Università degli Studi di Padova
Dipartimento di Scienze Statistiche

Corso di Laurea Magistrale in
Scienze Statistiche

MODELLO DI CLUSTERING PER PROFILI DI VARIAZIONE SPAZIALE DI DATI DI TRASCRITTOMICA

Relatore Prof. Davide Risso
Dipartimento di Scienze Statistiche

Correlatore Dott. Andrea Sottosanti
Dipartimento di Medicina

Correlatore Dott.ssa Stefania Pirrotta
Dipartimento di Biologia

Laureanda Sara Agavni' Castiglioni
Matricola 2026499

Anno Accademico 2022/2023

Indice

Introduzione

Negli ultimi anni, le tecnologie di sequenziamento del DNA sono notevolmente migliorate, dando origine a nuovi protocolli avanzati per il sequenziamento dell'RNA di una singola cellula (scRNA-seq) e, più recentemente, alla trascrittomica spaziale. Quest'ultima tecnologia è stata scelta da *Nature* come metodo dell'anno 2020 (Marx, 2021): permette di misurare l'espressione di migliaia di geni all'interno di un tessuto, fornendo informazioni non solo su quali geni sono espressi e in che quantità, ma anche su dove è stata effettuata la misurazione. Esplorare la distribuzione spaziale del profilo di espressione genica di un tessuto è di grande interesse scientifico perchè può rivelare preziose informazioni sui meccanismi biologici alla base della comunicazione cellula-cellula e sulle interazioni tumore-microambiente.

Esistono diversi protocolli di trascrittomica spaziale, e in questa tesi ci si concentra sui metodi *spot-based*, come definiti da Righelli et al. (2022). Una delle tecnologie più diffuse che segue questo tipo di protocolli è la piattaforma Visium di 10xGenomics (Rao et al., 2020). In breve, nella piattaforma Visium la sezione di tessuto viene posizionata su una griglia di celle, dette *spot*, dove ogni *spot* corrisponde ad un piccolo gruppo di cellule adiacenti di cui viene catturata l'espressione genica, ottenendo le misurazioni di decine di migliaia di geni per ogni *spot* e della relativa posizione spaziale.

La trascrittomica spaziale consente di integrare l'informazione spaziale nell'analisi dell'espressione genica, offrendo la possibilità di fare inferenza sui numerosi processi biologici che dipendono dall'organizzazione delle cellule nel tessuto. Grazie alla modellazione di questi dati è possibile identificare i geni che mostrano pattern di espressione

spaziale all'interno del tessuto, permettendo di acquisire nuove conoscenze sull'interazione tra le cellule, sulla regolazione dei geni e sulle eventuali deviazioni dei processi biologici in corso dai normali meccanismi di regolazione genica. Inoltre, individuare gruppi di geni co-regolati tra le cellule può portare alla scoperta di nuove vie metaboliche e alla predizione di funzioni per geni non ancora annotati (Chen et al., 2015).

In questo lavoro di tesi si presenta un modello e algoritmo di clustering per dati di espressione genica provenienti da un campione di tessuto processato con metodi di trascrittomica spaziale *spot-based*. Il metodo elaborato è una versione semplificata del metodo di co-clustering Spartaco (SPAtially Resolved TrAnscriptomics CO-clustering) proposto da Sottosanti & Risso (2022). Il metodo Spartaco originale individua gruppi di geni che presentano un'espressione simile all'interno di cluster di *spot* stimati. Nella nuova versione, invece, i gruppi di *spot* vengono definiti sulla base dell'annotazione manuale del tessuto, che spesso è già disponibile e presenta una fonte di conoscenza a priori. Questo approccio permette di ridurre notevolmente i tempi di calcolo del modello Spartaco, in cui i tempi di stima sui dataset di dimensioni standard in trascrittomica sono dell'ordine di diversi giorni e per questo il metodo risulta poco pratico nell'applicazione. Inoltre, si è studiata una seconda versione del metodo di clustering dei geni in cui si inserisce una penalizzazione nella massimizzazione della verosimiglianza del modello. L'aggiunta della penalizzazione costituisce un contributo originale che ha lo scopo di regolarizzare le stime dei parametri e ridurre la loro variabilità.

Per valutare la metodologia statistica adottata, si è condotto un estensivo studio di simulazione. Sono stati considerati diversi scenari di struttura spaziale dei dati, in cui si analizza l'accuratezza del modello nella formazione dei cluster dei geni e nella stima dei parametri, il contributo della penalizzazione alla performance del metodo, e infine, si indaga la robustezza del metodo nell'identificare correttamente i cluster dei geni in presenza di incertezze nell'annotazione degli *spot*.

I nuovi metodi di clustering dei geni formulati in questa tesi sono stati applicati ai dati di trascrittimica spaziale di un tessuto prostatico affetto da adenocarcinoma, mostrando concretamente come implementare la metodologia e l'utilità nel rispondere a specifiche domande biologiche rilevanti. Si è collaborato con il dipartimento di biologia

dell'Università di Padova per condurre un'ulteriore analisi sui dati di *signature*, le quali rappresentano misure sintetiche di alcune delle principali attività tumorali. Grazie all'applicazione del metodo di clustering, vengono individuate similitudini tra le distribuzioni delle diverse signature e viene quantificata la spazialità dell'attività che esse rappresentano.

Parte integrante del lavoro di tesi è stato lo sviluppo di una libreria R che implementasse la versione per il clustering dei geni del metodo Spartaco nella sua versione base e nella versione penalizzata. Il codice è consultabile nella cartella Github `https://github.com/sacastiglioni/ClusteringSpartaco`.

Capitolo 1

Contesto biologico

1.1 Trascrittomica spaziale

Un affascinante enigma della biologia molecolare riguarda come le cellule con lo stesso corredo genetico possano dare origine a cellule con caratteristiche diverse, ciascuna con una specifica funzione all'interno di un organismo multicellulare. Ad esempio, le cellule tumorali sono della stessa tipologia delle cellule sane corrispondenti, ma hanno perso il loro comportamento originale e hanno acquisito un nuovo comportamento, dannoso. Questa diversità fenotipica tra le cellule è stata attribuita alla capacità di attivare o disattivare geni diversi o in quantità diverse. Per questo motivo, la trascrittomica si occupa di studiare la correlazione tra destino e funzione delle cellule e i profili di espressione genica.

I geni sono porzioni di DNA che contengono le informazioni necessarie per codificare le proteine, elemento fondamentale della vita e responsabili della maggior parte delle attività cellulari. Studiare quali proteine sono presenti in una cellula aiuta a comprendere come funziona la cellula stessa, in quali meccanismi biologici è coinvolta, il suo destino e quale sia il suo ruolo all'interno di un organismo. Inoltre, alcune proteine sono associate a malattie o condizioni patologiche: l'identificazione di proteine che sono presenti in cellule malate e assenti in cellule sane può aiutare a identificare nuovi marcatori di malattie e a sviluppare nuove strategie diagnostiche e di cura (Gonzalez & Kann, 2012).

Il processo di trasformazione di un gene in una proteina avviene tramite un meccanismo di copia-incolla descritto dal dogma centrale della biologia molecolare di Crick (1970). In sintesi, il DNA viene prima convertito in molecole di RNA e poi tradotto in proteine. Questo processo è composto quindi da due fasi fondamentali: la trascrizione e la traduzione. Durante la trascrizione, che avviene nel nucleo, dei fattori di trascrizione trasformano la porzione di DNA che codifica la proteina desiderata in un filamento di RNA (mRNA), detto trascritto. I trascritti vengono poi portati fuori dal nucleo e nel citoplasma vengono tradotti in una sequenza di aminoacidi per formare la proteina finale. Nonostante il patrimonio genetico sia lo stesso in tutte le cellule di un organismo, i geni espressi, cioè quelli che vengono trascritti, variano a seconda dell'attività della cellula o del processo biologico in cui è coinvolta.

Il trascrittoma è la totalità dei trascritti presenti in una cellula, e la loro quantità. Studiare il trascrittoma permette di comprendere quali e quante copie di RNA vengono prodotte per ogni gene in una data cellula e questo fornisce una misura di abbondanza della presenza delle proteine codificate da tale gene.

L'espressione genetica è regolata da molteplici fattori, tra cui l'ambiente cellulare e lo stato della cellula. Differenze tra i trascritti delle cellule possono rappresentare differenze a livello di processi biologici in atto, oppure differenze a livello di genoma dovute alla presenza di mutazioni nel DNA. L'analisi dell'espressione genica può portare ad un aumento della conoscenza su processi ancora sconosciuti ed è sempre più rilevante nella ricerca biologica di base e medica (Kim et al., 2010).

Sono state sviluppate molteplici tecnologie per misurare i profili di espressione genica delle cellule. Una delle tecnologie più utilizzate è RNA-seq, un tecnica di sequenziamento dei trascritti di RNA. Di seguito viene presentata una sintesi del procedimento.

I trascritti sono lunghi filamenti costituiti da circa 10000 basi azotate, che sono di quattro tipi: Adenina, Guanina, Citosina, Uracile. Nessuna tecnologia è in grado di leggere l'intera sequenza codificante in modo efficiente. Per questo motivo, prima del sequenziamento, è necessario preparare il materiale: il frammento di RNA originario viene suddiviso in breve sequenze di 50-150 basi e l'insieme dei frammenti ottenuti costituiscono una libreria. A questo punto i frammenti della libreria possono essere sequenziati; si ottengono delle stringhe di quattro lettere, dette *reads*, corrispondenti alla

sequenza delle quattro basi azotate (Wang et al., 2009). Successivamente, è necessario identificare da quale gene provengono le *reads* lette; questo viene fatto con un procedura di allineamento delle *reads* sul genoma. Infine, si conta il numero di *reads* allineate che mappano su ogni gene. Queste conte rappresentano i livelli di espressione del gene: più il numero è alto, più il gene è espresso Pachter (2011). E' importante notare che il dato numerico ottenuto non è una misurazione assoluta del numero di trascritti presenti ma è una misura relativa di espressione, per un limite tecnico della tecnologia. Infatti, le conte dipendono dal numero totale di *reads* che è possibile sequenziare per campione e dalla lunghezza dei geni: più *reads* sono a disposizione, maggiore è la capacità di catturare i trascritti nelle cellule in studio; allo stesso modo, maggiore è la lunghezza di un gene, più sarà elevato il numero di *reads* (Oshlack & Wakefield, 2009) in cui verrà suddiviso e di conseguenza che verranno mappate su quel gene. Nella modellizzazione si tiene conto di queste criticità, ma rimane importante tenerlo a mente nell'interpretazione dei risultati.

Negli ultimi anni, la tecnologia di sequenziamento dell'RNA ha fatto enormi progressi, con nuovi protocolli che hanno permesso di aumentare la risoluzione dei dati. Inizialmente si era in grado di raccogliere i dati di profili di espressione solo a livello di interi gruppi di cellule, ad esempio di un tessuto o di un organismo (Wang et al., 2009). Successivamente si è potuto misurare l'espressione genica di una singola cellula (scRNA-seq), permettendo di individuare e confrontare i diversi tipi cellulari di uno stesso tessuto o stesso organismo (Hwang et al., 2018). Nel 2020 *Nature* ha scelto la trascrittomica spaziale come metodo dell'anno (Marx, 2021). La tecnologia di trascrittomica spaziale permette di catturare la distribuzione spaziale di espressione di migliaia di geni all'interno di un tessuto, fornendo informazioni non solo su quali geni sono espressi e in che quantità, ma anche su dove è stata effettuata la misurazione. La relazione tra le cellule e la loro posizione all'interno del tessuto è fondamentale per una comprensione più profonda di molti meccanismi biologici, come la comunicazione tra cellule e l'interazione tra tumore e microambiente circostante. E' un metodo rivoluzionario che consente agli scienziati di misurare l'attività genica in un campione di tessuto e mappare dove essa sta avvenendo.

Gli autori Righelli et al. (2022) classificano i protocolli di trascrittomica spaziale in

metodi *molecule-based* e metodi *spot-based*. I metodi molecule-based, seqFISH (Lubeck et al., 2014) e metodi simili, come MERFISH (Chen et al., 2015), forniscono l'espressione spaziale di migliaia di trascritti a livello subcellulare, ma le tecnologie necessarie per eseguire questo tipo di esperimenti spaziali sono spesso complesse e costose. I metodi spot-based, come Slide-seq (Rodriques et al., 2019) o la piattaforma 10XGenomics Visium (Rao et al., 2020), hanno una risoluzione significativamente inferiore, ma consentono di misurare quasi l'intero trascrittoma di cellule di un tessuto in modo relativamente semplice. I dati analizzati in questo progetto di tesi sono stati collezionati da 10xGenomics con la loro tecnologia 10xGenomics Visium, che viene descritta nel prossimo paragrafo.

1.1.1 Tecnologia 10X-Visium

La *Visium Spatial Gene Expression* di 10xGenomics è una tecnologia innovativa che permette la mappatura dell'espressione genica a livello quasi cellulare. La procedura di raccolta dei dati si compone di due fasi principali: una prima, in cui viene catturato l'mRNA dal campione di tessuto in studio ed una seconda, in cui i filamenti ottenuti vengono sequenziati. Mentre la seconda fase utilizza tecnologie di sequenziamento comuni, la novità sta nella preparazione della libreria. La tecnologia si basa sul posizionamento della sezione di tessuto su un'area di cattura di dimensioni 6.5mm x 6.5mm, suddivisa in una griglia di celle, ognuna delle quali è detta *spot* e rappresenta una piccola sezione del tessuto. Si riporta in Figura 1.1 un esempio di suddivisione in *spot*del tessuto. Il tessuto viene collocato sul vetrino e, attraverso un processo chimico, le cellule rilasciano l'RNA. In ciascuno *spot*sono presenti circa 200 milioni di catene di nucleotidi che catturano l'RNA liberato e sono dotate di un codice identificativo della cella, noto come *barcode* spaziale. Attraverso la trascrizione, vengono sintetizzati i filamenti di DNA complementare (cDNA) dall'RNA catturato, includendo il *barcode* spaziale, in preparazione per il sequenziamento. Poiché ogni *spot*consiste in una piccola quantità di cellule, le quantità di RNA catturate sono relativamente basse e devono essere amplificate per essere sequenziate con le tecnologie disponibili sul mercato. Tuttavia, l'amplificazione del DNA può distorcere i dati: ad esempio, alcuni geni con espressione bassa potrebbero non essere amplificati, producendo un eccesso di zeri, mentre altri geni con espressione

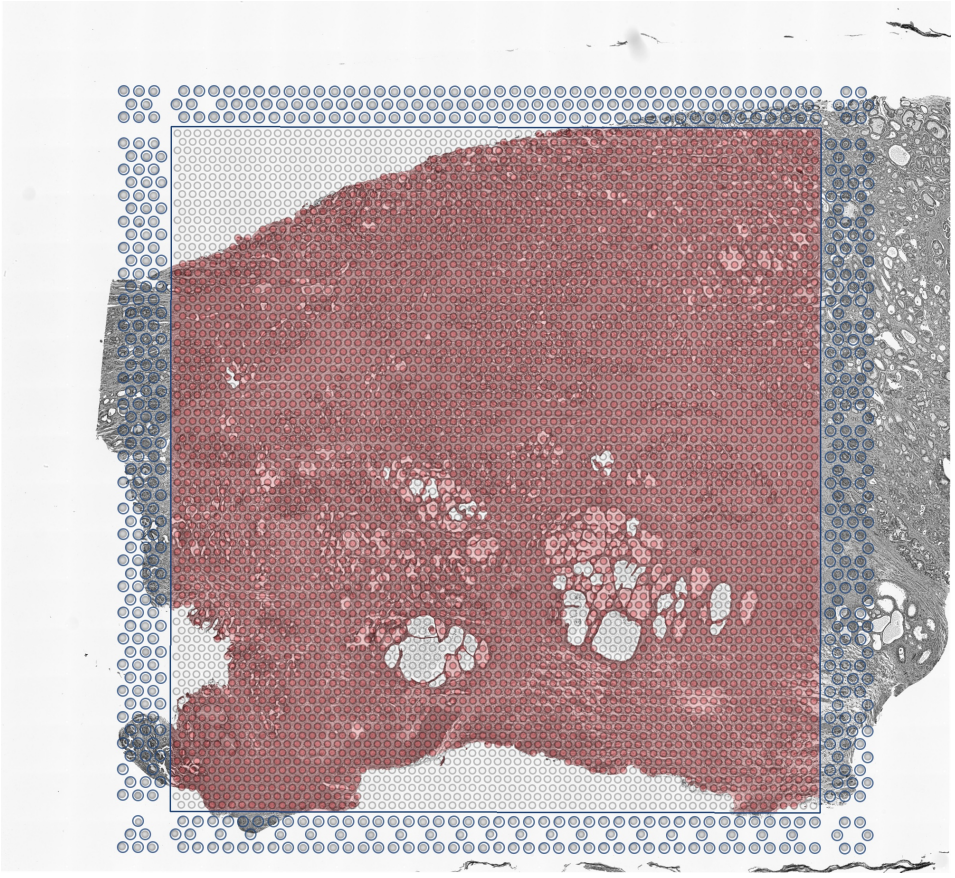


FIGURA 1.1: Campione di tessuto di prostata analizzato con la piattaforma 10X-Visium.

elevata potrebbero essere amplificati in modo eccessivo e diverso a seconda delle loro caratteristiche. Per ridurre questa distorsione, durante la retrotrascrizione, si aggiunge ai filamenti di cDNA una sequenza identificativa univoca, detta UMI (Unique Molecular Identifiers); in questo modo, se due sequenze hanno lo stesso UMI, si sa che sono una copia esatta dell'amplificazione e non due copie indipendenti di RNA. La libreria di filamenti così generata è adatta per l'analisi con sequenziatori di nuova generazione.

Dopo il sequenziamento, le posizioni spaziali delle *reads* possono essere identificate tramite i *barcode* spaziali. A questo punto l'allineamento sul genoma e la quantificazione possono essere effettuati come descritto nella procedura RNA-seq. Si noti che questa tecnologia non fornisce il profilo di espressione di ogni singola cellula nel tessuto, ma quello di tutte le cellule presenti nello *spot*, che possono includere da una a fino a dieci cellule tipicamente.

Supponendo di utilizzare una piattaforma che permette di sequenziare n *spot*e J

geni in K campioni di tessuto, il risultato del sequenziamento consiste in K matrici di conteggi di dimensione $J \times n$ con i geni in riga e gli *spot*in colonna e K matrici di dimensione $n \times 2$ che riportano le coordinate degli n *spot*analizzati. È possibile rappresentare graficamente i dati relativi all'espressione genica su una mappa che ricostruisce la sezione di tessuto, così da apprezzare come i livelli di espressione varino in diverse zone del campione.

1.2 Geni spazialmente variabili

La trascrittomica spaziale consente di analizzare i dati di espressione genica integrando nella modellizzazione anche l'informazione spaziale. Ciò permette di far inferenza sui numerosi processi biologici che dipendono dall'organizzazione delle cellule nel tessuto (Moses & Pachter, 2022).

In termini statistici, si parla di geni spazialmente variabili quando viene rilevata una significativa correlazione tra la distribuzione spaziale delle cellule e la relativa espressione di un determinato gene. Al contrario, quando l'espressione di un gene in uno *spot*è indipendente da quelli circostanti la variabilità spaziale è assente.

In letteratura esistono diverse procedure inferenziali per identificare questo tipo di geni, come SpatialDE (Svensson et al., 2018) Nel caso studio analizzato in questa tesi si è utilizzerà, invece, il metodo nnSVG (Weber et al., 2022) che consente di identificare i geni con livelli di espressione variabili in modo spaziale su tutto il tessuto o all'interno di aree definite a priori, utilizzando un'approssimazione del modello di processo gaussiano basato sui vicini più vicini in grado di catturare la dipendenza spaziale dell'espressione tramite la matrice di covarianza. Adatta il modello per ogni gene e poi ordina i geni in base al valore della statistica di rapporto di log-verosimiglianza per valutare il miglioramento dell'adattamento rispetto al modello lineare classico, in cui la struttura spaziale viene trascurata. Si sottolinea che questo tipo di approccio parametrico assume che il tipo di correlazione dell'espressione del gene e la posizione non cambi all'interno del tessuto, o delle aree in cui viene suddiviso. E' ragionevole pensare che i geni possano avere tipi di distribuzioni spaziali diverse nel tessuto, ad esempio in modo semplicistico, a seconda del tipo cellulare che compone quella zona, o più precisamente della natura istologica.

Identificare i geni che variano spazialmente contribuisce alla comprensione delle interazioni cellulari all'interno del tessuto, dell'effetto di tali interazioni sulla regolazione genica e della possibile deviazione dai normali meccanismi di regolazione genica che possono influenzare i processi biologici in corso nel tessuto. Inoltre, è possibile osservare come tali processi cambiano all'interno del tessuto. L'analisi delle co-variazioni dei livelli di espressione di diversi geni permette di individuare quali geni sono co-regolati, portando alla scoperta di nuove vie metaboliche ed a predire funzioni per geni non ancora annotati (Chen et al., 2015).

1.3 I dati

I dati analizzati nel caso studio di questa tesi sono stati raccolti e resi disponibili dall'azienda 10xGenomics. Si tratta di dati provenienti da una sezione di tessuto di prostata umana con una diagnosi originale di adenocarcinoma. L'adenocarcinoma è un tipo di tumore che si sviluppa nelle ghiandole che producono e secernono liquidi, come la prostata. È un tipo comune di tumore che può invadere i tessuti circostanti e diffondersi ad altre parti del corpo.

Il tessuto è stato processato tramite la piattaforma Visium descritta in precedenza. Il dataset contiene le misurazioni relative a 17931 geni in corrispondenza di 4371 spot; con una media di 5391 geni espressi per *spot*. I valori di espressione sono relativi ai conteggi UMI. Il vetrino corrispondente è stato manualmente annotato dal patologo dott. Esposito dell'Istituto Oncologico Veneto (IOV) tramite l'analisi delle immagini microscopio in cui viene considerata la citoarchitettura delle cellule, ovvero l'organizzazione spaziale e la disposizione delle cellule all'interno del tessuto. Sulla base di queste caratteristiche, il patologo ha suddiviso il tessuto in cinque macro categorie: vasi sanguigni, fibroblasti, ghiandole, stroma e tumore. Si riportano in Figura 4.4 l'immagini al microscopio della sezione del tessuto utilizzata per l'annotazione e la corrispondente annotazione manuale degli spot. Il tessuto stromale è una componente fondamentale del tessuto che circonda e sostiene le ghiandole prostatiche; è composto da tessuti connettivi tra cui cellule muscolari. Lo stroma svolge un ruolo importante nella regolazione della funzione prostatica e della crescita dell'eventuale tumore circostante (Sund & Kalluri, 2009). Inoltre, diversi studi suggeriscono che la composizione dello stroma è un fattore chiave nella prognosi

del tumore della prostata e nelle opzioni terapeutiche disponibili (Krušlin et al., 2015). I fibroblasti del tessuto prostatico sono tipi di cellule che costituiscono la matrice extracellulare nello stroma della prostata; svolgono un ruolo importante nella regolazione della proliferazione delle cellule epiteliali della prostata, dell'angiogenesi, della produzione di fattori di crescita e della risposta infiammatoria (Kalluri & Zeisberg, 2006). La presenza di fibroblasti anormali o l'alterazione del loro comportamento può contribuire allo sviluppo del tumore alla prostata e al suo decorso clinico (Kalluri & Zeisberg, 2006).

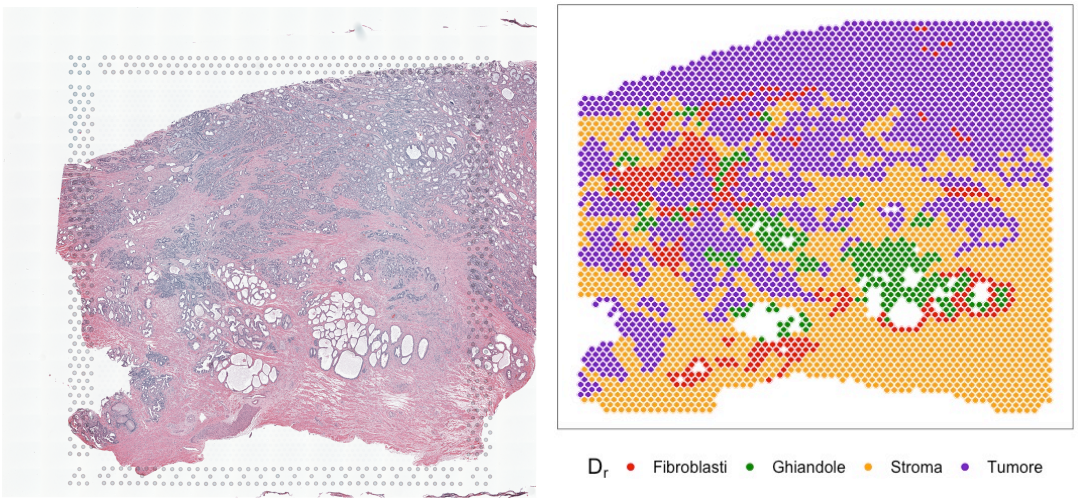


FIGURA 1.2: A sinistra: immagine al microscopio del campione di tessuto di prostata in analisi. A destra: mappa degli *spot*in cui viene suddiviso il tessuto colorati rispetto al cluster di appartenenza definito dall'annotazione patologica manuale.

Capitolo 2

Metodologia statistica

In questo Capitolo si presenta un possibile modello e algoritmo di clustering per dati di espressione genica provenienti da un campione di tessuto processato con metodi di trascrittomica spaziale *spot-based*. Nel metodo proposto si assume che la distribuzione spaziale di ogni gene sia il risultato di un processo stocastico di cui si stima un modello, così da catturare e quantificare ciò che vi è di strutturato nei dati. Il metodo elaborato è una versione semplificata del metodo di co-clustering Spartaco (SPAtially Resolved TrAnscriptomics CO-clustering) proposto da Sottosanti & Risso (2022). A differenza della versione originale, le etichette degli *spot* non sono stimate ma vengono definite sulla base dell'annotazione manuale degli *spot* del tessuto, spesso già disponibile, al fine di semplificare e ridurre drasticamente i tempi calcolo. Infine, viene introdotta un'originale modifica del metodo Spartaco come modello di clustering di geni, mediante l'inserimento di una penalizzazione nella massimizzazione della verosimiglianza del modello, con lo scopo di regolarizzare e di ridurre la varianza delle stime dei parametri.

I dati sono organizzati in una matrice con in riga i geni e in colonna gli *spot*, i cui elementi rappresentano il livello di espressione del gene di riga nello *spot* in colonna. Si noti che in questo contesto, il termine "spot" si riferisce alle celle della griglia in cui il tessuto viene suddiviso durante l'analisi. I "geni" sono le variabili misurate in ogni *spot*, rispetto alla terminologia utilizzata nei protocolli 10xVisium. Tuttavia, il metodo presentato è adatto anche a dati raccolti con altre tecnologie di trascrittomica spaziale *spot-based* e, più in generale, a qualsiasi dataset in cui le righe o le colonne sono osservazioni multivariate misurate in corrispondenza di una struttura spaziale.

2.1 Introduzione al metodo Spartaco

I metodi di clustering tradizionali raggruppano osservazioni basandosi sulla similarità delle loro caratteristiche. Tuttavia, per grandi insiemi di dati, come in trascrittomica, il numero di variabili è notevolmente elevato e queste tecniche diventano meno efficaci per discernerne la struttura. Per questo motivo, è importante "sintetizzare" anche le caratteristiche, il che può essere fatto riunendole in cluster, in parallelo al consueto raggruppamento delle osservazioni. Questo è l'obiettivo del co-clustering: data una matrice di dati, raggruppare contemporaneamente le colonne e le righe per individuare insiemi di osservazioni e insiemi di caratteristiche tale che tutti gli elementi all'interno del sottoinsieme siano simili (Tan & Witten, 2014). La matrice di dati, inizialmente di grandi dimensioni, può essere quindi sintetizzata da un numero limitato di blocchi, detti co-cluster, che risultano dalla combinazione di cluster di righe e cluster di colonne. I vantaggi di questo approccio al clustering sono due:

1. permette di identificare gruppi di elementi che altrimenti non potrebbero essere trovati calcolando la similarità sull'intero insieme di variabili a disposizione,

2. identifica quali caratteristiche sono rilevanti nella definizione di ogni cluster e facilita quindi l'interpretazione dei risultati.

Come per i metodi di clustering, esistono molte tecniche per eseguire il co-clustering. In letteratura si distinguono i metodi di co-clustering deterministici e quelli basati su modello. Quest'ultimi eseguono contemporaneamente sia la procedura di clustering che la ricostruzione del processo generativo probabilistico dei dati. La maggior parte della letteratura sulle tecniche di co-clustering basate su modello si basano su un modello chiamato Latent Block Model (Govaert & Nadif, 2013), denominato "LBM". Si tratta di un modello che amplia il modello di mistura standard quando si ipotizza che sia le righe che le colonne della matrice di dati derivino da cluster sottostanti non noti.

Il modello LBM si basa sull'ipotesi che le osservazioni all'interno dello stesso co-cluster siano indipendenti. Anche se questa ipotesi è vantaggiosa a livello computazionale, non è compatibile con l'evidenza di forti livelli di correlazione presenti nei dati di espressione genica tra cellule vicine (Efron, 2009). Il modello proposto da Sottosanti & Risso (2022) supera quest'ipotesi di indipendenza condizionata utilizzando un modello

LBM assume che ogni blocco si distribuisca come una distribuzione Gaussiana matriciale (Gupta & Nagar, 1999), la cui matrice di covarianza delle colonne è non-diagonale per tenere conto della dipendenza dell'espressione di un gene tra gli *spot* del blocco. Infatti, è ragionevole pensare che l'espressione di un certo gene possa dipendere da uno stesso fenomeno biologico che coinvolge cellule adiacenti. La sua caratteristica distintiva, rispetto ad altri metodi che utilizzano la stessa strategia, è che tale correlazione viene espressa in funzione della distanza tra le posizioni spaziali degli *spot* in cui viene misurata l'espressione genica. Di conseguenza, Spartaco divide la matrice dei dati in rettangoli sulla base delle medie, delle varianze e delle covarianze spaziali stimate. Inoltre, il modello include anche un effetto casuale gene-specifico per cogliere l'eventuale varianza residua con gli altri geni del blocco non spiegata dalla struttura spaziale comune.

Uno dei problemi principali dei modelli LBM è legato alla stima dei parametri. La massimizzazione della funzione di verosimiglianza è complessa e produce in molti casi infinite soluzioni per cui lo stimatore di massima verosimiglianza potrebbe non esistere, almeno a livello globale (Novais & Faria, 2021). L'algoritmo di Expectation-Maximization (EM) risolve questo problema, ma può essere lento nella convergenza e la sua efficienza dipende dai valori iniziali dei parametri per funzionare correttamente. Per superare i problemi dell'algoritmo EM, ne sono state presentate alcune generalizzazioni. Nell'ambito dei modelli di clustering, un esempio è l'algoritmo Classification EM (CEM) di Celeux & Govaert (1992) che utilizza solo una parte dei dati per calcolare le stime dei parametri, e non tutti i dati come nell'algoritmo EM. Inoltre, rispetto all'algoritmo EM, l'algoritmo CEM esegue un'operazione di classificazione delle osservazioni contemporaneamente alla stima dei parametri. In altre parole, mentre l'algoritmo EM si concentra solo sulla stima dei parametri del modello, l'algoritmo CEM consente anche di determinare a quale cluster appartiene ogni osservazione, un aspetto cruciale nell'ambito del co-clustering dove la classificazione delle righe e colonne è altrettanto di interesse come della stima dei parametri del modello. I dettagli dell'algoritmo CEM sono descritti nel paragrafo 2.2.2 di questa tesi. Nel caso del modello Spartaco, l'algoritmo richiede una modifica per l'aggiornamento delle etichette di colonna. Infatti, poiché queste sono correlate non è possibile scrivere esplicitamente la distribuzione condizionata richiesta

nel passo di classificazione. Per questo motivo, gli autori propongono di campionare la distribuzione condizionata tramite una catena di Markov e di eseguire un'allocazione stocastica delle etichette (passo SE) delle colonne. Questo problema non si presenta invece per la classificazione delle righe in gruppi, poiché si suppone siano indipendenti. L'algoritmo di stima alterna quindi: una mossa di classificazione (passo CE), una mossa di allocazione stocastica (passo SE) e una mossa di massimizzazione (passo M); gli autori lo chiamano algoritmo classification-stochastic EM (CS-EM). L'esecuzione di questa procedura presenta un elevato costo computazionale. Sottosanti & Risso (2022) sottolineano che è proprio il passo SE a rallentare l'algoritmo perchè richiede, ad ogni aggiornamento delle etichette di colonna, l'inversione delle matrici di covarianza di ogni co-cluster, che possono avere dimensioni molto elevate.

Con l'obiettivo di ridurre i lunghi tempi di calcolo del metodo Spartaco, in questa tesi si propone una sua versione semplificata. I dati in analisi sono gli stessi, anche la struttura del modello è la stessa ma ora si assumono note a priori le etichette di colonna e non più da determinare. Questa ipotesi permette di utilizzare l'algoritmo CEM per la stima del modello, eliminando il passo SE dall'algoritmo. In pratica, il metodo Spartaco diventa un metodo di clustering dei geni semi-supervisionato: è non supervisionato perché i geni vengono raggruppati sulla base di una similarità di distribuzione, e tale confronto non avviene tra tutti gli *spot* ma tra aree di *spot* con stessa etichetta.

Va evidenziato che richiedere la definizione a priori delle etichette per gli *spot* non costituisce una limitazione all'applicazione del metodo proposto. Al contrario, per questo tipo di dati sono spesso già disponibili annotazioni manuali che rappresentano informazioni indipendenti dai dati di espressione e costituiscono una fonte di conoscenza a priori. Grazie a questa nuova implementazione, il metodo è in grado di integrare queste informazioni nel processo di modellizzazione dei dati. Per valutare la robustezza del metodo all'errata specificazione delle etichette si esegue uno studio di simulazione. In questa versione semplificata del metodo, viene assunto che esistono gruppi di geni che presentano una distribuzione spaziale simile all'interno delle aree annotate, che sono definite a priori e non più calcolate sulla base dei dati. Quest'ipotesi può non essere vera e, inoltre, rappresenta una possibile criticità del metodo. Infatti potrebbero esserci *spot*

classificati in gruppi distinti nell'annotazione manuale ma coinvolti in uno stesso processo biologico, che presentano una similitudine nell'espressione di alcuni geni. Tuttavia, dal momento che il metodo confronta le distribuzioni dei geni solo all'interno delle aree definite dall'annotazione, non è in grado di rilevare queste similitudini potenzialmente interessanti.

Si noti, tuttavia, che i vantaggi del modello rimangono: i cluster di geni vengono creati confrontando un comportamento locale della distribuzione e non globale, in cui le similarità potrebbero non essere colte, e identifica quali gruppi di geni distinguono le aree di tessuto annotate. Semplicemente, le informazioni che si possono rilevare sono limitate a questa configurazione degli *spot*.

Nel caso studio presentato in questa tesi si utilizzerà come classificazione degli *spot* l'annotazione manuale fornita dal dott. Esposito dell'Istituto Oncologico Veneto, ottenuta dall'analisi delle immagini del tessuto a microscopio. L'annotazione patologica degli *spot* consiste nell'identificazione e nell'etichettatura in base alle caratteristiche morfologiche e biochimiche delle cellule che lo compongono.

2.2 Formulazione del modello

Come nel metodo Spartaco, anche nella versione proposta in questa tesi si suppone l'esistenza di una struttura a blocchi latente della matrice dei dati, dove ogni blocco segue una distribuzione Gaussiana matriciale. Si suppone, dunque, che i dati di espressione ottenuti dal sequenziamento siano già stati propriamente processati in modo tale da passare dalla scala naturale di conteggio ad una scala continua. I possibili metodi di trasformazione dei dati sono brevemente illustrati nella Sezione 4.1 del Capitolo 4. Sia $\mathbf{X} = (x_{ij})_{1\leq i\leq n, 1\leq j\leq p}$ la matrice di dimensioni $n \times p$ contenente l'espressione degli n geni rilevati nella griglia di p *spot* del tessuto analizzato, con $x_{ij} \in \mathbb{R}$ per ogni i e j. Sia $\mathrm{s}_j = (s_{jx}, s_{jy})$ la posizione dello *spot* j-esimo sulla superficie del vetrino, con $s_{jx}, s_{jyx} \in \mathbb{R}$, e $\mathbf{S} = (s_j)_{1\leq j\leq p}$ la matrice $p \times 2$ delle coordinate spaziali dei p *spot*.

Si assume che esistono K cluster di righe di X e R cluster di colonne di X, corrispondenti a una suddivisione della matrice in KR co-cluster rettangolari e distinti. Tale suddivisione comporta che le etichette di riga e di colonna siano assegnate in modo

indipendente le une dalle altre e che ogni elemento della matrice fa parte di un solo blocco. La non sovrapposizione dei co-cluster può rappresentare una limitazione in alcuni contesti, ma rende la modellizzazione più semplice e l'interpretazione dei risultati più agevole, come evidenziato da Tan & Witten (2014). Questa struttura può essere identificata tramite due vettori: siano $\boldsymbol{\mathcal{Z}_i} = (\mathcal{Z}_i)_{1\leq i\leq n}$ e $\boldsymbol{\mathcal{W}} = (\mathcal{W}_j)_{1\leq 1\leq j\leq p}$ che denotano, rispettivamente, il cluster di appartenenza delle righe e delle colonne. In questa versione del metodo Spartaco si suppone che le etichette di colonna $\boldsymbol{\mathcal{W}} = (\mathcal{W}_j)_{1\leq 1\leq j\leq p}$ siano note, ad esempio nel caso studio di questa tesi sono definite dall'annotazione patologica degli *spot*, come anticipato nel paragrafo precedente. Invece, le etichette di riga $\boldsymbol{\mathcal{Z}_i} = (\mathcal{Z}_i)_{1\leq i\leq n}$ sono considerate realizzazioni di una variabile casuale latente.

Si specifica un modello per ogni blocco della matrice e si ipotizza che non ci sia dipendenza tra le osservazioni appartenenti a blocchi diversi: geni/*spot* in un co-cluster sono correlati solo con geni/*spot* dello stesso co-cluster. Siano $\mathcal{C}_k = \{i = 1, \ldots, n : \mathcal{Z}_i = k\}$ il k-esimo cluster di riga, con $k = 1, \ldots, K$, e $\mathcal{D}_r = \{j = 1, \ldots, p : \mathcal{W}_j = r\}$ l' r-esimo cluster di colonna, con $r = 1, \ldots, R$. Si indicano con $n_h = |\mathcal{C}_k|$ e $p_r = |\mathcal{D}_r|$ le dimensioni del cluster di riga $\mathcal{C}_k$ e di colonna $\mathcal{D}_r$. Sia $\mathbf{X}^{kr} = (x_{ij})_{i\in\mathcal{C}_k, j\in\mathcal{D}_r}$ la sottomatrice di X contiene le osservazioni del kr-esimo blocco; $\mathbf{X}^{k\cdot} = (x_{ij})_{i\in\mathcal{C}_k, 1\leq j\leq p}$, di dimensioni $n_k \times p$, formata da tutte le righe in $\mathcal{C}_k$, e $\mathbf{X}^{\cdot r} = (x_{ij})_{1\leq i\leq n, j\in\mathcal{D}_r}$ la matrice di dimensioni $n \times p$ di tutte le colonne in $\mathcal{D}_r$. Sia $\mathbf{x}_{i\cdot}^{kr}$ il vettore contenente l'espressione del i-esimo gene nel cluster $\mathcal{C}_k$ tra i p_r *spot* nel cluster $\mathcal{D}_r$. Si definisce un modello per l'espressione $\mathbf{x}_{i\cdot}^{kr}$ del gene i negli *spot* del blocco kr:

$$\mathbf{x}_{i\cdot}^{kr} = \mu_{kr}\mathbf{1}_{p_r} + \sigma_{kr,i}\boldsymbol{\epsilon}_{i\cdot}^{kr}, \quad \boldsymbol{\epsilon}_{i\cdot}^{kr} \sim N_{p_r}(\mathbf{0}, \boldsymbol{\Delta}_{\mathbf{kr}}) \tag{2.1}$$

$$\boldsymbol{\Delta}_{\mathbf{kr}} = \tau_{kr}\mathcal{K}(\mathbf{S}_r; \boldsymbol{\phi}_r) + \xi_{kr}\mathbb{1}_{pr}, \tag{2.2}$$

dove μ_{kr} è un parametro scalare che rappresenta il valore medio di espressione nel blocco, $\mathbf{1}_{p_r}$ è un vettore di uni di dimensione p_r, $\sigma^2_{kr,i}$ è un parametro di varianza gene-specifico e $\boldsymbol{\Delta}_{kr}$ è la matrice di covarianza degli *spot*.

Nell'ambito della statistica spaziale e in altri modelli per dati di trascrittomia spaziale, è comune modellare la variabilità dell'espressione $\boldsymbol{\Delta}_{kr}$ come combinazione lineare di due termini: una matrice indentità $\mathbb{1}_{pr}$ di dimensione p_r e una matrice $\mathcal{K}(\mathbf{S}_r; \boldsymbol{\phi}_r) =$

$\left(k(\|\mathbf{s}_r^r - \mathbf{s}_{j\prime}^r\|; \boldsymbol{\phi}_r)\right)_{1\leq j,j\prime\leq p_r}$. La matrice $\mathcal{K}$ modella la correlazione dell'espressione tra *spot* ed è costruita utilizzando una funzione di covarianza spaziale isotropa $k(\cdot;\phi_r)$, dove ϕ_r è un parametro di scala e $\mathbf{S}^r = (s_j)_{j\in\mathcal{D}_r}$ è una sotto-matrice di S contenente le coordinate spaziali relative agli *spot* nel cluster $\mathcal{D}_r$. Il termine isotropo indica che il valore della funzione dipende solo dalla distanza tra le due posizioni $\left\|\mathbf{s}_j^r - \mathbf{s}_{j'}^r\right\|$; dunque la matrice di covarianza $\boldsymbol{\Delta}_{kr}$ è a-direzionale e indipendente dal valore dell'espressione del gene in quelle posizioni. La scelta della funzione sarà discussa più avanti. I parametri positivi τ_r e ξ_{kr} nella Formula(2.1) regolano la combinazione lineare tra la matrice $\mathcal{K}$ e la matrice $\mathbb{1}_{pr}$: il primo misura la dipendenza spaziale dei dati, il secondo è comunemente indicato in letteratura come *nugget effect* e misura la variabilità residua dell'espressione indipendente dalla posizione. Seguendo l'approccio adottato nel metodo Spartaco si considera la stessa funzione di covarianza $k(\cdot,\phi_r)$ per ogni cluster di colonne $\mathcal{D}_r$ con parametro ϕ_r specifico per cluster di *spot*.

La presenza di un parametro di varianza gene-specifico $\sigma^2_{kr,i}$ nel modello (2.1) permette di tenere conto dell'eventuale variabilità non spiegata dalla matrice di covarianza. Infatti, si noti che i parametri di media μ_{kr} e di covarianza $\boldsymbol{\Delta}_{kr}$ sono comuni a tutti i geni del co-cluster kr, e perciò descrivono soltanto il loro comportamento comune ai geni nei p_r *spot* di tale cluster. Questa variabilità ulteriore può essere legata sia ad un comportamento specifico del gene ma anche alla presenza di una dipendenza tra i geni. Il modello finora descritto considera solo la dipendenza dell'espressione di un gene in uno *spot* dalla sua espressione degli *spot* vicini, tuttavia è possibile che sia presente una dipendenza di espressione tra geni diversi nello stesso co-cluster, ad esempio perchè coinvolti nello stesso fenomeno biologico in atto nel gruppo di cellule. Il problema di modellare osservazioni con una dipendenza sistematica è comunemente affrontato nell'ambito dell'analisi di dati longitudinali, dove si utilizzano modelli ad effetti casuali. In linea con questo approccio, si assume che ogni $\sigma^2_{kr,i}$ sia realizzazione di una distribuzione gamma inversa $\mathcal{IG}(\alpha_{kr},\beta_{kr})$, dove α_{kr} e β_{kr} sono i parametri di forma e tasso. La gamma inversa ha come supporto i valori reali positivi ed è coniugata con la distribuzione gaussiana, risulta quindi adatta a modellare i valori del parametro $\sigma^2_{kr,i}$ e permette di ricavare esplicitamente l'espressione della densità di probabilità marginale di $\mathbf{x}_{i\cdot}^{kr}$:

$$f\left(\mathbf{x}_{i\cdot}^{kr};\boldsymbol{\theta}_{kr},\boldsymbol{\phi}_r\right)=\frac{1}{\sqrt{(2\pi)^{p_r}\det\left(\boldsymbol{\Delta}_{kr}\right)}}\frac{\Gamma\left(\alpha_{kr,i}^{*}\right)}{\Gamma\left(\alpha_{kr}\right)}\frac{\beta_{kr}^{\alpha_{kr}}}{{\beta_{kr,i}^{*}}^{\alpha_{kr,i}^{*}}} \tag{2.3}$$

dove $\det(\cdot)$ indica l'operatore determinante matriciale, $\alpha_{kr,i}^{*}=p_r/2+\alpha_{kr}$ e $\beta_{kr,i}^{*}=(\mathbf{x}_{i\cdot}^{kr}-\mu_{kr}\mathbf{1}_{p_r})\boldsymbol{\Delta}_{kr}^{-1}(\mathbf{x}_{i\cdot}^{kr}-\mu_{kr}\mathbf{1}_{p_r})/2+\beta_{kr}$. L'insieme dei parametri del modello specifico per il (k,r)-esimo co-cluster è dunque $\boldsymbol{\theta}_{kr}=\{\mu_k,\tau_r,\xi_{kr},\alpha_{kr},\beta_{kr}\}$, mentre ϕ_r è un parametro associato all'intero r-esimo cluster di colonne.

Condizionatamente al fatto che i geni appartenenti allo stesso cluster hanno un parametro di varianza con distribuzione a priori comune, essi si possono considerare indipendenti tra loro. Con questa assunzione, il modello può essere riformulato in termini di distribuzione di probabilità sull'intero blocco kr-esimo, $\mathbf{X}^{kr}\mid\Sigma_{kr}\sim\mathcal{MVN}\left(\mu_{kr}\mathbf{1}_{n_k x p_r},\Sigma_{kr},\Delta_{kr}\right)$, dove $\mathcal{MVN}$ denota la distribuzione normale matrice-variata e $\Sigma_{kr}=\operatorname{diag}\left(\sigma_{kr,1}^2,\ldots,\sigma_{kr,n_k}^2\right)$ è la matrice di covarianza (diagonale) dei geni. La scrittura del modello in termini di distribuzione normale matrice-variata implica che ogni riga, colonna e sottomatrice di X_{kr} è distribuita secondo una distribuzione gaussiana. Pertanto, il modello può essere formulato in modo equivalente in termini di colonne:

$$\mathbf{x}_{\cdot j}^{kr}\mid\boldsymbol{\Sigma}_{kr}\sim N_{n_k}\left\{\mu_{kr}\mathbf{1}_{n_k},\left(\tau_{kr}+\xi_{kr}\right)\boldsymbol{\Sigma}_{kr}\right\},\quad \operatorname{Cov}\left(\mathbf{x}_{\cdot j}^{kr},\mathbf{x}_{\cdot j'}^{kr}\right)=\tau_{kr}k\left(\|\mathbf{s}_j^r-\mathbf{s}_{j'}^r\|;\phi_r\right)\Sigma_{kr},$$

con $j,j'\in\mathcal{D}_r$.

2.3 Inferenza

2.3.1 Identificabilità

Il modello descritto nella Formula(2.1) non è identificabile nei parametri di covarianza. Infatti, per qualsiasi $a>0$, si ha che $\sigma_{kr,i}^2\cdot\boldsymbol{\Delta}_{kr}=a\cdot\sigma_{kr,i}^2\cdot\boldsymbol{\Delta}_{kr}/a=\tilde{\sigma}_{kr,i}^2\cdot\tilde{\boldsymbol{\Delta}}_{kr}$. Questo produce, nella pratica, un numero illimitato di soluzioni per la stima dei parametri.

Un modo comune per risolvere questo problema è assegnare a priori un valore ad alcuni parametri di covarianza. Nel modello in questione, bisognerebbe fissare $\sigma_{kr,i}^2=c$, per un $i\in\{1\ldots,n_k\}$, dove c è una costante positiva arbitraria, per ogni co-cluster kr. Tuttavia, quali geni appartengano al cluster $\mathcal{C}_k$ su cui applicare il vincolo per il co-cluster

kr non sono noti prima della stima, poiché le righe della matrice sono coinvolte in un processo di clustering; il che rende questa soluzione non praticabile.

Sottosanti & Risso (2022) hanno adottato una soluzione che pone il vincolo di identificabilità su $\boldsymbol{\Delta}_{kr}$: si può dimostrare che $\text{tr}\left(\boldsymbol{\Delta}_{kt}\right) = p_r\left(\tau_{kr} + \xi_{kr}\right)$ e si può dunque fissare la quantità $\left(\tau_{kr} + \xi_{kr}\right) = c_{\boldsymbol{\Delta}}$, dove $c_{\boldsymbol{\Delta}}$ è una costante positiva arbitraria. Tale vincolo ha anche un'importante conseguenza pratica: infatti, una volta che la stima $\hat{\tau}_{kr}$ è determinabile all'interno del dominio vincolato $(0, c_{\boldsymbol{\Delta}})$, allora $\hat{\xi}_{kr}$ si ottiene per differenza come $\hat{\xi}_{kr} = c_{\boldsymbol{\Delta}} - \hat{\tau}_{kr}$. Quindi, la stima $\hat{\tau}_{kr}$ e la stima $\hat{\xi}_{kr}$ sono interpretabili solo in relazione l'una all'altra e non in termini assoluti. Il termine $\frac{\hat{\tau}_{kr}}{\hat{\xi}_{kr}}$ è chiamato nell'articolo di Sottosanti & Risso (2022) *spatial signal-to-noise ratio*, e rappresenta la quantità di variazione di espressione dei geni in un cluster dovuta alla dipendenza spaziale rispetto alla variabilità residua. Una variazione elevata nell'espressione genica che è largamente spiegata dalla componente spaziale può indicare un'importante interazione tra le cellule di co-cluster kr.

2.3.2 Stima del modello con algoritmo semi-supervisionato

Per la stima del modello si ricorre alla massimizzazione della log-verosimiglianza di classificazione, che risulta avere la seguente espressione:

$$\log \mathcal{L}\left(\boldsymbol{\Theta}, \mathcal{Z} | \mathbf{X}, \mathcal{W}\right) = \sum_{i=1}^{n} \sum_{k=1}^{K} \mathbb{1}\left(\mathcal{Z}_i = k\right) \left\{ \sum_{r=1}^{R} \log f\left(\mathrm{x}_{i.}^{r}\right); \boldsymbol{\theta}_{kr}, \phi_r \right\} \tag{2.4}$$

dove $\boldsymbol{\Theta} = \bigcup_r \left\{ \bigcup_k \boldsymbol{\theta}_{kr}, \phi_r \right\}$, $\mathrm{x}_{i.}^{r}$ è la riga i-esima riga della matrice $\mathrm{X}^{\cdot r}$ e $f(\cdot;\cdot)$ è data dalla Formula(2.3). Si ricorda che il vettore $\boldsymbol{\mathcal{W}}$ delle etichette di colonna è supposto noto, mentre la componente latente del modello è data delle etichette di riga $\boldsymbol{\mathcal{Z}}$.

Come anticipato nella Sezione introduttiva di questo Capitolo, l'algoritmo di Classificazione EM (CEM), sviluppato da (Govaert & Nadif, 2013), permette una massimizzazione della log-verosimiglianza di classificazione che apparte in Formula(2.5). L'algoritmo funziona iterativamente: alterna un passo di classificazione (passo CE), in cui vengono aggiornate la stima delle etichette di riga $\boldsymbol{\mathcal{Z}}$, e un passo di massimizzazione (passo M), che aggiorna le stime dei parametri di $\boldsymbol{\Theta}$. Nel passaggio di classificazione, il CEM stima la probabilità a posteriori che una data osservazione sia generata da una

delle componenti della mistura, per ogni componente. Questa stima viene calcolare dalla distribuzione condizionata ottenuta della verosimiglianza di classificazione completa, che tiene conto non solo per della distribuzione delle osservazioni, ma anche della distribuzione delle etichette di clustering. Nel passaggio di massimizzazione, il CEM massimizza la verosimiglianza completa di classificazione per aggiornare i parametri della distribuzione delle misture. Questo processo viene iterato fino a quando la verosimiglianza non raggiunge un massimo o il processo non converge ad un punto stazionario.

Si noti che l'algoritmo CEM, nella sua versione originale, utilizza la verosimiglianza completa di classificazione del modello, che tiene conto sia della distribuzione delle osservazioni che della distribuzione delle etichette di clustering. Tuttavia, poiché qui come in Spartaco si assume implicitamente che $\Pr(\mathcal{Z}_i = k) = 1/K$ per qualsiasi k, non c'è alcuna differenza pratica tra la verosimiglianza di classificazione completa e la verosimiglianza di classificazione riportata in (2.5).

Si indichi con $(\boldsymbol{\Theta}, \boldsymbol{\mathcal{Z}})^{(t-1)}$ la stima dei parametri del modello e il vettore delle etichette di riga all'iterazione $t-1$. Al passo t, l'algoritmo esegue i seguenti passi:

- **Passo CE**: mantenendo fissi i parametri in $\boldsymbol{\Theta}^{(t-1)}$, calcola le probabilità a posteriori per ciascun cluster k di geni (E step), e poi aggiorna le etichette assegnando al ciascun gene ad il cluster con la massima probabilità a posteriori:

$$\mathcal{Z}_i^{(t)} = \underset{k=1,\dots,K}{\operatorname{argmax}} \frac{\prod_{r=1}^{R} f\left(\mathrm{x}_{i.}^{r}; \boldsymbol{\theta}_{kr}^{(t-1)}, \phi_r^{(t-1)}\right)}{\sum_{k'=1}^{K} \left\{\prod_{r=1}^{R} f\left(\mathrm{x}_{i.}^{r}; \boldsymbol{\theta}_{k'r}^{(t-1)}, \phi_r^{(t-1)}\right)\right\}}, \quad i = 1, \dots, n.$$

- **Passo M**: utilizzando le osservazioni delle righe in $\mathcal{C}_k^{(t)}$ al passo t, si aggiornano le stime dei parametri $\boldsymbol{\theta}_{kr}^{(t)}$ e $\phi_r^{(t)}$ tramite massimizzazione della verosimiglianza. Le derivate rispetto a $(\boldsymbol{\theta}_{kr}, \phi_r)$ non ammettono una soluzione esplicita per l'aggiornamento dei parametri del modello, per cui è necessario risolvere le equazioni di verosimiglianza per via numerica. Si utilizza l'algoritmo di Nelder & Mead (1965) implementato nel pacchetto `R stats`.

Si iterano questi due steps fino a convergenza della log-verosimiglianza, ottenendo così le stime finali di $(\hat{\boldsymbol{\Theta}}, \hat{\boldsymbol{\mathcal{Z}}})$ dove 2.5 è massima, almeno localmente.

Il metodo utilizzato per creare i co-cluster ha anche un'interpretazione geometrica. Analogamente a come l'algoritmo *k-means* minimizza la distanza euclidea tra le osservazioni e i centroidi, il metodo Spartaco e anche questa sua versione, minimizza la distanza di Mahalanobis tra le osservazioni e i centroidi dei blocchi, integrando la struttura spaziale dei dati nella matrice di covarianza. Matematicamente, la distanza di Mahalanobis si calcola come la radice quadrata della differenza tra due vettori, moltiplicata per l'inversa della matrice di covarianza. Questa distanza è utile per rilevare la separazione tra due gruppi di dati quando le variazioni in ogni dimensione non sono indipendenti. Questo garantisce la validità del modello in quanto algoritmo basato sulla distanza tra cluster, anche quando i dati non soddisfano pienamente le ipotesi probabilistiche.

2.3.3 Versione penalizzata del metodo

Si presenta una nuova versione del metodo di clustering che si basa sulla massimizzazione della verosimiglianza penalizzata del modello al fine di migliorare la stabilità delle stime dei parametri. La formazione dei gruppi di geni è basata sul confronto delle distribuzioni di espressione nei vari blocchi della matrice dei dati. Tuttavia, se le stime dei parametri di questi blocchi sono troppo variabili, ciò potrebbe compromettere l'individuazione della corretta configurazione in cluster dei geni. Inoltre, l'interpretazione dei cluster di geni individuati sono principalmente guidate dalle stime dei parametri di media e *spatial signal-to-noise ratio*, ed è dunque importante che queste siano affidabili.

Si considera quindi una stima penalizzata per i parametri di media e di *spatial signal-to-noise ratio* di ciascun co-cluster della matrice dei dati. Per il parametro di media di blocco μ_{kr} si utilizza una penalizzazione di tipo *ridge* (Hastie et al., 2009), poiché dal punto di vista biologico non ha un particolare significato trovare una media di blocco pari a zero. Per il parametro *spatial signal-to-noise ratio*, invece, si adotta una penalizzazione di tipo *lasso* (Tibshirani, 1996). Questa scelta è motivata dal fatto che è di interesse stimare a zero i valori di *spatial signal-to-noise ratio* effettivamente nulli, poiché questi indicano che il corrispondente co-cluster è privo di effetto di spazialità. Il modello viene stimato con lo stesso algoritmo descritto nel paragrafo precedente (2.3.2), dove la verosimiglianza di classificazione (2.5) viene sostituita dalla sua versione penalizzata:

$$\log \mathcal{L}\left(\boldsymbol{\Theta}, \mathcal{Z}|\mathbf{X}, \mathcal{W}\right) =$$
$$\sum_{i=1}^{n}\sum_{k=1}^{K} \mathbb{1}\left(\mathcal{Z}_i = k\right)\left\{\sum_{r=1}^{R} \log f\left(\mathrm{x}_{i,}^{r}; \boldsymbol{\theta}_{kr}, \phi_r\right)\right\} - \lambda_\mu \sum_{k=1}^{K}\sum_{r=1}^{R} \mu_{kr}^2 - \lambda_\tau \sum_{k=1}^{K}\sum_{r=1}^{R} |\tau_{kr}| \quad (2.5)$$

dove λ_μ e λ_τ sono parametri scalari non-negativi.

Nella versione penalizzata del metodo di clustering ci sono tre parametri che richiedono di essere fissati: il numero di cluster di riga K e i due parametri di penalizzazione, λ_{mu} e λ_{tau}. Per scegliere i valori ottimali dei parametri, si potrebbe utilizzare una procedura di cross-validation o generalizzare l'approccio utilizzato per la selezione di K per una selezione congiunta. Tuttavia, per il fine di regolarizzare le stime, si ritiene sufficiente una piccola penalità, e pertanto si è deciso di fissare i valori di λ_μ e λ_τ a priori rispettivamente a 1.5 e 0.3.

2.3.4 Misura di bontà del modello

Come per l'algoritmo EM, anche l'algoritmo CEM per la stima del modello non assicura la convergenza ad un punto di massimo globale, per questo motivo è importante eseguire più volte la procedura a partire da diverse posizioni iniziali. Nel contesto del clustering, si possono sfruttare le esecuzioni parallele per definire una misura di incertezza della struttura di co-clustering. In particolare, se le stime parallele generano cluster di righe simili, questo è un'evidenza a favore dell'esistenza di una singola configurazione a blocchi nei dati determinata da tale partizione stimata delle righe e la partizione predefinita delle colonne. Al contrario, se non c'è una chiara struttura di co-clustering nei dati, le diverse esecuzioni dell'algoritmo tenderanno a produrre soluzioni differenti ma egualmente valide. Si noti che in questo caso la valutazione è rispetto alla bontà del clustering di righe, ma questa dipende dalla partizione scelta per le colonne: se si ottengono più configurazioni di cluster di riga ugualmente probabili, questo significa che con tale suddivisione delle colonne non vi è una chiara suddivisione delle righe, ma non esclude che ne possa esistere una con una diversa partizione delle colonne. In questo senso si parla di evidenza a favore di una struttura di co-clustering. Per valutare la

precisione del metodo di clustering Sottosanti & Risso (2022) propongono una misura basata sull'indice CER, Clustering Error Rate, di Witten & Tibshirani (2010).

Si suppone di eseguire l'algoritmo di stima S volte sullo stesso insieme di dati e con etichette di colonna $\boldsymbol{\mathcal{W}}$: sia $(\hat{\boldsymbol{\Theta}}^{(s)}, \hat{\boldsymbol{\mathcal{Z}}}^{(s)})$ la stima dei parametri restituita dall'esecuzione s-esima, per $s = 1, \ldots, S$, e $\ell^s = \log \mathcal{L}(\hat{\boldsymbol{\Theta}}^{(s)}, \hat{\boldsymbol{\mathcal{Z}}}^{(s)}, \boldsymbol{\mathcal{W}})$. Si denoti con $s^* = \text{argmax}_s\, \ell^s$: $(\hat{\boldsymbol{\mathcal{Z}}}^{(s^*)}, \boldsymbol{\mathcal{W}})$ la partizione in blocchi con valore massimo della log-verosimiglianza di classificazione tra le S esecuzioni e sarà pertanto la soluzione finale restituita dall'algoritmo. Una misura di incertezza del co-clustering può essere definita in funzione delle distanze dalla configurazione ottimale $(\hat{\boldsymbol{\mathcal{Z}}}^{(s^*)}, \boldsymbol{\mathcal{W}})$ e delle altre stimate con evidenza inferiore $(\hat{\boldsymbol{\mathcal{Z}}}^{(s)}, \boldsymbol{\mathcal{W}})$, per $s \neq s^*$. Sia $\mathcal{I}_k = \mathbb{1}(\hat{\mathcal{Z}}_i^{(s^*)} = k)_{1 \leq i \leq n}$ il vettore binario che stabilisce quali righe appartengono al k-esimo cluster di riga all'esecuzione s^*-esima, per $k = 1, \ldots, K$, e sia $\mathcal{I}_{h_s(k)} = [\mathbb{1}(\hat{\mathcal{Z}}_i^{(s^*)} = h_s(k))]_{1 \leq i \leq n}$ è il vettore binario che indica quali righe appartengono al cluster $h_s(k)$ dato dal s-esimo run, dove $h_s(k) = \text{argmax}_{h=1,\ldots,K} \sum_{i=1}^{n} \mathbb{1}(\mathcal{Z}_i^{(s^*)} = k, \mathcal{Z}_i^{(s^*)} = h)$, e $s \neq s^*$. Inoltre, si considerino i pesi $\omega_s = 1/(\ell^{s^*} - \ell^{(s)})$. L'incertezza dei cluster di riga k può essere calcolata come

$$\epsilon_k^{\text{rows}} = \frac{\sum_{s \neq s^*} \omega_s \, \text{CER}\left(\mathcal{I}_k, \mathcal{I}_{h_s(k)}\right)}{\sum_{s \neq s^*} \omega_s}, \tag{2.6}$$

dove $\text{CER}(\cdot,\cdot)$ denota il tasso di errore di clustering citato precedentemente. L'indice quantifica la differenza tra due partizioni in due gruppi di uno stesso insieme di dati ed è definito come segue. Siano P e Q le due partizioni e sia $\mathbb{1}_{P(i,i')}$ un indicatore che indica se la partizione P colloca le osservazioni i e i' nello stesso gruppo, e sia $\mathbb{1}_{Q(i,i')}$ in modo analogo. Con questa notazione, l'indice CER è definito come:

$$\text{CER}\,(P, Q) = \sum_{i > i'} \left| \mathbb{1}_{P(i,i')} - \mathbb{1}_{Q(i,i')} \right| / \binom{n}{2}.$$

Quanto più è vicino a 0, tanto maggiore è l'accordo tra le due partizioni, mentre valori elevati indicano una mancanza di concordanza tra P e Q. In questo contesto, si calcola l'indice tra i gruppi della configurazione di riferimento e tutte le altre configurazioni stimate. Nel calcolo della discrepanza (2.6) si mediano i valori di CER con i pesi $\{\omega_s\}_{s \neq s^*}$ per tenere conto dell'evidenza a favore di tale configurazioni in esame: si attribuisce un peso elevato all'indice CER tra $\mathcal{I}_k$ e $\mathcal{I}_{h_s(k)}$ quando la differenza $\ell^{(s)} - \ell^{(s^*)}$ è piccola,

e vice-versa. Il motivo è intuitivo: se sia ω_S che CER($\mathcal{I}_k$,$\mathcal{I}_{h_s(k)}$) sono elevati, le due configurazioni di clustering sono considerevolmente diverse ma con valori simili di log-verosimiglianza. Questo indica che la struttura di clustering dei dati è incerta e non univoca. Al contrario, se ω_s è basso, la differenza tra $\hat{\boldsymbol{\mathcal{Z}}}^{(s^*)}$ e $\hat{\boldsymbol{\mathcal{Z}}}^{(s)}$ può essere considerata irrilevante, perché i dati indicano chiaramente che $\hat{\boldsymbol{\mathcal{Z}}}^{(s^*)}$ è la soluzione più probabile.

La misura di incertezza qui introdotta può essere interpretata in modo analogo all'indice CER; più i valori ϵ_k^{rows} sono prossimi a 0, maggiore è l'evidenza di un'unica struttura di co-clustering dei dati.

2.3.5 Selezione del kernel spaziale

La dipendenza spaziale di covarianza può esse modellata da diverse funzioni isotrope. Le più note sono:

- $k_1(d;\phi_1) = \theta_E) = \exp\left(-\frac{d}{\theta_{R,\alpha_R}}\right)$,
- $k_2(d;\phi_2 = \theta_R, \alpha_R) = \left(1 + \frac{d^2}{2\alpha_r\theta_R^2}\right)^{-\alpha_R}$,
- $k_3(d;\phi_3 = \theta_G) = \exp\left(-\frac{d^2}{2\theta_G^2}\right)$.

dove $k_1(\cdot;\phi_1)$ è il kernel *Esponenziale* con parametro di scala θ_E, $k_2(\cdot;\phi_2)$ è il kernel *Rational Quadratic* con parametri (θ_R, α_R) non negativi, e infine $k_3(\cdot;\phi_3)$ è il kernel *Gaussiano* con parametro di lunghezza di scala θ_G.

Per la selezione di un kernel di covarianza spaziale adatto per i dati in analisi un possibile approccio è quello di sfruttare la dipendenza spaziale empirica attraverso il variogramma (Cressie et al., 2022). Il variogramma è un grafico utilizzato in geostatistica per descrivere la dipendenza spaziale tra le misure di una variabile nello spazio. La forma del grafico fornisce informazioni sul tipo di dipendenza spaziale presente nei dati, ad esempio se la dipendenza è lineare o sferica, per scegliere la funzione più adatta tra una ampia lista di proposte (si veda per esempio Rasmussen & Williams, 2006). Nelle ipotesi del modello Spartaco questo approccio non è implementabile perché sarebbe necessario conoscere a priori i cluster di colonna, dato che si suppone un tipo di dipendenza spaziale per ogni blocco. Tuttavia, le simulazioni riportate nell'articolo di Sottosanti & Risso (2022) evidenziano come un kernel esponenziale riesca a modellare anche altri tipi di

interazione spaziale. Anche nell'implementazione di questa versione del metodo proposta in questa tesi si segue la stessa scelta, in modo tale da rimanere in linea con il metodo Spartaco.

2.3.6 Selezione del modello

Per stimare il modello è necessario specificare il numero di cluster di riga, sebbene questo non sia noto a priori. Inoltre si possono considerare diversi modelli di covarianza spaziale $k(\cdot,\cdot)$. Per selezionare il modello più adatto ai dati, sia per quanto riguarda il numero di cluster che la funzione di covarianza spaziale, si valutano diverse configurazioni confrontando i valori del criterio della log-likelihood integrata completa (ICL, Biernacki et al., 2000). Rispetto al modello formulato in (2.1)-(2.2), sua espressione risulta essere:

$$\text{ICL} = \log \mathcal{L}(\hat{\boldsymbol{\Theta}}, \hat{\boldsymbol{\mathcal{Z}}}) - n \log K - p \log R - \frac{4KR + \dim(\boldsymbol{\phi})R}{2} \log np, \tag{2.7}$$

dove $\dim(\phi)$ è la dimensione del vettore dei parametri ϕ_r, che non dipende da r. In pratica, il modello migliore è quello corrisponde a quello con il valore più elevato di (2.7).

Sotto la versione originale del modello Spartaco, dove anche le etichette di colonna sono da stimare, i criteri di informazione più comuni, AIC e BIC, non possono essere calcolati in quanto la verosimiglianza $p(\mathbf{X}; \boldsymbol{\Theta})$, marginalizzata rispetto alle variabili latenti $\boldsymbol{\mathcal{Z}}$, non è disponibile in forma chiusa. Per questo motivo è stato utilizzato il criterio ICL che non presenta tale criticità.

Anche se il modello qui presentato (2.5) non presenta il problema di calcolare la verosimiglianza marginalizzata, si sceglie comunque di utilizzare lo stesso criterio di informazione (ICL) per consentire eventuali confronti con la versione originale del metodo Spartaco in futuro. Per la stessa ragione, si mantiene anche la penalizzazione relativa ai parametri dei cluster di colonna sebbene siano costanti tra modelli confrontati nella selezione.

Capitolo 3

Studi di simulazioni

In questo capitolo si presenta uno studio di simulazione finalizzato a valutare diversi aspetti delle prestazioni del metodo di clustering proposto in questa tesi. Sia la versione base che quella penalizzata del metodo sono state esaminate analizzandone l'accuratezza sia nella formazione dei cluster dei geni che nella stima dei parametri del modello, e indagando il contributo della penalizzazione nella performance del metodo. Inoltre, è stata valutata la robustezza del metodo nell'identificare la corretta configurazione in cluster dei geni in presenza di incertezze nell'annotazione usata per definire le etichette degli *spot*.

3.1 Modello di simulazione

Il modello di simulazione si definisce seguendo l'approccio adottato nella simulazione 1 del metodo di co-clustering Spartaco nell'articolo di Sottosanti & Risso (2022). Si genera una matrice di dati con una struttura a blocchi latente, in cui ogni blocco segue una distribuzione normale matriciale. Assumendo $K = R^{=}3$ si generano 9 blocchi di dimensioni $n_k = 200$ x $p_r = 200$ per ogni $k = 1, 2, 3$ e $r = 1, 2, 3$. Siano $\boldsymbol{\mathcal{Z}}^{true}$ e $\boldsymbol{\mathcal{W}}^{true}$ le etichette di clustering di riga e colonna. Si definiscono gli insieme $C_k^{true} = i = 1, \ldots, n : \boldsymbol{\mathcal{Z}}_i^{true}$ e $D_r^{true} = i = 1, \ldots, n : \boldsymbol{\mathcal{W}}_i^{true}$. Le osservazioni del (k, r)-esimo blocco sono generate da:

$$\mathbf{X}^{kr} \sim \mathcal{MVN}(\mu_{kr}^{true} \mathbb{1}_{n_r * p_r}, \boldsymbol{\Sigma}_{kr}^{true}, \boldsymbol{\Delta}_{kr}^{true}), \quad \boldsymbol{\Delta}_{kr}^{true} = \tau_{kr}^{true} \boldsymbol{\mathcal{K}}_r^{true}(\mathbf{S}^r; \phi_r^{true}) + \xi_{kr}^{true} \mathbb{1}_{p_r} \quad (3.1)$$

dove $\boldsymbol{\mathcal{K}}_r^{true}(\mathbf{S}^r; \phi_r^{true}) = \left(k_r^{true}(\|\mathbf{s}_j^r - \mathbf{s}_{j'}^r)\|; \phi_r^{true})\right)_{\{1\leq j,j',\leq p_r\}}$ e $k_r^{true}(\cdot, \phi_r^{true})$ è una funzione di covarianza spaziale isotropa. Si definisce un kernel $\boldsymbol{\mathcal{K}}_r^{true}$ specifico per ogni cluster di *spot*. Nella simulazione si adottano tre diversi kernel spaziali per valutare se il metodo è robusto rispetto alla mis-specificazione del kernel, come avviene nelle simulazioni di Sottosanti & Risso (2022). Si utilizza un kernel *Esponenziale* con parametro di scala θ_E per le colonne in D_1^{true}, un kernel *Razionale Quadratico* di parametri (θ_R, α_R) per D_2^{true} e un kernel *Gaussiano* con parametro di scala θ_G per D_3^{true}. La specificazione dei parametri è descritta in seguito. Dalla distribuzione di $\mathbf{X}_{rk}$ in (3.1) si possono ricavare le distribuzioni marginali dell'espressione dei geni e degli *spot*, rispettivamente:

$$\mathbf{x}_{i\cdot}^{k\cdot}|\boldsymbol{\mathcal{Z}}^{true}, \boldsymbol{\mathcal{W}}^{true} \sim \mathcal{N}_p(\mu_{k1}^{true}\mathbb{1}_{p_1}, \ldots, \quad \mu_{k3}^{true}\mathbb{1}_{p_3}), \boldsymbol{\Sigma}_{ii}^{true}\,\mathrm{diag}(\boldsymbol{\Delta}_{kr}^{true})_{r=1,2,3}, \tag{3.2}$$

$$\mathbf{x}_{\cdot j}^{\cdot r}|\boldsymbol{\mathcal{Z}}^{true}, \quad \boldsymbol{\mathcal{W}}^{true} \sim \mathcal{N}_p(\mu_{1r}^{true}\mathbb{1}_{n_1}, \ldots, \mu_{3r}^{true}\mathbb{1}_{n_3}), c^{true}\,\mathrm{diag}(\boldsymbol{\Sigma}_k^{true})_{k=1,2,3}, \tag{3.3}$$

dove $\boldsymbol{\Sigma}_{ii}$ è la matrice di varianza del i-esimo gene e non dipende dal cluster di appartenenza k. La matrice di cross-covarianza tra le righe i e $i' \in C_k^{true}$ è $\mathrm{Cov}(\mathbf{x}_{i\cdot}^k, \mathbf{x}_{i'\cdot}^k) = \boldsymbol{\Sigma}_{k,ii}^{true}\,\mathrm{diag}(\boldsymbol{\Delta}_{kr}^{true})_{r=1,2,3}$ e la matrice di cross-covarianza tra due colonne $j, j' \in D_r^{true}$ è $\mathrm{Cov}(\mathbf{x}_{\cdot j}^r, \mathbf{x}_{\cdot j'}^r) = \mathrm{diag}\{\tau_{kr}^{true}k_r^{true}\left(\|\mathbf{s}_j^r - \mathbf{s}_{j'}^r)\|; \phi_r^{true}\right))\boldsymbol{\Sigma}_k^{true}\}_{k=1,2,3}$. Si assume che le matrici di varianza e covarianza rimangano costanti nelle le tre aree del tessuto, e varino solo tra i cluster di geni. Quindi $\boldsymbol{\Sigma}_{kr}^{true} = \boldsymbol{\Sigma}_k^{true}$ per ogni r. Si generano le matrici di varianza $\boldsymbol{\Sigma}_k^{true}$ come segue:

$$\boldsymbol{\Sigma}_1^{true} \sim \mathcal{W}i(210, 0.3\mathbb{1}_200), \boldsymbol{\Sigma}_2^{true} \sim \mathcal{W}i(230, 0.5\mathbb{1}_200), \boldsymbol{\Sigma}_3^{true} \sim \mathcal{W}i(200, \boldsymbol{\Sigma}_1^{true}/150) \tag{3.4}$$

dove $\mathcal{W}i(a, \mathbf{b})$ denota la distribuzione Wishart con a gradi di libertà e matrice di scala $\mathbf{b}$. La distribuzione inverse Wishart assicura che le matrici simulate siano definite positive. Per fare in modo che le simulazioni ricreino degli scenari che possono essere trovati nei dati reali, si utilizzano le coordinate spaziali di un dataset reale per definire la mappa di *spot* di cui i dati simulati rappresentano i relativi profili di espressione genica. Si utilizzano i dati di trascrittomica spaziale relativi ad una sezione tessuto della corteccia prefrontale dorsolaterale umana, processata con la tecologia Visium, raccolti da Maynard

et al. (2021). Il dataset è disponibile nel pacchetto R `spatialLIBD`. Per quest'analisi, si selezionano i dati relativi al soggetto con ID 151507. Analogamente ai dati in studio, anche gli *spot* di questo tessuto sono stati annotati manualmente. Vengono estratti 200 *spot* da ognuno dei tre strati presenti nella regione in alto a destra dell'immagine del tessuto riportata in Figura 3.1. Si ottiene così una mappa di 600 *spot*. Sebbene le etichette di clustering $\boldsymbol{\mathcal{W}}^{true}$ sono state assegnate sulla base della natura delle cellule che compongono lo *spot*, nelle ipotesi del modello di simulazione 3.1 hanno un diverso significato: corrispondono a gruppi di *spot* in cui si suppone che i gruppi di geni siano espressi con specifici profili di variazione spaziale. I parametri $(\theta_E, \theta_R, \alpha_R, \theta_G)$ sono stati fissati sulla base di quanto le relative aree nel tessuto sono estete: la funzione di covarianza di D_1^{true} è più schiacciata rispetto a quella di D_2^{true}, poiché D_1^{true} ricopre una porzione di tessuto più ristretta. Il cluster D_3^{true} è composto da due gruppi di *spot* separati, si fissano i parametri della funzione di covarianza $k_3^{true}(\cdot;\cdot)$ in modo tale che gli *spot* nello stesso gruppo siano spazialmente correlati, mentre *spot* in gruppi differenti siano poco correlati. Sulla base di queste considerazioni si selezionano i seguenti valori: $\theta_E = 50, \theta_R = 50, \alpha_R = 2$ e $\theta_G = 70$. Le funzioni di covarianza definite sono riportate nel grafico a sinistra in Figura 3.2. Si fissa il vincolo di identificabilità $c_{kr}^{true} = c^{true} = 100$. Si considerano tre diversi scenari per definizione della matrice dei *spatial signal-to-noise ratio* di blocco. In riferimento alla Figura in 3.3, si definiscono i valori della matrice in ogni scenario:

- **Scenario A)** : (i) effetto spaziale debole, $\tau_{kr}^{true}/\xi_{kr}^{true} = 0.2$; (ii) effetto spaziale che è la metà della variabilità residua, $\tau_{kr}^{true}/\xi_{kr}^{true} = 0.5$; e (iii) effetto spaziale comparabile alla variabilità residua, $\tau_{kr}^{true}/\xi_{kr}^{true} = 1$.

- **Scenario B)** : (i) effetto spaziale debole, $\tau_{kr}^{true}/\xi_{kr}^{true} = 0.2$; (ii) effetto spaziale comparabile all'effetto residuo, $\tau_{kr}^{true}/\xi_{kr}^{true} = 1$; e (iii) effetto spaziale maggiore rispetto alla variabilità residua, $\tau_{kr}^{true}/\xi_{kr}^{true} = 2$.

- **Scenario C)** : (i) nessun effetto spaziale, $\tau_{kr}^{true}/\xi_{kr}^{true} = 0$; (ii) effetto spaziale comparabile all'effetto residuo, $\tau_{kr}^{true}/\xi_{kr}^{true} = 1$; e (iii) effetto spaziale maggiore rispetto alla variabilità residua, $\tau_{kr}^{true}/\xi_{kr}^{true} = 2$.

Questi tre scenari corrispondono ad un progressivo aumento della spazialità presente nel dataset: nello scenario **A** non vi è nessun blocco con una spazialità elevata, ma nemmeno totalmente assente ($\tau_{kr}^{true}/\xi_{kr}^{true} > 0$). Nello scenario **B** si inseriscono per ogni riga dei blocchi con una spazialità sostanziale ($\tau_{kr}^{true}/\xi_{kr}^{true} > 1$). Infine, nello scenario **C** si inseriscono dei co-cluster con effetto spaziale nullo. Tramite il vincolo di identificabilità, dalla definizione della matrice *spatial signal-to-noise ratio* si ricavano anche i valori τ_{kr} e ξ_{kr}. Infine, si generano i valori di media μ_{kr} da una distribuzione uniforme in $(-3, 3)$. Le matrice dei dati generate dal modello vengono centrate prima della stima in modo tale che la penalizzazione sulla media sia uniforme tra i vari blocchi.

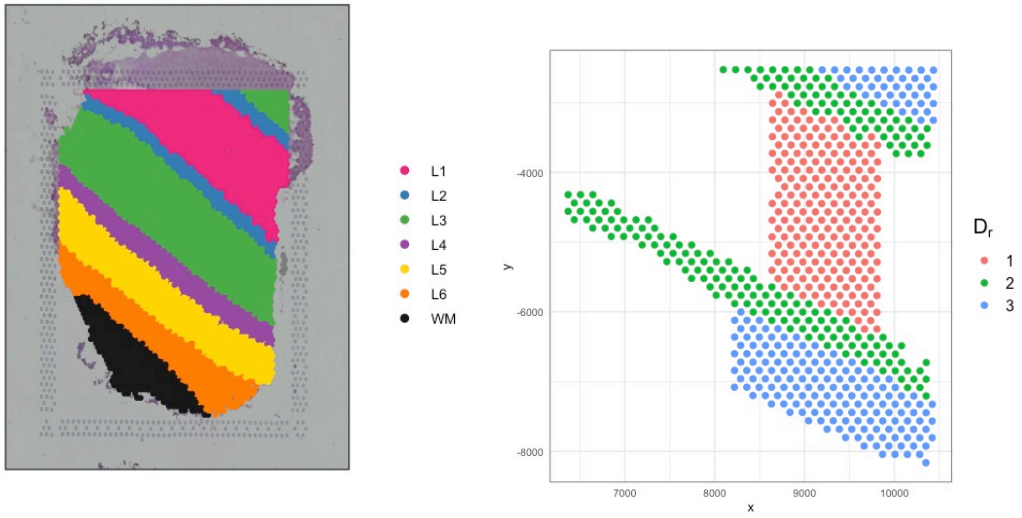


FIGURA 3.1: A sinistra: campione di tessuto di corteccia prefrontale dorsolaterale umana (DLPFC) processato con la piattaforma Visium e disponibile nel pacchetto R `spatialLIBD`. I punti rappresentano gli *spot* sulla superficie del chip. Colori diversi indicano l'annotazione manuale delle aree eseguita da Maynard et al. (2021): uno strato di materia bianca (WM) nella parte inferiore sinistra dell'immagine e 6 strati (da Layer6 a Layer1). A destra: mappa degli *spot* utilizzati per generare gli esperimenti di simulazione, estratti dal soggetto 151507 contenuto nel pacchetto R `spatialLIBD`. I cluster sono di dimensioni uguali, $p_1 = p_2 = p_3 = 200$.

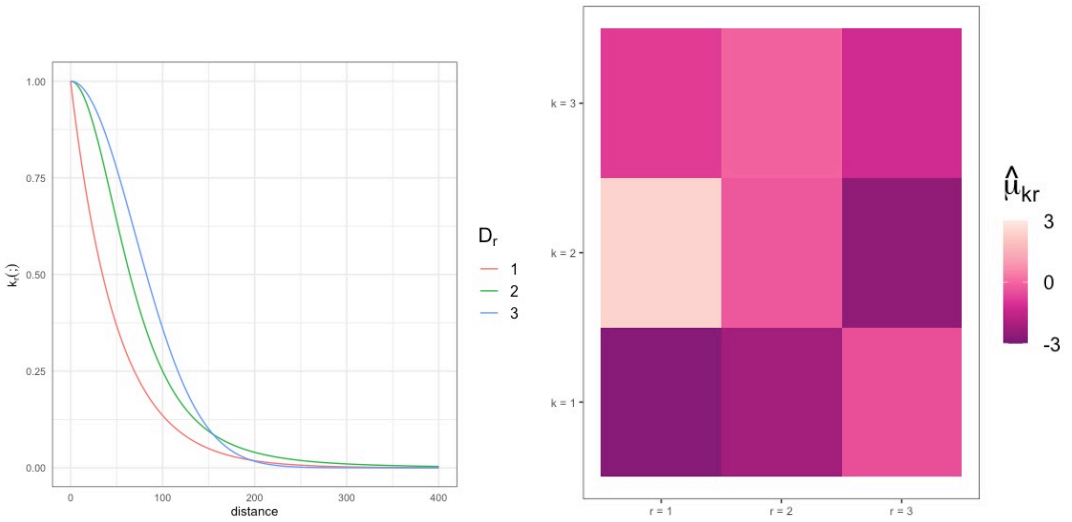
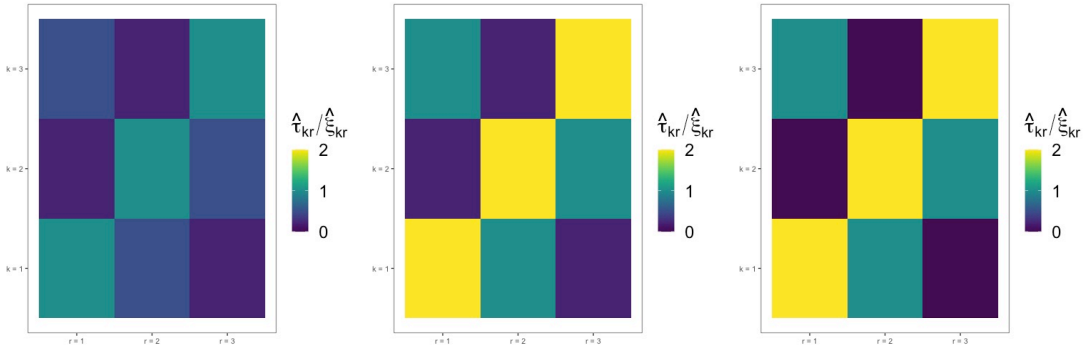


FIGURA 3.2: A sinistra: confronto delle funzioni di covarianza utilizzate nei tre cluster di *spot*. Quando $r = 1$, la covarianza è esponenziale con scala $\theta_E = 50$, quando $r = 2$, è razionale-quadratica con $\theta_R = 50 e \alpha_R = 2$, e quando $r = 3$ è gaussiana con scala $\theta_G = 70$. A destra: Rappresentazione delle struttura a blocchi latente utilizza per generare gli esperimenti di simulazione colorata in base al valore della media di ciascun blocco μ_{kr}^{true}.



(a) (b) (c)

FIGURA 3.3: Rappresentazione delle struttura a blocchi latente utilizza per generare gli esperimenti di simulazione. Tutti i blocchi nei pannelli (a)-(c) hanno la stessa dimensione e sono colorati in base al valore del *spatial signal-to-noise ratio* $\tau_{kr}^{true}/\xi_{kr}^{true}$ definiti rispettivamente nello scenario **A**, **B** e **C**.

3.2 Valutazione della performance del metodo di clustering

Si intende valutare la performance del metodo proposto rispetto a due aspetti: l'accuratezza della classificazione dei geni in gruppi e la precisione delle stime dei parametri di media e del *spatial signal-to-noise ratio* del modello.

Per ciascuno dei tre scenari descritti, sono state simulate 10 repliche della distribuzione descritta nel paragrafo precedente, su ognuna di esse è stato applicato il metodo di clustering con $K = 3$. Si usa il vero valore di K poiché la valutazione del criterio di selezione del modello tramite ICL non rientra negli obiettivi di questo studio. Si utilizza l'indice CER per valutare la discrepanza tra le configurazioni in cluster identificate dal metodo applicato alle 10 repliche dei dati rispetto alla classificazione corretta. La precisione delle stime della media e del parametro *signa-to-noise ratio* di ogni blocco viene valutata tramite l'Errore Relativo Medio (ERM). In particolare, per ognuna dei nove blocchi della matrice, si calcola la differenza tra il valore stimato e il corrispondente valore vero e si divide per tale valore. Successivamente, si prende il valore assoluto e si fa la media degli errori tra le 10 stime ottenute. Infine, i 9 errori ERM ottenuti vengono mediati per sintetizzarli in un'unica misura per favorire il confronto tra i 3 diversi scenari. Formalmente, si definisce l'ERM della stima di un parametro matriciale $\boldsymbol{\Theta}^{true} = (\theta_{kr})_{1\leq k\leq K, 1\leq r\leq R}$ di dimensioni K x R date N-stime $\hat{\boldsymbol{\Theta}}^{i} = (\hat{\theta}^{i}_{kr})_{1\leq k\leq K, 1\leq r\leq R}$, con $i = 1, \ldots, n$, nel seguente modo:

$$\text{ERM}(\boldsymbol{\Theta}) := \sum_{i=1}^{N} \left\{ \sum_{k=1}^{K} \sum_{r=1}^{R} |\hat{\theta}^{i}_{kr} - \theta^{true}_{kr}|/|\theta^{true}_{kr}| \right\} \tag{3.5}$$

In questo studio si è interessati a valutare l'ERM($\boldsymbol{\mu}$), dove $\boldsymbol{\mu} = (\mu_{kr})_{1\leq k\leq K, 1\leq r\leq R}$ matrice delle medie dei co-cluster, e l'ERM($\boldsymbol{\tau}/\boldsymbol{\xi}$), dove la matrice $\boldsymbol{\tau}/\boldsymbol{\xi} = (\tau_{kr}/\xi_{kr})_{1\leq k\leq K, 1\leq r\leq R}$ contiene i valori di *spatial signal-to-noise ratio* di ogni blocco. Nello scenario **C** vi sono dei valori di $\tau^{true}_{kr}/\xi^{true}_{kr}$ nulli, per rendere ben definita la Formula (3.5) vengono incrementati a 10^{-3}, considerato un valore abbastanza basso da rappresentare una variabilità spaziale nulla. Nonostante la penalità viene inserita sul parametro τ_{kr}, si sceglie di valutare dalla bontà della stima dello *spatial signal-to-noise ratio*, in quanto è questo il

parametro che viene fissato nella definizione del modello di simulazione; mentre τ si ricava tramite il vincolo di identificabilità. Inoltre, lo *spatial signal-to-noise ratio* di un blocco ha una chiara interpretazione biologica di interesse.

Si descrive l'impostazione adottata per l'algoritmo di stima. Per modellare la correlazione spaziale dell'espressione tra *spot* adiacenti si utilizza un kernel esponenziale, seguendo l'approccio adottato nelle simulazioni e nell'applicazione del modello Spartaco dagli autori Sottosanti & Risso (2022). Per ridurre le probabilità che l'algoritmo CEM converga a punti di massimo locali o punti di sella, si esegue la stima di ogni modello a partire da cinque punti iniziali scelti casualmente. Il numero massimo di iterazioni dell'algoritmo è stato fissato a 3000, con 10 passi di massimizzazione per ciascuna iterazione.

3.2.1 Risultati

In Tabella 3.5 sono riportati gli errori ERM($\boldsymbol{\mu}$) e ERM($\boldsymbol{\tau}/\boldsymbol{\xi}$), e i relativi standard error; mentre in Figura 3.4 si mostrano le distribuzioni dell'indice CER nei tre scenari di simulazione. Nel scenario **A**, si osserva una discrepanza tra le configurazioni in media pari a 0.01, mentre per **B** e **C** tutti gli indici CER sono uguali a 0. In tutti e tre gli scenari, l'errore ERM($\boldsymbol{\mu}$) risulta essere piuttosto elevato, soprattutto rispetto all'errore ERM($\boldsymbol{\tau}/\boldsymbol{\xi}$). Si osservi che nello scenario **C**, il valore di ERM($\boldsymbol{\tau}/\boldsymbol{\xi}$) è maggiore rispetto agli altri due scenari, il che potrebbe essere dovuto alla presenza di valori esattamente uguali a 0 nella matrice dei *spatial signal-to-noise ratio*.

	A	B	C
ERM($\boldsymbol{\mu}$)	6.321(3.101)	5.092(2.309)	6.031(3.015)
ERM($\boldsymbol{\tau}/\boldsymbol{\xi}$)	0.118(0.016)	0.112(0.018)	0.347(0.047)

TABELLA 3.1: Risultati del metodo di clustering su 10 repliche del modello di simulazione. Si riporta l'errore ERM($\boldsymbol{\mu}$) e ERM($\boldsymbol{\tau}/\boldsymbol{\xi}$) per ognuno dei tre scenari **A, B, C**, e il relativo standard error.

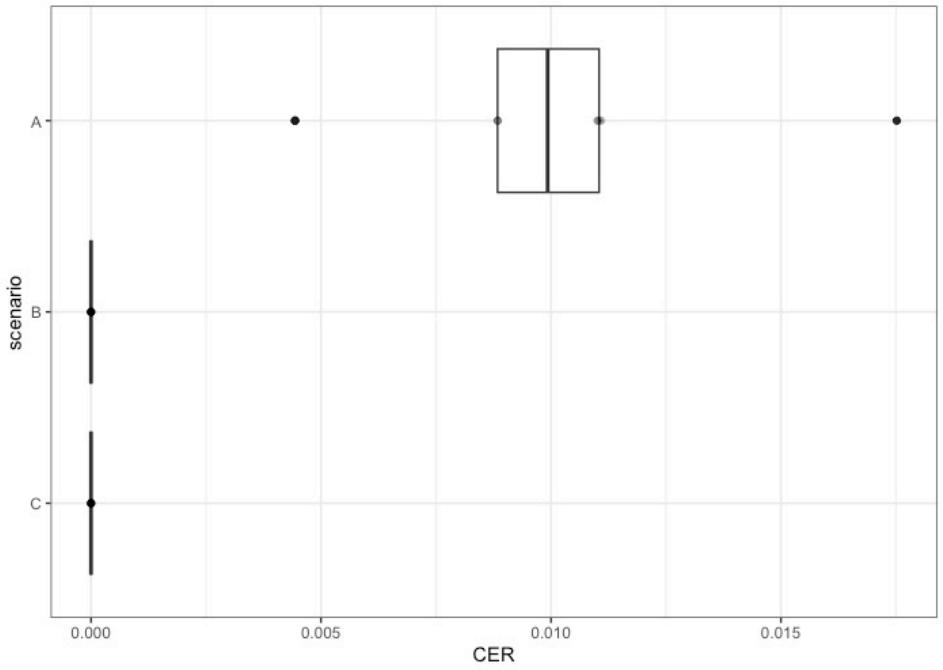


FIGURA 3.4: Boxplot dei valori dell'indice di discrepanza CER tra la corretta configurazione in cluster dei geni e la configurazione stimata dal metodo applicato alle 10 repliche del modello di simulazione per ognuno dei tra scenari.

3.3 Valutazione dell'effetto della penalizzazione nel metodo di clustering

Si applica il metodo di clustering penalizzato a 10 repliche del modello di simulazione per ognuno dei tre scenari descritti. Si valuta l'effetto della sola penalità *ridge* sulla media μ dei co-cluster, della sola penalità *lasso* sui parametri τ e una combinazione delle penalità: λ_μ e $\lambda_\tau \in \{0, 0.3, 0.5, 1, 1.5, 2\}$. Le impostazioni dell'algoritmo di stima sono quelle usate nel Paragrafo 3.2.

Per valutare l'impatto delle penalizzazioni utilizzate sui risultati del metodo di clustering, vengono adoperate diverse misure di sintesi. In particolare, per valutare l'accuratezza nella stima dei parametri di media μ e dello *spatial signal-to-noise ratio* di blocco, viene calcolato l'errore relativo medio come specificato nella Formula (3.5). Per quantificare la variabilità delle configurazioni dei cluster dei geni rispetto alla corretta suddivisione, si utilizza l'indice CER. Infine, per valutare la sparsità della matrice dei parametri $\hat{\tau}$ e la corretta identificazione degli elementi nulli e non nulli, si calcolano il tasso

di sparsità e il tasso di errore di sparsità. Il tasso di sparsità indica la proporzione di elementi prossimi a zero ($\hat{\tau}_{kr}/\hat{\xi}_{kr} < 10^{-3}$), mentre il tasso di errore di sparsità rappresenta la percentuale di elementi τ_{kr} che sono stati erroneamente stimati a zero o erroneamente stimati come non-zero ($\hat{\tau}_{kr}/\hat{\xi}_{kr} > 10^{-3}$ quando $\tau_{kr}/\xi_{kr} = 0$ e $\hat{\tau}_{kr}/\hat{\xi}_{kr} < 10^{-3}$ quando $\tau_{kr}/\xi_{kr} > 0$). E' particolarmente interessante considerare queste due misure di errore in quanto consentono di valutare se il metodo è in grado di identificare correttamente la struttura spaziale della matrice dei dati.

3.3.1 Risultati

I risultati delle simulazioni sono sintetizzati nelle Tabelle 3.2, 3.3 e (3.4). In tutti e tre gli scenari studiati risulta essere necessaria una penalizzazione sul parametro della media μ di blocco per ridurre il rispettivo ERM($\boldsymbol{\mu}$). Già con una penalità moderata ($\lambda_\mu = 0.3$), si verifica un significativo calo dell'ERM: da 6.32 a 1.23 nello scenario **A**; da 5.09 a 1.35 nello scenario **B**. Inoltre, l'errore progressivamente diminuisce con l'aumentare del valore di λ_μ. Invece, nello scenario **C** risulta necessaria una penalità maggiore per ridurre l'errore ERM($\boldsymbol{\mu}$) e una piccola penalità sulla matrice $\boldsymbol{\tau} = (\tau_{kr})_{1\leq k\leq K, 1\leq r\leq R}$ sembra favorirne la diminuzione ($\lambda_\tau = 0.3, 0.5$). Al contrario, penalità più elevate su $\boldsymbol{\tau}$ ($\lambda_\tau \geq 1$) peggiorano progressivamente l'errore ERM($\boldsymbol{\tau}/\boldsymbol{\xi}$), in ogni scenario. I valori CER rimangono a 0 negli scenari **B** e **C**, e per alcune combinazioni dei parametri di regolarizzazione diminuisce leggermente nello scenario **A** ($\lambda_\mu = 0.3$, $\lambda_\tau = 0.50$ e $\lambda_\mu = 0.3$, $\lambda_\tau = 0$). Dall'analisi dei tassi di sparsità e di errore di sparsità nel contesto **C**, si conferma che il metodo è in grado di stimare correttamente i valori di τ_{kr} che sono effettivamente nulli. Inoltre, la penalizzazione non porta a stimare erroneamente a zero i valori di *spatial signal-to-noise ratio* di blocco quando è invece presente un effetto spaziale.

Si sottolinea che lo studio è basato su sole 10 repliche per motivi computazionali, e quindi le considerazioni effettuate non possono essere considerate esaustive, ma rappresentano una prima analisi del modello di clustering penalizzato.

λ_μ	λ_τ	ERM($\boldsymbol{\mu}$)	ERM($\boldsymbol{\tau}/\boldsymbol{\xi}$)	Tasso sparsità	Tasso errore sparsità	CER
0.00	0.00	6.321(3.101)	0.118(0.016)	0(0)	-(-)	0.01(0.004)
0.00	0.30	5.458(3.561)	0.118(0.023)	0(0)	-(-)	0.01(0.004)
0.00	0.50	6.117(2.919)	0.131(0.033)	0(0)	-(-)	0.01(0.005)
0.00	1.00	6.206(3.357)	0.179(0.042)	0(0)	-(-)	0.011(0.005)
0.00	1.50	5.781(2.833)	0.23(0.042)	0(0)	-(-)	0.01(0.004)
0.00	2.00	5.694(2.559)	0.278(0.041)	0(0)	-(-)	0.01(0.003)
0.30	0.00	1.234(0.366)	0.116(0.016)	0(0)	-(-)	0.009(0.005)
0.30	0.30	1.474(0.356)	0.116(0.025)	0(0)	-(-)	0.01(0.004)
0.30	0.50	1.373(0.551)	0.127(0.034)	0(0)	-(-)	0.009(0.004)
0.30	1.00	1.598(0.671)	0.177(0.043)	0(0)	-(-)	0.011(0.004)
0.30	1.50	1.498(0.618)	0.227(0.044)	0(0)	-(-)	0.01(0.004)
0.30	2.00	1.465(0.675)	0.276(0.041)	0(0)	-(-)	0.01(0.005)
0.50	0.00	1.183(0.339)	0.12(0.017)	0(0)	-(-)	0.01(0.004)
0.50	0.30	1.051(0.135)	0.118(0.025)	0(0)	-(-)	0.009(0.004)
0.50	0.50	1.31(0.415)	0.128(0.034)	0(0)	-(-)	0.011(0.005)
0.50	1.00	1.06(0.054)	0.176(0.043)	0(0)	-(-)	0.01(0.004)
0.50	1.50	1.31(0.381)	0.227(0.044)	0(0)	-(-)	0.011(0.004)
0.50	2.00	1.261(0.359)	0.275(0.041)	0(0)	-(-)	0.01(0.003)
1.00	0.00	1.105(0.171)	0.119(0.017)	0(0)	-(-)	0.01(0.004)
1.00	0.30	1.113(0.213)	0.116(0.027)	0(0)	-(-)	0.011(0.004)
1.00	0.50	1.048(0.116)	0.125(0.033)	0(0)	-(-)	0.009(0.004)
1.00	1.00	1.108(0.205)	0.176(0.043)	0(0)	-(-)	0.009(0.002)
1.00	1.50	1.152(0.226)	0.227(0.044)	0(0)	-(-)	0.01(0.004)
1.00	2.00	1.148(0.291)	0.274(0.042)	0(0)	-(-)	0.01(0.004)
1.50	0.00	1.017(0.028)	0.12(0.018)	0(0)	-(-)	0.009(0.003)
1.50	0.30	1.015(0.021)	0.117(0.026)	0(0)	-(-)	0.01(0.004)
1.50	0.50	1.012(0.025)	0.128(0.035)	0(0)	-(-)	0.01(0.004)
1.50	1.00	1.098(0.146)	0.176(0.042)	0(0)	-(-)	0.01(0.003)
1.50	1.50	1.054(0.142)	0.227(0.044)	0(0)	-(-)	0.01(0.004)
1.50	2.00	1.07(0.12)	0.275(0.041)	0(0)	-(-)	0.01(0.003)
2.00	0.00	1.056(0.116)	0.119(0.015)	0(0)	-(-)	0.01(0.005)
2.00	0.30	1.019(0.044)	0.118(0.025)	0(0)	-(-)	0.011(0.004)
2.00	0.50	1.073(0.136)	0.127(0.035)	0(0)	-(-)	0.009(0.004)
2.00	1.00	1.045(0.12)	0.175(0.043)	0(0)	-(-)	0.011(0.004)
2.00	1.50	1.004(0.023)	0.226(0.044)	0(0)	-(-)	0.011(0.004)
2.00	2.00	0.941(0.054)	0.271(0.051)	0(0)	-(-)	0.009(0.002)

TABELLA 3.2: Risultati del metodo di clustering penalizzato con diversi valori di λ_μ e λ_τ nello scenario **A**. Si riporta ERM($\boldsymbol{\mu}$) (e standard error) e ERM($\boldsymbol{\tau}/\boldsymbol{\xi}$) (e standard error)calcolato sulle 10 repliche del modello di simulazione; il tasso di sparsità (e standard error) e del tasso di errore di sparsità (e stardard error) e l'indice CER (e standard error). In questo scenario non vi sono valori nulli nella matrice *spatial signal-to-noise ratio* $\boldsymbol{\tau}/\boldsymbol{\xi}$ per cui il tasso di errore di sparisità e il suo standard error non vengono calcolati.

λ_μ	λ_τ	ERM($\boldsymbol{\mu}$)	ERM($\boldsymbol{\tau}/\boldsymbol{\xi}$)	Tasso sparsità	Tasso errore sparsità	CER
0.00	0.00	5.092(2.309)	0.112(0.018)	0(0)	-(-)	0(0)
0.00	0.30	6.181(1.938)	0.119(0.024)	0(0)	-(-)	0(0)
0.00	0.50	6.116(2.84)	0.135(0.034)	0(0)	-(-)	0(0)
0.00	1.00	5.177(1.677)	0.188(0.043)	0(0)	-(-)	0(0)
0.00	1.50	6.46(2.687)	0.243(0.043)	0(0)	-(-)	0(0)
0.00	2.00	6.921(2.117)	0.292(0.042)	0(0)	-(-)	0(0)
0.30	0.00	1.349(0.295)	0.115(0.019)	0(0)	-(-)	0(0)
0.30	0.30	1.645(0.469)	0.117(0.024)	0(0)	-(-)	0(0)
0.30	0.50	1.434(0.629)	0.135(0.031)	0(0)	-(-)	0(0)
0.30	1.00	1.468(0.477)	0.185(0.042)	0(0)	-(-)	0(0)
0.30	1.50	1.56(0.724)	0.239(0.044)	0(0)	-(-)	0(0)
0.30	2.00	1.743(0.685)	0.29(0.042)	0(0)	-(-)	0(0)
0.50	0.00	1.17(0.289)	0.113(0.018)	0(0)	-(-)	0(0)
0.50	0.30	1.1(0.207)	0.118(0.024)	0(0)	-(-)	0(0)
0.50	0.50	1.112(0.217)	0.135(0.032)	0(0)	-(-)	0(0)
0.50	1.00	1.457(0.465)	0.184(0.044)	0(0)	-(-)	0(0)
0.50	1.50	1.26(0.433)	0.24(0.044)	0(0)	-(-)	0(0)
0.50	2.00	1.255(0.401)	0.289(0.043)	0(0)	-(-)	0(0)
1.00	0.00	1.116(0.164)	0.114(0.018)	0(0)	-(-)	0(0)
1.00	0.30	1.11(0.163)	0.118(0.025)	0(0)	-(-)	0(0)
1.00	0.50	1.128(0.233)	0.134(0.032)	0(0)	-(-)	0(0)
1.00	1.00	1.132(0.238)	0.185(0.045)	0(0)	-(-)	0(0)
1.00	1.50	1.057(0.083)	0.239(0.045)	0(0)	-(-)	0(0)
1.00	2.00	1.15(0.185)	0.29(0.043)	0(0)	-(-)	0(0)
1.50	0.00	1.006(0.02)	0.115(0.019)	0(0)	-(-)	0(0)
1.50	0.30	1.118(0.156)	0.118(0.024)	0(0)	-(-)	0(0)
1.50	0.50	1.048(0.159)	0.134(0.033)	0(0)	-(-)	0(0)
1.50	1.00	1.056(0.123)	0.184(0.043)	0(0)	-(-)	0(0)
1.50	1.50	1.059(0.104)	0.24(0.045)	0(0)	-(-)	0(0)
1.50	2.00	1.123(0.192)	0.289(0.042)	0(0)	-(-)	0(0)
2.00	0.00	1.072(0.104)	0.113(0.021)	0(0)	-(-)	0(0)
2.00	0.30	1.067(0.102)	0.119(0.025)	0(0)	-(-)	0(0)
2.00	0.50	1.042(0.085)	0.132(0.034)	0(0)	-(-)	0(0)
2.00	1.00	1.048(0.094)	0.184(0.044)	0(0)	-(-)	0(0)
2.00	1.50	1.076(0.106)	0.239(0.046)	0(0)	-(-)	0(0)
2.00	2.00	1.073(0.096)	0.289(0.043)	0(0)	-(-)	0(0)

TABELLA 3.3: Risultati del metodo di clustering penalizzato con diversi valori di λ_μ e λ_τ nello scenario **B**. Si riporta ERM($\boldsymbol{\mu}$) e ERM($\boldsymbol{\tau}/\boldsymbol{\xi}$) calcolato sulle 10 repliche del modello di simulazione; il tasso di sparsità (e standard error) e del tasso di errore di sparsità (e stardard error) e l'indice CER (e standard error). In questo scenario non vi sono valori nulli nella matrice *spatial signal-to-noise ratio* $\boldsymbol{\tau}/\boldsymbol{\xi}$ per cui il tasso di errore di sparisità e il suo standard error non vengono calcolati.

λ_μ	λ_τ	ERM($\boldsymbol{\mu}$)	ERM($\boldsymbol{\tau}/\boldsymbol{\xi}$)	Tasso sparsità	Tasso errore sparsità	CER
0.00	0.00	6.031(3.015)	0.347(0.047)	0.3(0.054)	0.1(0.161)	0(0)
0.00	0.30	5.434(2.319)	0.355(0.037)	0.3(0.054)	0.1(0.161)	0(0)
0.00	0.50	6.513(2.533)	0.369(0.04)	0.3(0.054)	0.1(0.161)	0(0)
0.00	1.00	5.865(3.006)	0.414(0.047)	0.3(0.054)	0.1(0.161)	0(0)
0.00	1.50	6.221(2.273)	0.455(0.044)	0.3(0.054)	0.1(0.161)	0(0)
0.00	2.00	6.474(2.591)	0.502(0.04)	0.3(0.054)	0.1(0.161)	0(0)
0.30	0.00	3.593(2.206)	0.362(0.045)	0.289(0.078)	0.133(0.233)	0(0)
0.30	0.30	3.17(2.417)	0.352(0.037)	0.3(0.054)	0.1(0.161)	0(0)
0.30	0.50	3.861(2.059)	0.347(0.049)	0.278(0.059)	0.167(0.176)	0(0)
0.30	1.00	2.362(1.427)	0.394(0.057)	0.289(0.057)	0.133(0.172)	0(0)
0.30	1.50	2.94(2.332)	0.444(0.048)	0.289(0.057)	0.133(0.172)	0(0)
0.30	2.00	3.589(2.219)	0.487(0.051)	0.3(0.054)	0.1(0.161)	0(0)
0.50	0.00	3.234(2.181)	0.337(0.061)	0.3(0.054)	0.1(0.161)	0(0)
0.50	0.30	3.2(2.23)	0.348(0.04)	0.289(0.057)	0.133(0.172)	0(0)
0.50	0.50	2.909(2.24)	0.348(0.049)	0.3(0.054)	0.1(0.161)	0(0)
0.50	1.00	3.259(2.154)	0.398(0.059)	0.289(0.057)	0.133(0.172)	0(0)
0.50	1.50	3.545(2.09)	0.46(0.042)	0.278(0.059)	0.167(0.176)	0(0)
0.50	2.00	3.2(2.196)	0.489(0.046)	0.289(0.057)	0.133(0.172)	0(0)
1.00	0.00	2.806(2.009)	0.337(0.062)	0.278(0.059)	0.167(0.176)	0(0)
1.00	0.30	2.726(1.832)	0.331(0.056)	0.278(0.059)	0.167(0.176)	0(0)
1.00	0.50	2.481(1.907)	0.324(0.06)	0.289(0.057)	0.133(0.172)	0(0)
1.00	1.00	3.178(1.541)	0.389(0.049)	0.289(0.057)	0.133(0.172)	0(0)
1.00	1.50	3.229(1.652)	0.431(0.06)	0.3(0.054)	0.1(0.161)	0(0)
1.00	2.00	2.773(1.935)	0.479(0.062)	0.289(0.057)	0.133(0.172)	0(0)
1.50	0.00	2.401(1.716)	0.323(0.065)	0.278(0.079)	0.167(0.236)	0(0)
1.50	0.30	2.601(1.642)	0.312(0.067)	0.3(0.054)	0.1(0.161)	0(0)
1.50	0.50	1.997(1.577)	0.329(0.082)	0.289(0.057)	0.133(0.172)	0(0)
1.50	1.00	2.464(1.729)	0.382(0.062)	0.278(0.059)	0.167(0.176)	0(0)
1.50	1.50	2.062(1.7)	0.437(0.052)	0.278(0.059)	0.167(0.176)	0(0)
1.50	2.00	2.234(1.734)	0.463(0.063)	0.3(0.054)	0.1(0.161)	0(0)
2.00	0.00	2.362(1.493)	0.316(0.074)	0.278(0.079)	0.167(0.236)	0(0)
2.00	0.30	2.048(1.354)	0.296(0.081)	0.267(0.078)	0.2(0.233)	0(0)
2.00	0.50	1.909(0.937)	0.333(0.067)	0.267(0.078)	0.2(0.233)	0(0)
2.00	1.00	2.061(1.517)	0.381(0.063)	0.256(0.075)	0.233(0.225)	0(0)
2.00	1.50	2.074(1.555)	0.406(0.068)	0.256(0.075)	0.233(0.225)	0(0)
2.00	2.00	1.785(0.962)	0.488(0.058)	0.267(0.078)	0.2(0.233)	0(0)

Tabella 3.4: Risultati del metodo di clustering penalizzato con diversi valori di λ_μ e λ_τ nello scenario **C**. Si riporta ERM($\boldsymbol{\mu}$) e ERM($\boldsymbol{\tau}/\boldsymbol{\xi}$) calcolato sulle 10 repliche del modello di simulazione; il tasso di sparsità (e standard error) e del tasso di errore di sparsità (e stardard error) e l'indice CER (e standard error).

3.4 Robustezza del metodo di clustering a variazioni delle etichette degli spot

Il metodo di clustering dei geni proposto in questa tesi si basa sul confronto delle distribuzioni spaziali delle relative espressioni in aree predefinite degli *spot*, tipicamente basate sui risultati di un annotazione patologica. Il processo di annotazione degli *spot* è spesso effettuato manualmente da patologi attraverso l'analisi di immagini al microscopio del tessuto in esame, come descritto in precedenza. Questo processo è complesso e comporta la valutazione di diverse caratteristiche del tessuto, per cui non è rara l'identificazione di diverse configurazioni degli *spot* a seconda di chi esegue l'annotazione. Pertanto, diventa importante valutare la robustezza del metodo rispetto alle perturbazioni delle etichette degli *spot*.

A tale scopo, vengono simulate 10 repliche dell'esperimento descritto nella prima sezione del capitolo, per ciascuno dei tre scenari per la struttura della matrice dei valori *spatial signal-to-noise ratio*. Successivamente, si introduce un errore di classificazione nelle etichette basate sull'annotazione patologica pari al 5% degli *spot* (per un totale di 30 spot), selezionate casualmente e indipendentemente dalla posizione e dal cluster di appartenenza. Si riporta in Figura 3.5 la nuova configurazione del tessuto. Utilizzando queste nuove etichette di colonna, si applica il metodo di clustering dei geni su ogni replica e si valuta la discrepanza dei risultati rispetto alla corretta configurazione dell'insieme dei geni, attraverso l'indice CER. Le impostazioni dell'algoritmo sono quelle usate nel Paragrafo 3.2.

3.4.1 Risultati

Sono riportate in Figura 3.4 le distribuzioni dell'indice CER per ogni scenario di simulazione. Nel primo scenario **A**, l'indice CER medio è pari a 0.02, il doppio rispetto a quello ottenuto dal metodo adattato con le etichette di colonna corrette, ma comunque basso. Nel secondo scenario **B**, si osserva una maggiore variabilità degli indici calcolati: in alcune repliche il metodo di clustering ottiene una discrepanza sostanziale rispetto alla configurazione corretta (con un massimo pari a 0.21), mentre in altre repliche la discrepanza è minima o nulla. Per questo motivo si sospetta che la discrepanza rilevata

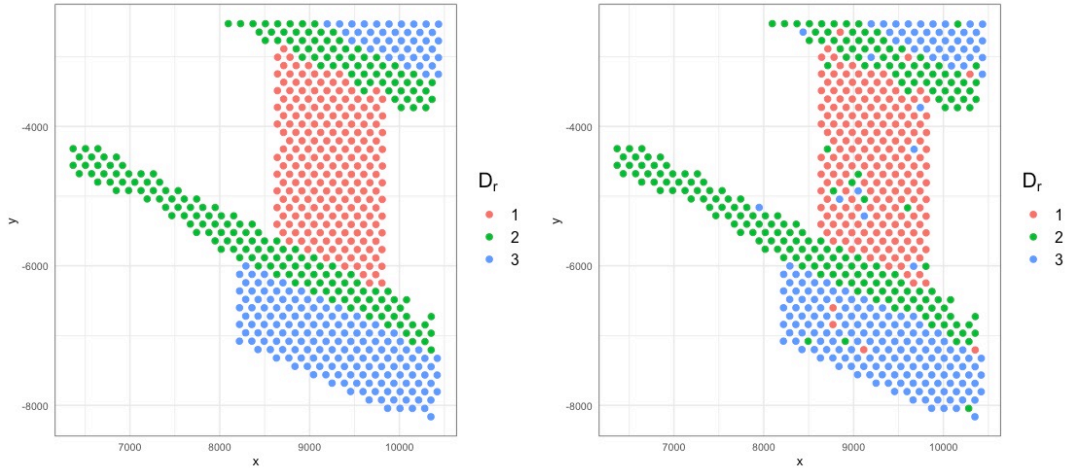


FIGURA 3.5: A sinistra: mappa degli *spot* colorati rispetto alla corretta configurazione in cluster dei geni. A destra: configurazione dei geni perturbata, utilizzata dal metodo negli studi di robustezza per ogni scenario.

in alcune repliche non sia dovuta all'errata specificazione delle etichette di colonna, ma piuttosto alla struttura spaziale specifica di tali repliche. Nel terzo scenario **C**, il metodo di clustering riesce sempre a identificare la classificazione corretta dei geni, nonostante le variazioni nelle etichette di colonna, indicando una forte robustezza del metodo in presenza di una importante struttura spaziale. Gli errori delle stime dei parametri del *spatial signal-to-noise ratio* aumentano come previsto, mentre quelli relativi ai parametri di media diminuiscono ma restano confrontabili con il modello stimato a partire dalle etichette di *spot* corrette (si veda Tabella 3.5).

	A	B	C
ERM($\boldsymbol{\mu}$)	5.636(2.915)	3.643(3.008)	5.49(1.904)
ERM($\boldsymbol{\tau}/\boldsymbol{\xi}$)	0.202(0.047)	0.729(0.525)	0.489(0.101)

TABELLA 3.5: Risultati del metodo di clustering applicato su 10 repliche del modello di simulazione e etichette di *spot* perturbate. Si riporta l'errore ERM($\boldsymbol{\mu}$) e ERM($\boldsymbol{\tau}/\boldsymbol{\xi}$) per ognuno dei tre scenari **A, B, C**, con i relativi standard error.

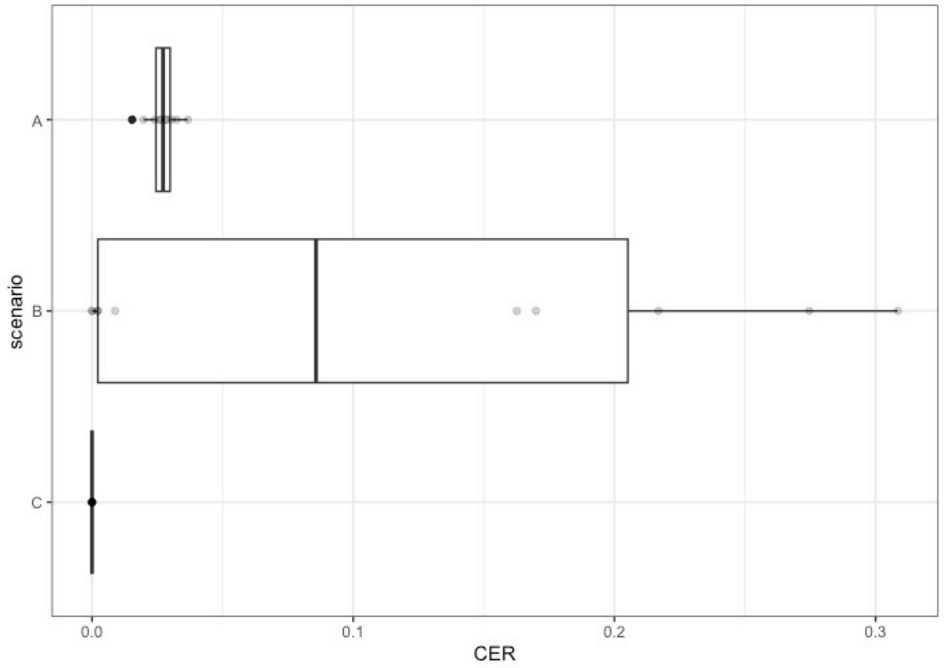


FIGURA 3.6: Boxplot dei valori dell'indice di discrepanza CER tra la corretta configurazione in cluster dei geni e la configurazione stimata dal metodo applicato alle 10 repliche del modello di simulazione per ognuno dei tre scenari e con etichette di *spot* perturbate

Capitolo 4

Caso studio

Si applica il metodo di clustering di geni proposto in questa tesi per analizzare i dati di trascrittomica spaziale relativi ad un tessuto di prostata affetto da adenocarcinoma, descritti nel Paragrafo 1.3. Nel corso dell'analisi, verrà illustrato come utilizzare il metodo e come interpretare i risultati ottenuti. Inoltre, verranno confrontati i risultati con quelli ottenuti dalla versione penalizzata del metodo al fine di valutare l'accuratezza delle stime del modello. Si conduce un'ulteriore analisi per valutare la presenza di una struttura spaziale nelle attività dei processi biologici noti nei tessuti di cancro, tramite l'applicazione del metodo di clustering ai punteggi delle *signature* ottenute dai dati di espressione genica. Infine, vengono mostrati i risultati principali dell'analisi approfondita sui cluster di geni identificati dal metodo di clustering, tenendo in considerazione anche i risultati sulle *signature*.

4.1 Analisi preliminari

Prima di procedere all'analisi dei dati di espressione genica è fondamentale eseguire delle operazioni preliminari di filtraggio e normalizzazione per ridurre la complessità dei dati e migliorare la qualità delle informazioni raccolte. In queste operazioni preliminari non viene considerata la struttura spaziale dei dati.

Il filtraggio consiste nella selezione dei geni più informativi. Questo passaggio consente di limitare le analisi ad un sottoinsieme di geni che sono più rilevanti per la domanda di ricerca o per il processo biologico in esame, eliminando i dati affetti da rumore o da

distorsioni tecniche, che potrebbero rendere meno evidenti delle differenze importanti tra i geni. Inoltre, la riduzione della dimensionalità del dataset aumenta il potere statistico di individuare differenze significative nell'espressione dei geni tra i campioni, e facilita l'interpretazione dei risultati perché si studia un insieme più piccolo di dati.

La normalizzazione dei dati mira a ridurre le distorsioni sistematiche dei profili di espressione genica tra gli *spot* non dovute a differenze biologiche. Queste variazioni possono essere causate da una serie di fattori, come ad esempio le variazioni del numero di cellule che compongono gli *spot* nel tessuto; le variazioni nell'efficienza durante la procedura amplificazione del materiale (PCR), e altri. Se i dati di espressione genica non vengono normalizzati, le variazioni nella quantità di RNA possono mascherare le reali differenze nell'espressione genica, rendendo difficile la comparazione tra *spot*, o evidenziando differenze che non sono biologicamente rilevanti.

Nella procedura di selezione dei geni più informativi comunemente adottata, si eseguono due operazioni indipendenti. La prima consiste nel filtraggio dei geni che presentano un livello di espressione inferiore a una soglia prestabilita, basata sui dati raccolti. La seconda consiste nella selezione dei geni con una alta variabilità di espressione. La scarsa espressione e la bassa variabilità sono considerati possibili indicatori di geni che potrebbero non avere un ruolo significativo nella biologia del tessuto in studio e quindi non fornire molte informazioni sui processi sottostanti. Quest'ipotesi sembra essere intuitiva, anche se in alcuni casi potrebbe non essere vera, ad esempio potrebbero esserci geni espressi solo in pochi *spot*, quindi con una media di espressione bassa, ma che sono marcatori di fenomeni rari o tipi di cellule rare. Sicuramente, però, dal punto di vista della modellazione statistica, questi tipi di misurazioni è più facile che siano meno affidabili: i conteggi relativi a questi geni è più frequente che siano affetti da rumore tecnico e che abbiano una variabilità maggiore; soprattutto è più difficile discernere quanto dell'espressione è dato dal rumore tecnico e quando da una variabilità biologica.

L'operazione di normalizzazione più diffusa nelle analisi di dati RNA-seq consiste in un riscalamento del totale dei conteggi grezzi di ogni *spot* per un fattore pari a 10^6 (*counts per million*, CPM) o con fattori che tengano conto della diversa quantità di molecole di RNA presenti negli *spot*. Tali fattori possono essere calcolati con diversi approcci, ad esempio utilizzando la funzione `estimateSizeFactorsFromMatrix` del pacchetto `R`

`DESeq2` (Anders Huber, 2010; Love et al., 2014) o la funzione `calcNormFactors` (Robinson Oshlack, 2010) del pacchetto `R edgeR`. Tipicamente si procede poi ad una trasformazione logaritmica per cercare di ridurre l'asimmetria della distribuzione. Queste procedure standard sono relativamente semplici da comprendere e calcolare, ma presentano alcune criticità nel momento in cui le si vogliono applicare ai conteggi UMI (Townes et al., 2019). Per questo, Townes et al. (2019) propongono un metodo alternativo basato sul modello multinomiale.

Nel metodo di Townes et al. (2019), la selezione dei geni più informativi avviene attraverso un confronto tra due modelli: un modello multinomiale nullo in cui si considera l'abbondanza relativa del gene costante tra gli *spot* del tessuto e un modello saturo che tiene conto della variabilità biologica nella stima. Per valutare la differenza tra i due modelli, viene utilizzata la statistica di devianza. Valori elevati della devianza indicano una forte evidenza contro il modello nullo e quindi la presenza di una variabilità biologica significativa. Il test viene eseguito per ogni gene. I risultati vengono ordinati rispetto i valori di devianza dei geni in senso decrescente e tramite un'analisi grafica, vengono selezionati i geni che presentano un valore di devianza alto prima del *plateau*. Il *plateau* rappresenta la maggior parte dei geni del database, che tipicamente ha un'espressione bassa e una variabilità moderata.

Successivamente, si calcolano i residui di devianza con segno, il cui valore assoluto rappresenta il contributo di ogni *spot* alla devianza totale di un gene, con segno. I residui di devianza con segno normalizzano e trasformano i dati in scala continua, rendendoli adatti per la modellizzazione nel metodo proposto in questa tesi. Questa trasformazione rende l'interpretazione dei risultati meno intuitiva, poiché la scala dei valori è basata sulla devianza e non più sul livello relativo di espressione genica. Dalle simulazioni fatte da Townes et al. (2019) risulta che valori dei residui prossimi a zero indicano poca variabilità o poca espressione, mentre valori più distanti indicano una maggiore variabilità e espressione. Nonostante la trasformazione non sia monotona, qualitativamente si mantiene la stessa scala dei dati di espressione: valori positivi dei residui indicano una sovra-espressione rispetto alla media nel tessuto; mentre valori negativi indicano una sotto-espressione.

4.1.1 Filtraggio e normalizzazione

Il dataset in studio comprende le misurazioni di espressione genica relative a 17943 geni in 4376 *spot*. Come prima selezione dei geni si individuano quelli che hanno almeno un conteggio UMI totale superiore ad una certa soglia. Per controllare che la scelta della soglia sia adeguata si esamina la distribuzione dei valori di espressione media in scala logaritmica tra tutti i geni. Secondo quanto suggerito da Lun et al. (2016), il picco della distribuzione rappresenta la maggioranza dei geni con espressione moderata, mentre la componente rettangolare corrisponde ai geni con espressione bassa. La soglia deve essere quindi impostata in un punto lungo la componente rettangolare per rimuovere la maggior parte dei geni che ha valori di abbondanza bassa. Dal grafico in Figura 4.1 sembra che una soglia pari a 50 sia adeguata, senza eliminare troppi geni in questa prima fase di filtraggio. Dopo questa selezione si mantengono i dati relativi a 13710 geni.

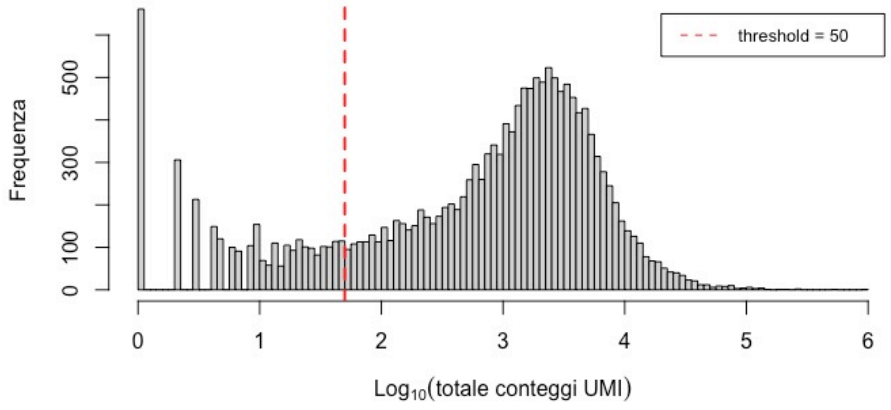


FIGURA 4.1: Istogramma dei conteggi log-media per tutti i geni nel dataset. La soglia per il filtraggio è rappresentata dalla linea rossa.

Si applica poi un secondo filtraggio con la procedura di selezione di geni descritta in Townes et al. (2019). Vengono calcolati i valori di devianza per ogni gene e successivamente ordinati in modo decrescente. Tramite l'analisi dei grafici in Figura 4.2 risulta sufficiente selezionare i primi 500 geni circa, ma si sceglie di seguire un approccio più conservativo e di includere nell'analisi i primi 1000 geni.

Infine, i dati di conteggio vengono normalizzati si calcolano i corrispondenti dei residui di devianza con segno utilizzando due diverse approssimazioni del modello multnomiale,

la distribuzione Binomiale e la distribuzione di Poisson. Si riportano in Figura 3.4 i boxplot dell'espressione media dei primi 100 geni negli *spot* del tessuto nella scala dei residui. Si osservi che non vi sono differenze significative tra le due approssimazioni. Di conseguenza, si opta per l'approssimazione basata sulla distribuzione Binomiale.

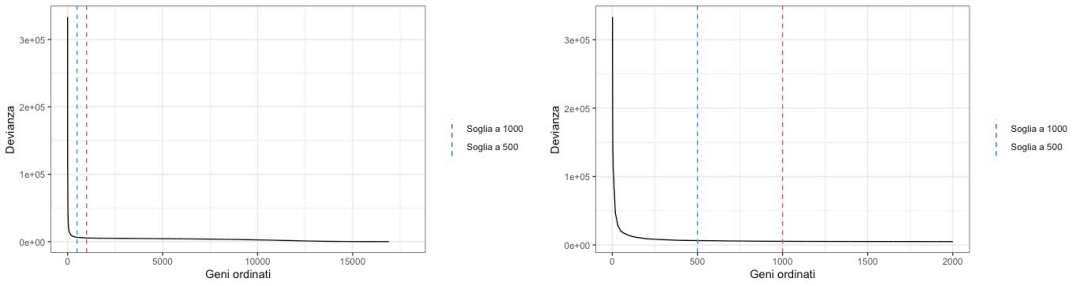


FIGURA 4.2: A sinistra: grafico dei geni presenti ordinati in senso decrescente in base al valore di devianza. Valori elevati di devianza sono associati a geni informativi. Anche se da una valutazione grafica il numero ideale di geni è di circa 500, si includono nell'analisi i 1000 geni con la devianza maggiore. A destra: stesso grafico, ma limitato ai primi 2000 geni sui 13710.

4.2 Analisi esplorative

Come descritto nel Paragrafo 1.3, gli *spot* del tessuto in studio sono stati annotati da un patologo. Si riportano in Figura 4.4 l'immagini al microscopio della sezione del tessuto utilizzata per l'annotazione e la corrispondente configurazione nelle categorie individuate. Si noti dalla Tabella 4.1 che la percentuale di *spot* identificati nella categoria di "Vasi sanguigni" è molto bassa. Dal momento che sono pochi *spot*, si sceglie di rimuoverli per non incorrere in problemi nella stima del modello.

Etichetta	**Totale spot**	**Percentuale** %
Vasi sanguigni	5	0.0011
Fibroblasti	371	0.0849
Ghiandole	287	0.0657
Stroma	1762	0.4031
Tumore	1946	0.4452

TABELLA 4.1: Elenco etichette di cluster definite dall'annotazione manuale del tessuto di prostata in analisi con il rispettivo numero totale di *spot* in ciascuno e la relativa frequenza percentuale

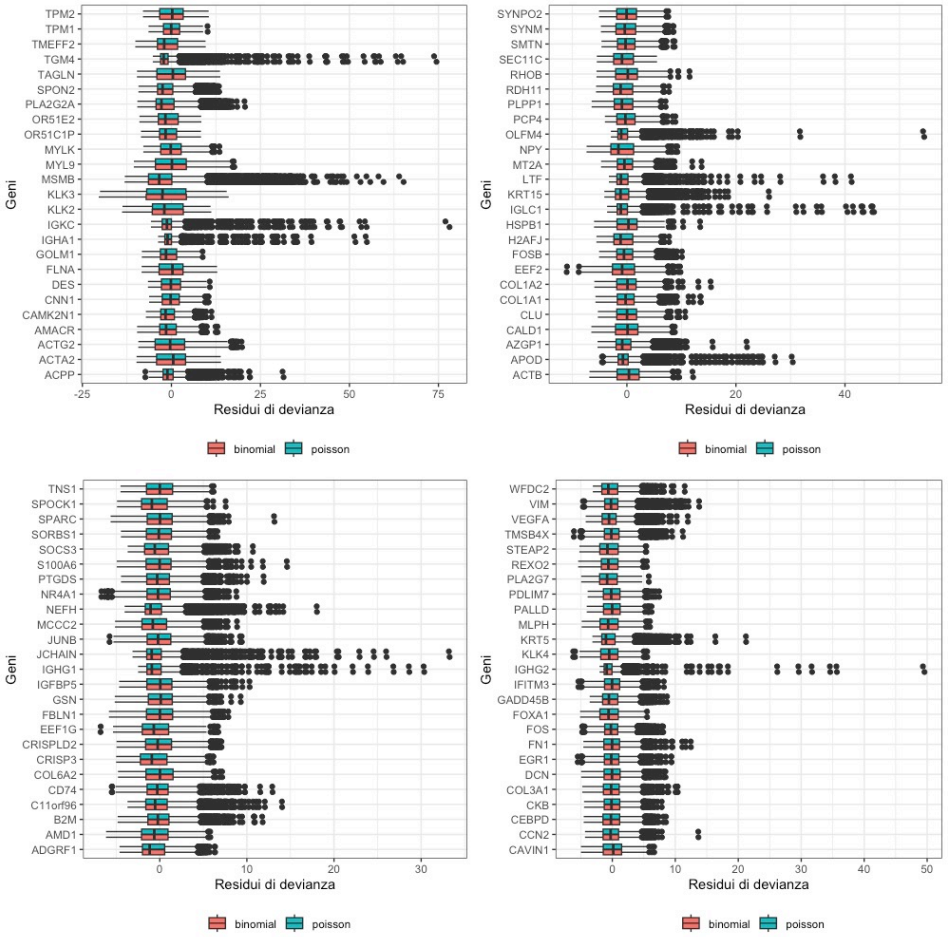


FIGURA 4.3: Boxplot dei residui della devianza dei primi 100 vettori riga $\mathbf{x}_i$ del campione di tessuto analizzato, calcolati utilizzando sia un'approssimazione del modello multinomiale basata che sulla distribuzione Binomiale che su quella di Poisson. Non si osservano eviednti differenze, dimostrando che i due metodi sono in pratica equivalenti su questo set di dati.

A scopo esplorativo si riportano in Figura 4.5 i grafici della distribuzione dei residui di devianza di quattro geni. Sono stati scelti i primi due geni con valori di devianza media su tutto il tessuto più alta (MSMB e KLK3), quello corrispondente al valore mediano di devianza (SRGN) e al valore minimo (SORL1), tra i 1000 geni in analisi. Si noti come la distribuzione del gene KLK3 è bimodale, indicando che il gene in alcuni *spot* è sotto-espresso, corrispondente ai valori negativi dei residui, e in altre zone è sovra-espresso, rispetto al valore medio su tutto il tessuto. Dai grafici degli istogrammi in Figura 4.5 si noti come i range di valori assunti siamo molto diversi tra i geni. Per questo motivo i grafici in Figura 4.6 delle relative espressioni nel tessuto vengono creati utilizzando una scala di colori specifica per ogni gene. Inoltre, per il gene MSMB e SORN si sceglie di eliminare gli *spot* che corrispondono a valori estremi della coda destra della distribuzione

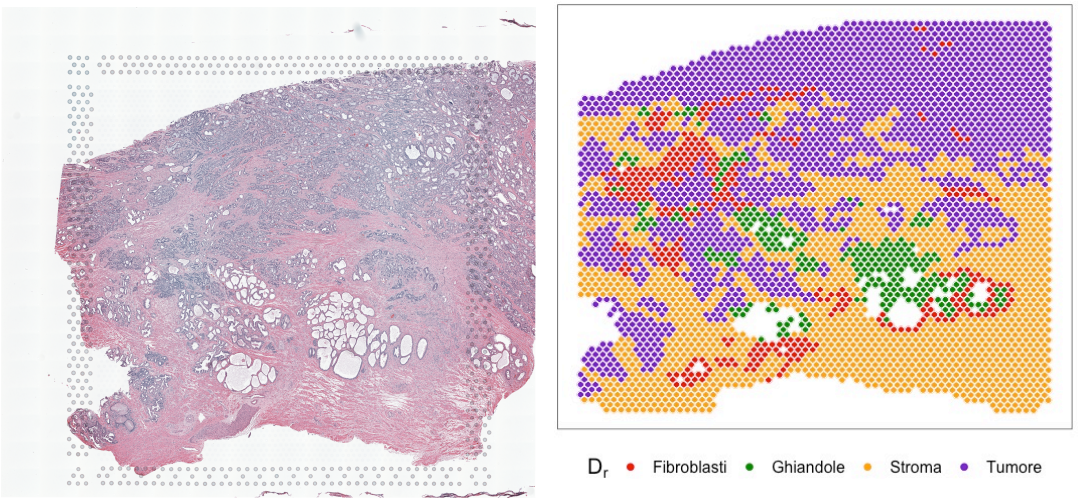


FIGURA 4.4: A sinistra: immagine al microscopio del campione di tessuto di prostata analizzato. A destra: mappa degli *spot* in cui viene suddiviso il tessuto colorati rispetto al cluster di appartenenza definito dall'annotazione patologica manuale.

(si eliminano 59 *spot* corrispondenti ai valori superiori a 3 volte il valore *upper-whiskers* per MSMB; mentre per SRGN si eliminano 15 *spot* corrispondenti ai valori superiori al relativo *upper-whiskers*). Si noti come per i geni SRGN e SORL1 il range di valori si riduce rispetto ai primi due geni presentati e la densità si concentra attorno lo zero.

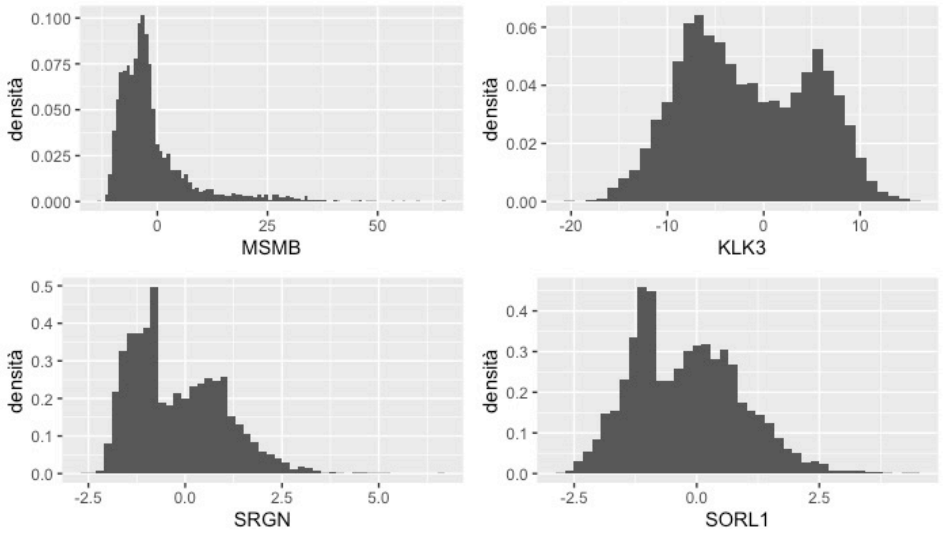


FIGURA 4.5: Istogrammi dei residui di devianza dei geni MSMB, KLK3, SRGN e SORL1 corrispondenti rispettivamente ai primi due geni con valore totale di devianza sul tessuto maggiore, al valore mediano e al valore minimo di devianza.

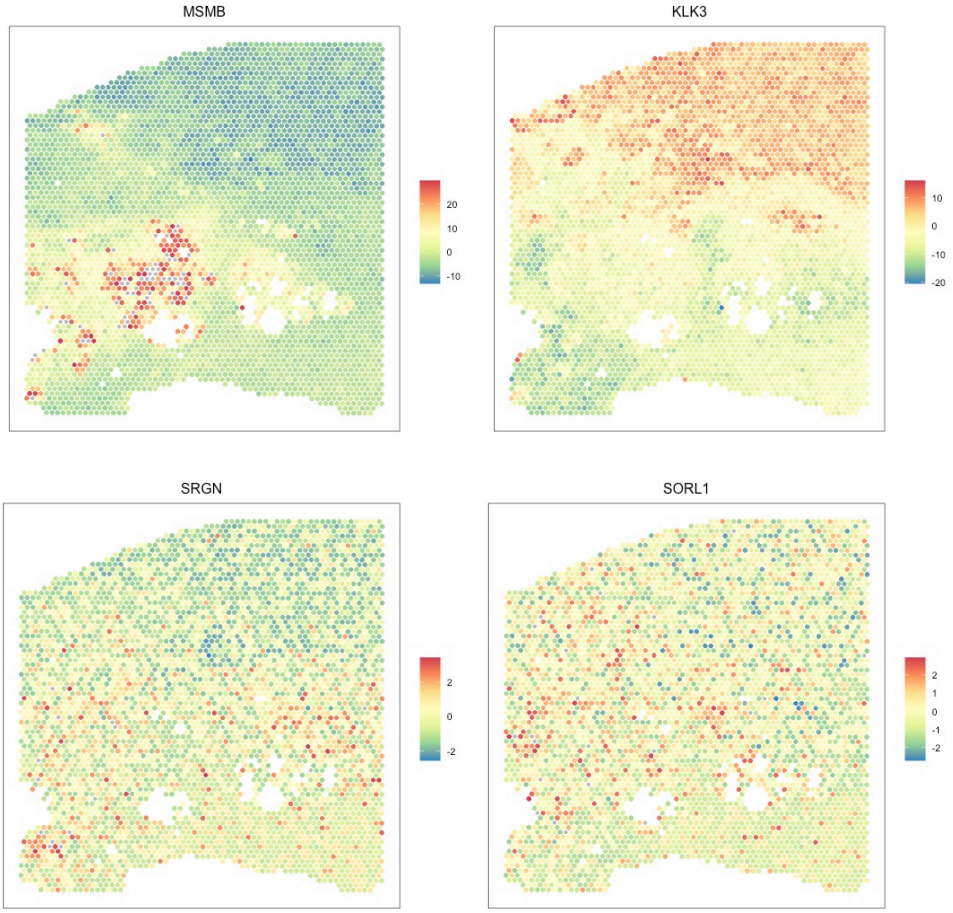


FIGURA 4.6: Grafico dell'espressione dei geni MSMB, KLK3, SRGN e SORL1 su tutto il tessuto nella scala dei residui di devianza.

4.3 Applicazione del metodo ai dati di espressione genica

Si applica il metodo di clustering dei geni presentato nel Capitolo 2 ai dati di espressione genica pre processati con le operazioni descritte nel paragrafo precedente. Le etichette di colonna utilizzate corrispondono all'annotazione patologica precedentemente descritta. Per selezionare il numero ottimale di cluster in cui suddividere l'insieme dei geni, vengono confrontate le configurazioni della matrice con un unico gruppo fino a 9 gruppi di geni, mantenendo la configurazione degli *spot* fissa.

Si descrive l'impostazione adottata per l'algoritmo di stima. Si utilizza un kernel esponenziale per modellare la dipendenza spaziale tra *spot*, seguendo l'approccio adottato nelle simulazioni e nell'applicazione del modello Spartaco dagli autori Sottosanti & Risso (2022). La stima di ogni modello è eseguita a partire da cinque punti iniziali scelti casualmente per controllare la convergenza dell'algoritmo. Il numero massimo di iterazioni dell'algoritmo è stato fissato a 3000, con 10 passi di massimizzazione per ciascuna iterazione. Nel caso in cui la verosimiglianza si stabilizzi prima, quando l'incremento all'ultimo passo è inferiore ad una soglia fissata a 10^{-4}, l'algoritmo o il passo di massimizzazione vengono interrotti. Uno studio preliminare del metodo su un dataset di dimensioni ridotte ha evidenziato la necessità di trasformare le coordinate spaziali in una scala tra 0 e 1 e aumentare il vincolo di identificabilità a 100, per evitare di incorrere in problemi di stabilità numerica.

Si ricorda che, nella notazione adottata nel capitolo 2, K rappresenta il numero di cluster in cui gli $n = 1000$ geni del dataset vengono partizionati; C_k indica il k-esimo gruppo all'interno di questa partizione e D_r rappresenta l'r-esimo gruppo della partizione degli *spot*.

4.3.1 Risultati

Le nove configurazioni stimate ($K \in \{1, \dots, 9\}$) sono confrontate utilizzando il criterio di informazione ICL, come descritto nella Sezione 2.3.6 del Capitolo 2. Il modello migliore, secondo questo criterio, è quello con valore maggiore dell'indice ICL. Dalla Tabella 4.2 e dal grafico in Figura 4.7 si osserva che i valori della log-verosimiglianza di

classificazione aumentano all'aumentare del numero di cluster di geni. I valori dell'ICL, invece, sembrano iniziare a stabilizzarsi dopo K = 5. Per $K = 1$ si ottiene un valore di ICL pari a -6908804, molto inferiore rispetto agli altri modelli, confermando la prensenza di una strutura in cluster dell'insieme dei geni. Per tenere in considerazione la variabilità della struttura di co-clustering individuata (nel senso descritto nella Sezione 2.3.4), vengono valutate anche le discrepanze tra configurazioni identificate dalle 5 stime di ogni modello, per lo stesso valore di K. Ciò è importante per evitare di selezionare un numero di cluster K in cui suddividere l'insieme dei geni che sembra buono, ma che corrisponde a molteplici partizioni plausibili. Nella Tabella 4.2 sono riportate tali misure di incertezza per ogni cluster, calcolate utilizzando la Formula (2.6) della Sezione 2.3.4, e il corrispondente valore medio utilizzato come misura di sintesi dell'incertezza della partizione a tale risoluzione. Dopo il modello con K = 2, il modello con media di errore minore è quello con K = 5.

Per favorire la comparazione tra i modelli che utilizzano diversi numeri di cluster di geni, si utilizza un "albero di clustering" implementato nel pacchetto `R clustree` di Zappia & Oshlack (2018). Questo tipo di grafico consente di esplorare visivamente l'impatto della variazione della risoluzione del clustering sulla partizione, mostrando le relazioni tra i cluster nelle diverse risoluzioni. L'albero viene costruito in modo che i nodi rappresentino i cluster individuati dal metodo, ordinati in livelli che indicano una diversa risoluzione di clustering. Gli archi, invece, rappresentano la transizione degli elementi attraverso tali risoluzioni. La grandezza dei nodi è definita in base al numero di elementi presenti nel gruppo, mentre il loro colore indica la risoluzione di appartenenza. Gli archi vengono colorati in base al numero di elementi che rappresentano. La trasparenza degli archi dipende dalla proporzione di elementi condivisi da due cluster che si sovrappongono, rispetto al numero totale di elementi in quei due cluster. Questa metrica, denominata metrica "in-proporzione", evidenzia gli archi più importanti: un'alta sovrapposizione tra due cluster rappresenta un raggruppamento ben definito di elementi, mentre una scarsa concordanza indica che il nuovo raggruppamento degli elementi è meno evidente. Gli alberi di clustering permetto quindi di individuare quali cluster rappresentano gruppi effettivamente simili di elementi e quali invece sono il risultato di una suddivisione eccessiva del dataset. Per i cluster che rimangono stabili mentre la

complessità del clustering aumenta, e dove ci sono archi a bassa in-proporzione, allora è probabile che si tratti di raggruppamenti "reali". Con queste analisi, si evita di considerare cluster che sono solo il risultato di una suddivisione troppo fine dei dati (*overclustering*). In sintesi, gli alberi di clustering permettono di visualizzare come i cluster si suddividono all'aumentare della risoluzione, quali cluster sono ben definiti e separati, quali sono correlati tra loro e come gli elementi si spostano da un gruppo all'altro durante il processo di clustering. Queste informazioni possono essere utili per la scelta della partizione migliore dell'insieme dei dati. E' importante sottolineare le relazioni tra i cluster a diverse risoluzioni sono individuate a posteriori della stima, ovvero la suddivisione dell'insieme dei geni in K gruppi non è influenzata dalle suddivisioni con risoluzione inferiore, a differenza invece di quanto accade per i metodi di clustering gerarchici. Il grafico in Figura 4.8 mostra l'albero di clustering relativo alle otto partizioni analizzate. Si può notare che, partendo dalla suddivisione iniziale in due cluster (K = 2), che ha misura di incertezza pari a 0, la stessa suddivisione viene mantenuta perfettamente fino a K = 4, perché le ulteriori suddivisioni avvengono all'interno dei cluster già individuati. A partire da K = 5, la suddivisione inizia ad alterarsi e vengono creati dei cluster che contengono osservazioni provenienti da cluster diversi, ad esempio il cluster $C_1^{K=5}$ contiene osservazioni che appartengono sia al cluster $C_2^{K=4}$ che al cluster $C_4^{K=4}$, mentre il cluster $C_4^{K=4}$ contiene osservazioni provenienti da $C_1^{K=4}$ e $C_3^{K=4}$, dove $C_i^{K=j}$ denota il cluster di geni i-esimo della partizione a j gruppi. Tuttavia, fino a $K = 7$, la configurazione complessiva dei geni in gruppi rimane sostanzialmente invariata, con un numero totale di spostamenti ridotto. A partire da $K = 8$, la configurazione in cluster inizia ad esserci più rumore rappresentando una possibile maggiore instabilità nella suddivisione. È interessante notare come l'albero individui tre rami principali di suddivisione, in cui cluster $C_2^{K=3}$ rimane sostanzialmente stabile all'aumentare della risoluzione.

Considerando i confronti fatti tra le diverse partizioni dei geni candidate, è stata selezionata la suddivisione dell'insieme dei geni in K = 5 cluster. Questa scelta è basata su diversi fattori: risulta esserci una buona evidenza a favore di un unica configurazione in 5 cluster, con un valore medio di incertezza tra i le stime parallele pari a 0.1; la bontà di adattamento ai dati è simile a quella dei modelli a risoluzione maggiore (ICL

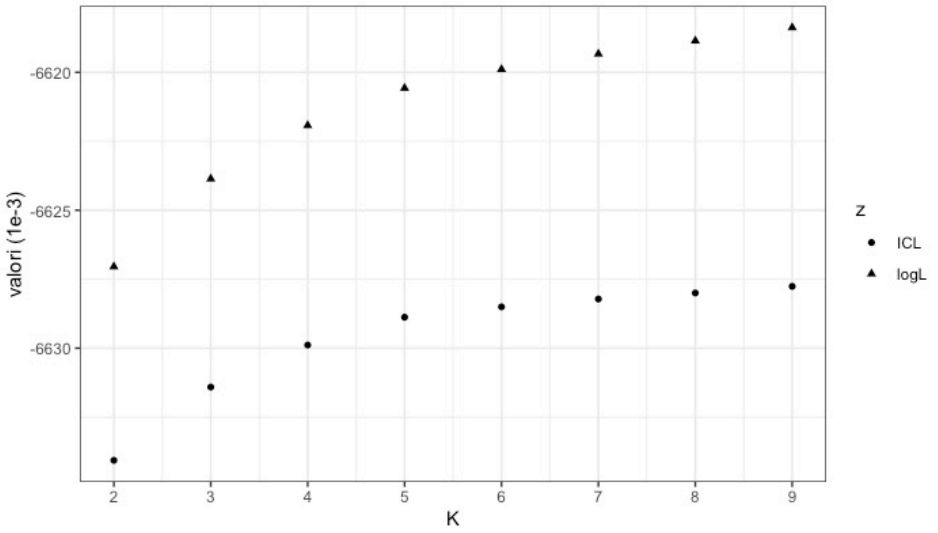


FIGURA 4.7: Si confrontano i valori di log-verosimigilianza di classificazione e l'ICL dei modelli con diverse risoluzioni di clustering dei geni $K = 2, \ldots, 9$

= -6628) e, infine, dall'albero di clustering emerge che questa suddivisione risulta essere complessivamente stabile, ma con un numero ridotto di cluster rispetto a quelli con risoluzione maggiore (K = 6 e K = 7), il che rende più agevole l'interpretazione dei risultati.

K	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	$\bar{C}_k$	ICL/10^3
2	0	0	-	-	-	-	-	-	-	0.00	-6634.07
3	0.012	0.012	0.023	-	-	-	-	-	-	0.02	-6631.41
4	0.008	0.083	0.039	0.063	-	-	-	-	-	0.05	-6629.89
5	0.023	0	0.004	0.006	0.016	-	-	-	-	0.01	-6628.88
6	0.009	0.002	0.03	0.027	0.013	0.057	-	-	-	0.02	-6628.50
7	0.024	0.043	0.1	0.062	0.07	0.053	0.1	-	-	0.06	-6628.22
8	0.101	0.076	0.105	0.058	0.052	0.215	0.019	0.062	-	0.09	-6628.00
9	0.061	0.005	0.143	0.086	0.311	0.153	0.009	0.105	0.07	0.10	-6627.76

TABELLA 4.2: Risultati dell'applicazione del metodo di clustering applicato ai dati del tessuto di prostata in analisi con diverse risoluzioni $K = 2, \ldots, 9$, K numero di cluster di geni. Per ciascun valore di K si stima il modello cinque volte in parallelo e si calcola la quantità ϵ_k^{row} con $k = 1, \ldots, K$, che misura l'incertezza del clustering; si riporta anche il valore medio di queste misure $\bar{C}_k$ e infine il valore del criterio d'informazione ICL utilizzato per la selezione del modello ottimale. Si noti che ogni modello è stimato indipendentemente dagli altri, per cui non vi è nessuna predefinita relazione tra i cluster C_k identificati.

Si ricorda che il metodo proposto in questa tesi non viene adattato direttamente ai dati di espressione genica, ma ai corrispondenti valori dei residui devianza con segno.

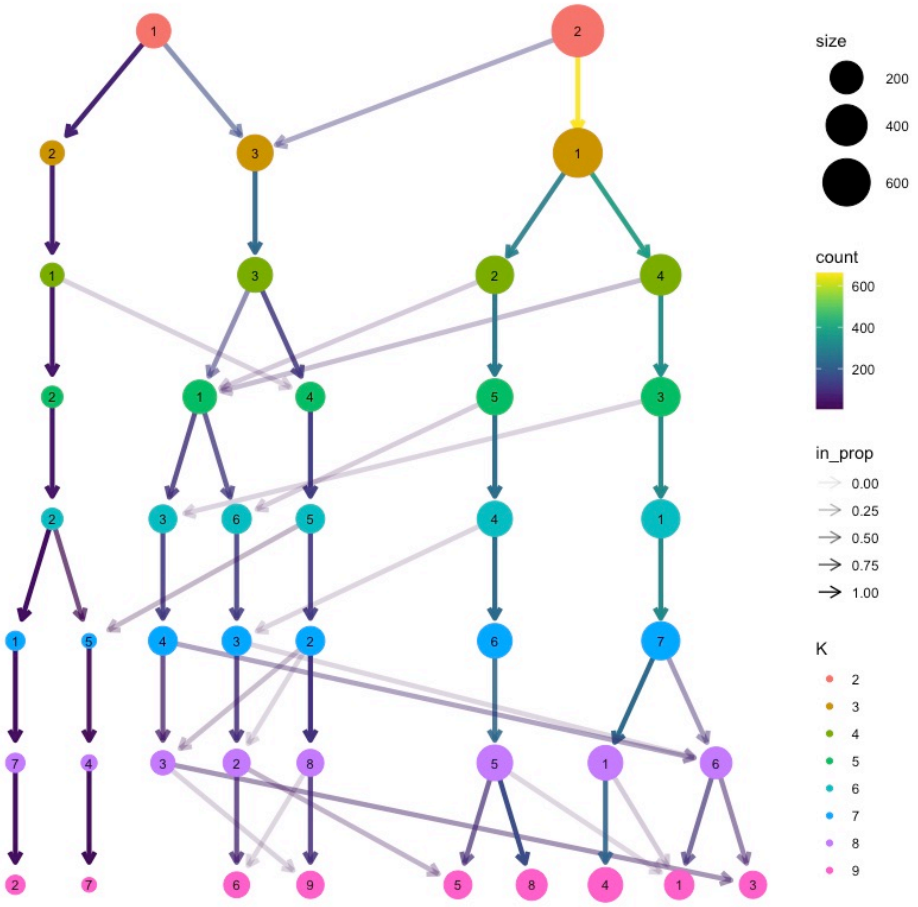


FIGURA 4.8: Albero di clustering relativo alle partizioni stimate sull'insieme dei geni dai metodi a diverse risoluzioni $K = 2, \ldots, 9$. Si può notare che, partendo dalla suddivisione iniziale in due cluster (K = 2), che ha misura di incertezza pari a 0, la stessa suddivisione viene mantenuta perfettamente fino a K = 4, perché le ulteriori suddivisioni avvengono all'interno dei cluster già individuati. A partire da K = 5, la suddivisione inizia ad alterarsi e vengono creati dei cluster che contengono osservazioni provenienti da cluster diversi

Tuttavia, qualitativamente, rimane l'ordinamento intuitivo dei dati di espressione: valori bassi dei residui in un certo *spot* corrispondono a geni poco espressi, mentre valori alti indicano geni molto espressi. Per questo motivo, si utilizzerà comunque il termine "espressione genica" per descrivere i risultati ottenuti, pur essendo consapevoli che non rappresenta un'interpretazione corretta dei dati di residui devianza, ma rende più scorrevole l'esposizione.

Il metodo di clustering utilizzato non si limita a fornire una semplice partizione dei geni in cinque gruppi. A differenza degli algoritmi tradizionali come *k-means*, il metodo proposto stima contemporaneamente al clustering un modello per i dati, il che

consente di esplorare le differenze a livello di distribuzione nelle quattro aree del tessuto (Fibroblasti, ghiandole, stroma e tumore), utilizzando misure di sintesi come la media e lo *spatial signal-to-noise ratio*, che hanno portato alla suddivisione dei geni in cluster.

	D_1	D_2	D_3	D_4
		Stima di μ		
C_1	-0.25	-0.14	-0.26	-0.28
C_2	-0.63	-1.00	-0.57	-0.67
C_3	0.05	0.09	-0.02	-0.34
C_4	-0.26	-0.05	-0.27	-0.33
C_5	-0.52	-0.48	-0.48	-0.08
		Stima di τ/ξ		
C_1	0.54	0.49	0.55	0.80
C_2	4.14	3.77	5.13	7.12
C_3	0.15	0.08	0.16	0.24
C_4	1.08	1.28	1.22	2.46
C_5	0.23	0.19	0.20	0.29
		Stima ϕ		
	0.51	0.42	0.49	0.45

TABELLA 4.3: Si riportano le stime dei parametri di media μ e *spatial signa-to-noise ratio* τ/ξ di ogni blocco (C_k, D_r) con $k = 1, \ldots, 5$ e $r = 1, \ldots, 4$. Infine si riportano le stime dei parametri di scala ϕ della funzione kernel esponenziale usata nella matrice di covarianza per modellare la correlazione spaziale.

Nella matrice dei dati è stata individuata una struttura in 20 blocchi, ottenuti dalla combinazione dei 5 gruppi di righe identificati e dei 4 gruppi di colonne corrispondenti all'annotazione patologica. Per ciascun blocco è stato stimato un modello gaussiano matriciale per la distribuzione corrispondente. Nella Tabella 4.3 sono riportate le stime dei parametri del modello per ciascun blocco: $\hat{\mu}_{kr}$, $\hat{\tau}_{kr}/\hat{\xi}_{kr}$, e $\hat{\phi}_r$, con $k = 1, \ldots, K = 5$ e $r = 1, \ldots, R = 4$. Si osservi che le stime del parametro ϕ sono simili nelle quattro aree del tessuto, suggerendo che il decadimento esponenziale nel kernel che modella la correlazione spaziale è praticamente costante nel tessuto. In altre parole, il numero di cellule che interagiscono non varia a seconda dell'istologia del tessuto, tuttavia può esserci una variazione di intensità. In Figura 4.9 si riporta la matrice dei dati con le righe e le colonne vengono ordinate in accordo con la sottostante struttura in blocchi. Nella matrice di sinistra i blocchi sono colorati rispetto al valore di media stimata $\hat{\mu}_{kr}$,

mentre nella matrice di destra rispetto al valore *spatial signal-to-noise ratio* stimato $\hat{\tau}_{kr}/\hat{\xi}_{kr}$. Questa suddivisione in blocchi e la relativa colorazione rendono la visualizzazione più chiara e immediata per l'analisi dei risultati. Dalla matrice di destra si nota una distinzione netta tra i gruppi di geni C_2, C_4 e C_3, C_4, C_5 in termini di importanza della variazione spaziale nel tessuto. Infatti, i geni appartenenti al cluster C_2 mostrano un'attività spaziale molto evidente, presente in tutte le aree del tessuto e particolarmente accentuata nella zona tumorale ($\hat{\tau}_{24}/\hat{\xi}_{24} = 7.12$). Anche per i geni in C_4, si rileva un'attività spaziale seppure più moderata. Si osservi che l'effetto spaziale stimato per entrambi i cluster segue la stessa progressione. Al contrario, per gli altri cluster si rileva una variabilità spaziale debole rispetto a quella residua ($\hat{\tau}_{kr}/\hat{\xi}_{kr} \leq 0.30$ per $k = 3, 5$ e $r = 1, \ldots, 4$). Infatti, i cluster C_3 e C_5 rappresentano una suddivisione di un cluster di origine e non presentano sovrapposizione con quello da cui proviene C_2 (si veda la Figura 4.8). Dall'analisi della matrice di sinistra emerge che la differenza principale dei cluster C_3 e C_5 è nella distribuzione dei valori di media: C_3 presenta valori di media bassi nella zona tumorale e valori elevati nel resto del tessuto, mentre C_5 mostra un andamento opposto, valori bassi nelle cellule tumorali e valori elevati nel resto del tessuto. Questo suggerisce la possibile presenza di geni marcatori per questo tipo di tumore all'interno di questi cluster. Analogamente, dal grafico in Figura 4.8 i cluster C_1 e C_4 possono essere visti che separazione dello stesso cluster originale, con valori di media simili in tutto il tessuto, ma si distinguono per l'intensità della variazione spaziale.

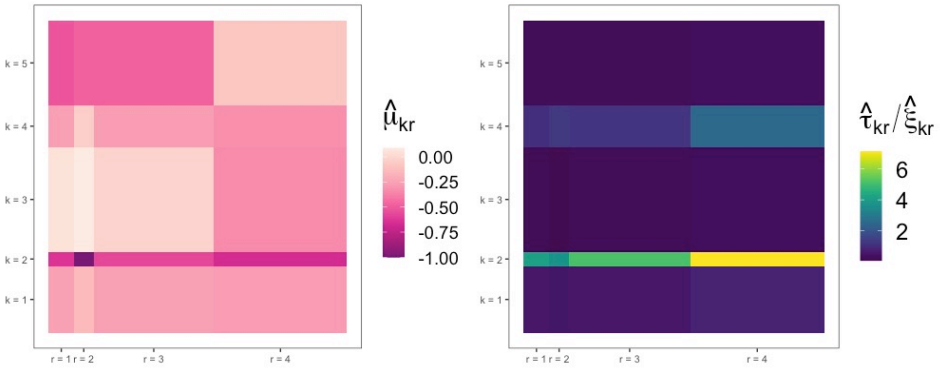


FIGURA 4.9: Matrice dei dati con geni e *spot* ordinati in accordo con la sottostante struttura in blocchi. I grafici sono colorati in base alla media stimata $\hat{\mu}_k r$ (a sinistra) e al rapporto segnale/rumore spaziale stimato $\hat{\tau}_{kr}/\hat{\xi}_{kr}$ (a destra).

I geni appartenenti al cluster C_2 mostrano una variazione spaziale comune così forte rispetto agli altri cluster che è già ben definita dal metodo con la suddivisione a 3 cluster e rimane stabile anche aumentando il numero di cluster in cui viene suddiviso il numero di geni. A scopo illustrativo, si riportano le espressioni di due geni, RRAS e MYLK, nella scala dei residui di devianza con segno. Il gene RRAS è coinvolto nella formazione di nuovi vasi sanguigni (angiogenesi) e sue alterazioni genetiche sono riscontrabili in numerosi tumori aggressivi (Sawada et al., 2015). Il gene appartiene al cluster C_3 e, infatti, i grafici in Figura 4.10 confermano quanto colto dal modello: la correlazione spaziale è debole e nella parte tumorale (figura destra) la media di espressione è inferiore rispetto al resto del tessuto (figura sinistra). La sotto-espressione di RRAS nella zona tumorale del campione di tessuto in studio potrebbe influire negativamente sulla capacità del tessuto di sviluppare nuovi vasi sanguigni, e quindi, nel prelievo di sostanze nutritive necessarie alla sua crescita. Il gene MYLK appartiene invece al cluster C_2 e infatti dai grafici in Figura 4.11 si può apprezzare la presenza di una forte correlazione spaziale, presente in tutto il tessuto. Anche questo gene risulta essere sotto-espresso nella zona tumorale e sovra-espresso nel resto del tessuto, come catturato del modello. Sebbene sia ancora un'ipotesi non confermata, diversi studi ritengono che il gene MYLK possa influenzare la capacità di invasione e di metastasi del cancro alla prostata (Dai et al., 2018). Il fatto che si osservi una sotto-espressione di due geni correlati positivamente con l'aggressività del tumore nella zona tumorale del tessuto, rispetto al resto, potrebbe fornire indicazioni sulla natura del tumore e rappresentare un punto di partenza per future ricerche biologiche.

Finora sono stati commentati i risultati relativi al comportamento comune dei geni appartenenti allo stesso cluster, in termini di valore medio di espressione e di variabilità spaziale media. Tuttavia, il metodo proposto in questa tesi ha come caratteristica distintiva quella di individuare i geni altamente variabili presenti in ogni area di *spot* annotata, ovvero quei geni che si discostano maggiormente dalla distribuzione media dei geni del cluster. Uno dei principali obiettivi della modellazione statistica applicata ai dati di trascrittomica è infatti quello di fornire un approccio automatico per la selezione dei geni più informativi, che possono rappresentare potenziali fonti di ispirazione per ulteriori studi biologici o per l'elaborazione di nuove ipotesi di ricerca. Nel modello

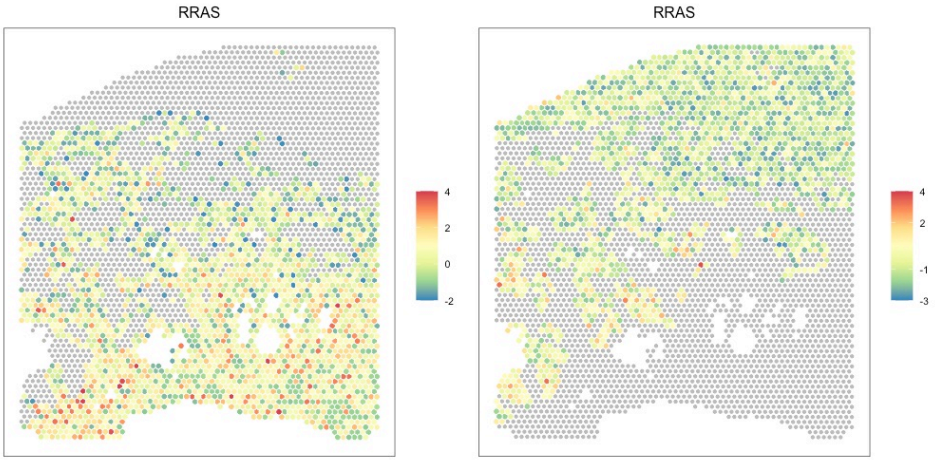


FIGURA 4.10: Grafico dell'espressione del gene RRAS nella zona dei fibroblasti, ghiandole e stroma a sinistra e nella zona tumorale a destra. Si nota una correlazione spaziale debole e una media di espressione inferiore nella parte tumorale rispetto al resto del tessuto.

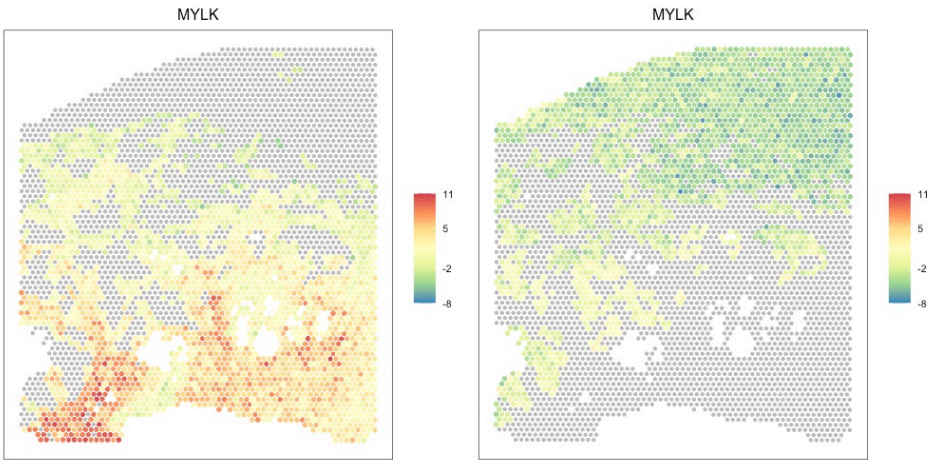


FIGURA 4.11: Grafico dell'espressione del gene MYLK nella zona dei fibroblasti, ghiandole e stroma a sinistra e nella zona tumorale a destra. Il gene risulta essere sotto-espresso nella zona tumorale e sovra-espresso nel resto del tessuto. Si noti la presenza di una forte correlazione spaziale, presente in tutto il tessuto.

proposto è presente un effetto casuale gene-specifico che mira a catturare la variabilità aggiuntiva rispetto a quella comune a tutti i geni del cluster di appartenenza per ogni area del tessuto. Ciò, oltre a favorire un adattamento migliore del modello ai dati, permette di quantificare quanto la distribuzione osservata dell'espressione del gene si discosti dalla distribuzione media tra tutti i geni del cluster. La variabile che modella la variabilità gene-specifica condizionata all'espressione osservata del gene i nel cluster di *spot* r-esimo è $\sigma^2_{\hat{\boldsymbol{z}}ir,i}|\mathbf{x}i\cdot^{\hat{\boldsymbol{z}}i}$. Si ritiene che i geni con valore di varianza gene-specifica attesa maggiore in una delle aree di *spot* annotate siano probabilmente informativi di

qualche meccanismo biologico che sta avvenendo, per via di una maggiore variabilità residua oppure di una variabilità spaziale con intensità maggiore rispetto a quella del gruppo. Grazie alla modellazione dell'espressione nelle sotto-aree del tessuto e non a livello globale, è possibile identificare geni altamente variabili anche solo in una delle quattro aree. Si osservi, ad esempio, la distribuzione dell'espressione dei geni VIM e ITGB8 nella zona stromale riportata in Figura 4.12. I due geni condividono un'evidente struttura spaziale, più accentuata per VIM, che mostra una variabilità maggiore di espressione. Infatti, appartengono entrambi allo stesso cluster C_2, ma VIM è tra i primi 20 geni più variabili nello stroma, mentre ITGB8 si trova al 49° posto. Questo esempio permette di cogliere anche un altro aspetto della modellazione: il modello stima la presenza di correlazione spaziale locale supposta costante in tutto il tessuto; da questi grafici si evince che in realtà il metodo è in grado di cogliere tale dipendenza anche se presente il pochi *spot* purchè sia molto accentuata.

Nella Figura 4.13 sono rappresentati i valori $\mathbb{E}(\sigma^2_{\hat{\boldsymbol{z}}ir,i}|\mathbf{x}i\cdot^{\hat{\boldsymbol{z}}_i})$ ordinati in senso decrescente, per ogni *spot* cluster, evidenziati in rosso i 20 valori più alti. Per motivi di chiarezza visiva, nella figura sono mostrati soltanto i primi 80 geni più variabili. Si noti che la maggior parte dei geni più variabili appartiene al cluster C_2 e C_4: su 133 geni altamente variabili per almeno uno *spot* cluster tra gli 80 selezionati, 45 appartengono a C_2 e 88 a C_4, nessuno al cluster C_1, C_3 e C_5. Si riportano nella Tabella 4.4 i venti geni con valore atteso di varianza gene-specifica a posteriori maggiore per ogni gruppo di *spot*. Si evidenziano i geni che risultano essere altamente variabili sono in un gruppo di *spot*. Nella tabella (4.5) si riportano i geni, e affianco del nome del gene, viene riportato tra parentesi l'etichetta corrispondente al/ai cluster/s di *spot* dove è stato identificato come gene altamente variabile. I risultati vengono confrontati con quelli ottenuti da altri metodi noti per la selezione dei geni altamente variabili: il metodo di Townes et al. (2019) e il metodo nnSVG (scalable identification of spatially variable genes using nearest-neighbor Gaussian processes) di Weber et al. (2022). Per ogni gene viene riportata la sua posizione rispetto all'ordinamento identificato da tali metodi.

La procedura di Townes et al. (2019), già utilizzata nelle analisi preliminari, ordina i geni sulla base esclusivamente della variabilità numerica senza tenere conto della struttura spaziale. Al contrario, il metodo nnSVG considera la struttura spaziale e ordina

i geni in base al valore della statistica di rapporto di log-verosimiglianza utilizzata per confrontare il modello spaziale, per la distribuzione dell'espressione su tutto il tessuto, con un modello lineare classico (si veda il Paragrafo 1.2 per maggiori dettagli). Da questo confronto emergono alcune importanti differenze tra la classificazione del geni più variabili rispetto alla distribuzione a posteriori della varianza gene-specifico e gli altri due metodi utilizzati: alcuni tra i geni i più variabili rispetto al metodo proposto in questa tesi non compaiono tra i primi 100 geni nelle altre classificazioni. È importante notare che la soglia che determina il numero da selezione in quanto più variabili è spesso arbitraria e limitata ad un gruppo piccolo di geni per semplificare le analisi successive. Tuttavia, ciò comporta il rischio di trascurare geni che potrebbero essere particolarmente informativi. Ad esempio, tra i primi 20 geni più variabili in D_3, rispetto al metodo di selezione proposto in questa tesi, compaiono il gene CD74, che è stato proposto come possibile bersaglio terapeutico in diversi tipi di cancro (Meyer-Siegler et al., 2006), e il gene VIM che è un importante marcatore della transizione epitelio-mesenchimale (EMT) nelle cellule tumorali (Zhang et al., 2019). Per il gruppo di *spot* D_2, tra i geni più variabili c'è il gene KRT7 che è associato a prognosi sfavorevole per pazienti con cancro alla prostata (Dariane et al., 2022). Nella Figura 4.14 si possono osservare le distribuzioni dei tre geni nelle rispettive aree del tessuto in cui risultano altamente variabili. Tuttavia, questi geni risultano, rispettivamente, al posto 65, 87, 188 nella classifica di Townes et al. (2019), e 94, 164 e 148 nella classifica di Weber et al. (2022). Tutti i tre geni risultano essere altamente variabili solo in un gruppo di *spot*. Questo evidenzia il fatto che non basta considerare la struttura spaziale se viene modellata sull'intero tessuto, senza limitarsi a zone specifiche. Infatti, i geni che risultano essere altamente variabili in tutte le quattro aree del tessuto sono nelle posizioni più alte della classifica di Weber et al. (2022), mentre i geni come VIM che presentano una variabilità elevata solo in una porzione del tessuto non vengono individuati.

In questo contesto in cui il tessuto è suddiviso in zone a priori della stima, si potrebbero applicare questi metodi di selezione su ciascun area separatamente, ma questo vuol dire fare un test per ogni area e quindi richiede un controllo della molteplicità del test. I vantaggi del metodo proposto per la selezione dei geni più informativi sono due: i geni più informativi vengono selezionati senza dover usare un test su ogni area; e inoltre si

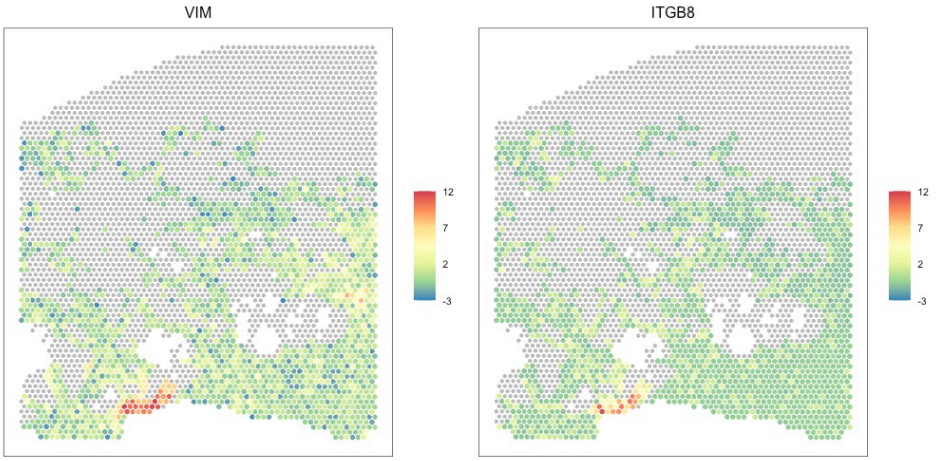


FIGURA 4.12: Distribuzione dell'espressione dei geni VIM (a sinistra) e ITGB8 (a destra) nella zona stromale. I due geni condividono un'evidente struttura spaziale, più accentuata per VIM che mostra una variabilità maggiore di espressione.

sfrutta il clustering dei geni per accorpare quelli che presentano un comportamento simile, il che permette di stimare in modo più affidabile la struttura spaziale e calcolare con maggiore precisione la variabilità intesa come scostamento dei geni dalla distribuzione comune. In conclusione si sottolinea come sia necessario prestare particolare attenzione alla scelta del metodo per la selezione geni più informativi, al fine di ottenere una selezione accurata che tenga conto dell'intera complessità del tessuto considerato.

	D_1	D_2	D_3	D_4
1	TGM4	TGM4	MSMB	MSMB
2	IGKC	MSMB	IGKC	KLK3
3	IGLC1	KLK3	ACTG2	ACPP
4	IGHA1	ACPP	MYL9	KLK2
5	MSMB	KRT15	ACTA2	**PLA2G2A**
6	MYL9	IGKC	KLK3	IGKC
7	KLK3	IGHA1	TAGLN	OLFM4
8	ACTA2	**KRT5**	FLNA	**LTF**
9	ACTG2	**NEFH**	TPM2	ACTA2
10	TAGLN	AZGP1	IGHA1	**SPON2**
11	FLNA	**KRT7**	MYLK	MYL9
12	KLK2	**MMP7**	APOD	IGHA1
13	TPM2	KLK2	DES	KRT15
14	MYLK	IGLC1	KLK2	**EEF2**
15	JCHAIN	OLFM4	ACPP	TAGLN
16	APOD	**IGHG1**	TPM1	APOD
17	DES	MYL9	**IGHG2**	**AMACR**
18	KRT15	ACTG2	**CD74**	TMSB4X
19	TPM1	TMSB4X	**VIM**	ACTG2
20	ACPP	JCHAIN	**FOSB**	**TMEFF2**

TABELLA 4.4: Si riportano i primi venti geni con valore atteso di varianza gene-specifica a posteriori maggiore per ogni gruppo di *spot* D_r con $r = 1, \ldots, 4$. Si evidenziano in grassetto i geni che risultano essere altamente variabili solo in un gruppo di *spot*.

Gene	**nnSVG**	**Townes**	**Gene**	**nnSVG**	**Townes**
ACP3 (1)(2)(3)(4)	1	20	IGLC1 (1)(2)	31	28
KLK2 (1)(2)(3)(4)	2	7	TPM1 (1)(3)	41	22
TGM4 (1)(2)	3	8	LTF (4)	45	42
TAGLN (1)(3)(4)	4	6	KRT15 (1)(2)(4)	46	32
IGKC (1)(2)(3)(4)	6	10	OLFM4 (2)(4)	49	50
ACTA2 (1)(3)(4)	7	5	IGHG2 (3)	59	79
MYL9 (1)(2)(3)(4)	9	3	APOD (1)(3)(4)	67	35
MSMB (1)(2)(3)(4)	12	1	FOSB (3)	69	43
DES (1)(3)	13	17	EEF2 (4)	71	30
KLK3 (1)(2)(3)(4)	14	2	MMP7 (2)	73	107
MYLK (1)(3)	15	15	AZGP1 (2)	85	36
ACTG2 (1)(2)(3)(4)	16	4	NEFH (2)	91	52
AMACR (4)	17	19	CD74 (3)	94	65
PLA2G2A (4)	18	9	KRT5 (2)	97	78
TPM2 (1)(3)	19	13	TMSB4X (2)(4)	115	83
IGHA1 (1)(2)(3)(4)	22	24	KRT7 (2)	135	118
TMEFF2 (4)	23	14	JCHAIN (1)(2)	146	61
FLNA (1)(3)	25	11	IGHG1 (2)	148	70
SPON2 (4)	27	12	VIM (3)	164	87

Tabella 4.5: Si riportano i geni identificati come altamente vairiabili dal metodo in almeno in un gruppo di *spot*, di cui l'etichetta viene riportata tra parentesi; si riportata la sua posizione rispetto all'ordinamento identificato dai metodi di Townes et al. (2019) e nnSVG di Weber et al. (2022).

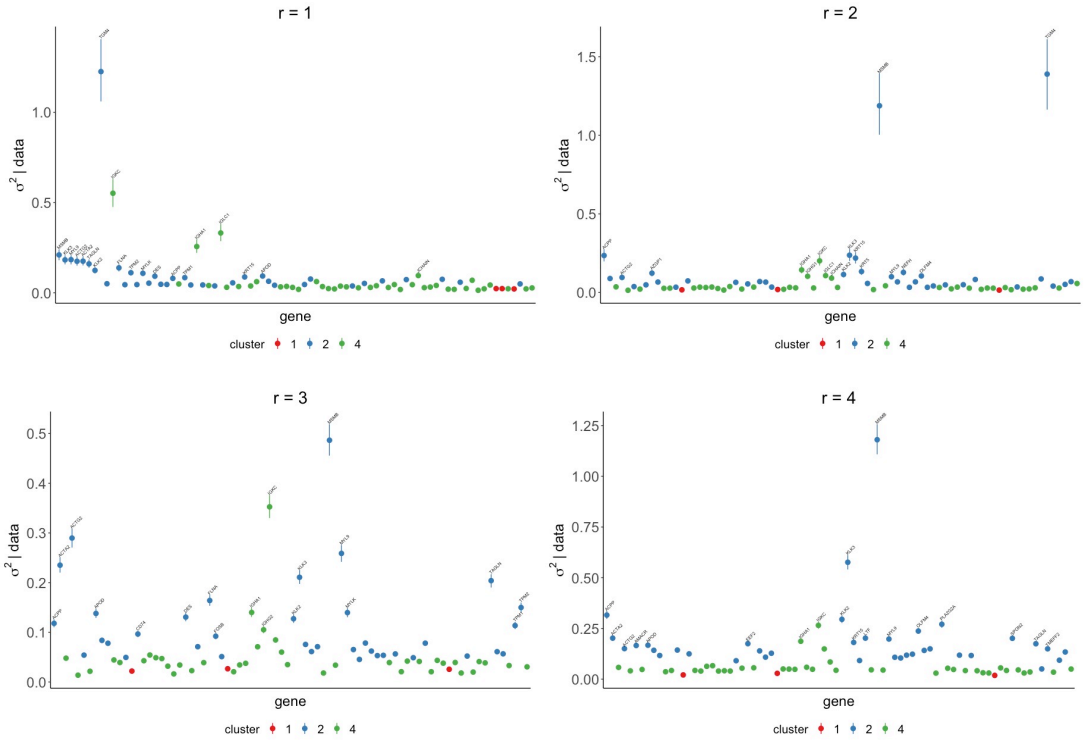


FIGURA 4.13: Ogni pannello fornisce la distribuzione di $\sigma^2_{\cdot r,i}|dati$. I punti indicano i valori attesi e le barre di errore gli intervalli di credibilità del 95%. Per ogni cluster di *spot*, si riportano i nomi dei venti geni con l'aspettativa maggiore. Tutti i punti sono colorati rispetto al cluster di geni di appartenenza.

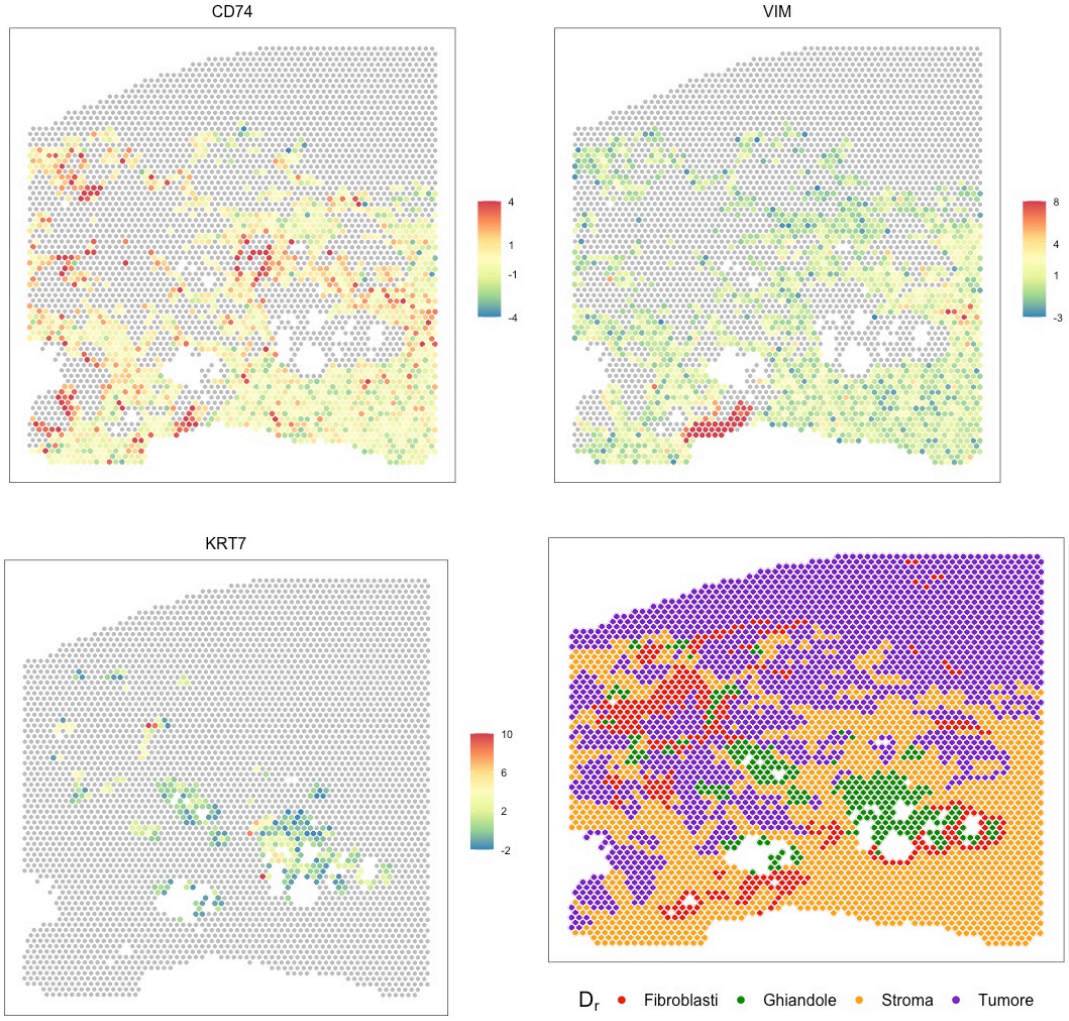


FIGURA 4.14: Distribuzione dell'espressione dei geni CD74 (in alto, a sinistra), VIM (in alto, a destra) e KRT7 (in basso, a sinistra) ognuno nella relativa area in cui risulta essere altamente variabile. In basso a destra si riporta la mappa degli *spot* colorata rispetto alla classificazione denotata dall'annotazione manuale del tessuto.

4.3.2 Risultati metodo di clustering penalizzato

Si applica il metodo di clustering dei geni penalizzato presentato nella Sezione 2.3.3 agli stessi dati di espressione genica pre-processati con una penalità $\lambda_\mu = 1.5$ e $\lambda_\tau = 0.3$. Si mantengono le stesse impostazioni dell'algoritmo utilizzate nel paragrafo precedente (4.3).

Si mostrano in Figura 4.15 la matrice dei dati in cui le righe e le colonne vengono ordinate in accordo con la sottostante struttura in blocchi. Nella matrice di sinistra i blocchi sono colorati rispetto al valore di media stimata $\hat{\mu}_{kr}$, mentre nella matrice di destra rispetto al valore *spatial signal-to-noise ratio* stimato $\hat{\tau}_{kr}/\hat{\xi}_{kr}$. Rispetto a quelle ottenute con il metodo di clustering non-penalizzato in Figura 4.9 non si osservano differenze evidenti, vengono accentuate le differenze che si erano già riscontrate tra i valori di media di blocco stimate. Inoltre, la struttura in cluster dei geni identificata dai due metodi è sostanzialmente la stessa (CER = 0.0110, che corrisponde a una differenza di etichette di 12 geni). Si può concludere che l'introduzione delle penalità sui parametri del modello di media e di *spatial signal-to-noise ratio* di blocco non alteri i risultati discussi nel paragrafo precedente e anzi confermi l'affidabilità dei parametri delle stime discusse.

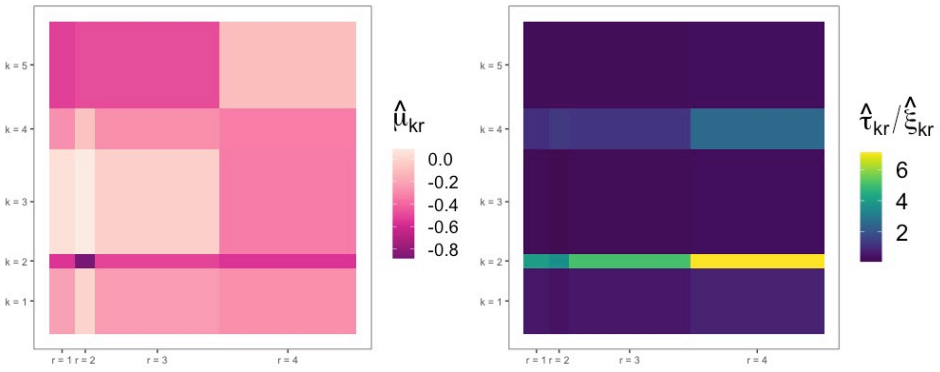


FIGURA 4.15: Matrice dei dati con geni e *spot* vengono ordinati in accordo con la sottostante struttura in blocchi. I grafici sono colorati in base alla media stimata $\hat{\mu}_k r$ (a sinistra) e al rapporto segnale/rumore spaziale stimato $\hat{\tau}_{kr}/\hat{\xi}_{kr}$ (a destra) dal metodo di clustering penalizzato con $\lambda_\mu = 1.5$ e $\lambda_\tau = 0.3$.

4.4 Applicazione del metodo alle *signature* oncologiche

In campo oncologico, le *signature* sono studiate per la loro capacità di evidenziare le attività tumorali (Pirrotta & Calura, 2023). Ogni stato biologico, sia esso uno stato di sviluppo, una risposta cellulare a segnali esterni o una malattia, si riflette in uno specifico profilo di espressione genica dell'organismo (Nevins & Potti, 2007). L'obiettivo di una *signature* di espressione genica è di sintetizzare in modo efficace una particolare attività biologica attraverso un insieme ridotto di geni, i cui livelli di espressione vengono combinati in un punteggio che quantifica la presenza dell'attività in questione.

Nel caso in studio si dispone dei dati di trascrittomica spaziale di un tessuto; avere l'informazione spaziale dell'espressione dei geni consente non solo di calcolare i punteggi, ma anche di mapparli sul tessuto e indagare eventuali schemi spaziali. È possibile che, come alcuni geni, anche per le *signature* possa esserci una dipendenza tra punteggi di *spot* adiacenti. Pertanto, può essere interessante applicare il metodo di clustering proposto in questa tesi per individuare gruppi di *signature* con distribuzioni spaziali simili e indagare quali hanno un significativo effetto spaziale. Si utilizzano le *signature* associate a processi oncologici raccolte nel pacchetto `R signifinder` di Pirrotta & Calura (2023). Sono state selezionate 19 *signature* relative a "Pan-tissue" e "Prostate", specifiche per i dati di RNAseq. Di queste 19, due sono state escluse per via dell'eccessivo numero di geni mancanti necessari per il calcolo del punteggio delle *signature* (superiore al 30%). Nella Tabella 4.6 sono riportati i nomi delle *signature* e il numero di geni che le costituiscono.

Si calcolano i punteggi delle *signature* selezionate per ogni *spot* del tessuto in studio. Su questo dataset si adatta il metodo di clustering per individuare eventuali similitudini nella distribuzione delle diverse *signature* e quantificare la spazialità dell'attività che rappresentano.

	Nome signature	N. geni		Nome signature	N. geni
1	EMT_Mak	77	10	CIN_Carter_70	70
2	Hypoxia_Buffa	49	11	CellCycle_Lundberg	463
3	ImmunoScore_Roh	41	12	CellCycle_Davoli	10
4	MitoticIndex_Yang	9	13	ASC_Smith	50
5	IFN_Ayers	6	14	ImmuneCyt_Davoli	7
6	ExpandedImmune_Ayers	18	15	ECM_Chakravarthy_up	30
7	Tinflam_Ayers	18	16	ECM_Chakravarthy_down	28
8	StemCellCD49f_Smith	90	17	VEGF_Hu	12
9	CIN_Carter_25	25			

TABELLA 4.6: Elenco delle *signature* selezionate per l'analisi del tessuto di prostata in studio con il numero di geni che si considerano nel calcolo del rispettivo punteggio.

4.4.1 Risultati

Si applica il metodo di clustering sulla matrice di dati contenente per riga valori dei punteggi di ciascuna *signature* in ognuno degli *spot* del tessuto. I punteggi vengono calcolati sui residui di devianza con segno e non direttamente sui conteggi di espressione dei geni in modo tale da normalizzare i dati da eventuali distorsioni tecniche. Inoltre, poiché le scale dei punteggi possono essere molto diverse tra loro, i valori vengono scalati a tra 0 e 1 per renderli confrontabili. Si utilizzano le etichette di colonna fornite dall'annotazione patologica. Le impostazioni del modello sono le stesse utilizzate nelle altre analisi fatte, ad esempio si veda il Paragrafo 4.3.

Si seleziona il modello di clustering confrontando diversi valori di $K \in \{1, 2, 3\}$. Il modello con $K = 2$ presenta un valore di ICL maggiore rispetto a $K = 1$ (ICL = -134154.1), confermando la presenza di un raggruppamento in cluster delle *signature*. Invece, con $K = 3$, l'algoritmo di stima si è arrestato poichè uno dei tre cluster si riduce fino a contenere una sola *signature*, indicando la forte evidenza contro la presenza di un terzo cluster. La configurazione in cluster sembra essere abbastanza chiara: tra le cinque stime parallele del modello si ottiene una misura di incertezza ϵ_{row} inferiore a 0.002. Si seleziona dunque il modello con due cluster, C_1 e C_2, con rispettivamente 13 e 4 *signature*.

Nella Figura 4.16 è mostrata la matrice dei dati, con le righe e le colonne ordinate in accordo con una struttura a blocchi sottostante. I blocchi sono colorati in base ai valori di media stimata $\hat{\mu}_{kr}$ nella matrice di sinistra, e in base al valore di *spatial signal-to-noise ratio* stimato $\hat{\tau}_{kr}/\hat{\xi}_{kr}$ nella matrice di destra. Il modello cattura una chiara

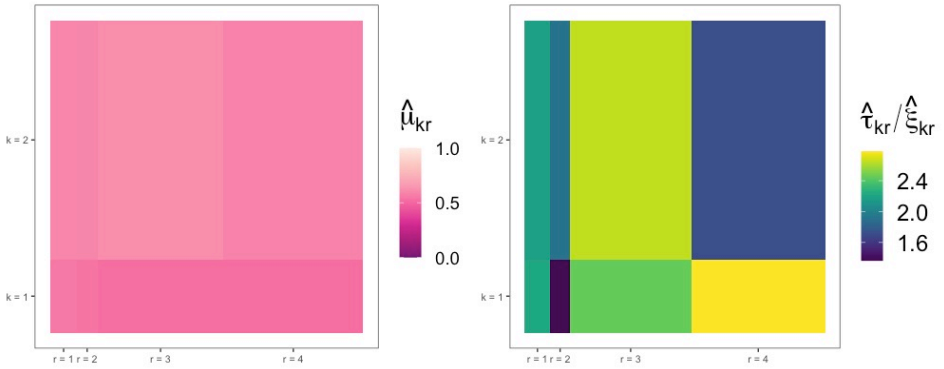


FIGURA 4.16: Matrice dei dati con *signature* e *spot* ordinati in accordo con la sottostante struttura in blocchi. I grafici sono colorati in base alla media stimata $\hat{\mu}_k r$ (a sinistra) e al rapporto segnale/rumore spaziale stimato $\hat{\tau}_{kr}/\hat{\xi}_{kr}$ (a destra).

struttura spaziale per entrambi i cluster di *signature* con spazialità sostanziale in tutte le quattro aree annotate del tessuto ($\hat{\tau}_{kr}/\hat{\xi}_{kr} > 1.36$). In particolare, il cluster C_2 mostra un maggiore effetto spaziale nella zona tumorale e inferiore nella zona stromale, mentre il cluster C_1 ha un comportamento complementare. Al contrario, non emergono differenze importanti nei valori di media, che risultano essere tutti prossimi a 0.5. Questo risultato sottolinea l'importanza di considerare la struttura spaziale delle *signature* nel processo di clustering: la mancata considerazione della struttura spaziale potrebbe portare a conclusioni errate e a una scarsa comprensione delle caratteristiche dei dati.

4.5 Caratterizzazione dei cluster di geni tramite analisi di arricchimento

In questo paragrafo si presentano i risultati più rilevanti di un'analisi estensiva sui cluster di geni individuate dal metodo di clustering. Per confrontare e confermare tali risultati, si sono utilizzati anche i risultati delle analisi sulle *signature*.

Una volta individuati i cluster all'interno dell'insieme di geni, l'obiettivo successivo è quello di fornire un'interpretazione di tali cluster. Tale interpretazione consiste nell'identificare se delle caratteristiche biologiche che accomunano i geni all'interno di ogni cluster, quindi rispondere alla domanda fondamentale: la somiglianza nella distribuzione spaziale dei geni è dovuta alla presenza di un qualche fattore biologico comune? E se sì, quale?

Per identificare le caratteristiche biologiche che accomunano i geni all'interno di un cluster, si può utilizzare l' *Over Representation Analysis* (ORA), una tecnica di analisi statistica che confronta la frequenza dei geni di un insieme specifico con quella di un insieme di controllo. Quest'ultimo è costituito da geni annotati con una categoria funzionale comune, come un processo biologico o una localizzazione cellulare, e raccolti nelle ontologie di geni. L'obiettivo è di identificare quali categorie dell'ontologia sono maggiormente associate ai geni del cluster in esame e ipotizzare quali processi biologici potrebbero essere responsabili per l'espressione simile dei geni. Si utilizza un test chi-quadro per ciascun *gene-set* dell'ontologia, calcolando i p-value aggiustati con la correzione di Benjamini e Hochberg (Benjamini & Hochberg, 1995) per controllare i problemi di molteplicità del test. In questo modo, si possono identificare i *gene-set* con maggiore evidenza di arricchimento e capire se un cluster è associato a una specifica categoria biologica.

Si conducono i test considerando diverse ontologie: Gene Ontology, KEGG, MSIGDBR considerando la collezione relativa all'oncogenetica C6, che si distinguono per il tipo di informazioni riportate. Gene Ontology si concentra sulla descrizione della funzione dei geni, KEGG sui *pathway* metabolici e molecolari e MSigDB C6 sugli insiemi di geni associati a *pathway* biologici specifici di oncogenetica. Le analisi di arricchimento sui cluster dei geni evidenziano due risultati principali. Il primo riguarda i geni del cluster C_3, questi sono maggiormente espressi fuori dalla massa tumorale, nel microambiente (TME) che la circonda, e hanno invece espressione più bassa al suo interno (nel tumore $\hat{\mu}_{34} = -0.34$; nei fibroblasti, nelle ghiandole e in stroma rispettivamente $\hat{\mu}_{31} = 0.05,\ \hat{\mu}_{32} = 0.09, \hat{\mu}_{33} = -0.02$). L'analisi di arricchimento di questi geni fatta sul database di Gene Ontology nella categoria Biological Proces mostra un forte coinvolgimento di questi geni in processi per lo più legati alla riorganizzazione morfologica delle cellule (p-value = 0.03). Questi processi includono la transizione epitelio-mesenchimale (EMT), l'organizzazione del citoscheletro e della matrice extracellulare e il riarrangiamento delle tube epiteliali. Questi risultati suggeriscono uno stato morfologico fortemente mesenchimale delle cellule che compongono il microambiente tumorale. Questo stato infatti caratterizza le cellule connettivali che qui troviamo associate al tumore (cancer associated fibroblasts, CAF) e che circondano la massa tumorale, di cui una

delle principali funzioni consiste infatti nella modifica del microambiente del tumore rimodellando la matrice extracellulare (Belhabib et al., 2021) e così favorendo la progressione del tumore. Un risultato analogo si ottiene analizzando la distribuzione della *signature* per la caratterizzazione dello stato di EMT proposta da Mak et al. (2016) nel tessuto, dove i punteggi positivi rispecchiano uno stato mesenchimale e quelli negativi uno più epiteliale. Si nota infatti come questa si accentui con punteggi più alti nelle zone dei CAF e dello stroma e presenti invece punteggi negativi principalmente all'interno della massa tumorale. Dalla Figura 4.17 si osserva inoltre come questi si distribuiscono diversamente tra le varie zone: zone più uniformemente stromali o tumorali tendono ad avere i punteggi più alti e più bassi rispettivamente, mentre le aree più eterogenee e miste tra tumore e stroma tendono ad avere punteggi più intermedi e più vicini a valore zero. Questa *signature* è tra quelle nel cluster 1 secondo i risultati del modello, e che infatti riconosce in questo cluster una forte dipendenza spaziale proprio all'interno delle cellule stromali e tumorali, inferiore invece nella aree di fibroblasti e ghiandole. Da questi risultati si ipotizza una eterogeneità dell'attività delle cellule tumorali dettata dall'interazione più o meno forte con il suo microambiente.

Il secondo risultato delle analisi di arricchimento riguarda i geni del cluster C_5, al contrario di quelli del cluster C_3, sono maggiormente espressi dalle cellule tumorali e meno nel TME (nel tumore $\hat{\mu}_{34} = -0.08$; nei fibroblasti, nelle ghiandole e in stroma rispettivamente $\hat{\mu}_{31} = -0.52,\ \hat{\mu}_{32} = -0.48, \hat{\mu}_{33} = -0.48$). I risultati delle analisi di arricchimento suggeriscono un coinvolgimento di questi geni nei processi di metabolismo e processamento delle proteine. Questi geni risultano arricchiti nei *pathway* di KEGG che coinvolgono il processamento delle proteine, *pathway* metabolici e la termogenesi (p-value < 0.01). Nelle analisi di arricchimento di questi geni fatta sul database di ene Ontology nelle categorie Biological Process e Molecular Function mostrano un coinvolgimento di questi geni nel trasporto intracellulare delle proteine, nella regolazione del processo di traduzione e nella respirazione aerobica (p-value < 0.05). Questi risultati rispecchiano la condizione delle cellule tumorali, che ottimizzano il loro metabolismo per venire incontro ai loro bisogni energetici. Infatti, per soddisfare le loro elevate esigenze nutrizionali ed energetiche, solitamente le cellule tumorali sfruttano le vie di traffico intracellulare per assorbire i nutrienti e le macromolecole dal microambiente tumorale

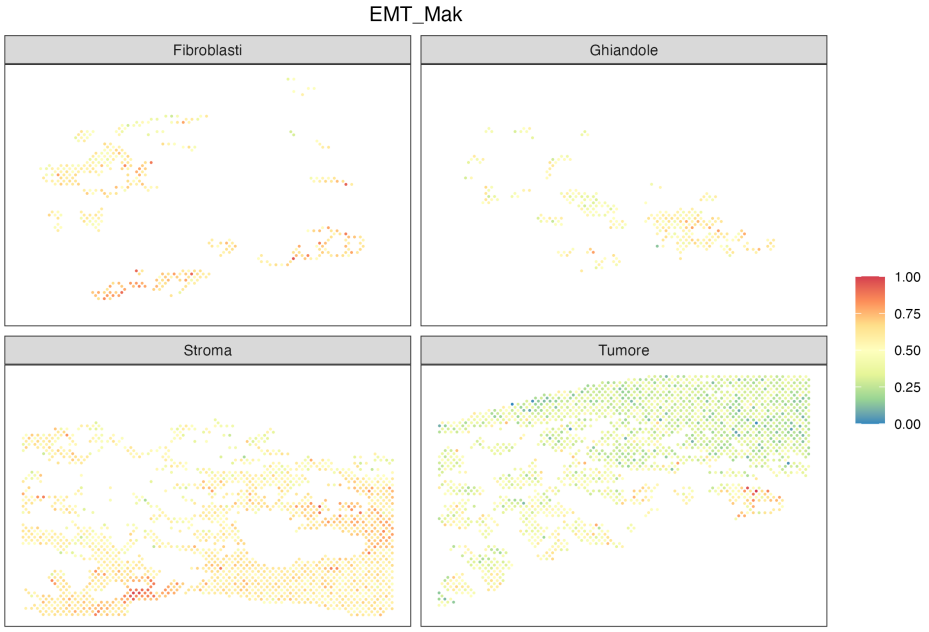


FIGURA 4.17: Distribuzione spaziale del punteggio della *signature* EMT nelle quattro aree annotate del tessuto. Si osservano che le zone più uniformemente stromali o tumorali tendono ad avere i punteggi più alti e più bassi rispettivamente, mentre le aree più eterogenee e miste tra tumore e stroma tendono ad avere punteggi più intermedi e più vicini a valore zero.

(Sneeggen et al., 2020). Questo fenomeno è inoltre supportato dalla regolazione del processo di traduzione, che viene ampiamente sfruttato dalle cellule tumorali per i propri bisogni (Robichaud et al., 2019). Per gli altri tre cluster di geni individuati dal metodo sui dati in studio non sono emerse evidenze significative. Tuttavia, sarebbe opportuno esplorare tali cluster in futuro utilizzando Ontologie più specifiche per i dati di proposta e di tumore alla prostata.

Conclusioni

Sebbene le caratteristiche del metodo di co-clustering Spartaco di Sottosanti & Risso (2022) possano rappresentare un valido strumento per l'analisi dei dati di trascrittomica, i suoi lunghi tempi di calcolo lo rendono poco pratico nelle applicazioni. La versione semplificata, proposta in questa tesi, supera questo problema riducendo drasticamente il costo computazionale. In particolare, la stima del modello originario richiede diversi giorni di elaborazione, mentre ora può essere completata in meno di un giorno.

L'introduzione della penalizzazione nella stima del modello rappresenta un importante contributo di questa tesi: gli studi di simulazione evidenziano una sostanziale riduzione della variabilità delle stime, in tutti i diversi scenari di struttura spaziale considerati, soprattutto per i parametri di media. Inoltre, si è constatato che l'introduzione della penalità non compromette la capacità del metodo nell'individuare correttamente i gruppi di geni e le aree in cui la variabilità spaziale è presente o assente, che è uno dei principali obiettivi del metodo.

Va evidenziato che richiedere la definizione a priori delle etichette per gli *spot* non costituisce una limitazione all'applicazione del metodo proposto. Al contrario, nell'ambito della trascrittomica spaziale è una pratica comune rivolgersi ad un patologo per annotare manualmente il tessuto sulla base della natura istologica. Queste informazioni costituiscono una fonte di conoscenza a priori, e grazie a questa nuova implementazione, si è in grado di integrarle nel processo di modellizzazione dei dati. Con questa modifica, il metodo Spartaco diventa un metodo di clustering semi-supervisionato dei geni: sebbene sia ancora basato sulla similarità di distribuzione, il confronto avviene solo tra aree di *spot* con stessa etichetta, anziché tra tutti gli *spot*. Si sottolinea che i vantaggi derivanti dalla formulazione del modello Spartaco rimangono invariati. In particolare, i cluster di geni vengono creati confrontando il comportamento locale dell'espressione,

il che consente di catturare le similarità di distribuzione che altrimenti potrebbero non essere colte in un confronto globale su tutto il tessuto. Inoltre, contemporaneamente al clustering dei geni, viene stimato un modello per i dati, il quale che consente di esplorare le differenze a livello di distribuzione nelle aree annotate che hanno portato alla suddivisione dei geni in cluster. Un altro grande vantaggio di questo metodo è che individua geni che sono spazialmente variabili anche solo in alcune regioni del tessuto, senza richiedere un controllo della molteplicità del test. Ciò rappresenta un'innovazione rispetto ad altri noti metodi, come SpatialDE di Svensson et al. (2018) o nnSVG di Weber et al. (2022), in cui per rispondere alla stessa domanda è necessario applicarli separatamente su ogni regione annotata. Infine, il metodo proposto permette di separare la variabilità spaziale, comune a tutti i geni all'interno di uno stesso cluster, dalla variabilità totale di ciascun gene, individuando così quelli maggiormente variabili e quindi, in un certo senso, più informativi in ogni area di *spot*.

La modifica apportata al metodo Spartaco potrebbe presentare una potenziale criticità legata alla dipendenza dall'annotazione degli *spot*. Tuttavia, gli studi di simulazione condotti hanno evidenziato una buona robustezza del metodo alle variazioni dell'annotazione: le capacità del metodo nell'individuare correttamente la configurazione in cluster dei geni non vengono compromesse.

Si è applicato il modello di clustering a dati di tessuto prostatico e si è descritto come interpretare i risultati ottenuti per fornire utili informazioni alla ricerca biologica. Inoltre, in questa applicazione è stata proposta una metodologia per rendere più completa la procedura di selezione del modello di clustering, integrando il criterio di informazione utilizzato dal metodo Spartaco con le misure di incertezza del clustering e l'uso degli alberi di clustering. Il metodo individua 5 cluster di geni con comportamenti distinti nelle 4 aree annotate, evidenziando differenze sia in termini di media che di dipendenza spaziale. In particolare, si distinguono 2 cluster per una marcata struttura spaziale che varia tra le diverse aree, mentre i restanti 3 mostrano differenze sostanziali solo a livello di media.

Si selezionano i geni altamente variabili per ogni area di *spot*. Si osserva che hanno tutti una variabilità spaziale rilevante ed alcuni sono specifici solo di alcune aree. La forza del metodo di selezione emerge chiaramente nel confronto con il metodo basato

sulla devianza di Townes et al. (2019), che non tiene conto della struttura spaziale, e nnSVG di Weber et al. (2022), in cui la modellizzazione avviene a livello di intero tessuto. Infatti, questi metodi considerano poco informativi alcuni dei geni tra i 20 più variabili in almeno una determinata area di tessuto.

I risultati ottenuti in questo caso studio sottolineano l'importanza di modellare una struttura spaziale specifica per ciascun gruppo di *spot* dell'espressione genica per individuare importanti differenze di distribuzioni tra geni e ottenere così una selezione più accurata dei geni informativi.

Infine, grazie alla collaborazione con il Dipartimento di Biologia dell'Università di Padova, è stato condotta un'analisi congiunta dei dati di espressione genica e dalle *signature* molecolari, portando a due importanti risultati. In primo luogo, è stata riscontrata un'eterogeneità nell'attività delle cellule tumorali, la quale è influenzata dall'interazione più o meno forte con il microambiente circostante. In secondo luogo, è stato individuato un coinvolgimento di un cluster di geni nei processi di metabolismo e processamento delle proteine.

Questa tesi si concentra sulla formulazione e sull'applicazione dei modelli proposti per la modellazione dei dati di trascrittomica spaziale. Tuttavia, i metodi possono essere facilmente generalizzati per qualsiasi dataset in cui le righe o le colonne rappresentano osservazioni multivariate acquisite in corrispondenza di una struttura spaziale.

In futuro, si propone di condurre ulteriori studi di simulazione considerando variazioni delle etichette degli *spot* situati ai bordi delle aree annotate, ricreando uno scenario così più realistico. Inoltre, sarebbe utile valutare se la versione penalizzata del metodo porti a miglioramenti nei casi di errata specificazione, sviluppando anche un criterio per la selezione dei parametri di penalizzazione in modo tale di scegliere i valori ottimali a seconda dei dati in studio. Infine, si ritiene interessante lo sviluppo di una versione di Spartaco intermedia tra quella proposta in questa tesi e quella originale, in cui l'annotazione manuale viene considerata come configurazione iniziale e che poi il metodo può modificare per migliorare l'adattamento del modello, ad esempio scambiando le etichette degli *spot* al confine delle aree annotate. Si prevede che questa soluzione permetta di mantenere un costo computazionale moderato e di ottenere un sostanziale miglioramento nella robustezza del metodo.

Bibliografia

Belhabib, I., Zaghdoudi, S., Lac, C., Bousquet, C. & Jean, C. (2021). Extracellular Matrices and Cancer-Associated Fibroblasts: Targets for Cancer Diagnosis and Therapy? *Cancers* **13**, 3466.

Benjamini, Y. & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)* **57**, 289–300. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.2517-6161.1995.tb02031.x.

Biernacki, C., Celeux, G. & Govaert, G. (2000). Assessing a Mixture Model for Clustering with the Integrated Completed Likelihood. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* **22**, 719–725.

Celeux, G. & Govaert, G. (1992). A classification EM algorithm for clustering and two stochastic versions. *Computational Statistics & Data Analysis* **14**, 315–332.

Chen, K. H., Boettiger, A. N., Moffitt, J. R., Wang, S. & Zhuang, X. (2015). Spatially resolved, highly multiplexed RNA profiling in single cells. *Science (New York, N.Y.)* **348**, aaa6090.

Cressie, N., Sainsbury-Dale, M. & Zammit-Mangion, A. (2022). Basis-Function Models in Spatial Statistics. ArXiv:2202.03660 [stat].

Crick, F. (1970). Central Dogma of Molecular Biology. *Nature* **227**, 561–563. Number: 5258 Publisher: Nature Publishing Group.

Dai, Y., Li, D., Chen, X., Tan, X., Gu, J., Chen, M. & Zhang, X. (2018). Circular RNA Myosin Light Chain Kinase (MYLK) Promotes Prostate Cancer Progression through Modulating Mir-29a Expression. *Medical Science Monitor : International Medical Journal of Experimental and Clinical Research* **24**, 3462. Publisher: International Scientific Information, Inc.

Dariane, C., Clairefond, S., Péant, B., Communal, L., Thian, Z., Ouellet, V., Trudel, D., Benzerdjeb, N., Azzi, F., Méjean, A., Timsit, M.-O., Baurès, M., Guidotti, J.-E., Goffin, V., Karakiewicz, P. I., Mes-Masson, A.-M. & Saad, F. (2022). High Keratin-7 Expression in Benign Peri-Tumoral Prostatic Glands Is Predictive of Bone Metastasis Onset and Prostate Cancer-Specific Mortality. *Cancers* **14**, 1623.

Efron, B. (2009). Are a set of microarrays independent of each other? *The Annals of Applied Statistics* **3**. ArXiv:0910.1426 [stat].

Gonzalez, M. W. & Kann, M. G. (2012). Chapter 4: Protein Interactions and Disease. In *PLoS Computational Biology*, vol. 8. p. e1002819.

Govaert, G. & Nadif, M. (2013). Co-Clustering: Models, Algorithms and Applications | Wiley.

Gupta, A. K. & Nagar, D. K. (1999). *Matrix Variate Distributions*. New York: Chapman and Hall/CRC.

Hastie, T., Tibshirani, R. & Friedman, J. H. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer. Google-Books-ID: eBSgoAEACAAJ.

Hwang, B., Lee, J. H. & Bang, D. (2018). Single-cell RNA sequencing technologies and bioinformatics pipelines. *Experimental & Molecular Medicine* **50**, 1–14. Number: 8 Publisher: Nature Publishing Group.

Kalluri, R. & Zeisberg, M. (2006). Fibroblasts in cancer. *Nature Reviews Cancer* **6**, 392–401. Number: 5 Publisher: Nature Publishing Group.

Kim, K., Zakharkin, S. O. & Allison, D. B. (2010). EXPECTATIONS, VALIDITY, AND REALITY IN GENE EXPRESSION PROFILING. *Journal of clinical epidemiology* **63**, 950–959.

Krušlin, B., Ulamec, M. & Tomas, D. (2015). Prostate cancer stroma: an important factor in cancer growth and progression. *Bosnian Journal of Basic Medical Sciences* **15**, 1–8.

Lubeck, E., Coskun, A. F., Zhiyentayev, T., Ahmad, M. & Cai, L. (2014). Single cell in situ RNA profiling by sequential hybridization. *Nature methods* **11**, 360–361.

Lun, A. T. L., McCarthy, D. J. & Marioni, J. C. (2016). A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. Tech. Rep. 5:2122, F1000Research. Type: article.

Mak, M. P., Tong, P., Diao, L., Cardnell, R. J., Gibbons, D. L., William, W. N., Skoulidis, F., Parra, E. R., Rodriguez-Canales, J., Wistuba, I. I., Heymach, J. V., Weinstein, J. N., Coombes, K. R., Wang, J. & Byers, L. A. (2016). A Patient-Derived, Pan-Cancer EMT Signature Identifies Global Molecular Alterations and Immune Target Enrichment Following Epithelial-to-Mesenchymal Transition. *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research* **22**, 609–620.

Marx, V. (2021). Method of the Year: spatially resolved transcriptomics. *Nature Methods* **18**, 9–14. Number: 1 Publisher: Nature Publishing Group.

Maynard, K. R., Collado-Torres, L., Weber, L. M., Uytingco, C., Barry, B. K., Williams, S. R., Catallini, J. L., Tran, M. N., Besich, Z., Tippani, M., Chew, J., Yin, Y., Kleinman, J. E., Hyde, T. M., Rao, N., Hicks, S. C., Martinowich, K. & Jaffe, A. E. (2021). Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex. *Nature Neuroscience* **24**, 425–436.

Meyer-Siegler, K. L., Iczkowski, K. A., Leng, L., Bucala, R. & Vera, P. L. (2006). Inhibition of macrophage migration inhibitory factor or its receptor (CD74) attenuates growth and invasion of DU-145 prostate cancer cells. *Journal of Immunology (Baltimore, Md.: 1950)* **177**, 8730–8739.

Moses, L. & Pachter, L. (2022). Museum of spatial transcriptomics. *Nature Methods* **19**, 534–546. Number: 5 Publisher: Nature Publishing Group.

Nelder, J. & Mead, R. (1965). Simplex Method for Function Minimization | The Computer Journal | Oxford Academic.

Nevins, J. R. & Potti, A. (2007). Mining gene expression profiles: expression signatures as cancer phenotypes. *Nature Reviews Genetics* **8**, 601–609. Number: 8 Publisher: Nature Publishing Group.

Novais, L. & Faria, S. (2021). Comparison of the EM, CEM and SEM algorithms in the estimation of finite mixtures of linear mixed models: a simulation study. *Computational Statistics* **36**, 2507–2533.

Oshlack, A. & Wakefield, M. J. (2009). Transcript length bias in RNA-seq data confounds systems biology. *Biology Direct* **4**, 14.

Pachter, L. (2011). Models for transcript quantification from RNA-Seq. ArXiv:1104.3889 [q-bio, stat].

Pirrotta, S. & Calura, E. (2023). signifinder: Implementations of transcriptional cancer signatures.

Rao, N., Clark, S. & Habern, O. (2020). Bridging Genomics and Tissue Pathology. *Genetic Engineering & Biotechnology News* **40**, 50–51. Publisher: Mary Ann Liebert, Inc., publishers.

Rasmussen, C. E. & Williams, C. K. I. (2006). *Gaussian processes for machine learning*. Adaptive computation and machine learning. Cambridge, Mass: MIT Press. OCLC: ocm61285753.

Righelli, D., Weber, L. M., Crowell, H. L., Pardo, B., Collado-Torres, L., Ghazanfar, S., Lun, A. T. L., Hicks, S. C. & Risso, D. (2022). SpatialExperiment: infrastructure for spatially-resolved transcriptomics data in R using Bioconductor. *Bioinformatics* **38**, 3128–3131.

Robichaud, N., Sonenberg, N., Ruggero, D. & Schneider, R. J. (2019). Translational Control in Cancer. *Cold Spring Harbor Perspectives in Biology* **11**, a032896.

Rodriques, S. G., Stickels, R. R., Goeva, A., Martin, C. A., Murray, E., Vanderburg, C. R., Welch, J., Chen, L. M., Chen, F. & Macosko, E. Z. (2019). Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. *Science (New York, N.Y.)* **363**, 1463–1467.

Sawada, J., Li, F. & Komatsu, M. (2015). R-Ras protein inhibits autophosphorylation of vascular endothelial growth factor receptor 2 in endothelial cells and suppresses receptor activation in tumor vasculature. *The Journal of Biological Chemistry* **290**, 8133–8145.

Sneeggen, M., Guadagno, N. A. & Progida, C. (2020). Intracellular Transport in Cancer Metabolic Reprogramming. *Frontiers in Cell and Developmental Biology* **8**, 597608.

Sottosanti, A. & Risso, D. (2022). Co-clustering of Spatially Resolved Transcriptomic Data. *The Annals of Applied Statistics* In Press.

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

Svensson, V., Teichmann, S. & Stegle, O. (2018). SpatialDE: identification of spatially variable genes | Nature Methods.

Tan, K. M. & Witten, D. M. (2014). Sparse Biclustering of Transposable Data. *Journal of computational and graphical statistics : a joint publication of American Statistical Association, Institute of Mathematical Statistics, Interface Foundation of North America* **23**, 985–1008.

Tibshirani, R. (1996). Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)* **58**, 267–288. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.2517-6161.1996.tb02080.x.

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 Biology* **20**, 295.

Wang, Z., Gerstein, M. & Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews. Genetics* **10**, 57–63.

Weber, L. M., Saha, A., Datta, A., Hansen, K. D. & Hicks, S. C. (2022). nnSVG: scalable identification of spatially variable genes using nearest-neighbor Gaussian processes. *bioRxiv* , 2022.05.16.492124.

Witten, D. M. & Tibshirani, R. (2010). A framework for feature selection in clustering. *Journal of the American Statistical Association* **105**, 713–726.

Zappia, L. & Oshlack, A. (2018). Clustering trees: a visualization for evaluating clusterings at multiple resolutions. *GigaScience* **7**, giy083.

Zhang, Y., Zhang, J., Liang, S., Lang, G., Liu, G., Liu, P. & Deng, X. (2019). Long non-coding RNA VIM-AS1 promotes prostate cancer growth and invasion by regulating epithelial-mesenchymal transition. *Journal of B.U.ON.: official journal of the Balkan Union of Oncology* **24**, 2090–2098.