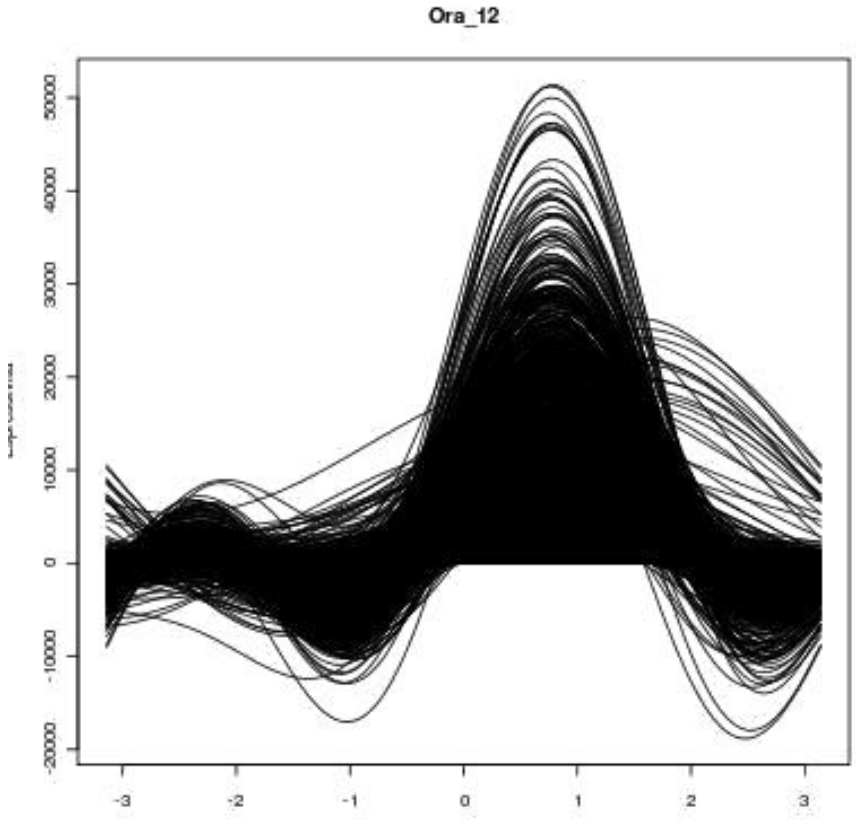

Figura 10: Diagramma di Andrews relativo alle quattro misurazioni (in condizioni diverse) effettuate all'ora dodici

Pur essendo discretamente confuso a causa del numero elevato di funzioni disegnate, una rapida osservazione ci porta ad osservare che nella parte destra si notano alcune funzioni la cui traiettoria sembra deviare da quella comunemente seguita dalla maggior parte dei geni.

Alcune sembrano avere un comportamento addirittura opposto nella parte iniziale del diagramma e per altre il picco più alto si trova leggermente spostato rispetto a quello tracciato dalla maggioranza dei probe-set.

Vengono quindi osservati i grafici relativi alle altre ore di misurazione e pur non presentandoli si rende noto che si possono applicare ad essi le stesse considerazioni appena fatte.

Ma sono sempre gli stessi geni che si comportano in maniera anomala ad ogni ora? Se così fosse avremmo individuato un gruppo di probe set che risponde uniformemente in modo diverso alle varie condizioni in cui vengono effettuate le misurazioni.

Si cerca di conseguenza di individuarli sfruttando ancora una volta tecniche di analisi dei gruppi. Purtroppo la grande variabilità nei valori assunti fa si che utilizzando semplicemente una prima suddivisione tramite i metodi non gerarchici essi non risultino facilmente separabili dagli altri, neppure con un numero elevato di gruppi.

Viene presentato il risultato anche di questa fase di classificazione poiché, al contrario di quanto accadeva nella parte precedente dell'indagine, stavolta i grafici ci indicano con ragionevole certezza che i gruppi individuati utilizzando un ora di riferimento si mantengono anche per le altre (cfr Figura 11).

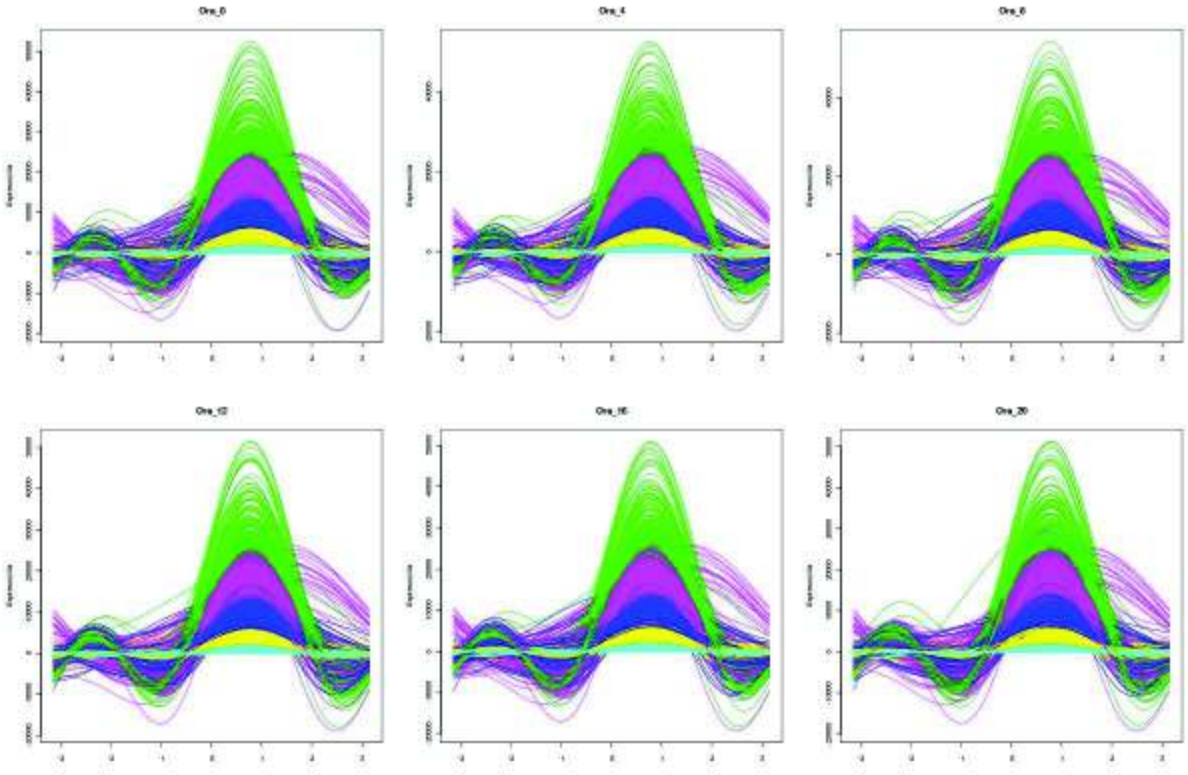

Figura 11: Curve di Andrews per i valori di espressione misurati ad ogni orario con evidenziati i risultati di una cluster analysis ottenuta sui livelli misurati nella prima condizione

Questo induce un certo grado di fiducia che anche gli anomali siano sempre gli stessi.

Si procede individuando i gruppi di appartenenza di quelle funzioni che vengono giudicate comportarsi diversamente dal resto, ed effettuo su di essi un analisi cluster, gerarchica stavolta, cercando di individuare quali traiettorie vengono ad essere aggregate dopo rispetto alle altre.

Vengono così identificati principalmente tre gruppi di geni, uno che comprende la quasi totalità dei probe set e due molto più esigui in dimensioni, rispettivamente di 6 e 17 geni, che si uniscono rapidamente tra loro ma in un certo senso faticano ad associarsi al gruppo più numeroso.
Colorare quindi questi 23 valori in un modo diverso rispetto al resto nei grafici presentati poc'anzi permette di vedere che essi corrispondono proprio a quelle funzioni che erano state individuate come anomale e che esse si presentano circa allo stesso modo qualsiasi sia l'ora di riferimento (cfr Figura 12).

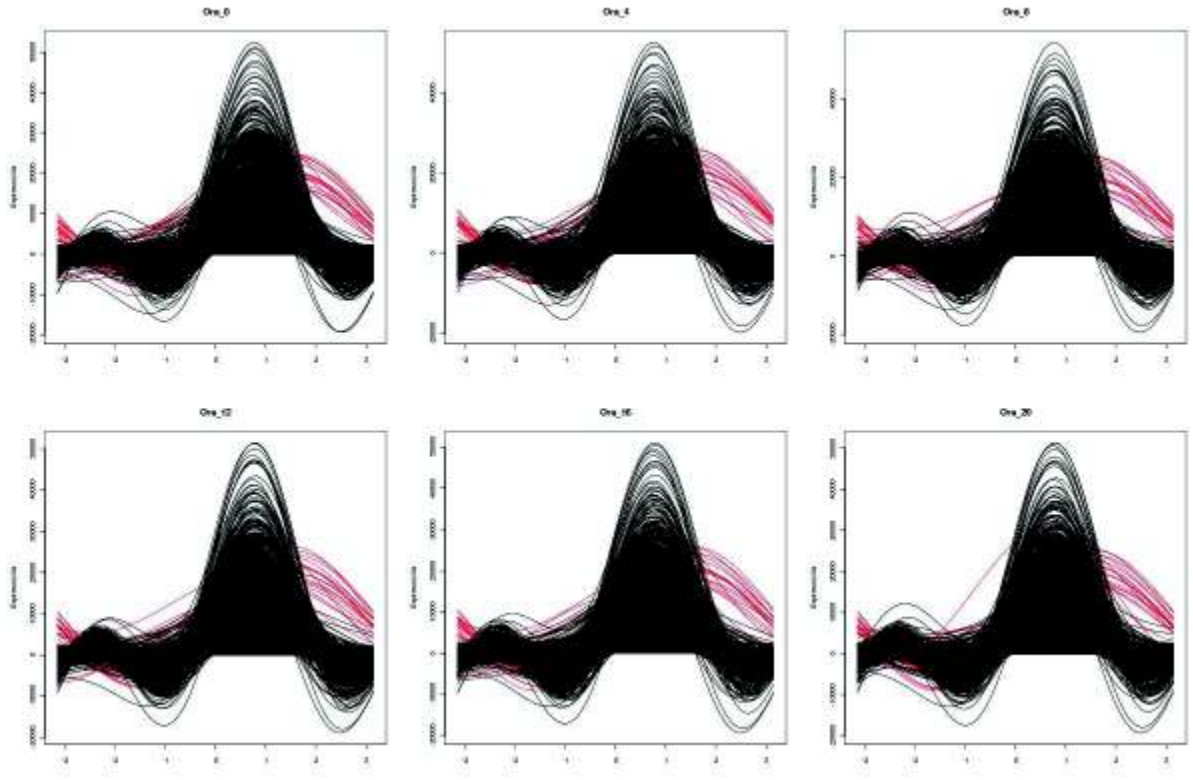

Figura 12: Curve di Andrews per i valori di espressione misurati ad ogni orario con evidenziati i probe-set ritenuti anomali

Avendo quindi osservato che l'ora della misurazione non influisce pesantemente, è possibile avere un'ulteriore riprova della stranezza di questi geni evidenziandoli in un diagramma di Andrews complessivo su tutte le ventiquattro osservazioni. Anche in questo caso è evidente che, fatto salvo qualche valore più intenso, la maggior parte delle traiettorie si muove in modo sincrono, mentre per i ventitre geni definiti anomali (ed ancora una volta evidenziati in rosso) si nota facilmente più d'una inversione di tendenza e,

come già osservato in precedenza, il picco centrale meno elevato e leggermente sfasato (cfr Figura 13).

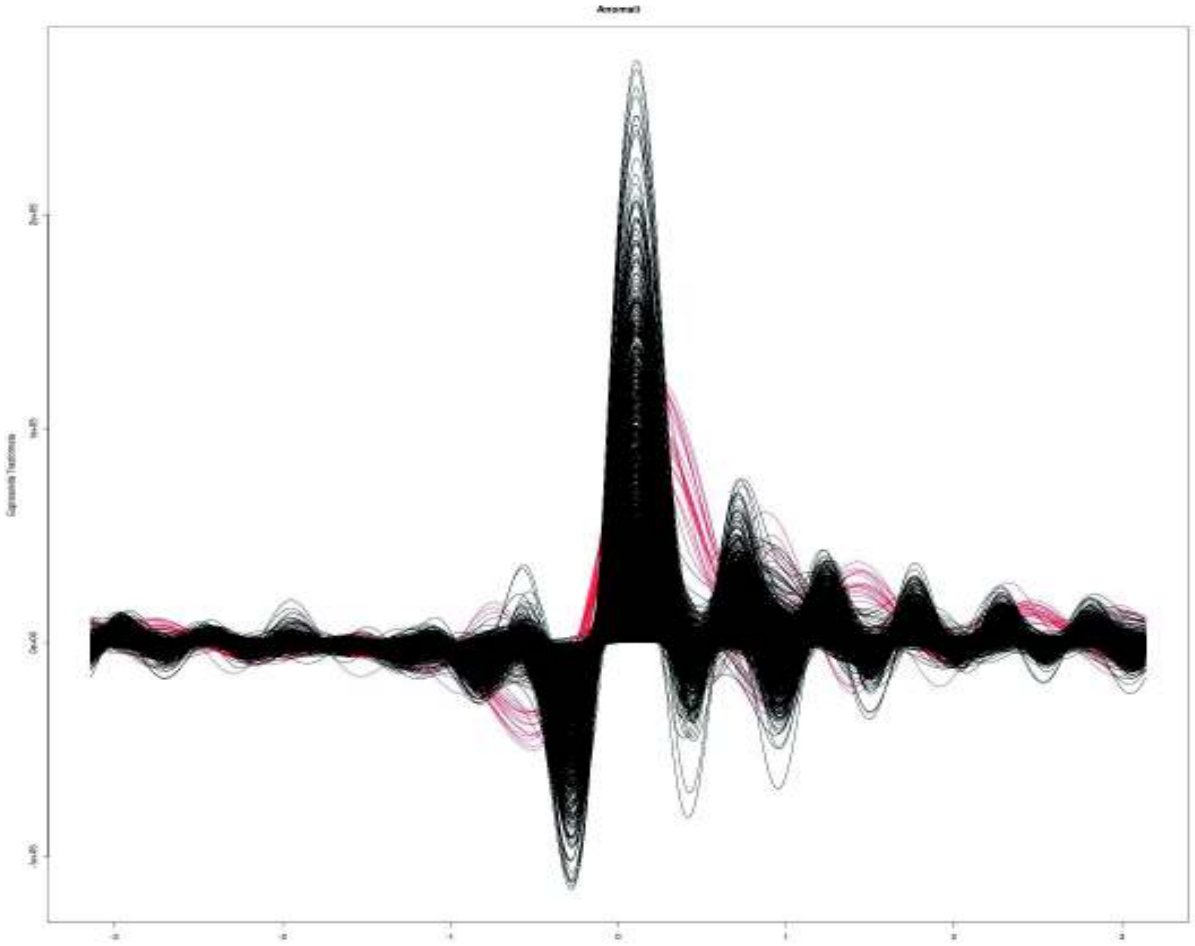

Figura 13: Curve di Andrews per tutti i ventiquattro valori di espressione misurati, sei valori giornalieri in ognuna delle quattro condizioni

Concludiamo quindi che varrebbe la pena concentrare l'attenzione su questo ristretto gruppo di geni, non solo dal punto di vista statistico, per indagare le cause della loro anormalità.

3.4 Il Gene Medio

Ora facciamo però un passo indietro. Tutti i grafici mostrati sono fitti di linee vista l'alta numerosità dei dati. Abbiamo visto in precedenza come questo ci porti talvolta a conclusioni fuorvianti come nel caso dell'analisi dei gruppi sulle variazioni orarie all'interno di una singola condizione.
Sempre in quell'ambito si notava inoltre, dalla matrice di confusione presentata come esempio, che la stragrande maggioranza dei probe set assume valori di espressione molto bassi: il gruppo con valori più contenuti raccoglieva infatti oltre tredicimila geni su diciassettemila.
Lo stesso dataset ristretto fornitoci inizialmente cercava di limitare questo fenomeno introducendo una soglia minima all'espressione e avevamo notato come riducesse a circa il 3% la grandezza del campione.
Pare quindi sensato pensare che una grossa percentuale dei probe set in analisi si comporti in maniera simile, assumendo valori contenuti e oscillando in valore assoluto (qui sta proprio la grossa differenza con i criteri di selezione del dataset ridotto) in modo altrettanto contenuto.
Per selezionare ed accorpare in qualche modo questi geni ancora una volta verranno in aiuto i grafici di Andrews, non tanto con le loro qualità di visualizzazione, stavolta, ma grazie alle loro proprietà statistiche.
Lo stesso Andrews, infatti, nell'articolo in cui li presenta, ci fornisce delle indicazioni su come costruire per ogni punto di una curva un intervallo di confidenza basato sulla variabilità dei dati presi in considerazione e, spostando l'ottica più sul generale, una banda di confidenza per l'intera funzione che, pur non essendo altro che un'approssimazione da non confondere con un vero intervallo di confidenza per la curva nel suo insieme, fornisce un'idea di quali funzioni non siano statisticamente diverse tra loro nel complesso.
Si è cercato quindi di identificare una sorta di gene medio, ovvero la media sulle quattro condizioni della media dei livelli misurati ad ogni rilevazione.

Individuato quindi questo gene, nel grafico di Andrews sono state tracciate le relative bande di confidenza (si rimanda alla parte precedente per la trattazione formale dell'argomento) per testare l'ipotesi di uguaglianza complessiva degli altri geni con il gene medio (cfr Figura 14).

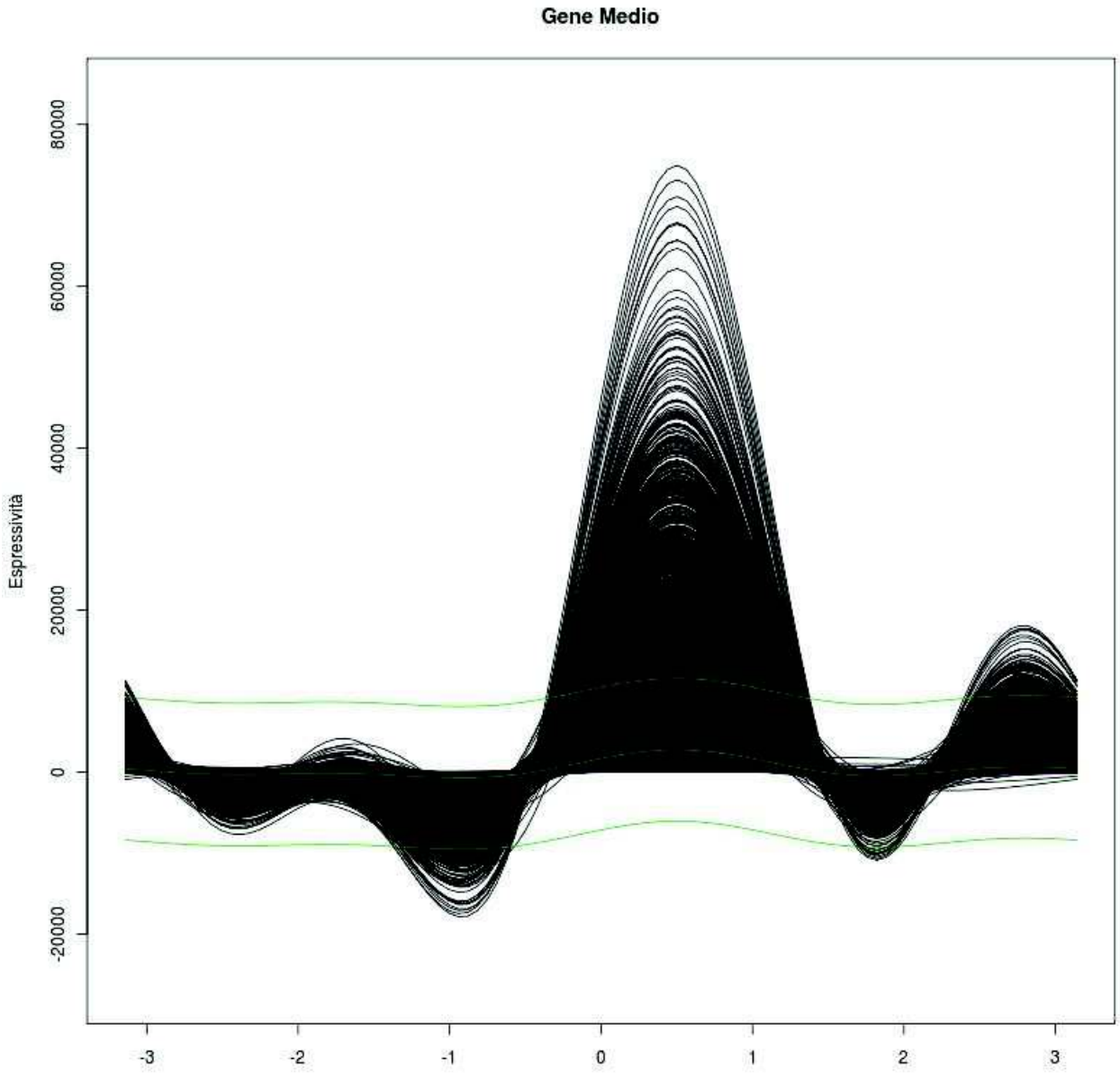

Figura 14: curve di Andrews e bande di confidenza per il test di uguaglianza complessiva al gene medio

Si afferma quindi che tutti quei geni che non possiedono almeno un valore della loro funzione al di fuori da queste bande non si comportano in maniera

significativamente differente dal 'gene medio'. È possibile quindi creare un nuovo insieme di dati ristretto, che risulta essere formato da circa ottocento unità, con tutti i geni che invece sono “diversi” dal gene medio.

È inoltre sicuramente interessante notare come siano molto poche le corrispondenze con i probe set definiti come oscillanti all'inizio: solo una trentina di geni sono presenti in entrambi gli insiemi, mentre tutti i geni classificati come anomali al punto precedente sono presenti anche in quest'ultimo.

Ora i dati risultano essere sicuramente più maneggevoli per il calcolatore e, pur sapendo che molti di questi probe set si differenziano dalla media solo perchè risultano esprimersi con valori molto più elevati e non perchè seguano andamenti effettivamente diversi dagli altri, sembra ragionevole basare ogni ulteriore considerazione su questi dati restringendo ogni futura possibile analisi ad usare solo questo insieme ed il 'gene medio'.

Conclusioni

Il gruppo dei geni che i biologi definisce 'oscillanti' non pare, alla luce delle analisi condotte, distinguibile dal gruppo di geni nel dataset completo che si assestano attorno a valori simili di espressione.

Si è scelto quindi di effettuare l'analisi sulla totalità dei 17313 probe set, arrivando a concludere che il modificarsi nella giornata dell'espressione presenta un'evidente sincronicità tra i geni, dovuta probabilmente all'appartenenza degli stessi geni al gruppo che regola il medesimo ciclo biologico.

Le quattro condizioni in esame inoltre non contibuiscono a modificarne il comportamento in maniera significativa, se non per una manciata di essi. Solo per una piccola percentuale di geni infatti, in risposta ad una condizione, il livello di espressione si assesta su valori sistematicamente più alti o più bassi ma, risultato di particolare interesse, è esclusivamente per una ristretta minoranza che, al passare delle ore, questa varizione è chiaramente osservabile dal confronto di più esperimenti. Entrando quindi più nel dettaglio, tramite alcune tecniche di classificazione statistica, sono stati individuati questi geni, ventitre, e si è scoperto che per essi, con riferimento ai diagrammi di Andrews, la traiettoria tracciata dalla funzione relativa alla risposta nelle quattro condizioni, è costantemente differente da quelle tracciate dagli altri geni, qualsiasi sia l'ora di riferimento fissata. Essi vengono quindi classificati come anomali e varrebbe forse la pena di analizzare in dettaglio, dal punto di vista biologico più che statistico, le ragioni che li portano a rispondere in maniera così stabilmente differente in tutte le condizioni indotte.

Infine, seguendo le intenzioni di chi ha fornito i dati ma utilizzando diversi criteri, è stato individuato un dataset ristretto contenente un numero minore di probe set, per evitare l'insorgenza di grossi problemi computazionali nelle analisi. Sono stati sintetizzati in un cosiddetto gene medio, costruito appositamente, tutti quei geni che possiedono valori di espressività molto bassi, mediamente attorno al 100, ed oscillano in maniera lieve al passare

delle ore. Tali geni rappresentano circa il 95% del campione in analisi e di conseguenza il dataset formato dal gene medio e dai restanti, non accomunabili ad esso, risulta essere di circa 800 unità.

È inoltre notevole che nel sottocampione così costruito siano presenti solo circa 30 geni dei 657 geni che i biologi avevano separato per formare il *dataset ristretto*, avvalorando l'ipotesi che fosse una divisione poco sensata dal punto di vista statistico, mentre vi si presentano tutti i 23 geni individuati come anomali.

Appendice: Codice R

A.1 Trasformata di Andrews

```
f_xi <- function(xi,evalPoints){

k <- length(xi)

# y è il vettore con l'ascissa di ogni punto in evalPoints,
# ytemp è di appoggio per creare ogni punto di y

   y <- vector()
   ytemp <- 0

# creo per ogni punto in evalPoints..
   for(t in evalPoints){
      ytemp <- xi[1]/sqrt(2)

# ...il valore dell'ascissa tale che
# f_xi(t) = x1/sqrt(2) + x2sin(t) + x3cos(t)+ x4sin(2t) + ...

      for(p in 2:k){
            if (p %% 2 == 0) { ytemp <- ytemp + xi[p]*sin((p %/% 2)*t)
                  }else ytemp <- ytemp + xi[p]*cos((p %/% 2)*t)
      }
      y <- c(y,ytemp)
   }
# restituisco il vettore y
   return(y)
}
```

A.2 Grafico di Andrews

```
andrewsPlot <- function(x,group=rep(-1,nrow(x)),
nEval=101,labx="",laby="",main_title="Andrews' Plot"){

stopifnot(length(group)==nrow(x))
# controllo che il vettore dei raggruppamenti fornito sia
# in corrispondenza biunivoca
# (di default sfrutta nrow(x) quindi nessun problema)

   n=nrow(x)

# sequenza di punti in cui la funzione è valutata
   evalPoints <- seq(-pi, pi, length=nEval)

   Y <- NULL # matrice delle ordinate delle funzioni

   for (i in 1:n) {

# elimina ogni struttura presente nel vettore delle caratteristiche
        xi <- as.numeric(unname(unlist(x[i,])))

# trasformata di andrews, vettore di ordinate per la funzione f_xi
        yi <- f_xi(xi,evalPoints)
```

```
        Y <- cbind(Y, yi)
    }

    # limiti del grafico
    ymin <- min(Y[, 1:n])
    ymax <- max(Y[, 1:n])


   # grafico vuoto a cui aggiungerò le curve
   plot(0, 0, type="n", xlim=c(-pi, pi), ylim=c(ymin,
ymax),main=main_title, xlab=labx, ylab=laby)

   # disegno le curve
   for (i in 1:n) lines(evalPoints, Y[, i],col=(group[i])+2)

}
```

A.3 Grafico di Andrews con bande di confidenza

```
## NB band è un vettore contentente le posizioni (indice) delle unità  per
cui si vuole costruire la banda di confidenza

andrewsPlot_band <- function(x,nEval=101,alpha=0.05,labx="",laby="",
main_title="Andrews' Plot",band=NULL){

stopifnot(length(band)<=10)
# controllo che non si vogliano disegnare più di 10 bande,
# altrimenti risulta troppo confuso

   n=nrow(x)

# x è una matrice di k caratteristiche, di cui ognuna osservata n volte
   k<-ncol(x)

# sequenza di punti in cui la funzione è valutata
   evalPoints <- seq(-pi, pi, length=nEval)
   Y <- NULL # matrice delle ordinate delle funzioni
   upBand <- NULL
   dwBand <- NULL

# varianza delle k caratteristiche
   sigmasq <- vector(length=k)
   for (i in 1:k) sigmasq[i] <- var(x[,i])

# varianza campionaria
   sigma2 <- mean(sigmasq)

# calcolo le f_xi
   for (i in 1:n) {

# elimina ogni struttura presente nel vettore delle caratteristiche
  xi <- as.numeric(unname(unlist(x[i,])))

# trasformata di andrews, vettore di ordinate per la funzione f_xi
        yi <- f_xi(xi,evalPoints)
        Y <- cbind(Y, yi)
    }

# calcolo le bande di confidenza per i geni selezionate
    for (i in band) {
```

```
       up_i <- Y[,i]+(sqrt(qchisq(1-alpha,k)*sigma2*(k+1)/2))
   dw_i <- Y[,i]-(sqrt(qchisq(1-alpha,k)*sigma2*(k+1)/2))
   upBand <- cbind(upBand,up_i)
         dwBand <- cbind(dwBand,dw_i)
    }

    # limiti del grafico
    ymin <- min(min(Y),min(dwBand))
    ymax <- max(max(Y),max(upBand))

    # grafico vuoto a cui aggiungerà² le curve
    plot(0, 0, type="n", xlim=c(-pi, pi), ylim=c(ymin,ymax),
main=main_title, xlab=labx, ylab=laby)

   #disegno le curve
   for (i in 1:n) lines(evalPoints, Y[, i],col=(group[i])+2)

   #disegno le bande per i geni selezionati
   for (i in band) {
        lines(evalPoints, Y[, i],col=i)
        lines(evalPoints, upBand[,i],col=i)
        lines(evalPoints, dwBand[,i],col=i)
   }
}
```

Bibliografia

D. F. Andrews (1972). "Plots of High-Dimensional Data". International Biometric Society 18 (1): pages 125–136.

R. Moustafa, E. Wegman (2002). "On Some Generalization to Parallel Coordinate Plot. Seeing a million, A Data Visualization Workshop", Rain am Lech (nr.), Germany.

R. Moustafa, E. Wegman (2006). "Multivariate continuous data - Parallel Coordinates". In: Unwin, A., Theus M., Hofmann, H. (Eds.), Graphics of Large Datasets: Visualizing a Million, Springer: 143–156.

S. Zani (2000). "Analisi dei dati stitistici vol.2, osservazioni multidimensionali", Giuffrè, Milano.

A. Azzalini, B. Scarpa (2004). "Analisi dei Dati e Data Mining". Springer-Verlag Italia, Milano.