Università degli Studi di Padova
Dipartimento di Scienze Statistiche
Corso di Laurea Triennale in
Statistica per le Tecnologie e le Scienze

Relazione Finale

Comparazione di metodi di clustering per l'analisi di esperimenti di trascrittomica spaziale

Relatore Dott. Andrea Sottosanti
Dipartimento di Medicina

Laureando Acazi Davide
Matricola 2001311

Anno Accademico 2023/2024

Indice

Introduzione

Le scienze omiche sono discipline che si basano su tecnologie di analisi che consentono la produzione di dati in numero molto elevato, utili per l'interpretazione e la descrizione dei vari sistemi biologici analizzati. Sono essenzialmente quattro: la genomica che si occupa di struttura, funzione, contenuto ed evoluzione del DNA, la trascrittomica che si occupa di studiare la codifica, la regolazione e l'espressione dei geni, più banalmente del "dialogo" fra il DNA e le proteine, la proteomica che tratta gli agglomerati di proteine prodotte o alterate di un organismo, la metabolomica che studia le funzioni e i comportamenti che hanno interi organismi, cellule, organi o tessuti quando assimilano una sostanza.

La trascrittomica ha lo scopo di comprendere più dettagliatamente la possibile origine di un processo anomalo nell'organismo e di conseguenza anche nell'individuazione di possibili terapie per la guarigione. Questa scienza studia l'analisi della totalità degli RNA trascritti di un genoma e si riferisce a tecnologie sviluppate che misurano l'espressione genica fino a migliaia di geni.

L'espressione genica è una quantità che rappresenta la trasformazione delle sequenze di DNA, contenute nei geni, in molecole più efficienti come le proteine. Nel processo di trascrizione il DNA viene trascritto in RNA messaggero (mRNA), e nel processo di traduzione l'mRNA viene tradotto in una sequenza di amminoacidi, che formano una proteina.

Quindi, capire come i geni vengono espressi in un tessuto cellulare è fondamentale per la ricerca scientifica e per vari motivi legati alla salute. Questa comprensione aiuta a capire quali geni sono più attivi in un determinato tessuto e come contribuiscono alla sua funzione specifica.

Capitolo 1

Ambito biologico e strumenti utilizzati

1.1 Trascrittomica Spaziale

La trascrittomica spaziale è una classe di tecnologie che fa parte delle scienze omiche. Essa permette di misurare e localizzare l'attività di migliaia di geni in un campione di tessuto, riuscendone a mappare i frammenti nei quali si verifica l'attività. La comprensione del comportamento dei geni nelle diverse aree del tessuto è di principale interesse nell'ambito dei meccanismi biologici chiave, come la comunicazione cellula-cellula o l'interazione tumore-microambiente (Sottosanti & Risso, 2023).

La trascrittomica intende quindi studiare i meccanismi che legano i profili di espressione genica con la funzione delle cellule. Nel processo di trascrizione una porzione di DNA viene trasformata in un filamento di RNA (mRNA) che viene chiamato trascritto: è la prima fase di trasformazione di un gene in una proteina. In ogni trascritto i geni espressi variano e dipendono dall'attività della cellula. La quantità e l'insieme dei trascritti presenti in una cellula viene detto trascrittoma, esso aiuta a comprendere come avviene la produzione delle copie di RNA, per ogni gene, in una data cellula. Il processo di traduzione può avvenire attraverso diverse tecnologie che sono state sviluppate negli ultimi decenni, una delle più adoperate è RNA-seq (RNA-sequencing) che permette: l'isolamento degli RNA totali dalla cellula o dal tessuto analizzati, la conversione in molecole di DNA tramite una reazione ed infine il sequenziamento dello stesso DNA.

La trascrittomica spaziale è in grado di rilevare la distribuzione spaziale di migliaia di geni che sono contenuti in un tessuto, riuscendo anche a definire quanti e quali geni

sono presenti in esso. La collocazione spaziale delle varie cellule è di principale interesse per lo studio e l'analisi per la comprensione dei meccanismi biologici che le caratterizza (Castiglioni, 2022/2023).

1.1.1 Tecnologia Visium

Una delle tecnologie più innovative degli ultimi anni riguardante la trascrittomica spaziale è la *Visium Spatial Gene Expression*, sviluppata da 10x Genomics, essa consente la mappatura dell'espressione genica quasi totalmente a livello cellulare.

Questa procedura si basa sul posizionamento del tessuto selezionato su un area pari a 6.5mm x 6.5mm, letteralmente una griglia di celle dove ognuna delle quali viene detta spot. Dentro ad ogni cella "cade" del frammento di tessuto, un esempio di campione di tessuto e di un raggruppamento degli spot effettuato dalla tecnologia è visibile in 1.1. Idealmente si vorrebbe che in ogni spot cadesse una sola cellula, ma questa tecnologia non lo permette non essendo abbastanza precisa, infatti, per ogni spot si hanno all'incirca 10/15 elementi. Ogni foro ha un identificativo detto *barcode spaziale* che funge da nominativo per identificare dove si colloca spazialmente il materiale trascritto.

Il processo avviene attraverso una reazione chimica che permette alle cellule di rilasciare l'RNA che, attraverso la trascrizione, viene sintetizzato in filamenti di DNA. Prima della fase di sequenziamento le quantità di RNA intrappolate da ogni spot devono essere amplificate essendo relativamente basse per la tecnologia Visium, questa fase di amplificazione potrebbe creare alcuni problemi: i geni con espressione molto bassa potrebbero risultare nulli, mentre quelli con espressione alta potrebbero essere amplificati diversamente a seconda delle loro caratteristiche.

Per sviare questo problema si aggiunge un identificativo univoco detto UMI (*Unique Molecular Identifiers*), cosi facendo, due sequenze di DNA con lo stesso codice non sono due copie indipendenti di RNA ma risultano come una copia della stessa amplificazione (Castiglioni, 2022/2023).

La tecnologia Visium è stata sviluppata da una società di tecnologia biomedica: la 10x Genomics. L'idea iniziale era cercare di mantenere intatta la struttura spaziale dei tessuti campionati mentre si studiavano le informazioni sull'espressione genica. Resa disponibile solo nel 2017, ha fin da subito riscontrato molto interesse nell'ambito di sviluppo in diversi campi scientifici: dalle neuroscienze all'oncologia ma anche nella biologia in generale. Essendo un'innovazione è in continuo e repentino sviluppo con perfezionamenti nella sensibilità, nella capacità di analizzare tessuti di varie grandezze ma anche nella risoluzione spaziale, con un interesse ampliatosi in modo esponenziale

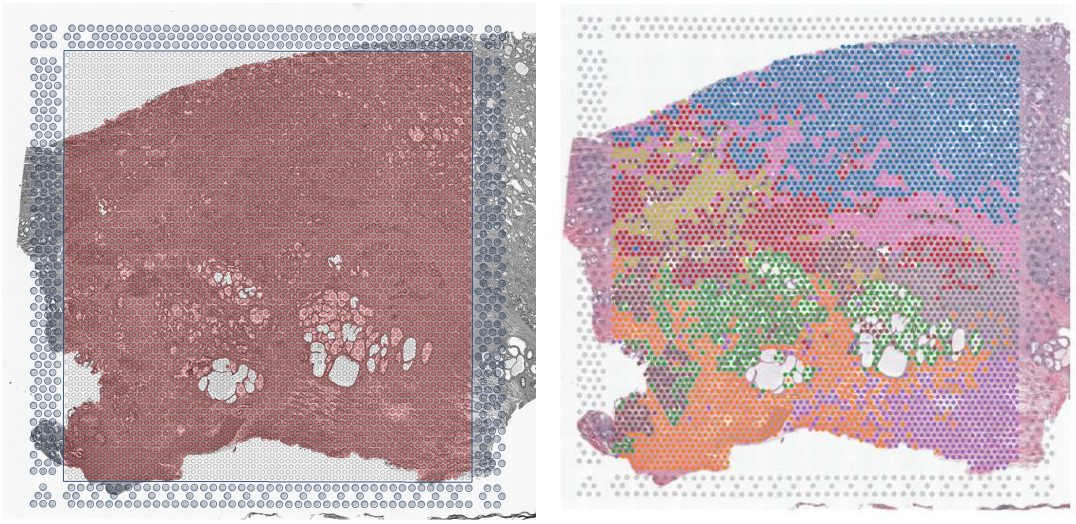

FIGURA 1.1: A sinistra un campione di tessuto cellulare di una prostata analizzato con la tecnologia 10X-Visium. A destra, un esempio di raggruppamento dei geni presente nel sito 10x Genomics (Sommario classificazione).

per le innumerevoli possibilità che ha nella ricerca e nel diagnosticare malattie, vanta anche molti investitori in tutto il mondo.

1.2 Dati utilizzati e obiettivi

Il dataset analizzato è un campione di tessuto umano maschile concesso dalla Indivumed Therapeutics, un'azienda tedesca che studia la biotecnologica incentrata sull'oncologia di precisione, infatti l'azienda vanta di aver realizzato protocolli e procedure tecnologicamente molto avanzate per il campionamento di tessuti biologici di alta qualità.

Nel tessuto selezionato è stato individuato un adenocarcinoma: un tumore maligno invasivo, comune negli organi ghiandolari come pancreas, mammelle o come nel nostro caso, nella prostata. Il tessuto è stato sezionato come descritto nel protocollo di *Visium Spatial Gene Expression for FFPE Demonstrated Protocol* che analizza i livelli di mRNA usando delle sonde nell'intero trascrittoma. Il mantenimento del tessuto e dell'alta qualità dell'RNA è di fondamentale importanza per l'interpretazione dei risultati, infatti il processo di conservazione prevede di fissare dei campioni di tessuto nella formalina e incorporati in paraffina, ovvero in della cera. Successivamente una sezione pari a 5µm è stata posizionata nel vetrino Visium e fungerà da input per la processione dei dati.

Si ottiene una matrice formata da 1000 righe in cui sono presenti i geni ordinati per espressione genica e da 4366 colonne in cui sono presenti i barcode degli spot della tecnologia Visium. Una rappresentazione del campione del tessuto preso in esame è

presente nella sezione precedente 1.1, dalla quale si può notare come siano già presenti delle classificazioni del tessuto, in particolare del tumore stesso raffigurato dalla parte superiore colorata in blu (Sommario Classificazione).

Del dataset a disposizione si studia la matrice di geni verso gli spot contenente dei valori già normalizzati in modo appropriato per l'analisi tramite i modelli mistura come si spiega in sezione 3.1, si ha poi una suddivisione in cluster eseguita manualmente dal patologo Dr.Esposito dell'Istituto Oncologico Veneto (IOV), ed infine si hanno le coordinate di ogni spot rilevate dalla tecnologia Visium che permettono di avere una mappa spaziale di dove sono collocati.

Gli obiettivi della seguente tesi si concentrano sullo studio e sull'approfondimento di alcune tipologie di clustering, in particolar modo su come i modelli mistura classificano in modo corretto i vari profili di espressione genica rilevati negli spot ma anche sulla comprensione della capacità degli stessi di identificare diverse tipologie di cellule nei tessuti di trascrittomica spaziale.

Capitolo 2

Metodi

2.1 Introduzione ai modelli mistura

I modelli mistura sono una classe di modelli statistici usati per descrivere dati che provengono da una popolazione composta da diverse sottopopolazioni. L'idea di base è che i dati presi in analisi possano provenire da più di una fonte, ciascuna con una distribuzione di probabilità differente.

Uno dei casi più comuni è il modello mistura Gaussiana (2.1), caratterizzato da diverse assunzioni: i dati provengono da più distribuzioni normali con parametri diversi, ad esempio in figura 2.1 si nota come la distribuzione rappresentata in blu abbia media e varianza differenti rispetto alla distribuzione rappresentata in rosso; le osservazioni che fanno parte di una certa componente sono incorrelate con le osservazioni di una componente diversa, si parla di indipendenza fra distribuzioni, ad esempio un'osservazione ottenuta dalla distribuzione rappresentata in rosso in figura 2.1 non è influenzata dalla distribuzione colorata in blu; la distribuzione mistura complessiva rappresentata in arancione in figura 2.1 è una combinazione lineare delle distribuzioni che la compongono, ognuna con un peso specifico, nella figura 2.1 sottostante la distribuzione mistura arancione è combinazione lineare delle distribuzioni colorate in rosso e blu.

Si considera una popolazione Ω formata da k sottogruppi, quindi si ha che $\Omega = \Omega_1 \cup ... \cup \Omega_k$, le osservazioni di Ω sono estratte casualmente con probabilità rispettivamente uguali a α_1, ..., α_k che si riferiscono al peso di ogni sottopopolazione Ω_j. Ogni unità statistica $y \in \Omega$ appartiene ad un solo cluster Ω_j $(j = 1, ..., k)$. Si è interessati ad una certa quantità x realizzazione della variabile casuale X, che sia eterogenea fra i diversi gruppi ed omogenea all'interno di ogni gruppo, in questo caso allora la variabile aleatoria X che ci interessa, condizionatamente al gruppo di appartenenza avrà una distribuzione di probabilità diversa. Si può assumere che ogni gruppo abbia le distribuzioni di

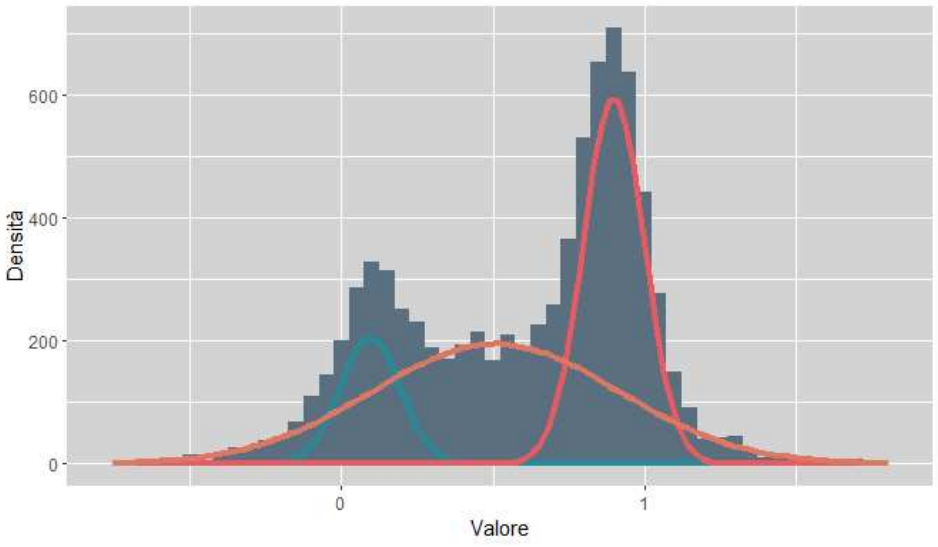

FIGURA 2.1: In arancione un esempio di densità mistura formata dalla somma di due densità Gaussiane in blu e rosso. L'istogramma rappresenta i dati osservati provenienti dalle distribuzioni colorate.

probabilità che appartengono alla stessa famiglia parametrica $f(\mathrm{x};\theta)$ e che il parametro $\theta \in \Theta$ sia variabile gruppo per gruppo. I diversi cluster possono essere identificati da una variabile discreta S che ha valori compresi in $\{1, \ldots, k\}$. Quando viene effettuato un campionamento si osserva la quantità X e la variabile S che indica a quale gruppo fanno parte le unità campionate, di cui la probabilità di appartenenza ad ogni gruppo è α_j con $(j = 1, \ldots, k)$.
Nei casi in cui non si hanno a disposizione il numero di gruppi si osserva soltanto la variabile aleatoria X, allora si crea un modello mistura finita di distribuzioni e la densità marginale di X è:

$$p(x;\psi) = \alpha_1 f(x \mid \theta_1) + ... + \alpha_k f(x \mid \theta_k) \tag{2.1}$$

dove con ψ si indica il vettore dei parametri del modello, le probabilità di appartenenza ai vari cluster α_1,...,α_k sono chiamate pesi della mistura e le $f_{(.)}$ sono le varie densità componenti della mistura.

Tra i vari modelli mistura vi è l'applicazione dei modelli di mistura Gaussiani, cioè con tutte le componenti della mistura sono funzioni di densità che provengono da distribuzioni normali. Nel caso in cui si è in presenza di misture di distribuzioni Gaussiane multivariate si ottiene:

$$f(x;\mu_j,\Sigma_j) = \frac{1}{|2\pi\Sigma_j|^{\frac{1}{2}}} \exp\{(x-\mu_j)'\Sigma_j^{-1}(x-\mu_j)\} \tag{2.2}$$

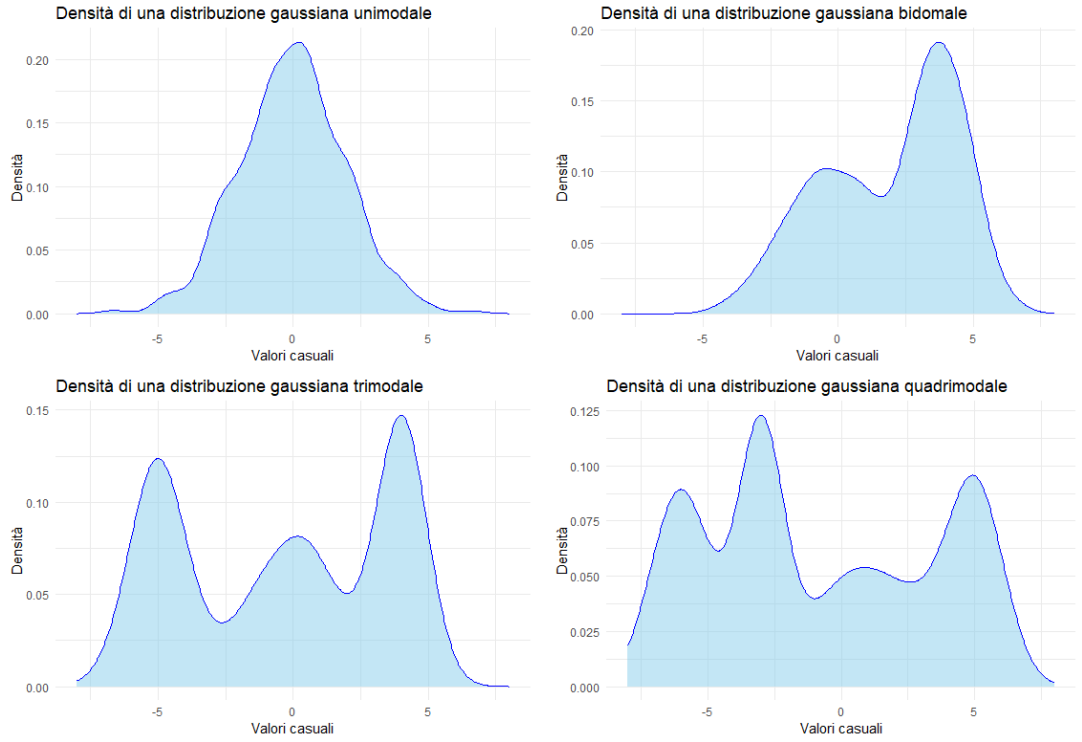

FIGURA 2.2: In alto a sinistra è rappresentata una densità normale unimodale, in alto a destra una densità normale bimodale, in basso a sinistra una densità normale trimodale e in basso a destra una densità normale quadrimodale.

con media il vettore delle medie μ_j e matrice di varianze e covarianze la matrice Σ_j, con $j = (1, ..., k)$. I parametri sono:

$$\psi = (\alpha_1, \ldots, \alpha_{(k-1)}; \mu_1, \ldots, \mu_k; \Sigma_1, \ldots, \Sigma_k) \tag{2.3}$$

si riportano in figura 2.2 degli esempi di densità di distribuzioni normali univariate.

Si possono considerare due approcci applicativi per i modelli mistura: la costruzione di modelli di clustering, che verrà applicato nelle analisi riportate nel seguente elaborato, e la stima di densità non note.

Nel contesto che ci interessa la densità $f_j(\cdot)$ rappresenta la distribuzione di X e α_j la probabilità che l'unità statistica y provenga da questa distribuzione. Si ragiona con la variabile categoriale Z_n con $n = (1, \ldots, N)$ che funge da etichetta di X_n per identificare il gruppo da cui proviene l'osservazione. A questo punto si considera Z_n come un vettore k-dimensionale $Z_n = (Z_{nj},\ j = 1, \ldots, k)$ che assume valore 1 se proviene dal j-esimo gruppo e 0 altrimenti. Si ottiene quindi che Z_n ha una distribuzione multinomiale per tutti i gruppi presenti $j = (1, \ldots, k)$ con probabilità assegnate $\alpha_1, \ldots, \alpha_k$ ovvero:

$$P(Z_n = x_n) = (\alpha_1)^{z_{1n}} (\alpha_2)^{z_{2n}} \ldots (\alpha_k)^{z_{kn}} \tag{2.4}$$

e si avrà una distribuzione multinomiale di dimensione k:

$$Z_n \sim M_k(1, \alpha) \quad \text{con} \quad \alpha = (\alpha_1, \ldots, \alpha_k)'. \tag{2.5}$$

Segue che i dati possono essere classificati nella seguente maniera:

$$\{y_n = (x_n, z_n) : n = 1, \ldots, N\} \tag{2.6}$$

in cui ogni $z_n = (z_{nj} : j = 1, \ldots, k)$ è il vettore k-dimensionale formato da valori 1 se il dato appartiene al gruppo e 0 altrimenti.

2.1.1 Parametrizzazioni possibili nei modelli Gaussiani

In Banfield & Raftery (1993) viene spiegato come sviluppare dei modelli in cui non si assumono matrici di varianza e covarianza uguali, ovvero dove passo dopo passo vengono modificate forma, volume e orientamento mantenendole costanti in modo alternato. Si ricorda che ogni matrice di varianza e covarianza è scomponibile tramite decomposizione spettrale (Reed & Simon, 1980) nel seguente modo:

$$\Sigma_j = \lambda_j D_j A_j D_j^{'} \tag{2.7}$$

dove $\lambda_j = |\Sigma_j|^{\frac{1}{k}}$ determina il volume del j-esimo gruppo, D_j è la matrice ortogonale degli autovettori di Σ_j e ne determina l'orientamento e infine A_j è una matrice diagonale che ha come elementi gli autovalori di Σ_j normalizzati e ne determina la forma, si ha che $|A_j| = 1$.

Per descrivere i modelli che derivano dalla parametrizzazione descritta nella formula 2.7, nella situazione unidimensionale i modelli etichettati con sigla $\boldsymbol{E}$ indicano il caso di elementi omoschedastici, mentre quelli contrassegnati con $\boldsymbol{V}$ denotano elementi eteroschedastici. Nei casi multidimensionali un primo esempio di modelli si basa su elementi di forma sferica, cioè dove $A_j = I$ con $\boldsymbol{I}$ matrice identità.

Il secondo esempio più comune dei modelli si radica sull'assunzione che le matrici di varianza e covarianza Σ_j siano diagonali. Questo indica che nella formula 2.7 le matrici che definiscono l'orientamento D_j sono tutte matrici di permutazione, quindi non sono rilevanti e si ottengono quattro diversi modelli. Un altro caso è quello elissoidale dove si possono ottenere volumi variabili indicati con la sigla $\boldsymbol{V}$, forme costanti citate come $\boldsymbol{E}$, e di nuovo orientamenti variabili $\boldsymbol{V}$, ne consegue che il modello in questione avrà la sigla $\boldsymbol{VEV}$. Nel caso in cui le matrici che determinano forma, orientamento e volume assumano la forma di una matrice identità, nella sigla viene riportata la lettera $\boldsymbol{I}$. In

tabella 2.1 sono riportate le possibili strutture che si possono creare delle matrici di varianze e covarianze per i modelli con diverse parametrizzazioni.

Nella prima colonna sono riportate le sigle delle possibili caratteristiche che possono assumere il volume, la forma e l'orientamento dei vari gruppi. Si possono avere la sigla ***E*** che sta ad indicare che i gruppi formati da quel modello hanno quella particolare caratteristica uguale *(equal)*, oppure ***V*** cioè con quella caratteristica variabile *(variable)*, o in caso ci fosse la ***I*** quella particolare caratteristica non è rilevante dato che quella matrice è una matrice identità *(identity)*. Nella seconda colonna si riporta la parametrizzazione assunta dalla decomposizione spettrale della matrice di varianza e covarianza Σ_j, riportata nella formula 2.7, in quella particolare situazione, nella terza si indica il tipo di distribuzione corrispondente, dalla quarta colonna alla sesta vengono riportate le corrispondenti definizioni delle sigle presenti nella prima colonna, ed infine nella colonna finale dei parametri è riportata la quantità di parametri da stimare in quel preciso caso. In figura 2.3 vengono riportati i 14 modelli appena descritti in tabella 2.1 nel caso di tre gruppi per capire graficamente come sarebbero rappresentati i dati.

Sigla	Modello	Distribuzione	Volume	Forma	Orientamento	N. parametri
E		univariata	uguale			1
V		univariata	variabile			k
EII	λI	sferica	uguale	uguale	NA	$\alpha + 1$
VII	$\lambda_j I$	sferica	variabile	uguale	NA	$\alpha + q$
EEI	λA	diagonale	uguale	uguale	assi coord	$\alpha + q$
VEI	$\lambda_j A$	diagonale	variabile	uguale	assi coord	$\alpha + q + g - 1$
EVI	λA_j	diagonale	uguale	variabile	assi coord	$\alpha + qk - k + 1$
VVI	$\lambda_j A_j$	diagonale	variabile	variabile	assi coord	$\alpha + qk$
EEE	$\lambda DAD'$	elissoidale	uguale	uguale	uguale	$\alpha + \beta$
VEE	$\lambda_j DAD'$	elissoidale	variabile	uguale	uguale	$\alpha + \beta + k - 1$
EVE	$\lambda DA_j D'$	elissoidale	uguale	variabile	uguale	$\alpha + \beta + (k-1)(q-1)$
VVE	$\lambda_j DA_j D'$	elissoidale	variabile	variabile	uguale	$\alpha + \beta + (k-1)q$
EEV	$\lambda D_j AD_j'$	elissoidale	uguale	uguale	variabile	$\alpha - k\beta + (k-1)q$
VEV	$\lambda_j D_j AD_j'$	elissoidale	variabile	uguale	variabile	$\alpha - k\beta + (k-1)(q-1)$
EVV	$\lambda D_j A_j D_j'$	elissoidale	uguale	variabile	variabile	$\alpha - k\beta + (k-1)$
VVV	$\lambda_j D_j A_j D_j'$	elissoidale	variabile	variabile	variabile	$\alpha - k\beta$

TABELLA 2.1: Parametrizzazione delle matrici di varianza e covarianza Σ_j.

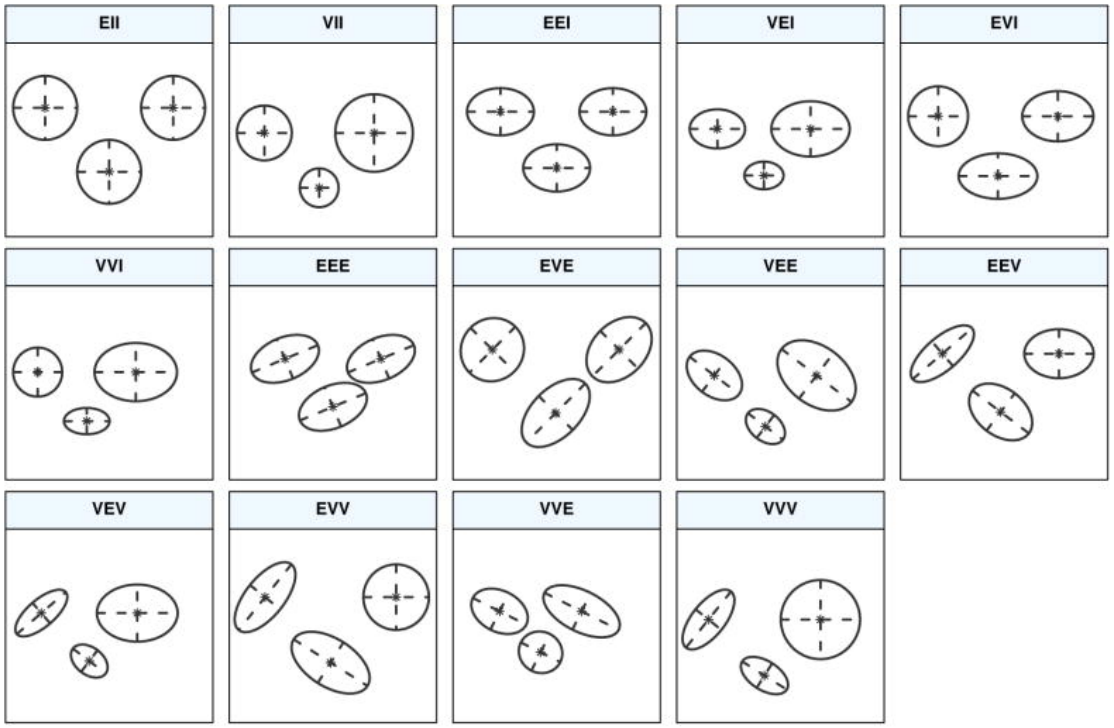

FIGURA 2.3: Grafico dei 14 modelli mistura Gaussiani nel caso di tre gruppi.

2.1.2 Stime di massima verosimiglianza per modelli mistura

Dato un campione $\underline{X} = (x_1, \ldots, x_N)$ di N unità indipendenti dal modello mistura 2.1, dove la funzione di verosimiglianza è:

$$L(\psi; \underline{X}) = \prod_{i=1}^{n} p(\underline{X}; \psi) = \prod_{n=1}^{N} \left[\sum_{j=1}^{k} \alpha_j f_i(x_n; \theta_j) \right] \tag{2.8}$$

dove gli α_j sono le probabilità di appartenere ad ogni gruppo, mentre le $f_i(x_n; \theta_j)$ sono le densità di ogni componente. Per ottenere il vettore di stime di massima verosimiglianza per i modelli mistura bisogna massimizzare la funzione di verosimiglianza 2.8 appena descritta (Johnson et al., 2014).

Se invece si considerano i dati classificati come in 2.6 allora la verosimiglianza corrispondente a $\underline{Y} = (y_1, \ldots, y_N)$ diventa:

$$L(\psi; \underline{X}) = \prod_{n=1}^{N} \prod_{j=1}^{k} \alpha_j{}^{z_{nj}} f_j(x_n; \theta_j)^{z_{nj}} \tag{2.9}$$

dove z_{nj} assume valore 1 se x_n proviene dal gruppo j-esimo e 0 altrimenti.

2.1.3 Algoritmo EM

L'algoritmo *Expectation-Maximization* è un processo iterativo per calcolare la stima di massima verosimiglianza quando non si hanno tutti i dati a disposizione. Ad ogni iterazione questo algoritmo compie due fasi: una fase chiamata *expectation* (fase E) e una fase detta *maximization* (fase M) (Dempster et al., 1977).

Sia allora $X = (x_1, \ldots, x_N)$ un campione di dati generati da una combinazione di diverse distribuzioni Gaussiane. Non sapendo da quale distribuzione è stata generata ogni osservazione, i dati mancanti sono le variabili indicatrici Z_i dove $Z = (Z_1, \ldots, Z_N)$. La funzione di densità congiunta dei dati osservati X e delle variabili latenti Z è:

$$f(X, Z|\theta) = f(Z|\theta)f(X|Z, \theta) \tag{2.10}$$

dove θ è il vettore di parametri del modello composto dalle medie μ_k, le matrici di varianza e covarianza Σ_k e le probabilità di appartenenza ad ogni gruppo α_k.

L'algoritmo EM utilizza la funzione di verosimiglianza marginale dei dati osservati X che si ottiene sommando tutte le possibili variabili latenti Z:

$$L(X|\theta) = \sum_{Z} f(X|Z, \theta)f(z|\theta) \tag{2.11}$$

L'algoritmo permette di trovare le stime di massima verosimiglianza della funzione di verosimiglianza marginale appena descritta (2.11) applicando in modo iterativo i seguenti passi:

E-step Si calcola la quantità $Q(\theta; \theta^{(i)})$ dove:

$$Q(\theta; \theta^{(i)}) = \mathbb{E}_{\theta^{(i)}}\{logL_c(\theta|\underline{X})\} \tag{2.12}$$

M-step A questo punto si calcola un valore di $\theta^{(i+1)} \in \Theta$ che massimizza $Q(\theta; \theta^{(i)})$:

$$Q(\theta^{(i+1)}; \theta^{(i)}) \geq Q(\theta; \theta^{(i)}) \quad \forall\theta \in \Theta \tag{2.13}$$

Queste due fasi vengono eseguite dall'algoritmo in maniera alternata fino a quando la differenza

$$L(\theta^{(i+1)}) - L(\theta^{(i)})$$

risulta minore di una prefissata quantità. Questo algoritmo è utile soprattutto quando ci sono variabili che non sono osservabili in modo diretto o quando si è in presenza di dati incompleti. Si usa nella stima dei modelli mistura nei casi in cui non si conoscono le distribuzioni da cui derivano le osservazioni.

2.1.4 Stima del numero di componenti nei modelli di clustering

Nella costruzione di modelli di clustering mediante i modelli mistura la stima della quantità delle componenti "k" è un importante problema per la loro formulazione. I criteri più conosciuti ed utilizzati sono basati sulla funzione di log-verosimiglianza, il primo tiene conto sia della bontà di adattamento che della complessità del modello ed è *Akaike information criterion* (AIC) (Akaike, 1974) mentre il secondo *Schwartz's Baysian criterion* (BIC) (Schwarz, 1978) ha una penalizzazione più forte rispetto al primo per i modelli con più parametri, quest'ultimo criterio cerca di dare più importanza ai modelli meno complessi.

Si nota che nella formazione dei modelli di clustering la stima del numero "k" di componenti della mistura è di primaria importanza dato che corrisponde esattamente con la stima del numero di gruppi dei dati presi in considerazione. Quindi il problema principale è quello di trovare una buona stima del numero di gruppi che si adatti in modo coeso ai dati osservati.

I criteri di selezione dei modelli in generale hanno una formulazione comune che comprende la funzione di log-verosimiglianza e una penalizzazione che aumenta all'aumentare dei parametri. La forma che assumono è

$$-2L(\hat{\psi}) + C_K \tag{2.14}$$

dove L indica la funzione di log-verosimiglianza, $(\hat{\psi})$ è la stima di massima verosimiglianza in cui viene calcolata la log-verosimiglianza ed infine C_K è la penalità che viene sommata al valore ottenuto precedentemente sulla log-verosimiglianza, che è equivalente al numero di parametri di cui il modello necessita. Valori grandi positivi della formula 2.14 indicano che il modello è formulato malamente ed è di difficile interpretazione, invece valori piccoli positivi o negativi suggeriscono che è stato formulato nella maniera giusta e si adatta bene ai dati.

Come descritto in precedenza uno dei criteri più comuni ed utilizzati per la selezione del modello è l'indice di BIC (*Bayesian Information Criterion*) (Schwarz, 1978), questo indice è utile quando si è nelle condizioni di dover considerare sia la complessità del modello sia la sua bontà di adattarsi in modo coerente ai dati, in più, pondera molto positivamente la semplicità dei modelli e quindi l'interpretazione degli stessi. La sua formulazione è:

$$BIC = -2logL(\hat{\psi}) + qlog(N) \tag{2.15}$$

dove $logL(\hat{\psi})$ è il valore della log-verosimiglianza calcolata nella stima di massima verosimiglianza ottenuta, N è il numero delle unità considerate, infine q è il numero di parametri stimati nel modello. Quando si deve selezionare un modello si preferisce sempre quello con il minor indice BIC.

Un altro indice molto usato e il più conosciuto è il criterio di Akaike (Akaike, 1974), uno dei più importanti criteri di selezione dei modelli. Anche in questo caso sono due le principali componenti di cui tiene conto il criterio e sono la bontà di adattamento del modello ai dati che viene descritto dalla funzione di log-verosimiglianza che calcolata nella stima di massima verosimiglianza indica quanto il modello si adatta bene ai dati osservati, l'altra componente importante è la complessità del modello che viene valutata dal numero di parametri presenti in esso, ovviamente più parametri avrà il modello e più sarà di difficile interpretazione. La formula del criterio di Akaike è

$$AIC = -2logL(\hat{\psi}) + 2q \tag{2.16}$$

dove $logL(\hat{\psi})$ è il valore della log-verosimiglianza calcolata nella stima di massima verosimiglianza e "q" è il numero di parametri liberi nel modello. Anche l'AIC penalizza il modello per il doppio del numero dei parametri liberi ottenuti facendo si che i modelli più semplici siano preferiti ai più complessi a patto che siano uguali per quanto riguarda l'adattamento ai dati. Anche in questo caso quindi si seleziona il modello con l'indice AIC inferiore.

Dopo aver individuato un possibile numero di gruppi con i criteri sopracitati BIC (Schwarz, 1978) o AIC (Akaike, 1974) si può confrontare la stima ottenuta con un altro metodo per la stima delle componenti che si basa sulla verifica d'ipotesi: il test rapporto di verosimiglianza. Il test del rapporto di verosimiglianza essendo basato su una verifica d'ipotesi necessita di un'ipotesi nulla e un'ipotesi alternativa. Dopo aver ottenuto un campione $X_1, \ldots, X_N$ dalla formula 2.1 si formula l'ipotesi seguente

$$H_0 : k = k_0 \quad vs \quad H_1 : k = k_1$$

dove l'ipotesi nulla indica che il modello mistura è formato da k_0 elementi rispetto all'alternativa dove si afferma che il modello è formato da k_1 elementi e si ha che $k_1 \geq k_0$. Questo test si basa sul fatto che la distribuzione asintotica sotto l'ipotesi nulla sia una chi-quadrato ma non essendoci le condizioni di regolarità di cui necessita il test nei modelli mistura non è possibile applicarlo. Con riferimento al lavoro svolto in Lo et al. (2001), dove riferendosi alle distribuzioni Gaussiane univariate con varianza uguale, si è giunti ad un approccio che sotto le dovute ipotesi di regolarità porta ad avere la statistica

del rapporto di verosimiglianza che segue asintoticamente la distribuzione chi-quadrato e rendendo di fatto il test utilizzabile.

2.2 Metodi di clustering per la classificazione cellulare

L'analisi dei gruppi o cluster analysis è una tecnica di analisi dei dati adoperata per raggruppare insiemi di osservazioni o dati simili fra loro in cluster omogenei, utilizzando una serie di metodi statistici multivariati. L'obiettivo principale è quindi quello di individuare vari tipi di gruppi fra i dati senza conoscere a priori eventuali variabili di classificazione o somiglianze fra essi. Si cerca di avere delle osservazioni molto simili all'interno di un gruppo ed invece avere delle differenze sostanziali fra i diversi gruppi ottenuti. Questa metodologia serve a mostrare delle relazioni dei dati che prima non erano visibili e quindi a facilitarne l'analisi e lo studio, per fare ciò si utilizzano degli algoritmi di classificazione che sono basati su diverse misure di similarità o distanza dei dati che consentiranno la formazione di gruppi molto diversi fra loro ma con unità simili al loro interno. Questa tecnica viene utilizzata in svariati rami scientifici e ci sono diversi approcci per effettuare l'analisi dei gruppi, tra cui:

- Metodi gerarchici
- Metodi basati su partizioni
- Metodi basati su modelli

di seguito verranno esposte le varie metodologie di clustering per la classificazione cellulare utilizzate in questo elaborato.

2.2.1 Metodo delle k-means

I metodi basati su partizioni necessitano di un numero di cluster specificato a priori in cui le diverse unità statistiche verranno poi suddivise dal criterio su cui è basato questo metodo. Il più famoso metodo non gerarchico è l'algoritmo k-means.

Questo algoritmo consiste nel suddividere le osservazioni nei k gruppi prescelti. La selezione del gruppo di cui faranno parte le unità statistiche è sostanzialmente dovuta alla distanza fra il punto stesso e il punto medio di ogni cluster, anche detto centroide. Si basa principalmente su tre passi:

1. Le osservazioni vengono suddivise in k gruppi di partenza;

2. L'algoritmo calcola "k" medie, una per ogni gruppo che fungerà da centroide, a questo punto l'algoritmo calcola la distanza (solitamente distanza euclidea) tra l'osservazione e ciascun centroide. L'osservazione a questo punto viene classificata nel gruppo il cui centroide è più vicino ad essa;

3. Si ripete il punto precedente finché non ci sono più spostamenti e nei cluster non avvengono più modifiche.

Si nota che il metodo k-means è molto sensibile al numero di gruppi assegnato inizialmente, quindi tipicamente si esegue l'algoritmo più volte con diversi numeri di centroidi e si seleziona il risultato migliore.

2.2.2 Metodi basati su modelli probabilistici

Come anticipato nel paragrafo precedente un possibile approccio all'analisi cluster può essere effettuato mediante i modelli, in questo caso si tratta della classificazione mediante dei modelli probabilistici, quello che viene utilizzato in questo elaborato è il modello mistura Gaussiana. In questo approccio il problema di scegliere un metodo di classificazione e di determinare un numero adeguato di cluster può tradursi come un problema di scelta di un modello.

Due approcci complementari messi insieme da Fraley & Raftery (2002) che hanno congiunto l'agglomerazione gerarchica basata sulla verosimiglianza di classificazione (Murtagh & Raftery, 1984) e (Banfield & Raftery, 1993) con l'algoritmo EM utilizzato per la stima di massima verosimiglianza dei modelli mistura Gaussiani, dove le matrici di varianza e covarianza sono parametrizzate tramite la scomposizione spettrale (Hand, 2018) e (Celeux & Govaert, 1995). La combinazione dei due approcci funziona bene perché l'agglomerazione gerarchica produce delle partizioni delle osservazioni senza dover specificare il numero di cluster nell'inizializzazione, mentre il secondo passo dell'algoritmo EM funziona bene quando è avviato con accortezza, quindi avendo una stima del numero di gruppi presenti nei dati i due processi si completano a vicenda.

Nell'approccio basato sui modelli mistura si ipotizza che i dati siano generati da diverse distribuzioni Gaussiane ognuna delle quali rappresenta un gruppo dei dati, l'obiettivo è quello di massimizzare la funzione di verosimiglianza rispetto ai parametri della mistura quindi di modellare la distribuzione dei dati. Nell'approccio basato sulla classificazione si cerca di dividere i dati in gruppi differenti in base alla probabilità che ogni punto appartenga a ciascun cluster, questa probabilità si basa sui parametri delle distribuzioni Gaussiane che li compongono.

2.2.3 Analisi delle componenti principali

La *Principal Component Analysis* (PCA) (Bolasco, 2022) è una metodologia statistica che viene utilizzata per la riduzione della dimensionalità dei dati e per permetterne poi l'analisi semplificata. Gli obiettivi di questa procedura sono la riduzione della dimensionalità mantenendo i dati in modo più simile agli originali e l'eliminazione della ridondanza cercando di avere meno variabili che spiegano le stesse informazioni. I principali vantaggi sono la riduzione delle variabili in esame, con conseguente miglioramento dell'interpretabilità, e il filtraggio delle componenti che hanno poca varianza quindi una riduzione del rumore.

Generalmente si usa quando si hanno nel dataset più variabili molto simili fra loro, quindi molto correlate, e se ne vogliono avere una quantità minore che sia in grado di descrive e sintetizzare le altre senza però perdere troppa informazione. L'analisi delle componenti principali ha alcuni semplici passaggi che però rendono molto l'idea di come essa funzioni. Per prima cosa c'è la standardizzazione di tutte le variabili in modo di renderle tutte uguali, con media pari a zero e varianza unitaria.

Il passo successivo è quello di eseguire la decomposizione spettrale della matrice di varianza e covarianza in modo tale da ottenere autovalori e autovettori che saranno degli "indici" rispetto a dove i dati variano maggiormente e serviranno per ottenere le componenti principali tramite una combinazione lineare. Si ottengono delle componenti principali ordinate per importanza, ovvero per varianza totale spiegata da ogni componente. Solitamente si mantengono componenti per un ammontare del 70-80 % di variabilità totale. Le componenti principali ottenute non sono altro che delle combinazioni lineari delle variabili originali che sono incorrelate e hanno varianza massima.

Per la scelta delle componenti principali, oltre a cercare di avere un minimo di varianza totale spiegata di circa il 70-80 %, un altro criterio è quello di mantenere tutte le componenti che hanno gli autovalori uguali o maggiori di 1. Un ultimo criterio può essere quello di visualizzare lo scree-plot dove si mettono nell'asse delle ordinate i valori della varianza spiegata da ogni singolo componente e nell'asse delle ascisse il numero corrispondente alla componente principale. Unendo tutti i punti ottenuti si ottiene una linea continua e spezzata come si può vedere dal grafico 3.5 per ogni nuovo componente, il criterio si basa sul cercare di individuare un "gomito" della linea che starebbe a significare un importante decremento della varianza spiegata che dopo il quale la linea tende a rimanere piatta (Johnson et al., 2014).

Capitolo 3

Risultati

3.1 Dati

Si spiegano ora più dettagliatamente i dati descritti nel primo capitolo 1.2 dove, il Dr. Esposito, riferendosi alla citoarchitettura ovvero all'organizzazione spaziale delle cellule, ha separato il campione di tessuto in 4 gruppi differenti: tumore, fibroblasti, ghiandole e stroma. Rispettivamente si parla di zona malata caratterizzata dal tumore, zona di cellule con funzione connettiva ovvero molecole destinate a sostenere altri tessuti, zona formata da ghiandole ovvero formata da cellule specializzate nello secernere sostanze di diversa composizione chimica ed infine una zona formata da tessuto connettivo fibroso con funzione strutturale (Sund & Kalluri, 2009) e (Kalluri & Zeisberg, 2006).

In Figura 3.1 è riportato il raggruppamento appena descritto con la rispettiva tabella di frequenze assolute e relative percentuali degli spot (3.1), si osserva che il totale delle frequenze percentuali somma ad un valore pari a 0.9992 dato che 5 elementi classificati come vasi sanguigni che erano presenti nel tessuto sono stati esclusi in precedenza perché erano troppo pochi rispetto alle numerosità degli altri quattro gruppi individuati.

I dati in questione sono stati filtrati e normalizzati per ridurne la complessità e per fare in modo di aumentare la qualità delle informazioni. La prima procedura è stata filtrare le colonne, ovvero i geni, dei quali sono stati selezionati i più espressi, questo per consentire un'analisi più precisa in un sottoinsieme più indicativo per la ricerca di raggruppamenti che si sta svolgendo. Inoltre, la riduzione del dataset effettuata rende i dati più puliti e chiari per aiutare le metodologie statistiche, in questo caso il clustering, ad individuare più facilmente delle differenze significative nell'espressione dei geni (Castiglioni, 2022/2023).

La normalizzazione aiuta a diminuire la distorsione dei profili di espressione genica presenti tra i vari spot, in pratica si cerca di pulire i dati dalla distorsione che non è

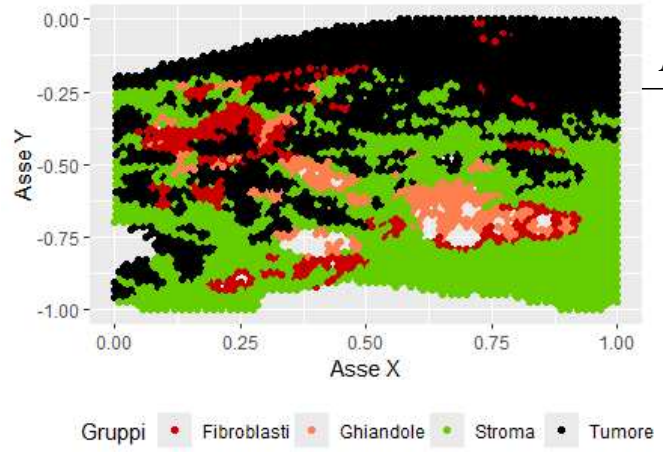

FIGURA 3.1: Figura del tessuto analizzato con i raggruppamenti eseguiti dal patologo Dr. Esposito.

Annotazione	Spot	Percentuale %
Tumore	1946	0.4457
Fibroblasti	371	0.085
Ghiandole	287	0.065
Stroma	1762	0.4035

TABELLA 3.1: Nella tabella viene riportata la suddivisione in gruppi eseguita dal Dr. Esposito. Da sinistra verso destra si hanno: il nominativo del gruppo di appartenenza, la numerosità degli spot presenti e la frequenza percentuale.

dovuta dalle differenze biologiche che si ricercano tra le espressioni geniche presenti negli spot. Se questa procedura non viene eseguita si complica lo studio della relazione fra gli spot, rendendo più evidenti delle differenze che non sono in realtà rilevanti. Si esegue una procedura di selezione dove vengono identificati i geni più significativi, per svolgerla sono presenti due procedimenti: in primo luogo si individua una soglia secondo la quale i geni che la superano sono più rilevanti ed entrano a far parte dell'analisi, mentre quelli inferiori vengono scartati. Nella seconda fase si selezionano i geni con un elevata variabilità di espressione. Quindi a questo punto i geni con una bassa espressione e bassa variabilità vengono esclusi dallo studio, venendo considerati come non informativi per l'analisi che si sta svolgendo.

Per la normalizzazione dei dati in Townes et al. (2019) viene proposto un metodo basato sul modello multinomiale, in cui la selezione dei geni più significativi viene eseguita dal confronto tra due modelli: un modello nullo e un modello saturo. Nel modello nullo si ipotizza che le osservazioni siano distribuite in maniera uniforme fra le varie categorie, ovvero che non ci siano differenze fra i gruppi che si assumono nel modello, questo caso viene preso come riferimento nel confronto con il modello saturo. Nel modello saturo si ipotizza il caso più complesso possibile dove si hanno tanti parametri quante le osservazioni e si ottiene una flessibilità complessiva nel rappresentare i dati. Con i dati in possesso per l'esecuzione delle analisi con il modello nullo si esamina l'abbondanza relativa del gene costante tra gli spot del tessuto, mentre applicando il modello saturo si esamina la variabilità biologica della stima eseguita.

La differenza fra i due modelli viene misurata attraverso la devianza e viene eseguita per ogni gene, si ha che un valore alto di essa indica un'elevata variabilità biologica

quindi è favorito il modello saturo e sfavorito il modello nullo. Dopo aver ottenuto tutti i risultati dei test, i geni vengono ordinati dal valore più alto al più basso, in maniera decrescente, per poi selezionare quelli con valore di devianza più alto.

Il passo successivo è il calcolo dei residui di devianza con segno, che, presi in valore assoluto, forniscono il contributo che ogni spot dà alla devianza totale di un gene, con il segno. La trasformazione utilizzata rende i dati non più interpretabili. Infatti, inizialmente la scala dei dati era basata sul livello di espressione genica mentre ora è basata sulla devianza. Sempre in Townes et al. (2019) si dimostra come valori attorno allo zero dei residui indicano poca espressione o poca variabilità, mentre valori che si discostano dallo zero sono caratterizzati da valori bassi di variabilità ed espressione (Castiglioni, 2022/2023).

3.2 Analisi esplorativa

	Pos. Gene									
	1	2	3	4	5	36	37	38	39	40
Media	-1.2	-1.7	0.2	0	0.2	-0.3	-0.6	0	0	0
Varianza	74.3	44.2	29.8	28.5	23.8	6.2	5.8	5.6	5.5	5.3
Val. Max.	65.2	16.1	17.7	19.8	13.9	22	5.7	7.6	10.7	13.4
Val. Min.	-13.1	-20.4	-10.5	-9.3	-9.8	-5.4	-5.7	-5.1	-5.3	-6

Tabella 3.2: Tabella dei primi e ultimi cinque geni dei quaranta presi in considerazione nella figura dei boxplot.

Nelle precedenti sezioni si è spiegato come i geni siano ordinati per importanza relativamente all'espressione che è stata rilevata nel tessuto. Si procede ora con un'analisi esplorativa dei primi quaranta geni per studiarne le principali caratteristiche e per evidenziarne le differenze maggiori. I boxplot dei geni rispetto ai valori di espressione genica che sono state rilevate in tutti i 4366 spot sono riportati in figura 3.2.

Come si può vedere dal grafico i primi geni sono quelli con le variabilità più elevate infatti i primi tre hanno rispettivamente varianze pari a 74.3 44.2 e 29.8, questo indica come i geni più importanti abbiano una variabilità di espressione maggiore rispetto ai successivi. Il primo gene come si può vedere in tabella 3.2 ha valore massimo pari a 65.2 e minimo pari a -13.1, diverso rispetto al quinto gene che ha valore massimo 13.9 e minimo -9.8 con una varianza pari soltanto a 23.8. Valutando anche le quantità del quarantesimo gene si ha un valore massimo di 13.4 e minimo di -6 che rende l'idea di come la variabilità dell'espressione genica, e quindi l'importanza che questo gene ha

nel tessuto analizzato, diminuisca notevolmente dopo solo 40 geni. Osservando i valori presenti nel primo gene si può vedere come avendo una media di valori pari a -1.2, la scatola del boxplot è bassa rispetto alle altre e la conseguenza è che tanti valori alti positivi diventano outlier. Da questa analisi si evidenzia come i geni più espressi nell'intero tessuto sono sopratutto quelli con valori più variabili.

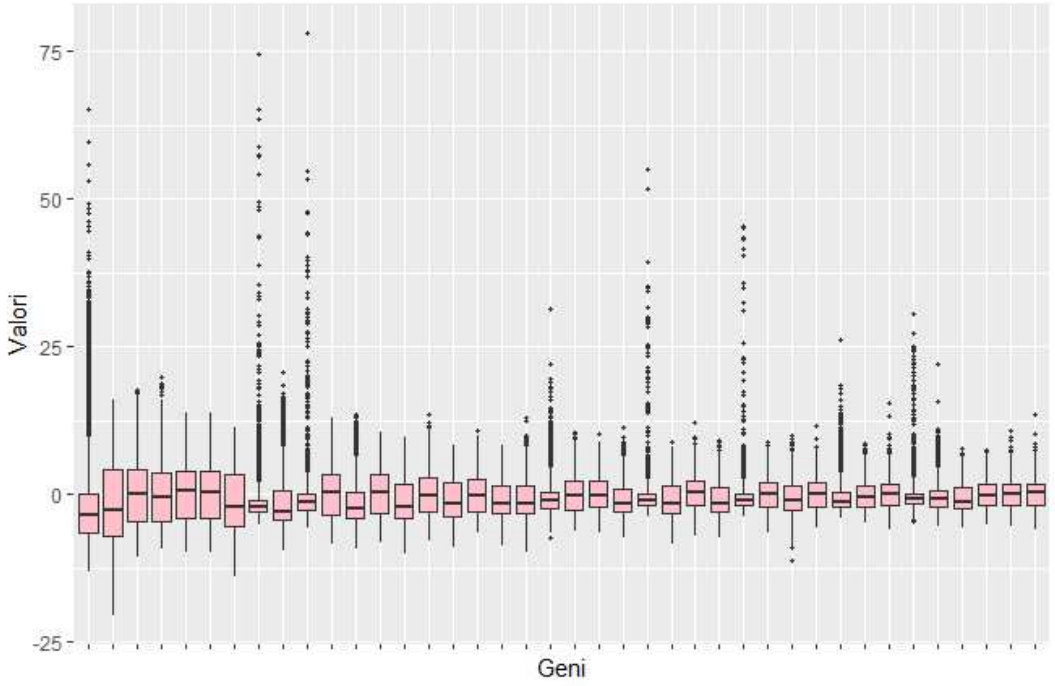

FIGURA 3.2: Boxplot dei primi 40 geni.

Si osserva che, provando a diminuire il numero di spot in cui i geni venivano rilevati, il grafico 3.2 appena discusso, rimaneva pressoché immutato, oltre che a dei minimi cambiamenti del valore medio, si osservava solo un aumento degli outlier. Questo indica che selezionare casualmente 300, 1000 o tutti gli spot della tecnologia Visium non altera i valori di espressione genica rilevati. Si ha che i geni sono espressi spazialmente in modo uniforme, ovvero, per spiegare in termini più semplici, se si seleziona metà del tessuto analizzato, il primo gene denominato "ENSG00000263639" rimane comunque quello più osservabile e più comune in tutti gli spot, e questo accade per tutti i successivi geni rilevati.

3.3 Analisi di robustezza

Si è svolta un'analisi di robustezza per valutare la sensibilità dei metodi mistura nel classificare gli spot nei casi di 2, 3 e 4 cluster al variare dei geni selezionati. Più semplicemente, si sono confrontati il raggruppamento eseguito manualmente dal Dr. Esposito

e il raggruppamento eseguito tramite modelli mistura nelle situazioni di 2, 3 e 4 cluster per tre diverse numerosità dei geni rilevati nel campione. In tabella 3.3 si osservano nella prima colonna le tre casistiche dei gruppi in cui si vogliono suddividere gli spot, mentre nella 2 colonna si confronta il caso in cui si mantengono solo i primi 150 e 500 geni più variabili di tutto il campione, nella terza si verifica la differenza fra mantenerne solo 150 geni oppure selezionarne 1000, ed infine nell'ultima colonna si confronta il caso di mantenimento di tutti i geni con il caso dei primi 500 più rilevati nel tessuto.

Si utilizza la funzione "rand.index" (Rand, 1971) della libreria R "*fossil*" che misura l'indice di Rand, ovvero una misura di similarità tra due gruppi di osservazioni. L'indice varia tra 0 e 1 dove una misura pari a 1 indica che il raggruppamento effettuato in due cluster ha classificato nello stesso modo le osservazioni mentre un valore dell'indice più vicino allo 0 indica un raggruppamento molto differente tra i due gruppi analizzati.

N.gruppi	Confronti dei geni più espressi		
	150-500	150-1000	500-1000
k = 2	0.6455	0.6490	0.9908
k = 3	0.5463	0.6257	0.8175
k = 4	0.9677	0.8179	0.8323

Tabella 3.3: Indici di Rand, in riga il numero di cluster stimati e in colonna i due casi di geni utilizzati per il confronto dei raggruppamenti degli spot.

Nella tabella a doppia entrata 3.3 vengono riportati i vari punteggi dell'indice di Rand ottenuti confrontando i vari cluster che si erano formati con la funzione Mclust (Scrucca et al., 2023) che con la selezione del BIC esegue i raggruppamenti tramite i modelli mistura gaussiani. Sono quindi stati svolti tre tipi di raggruppamenti degli spot per tre diverse numerosità dei geni: 150, 500 e 1000.

Si può vedere come nel caso di 2 e 3 gruppi degli spot fra 150-500 e 500-1000 geni non si abbia un punteggio molto elevato dell'indice di Rand, dato che non si supera lo 0.65, questo sta ad indicare come ci sia una notevole differenza fra il raggruppamento degli spot in 2 o 3 gruppi nella situazione in cui si selezionano solo 150 geni e 500 geni, oppure nella situazione in cui si selezionano 150 geni e 1000 geni, in sintesi i gruppi ottenuti classificano i 4366 spot in modo significativamente differente. Questa evidenza svanisce utilizzando 4 gruppi dove sembra che questa diversità non sia più cosi marcata, infatti il raggruppamento dei 4366 spot in 4 diversi gruppi nel caso di 150 e 500 geni ottiene un punteggio di Rand pari allo 0.96, quindi i gruppi in questi due casi sono molto simili, e peggiora leggermente nel caso di 150 e 1000 geni in cui, sempre valutando il raggruppamento degli spot in 4 cluster diversi si ottiene un indice di Rand pari allo 0.81.

Ottenuto questo risultato si scarta l'ipotesi di mantenere i 150 geni avendo dedotto che sono troppo pochi e si passa a valutare la differenza fra 500 e 1000 geni.

Quindi osservando la 3 colonna si nota come gli indici molto alti (tutti maggiori di 0.80) nei tre casi di raggruppamenti degli spot, ovvero con 2, 3 e 4 gruppi, portino alla conclusione che utilizzare 1000 geni o la metà degli stessi sia pressoché la stessa cosa, dopo questa valutazione, il restante documento si riferisce all'uso dei soli primi 500 geni più informativi dell'intero tessuto analizzato.

3.4 Classificazione delle aree del tessuto tramite modelli mistura

Si procede ora con l'analisi dei gruppi utilizzando la funzione mclustBIC (Scrucca et al., 2023) per il calcolo del *Bayesian Information Criterion* (BIC) (Schwarz, 1978) di diversi modelli mistura gaussiani con differenti strutture di volume, forma e orientamento. Molti modelli non vengono stimati a causa dell'alto numero di geni utilizzati, infatti la funzione riesce a stimare per nove gruppi solo i tre casi in cui si ipotizzano le componenti con distribuzione sferica, diagonale ed elissoidale e più precisamente i casi in cui il volume e la forma sono uguali in tutti i gruppi, ovvero quando il parametro λ e la matrice scomposta "A" sono unici. Le parametrizzazioni appena elencate sono riportate nella tabella 2.1. Si precisa inoltre, che diminuendo il numero dei geni a 70 vengono stimati tutti i tipi di modelli presenti in tabella con 2 o 3 cluster, diminuendo fino a 50 geni si ottengono tutti i modelli in tutti i casi presenti di 9 cluster. Di seguito in figura 3.3 si riportano i risultati del BIC ottenuti dalla funzione precedentemente descritta, cosi da poter valutare il miglior modello per il successivo raggruppamento degli spot.

Il modello con il valore di BIC più alto che ci viene consigliato è quello con forma "EEI" e nove cluster, successivamente quello con otto e sette cluster. Più precisamente il modello preferito è quello che si ipotizza abbia distribuzione diagonale, più precisamente è il modello in cui la matrice di varianza e covarianza è diagonale, con uguale volume e forma, quindi avrà le sembianze dei raggruppamenti riportati in figura 2.3.

A questo punto, si procede con il clustering dei dati con la funzione Mclust alla quale si specifica a quale criterio di selezione si vuole far riferimento che, in questo caso, è il BIC appena calcolato. Si indica anche in quanti gruppi si vogliono dividere gli spot, che dopo vari tentativi si è deciso di mantenere dai due ai quattro gruppi come era stato fatto nella classificazione manuale eseguita dal patologo sopracitato.

Dopo aver eseguito i tre casi di raggruppamento delle osservazioni, si sono ottenuti i

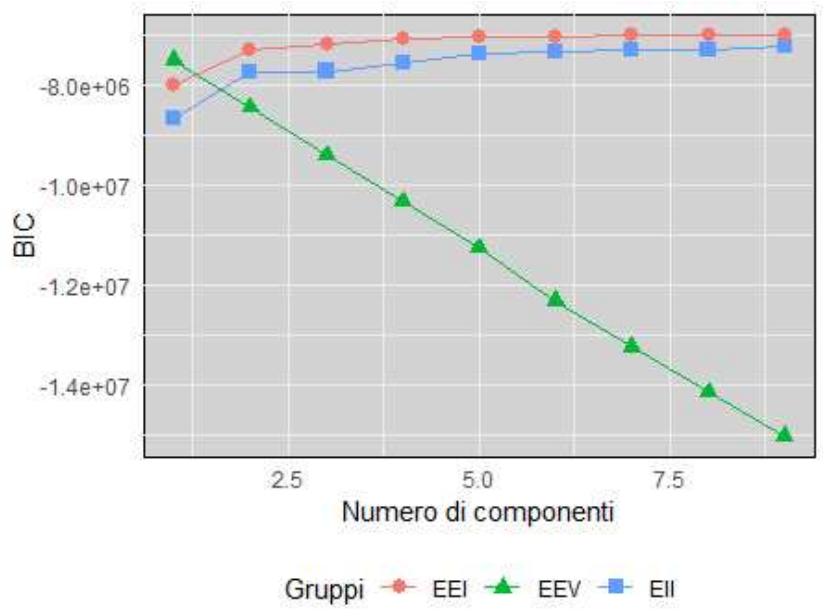

FIGURA 3.3: Grafico dei valori del BIC in funzione dei gruppi.

risultati riportati in figura 3.4 dove si evidenzia nella situazione di quattro cluster che il quarto gruppo è il meno esteso, mentre che il primo e il secondo gruppo sono i più numerosi e più nitidi, come invece non è il terzo gruppo che si dirada un po' su tutto il tessuto. Si valutano ora le classificazioni in due, tre e quattro cluster eseguite mediante modelli mistura rispetto a quella eseguita manualmente dal Dr. Esposito.

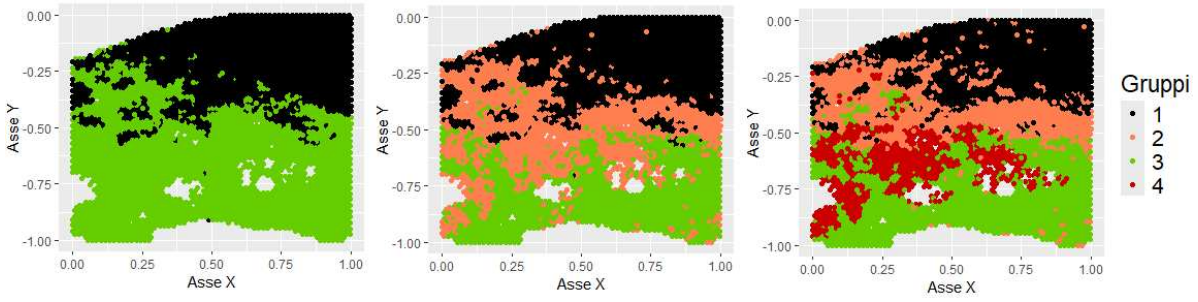

FIGURA 3.4: Da sinistra verso destra i tre raggruppamenti effettuati tramite modelli mistura, rispettivamente con 2, 3 e 4 gruppi.

Il primo confronto che si esegue è il caso in cui si hanno due gruppi formati tramite modelli mistura, più precisamente il gruppo in cui si classificano gli spot come parte tumorale e il gruppo in cui gli spot vengono classificati come Stroma, con la classificazione eseguita manualmente. In tabella 3.4 sono riportate le frequenze assolute dei cluster e nella quarta colonna il punteggio di Rand ottenuto. Sembra che il raggruppamento in 2

Annotazione cluster	Metodologia di raggruppamento		Indice di Rand
	Mclust	Citoarchitettura	
Tumore	1695	1946	0.63
Stroma	2671	1762	
Ghiandole	/	287	
Fibroblasti	/	371	

Tabella 3.4: Tabella di confronto delle frequenze assolute nel caso di 2 gruppi fra il raggruppamento eseguito mediante modelli mistura e il raggruppamento manuale eseguito mediante citoarchitettura dal Dr.Esposito. Nella quarta riga viene riportato l'indice di Rand che misura la similarità dei due raggruppamenti.

gruppi non sia sufficientemente adeguato, infatti i due diversi metodi di raggruppamento risultano molto diversi fra loro, con un punteggio di Rand pari solo a 0.63.

Annotazione cluster	Metodologia di raggruppamento		Indice di Rand
	Mclust	Citoarchitettura	
Tumore	1507	1946	
Stroma	1379	1762	0.70
Ghiandole	1480	287	
Fibroblasti	/	371	

Tabella 3.5: Tabella di confronto delle frequenze assolute nel caso di 3 gruppi fra il raggruppamento eseguito mediante modelli mistura e il raggruppamento manuale eseguito mediante citoarchitettura dal Dr.Esposito. Nella quarta riga viene riportato l'indice di Rand che misura la similarità dei due raggruppamenti.

Nel secondo confronto si prendono in considerazione tre diversi gruppi formati mediante il modello mistura, caratterizzati da parti del tessuto classificate come tumore, stroma e ghiandole, con la classificazione eseguita manualmente. In tabella 3.5 si riportano le frequenze assolute dei tre diversi cluster. Si nota che l'indice di Rand ora ha un valore pari a 0.70, migliorato rispetto al caso precedente ma non ancora un punteggio che si può ritenere soddisfacente.

Nella classificazione eseguita in 4 diversi gruppi eseguita mediante modello mistura, si aggiunge anche il gruppo in cui gli spot vengono classificati come dei fibroblasti. In questo caso si ottiene un indice di Rand pari a 0.72 quindi migliora poco rispetto al precedente caso in cui si confrontava il caso di 3 cluster. La differenza di numerosità si nota in tutti i cluster, infatti gli spot che vengono classificati come parte tumorale del tessuto differiscono per ben 700 unità. Il patologo classifica molti più spot del modello mistura nei primi due gruppi, tumore e stroma, con delle numerosità pari a 1946 e 1762, e molti meno per i gruppi formati da ghiandole (287) e fibroblasti (371), quando invece il modello mistura sembra fare meno differenze per i quattro gruppi, classificando più spot

Annotazione cluster	Metodologia di raggruppamento		Indice di Rand
	Mclust	Citoarchitettura	
Tumore	1253	1946	0.72
Stroma	1395	1762	
Ghiandole	993	287	
Fibroblasti	725	371	

TABELLA 3.6: Tabella di confronto delle frequenze assolute nel caso di 4 gruppi fra il raggruppamento eseguito mediante modelli mistura e il raggruppamento manuale eseguito mediante citoarchitettura dal Dr.Esposito. Nella quarta riga viene riportato l'indice di Rand che misura la similarità dei due raggruppamenti.

nel 3 e 4 gruppo (ghiandole e fibroblasti) con delle numerosità pari rispettivamente a 993 e 725. Questa differenza nel classificare un diverso numero di spot è visibile direttamente dalle figure di tessuto suddiviso nei 4 cluster ottenuti prima dal Dr. Esposito che si trova in figura 3.1, e dal modello mistura che si nota poco sopra in figura 3.4.

3.5 Classificazione delle aree del tessuto tramite k-means e PCA

In questa sezione si individuano il numero di cluster da utilizzare per eseguire il raggruppamento tramite k-means e si svolge un'analisi per la selezione delle componenti principali, che poi verrà utilizzata per un ulteriore raggruppamento tramite il modello mistura. Un passo fondamentale per l'utilizzo del metodo delle k-means è, come si è detto nella sezione precedente 2.2.1, la selezione del numero di cluster in cui si vogliono partizionare i dati. Da una prima analisi di raggruppamento eseguita con il metodo delle k-means sul dataset selezionato, si è evidenziato come l'innalzamento del numero di gruppi non porti ad un grosso vantaggio in termini di varianza totale spiegata dai raggruppamenti, dato che uno dei metodi più usuali per la selezione del numero di gruppi è quello di ottenere una varianza spiegata fra il 70% e l'80%.

Come si può vedere dalla figura 3.5 fino a 4 gruppi c'è un incremento significativo della varianza spiegata, poi, all'aumentare dei cluster non c'è un'evidente crescita quindi risulta pressoché inutile selezionarne di più. Si può notare come dopo il quarto gruppo ci sia un incremento del solo 2% o anche minore nei sei gruppi successivamente aggiunti. Dunque si è deciso di assumere come più valido il raggruppamento con 4 cluster, avendo un valore di varianza spiegata totale pari al 44.7%.
Prima di procedere al confronto fra i vari metodi di clustering si valutano graficamente le componenti principali e l'importanza di ognuna di esse. Si è applicata la funzione

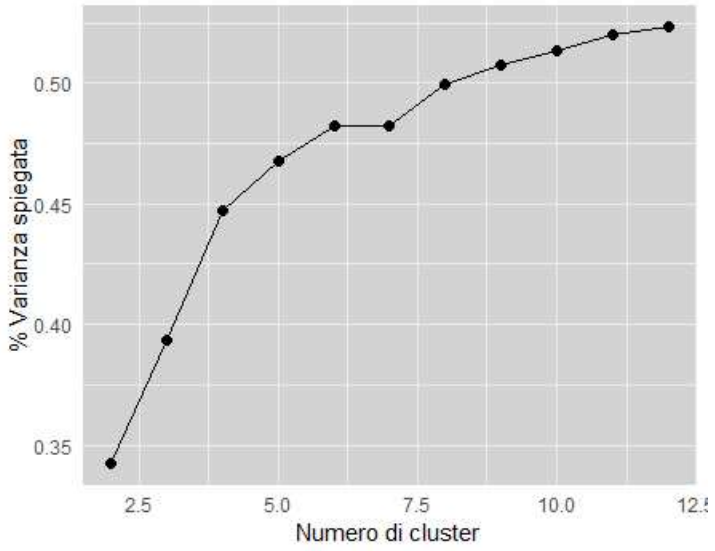

FIGURA 3.5: Grafico dei vari raggruppamenti eseguiti con k-means.

"prcomp", presente nel software "RStudio" (Becker et al., 1988), che effettua l'analisi delle componenti principali sul dataset ridotto composto da 500 geni. Si hanno ora a disposizione tre componenti principali che riassumono l'espressione dei geni attraverso delle combinazioni lineari delle variabili originali che riassumono il 53% di varianza spiegata dell'intero dataset considerato.

Dando uno sguardo alle nuove variabili, gli spot sono ora rappresentati nello spazio delle componenti principali che vengono riportati in un grafico rispetto a due componenti prese contemporaneamente in considerazione. Applicando la classificazione degli spot eseguita tramite modelli mistura, nel caso di tre gruppi, riportata in figura 3.6, si può notare come in questa nuova "dimensione" la prima componente principale, posta nell'asse delle ascisse nel 1 e 2 grafico, riesca a discriminare molto vistosamente gli spot per il gruppo di cui fanno parte. Infatti gli spot che hanno valori alti positivi, presenti nella 1 componente principale, vengono classificati come parte sana del tessuto.

I valori attorno allo zero che sono presenti nella prima componente principale vengono classificati in modo abbastanza chiaro nel terzo gruppo, quello composto dalle ghiandole, ed infine i valori negativi vengono classificati nel gruppo che riguarda il tumore.

La notevole capacità di questa componente di suddividere gli spot nei tre gruppi in modo chiaro dipende dall'importanza della prima componente principale, come è emerso anche dalle analisi, in quanto da sola spiega il 42% della proporzione di varianza presente nei dati originali, risulta molto significativa perché funge da variabile riassuntiva di tutti i geni con dei pesi differenti per importanza.

Queste evidenti diversificazioni non sono cosi marcate ed osservabili nella seconda componente principale, infatti come si nota dal primo grafico ma guardando l'asse delle

ordinate i geni hanno tutti valori attorno allo zero, si può però dire che i valori alti positivi presenti in questa componente vengono classificati nel gruppo caratterizzato delle ghiandole.

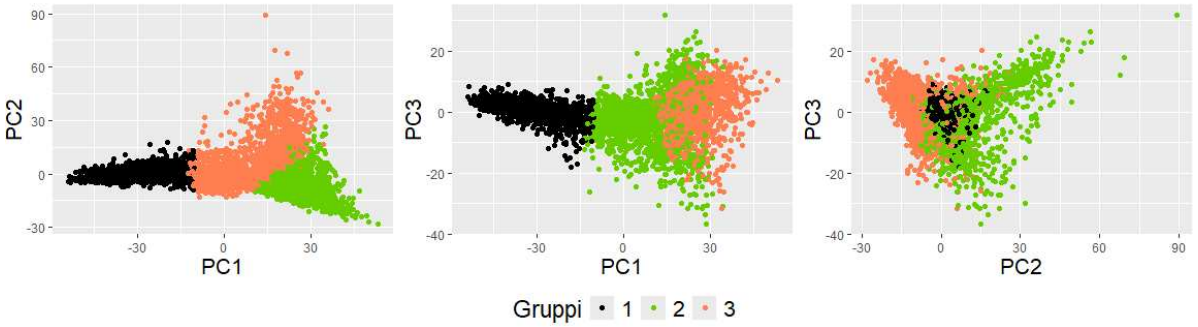

FIGURA 3.6: Confronto a coppie fra le prime tre componenti principali con un totale di 53% di varianza spiegata totale, i raggruppamenti sono ottenuti con il modello "EEI" eseguito tramite modello mistura.

Nel secondo grafico della figura 3.6 si mettono a confronto la prima e la terza componente principale. Si nota, di nuovo, la differenza molto marcata dei tre cluster nell'asse delle ascisse dove è presente la prima componente principale, anche se in maniera meno rilevante. Rispetto alla terza componente principale invece, si osserva che per valori alti positivi o bassi negativi gli spot vengono classificati nei gruppi diversi dal tumore, quindi tendenzialmente sani, mentre per valori vicini allo zero si caratterizzano unità classificate come malate.

Infine mettendo in relazione la seconda e la terza componente principale, che si può osservare nel terzo riquadro presente in figura 3.6, non si notano grandi differenze di valori che gli spot assumono in funzione di gruppi di appartenenza diversi, questo è anche conseguenza di ciò che era emerso sulla scelta delle componenti, infatti con solo 8% e 3% rispettivamente, di varianza dei dati originali spiegata queste due componenti non sono molto rilevanti. Si può dire che i valori nulli e vicini allo zero, in questo caso vengono classificati in entrambe le nuove variabili come tumorali, mentre per valori che si discostano dallo zero si hanno dei gruppi alternativi come stroma e ghiandole.
Avendo capito che la prima componente principale è quella che riassume meglio i dati originali, se ne prendono ora in considerazione i valori della matrice dei pesi, in cui si hanno dei valori che fungono da pesi per ogni gene, in base a quanto è rappresentato. Avendo i valori che vanno da -1 a 1, i valori alti positivi o bassi negativi presenti in questa matrice indicano che il gene è significativo e avrà un ruolo importante nell'interpretazione di tale componente, mentre valori attorno allo 0 indicano che quel gene non

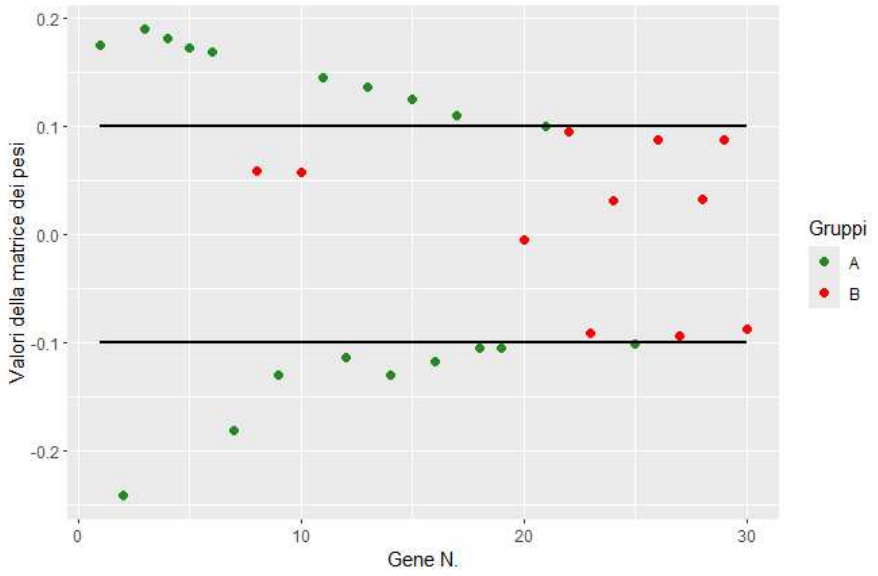

FIGURA 3.7: Grafico dei primi 30 valori dei geni della matrice di rotazione riferita alla prima componente principale, il gruppo A di colore verde si riferisce ai valori esterni alle linee di riferimento, il gruppo B di colore rosso ai valori interni.

è significativo nell'interpretazione. Ricordando che i geni sono ordinati per importanza, si selezionano soli i primi trenta che compongono la prima componente principale, e che si possono osservare in figura 3.7.

Come si può vedere dalla figura 3.7 sono presenti due linee orizzontali nei valori di 0.1 e -0.1 che caratterizzano una soglia di importanza, ovvero se ne considerano soltanto i geni con pesi maggiori o minori. In pratica, si stanno selezionando i geni che hanno dei valori nella matrice dei pesi più alti o bassi di una soglia designata, per fare in modo di avere soltanto i geni che rappresentano maggiormente la prima componente principale. I geni colorati in verde sono quelli con valori alti positivi o bassi negativi, più precisamente maggiori/uguali o minori/uguali rispettivamente di 0.1 e di -0.1.

Da questa analisi emerge che i geni più rappresentativi della prima componente principale se hanno valore basso negativo vengono classificati nella zona tumorale, invece i geni con valori dei pesi alti positivi vengono classificati nella parte sana del tessuto, in più si nota questa compensazione tra valori dei pesi alti positivi e bassi negativi che rendono ancor più l'idea della diversificazione delle due aree di tessuto.

3.6 Comparazione dei vari metodi di raggruppamento

Si procede in questa sezione con i confronti tra il metodo di raggruppamento tramite modelli mistura con il metodo delle k-means e con le componenti principali illustrato precedentemente. Il primo confronto che si effettua è quello fra il metodo di raggruppamento tramite modello mistura e lo stesso ma usato sulle sue componenti principali.

Si specifica che nei seguenti 3 confronti fra raggruppamenti, nei grafici rappresentanti il tessuto, precisamente nelle figure: 3.8, 3.9 e 3.10, la notazione che si riferisce al gruppo 1 fa riferimento agli spot classificati come tumorali, il gruppo 2 agli spot classificati come stroma, il gruppo 3 agli spot classificati come ghiandole ed infine il gruppo 4 agli spot classificati come fibroblasti.

3.6.1 Mclust VS MclustPCA

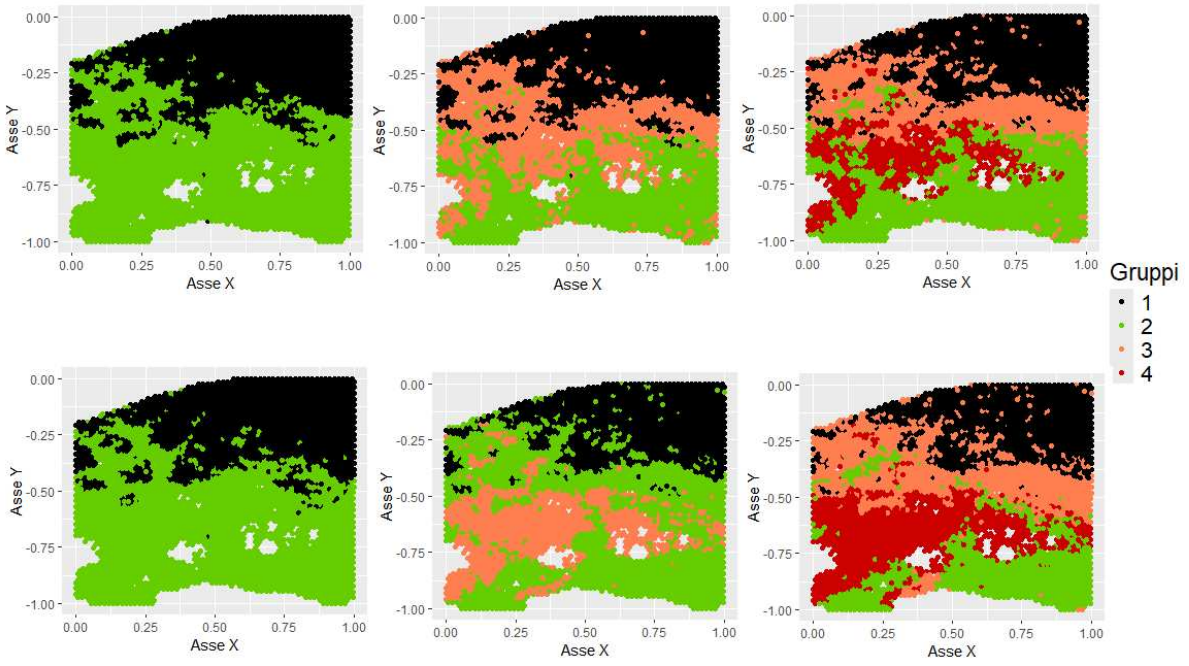

FIGURA 3.8: Nella prima riga sono rappresentati i raggruppamenti effettuati sui dati originali, mentre nella seconda riga quelli effettuati sulle componenti principali, in entrambi i casi si è utilizzato il metodo basato sui modelli mistura.

In tabella 3.7 sono riportati i risultati del clustering effettuato con il procedimento mediante modelli mistura prima sui dati originali e poi sui dati trasformati mediante PCA. Dalla prima colonna si nota che nel caso di due gruppi le due procedure classificano nello stesso modo gran parte degli spot, infatti l'indice di Rand, che si può osservare

nella prima riga in tabella 3.8, di 0.92 è molto alto, mentre nel caso di tre e quattro gruppi la classificazione peggiora.

Come si nota nel caso di tre gruppi l'indice è pari a 0.71, questo sta ad indicare che in questa situazione diversi spot non vengono classificati in gruppi differenti. Nel caso di quattro gruppi l'indice raggiunge un valore pari a 0.85, quindi migliora rispetto al caso di tre gruppi, questo si vede chiaramente dalla seconda e terza colonna della figura 3.8, in cui sono raffigurati rispettivamente i raggruppamenti con 3 e 4 cluster.

Il risultato ottenuto sembra indicare che queste procedure di raggruppamento siano in accordo, ciò si poteva prevedere quando nella scelta delle componenti principali la prima variabile trasformata rappresentava la gran parte dei geni più informativi del campione. Le differenze poco marcate nel caso di due gruppi si possono osservare più facilmente dai grafici 3.8, mentre si nota come nelle situazioni con tre e quattro gruppi siano presenti maggiori differenze sopratutto nei gruppi 3 e 4, rispettivamente di colore arancione e rosso.

Gruppi	Mclust			MclustPCA		
	2	3	4	2	3	4
Tumore	1695	1507	1253	1613	1220	1103
Stroma	2671	1379	1395	2753	2059	908
Ghiandole	/	1480	995	/	1087	1225
Fibroblasti	/	/	723	/	/	1130

Tabella 3.7: Tabella di classificazione con due, tre e quattro gruppi eseguita mediante modelli mistura contro lo stesso metodo eseguito su tre componenti principali.

N. Gruppi a confronto	Indice di Rand
k = 2	0.92
k = 3	0.71
k = 4	0.85

Tabella 3.8: Tabella contenente l'indice di Rand per la valutazione dei due casi di raggruppamento eseguiti con 2, 3 e 4 cluster, fra il metodo mediante modello mistura prima coi dati originali e poi sulle componenti principali.

Si passa ora alla valutazione tra il metodo di raggruppamento effettuato tramite modello mistura sui dati originali con il metodo delle k-means.

3.6.2 Mclust VS k-means

In questo caso sembra che i due metodi lavorino molto bene e sicuramente meglio del caso precedente. Anche in questo confronto si osserva che nel caso di tre gruppi sono presenti dei disaccordi fra le metodologie utilizzate, nel caso di due gruppi si nota come le coppie di osservazioni collocate nel gruppo 1 e gruppo 2 siano molto simili, infatti in questo primo caso l'indice di Rand raggiunge il valore di 0.98. Anche nella situazione di quattro gruppi si può osservare come tutti i cluster siano in accordo fra loro raggiungendo in questo caso un punteggio di Rand pari a 0.96, tutti gli indici di Rand di questo confronto sono riportati in tabella 3.10.

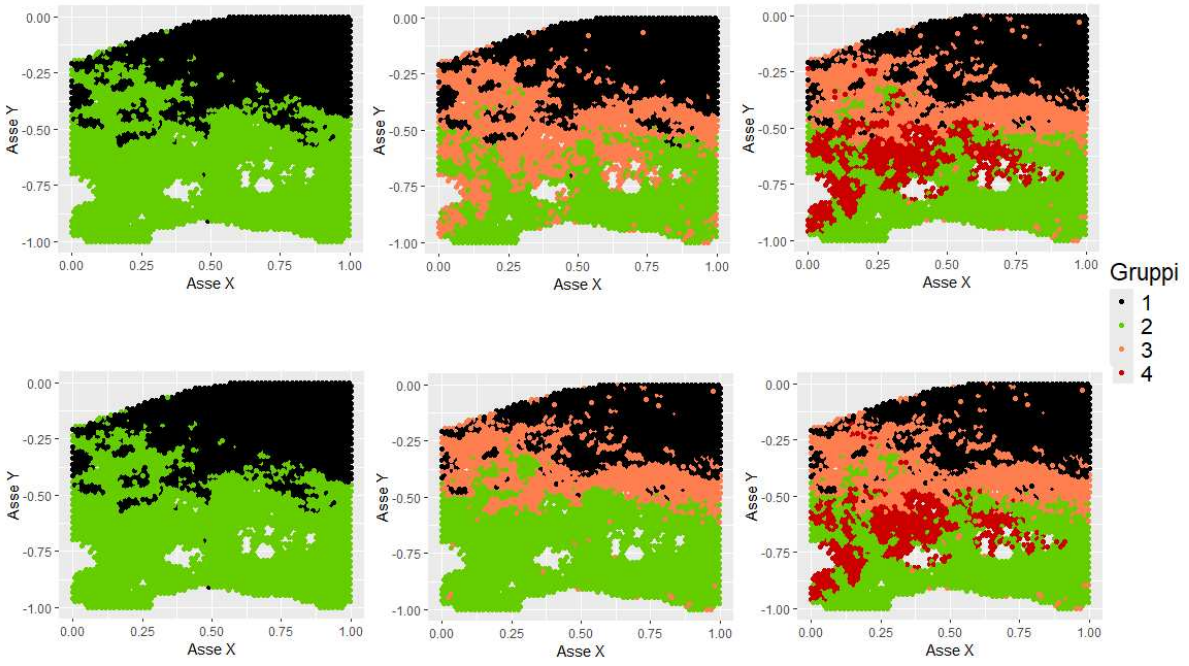

FIGURA 3.9: Nella prima riga sono rappresentati i raggruppamenti effettuati sui dati originali ottenuti tramite modelli mistura, mentre nella seconda riga quelli eseguiti con il metodo k-means.

Come descritto in precedenza, nelle situazioni con due o quattro gruppi questi due metodi classificano in modo molto simile la maggior parte delle coppie di osservazioni, questo si può osservare anche dal numero di unità presenti nei gruppi che non differiscono mai di più di 100. Questa concordanza non è cosi marcata nel caso di tre gruppi in cui l'indice di Rand è pari a 0.77 e sono presenti diverse differenze nelle classificazioni. Queste differenze poco marcate nei casi di due e quattro gruppi si possono osservare nella figura 3.9 che sono rappresentate nella prima e terza colonna, mentre la discordanza nel caso di tre gruppi è osservabile nella colonna centrale dove si nota molto bene come le due procedure classifichino diversamente gli spot nei gruppi 2 e 3 di colore verde e arancione.

Gruppi	Mclust			k-means		
	2	3	4	2	3	4
1	1695	1507	1253	1687	1219	1230
2	2671	1379	1395	2679	2123	1470
3	/	1480	993	/	1023	1043
4	/	/	725	/	/	623

Tabella 3.9: Tabella di classificazione con due, tre e quattro gruppi eseguita mediante modelli mistura contro il raggruppamento effettuato con le k-means.

N. Gruppi a confronto	Indice di Rand
k = 2	0.98
k = 3	0.77
k = 4	0.96

Tabella 3.10: Tabella contenente l'indice di Rand per la valutazione dei due metodi di raggruppamento eseguiti con 2, 3 e 4 cluster, fra il metodo mediante modello mistura sui dati originali e il metodo delle k-means.

Infine si valutano il metodo tramite modello mistura utilizzato sulle componenti principali dei dati originali contro il metodo delle k-means.

3.6.3 MclustPCA VS k-means

In questo caso si confrontato i raggruppamenti dei dati trasformati mediante componenti principali e raggruppati mediante modelli mistura con i dati originali raggruppati mediante k-means, si nota come nella situazione di due gruppi ci sia sempre molta concordanza. Questo è emerso anche nelle precedenti comparazioni, infatti anche qui i metodi ottengono un punteggio di Rand, riportato in tabella 3.11 pari a 0.92.

N. Gruppi a confronto	Indice di Rand
k = 2	0.92
k = 3	0.73
k = 4	0.84

Tabella 3.11: Tabella contenente l'indice di Rand per la valutazione dei due casi di raggruppamento eseguiti con 2, 3 e 4 cluster, fra il metodo mediante modello mistura eseguito sulle componenti principali e il metodo k-means.

La situazione cambia e peggiora nel caso di tre gruppi, in cui si ottiene un punteggio di Rand di 0.73 che sottolinea ancora una volta come nel caso di tre gruppi i metodi

facciano fatica a distinguere fra il gruppo 2 e 3, ovvero quelli caratterizzati dalla parte sana, formata da cellule classificate come ghiandole o stroma.

Gruppi	MclustPCA			k-means		
	2	3	4	2	3	4
1	1613	1220	1094	1687	1219	1230
2	2753	2059	939	2679	1023	1470
3	/	1087	1185	/	2124	1043
4	/	/	1148	/	/	623

TABELLA 3.12: Tabella di classificazione con due, tre e quattro gruppi eseguita mediante modelli mistura sulle prime tre componenti principali contro il metodo delle k-means.

Nel caso in cui si ipotizzino quattro gruppi la situazione sembra migliorare, si ottiene un punteggio di Rand pari 0.84 e qui la criticità sembra risiedere sempre nella diversificazione della fascia centrale del tessuto, come si può osservare dalla figura 3.10 nella terza colonna la parte rossa e arancione sono molto differenti.

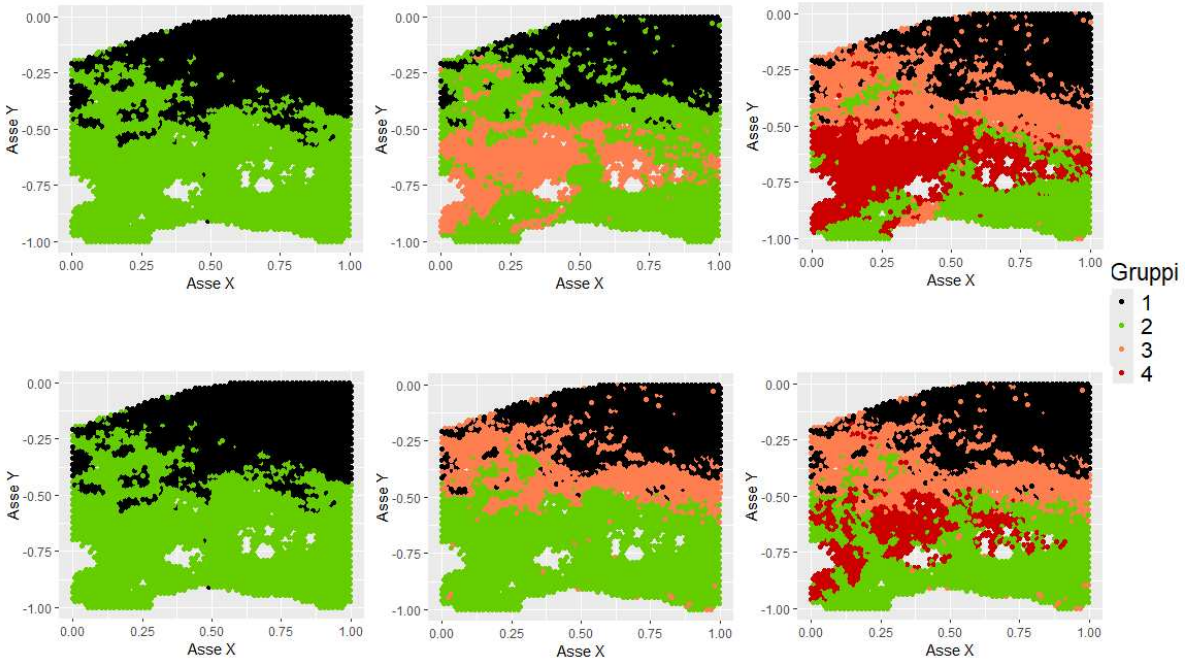

FIGURA 3.10: Nella prima riga sono rappresentati i raggruppamenti effettuati sulle componenti principali tramite modelli mistura, mentre nella seconda riga quelli tramite il metodo delle k-means.

Anche nella seconda colonna della figura 3.9 sono molto evidenti le differenze nel caso di tre gruppi descritte poco sopra, infatti si nota facilmente come la zona verde nella

prima riga, che si riferisce al raggruppamento eseguito sulle componenti principali mediante modello mistura, sia più sparsa e ampia rispetto alla seconda riga, che si riferisce ai dati originali raggruppati con k-means, che in questo caso è più arancione/nera.

Conclusioni

In questo elaborato sono stati analizzati i metodi di raggruppamento eseguiti tramite modello mistura e tramite k-means, per la classificazione cellulare.

Il dataset ridotto, formato da una matrice contenente 4366 righe, in cui si avevano gli spot della tecnologia Visium utilizzata, ognuno con il suo *barcode* identificativo e 500 colonne in cui si avevano i geni ordinati in modo da avere i più importanti nelle prime posizioni e quelli meno significativi nelle ultime.

L'obiettivo era quello di valutare e comparare diverse metodologie di clustering per capire se erano utili al rilevamento dei diversi gruppi di cellule presenti nel tessuto scelto. Le metodologie di clustering utilizzate sono state il pacchetto Mclust, che si basa sui modelli mistura, sempre Mclust però applicato sulle componenti principali ed infine il metodo delle k-means.

Si aveva a disposizione una classificazione degli spot eseguita manualmente dal Dr. Esposito, presa come riferimento, che si è confrontata con il raggruppamento eseguito mediante modello mistura. Da questo primo confronto è risultato che i due metodi non erano molto in accordo, utilizzando l'indice di Rand per misurare la similarità dei diversi gruppi, si erano ottenuti dei risultati poco soddisfacenti.Dopo aver studiato brevemente il numero di gruppi da utilizzare nel metodo di clustering k-means, ed avendo individuato 4 gruppi come accettabili, si è studiato il numero di componenti principali adeguato per ridurre i 500 geni di partenza, ottenendo tre componenti sufficienti per rappresentarli.

Il raggruppamento eseguito mediante modelli mistura risulta un buon metodo di raggruppamento in questo caso di dati di espressione genica normalizzati, con gli unici problemi che in questi tipi di dataset il calcolo del BIC risulta complicato in termini computazionali date le numerose variabili presenti. Il primo confronto aveva mostrato che i raggruppamenti fra dati originali e componenti principali mediante modello mistura era accettabile, ma, dopo aver visto che la procedura tramite k-means è computazionalmente più rapida ed esegue dei cluster pressoché uguali in 2 casi su 3 rispetto a Mclust, si ritiene il migliore dei tre casi di raggruppamento nel caso di dataset di grosse dimensioni e con dati per la classificazione cellulare.

Bibliografia

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control* **19**, 716–723.

Banfield, J. D. & Raftery, A. E. (1993). Model-based gaussian and non-gaussian clustering. *Biometrics* **49**, 803–821.

Becker, R., Chambers, J. & Wilks, A. (1988). *The New S Language: A Programming Environment for Data Analysis and Graphics*. Computer science series. Wadsworth & Brooks/Cole Advanced Books & Software.

Bolasco, S. (2022). *Analisi multidimensionale dei dati. Metodi, strategie e criteri d'interpretazione*. Carocci.

Castiglioni, S. A. (2022/2023). Modello di clustering per profili di variazione spaziale di dati di trascrittomica .

Celeux, G. & Govaert, G. (1995). Gaussian parsimonious clustering models. *Pattern Recognition* **28**, 781–793.

Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)* **39**, 1–22.

Fraley, C. & Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association* **97**, 611–631.

Hand, D. J. (2018). Mixture Models: Inference and Applications to Clustering. *Journal of the Royal Statistical Society Series C: Applied Statistics* **38**, 384–385.

Johnson, R. A., Wichern, D. W. & Johnson, R. A. (2014). *Applied Multivariate Statistical Analysis / Richard Johnson, Dean Wichern*. Edinburgh: Pearson Education Limited, 6th ed.

Kalluri, R. & Zeisberg, M. (2006). Fibroblasts in cancer. *Nat. Rev. Cancer* **6**, 392–401.

Lo, Y., Mendell, N. R. & Rubin, D. B. (2001). Testing the number of components in a normal mixture. *Biometrika* **88**, 767–778.

Murtagh, F. & Raftery, A. (1984). Fitting straight lines to point patterns. *Pattern Recognition* **17**, 479–483.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association* **66**, 846–850.

Reed, M. & Simon, B. (1980). *Methods of Modern Mathematical Physics: Functional analysis*. No. v. 1 in Methods of Modern Mathematical Physics. Academic Press.

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics* **6**, 461 – 464.

Scrucca, L., Fraley, C., Murphy, T. B. & Raftery, A. E. (2023). *Model-Based Clustering, Classification, and Density Estimation Using mclust in R*. Chapman and Hall/CRC.

Sottosanti, A. & Risso, D. (2023). Co-clustering of spatially resolved transcriptomic data. *The Annals of Applied Statistics* **17**, 1444 – 1468.

Sund, M. & Kalluri, R. (2009). Tumor stroma derived biomarkers in cancer. *Cancer Metastasis Rev.* **28**, 177–183.

Townes, F. W., Hicks, S. C., Aryee, M. J. & Irizarry, R. A. (2019). Feature selection and dimension reduction for single-cell RNA-Seq based on a multinomial model. *Genome Biol* **20**, 295.