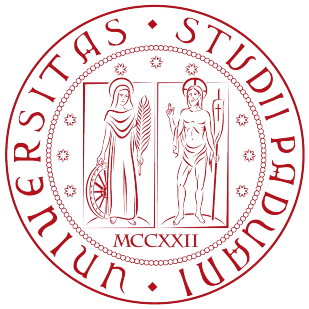
Università degli studi di Padova

Dipartimento di Scienze Statistiche

Corso di Laurea Magistrale in
Scienze Statistiche



Analisi di Procuste
per dati di trascrittomica spaziale

Relatore: Prof. Livio Finos
Dipartimento di Psicologia dello Sviluppo e della Socializzazione

Correlatori: Prof. Davide Risso
Dipartimento di Scienze Statistiche

Dott.ssa Angela Andreella
Dipartimento di Economia, Università Ca' Foscari Venezia

Laureanda: Daniela Corbetta
Matricola N. 1237312

Anno Accademico 2021/2022

Indice

Introduzione

L'analisi di dati di espressione genica negli ultimi anni si è rilevata fondamentale nella ricerca medica: ha infatti permesso di sviluppare farmaci mirati ed efficaci, di diagnosticare precocemente alcune patologie e di conseguenza di intervenire tempestivamente ed agevolarne il decorso. Le tecniche fino ad oggi più utilizzare si basano sul sequenziamento del codice genetico e permettono di ottenere i dati di espressione genica di una determinata cellula o di un campione. Non preservano però l'informazione relativa alla localizzazione spaziale della cellula analizzata, e questo diventa problematico in alcune applicazioni. Uno degli esempi in tal senso più noti è rappresentato dagli studi sul cervello umano: in questo caso, in virtù della struttura laminare ben definita propria del cervello, conoscere la posizione esatta di una sua cellula diventa decisivo, considerando ad esempio che per alcuni disturbi neuropsichiatrici sono state trovate alterazioni nell'espressione genica in punti specifici della corteccia. Per far fronte a questa difficoltà è nata la trascrittomica spaziale, che comprende un insieme di tecniche che permettono di ottenere i dati di espressione genica mantenendo l'informazione relativa alla localizzazione spaziale del campione analizzato. Uno degli approcci più promettenti per la trascrittomica spaziale è stato proposto da Ståhl et al. (2016) e permette di utilizzare i sequenziatori di seconda generazione già impiegati per l'analisi del sequenziamento dell'RNA.

Un problema ben noto nel contesto delle neuroscienze è che i cervelli di diversi individui non sono allineati dal punto di vista funzionale. In altre parole significa che, anche supponendo che i cervelli siano allineati dal punto

di vista anatomico, la stessa zona svolge, in individui diversi, una funzione diversa. Se dunque si vuole sfruttare l'informazione relativa alla localizzazione spaziale per confrontare campioni di soggetti diversi in diverse condizione biologiche o patologiche (ad esempio sani contro malati), si rischia di utilizzare coordinate sbagliate. Si può quindi pensare di far precedere all'analisi differenziale un passo preliminare in cui si allineano le immagini ottenute dal punto di vista funzionale. Prendendo spunto dalle neuroscienze, dove viene già regolarmente impiegata, l'analisi di Procuste sembra essere uno strumento promettente anche per l'allineamento di dati di trascrittomica spaziale.

Procuste nella mitologia greca è il soprannome di un brigante di nome Damaste, che sulla via da Megara ad Atene assaltava i viandanti e li stendeva su un letto facendoli combaciare perfettamente alla forma di esso, stirandoli a forza se troppo bassi e amputandoli se troppo alti. Così come il brigante greco menomava le sue vittime per far sì che il loro corpo si adattasse il meglio possibile ad un letto, l'analisi di Procuste in algebra ha l'obiettivo di rendere il più simile tra loro diverse matrici tramite similitudini (traslazioni, rotazioni, riflessioni e *scaling*). A termine dell'allineamento Procuste quindi ogni punto delle matrici allineate diventa una combinazione lineare degli altri punti.

Obiettivo di questa tesi è applicare metodi di allineamento Procuste a dati di trascrittomica spaziale di campioni di tessuto della corteccia cerebrale. Oltre a rendere più simili le matrici, ci si aspetta che l'allineamento Procuste produca dei benefici dal punto di vista dell'analisi differenziale. Si può infatti pensare che la variabilità dei dati di trascrittomica spaziale sia scomponibile in due parti: una componente di variabilità biologica e una componente di disturbo dovuta al mancato allineamento. Ci si aspetta che l'allineamento assorba la variabilità di disturbo e di conseguenza che elimini falsi positivi dovuti all'effetto di soggetto.

La struttura della tesi è come segue: nel Capitolo 1 si introdurrà il contesto biologico, quindi si spiegherà come funziona il sequenziamento del DNA, l'importanza degli studi di espressione genica sul cervello, il funzionamento

della tecnologia della trascrittomica spaziale e si introdurranno i dati; nel Capitolo 2 si discuteranno nel dettaglio i metodi proposti per risolvere il problema di Procuste. Si partirà dalla soluzione esplicita per l'allineamento di due sole matrici, si descriveranno le tecniche proposte per l'allineamento di più matrici evidenziandone limiti e criticità e infine si presenteranno due modelli recentemente sviluppati, il modello ProMises e il modello ProMises Efficiente, che risolvono i problemi che le altre tecniche presentano. Per quanto riguarda il modello ProMises Efficiente, se ne presenteranno due versioni e si mostrerà come sia facilmente applicabile anche al caso in cui le matrici da allineare abbiano diverso numero di colonne. Nel Capitolo 3, vera novità di questo elaborato, si passerà all'applicazione del problema di Procuste a dati ottenuti tramite trascrittomica spaziale del cervello umano. L'obiettivo del Capitolo 3 è duplice: si vuole da un lato mostrare che l'allineamento rende le immagini più omogenee e di conseguenza rende confrontabili immagini derivanti da soggetti diversi, dall'altro si vuole dare una motivazione biologica.

Parte integrante del lavoro di tesi è stato lo sviluppo di una libreria `R` che implementasse il modello ProMises nella sua versione base e nella versione Efficiente. Il codice è consultabile nella cartella Github `https://github.com/angeella/alignProMises`. La libreria contiene quattro funzioni: `GPASub` che implementa il problema di Procuste per l'allineamento a una matrice di riferimento nota, `ProMisesModel` che implementa l'algoritmo del modello ProMises per l'allineamento in uno spazio comune ignoto, `EfficientProMises` che implementa la versione Efficiente del modello ProMises con decomposizione della matrice media e `EfficientProMisesSubj` che implementa la versione Efficiente con decomposizione delle matrici X_i.

Le analisi sono state eseguite con il software `R`, versione 4.1.0.

Capitolo 1

Contesto biologico

Lo studio dell'espressione genica è essenziale per comprendere a fondo i processi biologici che regolano l'organismo. Un gene è una sequenza di DNA che codifica una proteina, elemento costitutivo della vita e responsabile di tutte le attività di una cellula. Nelle cellule eucariote i geni sono organizzati in maniera discontinua: la sequenza codificante è spezzettata in piccoli pezzi chiamati esoni inframmezzati da pezzi non codificanti che prendono il nome di introni. Il meccanismo con cui avviene il passaggio da un gene a una proteina è stato racchiuso da Crick (1970) nel dogma centrale della biologia molecolare, per il quale il DNA viene trascritto in RNA e l'RNA viene tradotto in proteine. Ci sono quindi due fasi fondamentali: la trascrizione e la traduzione. Nella prima, che avviene nel nucleo, il filamento di DNA viene trascritto in un filamento di RNA complementare (RNA messaggero, mRNA) grazie all'enzima RNA-polimerasi e ad altre proteine chiamate fattori di trascrizione. Nella seconda, che avviene invece nel citoplasma, le triplette nucleotidiche del filamento di mRNA vengono tradotte negli amminoacidi corrispondenti, formando così la proteina finale. Tra le due fasi vi è un passaggio intermedio, lo *splicing*, nel quale vengono eliminati gli introni; la traduzione opera quindi solamente sulla parte codificante del trascritto.

Nonostante tutte le cellule di un organismo condividano lo stesso patrimonio genetico, i geni espressi, ossia i geni che vengono trascritti, sono diversi

in base al tipo di cellula o ad alcune condizioni biologiche o patologiche. Lo studio dell'espressione genica è quindi di fondamentale importanza anche nella medicina perché permette di individuare tali condizioni patologiche e mettere in atto cure mirate ed efficaci.

Per misurare l'espressione genica si può ricorrere alla tecnologia dell'RNA *sequencing* (RNA-*Seq*), basata sul sequenziamento dell'RNA. Questa tecnica permette di ottenere delle *reads*, ossia delle brevi sequenze di 50-150 basi azotate corrispondenti a un tratto del filamento di RNA originario. I geni sono mediamente lunghi 10.000 basi e le tecnologie moderne non sono in grado di sequenziarli interamente. Per capire quindi a quale gene appartiene una determinata *read* bisogna allinearla al genoma. Una volta allineate le *reads* sul genoma, il livello di espressione del gene sarà dato dal numero di *reads* che mappano su tale gene. A livello concettuale, più *reads* mappano su un gene, più proteine quel gene produce.

Ipotizzando di analizzare J geni in n campioni, i dati di espressione genica possono quindi essere rappresentati da una matrice di conteggi di dimensione $J \times n$ in cui in riga si hanno i geni e in colonna le unità statistiche, che possono essere campioni, singole cellule, singoli nuclei o piccole porzioni di tessuto. Sia y_{ij} il conteggio relativo all'espressione del gene j nel campione i. Un esempio di matrice dei dati è riportato in Tabella 1.1.

	Unità 1	Unità 2	...	Unità n
Gene 1	y_{11}	y_{21}	...	y_{n1}
Gene 2	y_{12}	y_{22}	...	y_{n2}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
Gene J	y_{1J}	y_{2J}	...	y_{nJ}

Tabella 1.1: Matrice dei dati di un esperimento di RNA *sequencing* su n campioni e J geni

Spesso, oltre alla matrice dei conteggi, si hanno a disposizione informazio-

ni riguardo i geni e i campioni che prendono il nome di meta-dati. I meta-dati per i campioni possono per esempio includere il tipo cellulare, la condizione biologica, il lotto o la data in cui il campione è stato sequenziato.

Per avere un'idea dell'ordine di grandezza di n bisogna fare una distinzione tra le due principali tipologie di dati di RNA-*Seq*: i dati *bulk* RNA-*Seq* e i dati *single cell* RNA-*Seq* (*sc*RNA-*Seq*). La differenza principale sta nel fatto che con i dati di *bulk* RNA-*Seq* si analizza l'espressione media di un gene in un gruppo di cellule, ad esempio di un tessuto, di un organismo o di una linea cellulare, mentre con i dati *sc*RNA-*Seq* si misura l'espressione genica di una sola cellula. Con i dati *sc*RNA-*Seq* si hanno quindi generalmente centinaia o migliaia di cellule, mentre con i dati *bulk* il numero di unità statistiche dipende dal numero di campioni che si hanno a disposizione ed è di norma di molto inferiore. Va sottolineato inoltre che i dati *bulk* non sono adatti quando si vogliono studiare tessuti particolarmente complessi ed eterogenei come ad esempio il cervello.

Un metodo alternativo al *sc*RNA-*Seq* è il *single nucleus* RNA-*Seq* (*sn*RNA-*Seq*), che differisce dal primo in quanto sequenzia solo il nucleo di una cellula e non la cellula intera. Se nel *sc*RNA-*Seq* vengono sequenziati sia i filamenti di mRNA che si trovano nel nucleo sia quelli che si trovano nel citoplasma, con il *sn*RNA-*Seq* vengono sequenziati quindi solamente i trascritti nucleari. Uno dei vantaggi del sequenziamento *single nucleus* è che i nuclei necessari possono essere facilmente ottenuti anche da tessuti congelati o conservati, mentre per isolare le singole cellule necessarie per il *sc*RNA-*Seq* sarebbero richiesti un complicato processo e una lunga incubazione che possono stressare le cellule, risultando in un'espressione genica falsata. Il sequenziamento *single nucleus* è inoltre preferibile al sequenziamento a singola cellula quando si tratta di analizzare cellule difficili da isolare, come i neuroni o gli adipociti (Bakken et al., 2018). Al contrario, uno degli svantaggi di questa tecnologia è che per lo studio di determinate patologie l'informazione racchiusa nei trascritti citoplasmatici, che viene inevitabilmente tralasciata, risulta importante.

1.1 Studi di espressione genica nel cervello umano

Studi genetici sul cervello umano sono fondamentali per identificare geni collegati ad un maggior rischio di sviluppare alcune patologie cerebrali, come schizofrenia e disturbi nello spettro dell'autismo, e permettere di agire precocemente per ritardarne l'insorgenza e gestirne al meglio il decorso.

Prima di entrare nello specifico dello studio dell'espressione genica nel cervello umano, è necessaria una breve digressione anatomica. Il cervello controlla tutte le attività umane grazie all'elettricità in esso presente, che viene trasmessa dal cervello agli organi interessati sotto forma di impulso nervoso. È composto da due tipi di cellule: i neuroni, responsabili della generazione dell'elettricità e della trasmissione degli impulsi, e le cellule gliali, che hanno funzione di supporto strutturale dei neuroni. I neuroni sono composti dal soma, il corpo centrale che contiene il nucleo che prende il nome di neutrone, e da prolungamenti citoplasmatici. Gli impulsi nervosi si generano nel soma e si propagano grazie ai prolungamenti, che si differenziano in dendriti e assone. I dendriti sono ramificazioni corte che ricevono il segnale da altri neuroni mentre l'assone è un ramo molto più lungo, responsabile, tramite il proprio terminale, della trasmissione del segnale elettrico. Il contatto tra diversi neuroni è chimico e prende il nome di sinapsi: la parte pre-sinaptica è formata dal terminale del neurone che trasmette l'impulso mentre la parte post-sinaptica è formata dal dendrite del neurone che lo riceve. La generazione degli impulsi è possibile grazie al fatto che un neurone a riposo contiene energia, chiamata potenziale di membrana. Variazioni del potenziale di membrana producono tali impulsi nervosi.

Gli assoni sono ottimi conduttori dei segnali elettrici grazie al fatto che sono ricoperti da una membrana isolante, la mielina, che velocizza la propagazione del segnale. La mielina nel sistema nervoso centrale è generata da un particolare tipo di cellule gliali, gli oligodendrociti, e conferisce agli assoni il caratteristico colore bianco. Gli assoni rivestiti di mielina costituiscono

uno dei due principali tessuti da cui è formato il cervello, la materia bianca, mentre i soma dei neuroni costituiscono la materia grigia. Lo spazio privo di corpi cellulari, denso quindi di dendriti, terminali assonici e connessioni sinaptiche, prende il nome di neuropilo.

Al giorno d'oggi, a causa della difficoltà insita nell'isolare i neuroni e del fatto che gli studi sul cervello umano vengono prevalentemente eseguiti post-mortem su tessuti congelati, la maggior parte dei dati disponibili sul cervello umano è stata ottenuta tramite sequenziamento *single nucleus*. Non vengono quindi sequenziati compartimento citoplasmatico, assoni e dendriti e si perde così l'informazione riguardo l'espressione genica dei trascritti citoplasmatici e neuropili. Ciò è un problema per lo studio di alcuni disturbi psichiatrici come la schizofrenia in quanto recenti studi (Skene et al., 2018) hanno evidenziato come geni associati a queste malattie siano arricchiti da trascritti neuropili.

Un'altra limitazione delle tecnologie *sc*RNA-*Seq* e *sn*RNA-*Seq* se utilizzate in questo contesto è che non forniscono l'esatta posizione anatomica della cellula o del nucleo analizzato, per quanto essa possa essere prevista sulla base dell'espressione di alcuni geni, detti marcatori, la cui espressione è specifica di un determinato tipo cellulare. Il cervello umano ha un'organizzazione spaziale strettamente collegata alla propria funzione. In particolare, la corteccia cerebrale ha una struttura laminare suddivisa in sei strati, numerati da I, il più esterno, a VI, il più vicino allo strato di materia bianca, che differiscono sia per la morfologia che per la funzione delle cellule neuronali di cui sono composti. Per alcuni disturbi neuropsichiatrici sono emerse differenze in strutture neuronali e sinaptiche in specifici strati corticali (Velmeshev et al., 2019; Sweet, Fish e Lewis, 2010). Avere a disposizione l'informazione relativa alla localizzazione spaziale, e poter di conseguenza mappare l'espressione genica delle varie regioni cerebrali, porterebbe quindi a identificare e diagnosticare precocemente tali patologie.

A questo scopo è recentemente nata la trascrittomica spaziale, tecnologia che, oltre all'espressione genica, permette di risalire alla localizzazione spaziale della cellula sequenziata.

1.2 Trascrittomica spaziale

La trascrittomica spaziale comprende un insieme di tecniche che permettono di sequenziare il genoma delle cellule preservando l'informazione relativa alla loro localizzazione spaziale. I primi approcci proposti per la trascrittomica spaziale si basano o sull'isolamento di microsezioni di tessuto e sul conseguente sequenziamento (tecniche basate sulla microdissezione), o sulla visualizzazione *in situ* delle molecole di RNA tramite ibridazione o sequenziamento (Fluorescent *in situ* hybridization, *in situ* sequencing) (Asp, Bergenstråhle e Lundeberg, 2020). Queste tecniche hanno però alcune limitazioni: le tecniche basate sulla microdissezione richiedono infatti di sezionare tutto il tessuto da analizzare, rendendo di fatto impossibile sequenziare l'intero campione, mentre le tecniche basate sull'ibridazione richiedono di osservare la fluorescenza al microscopio e questo comporta problemi legati alla visualizzazione.

Un'alternativa che evita tali problemi è stata proposta da Ståhl et al. (2016) e consiste nel catturare i trascritti *in situ* per poi sequenziarli *ex situ* tramite sequenziatori di seconda generazione già usati per il *sc*RNA-*seq*. Il metodo di Ståhl et al. sfrutta oligonucleotidi *barcoded*, contenenti cioè una sequenza nucleotidica nota che funge da tag spaziale, che hanno la capacità di legarsi ai filamenti di mRNA da sequenziare. Sequenziando quindi anche tali oligonucleotidi è possibile risalire alle coordinate della cellula analizzata.

1.2.1 Piattaforma Visium

Attualmente, una delle piattaforme più usate per la trascrittomica spaziale tramite l'approccio di Ståhl et al. (2016) è la *Visium Spatial Gene Expression* messa a punto dall'azienda 10x Genomics.

La tecnologia Visium utilizza dei vetrini per la preparazione delle librerie, rappresentati in Figura 1.1, sui quali sono presenti o due o quattro aree di cattura di dimensione $6.5mm \times 6.5mm$. Ogni area di cattura è formata da circa 5.000 punti di cattura con tag spaziali, ognuno dei quali contenente milioni di oligonucleotidi di cattura *barcoded* che si legano ai filamenti di

mRNA delle cellule da sequenziare. Si sottolinea come con questa tecnologia non si sequenzi il genoma di una singola cellula ma quello di tutte le cellule presenti nel punto di cattura, che ne comprende generalmente da una a dieci.

I passi dell'analisi di trascrittomica spaziale tramite la tecnologia Visium sono riassunti in Figura 1.2. Innanzitutto, si richiede di preparare il campione posizionando il tessuto da analizzare su una delle aree di cattura del vetrino. Si passa poi alla costruzione dell'immagine utilizzando tecniche standard di colorazione e fissaggio per visualizzare sezioni del tessuto sul vetrino o per identificare proteine presenti nel tessuto. Vi è poi il passo di costruzione della libreria: mentre si trova ancora sul vetrino, il tessuto viene permeabilizzato in maniera che rilasci l'mRNA che si lega agli oligonucleotidi spazialmente taggati presenti sull'area di cattura. I tag spaziali vengono aggiunti all'RNA da sequenziare durante il passo di retrotrascrizione, ossia il processo per cui da una sequenza di RNA si risale ad una sequenza di DNA che prende il nome di DNA complementare (cDNA). Si ottengono così molecole di cDNA taggate, cioè contenenti sia il genoma da sequenziare che il nucleotide che identifica la localizzazione spaziale, che costituiscono la libreria da sequenziare. La libreria così ottenuta è compatibile con alcuni sequenziatori di seconda generazione, come per esempio Illumina. Al termine del sequenziamento si risale quindi alla posizione spaziale del codice genetico sequenziato grazie agli oligonucleotidi *barcoded*.

Supponendo di avere una piattaforma che permette di sequenziare n punti e J geni in k campioni, il dato finale che si ottiene con la piattaforma Visium comprende k matrici di conteggi di dimensione $J \times n$ con i geni in riga e i punti in colonna (Tabella 1.1) e k matrici di dimensione $n \times 2$ che riportano le coordinate degli n punti analizzati. Si possono quindi visualizzare i dati di espressione genica in una mappa che ricostruisce la sezione di tessuto (Figura 1.3) e osservare come un gene abbia diversi livelli di espressione in diverse zone del campione.

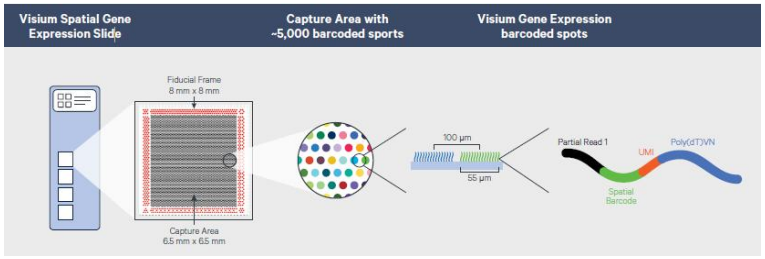


Figura 1.1: Composizione di un vetrino della piattaforma *Visium Spatial Gene Expression*, immagine presa da `https://www.10xgenomics.com/spatial-transcriptomics`.

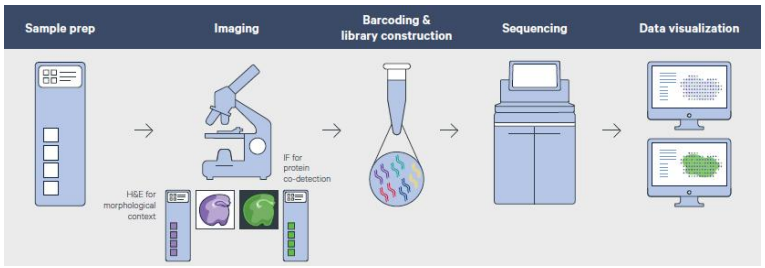


Figura 1.2: Schema riassuntivo del sequenziamento tramite trascrittomica spaziale del *Visium Spatial Gene Expression*, immagine presa da `https://www.10xgenomics.com/spatial-transcriptomics`.

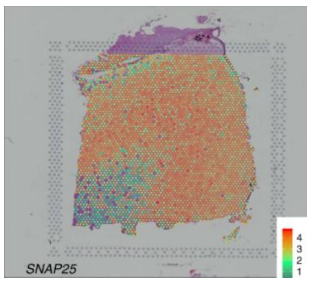


Figura 1.3: Esempio di visualizzazione dell'espressione del gene SNAP25 in un campione di tessuto della corteccia frontale dorso laterale (Maynard et al., 2021). Si può notare come il gene abbia diversi livelli di espressione nei diversi strati della corteccia.

1.3 I dati

I dati analizzati in questa tesi sono stati raccolti da Maynard et al. (2021) utilizzando la piattaforma per la trascrittomica spaziale 10x Genomics Visium descritta in precedenza. Derivano da sezioni di tessuto della corteccia dorso-laterale pre-frontale di tre soggetti adulti sani, due uomini e una donna, di età compresa tra i 30 e i 46 anni. Per ogni soggetto sono stati raccolti due paia di repliche spaziali, posizionate a 300 μm le une dalle altre. Si hanno quindi a disposizione quattro immagini per soggetto, per un totale di 12 immagini.

Per ogni immagine sono stati sequenziati circa 4000 punti, chiamati *spot* (47.681 *spot* in totale). Mediamente, ogni *spot* contiene 3.3 cellule, con il 15% di *spot* che contengono una sola cellula e il 9.7% di *spot* neurofili, che non contengono quindi alcun corpo cellulare. Ogni *spot* è stato manualmente assegnato ad uno dei sei strati corticali o alla materia bianca tramite un approccio supervisionato che considera sia la citoarchitettura delle cellule in esso presenti sia i geni marcatori. Le immagini con le annotazioni degli strati sono riportate in Figura 1.4.

Per dimostrare l'efficacia dei metodi Procuste sull'analisi differenziale, andrebbero allineate matrici provenienti dal sequenziamento di campioni di tessuto in diverse condizioni biologiche per poi adattare modelli che considerino come covariata la condizione biologica e confrontarne i risultati con quelli derivanti da modelli adattati sui dati non allineati. Tuttavia, il *dataset* analizzato in questo elaborato è uno dei primi *dataset* resi pubblici prodotti con la piattaforma *Visium* e contiene solamente dati derivanti da individui sani. Per poter valutare l'effetto dell'allineamento nell'analisi si considererà pertanto l'individuo come condizione biologica e si confronteranno quindi le quattro immagini di un individuo contro le quattro immagini di un altro, aspettandosi che l'allineamento assorba una quota di variabilità imputabile al mancato allineamento delle immagini di partenza.

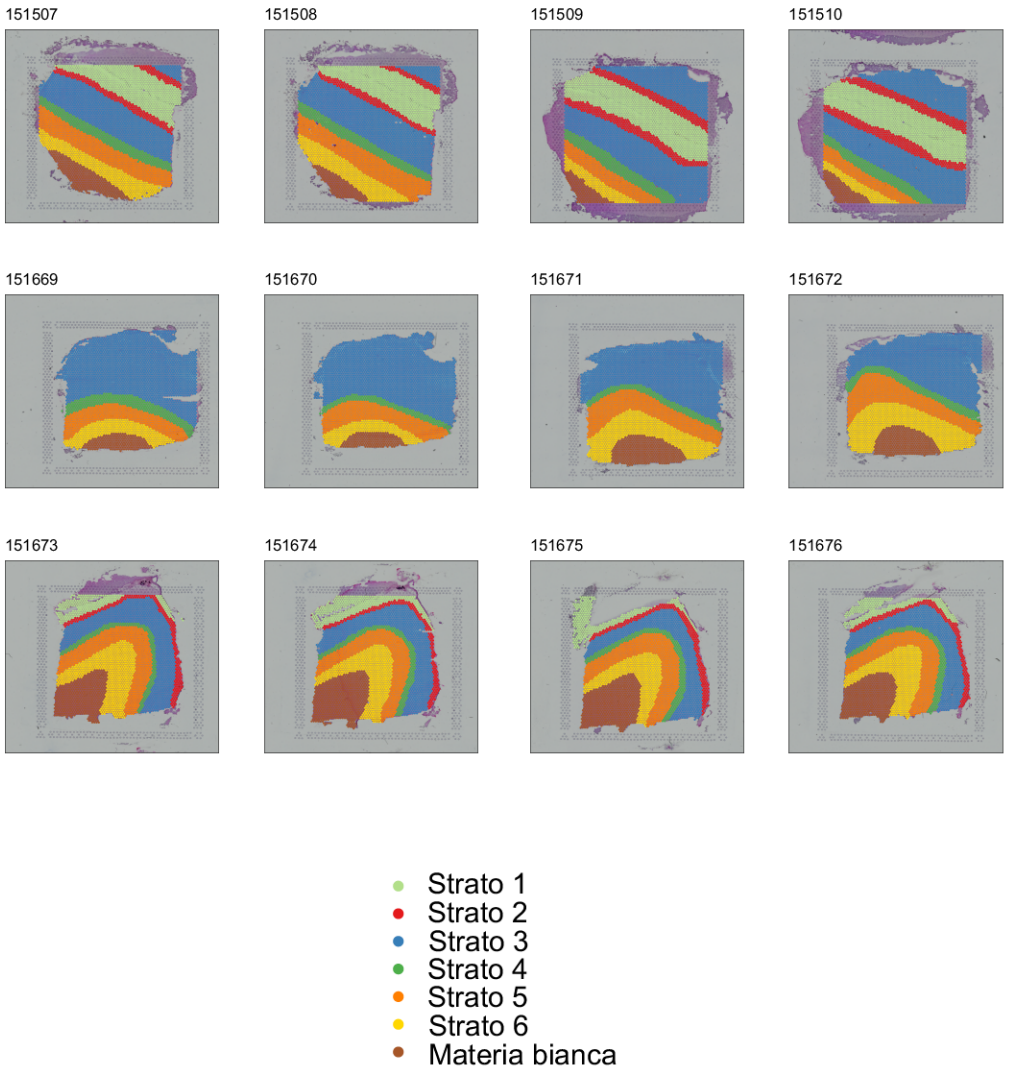


Figura 1.4: Immagini dei campioni di corteccia pre-frontale dorso-laterale analizzati con le annotazioni degli strati. Le immagini di ogni soggetto sono riportate in riga. Le prime due immagini di ogni riga rappresentano la prima coppia di repliche, le seconde due la seconda coppia.

Capitolo 2

Analisi di Procuste

2.1 Soluzione per due matrici

Come accennato nell'Introduzione, il metodo di Procuste ha lo scopo di rendere il più simili tra loro diverse matrici, allineandole alle coordinate di uno spazio di riferimento comune tramite similitudini. Nel caso in cui si vogliano allineare due matrici, X_1 e X_2, questo si traduce nel cercare una matrice T per cui post-moltiplicare X_1 in maniera da renderla il più simile possibile a X_2. La matrice X_2 assume il ruolo di matrice di riferimento mentre la matrice X_1 viene allineata ad essa. Esprimendo il concetto di similitudine tramite il criterio dei minimi quadrati, la forma più generale del problema di Procuste è quindi

$$T = \underset{T^*}{\operatorname{argmin}} \|X_1 T^* - X_2\|_F^2 \tag{2.1}$$

dove X_1 e X_2 sono due matrici di dimensione rispettivamente $n \times m_1$ e $n \times m_2$, T è una matrice di dimensione $m_1 \times m_2$ e $\|\cdot\|_F^2$ indica la norma di Frobenius, che è pari alla traccia della matrice delle differenze al quadrato ($\|A_1 - A_2\|_F^2 = \mathrm{Tr}\big((A_1 - A_2)(A_1 - A_2)^\top\big)$).

In letteratura sono state proposte diverse soluzioni al problema Procuste che dipendono dalla dimensione delle due matrici considerate e dal tipo di trasformazione richiesta. La matrice T può infatti essere una matrice orto-

gonale, una matrice ortonormale o una trasformazione obliqua se le matrici sono di uguale dimensione oppure una matrice semi-ortogonale se le matrici hanno dimensioni diverse. Il problema può inoltre essere formulato in una versione simmetrica in cui le due matrici assumono lo stesso ruolo, ossia $\|X_1T_1 - X_2T_2\|_F^2$. In questa tesi ci si focalizzerà inizialmente sul caso in cui T è una matrice ortogonale di rotazione o riflessione, che verrà indicata con R, e sulla versione asimmetrica del problema Procuste definito dalla formula (2.1). In questo caso quindi, le matrici X_1 e X_2 sono di uguale dimensione $n \times m$ e la matrice $R \in \mathcal{O}(m)$, dove $\mathcal{O}(m)$ è il gruppo delle matrici ortogonali di dimensione m. Si discuterà in seguito, nel paragrafo 2.2.4, un metodo che permette di applicare l'allineamento Procuste a matrici con diverso numero di colonne.

La soluzione nel caso di due matrici è esplicita ed è stata proposta inizialmente da Green (1952) che ha considerato il caso in cui le matrici sono a rango pieno e poi estesa da Schönemann (1966) al caso in cui le matrici non siano a rango pieno. Si riportano di seguito la soluzione e la dimostrazione presenti in Gower, Dijksterhuis et al. (2004). Innanzitutto, minimizzare $\|X_1R - X_2\|_F^2$ è equivalente a massimizzare $\mathrm{Tr}(X_2^\top X_1 R)$. Sia $X_2^\top X_1 = UDV^\top$ la decomposizione a valori singolari di $X_2^\top X_1$. Si ha quindi che $\mathrm{Tr}(X_2^\top X_1 R) = \mathrm{Tr}(UDV^\top R) = \mathrm{Tr}(DV^\top RU) = \mathrm{Tr}(DH)$, dove $H = V^\top RU$ è una matrice ortogonale in quanto prodotto di matrici ortogonali. Dato che $\mathrm{Tr}(DH) = \sum_{i=1}^m h_{ii}d_i$ e sapendo che gli elementi d_i sono non negativi, si ha che il massimo si ha quando gli elementi h_{ii} sono pari a 1, e quindi per $H = V^\top RU = I$. Segue che la soluzione per R è data da

$$\hat{R} = \underset{R}{\mathrm{argmin}} \|X_1R - X_2\|_F^2 = VU^\top \tag{2.2}$$

dove U è la matrice di dimensione $m \times m$ contenente i vettori singolari sinistri di $X_2^\top X_1$ mentre V è la matrice di dimensione $m \times m$ contenente i vettori singolari destri di $X_2^\top X_1$. Condizione necessaria e sufficiente affinché la soluzione trovata sia un minimo globale per $\|X_1R - X_2\|_F^2$ è che $X_2^\top X_1 R$ sia simmetrica e semi-definita positiva (Berge, 1977).

Se richiesto, si può anche applicare un fattore di *scaling* isotropico per minimizzare

$$\|\alpha^{-1}X_1R - X_2\|_F^2. \tag{2.3}$$

In questo caso, la stima di R rimane invariata mentre la stima di α viene sempre ottenuta tramite il criterio dei minimi quadrati ed è pari a

$$\alpha = \frac{\|X_1\|_F^2}{\mathrm{Tr}(X_2^\top X_1 R)} = \frac{\|X_1\|_F^2}{\mathrm{Tr}(D)}. \tag{2.4}$$

2.2 Soluzione per N matrici

2.2.1 Generalized Procustes Analysis

Si consideri ora il caso in cui si vogliono allineare N matrici, $X_1, \ldots, X_N$, tutte di uguale dimensione $n \times m$. In questo caso, il criterio dei minimi quadrati viene riformulato in maniera da trovare N matrici $R_1, \ldots, R_N \in \mathcal{O}(m)$ che minimizzino (Gower, Dijksterhuis et al., 2004)

$$\sum_{i<j}^{N} \|X_iR_i - X_jR_j\|_F^2. \tag{2.5}$$

L'espressione (2.5) può essere riscritta come somma di deviazioni dalla media delle matrici ruotate: $\sum_{i<j}^{N} \|X_iR_i - X_jR_j\|_F^2 = N\sum_{i=1}^{N} \|X_iR_i - M\|_F^2$, con $M = \frac{1}{N}\sum_{i=1}^{N} X_iR_i$. Questa riformulazione suggerisce come il problema Procuste nel caso con più di due matrici possa essere visto come la somma di singoli problemi Procuste di allineamento a una matrice di riferimento comune.

Non si ha più una soluzione esplicita; un algoritmo iterativo per calcolare gli elementi delle matrici di rotazione è stato proposto da Gower (1975) e prende il nome di *Generalised Procustes Analysis* (GPA). L'algoritmo consiste nel calcolare, ad ogni iterazione, la media delle matrici ruotate, $\hat{M}$, e, per $i = 1, \ldots, N$, risolvere il problema di allineamento Procuste $\|X_iR - \hat{M}\|_F^2$. Ci si ferma quando la distanza tra le matrici medie di due iterazioni successive è inferiore ad un valore soglia scelto a priori.

Analogamente al caso con due matrici, si può richiedere di stimare per ogni matrice X_i un fattore di *scaling* isotropico α_i e minimizzare

$$\sum_{i<j}^{N} \|\alpha_i^{-1} X_i R_i - \alpha_j^{-1} X_j R_j\|_F^2 = N \sum_{i=1}^{N} \|\alpha_i^{-1} X_i R_i - M\|_F^2 \tag{2.6}$$

Anche in questo caso, tramite il criterio dei minimi quadrati, si risale a

$$\hat{\alpha}_{i_{\hat{R}_i}} = \frac{\|X_i\|_F^2}{\mathrm{Tr}(M^\top X_i \hat{R}_i)} = \frac{\|X_i\|_F^2}{\mathrm{Tr}(D_i)}. \tag{2.7}$$

α_i viene inizializzato ad 1 e ad ogni iterazione il suo valore viene aggiornato tramite la formula (2.7). L'algoritmo per la *Generalized Procustes Analysis* proposto da Gower è quindi il seguente:

Algoritmo 1 *Generalized Procustes Analysis* (Gower, 1975). `T` è un valore soglia per la distanza tra matrici medie di due iterazioni successive scelto a priori e `maxIt` è il numero massimo di iterazioni.

Input: $\boldsymbol{X}_i$, `T` , `maxIt`, $\forall i = 1, \ldots, N$
Output: $\hat{\boldsymbol{X}}_i$ $\forall i = 1, \ldots, N$

$\hat{\boldsymbol{M}} = \sum_{i=1}^{N} \boldsymbol{X}_i / N$, $\hat{\alpha}_i = 1$
`count` $= 0$, `dist` $=$ `Inf`
while `dist` $>$ `T` OR `count` $<$ `maxIt` **do**
  for $i = 1$ to N **do**
    $\boldsymbol{U}_i \boldsymbol{D}_i \boldsymbol{V}_i^\top = \mathtt{SVD}(\boldsymbol{X}_i^\top \hat{\boldsymbol{M}})$ ▷ Decomposizione a valori singolari
    $\hat{\boldsymbol{R}}_i = \boldsymbol{U}_i \boldsymbol{V}_i^\top$
    $\hat{\boldsymbol{X}}_i = \boldsymbol{X}_i \hat{\boldsymbol{R}}_i$
    $\hat{\alpha}_{i_{\hat{\boldsymbol{R}}_i}} = \|\boldsymbol{X}_i^\top\|_F^2 / \mathrm{Tr}(\boldsymbol{D}_i)$
    $\hat{\boldsymbol{X}}_i = \hat{\alpha}_{i_{\hat{\boldsymbol{R}}_i}}^{-1} \hat{\boldsymbol{X}}_i$ ▷ Aggiorna $\boldsymbol{X}_i$
  end for
  $\hat{\boldsymbol{M}}_{\text{old}} = \hat{\boldsymbol{M}}$ ▷ Salva $\hat{\boldsymbol{M}}$
  $\hat{\boldsymbol{M}} = \sum_{i=1}^{N} \hat{\boldsymbol{X}}_i / N$ ▷ Aggiorna $\hat{\boldsymbol{M}}$
end while

Berge (1977) ha evidenziato una limitazione dell'algoritmo proposto da Gower: se la matrice M è nulla, allora l'allineamento Procuste non ha alcun effetto per nessuna X_i. Per superare questo problema ha suggerito una

leggera modifica all'algoritmo, sostituendo alla media globale nel passo 2 la media calcolata escludendo la matrice i, $M_{(i)}$. In questo modo, anche nel caso in cui la media globale sia nulla, ci sarà almeno una matrice $M_{(i)}$ non nulla. L'algoritmo avrà quindi effetto almeno su una X_i.

Un aspetto critico della GPA è che non fornisce una soluzione unica, infatti, per ogni matrice $Q \in \mathcal{O}(m)$, si ha che, per $i = 1, \dots, N$, $R_i Q$ sono ancora soluzioni valide (Andreella, 2021):

$$\begin{aligned}
\min_{R_i} \sum_{i=1}^{N} \|X_i R_i Q - MQ\|_F^2 &= \min_{R_i} \sum_{i=1}^{N} \mathrm{Tr}\big((X_i R_i Q - MQ)^\top (X_i R_i Q - MQ)\big) \\
&= \min_{R_i} \sum_{i=1}^{N} \mathrm{Tr}\big(Q^\top (X_i R_i - M)^\top (X_i R_i - M) Q\big) \\
&= \min_{R_i} \sum_{i=1}^{N} \mathrm{Tr}\big((X_i R_i - M)^\top (X_i R_i - M)\big) \\
&= \min_{R_i} \sum_{i=1}^{N} \|X_i R_i - M\|_F^2
\end{aligned} \tag{2.8}$$

Nel caso in cui il numero di colonne sia $m \leq 3$, Q modifica solamente l'orientamento di R_i. mentre, nei casi ad alta dimensionalità, Q porta ad infinite possibili interpretazioni. Questo è un problema grave in applicazioni in cui le matrici hanno un'interpretazione spaziale, come nelle neuroscienze o nella trascrittomica spaziale, perché Q modifica l'orientamento spaziale delle immagini allineate. Un modello che tiene in considerazione questo aspetto e che fornisce una soluzione unica è discusso nel paragrafo 2.2.3.

Goodall (1991) ha riformulato il problema di Procuste sotto forma di modello statistico proponendo il *perturbation model*. Per risolvere il problema della non unicità della soluzione, sono state proposte estensioni da un punto di vista bayesiano del *perturbation model* (Green e Mardia, 2006; Andreella e Finos, 2022).

2.2.2 Perturbation Model

Nel *perturbation model* le matrici $X_1, \dots, X_N \in \mathbb{R}^{n \times m}$ da allineare vengono espresse come perturbazioni casuali di una matrice di riferimento comune M. In particolare, vengono definite come similitudini (rotazioni, riflessioni, traslazioni e *scaling*) della matrice M alla quale viene aggiunto un termine di errore casuale matriciale E_i. Il modello descritto è quindi

$$X_i = \alpha_i (M + E_i) R_i^\top + 1_n t_i^\top, \tag{2.9}$$

dove $E_i \sim \mathcal{MN}_{n,m}(0, \Sigma_n, \Sigma_m)$ è il termine d'errore con distribuzione normale matriciale, $\alpha_i \in \mathbb{R}^+$ è il termine di *scaling* isotropico, $t_i \in \mathbb{R}^{m \times 1}$ è il vettore di traslazione e $R_i \in \mathcal{O}(m)$. Nel seguito si assumerà che le matrici siano centrate, e di conseguenza si tralascerà il vettore di traslazione.

Per semplificare i calcoli, si può assumere che gli errori E_i siano gaussiani al primo ordine e che quindi $\Sigma_n \otimes \Sigma_m = \sigma^2 I_n \otimes I_m$. L'assunzione $\Sigma_m = I_m$ è giustificata se le colonne vengono normalizzate. Assumere $\Sigma_n = \sigma^2 I_n$ non implica che l'immagine complessiva abbia errori indipendenti perché il termine di errore e la media vengono poi ruotati, introducendo una dipendenza.

La matrice M rappresenta le coordinate dello spazio comune a cui le matrici vengono allineate ed è quindi pari alla media aritmetica elemento per elemento delle matrici allineate: $M = \frac{1}{N} \sum_{i=1}^{N} \alpha_i^{-1} X_i R_i$.

La matrice X_i ha distribuzione

$$\mathrm{vec}(X_i | R_i, \alpha_i) \sim \mathcal{N}_{nm}(\alpha_i M R_i^\top, \alpha_i^2 \sigma^2 I_{nm}). \tag{2.10}$$

Sotto questo modello, per una matrice X_i, è possibile scrivere la log-verosimiglianza per α_i e R_i che è pari a

$$\ell(\alpha_i, R_i) \propto -\frac{1}{2\sigma^2 \alpha_i^2} \mathrm{Tr}\big((X_i - \alpha_i M R_i^\top)^\top (X_i - \alpha_i M R_i^\top)\big). \tag{2.11}$$

Assumendo indipendenza tra le matrici $X_1, \dots, X_N$ da allineare, la log-verosimiglianza congiunta è data dalla somma delle log-verosimiglianze definite in equazione (2.11):

$$\ell(\alpha_i, R_i) \propto \sum_{i=1}^{N} \left(-\frac{1}{2\sigma^2 \alpha_i^2} \mathrm{Tr}\big((X_i - \alpha_i M R_i^\top)^\top (X_i - \alpha_i M R_i^\top)\big) \right). \tag{2.12}$$

Lo stimatore di massima verosimiglianza per R_i si ottiene risolvendo il problema di massimizzazione

$$\hat{R}_i = \underset{R_i}{\operatorname{argmax}} \left\{ -\sum_{i=1}^{N} \frac{1}{2\alpha_i^2} \|X_i^\top - \alpha_i R_i M^\top\|_F^2 \right\} \tag{2.13}$$

che, nel caso in cui M è nota, ha una soluzione esplicita che deriva dalla soluzione del singolo problema di massimizzazione

$$\begin{aligned}
\hat{R}_i &= \underset{R_i}{\operatorname{argmax}} \left\{ -\|X_i^\top - \alpha_i R_i M^\top\|_F^2 \right\} \\
&= \underset{R_i}{\operatorname{argmax}} \left\{ -\operatorname{Tr}\big((X_i^\top - \alpha_i R_i M^\top)(X_i^\top - \alpha_i R_i M^\top)^\top\big) \right\} \\
&= \underset{R_i}{\operatorname{argmax}} \left\{ \operatorname{Tr}(X_i^\top M R_i^\top) \right\} = U_i V_i^\top,
\end{aligned}$$

con $X_i^\top M = U_i D_i V_i^\top$.

Per quanto riguarda lo stimatore di massima verosimiglianza del parametro di *scaling*, si consideri la log-verosimiglianza profilo per α_i,

$$\begin{aligned}
\ell_p(\alpha_i) &= -\frac{1}{2\alpha_i^2} \|X_i^\top - \alpha_i \hat{R}_i M^\top\|^2 \\
&= -\frac{1}{2\alpha_i^2} \|X_i^\top\|^2 + \frac{1}{\alpha_i} < X_i^\top M, \hat{R}_i >
\end{aligned}$$

e se ne calcoli la derivata prima:

$$\ell_p'(\alpha_i) = \frac{1}{\alpha_i} \|X_i^\top\|^2 - < X_i^\top M, \hat{R}_i > .$$

Ponendo $\ell_p'(\alpha_i) = 0$, si ottiene

$$\hat{\alpha}_{i_{\hat{R}_i}} = \frac{\|X_i\|_F^2}{\operatorname{Tr}(D_i)}.$$

Gli stimatori di massima verosimiglianza per R_i e α_i sono quindi equivalenti agli stimatori ai minimi quadrati precedentemente ottenuti. Per quanto, considerando M nota, gli stimatori siano disponibili in forma chiusa, la soluzione fornita non è unica. Il problema Procuste può infatti essere riformulato come

$$\max \operatorname{Tr}(A_i^\top R_i) = \max \operatorname{Tr}(X_i^\top M R_i) \tag{2.14}$$

che è stato dimostrato avere soluzione unica solo nel caso in cui la matrice A_i sia a rango pieno (Trendafilov e Lippert, 2002; Myronenko e Song, 2009). Avendo quindi $n < m$, $X_i^\top M$ ha rango n e di conseguenza la soluzione non è unica.

Nel caso invece in cui M non sia nota, la massimizzazione in equazione (2.13) non ha una soluzione analitica esplicita. Si ricorre alla GPA descritta al paragrafo precedente, ottenendo nuovamente, come dimostrato con la serie di equivalenze in (2.8), una soluzione non unica. Andreella e Finos (2022) hanno proposto il modello von Mises-Fisher-Procrustes (modello ProMises) che rivisita il *perturbation model* in chiave bayesiana e dal quale si deriva un algoritmo per il calcolo delle matrici di rotazione che fornisce sempre una soluzione unica.

2.2.3 Modello ProMises

Il punto cardine del modello ProMises è la specificazione di una distribuzione a priori per il parametro di rotazione R_i che rifletta la struttura di orientamento spaziale dei dati di partenza. Dato che R_i è un parametro ortogonale, si richiede che la distribuzione a priori prenda valori in una varietà di Manifold ($V_m(\mathbb{R}^m)$).

2.2.3.1 Distribuzione von Mises-Fisher matriciale

La distribuzione proposta da Andreella e Finos (2022) è la distribuzione von Mises-Fisher matriciale (Downs, 1972), definita come

$$f(R_i) = C(F, k) \exp\left\{\mathrm{Tr}(kF^\top R_i)\right\}, \tag{2.15}$$

dove $C(F, k)$ è una costante di normalizzazione, $F \in \mathbb{R}^{m \times m}$ è il parametro di posizione matriciale e $k \in \mathbb{R}^+$ è il parametro di concentrazione.

La distribuzione scelta ha diverse proprietà interessanti sia dal punto di vista interpretativo che computazionale. Permette infatti di mantenere l'informazione relativa alla struttura anatomica dei dati da allineare tramite un'opportuna specificazione del parametro F, è un membro della fami-

glia esponenziale (Barndorff Nielsen, 1973) ed è una priori coniugata per la distribuzione normale matriciale (Green e Mardia, 2006).

Per capire perché il parametro F sia responsabile dell'orientamento spaziale delle matrici allineate, se ne considerino la decomposizione polare e la decomposizione a valori singolari:

$$F = PK = L\Sigma B^\top = LB^\top B\Sigma B^\top \tag{2.16}$$

con P, L, B matrici ortogonali di dimensione $m \times m$, K matrice semidefinita positiva di dimensione $m \times m$ e $\Sigma \in \mathbb{R}^{m\times m}$ matrice diagonale con numeri reali non negativi sulla diagonale. Jupp e Mardia (1979) hanno dimostrato che la moda della distribuzione von Mises-Fisher si ha in corrispondenza di P che, come si può notare dalle equivalenze (2.16), è pari a $LB^\top$, quindi al prodotto tra i vettori singolari destri e sinistri di F. Inoltre, se si definisce F a rango pieno, allora anche Σ è a rango pieno. Segue che la decomposizione polare, e di conseguenza la moda della distribuzione, sono uniche e pertanto P è un massimo globale. Definire F a rango pieno permette quindi di ottenere una soluzione unica per $\hat{R}_i$ che riflette l'orientamento spaziale dato da F. Se si hanno a disposizione le coordinate dei punti che si vogliono allineare, una valida proposta per F è

$$F = \exp\big\{-D\big\}, \tag{2.17}$$

dove $D \in \mathbb{R}^{m\times m}$ è la matrice delle distanze euclidee tra i punti. In questo modo, nella rotazione, verrà dato più peso ai punti che erano vicini nello spazio originario.

Si dimostra ora che la distribuzione von Mises-Fisher (2.15) è una priori coniugata per la distribuzione normale matriciale, e che il parametro di

posizione della distribuzione a posteriori è pari a $F^* = X_i^\top M + k\sigma^2 F$:

$$\begin{aligned}\prod_{i=1}^{N} f(R_i|X_i,k,F) &\propto f(X_i|R_i)f(R_i)\\ &= \prod_{i=1}^{N}\Big\{ -\frac{1}{2\sigma^2}\mathrm{Tr}\big((X_i - MR_i^\top)(X_i - MR_i^\top)^\top\big)\Big\}\\ &\quad\cdot \exp\big\{k\mathrm{Tr}(F^\top R_i)\big\}\\ &= \exp\big\{-\sum_{i=1}^{N}\frac{1}{2}f(X_i)\big\}\exp\big\{\sum_{i=1}^{N}\mathrm{Tr}((X_i^\top M + k\sigma^2 F)^\top R_i)\big\}\end{aligned}$$

Sia $k^* = k\sigma^2$. L'ultima espressione è il nucleo di una distribuzione von Mises-Fisher con parametro di posizione $F^* = X_i^\top M + k^* F$. Il parametro a posteriori è quindi la somma di due quantità, $X_i^\top M$ e $k^* F$. Si considerino la decomposizione a valori singolari di $X_i^\top M$ e la decomposizione polare di F. Le due quantità possono essere scritte come $X_i^\top M = U_i D_i V_i^\top = U_i V_i^\top V_i D_i V_i^\top$ e $F = PK$. Focalizzandosi sulla prima equivalenza, si può notare come $X_i^\top M$ sia espresso dal prodotto tra lo stimatore di massima verosimiglianza di R_i, $U_i V_i^\top$, e $V_i D_i V_i^\top$ che, in quanto parte ellittica della decomposizione polare di $X_i^\top M$, ne rappresenta una misura di variabilità. Nella decomposizione polare di F invece, si ha che P è la moda della distribuzione a priori di R_i e K ne è una misura di variabilità. Riscrivendo F^* come

$$F^* \propto U_i V_i^\top V_i D_i V_i^\top + PK$$

si può notare quindi come il parametro di posizione della distribuzione a posteriori sia dato dalla somma della stima di massima verosimiglianza di R_i e dalla moda della distribuzione a priori, entrambe moltiplicate per una misura della propria variabilità.

Il parametro $k \in \mathbb{R}^+$ definisce invece la concentrazione attorno ad F. Se $k = 0$ la distribuzione a priori è una uniforme e si torna di conseguenza alla GPA, mentre se $k \to +\infty$ la distribuzione a priori tende ad una Dirac.

2.2.3.2 Stima della matrice di rotazione R_i

Dato che la distribuzione $f(X_i|R_i, \alpha_i)$ dipende solamente dal prodotto $\alpha_i R_i$, in seguito si considererà la distribuzione $f(X_i|R_i\alpha_i)$. Di conseguenza, si considererà come distribuzione a priori per $\alpha_i R_i$ la distribuzione von Mises-Fisher

$$f(\alpha_i R_i) \sim \frac{1}{\alpha_i} \exp\big\{\frac{k}{\alpha_i}\mathrm{Tr}(F^\top R_i)\big\}. \tag{2.18}$$

Il modello ProMises definisce quindi le matrici $X_1, \ldots, X_N$ tramite il *perturbation model* descritto in equazione (2.9), assumendo $\Sigma_n \otimes \Sigma_m = \sigma^2 I_{nm}$, e assegna al parametro $\alpha_i R_i$ la distribuzione a priori (2.18).

Si derivano di seguito gli stimatori per i parametri R_i e α_i tramite la stima del massimo della probabilità a posteriori (MAP).

Si consideri la distribuzione a posteriori congiunta di α_i e R_i per una matrice X_i

$$\begin{aligned} f(\alpha_i R_i|X_i, k, F) \propto & f(X_i|\alpha_i, R_i) f(\alpha_i R_i) \\ = & \exp\big\{ -\frac{1}{2\sigma^2\alpha_i^2}\mathrm{Tr}\big((X_i - \alpha_i M R_i^\top)^\top (X_i - \alpha_i M R_i^\top)\big)\big\} \\ & \cdot \exp\big\{\frac{k}{\alpha_i}\mathrm{Tr}(F^\top R_i)\big\}\alpha_i^{-1} \end{aligned} \tag{2.19}$$

da cui si ricava la log-posteriori

$$\begin{aligned} \log f(\alpha_i R_i|X_i, k, F) \propto & \frac{1}{2\sigma^2\alpha_i^2}\big\{ -\mathrm{Tr}\big((X_i - \alpha_i M R_i^\top)^\top (X_i - \alpha_i M R_i^\top)\big) \\ & + 2\alpha_i k\sigma^2 \mathrm{Tr}(F^\top R_i)\big\} - \log(\alpha_i). \end{aligned}$$

Sia $k^* = k\sigma^2$. Per α_i, M e k^* fissati, la stima MAP per R_i si ricava massi-

mizzando la log-posteriori:

$$\begin{aligned}\hat{R}_i &= \underset{R_i}{\text{argmax}} \Big\{ -\text{Tr}\big((X_i - \alpha_i M R_i^\top)^\top (X_i - \alpha_i M R_i^\top)\big) + 2\alpha_i k^* \text{Tr}(F^\top R_i) \Big\} \\ &= \underset{R_i}{\text{argmax}} \Big\{ -\text{Tr}(X_i^\top X_i) - \alpha_i^2 \text{Tr}(M R_i^\top R_i M) + 2\alpha_i \text{Tr}(M R_i^\top X_i^\top) \\ &\qquad + 2\alpha_i k^* \text{Tr}(F^\top R_i) \Big\} \\ &= \underset{R_i}{\text{argmax}} \Big\{ \text{Tr}(R_i^\top X_i^\top M) + \text{Tr}(k^* F^\top R_i) \Big\} \\ &= \underset{R_i}{\text{argmax}} \Big\{ \text{Tr}(R_i^\top (X_i^\top M + k^* F)) \Big\}\end{aligned}$$

Sia $X_i^\top M + k^* F = U_i D_i V_i^\top$ la SVD di $X_i^\top M + k^* F$. Allora

$$\underset{R_i}{\text{argmax}} \Big\{ \text{Tr}(R_i^\top (X_i^\top M + k^* F)) \Big\} = \underset{R_i}{\text{argmax}} \Big\{ \text{Tr}(R_i^\top U_i D_i V_i^\top) \Big\} = U_i V_i^\top$$

per calcoli analoghi a quelli fatti per la soluzione esplicita per il caso con due matrici riportati al paragrafo 2.1.

Fissato $\hat{R}_i$, il massimo a posteriori per α_i, $\hat{\alpha}_{i_{\hat{R}_i}}$, è dato da

$$\begin{aligned}\hat{\alpha}_{i_{\hat{R}_i}} &= \text{argmax} \Big\{ \frac{1}{2\alpha_i^2} \big(-\|X_i^\top - \alpha_i \hat{R}_i M^\top\|_F^2 + 2k^* \alpha_i \text{Tr}(F^\top \hat{R}_i) \big) - \log(\alpha_i) \Big\} \\ &= \text{argmax} \Big\{ -\frac{1}{2\alpha_i^2} \|X_i^\top\|_F^2 - \frac{1}{2\alpha_i^2} \|\alpha_i \hat{R}_i M\|^2 + \frac{1}{\alpha_i^2} < X_i^\top M, \alpha_i \hat{R}_i > \\ &\quad + \frac{k^*}{\alpha_i} < F, \hat{R}_i > - \log(\alpha_i) \Big\} \\ &= \text{argmax} \Big\{ -\frac{1}{2\alpha_i^2} \|X_i^\top\|_F^2 + \frac{1}{\alpha_i} < X_i^\top M + k^* F, \hat{R}_i > - \log(\alpha_i) \Big\}\end{aligned}$$

da cui si ricava la derivata prima

$$\begin{aligned}\alpha_i^{-2} \|X_i\|_F^2 - \alpha_i^{-1} < X_i^\top M + k^* F, \hat{R}_i > -1 = 0 \\ \alpha_i^{-2} \|X_i\|_F^2 - \alpha_i^{-1} \text{Tr}(D_i) - 1 = 0.\end{aligned}$$

Applicando il teorema di Viéte si ha che, sotto la condizione $\text{Tr}(D_i) \gg 1/\text{Tr}(D_i)$,

$$\hat{\alpha}_{i_{\hat{R}_i}} \approx \frac{\|X_i^\top\|_F^2}{\text{Tr}(D_i)}. \tag{2.20}$$

Considerando ora le N matrici $X_1, \ldots, X_N \in \mathbb{R}^{n\times m}$, assumendo indipendenza si ha che la distribuzione a posteriori congiunta è data dal prodotto delle distribuzioni a posteriori definite in equazione (2.19). La log-posteriori è quindi proporzionale a

$$\log f(\alpha_i, R_i|X_i, k, F) \propto \sum_{i=1}^{N} \mathrm{Tr}\bigg(-\frac{1}{2\alpha_i^2\sigma}(X_i - \alpha_i M R_i^\top)(X_i - \alpha_i M R_i^\top)\bigg) + k\sum_{i=1}^{N}\frac{1}{\alpha_i}\mathrm{Tr}(F^\top R_i) + C_{\alpha_i},$$

dove C_{α_i} è una funzione dei parametri α_i. Segue quindi che gli stimatori per R_i si ottengono risolvendo

$$\hat{R}_i = \underset{R_i}{\mathrm{argmax}} \left\{ \sum_{i=1}^{N} \frac{1}{2\alpha_i^2}\|X_i^\top - \alpha_i R_i M^\top\|_F^2 + k^* \sum_{i=1}^{N}\frac{1}{\alpha_i}\mathrm{Tr}(F^\top R_i)\right\}. \quad (2.21)$$

Nel caso in cui la matrice M sia nota, la massimizzazione in equazione (2.21) può essere vista come la somma dei singoli problemi di massimizzazione $\hat{R}_i = \mathrm{argmax}_{R_i}\left\{\frac{1}{2\alpha_i^2}\|X_i^\top - \alpha_i R_i M^\top\|_F^2 + k^*\frac{1}{\alpha_i}\mathrm{Tr}(F^\top R_i)\right\}$. In questo caso quindi, si ottengono soluzioni esplicite per $\hat{R}_i$ e $\hat{\alpha}_i$, pari rispettivamente a $\hat{R}_i = U_i V_i^\top$ e $\hat{\alpha}_i = \|X_i^\top\|_F^2/\mathrm{Tr}(D_i)$, dove le matrici U_i, V_i e D_i derivano dalla decomposizione a valori singolari di $X_i^\top M + k^* F = U_i D_i V_i^\top$.

Nel caso invece in cui M non sia nota, la soluzione non è più esplicita e si ricorre ad un algoritmo iterativo i cui passi sono i seguenti:

Algoritmo 2 Modello ProMises (Andreella e Finos, 2022). `T` è un valore soglia per la distanza tra matrici medie di due iterazioni successive scelto a priori e `maxIt` è il numero massimo di iterazioni.

Input: $\boldsymbol{X}_i$, `T` , `maxIt`, $\forall i = 1, \ldots, N$

Output: $\hat{\boldsymbol{X}}_i \ \forall i = 1, \ldots, N$

$\hat{\boldsymbol{M}} = \sum_{i=1}^{N} \boldsymbol{X}_i / N$, $\hat{\alpha}_i = 1$

`count` $= 0$, `dist` $=$ `Inf`

while `dist` $>$ `T` OR `count` $<$ `maxIt` **do**

  for $i = 1$ to N **do**

    $\boldsymbol{U}_i \boldsymbol{D}_i \boldsymbol{V}_i^\top = \texttt{SVD}(\boldsymbol{X}_i^\top \hat{\boldsymbol{M}} + k^* \boldsymbol{F})$ ▷ Decomposizione a valori singolari

    $\hat{\boldsymbol{R}}_i = \boldsymbol{U}_i \boldsymbol{V}_i^\top$

    $\hat{\boldsymbol{X}}_i = \boldsymbol{X}_i \hat{\boldsymbol{R}}_i$

    $\hat{\alpha}_{i\hat{\boldsymbol{R}}_i} = \|\boldsymbol{X}_i^\top\|_F^2 / \mathrm{Tr}(\boldsymbol{D}_i)$

    $\hat{\boldsymbol{X}}_i = \hat{\alpha}_{i\hat{\boldsymbol{R}}_i}^{-1} \hat{\boldsymbol{X}}_i$ ▷ Aggiorna $\boldsymbol{X}_i$

  end for

  $\hat{\boldsymbol{M}}_{\text{old}} = \hat{\boldsymbol{M}}$ ▷ Salva $\hat{\boldsymbol{M}}$

  $\hat{\boldsymbol{M}} = \sum_{i=1}^{N} \hat{\boldsymbol{X}}_i / N$ ▷ Aggiorna $\hat{\boldsymbol{M}}$

end while

L'algoritmo derivante dal modello ProMises risulta quindi in una lieve modifica della GPA per cui, invece di calcolare la SVD di $X_i^\top M$, si decompone $X_i^\top M + k^* F$.

A differenza della GPA, nel caso in cui il parametro di posizione F è a rango pieno, si ottiene una soluzione unica. Si riscriva il problema Procuste definito dal modello ProMises come $\max \sum_{i=1}^{N} \mathrm{Tr}\big((X_i^\top M + k^* F) R_i\big)$. F a rango pieno implica che $F^* = X_i^\top M + k^* F$ sia a rango pieno e di conseguenza che la decomposizione polare di F^* sia unica. Dato che la funzione obiettivo è convessa, e ricordando che la moda della distribuzione von Mises-Fisher è data dalla matrice P della decomposizione polare del parametro F, si ha che $\hat{R}_i$ è un massimo globale. Inoltre, considerando il modello ProMises, non vale la proprietà di permutazione ciclica della traccia che è stata sfruttata nella dimostrazione (2.8), quindi, definita Q come una matrice ortogonale

di dimensione m, si ha che moltiplicare la matrice di rotazione R per Q non porta allo stesso problema di massimizzazione definito in (2.21), cioè $\mathrm{Tr}(F^\top R_i) \neq \mathrm{Tr}(F^\top R_i Q)$.

Il modello ProMises fornisce quindi una soluzione unica del problema di Procuste, rendendo possibile mantenere l'interpretazione spaziale delle matrici allineate.

2.2.4 Modello ProMises Efficiente

Una limitazione del modello ProMises discusso nel paragrafo precedente è che richiede la decomposizione a valori singolari di matrici quadrate di dimensione pari al numero di colonne delle matrici da allineare, che ha una complessità computazionale pari a $O(m^3)$. Nel caso dell'applicazione a matrici con un alto numero di colonne, come può essere il caso di applicazioni a dati di *sc*RNA-*Seq* o a dati di trascrittomica spaziale, la complessità computazionale è quindi molto elevata e di conseguenza è necessaria molta memoria e i tempi di esecuzione diventano insostenibili. Una soluzione a questo problema è stata proposta da Andreella e Finos (2022), che hanno esteso il modello ProMises da loro formulato aggiungendo un passo preliminare in cui, tramite una trasformazione semi-ortogonale, viene ricavata una rappresentazione a più bassa dimensionalità dei dati senza perdita di informazione. In questo contesto con matrice semi-ortogonale si intende una matrice non quadrata le cui colonne sono ortonormali. Il modello risultante prende il nome di modello ProMises Efficiente.

La soluzione proposta, vantaggiosa nel caso in cui $m \gg n$, sfrutta la *thin*-SVD (Bai et al., 2000) delle matrici X_i. Con *thin*-SVD si intende la decomposizione $X_i = L_i S_i Q_i^\top$, dove $L_i \in \mathbb{R}^{n \times n}$ è una matrice ortogonale di dimensione $n \times n$ contenente gli n autovalori sinistri, $S_i \in \mathbb{R}^{n \times n}$ è una matrice diagonale che contiene gli n valori singolari non nulli e $Q_i \in \mathbb{R}^{m \times n}$ è una matrice semi-ortogonale che contiene i primi n vettori singolari destri. Per ricavare la rappresentazione a più bassa dimensionalità dei dati da allineare, le matrici $X_1, \ldots, X_N$ vengono moltiplicate per le matrici $Q_1, \ldots, Q_N$. Si

applica poi l'algoritmo del modello ProMises alle matrici quadrate di dimensione n $X_i^* = X_i Q_i$ e si ottengono le matrici ruotate $\hat{X}_1^*, \ldots, \hat{X}_N^* \in \mathbb{R}^{n \times n}$. Per riportarle nello spazio originale, viene applicata alle matrici $\hat{X}_i^*$ la trasformazione inversa $Q_i^\top$. In questo modo il problema Procuste richiede solamente la decomposizione a valori singolari di matrici quadrate di dimensione pari al numero di righe delle matrici da allineare, riducendo quindi nettamente la complessità computazionale nel caso in cui $n \ll m$.

La formulazione Efficiente non causa una perdita di informazione perché il problema Procuste, nel caso in cui le matrici da scomporre siano di rango n, analizza le prime $n \times n$ dimensioni di R_i. Infatti, se si moltiplicano le matrici X_i per Q_i, il massimo definito in equazione 2.14 rimane invariato (Andreella e Finos, 2022):

$$\max_{R_i \in \mathcal{O}(m)} \operatorname{Tr}(R_i^\top X_i^\top X_j) = \max_{R_i^* \in \mathcal{O}(n)} \operatorname{Tr}(R_i^{*\top} Q_i^\top X_i^\top X_j Q_j). \tag{2.22}$$

Per dimostrarlo, si scriva $\operatorname{Tr}(R_i^\top X_i^\top X_j)$ come $\operatorname{Tr}(X_i R_i X_j^\top)$ e si consideri la decomposizione a valori singolari di $X_i = L_i S_i C_i^\top$, con $S_i \in \mathbb{R}^{n \times m}$. La matrice S_i può essere definita come una matrice a blocchi,

$$S_i = \begin{bmatrix} S_i^* & 0 \end{bmatrix},$$

dove $S_i^* \in \mathbb{R}^{n \times n}$ e 0 è una matrice di zeri di dimensione $n \times (m - n)$. Segue che

$$\begin{aligned} \operatorname{Tr}(X_i R_i X_j^\top) &= \operatorname{Tr}(L_i S_i C_i^\top R_i C_j S_j^\top L_j^\top) \\ &= \operatorname{Tr}(L_i S_i R_i^o S_j^\top C_j^\top), \end{aligned}$$

con $R_i^o = C_i^\top R_i C_j \in \mathcal{O}(m)$ perché prodotto di matrici ortogonali. Si scriva R_i^o come una matrice a blocchi,

$$R_i^o = \begin{bmatrix} R_{11i}^o & R_{12i}^o \\ R_{21i}^o & R_{22i}^o \end{bmatrix}$$

con $R^o_{11i} \in \mathbb{R}^{n\times n}$, $R^o_{12i} \in \mathbb{R}^{n\times m-n}$, $R^o_{21i} \in \mathbb{R}^{m-n\times n}$, $R^o_{22i} \in \mathbb{R}^{m-n\times m-n}$. Si ha quindi

$$\begin{aligned} L_i S_i R_i^o S_j^\top L_j^\top &= L_i \begin{bmatrix} S_i^* & 0 \end{bmatrix} \begin{bmatrix} R^o_{11i} & R^o_{12i} \\ R^o_{21i} & R^o_{22i} \end{bmatrix} \begin{bmatrix} S_j^* \\ 0 \end{bmatrix} L_j^\top \\ &= \begin{bmatrix} L_i S_i^* & 0 \end{bmatrix} \begin{bmatrix} R^o_{11i} & R^o_{12i} \\ R^o_{21i} & R^o_{22i} \end{bmatrix} \begin{bmatrix} S_j^{*\top} L_j^\top \\ 0 \end{bmatrix}. \\ &= L_i S_i^* R^o_{11i} S_j^* L_j^\top \end{aligned}$$

Sostituendo quest'ultima espressione nella massimizzazione iniziale, si ha

$$\max_{R_i} \operatorname{Tr}(X_i R_i X_j^\top) = \max_{R^o_{11i}} \operatorname{Tr}(L_i S_i^* R^o_{11i} S_j^{*\top} L_j^\top).$$

Sia $X_i = L_i S_i^* Q_i$ la *thin*-SVD di X_i. Si ha allora

$$\max_{R_i} \operatorname{Tr}(X_i R_i X_j^\top) = \max_{R_i^*} \operatorname{Tr}(X_i Q R_i^* Q_j^\top X_j^\top),$$

con $R_i^* = R^o_{11i} \in \mathbb{R}^{n\times n}$.

L'equivalenza in 2.22 vale anche se si considera la massimizzazione del modello ProMises,

$$\max_{R_i \in \mathcal{O}(m)} \operatorname{Tr}\big(R_i^\top (X_i^\top X_j + kF)\big) = \max_{R_i^* \in \mathcal{O}(n)} \operatorname{Tr}\big(R_i^{*\top} (Q_i^\top X_i^\top X_j Q_j + kF^*)\big),$$

con $F \in \mathbb{R}^{m\times m}$ e $F^* \in \mathbb{R}^{n\times n}$. La trasformazione semi-ortogonale Q_i non causa quindi una perdita di informazione perché il modello Promises, nel caso in cui $n < m$, lavora sulle prime n componenti di R_i.

Nonostante il massimo sia equivalente, proiettare la soluzione del modello ProMises Efficiente nello spazio di dimensione $m \times m$, $Q_i R_i^* Q_i^\top$, non restituirà la stessa soluzione del modello ProMises, R_i, perché i due modelli sono definiti sotto vincoli diversi. Infatti, $R_i \in \mathcal{O}(m)$ mentre $R_i^* \in \mathcal{O}(n)$.

Formalmente, siano $X_1, \ldots, X_N \in \mathbb{R}^{n\times m}$ le matrici da allineare di rango n. Il modello ProMises Efficiente è quindi espresso come

$$X_i Q_i = \alpha_i (M^* + E_i) R_i^{*\top}, \tag{2.23}$$

dove $E_i \sim \mathcal{MN}_{n,n}(0, \sigma^2 I_n, I_n)$, R_i^* ha distribuzione von Mises-Fisher con parametro di posizione $F^* \in \mathbb{R}^{n \times n}$ e parametro di concentrazione k e $M^* \in \mathbb{R}^{n \times n}$. Per ridefinire F^* valgono considerazioni analoghe a quelle fatte nel paragrafo 2.2.3.1 per la definizione di F: si vuole che F^* rifletta l'informazione riguardante l'organizzazione spaziale dei dati. F^* può quindi essere definito come una matrice identità di ordine n, come una matrice simmetrica avente 1 sulla diagonale e valori decrescenti fuori dalla diagonale o come un'approssimazione a rango ridotto della matrice F definita in equazione (2.17).

L'algoritmo per il calcolo dei parametri di rotazione e *scaling* con il modello ProMises Efficiente è il seguente:

Algoritmo 3 Modello ProMises Efficiente (Andreella e Finos, 2022). `T` è un valore soglia per la distanza tra matrici medie di due iterazioni successive scelto a priori e `maxIt` è il numero massimo di iterazioni.

Input: $\boldsymbol{X}_i$, `T` , `maxIt`, $\forall i = 1, \ldots, N$

Output: $\hat{\boldsymbol{X}}_i$ $\forall i = 1, \ldots, N$

for $i = 1$ to N **do**
  $\boldsymbol{L}_i \boldsymbol{S}_i \boldsymbol{Q}_i^\top = \texttt{SVD}(\boldsymbol{X_i})$ ▷ *thin*-SVD
  $\boldsymbol{X_i^*} = \boldsymbol{X_i Q_i}$
end for
$\hat{\boldsymbol{M}} = \sum_{i=1}^{N} \boldsymbol{X}_i^* / N$, $\hat{\alpha_i} = 1$
`count` $= 0$, `dist` $=$ `Inf`
while `dist` $>$ `T` OR `count` $<$ `maxIt` **do**
  for $i = 1$ to N **do**
    $\boldsymbol{U}_i \boldsymbol{D}_i \boldsymbol{V}_i^\top = \texttt{SVD}(\boldsymbol{X_i}^{*\top} \hat{\boldsymbol{M}} + k^* \boldsymbol{F}^*)$ ▷ Decomposizione a valori singolari
    $\hat{\boldsymbol{R}}_i^* = \boldsymbol{U}_i \boldsymbol{V}_i^\top$
    $\hat{\boldsymbol{X}}_i^* = \boldsymbol{X}_i^* \hat{\boldsymbol{R}}_i^*$
    $\hat{\alpha_i}_{\hat{\boldsymbol{R}}_i} = \|\boldsymbol{X}_i^{*\top}\|_F^2 / \mathrm{Tr}(\boldsymbol{D}_i)$
    $\hat{\boldsymbol{X}}_i^* = \hat{\alpha_i}_{\hat{\boldsymbol{R}}_i}^{-1} \hat{\boldsymbol{X}}_i^*$ ▷ Aggiorna $\boldsymbol{X}_i^*$
  end for
  $\hat{\boldsymbol{M}}_{\text{old}} = \hat{\boldsymbol{M}}$ ▷ Salva $\hat{\boldsymbol{M}}$
  $\hat{\boldsymbol{M}} = \sum_{i=1}^{N} \hat{\boldsymbol{X}}_i^* / N$ ▷ Aggiorna $\hat{\boldsymbol{M}}$
end while
$\hat{\boldsymbol{X}}_i = \hat{\boldsymbol{X}}_i^* \boldsymbol{Q}_i^\top$

Oltre alla ridotta complessità computazionale, un aspetto vantaggioso di questa versione dell'algoritmo è che il fatto che sia applicabile a matrici con diverso numero di colonne. Siano $X_1, \ldots, X_N$ matrici di dimensione rispettivamente $n \times m_1, \ldots, n \times m_N$. Le matrici $Q_i^\top$ derivanti dalla *thin*-SVD delle X_i sono anch'esse di dimensione $n \times m_i$ pertanto, moltiplicando ogni X_i per la Q_i corrispondente, si ottengono matrici di uguale dimensione $n \times n$ alle quali si può applicare l'algoritmo del modello ProMises. Una volta ottenute le matrici ruotate $\hat{X}_i^* = X_i^* R_i^* \in \mathbb{R}^{n \times n}$, esse vengono moltiplicate per la trasformazione inversa $Q_i^\top \in \mathbb{R}^{n \times m_i}$ per ottenere le matrici allineate

$\hat{X}_i \in \mathbb{R}^{n \times m_i}$. In questo caso, per tenere conto della localizzazione spaziale, si può assumere una diversa distribuzione a priori per ogni R_i^* considerando un diverso parametro di posizione F_i^*, sempre con il vincolo che $\mathrm{Tr}(F_i^*) = n$. Sia Q_i la matrice semi-ortogonale derivante dalla decomposizione della matrice i e sia C_i la matrice di dimensione $m_i \times 2$ contenente le coordinate in due (o in tre) dimensioni dei punti della matrice X_i. Una valida proposta per la definizione di F_i^* consiste nell'applicare a C_i la trasformazione semi-ortogonale Q_i, ottenendo $C_i^* = Q_i^\top C \in \mathbb{R}^{n \times 2}$, calcolare la matrice delle distanze euclidee D_i^* utilizzando le coordinate in C_i^* e porre $F_i^* = \exp\{-D_i^*\}$.

Nel caso in cui le matrici X_i abbiano la stessa dimensione, per ridurre ulteriormente il tempo necessario, si può considerare come trasformazione semi-ortogonale la matrice Q derivante dalla *thin*-SVD della matrice media aritmetica elemento per elemento delle matrici da allineare, $M = \sum_{i=1}^{N} X_i/N$. In questo modo, al posto di decomporre N matrici di dimensione $n \times m$, ne viene decomposta solamente una. L'algoritmo è quindi identico all'Algoritmo 3, ad eccezione delle prime 3 righe che diventano:

Algoritmo 4 Modifica modello ProMises Efficiente (Andreella, 2021)

Input: $\boldsymbol{X}_i$, `T` , `maxIt`, $\forall i = 1, \ldots, N$

Output: $\hat{\boldsymbol{X}}_i$ $\forall i = 1, \ldots, N$

1: $\hat{\boldsymbol{M}} = \sum_{i=1}^{N} \boldsymbol{X}_i/N$

2: $\boldsymbol{LSQ}^\top = \texttt{SVD}(\hat{\boldsymbol{M}})$ ▷ *thin*-SVD

3: $\boldsymbol{X_i^* = X_i Q}$

2.3 Conclusioni

Riassumendo, nel caso in cui si vogliano allineare due sole matrici il problema di Procuste fornisce una soluzione esplicita per i parametri di rotazione e di *scaling*, ottenuta tramite il criterio dei minimi quadrati, che è unica nel caso in cui la matrice $X_1^\top X_2$ sia a rango pieno. Nel caso invece in cui si vogliano allineare più matrici, è stata proposta la *Generalized Procustes Analysis*

(GPA). Sia M la matrice che rappresenta lo spazio comune. La GPA, nel caso in cui M sia nota, fornisce una soluzione unica solo se le matrici da decomporre sono a rango pieno mentre, nel caso in cui M non sia nota, accetta come soluzione valida qualsiasi rotazione della soluzione trovata. Il fatto che siano ammissibili diverse soluzioni è un aspetto critico nell'analisi di dati di trascrittomica spaziale perché fa sì che si perda l'interpretazione spaziale.

Il problema di Procuste è stato riformulato sotto forma di modello statistico da Goodall (1991) che ha definito il *perturbation model*. A partire dal *perturbation model*, Andreella e Finos (2022) hanno proposto due modelli, il modello ProMises e il modello ProMises Efficiente, che rivisitano l'analisi in contesto bayesiano. Entrambi i modelli assumono una distribuzione normale matriciale per le matrici da allineare e una distribuzione von Mises-Fisher come distribuzione a priori per il parametro di rotazione. La distribuzione von Mises-Fisher è una priori coniugata per la distribuzione normale matriciale, quindi è facile risalire alla distribuzione a posteriori. Aspetti vantaggiosi di questi modelli sono il fatto che forniscono algoritmi per il calcolo degli stimatori dei parametri di rotazione e *scaling* che portano a soluzioni uniche e il fatto che permettono di inserire l'informazione riguardante la localizzazione spaziale dei punti da allineare tramite un'opportuna specificazione del parametro di posizione della a priori. Gli algoritmi che derivano da questi modelli risultano in una lieve modifica della GPA. Il modello ProMises Efficiente riduce nettamente la complessità computazionale nel caso in cui $n \ll m$ ed è adattabile, nel caso i cui le matrici abbiano rango n, al caso in cui si vogliano allineare matrici con diverso numero di colonne.

Capitolo 3

Allineamento Procuste dei dati di trascrittomica spaziale

Si riportano in questo capitolo i risultati ottenuti dall'applicazione del problema di Procuste e del modello ProMises a dati di trascrittomica spaziale. Si sono seguite due strade: in un primo momento sono state allineate due sole immagini in maniera da poter sfruttare la soluzione esplicita riportata in equazione (2.2), successivamente sono state allineate più immagini tramite il modello ProMises Efficiente. È stato scelta la versione efficiente del modello ProMises per motivi computazionali.

L'obiettivo dell'applicazione è duplice. Innanzitutto, si punta a mostrare che in questo contesto il metodo funziona e rende le immagini allineate effettivamente più simili tra loro rispetto alle immagini non allineate. In un secondo momento, si vuole mostrarne l'efficacia da un punto di vista biologico. A questo scopo si condurranno due tipi di analisi. Inizialmente, prendendo spunto dall'articolo di riferimento (Maynard et al., 2021), si sfrutterà l'informazione relativa agli strati corticali. Successivamente, come già accennato nel paragrafo di presentazione dei dati, si considererà l'appartenenza dell'immagine a un individuo come condizione biologica e si applicheranno modelli di espressione differenziale.

3.1 Analisi preliminari

I dati di espressione genica necessitano generalmente di alcune analisi preliminari. In particolare, due delle operazioni più comunemente eseguite consistono nel filtraggio dei geni poco espressi e nella normalizzazione, processo che elimina errori sistematici e distorsioni. Si hanno infatti normalmente a disposizione i conteggi relativi a decine di migliaia di geni, di cui solo una piccola parte espressa, e di conseguenza si preferisce eliminare i geni non espressi. Inoltre, ogni campione può differire dagli altri per il numero di *reads* sequenziate, per differenze nella preparazione delle librerie e nei protocolli di sequenziamento, e tutto ciò porta ad osservare dati distorti. Si vuole quindi normalizzare i dati in maniera che riflettano unicamente le differenze biologiche tra i diversi campioni e non gli errori sistematici dovuti ad altri fattori. Si descrivono di seguito le operazioni preliminari che sono state applicate ai dati di Maynard et al. (2021).

3.1.1 Filtraggio

Inizialmente, i dati di Maynard et al. (2021) presentavano i conteggi dell'espressione genica di 33538 geni. Nell'analisi si è deciso di considerare solamente i 1000 geni più variabili (HVG). Per selezionare i 1000 geni HVG, la varianza dei profili di log-espressione genica è stata suddivisa in varianza tecnica e varianza biologica, che è la componente di interesse. Per fare ciò è stato stimato sulla base dei dati un trend media-varianza a partire dal quale, per ogni valore medio di log-espressione, è stato calcolato il valore di varianza stimato. Il valore stimato è stato considerato come componente di varianza tecnica, mentre la differenza tra il valore stimato e il valore osservato è stata considerata come componente di varianza biologica. I geni sono stati ordinati per varianza biologica crescente e sono stati selezionati i 1000 geni per cui essa era più alta.

3.1.2 Trasformazione pre-allineamento e normalizzazione

Nel seguito del Capitolo si metteranno a confronto i risultati ottenuti dai dati allineati con i risultati ottenuti dai dati normalizzati seguendo una *pipeline* standard. Di conseguenza, sono state eseguite due normalizzazioni differenti, una per preparare i dati alla rotazione Procuste e una per ottenere i dati con cui confrontarli.

Venendo ora alla trasformazione pre-allineamento, si è visto nel Capitolo 2 che il modello ProMises assume che le matrici da allineare siano centrate e che seguano una distribuzione normale matriciale. Si è quindi deciso di applicare delle trasformazioni ai conteggi grezzi in maniera da ottenere dati che rispettassero tali assunzioni. In particolare, per ogni *spot*

- è stato assegnato ad ognuno dei 1000 geni il rango corrispondente;
- il rango è stato diviso per 1000;
- è stato assegnato ad ogni valore il quantile corrispondente di una normale standard.

Un problema sorto nella trasformazione in ranghi è stato che, per ogni *spot*, molti geni non risultavano espressi, pertanto il valore relativo al conteggio è zero e si hanno di conseguenza molti *ties*. Per assegnare i ranghi ai *ties* sono stati adottati due approcci: assegnare ai valori uguali un rango casuale o calcolare la media degli indici ordinati. I risultati ottenuti con i due metodi portano alle stesse conclusioni e di conseguenza si riportano nella tesi solamente i risultati derivanti dall'approccio casuale.

Per quanto riguarda invece la normalizzazione classica, è stato seguito il metodo più in voga per l'analisi di dati di *sc*RNA-*Seq* normalizzando i conteggi tramite la funzione `logNormCounts` del pacchetto `scuttle` (McCarthy et al., 2017). Ogni conteggio viene diviso per un *size factor* cellula specifico, dato dalla *library size* normalizzata in maniera che la media su tutte le cellule sia pari ad 1, e ne viene preso il logaritmo in base 2. Quando nel

seguito si parlerà di dati non allineati e si applicheranno a tali dati metodi statistici per il *clustering* o per l'espressione differenziale si intenderanno i dati log-normalizzati con il procedimento appena descritto.

3.2 Allineamento e confronto con i dati non allineati

3.2.1 Allineamento di due immagini

Il primo passo dell'applicazione è stato considerare l'allineamento di due sole immagini, in maniera da poter contare su una soluzione in forma chiusa. Sono inizialmente state scelte due immagini di uno stesso soggetto, già simili in partenza. In particolare, le immagini scelte sono l'immagine 151673 e l'immagine 151674. L'immagine 151673 viene ruotata mentre l'immagine 151674 funge da riferimento. Sono stati esclusi dall'analisi gli *spot* non presenti in entrambe le immagini e si sono di conseguenza ottenute matrici di dimensione 1000×2967.

Per verificare che l'allineamento aumenti la somiglianza tra le due immagini, è stata innanzitutto valutata la distanza tra le due matrici pre e post allineamento tramite il calcolo della norma di Frobenius della matrice delle differenze ($\|X_1\hat{R} - X_2\|_F$ contro $\|X_1 - X_2\|_F$). Per definizione infatti, la matrice $\hat{R}$ deve rendere tale distanza minima. Si osserva come la distanza cali, infatti $\|X_1 - X_2\|_F^2 = 1886.148$ mentre $\|X_1\hat{R} - X_2\|_F^2 = 573.468$.

Per valutare poi che i dati riflettessero effettivamente una maggiore omogeneità, sono stati applicati algoritmi di *clustering* separatamente ad entrambe le immagini. Le partizioni ottenute sono state confrontate tramite indice di Rand aggiustato (Hubert e Arabie, 1985) considerando come stessa unità statistica gli *spot* con le stesse coordinate. Prendendo spunto dall'articolo di riferimento (Maynard et al., 2021) si vuole che la partizione finale sia formata da 8 *cluster*, uno per ogni strato (6 strati corticali e la materia bianca) più uno per eventuali *outliers*. Sono stati quindi scelti algoritmi che permettessero di

controllare il numero finale di *cluster*. In particolare, sono stati applicati un algoritmo basato sulle reti, il *walktrap* (Pons e Latapy, 2005), e due metodi partizionali, l'algoritmo delle K-medie (Hartigan e Wong, 1979) e il *Partitioning Aroung Medoid* (Kaufman e Rousseeuw, 1990), che verrà abbreviato in PAM. Dato che a seguito del filtraggio dei geni sono state mantenute 1000 variabili, per evitare di incorrere nella maledizione della dimensionalità tali algoritmi sono stati applicati sulle prime 50 componenti principali. In Tabella 3.1 sono riportati gli indici di Rand aggiustati calcolati tra le partizioni ottenute con i tre algoritmi pre e post allineamento Procuste. Si può notare come la concordanza sia sempre più alta quando calcolata sulle immagini allineate. L'incremento maggiore si ha con l'algoritmo *walktrap*: l'indice di Rand sale infatti da 0.149 a 0.220.

In Figura 3.1 è rappresentato un confronto grafico tra la partizione derivante dall'applicazione dell'algoritmo *walktrap* all'immagine 151674 (immagine di riferimento) e le partizioni derivanti dall'applicazione dell'algoritmo *walktrap* all'immagine 151673 prima e dopo l'allineamento Procuste. Gli *spot*, rappresentati in colonna, sono colorati sulla base del *cluster* a cui appartengono. Osservando il *cluster* verde, il *cluster* arancione e il *cluster* viola chiaro appare evidente come la partizione ottenuta a seguito dell'allineamento sia più simile a quella ottenuta sull'immagine di riferimento rispetto a quella ottenuta prima dell'allineamento.

La maggior somiglianza emerge anche dalle immagini in Figura 3.2. Ogni immagine rappresenta gli *spot* analizzati nello spazio definito dalle proprie coordinate. Gli *spot* sono colorati sulla base del *cluster* a cui appartengono. Le due immagini sulla prima riga rappresentano i *cluster* ottenuti prima dell'allineamento, mentre l'immagine sulla seconda riga rappresenta i *cluster* ottenuti sull'immagine 151673 allineata. Appare evidente come le partizioni ottenute a seguito dell'allineamento siano più simili.

Tabella 3.1: Indice di Rand aggiustato tra le partizioni ottenute con diversi algoritmi di *clustering* per le immagini 151673 e 151674 pre e post allineamento Procuste.

Algoritmo	Immagini non allineate	Immagini allineate
Walktrap	0.149	0.220
k-medie	0.219	0.242
PAM	0.110	0.168

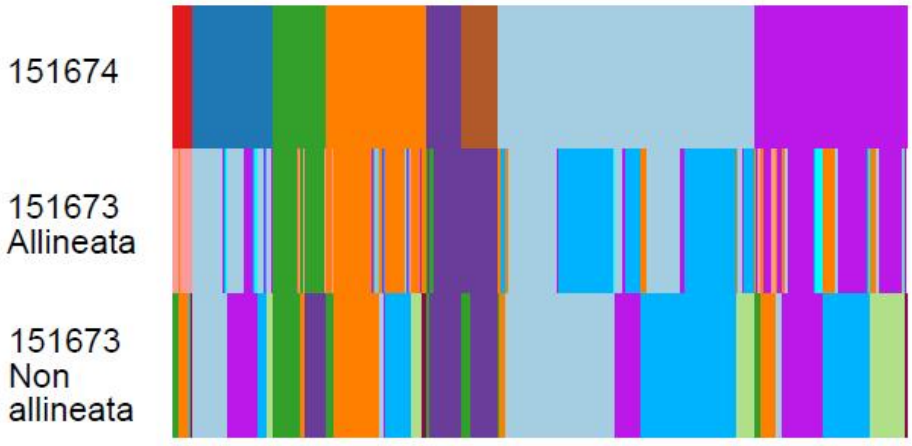


Figura 3.1: Confronto grafico fra la partizione derivante dall'immagine 151674 (immagine di riferimento) e le partizioni derivanti dall'immagine 151673, prima e dopo l'allineamento.

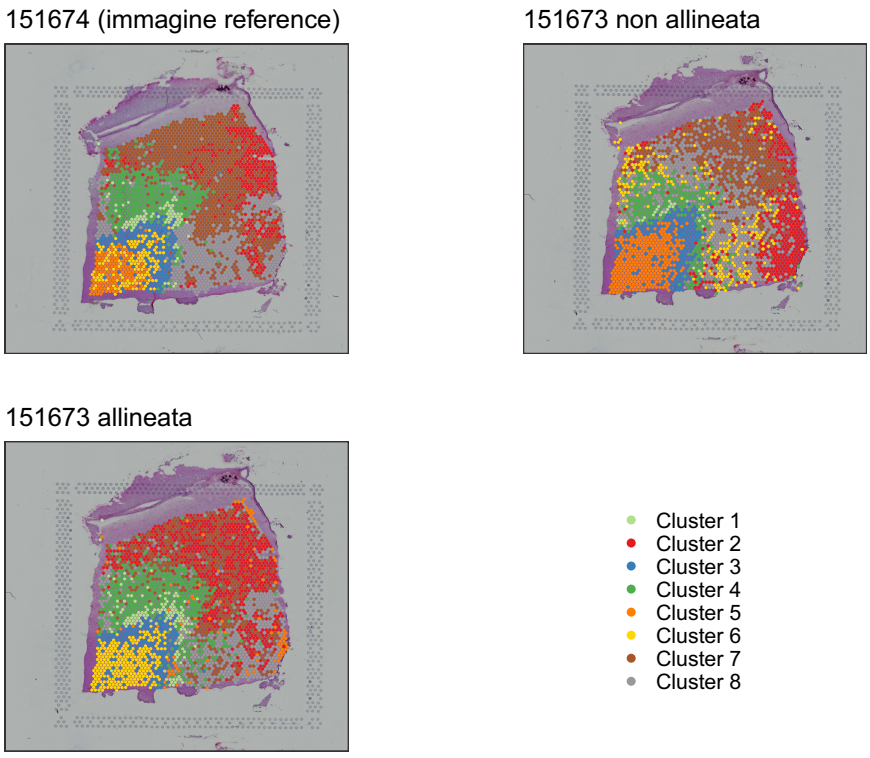


Figura 3.2: Immagini 151673 e 151674 con i *cluster* derivanti dal *walktrap* applicato ai dati non allineati (prima riga) e allineati (seconda riga).

Per verificare che l'incremento non fosse dovuto al caso l'analisi è stata ripetuta mantenendo fissa l'immagine 151674 come immagine di riferimento e cambiando l'immagine da allineare. Sono sempre stati tenuti solamente gli *spot* presenti in entrambe le immagini. In Tabella 3.2 sono riportate le distanze, sempre calcolate tramite la norma di Frobenius della matrice delle differenze, tra la matrice 151674 e le altre matrici considerate prima e dopo l'allineamento, mentre in Tabella 3.3 sono riportati gli indici di Rand aggiustati tra le partizioni ottenute sull'immagine di riferimento e sulle altre immagini prima e dopo l'allineamento, sempre tenendo fisso ad 8 il numero finale di *cluster*. La distanza tra matrici è sempre inferiore quando calcolata sulle immagini allineate. Si noti inoltre che la distanza tra le matrici prima dell'allineamento è superiore se calcolata per le immagini 151507 e 151669, a

testimonianza del fatto che le matrici dello stesso soggetto fossero in origine più simili. Questo si verifica anche dopo l'allineamento, infatti la distanza dalla matrice di riferimento per le matrici 151673 e 151674 è inferiore rispetto alla stessa calcolata per le altre due matrici considerate. Inoltre, la concordanza tra le partizioni tra le due immagini è sempre maggiore a seguito della rotazione Procuste.

Si conclude quindi che allineare dati di trascrittomica spaziale tramite il metodo di Procuste rende effettivamente le immagini più omogenee.

Tabella 3.2: Norma di Frobenius tra diverse matrici e la matrice 151674 pre e post-allineamento.

Immagine		Distanza
151675	Pre allineamento	1933.577
	Post Allineamento	609.217
151507	Pre allineamento	2167.698
	Post allineamento	694.156
151669	Pre allineamento	2053.806
	Post allineamento	644.0647

Tabella 3.3: Indice di Rand aggiustato per il confronto tra le partizioni ottenute su varie immagini e la partizione ottenuta sull'immagine 151674 considerando sia i dati allineati che i dati non allineati. L'immagine 151674 è sempre presa a riferimento.

Immagine		Walktrap	K-medie	PAM
151675	Non allineate	0.168	0.148	0.077
	Allineate	0.185	0.218	0.107
151507	Non allineate	0.075	0.140	0.057
	Allineate	0.145	0.247	0.110
151669	Non allineate	0.108	0.149	0.033
	Allineate	0.153	0.205	0.083

3.2.2 Allineamento di più immagini

Si procede ora con l'allineamento di più immagini. Analogamente a quanto fatto nel paragrafo precedente, si vuole dimostrare che la rotazione rende le immagini più simili.

Sono state eseguite due prove: inizialmente sono state allineate le 4 immagini di uno stesso individuo (immagini 151673, 151674, 151675 e 151676), successivamente 3 immagini di individui diversi (immagini 151673, 151669 e 151507). Per motivi computazionali, le immagini sono state allineate tramite il modello ProMises Efficiente. In particolare, è stata applicata la versione dell'algoritmo che moltiplica le matrici da allineare per la matrice semi-ortogonale Q^T derivante dalla *thin*-SVD della matrice media $M = \sum_{i=1}^{N} X_i/N = LSQ^T$ (Algoritmo 4). Dato che questa versione dell'algoritmo richiede che le immagini da allineare siano della stessa dimensione, in entrambi i casi è stata selezionata l'intersezione delle immagini, scartando gli *spot* per cui l'informazione relativa ai livelli di espressione genica non era

presente in tutte le matrici. Nel primo caso sono state ottenute matrici di dimensione 1000×2967, nel secondo caso matrici di dimensione 1000×3227. Si è considerata uguale distribuzione a priori per tutti i parametri ortogonali R_i. Per definire il parametro di posizione $F^* \in \mathbb{R}^{1000\times1000}$ è stata moltiplicata la matrice delle coordinate per la matrice semi-ortogonale Q^T, è stata calcolata la matrice delle distanze euclidee D^* tra gli *spot* considerando queste coordinate ridotte e infine il parametro di posizione è stato posto uguale a $F^* = \exp(-D^*)$. In Figura A.1 in Appendice sono riportate le *heatmap* del parametro di posizione F^* ottenuto per l'allineamento delle quattro immagini dello stesso individuo (grafico a sinistra) e delle tre immagini di individui diversi (grafico a destra). Si noti come i valori nella diagonale principale siano tutti pari a 1. Il valore soglia `T` per la distanza tra le matrici medie di due iterazioni successive, calcolata come la norma di Frobenius della matrice differenza, $\|M - M_{old}\|_F^2$, è stato posto uguale a 1. In entrambe le prove l'algoritmo ha impiegato 7 iterazioni per arrivare a convergenza (Figura 3.3).

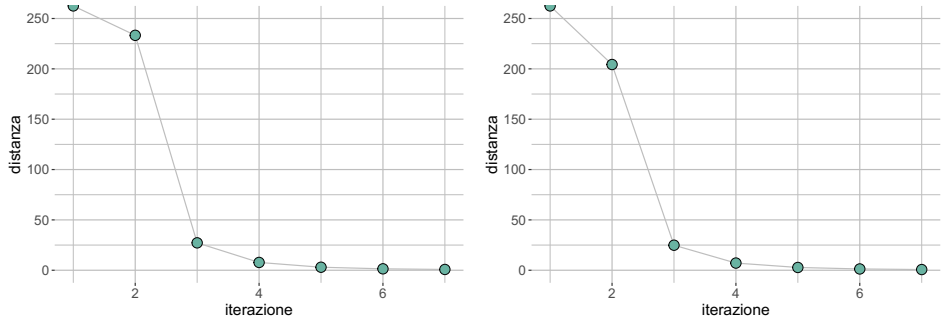


Figura 3.3: Distanza tra le matrici medie di due iterazioni successive per l'allineamento delle tre immagini di individui diversi (grafico a destra) e per le quattro immagini dello stesso individuo (grafico a sinistra).

Anche in questo caso si vuole verificare che la definizione sia soddisfatta. In Figura 3.4 e in Figura 3.5 sono riportate le distanze, calcolate tramite la norma di Frobenius della matrice differenza, tra le varie matrici prima (grafici a sinistra) e dopo (grafici a destra) l'allineamento Procuste. In Figura 3.4 sono rappresentate le distanze tra le quattro matrici dello stesso soggetto, in

Figura 3.5 le distanze tra le tre matrici di soggetti diversi. In entrambi i casi appare evidente come la rotazione Procuste renda più omogenee le diverse immagini, infatti le distanze tra le matrici a seguito dell'allineamento sono sempre inferiori rispetto alle distanze tra le matrici prima dell'allineamento. Confrontando il grafico a sinistra in Figura 3.4 e il grafico a sinistra in Figura 3.5 si nota inoltre che le tre immagini di individui diversi erano in origine più diverse rispetto alle quattro immagini dello stesso individuo e rimangono più diverse anche a seguito dell'allineamento. La distanza tra esse è infatti sempre superiore rispetto alla distanza tra le immagini allineate (grafici a destra) dello stesso soggetto.

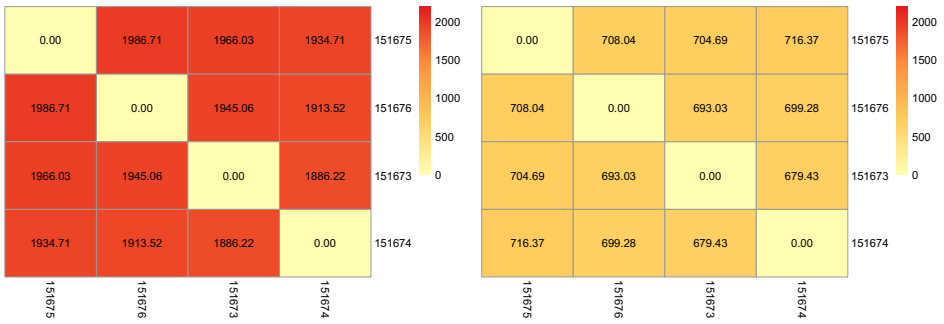


Figura 3.4: Norma di Frobenius della differenza tra le matrici prima e dopo l'allineamento tramite il modello ProMises Efficiente.

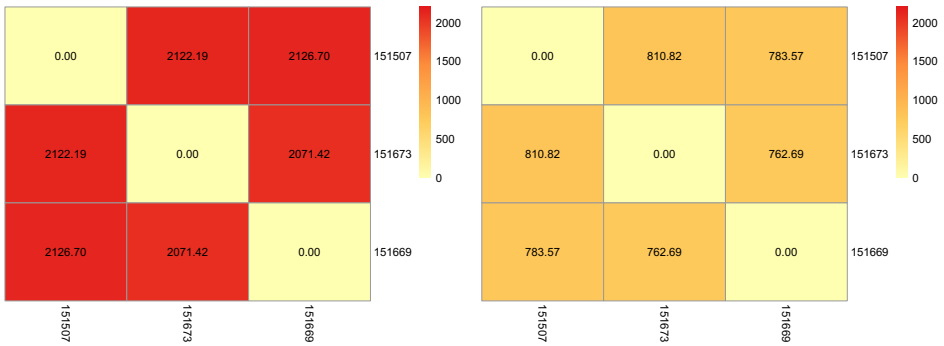


Figura 3.5: Norma di Frobenius della differenza tra le matrici prima e dopo l'allineamento tramite il modello ProMises Efficiente.

Per mostrare ora che l'allineamento Procuste rende le immagini più omogenee rispetto alle matrici normalizzate secondo la *pipeline* standard, si è deciso di calcolare e analizzare la matrice media delle immagini log-normalizzate e la stessa delle immagini ruotate. Ogni punto di questa nuova matrice è dato dalla media aritmetica dei punti nella stessa posizione. Supponendo che immagini derivanti da diversi campioni non siano allineate dal punto di vista funzionale e quindi non confrontabili, segue che prima dell'allineamento punti con le stesse coordinate di immagini diverse non siano in realtà punti corrispondenti mentre dopo l'allineamento sì. Fare la media delle matrici non allineate quindi, dato che le coordinate degli *spot* non sono veritiere, porta ad ottenere una matrice casuale. Al contrario, dopo l'allineamento, si calcola la media di oggetti omogenei.

Sulle immagini medie sono stati applicati algoritmi di *clustering*. Sono stati applicati due algoritmi basati sulle reti, l'algoritmo di Louvain (Blondel et al., 2008) e il *walktrap*, al quale in questo caso non viene applicata una potatura. In questo modo non è necessario specificare a priori il numero di *cluster* desiderati. Per la costruzione del grafo si adotta l'approccio *shared nearest neighbour*, cioè due punti vengono considerati connessi se hanno un numero di vicini più vicini in comune che è superiore a un iperparametro da definire, k. Dato che se si applicasse l'algoritmo considerando come variabili i 1000 geni si incorrerebbe nella maledizione della dimensionalità, vengono calcolate, sia per la matrice media delle immagini allineate sia per la matrice media delle immagini log-normalizzate, le prime 50 componenti principali e si applica il *clustering* su di esse. Come misura di bontà delle partizioni ottenute si considerano la *silhouette* media globale e la *silhouette* media per *cluster* (Rousseeuw, 1987). Dato che la media delle immagini prima dell'allineamento è una media di punti non corrispondenti mentre la media dopo l'allineamento sì, ci si aspetta che i *cluster* ottenuti sulla media delle matrici allineate siano più omogenei rispetto a quelli ottenuti sulla media delle matrici non allineate. Si riportano in Figura 3.6 i valori della *silhouette* media al variare dell'iperparametro k per l'algoritmo di Louvain e in Figura 3.7 i valori

della *silhouette* media al variare dell'iperparametro k per il *walktrap*. I grafici a sinistra riportano i risultati derivanti dall'allineamento delle 4 immagini dello stesso soggetto, i grafici a destra i risultati derivanti dall'allineamento delle tre immagini di soggetti diversi. La *silhouette* è stata calcolata considerando come matrice delle distanze la matrice delle distanze euclidee calcolate sulle componenti principali derivanti dalla media delle matrici non allineate per le immagini non allineate e la matrice delle distanze euclidee calcolate sulle componenti principali derivanti dalla media delle matrici allineate per le immagini allineate. Nel caso dell'allineamento delle 4 immagini dello stesso individuo si può notare come per l'algoritmo di Louvain, a partire da $k = 80$, la *silhouette* calcolata sulla media delle immagini allineate sia sempre superiore a quella calcolata sulla media delle immagini non allineate. Per quanto riguarda invece il *walktrap* si ha che la *silhouette* massima ottenuta dalla media dei dati allineati, che si ha per $k = 160$, è superiore rispetto alla *silhouette* massima ottenuta dalla media dei dati non allineati, che si ha per $k = 70$, mentre per gli altri valori dell'iperparametro le spezzate sono confrontabili. Per quanto riguarda invece l'allineamento delle tre immagini di individui diversi, con l'algoritmo di Louvain si ha che la *silhouette* post-allineamento è sempre superiore, mentre con il *walktrap* si ha di nuovo che la *silhouette* massima ottenuta dalla media dei dati allineati, che si ha per $k = 140$, è superiore rispetto alla *silhouette* massima ottenuta dalla media dei dati non allineati, che si ha per $k = 40$. Per gli altri valori dell'iperparametro, la spezzata dei valori di *silhouette* media dei dati allineati è sempre superiore a partire da $k = 120$, mentre per valori inferiori le due spezzate si intersecano. Si ottengono quindi risultati più forti per l'allineamento di tre immagini di individui diversi, a causa del fatto che le quattro immagini dello stesso individuo fossero già in origine relativamente ben allineate.

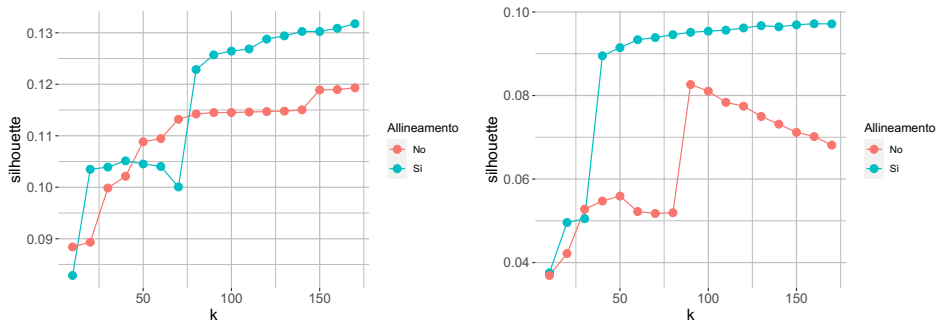


Figura 3.6: Silhouette medie globali derivanti dall'applicazione dell'algoritmo di Louvain al variare dei valori dell'iperparametro k che rappresenta il numero di vicini più vicini condivisi. La spezzata rosa deriva dall'applicazione dell'algoritmo alla media delle immagini non allineate, la linea azzurra invece deriva dall'applicazione dell'algoritmo alle immagini allineate. Il grafico sulla sinistra deriva dall'applicazione alla media delle quattro immagini dello stesso individuo, il grafico sulla destra dall'applicazione alla media delle tre immagini di individui diversi.

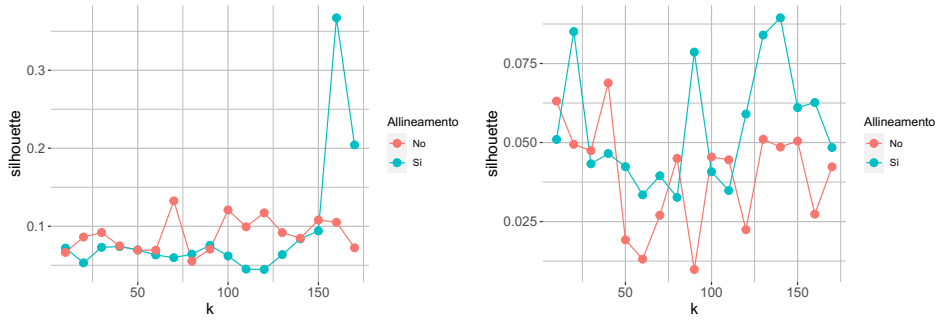


Figura 3.7: Silhouette medie globali derivanti dall'applicazione dell'algoritmo *walktrap* al variare dei valori dell'iperparametro k che rappresenta il numero di vicini più vicini condivisi. La spezzata rosa deriva dall'applicazione dell'algoritmo alla media delle immagini non allineate, la linea azzurra invece deriva dall'applicazione dell'algoritmo alle immagini allineate. Il grafico sulla sinistra deriva dall'applicazione alla media delle quattro immagini dello stesso individuo, il grafico sulla destra dall'applicazione alla media delle tre immagini di individui diversi.

Si riportano inoltre in Appendice i grafici delle *silhouette* punto per punto, con l'indicazione della *silhouette* media globale e media per *cluster*, per le partizioni derivanti dall'algoritmo di Louvain e dal *walktrap* applicati all'immagine media delle quattro immagini dello stesso individuo (rispettivamente Figura A.2 e Figura A.4) e all'immagine media delle tre immagini di individui diversi (rispettivamente Figura A.3 e Figura A.5), calcolate per i valori dell'iperparametro k che portano una *silhouette* maggiore. Si nota nuovamente che le partizioni ottenute sulle matrici medie delle immagini allineate sono migliori.

Si conclude quindi che sia la soluzione esplicita del metodo di Procuste per l'allineamento di due immagini, sia l'algoritmo del modello ProMises Efficiente funzionano e riducono le distanze tra le immagini di trascrittomica spaziale. Inoltre, da un confronto con i dati log-normalizzati secondo una *pipeline* standard, si nota che i dati trasformati con una rotazione Procuste risultano più omogenei.

3.3 Applicazione biologica

Dimostrato che l'allineamento Procuste rende i dati di trascrittomica spaziale di diversi campioni più omogenei, si vuole ora mostrare che ha anche un beneficio dal punto di vista biologico.

3.3.1 Concordanza con gli strati

Si è inizialmente pensato di sfruttare l'informazione relativa allo strato corticale a cui gli *spot* appartengono lavorando su due sole immagini. In particolare, sono nuovamente state scelte l'immagine 151673 e l'immagine 151674. Uno degli obiettivi dell'articolo di riferimento (Maynard et al., 2021) è stato infatti applicare algoritmi di *clustering* con diversi sottoinsiemi di geni (geni HVG, geni spazialmente variabili e geni marcatori noti in letteratura) per trovare le partizioni che restituiscano maggiore concordanza con i *cluster* definiti dagli strati. In tutti i casi non hanno raggiunto risultati particolarmente

buoni, infatti gli indici di Rand aggiustati tra le loro partizioni e le etichette che identificano gli strati sono pari a 0.2. Nel momento in cui si allineano due immagini, l'immagine che rimane fissa e funge da *reference* assume il ruolo di spazio di riferimento. Si può quindi pensare di prendere le etichette di appartenenza agli strati degli *spot* dell'immagine 151674 come riferimento. Ci si aspettava quindi che, se i *cluster* ottenuti riflettono effettivamente l'appartenenza agli strati, l'indice di Rand tra i *cluster* ottenuti a partire dai dati per l'immagine che viene ruotata e la partizione data dalle etichette degli strati dell'immagine *reference* aumenti a seguito dell'allineamento, ma questo non si verifica. In Tabella 3.4 sono riportati gli indici di Rand aggiustati per le partizioni derivanti dal *clustering* applicato all'immagine 151673 prima e dopo l'allineamento e le etichette di appartenenza agli strati assegnate manualmente all'immagine 151674. Si può notare come non ci sia un netto incremento dell'indice di Rand aggiustato e che in entrambi i casi, per tutti gli algoritmi considerati, sia basso.

Tabella 3.4: Indice di Rand aggiustato tra le etichette assegnate manualmente all'immagine 151674 e le partizioni ottenute con diversi algoritmi di *clustering* per l'immagine 151673 pre e post allineamento Procuste.

Algoritmo	Immagini non allineate	Immagini Allineate
Walktrap	0.173	0.190
k-medie	0.211	0.210
PAM	0.115	0.144

Si è quindi pensato di calcolare la matrice data dalla media aritmetica punto per punto delle due immagini prima e dopo l'allineamento e applicare su di essa algoritmi di *clustering*, sempre tenendo fisso il numero finale di *cluster* cercati a otto. Anche in questo caso ci si aspettava che la concordanza tra la partizione ottenuta e le etichette assegnate manualmente all'immagine

151674, sempre misurata tramite l'indice di Rand aggiustato, aumentasse a seguito della rotazione. I risultati ottenuti sono riassunti in Tabella 3.5. L'indice di Rand post rotazione è sempre inferiore rispetto all'indice di Rand prima della rotazione. Si conclude quindi che i *cluster* ottenuti a seguito dell'allineamento non riflettono l'organizzazione laminare dell'immagine di riferimento.

Tabella 3.5: Indice di Rand aggiustato tra le etichette assegnate manualmente all'immagine 151674 e le partizioni ottenute con diversi algoritmi di *clustering* per l'immagine media delle immagini 151673 e 151674 pre e post allineamento Procuste.

Algoritmo	Immagini non allineate	Immagini Allineate
Walktrap	0.218	0.166
k-medie	0.224	0.181
PAM	0.220	0.140

Per verificare inoltre la qualità delle partizioni ottenute tenendo fisso a 8 il numero finale di *cluster*, è stata calcolata la *silhouette* dei punti per i *cluster* ottenuti tramite algoritmo *walktrap* applicato alle immagini 151674, 151673 e 151673 dopo l'allineamento. Si può notare come le partizioni ottenute non siano soddisfacenti (Figura 3.8). La silhouette media globale è infatti sempre bassa pari in particolare a -0.01 nel caso dell'immagine 151673 prima dell'allineamento (grafico in alto a sinistra), 0.04 per l'immagine 151673 ruotata (grafico in alto a destra) e 0.03 per l'immagine 151674 (grafico in basso). Inoltre, anche la silhouette media per *cluster* è bassa per molti *cluster*, e diversi *spot* hanno silhouette negativa.

Al termine di questa applicazione si conclude quindi che allineare una matrice non fa sì che aumenti la concordanza con le etichette degli strati dell'immagine presa a riferimento e che tenere fisso ad otto il numero di

cluster non porta ad ottenere partizioni buone. I *cluster* che si ottengono non riflettono quindi l'organizzazione in strati della corteccia.

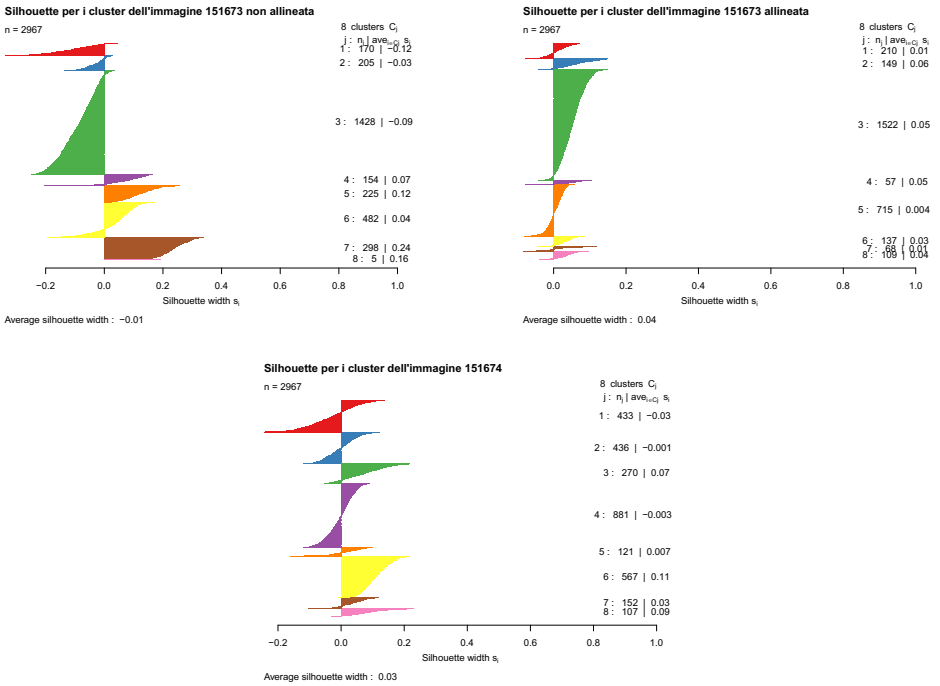


Figura 3.8: Silhouette per le partizioni derivanti dal *walktrap* con 8 *cluster* per l'immagine 151673 (figura in alto a sinistra), l'immagine 151673 allineata (figura in alto a destra) e l'immagine 151674 (figura in basso).

3.3.2 Inferenza

Si decide quindi di abbandonare la strada suggerita dall'articolo di riferimento e indagare come l'allineamento Procuste si comporta rispetto all'analisi differenziale. Idealmente, in un *dataset* con diversi individui in diverse condizioni biologiche, si vuole che l'allineamento assorba l'effetto di individuo e mantenga la variabilità biologica. La variabilità delle immagini non allineate può infatti essere scomposta in due parti: una componente di disturbo data dalla variabilità tecnica dovuta al mancato allineamento e una componente di interesse data dalla variabilità biologica. A seguito dell'allineamento, ci si aspetta che la varianza tecnica sia attutita.

I dati raccolti da Maynard et al. (2021) derivano tutti da cervelli di soggetti sani. Si può quindi assumere che i geni differenzialmente espressi che si trovano conducendo analisi differenziale considerando come covariata il soggetto di appartenenza siano falsi positivi o imputabili ad effetti di *batch*. Si vogliono quindi confrontare i risultati derivanti dall'applicazione di modelli per l'analisi differenziale sulle immagini non allineate, che ci si aspetta risentano fortemente di un effetto di soggetto, e sulle immagini allineate. A questo scopo verranno allineate 8 immagini, 4 di un soggetto e 4 di un altro (immagini 151673, 151674, 151675, 151676 e immagini 151507, 151508, 151509, 151510). Sono state scelte queste immagini perché attraversano tutti gli strati corticali (Figura 1.4). Per non scartare osservazioni, le immagini sono state allineate tramite la versione dell'algoritmo ProMises Efficiente descritta nell'Algoritmo 3. Dato che le matrici hanno quindi diverso numero di colonne, è stato definito un parametro di posizione diverso per la distribuzione a priori di ogni parametro di rotazione R_i^*, pari in particolare a

$$F_i^* = \exp(-D_i^*),$$

dove D_i^* è la matrice delle distanze euclidee calcolate a partire dalla matrice $C_i^* = C_i Q_i$, con C_i matrice con le coordinate in due dimensioni per la matrice i e Q_i matrice semi-ortogonale derivante dalla *thin*-SVD di X_i.

Il valore soglia per la distanza tra le matrici medie in due iterazioni successive è stato nuovamente posto pari a 1. L'algoritmo ha impiegato 5 iterazioni per arrivare a convergenza (Figura 3.9).

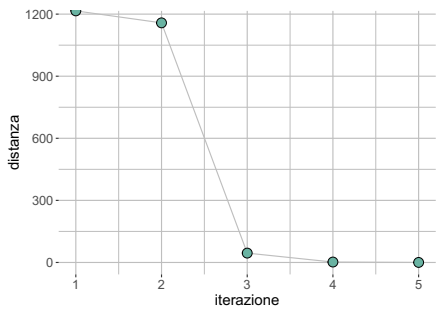


Figura 3.9: Distanza tra le matrici medie di due iterazioni successive per l'allineamento di 8 immagini.

Innanzitutto, è stata condotta un'analisi esplorativa per vedere se gli *spot* prima dell'allineamento esibiscono un marcato effetto di soggetto e quanto l'allineamento lo assorbe. È stato costruito un unico oggetto di dimensione 1000×32397 concatenando le otto matrici e sono state calcolate le prime 50 componenti principali a partire dai dati log-normalizzati e dai dati allineati. Per permettere un'adeguata visualizzazione della variabilità degli *spot* in due dimensioni, a partire dalle componenti principali è stato costruito il grafico *t-distributed stochastic neighbour embedding* (Van der Maaten e Hinton, 2008), che verrà nel seguito abbreviato in t-SNE. In Figura 3.10 si riportano i grafici t-SNE costruiti a partire dalle immagini log-normalizzate (grafico a sinistra) e dalle immagini allineate (grafico a destra). Si può notare come prima dell'allineamento ci sia una differenza, per quanto non molto marcata, tra i due soggetti che viene eliminata a seguito della rotazione Procuste.

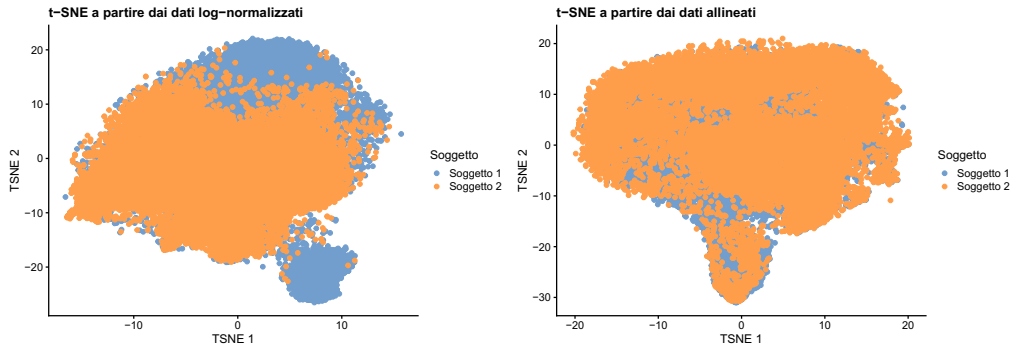


Figura 3.10: Rappresentazione *t*-SNE a partire dai dati log-normalizzati (grafico a sinistra) e dai dati ruotati tramite allineamento Procuste (grafico a destra).

3.3.2.1 Analisi differenziale

Venendo ora all'analisi differenziale, è stato deciso di adottare un approccio *pseudo-bulk*, conducendola quindi a livello di campione (cioè a livello di immagine) e non a livello di singolo *spot*. L'approccio *pseudo-bulk* necessita di due variabili: il campione, quindi in questo caso l'immagine, e un *cluster* di appartenenza. Tutti i dati di un campione in uno stesso *cluster* vengono sommati e fatti collassare in un'unica osservazione e l'analisi viene eseguita separatamente per ogni *cluster*. Si ottengono in questo modo matrici di dati simili ai dati *bulk* RNA-*Seq* introdotti nel Capitolo 1, con stesso numero di geni dei dati di partenza e numero di unità statistiche pari al numero di campioni. In questo caso quindi, avendo otto immagini e 1000 geni, si otterranno c matrici di dimensione 1000×8, dove c è il numero di *cluster* considerati.

L'analisi verrà ripetuta considerando tre diversi approcci. Inizialmente non verrà considerata alcuna partizione; tutti i dati di una stessa immagine verranno quindi sommati e si lavorerà con una sola matrice di dimensione 1000×8. Successivamente, dato che come detto nel Capitolo 1 per determinate patologie è di interesse trovare geni differenzialmente espressi in specifici strati corticali, verranno considerati i 6 strati più la materia bianca e si ripeterà quindi l'analisi separatamente su 7 matrici, una per strato,

di dimensione 1000×8. Infine gli *spot* verranno assegnati al tipo cellulare corrispondente tramite l'approccio `SingleR` (Aran et al., 2019) e si condurrà l'analisi separatamente per ogni tipo cellulare. L'approccio `SingleR` consiste nello scegliere un *dataset* di riferimento che contenga i dati di espressione genica di cellule derivanti dallo stesso tessuto del *dataset* da analizzare e utilizzare tale *dataset* già annotato per etichettare le nuove cellule. In particolare, viene calcolata la correlazione di Spearman tra ogni cellula del *dataset* da annotare e ogni cellula del *dataset* di riferimento, utilizzando solamente i geni marcatori derivanti da tutti i confronti a coppie tra i tipi cellulari del *dataset* di riferimento. Per ogni cellula da etichettare viene poi calcolato, per ogni etichetta, un quantile, generalmente il quantile 0.8, delle correlazioni tra tale cellula e tutte le cellule annotate con quell'etichetta. Infine, viene assegnato alla cellula il tipo cellulare per cui il quantile scelto delle correlazioni risulta più alto. Come *dataset* di riferimento è stato scelto il *dataset* `DarmanisBrainData` (Darmanis et al., 2015) della libreria `scRNAseq` (Risso e Cole, 2021), che contiene i dati di espressione genica di 466 cellule di cervelli adulti e fetali. Dato che gli individui del *dataset* da analizzare sono adulti, vengono eliminate dal *dataset* di riferimento le cellule fetali, ottenendo così un *dataset* di 332 cellule. I tipi cellulari presenti sono neuroni, astrociti, cellule endoteliali, cellule ibride, microglia, neuroni, oligodendrociti e cellule precursori degli oligodendrociti (OPC).

Dato che l'obiettivo è sempre il confronto con i dati non allineati, per ogni approccio verranno costruiti due insiemi di dati *pseudo-bulk*, derivanti o dai dati log-normalizzati o dai dati ruotati.

Un aspetto critico di questa analisi è stata la scelta del modello. I dati di RNA-*Seq* sono infatti dati di conteggio, e quindi molti degli approcci proposti per l'analisi di questo tipo di dati, come `edgeR` (Robinson, McCarthy e Smyth, 2010; McCarthy, Chen e Smyth, 2012) e `DESeq2` (Love, Huber e Anders, 2014), si basano su un modello lineare generalizzato binomiale negativo e inseriscono il fattore di normalizzazione come termine di *offset*. Applicando ai dati una rotazione, un modello per dati di conteggio non risulta

più appropriato. Un'alternativa, applicabile anche ai dati ruotati, consiste nel considerare dei modelli lineari. Modelli lineari per l'analisi di dati genomici sono implementati nella libreria `R limma` (Ritchie et al., 2015), inizialmente sviluppata per l'analisi di dati di *microarray* (Smyth, 2004) ma che funziona bene anche con dati di RNA-*Seq*. Sia $Y_j \in \mathbb{R}^{n\times 1}$ il vettore contenente i valori di espressione log-normalizzati del gene j negli n campioni. I modelli lineari implementati nella libreria `limma` assumono, per ogni gene j,

$$E[Y_j] = X\beta_j, \tag{3.1}$$

con $Var(Y_j) = \sigma_j^2$, X matrice del disegno e β_j vettore dei coefficienti. Dato che si stima un modello per ogni gene, l'idea alla base dei modelli della libreria `limma` consiste nel prendere in prestito l'informazione degli altri geni per la stima dei parametri relativi al gene j, sfruttando un approccio bayesiano empirico. Sia $V = X^\top X$. In particolare, si assume una distribuzione a priori normale per i coefficienti β_j differenzialmente espressi, $\beta_j|\beta_j \neq 0 \sim N(0, V_0\sigma_j^2)$ e una distribuzione a priori χ^2 per l'inverso delle varianze σ_j^2, $\frac{1}{\sigma_j^2} \sim \frac{1}{d_0 s_0^2}\chi^2_{d_0}$, dove s_0^2 è uno stimatore della varianza a priori con d_0 gradi di libertà. Una criticità nell'applicazione dei modelli lineari a dati di conteggio è che, anche a seguito di un'opportuna trasformazione, esibiscono una forte eteroschedasticità. Infatti, i geni con conteggi bassi esibiscono, oltre alla varianza biologica, una componente di varianza tecnica da non ignorare. È quindi di fondamentale importanza modellare la relazione media-varianza e inserirla nel modello. In questo elaborato si applicherà l'approccio *limma-trend* (Law et al., 2014; Phipson et al., 2016), che stima in maniera non parametrica tramite delle *spline* di regressione il trend media varianza per ogni gene e lo incorpora nel calcolo della media della varianza a posteriori, schiacciando le stime a posteriori verso la curva stimata. Per valutare la significatività dei coefficienti non verrà utilizzata la statistica t classica ma una sua versione moderata, pari a

$$t_{kj} = \frac{\hat{\beta}_{kj}}{\widetilde{s}_j\sqrt{v_k}},$$

dove v_k e il k-esimo elemento diagonale di $X^\top X$ e $\widetilde{s}_j$ è la media a posteriori della varianza σ_j^2.

Dato che in questa analisi è di interesse il confronto tra le immagini dei due individui, la matrice del disegno sarà pari a

$$X = \begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \end{bmatrix},$$

dove la prima colonna rappresenta l'intercetta e la seconda l'appartenenza al soggetto, e il vettore dei coefficienti sarà $\beta_j^\top = \begin{bmatrix} \beta_{0j} & \beta_{1j} \end{bmatrix}$. Si vuole testare l'ipotesi nulla $H_0 : \beta_{1j} = 0$ contro l'ipotesi alternativa $H_1 : \beta_{1j} \neq 0$, per $j = 1, \ldots, 1000$, ad un livello di significatività $\alpha = 0.05$. Si decide di controllare l'errore di primo tipo tramite il *False Discovery Rate*, fissando la proporzione di falsi positivi accettata al 5%.

Analisi con un solo gruppo

Si riportano ora i risultati derivanti dall'analisi che considera un unico gruppo. Si confrontano quindi i risultati derivanti dall'applicazione del modello *limma-trend* alla matrice 1000×8 di dati *pseudo-bulk* costruita a partire dai dati log-normalizzati e i risultati derivanti dall'applicazione dello stesso modello alla matrice 1000×8 di dati *pseudo-bulk* costruita a partire dai dati ruotati. Innanzitutto, a scopo esplorativo, sono state calcolate le componenti principali di entrambe le matrici e i dati sono stati rappresentati nello spazio delle prime due (Figura 3.11). Si può notare come, in entrambi i casi, la variabilità dei dati sia in gran parte spiegata dal soggetto a cui le immagini

appartengono. In particolare, per i dati non allineati (grafico a sinistra), l'effetto di soggetto si riflette nella prima componente principale, mentre per i dati allineati (grafico a destra) l'effetto di soggetto è riassunto dalla seconda componente principale.

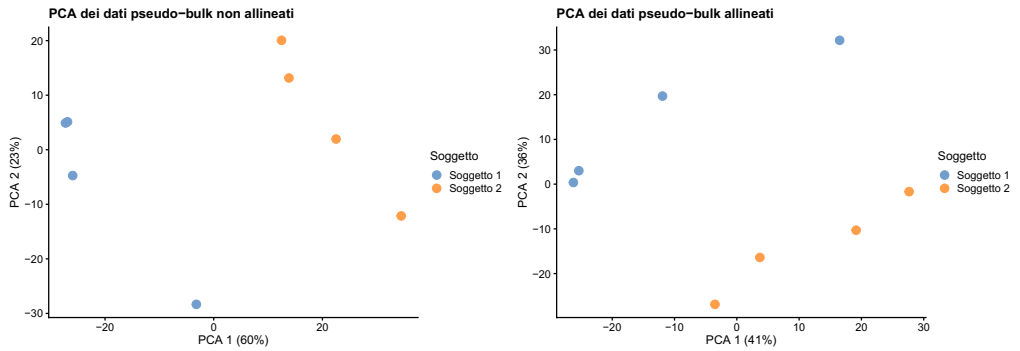


Figura 3.11: Dati *pseudo-bulk* rappresentati nello spazio delle prime due componenti principali ottenute a partire dai dati non allineati (grafico a sinistra) e a partire dai dati allineati (grafico a destra). I punti sono colorati per soggetto.

Il modello applicato ai dati non allineati rileva in tutto 561 geni differenzialmente espressi, mentre il modello applicato ai dati allineati ne rileva 333. Come ci si aspettava quindi, il numero di geni differenzialmente espressi cala se si considerano i dati allineati, quindi si conferma che, almeno in parte, la rotazione Procuste assorbe l'effetto di soggetto rendendo confrontabili dati derivanti da campioni diversi. Inoltre, guardando la tabella a doppia entrata (Tabella 3.6), si nota che dei 331 geni differenzialmente espressi identificati lavorando con i dati allineati, 242 erano stati identificati anche dal modello applicato ai dati non allineati.

Tabella 3.6: Tabella a due entrate per il numero di geni differenzialmente espressi con i due approcci.

	Immagini non allineate		
Immagini allineate		$p < 0.05$	$p > 0.05$	Totale
	$p < 0.05$	242	91	333
	$p > 0.05$	319	348	667
	Totale	561	439	1000

Si rimanda all'appendice per i grafici con i trend media-varianza stimati dai modelli (Figura A.6) e per gli istogramma dei *p-value* ottenuti (Figura A.7), che come ci si aspettava presentano un picco in corrispondenza dello zero.

Approccio con un *cluster* per strato

Si riportano in Tabella 3.7 il numero di geni differenzialmente espressi identificati in ogni strato per il modello applicato ai dati allineati e per il modello applicato ai dati non allineati. Si può notare come con le immagini allineate il numero di geni differenzialmente espressi rilevati cali drasticamente. È interessante la differenza nello strato 1, infatti il modello applicato ai dati *pseudo-bulk* derivanti dai dati log-normalizzati identifica 953 geni differenzialmente espressi, numero molto alto soprattutto considerato che lo strato 1 è composto da pochi corpi cellulari, mentre il modello applicato ai dati *pseudo-bulk* derivanti dai dati non allineati rileva solamente un gene differenzialmente espresso. Rimane elevato il numero di geni differenzialmente espressi nella materia bianca.

Tabella 3.7: Numero di geni differenzialmente espressi in ogni strato.

Strato	Immagini non allineate	Immagini allineate
Strato I	953	1
Strato II	15	3
Strato III	747	313
Strato IV	413	128
Strato V	60	119
Strato VI	531	561
Materia bianca	879	834

Osservando il grafico t-SNE in Figura 3.12 nel quale si proiettano solamente i punti di materia bianca, si può osservare che, sia con i dati log-normalizzati sia con i dati allineati, i due soggetti sono separati. C'è quindi un'effettiva differenza biologica nella materia bianca dei due soggetti.

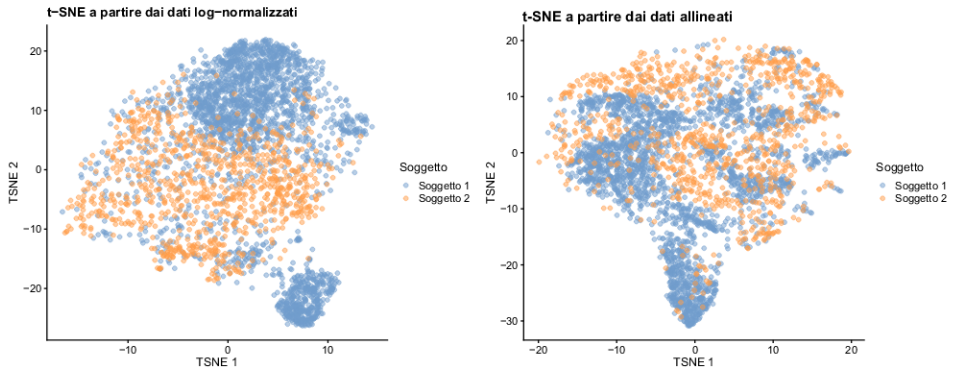


Figura 3.12: Rappresentazione t-SNE dei dati log-normalizzati (grafico a sinistra) e dei dati allineati (grafico a destra) dei punti di materia bianca.

Approccio con un *cluster* per tipo cellulare

Si procede ora all'analisi per tipo cellulare. Si riportano rispettivamente in

Tabella 3.8 e in Tabella 3.9 le tabelle di frequenza dei tipi cellulari ottenuti dall'applicazione dell'algoritmo `SingleR` ai dati log-normalizzati e ai dati ruotati. Le due classificazioni sono abbastanza concordi, infatti l'indice di Rand aggiustato è pari a 0.53.

Tabella 3.8: Tabella di frequenza dei gruppi cellulari assegnati tramite approccio `SingleR` applicato ai dati log-normalizzati.

Astrociti	Cellule endoteliali	Cellule ibride	Microglia	Neuroni	Oligodendrociti	OPC
546	209	444	63	29123	2005	7

Tabella 3.9: Tabella di frequenza dei gruppi cellulari assegnati tramite approccio `SingleR` applicato ai dati ruotati.

Astrociti	Cellule endoteliali	Cellule ibride	Microglia	Neuroni	Oligodendrociti	OPC
192	140	359	47	30013	1628	18

Anche in questo caso, si ha che l'applicazione del modello ai dati allineati fa sì che il numero di geni differenzialmente espressi rilevati cali per ogni tipo cellulare. Rimane alto il numero di geni differenzialmente espressi per gli oligodendrociti, risultato coerente con l'analisi svolta al paragrafo precedente in quanto gli oligodendrociti sono le cellule di cui è composta la materia bianca. Si può infatti notare dalle *heatmap* in Figura 3.13, che rappresentano il confronto tra gli strati corticali e i tipi cellulari per i dati log-normalizzati (grafico a sinistra) e per i dati allineati (grafico a destra), che c'è una forte corrispondenza tra le cellule etichettate come oligodendrociti e la materia bianca.

Tabella 3.10: Numero di geni differenzialmente espressi tra i due soggetti per ogni tipo cellulare.

Tipo cellulare	Immagini non allineate	Immagini allineate
Astrociti	512	0
Neuroni	539	340
Cellule endoteliali	0	0
Microglia	0	0
Cellule ibride	286	0
OPC	0	0
Oligodendrociti	982	695

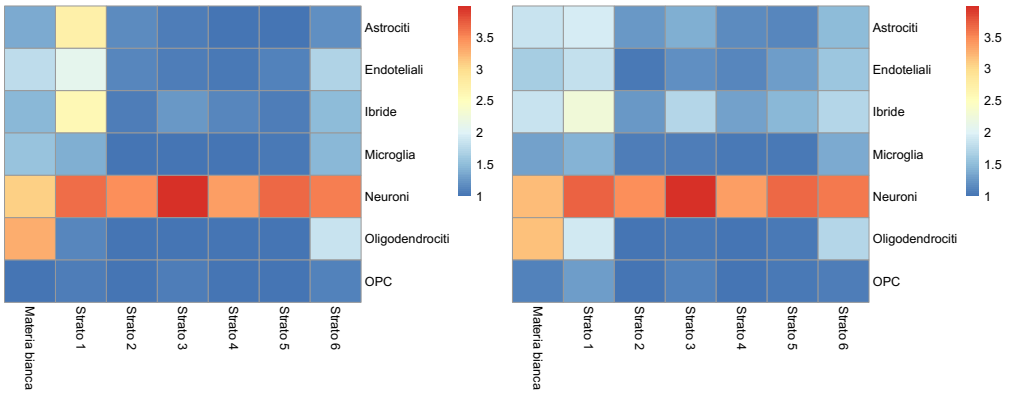


Figura 3.13: Confronto tra gli strati e i tipi cellulari ottenuti dai dati non allineati (grafico a destra) e dai dati allineati (grafico a sinistra).

In conclusione quindi, l'allineamento assorbe una quota di variabilità tra i due gruppi, imputabile al mancato allineamento. Rimane, in tutte le analisi svolte, una differenza biologica, particolarmente rilevante nella materia bianca.

Una difficoltà di questa analisi è che, essendo dati reali, non si può sapere quali geni siano effettivamente differenzialmente espressi e quali lo risultino a causa di falsi positivi. Inoltre, dato che i due soggetti sono entrambi sani, non si possono sfruttare ontologie di geni per vedere se i geni differenzialmente espressi hanno un senso biologico. Sarebbe necessario condurre uno studio di simulazione per valutare al meglio l'effetto dell'allineamento Procuste, tenendo presente che simulare dati realistici in questo contesto è complicato.

Conclusioni

In conclusione sono stati per la prima volta applicati algoritmi per l'allineamento Procuste a dati di trascrittomica spaziale della corteccia cerebrale umana e si sono ottenuti risultati soddisfacenti.

Ripercorrendo in ordine le analisi effettuate, sono state inizialmente allineate due sole immagini per sfruttare la soluzione esplicita proposta da Green (1952) ed è stato verificato che fossero effettivamente più omogenee. Per fare ciò sono stati applicati algoritmi di *clustering* separatamente sulle due immagini non allineate e sulle due immagini allineate e si è valutata la concordanza tra le partizioni ottenute tramite l'indice di Rand aggiustato, considerando i punti nelle stesse coordinate come stesse unità statistiche. Sono stati applicati diversi algoritmi di *clustering* che permettessero di controllare il numero finale di *cluster* desiderato, che è stato posto pari ad 8 seguendo quanto fatto dagli autori dell'articolo di riferimento (Maynard et al., 2021). È risultato evidente come l'allineamento rendesse le partizioni ottenute sulle due immagini più simili, infatti l'indice di Rand è sempre superiore se calcolato post rotazione Procuste. Inoltre, tra gli algoritmi considerati, il *walktrap* ha dato risultati migliori, producendo un incremento maggiore.

Sono state successivamente allineate più immagini utilizzando, per motivi computazionali e per garantire l'unicità della soluzione, il modello ProMises Efficiente, con l'obiettivo di dimostrare che l'allineamento Procuste rende le diverse immagini più omogenee. A questo scopo sono state allineate 4 immagini dello stesso individuo e 3 immagini di individui diversi. Le immagini di individui diversi erano in origine più diverse rispetto alle quattro dello stesso

individuo. A termine dell'allineamento Procuste, sono state calcolate le distanze tra le immagini e si è osservato come fossero di molto inferiori rispetto a quelle calcolate tra le immagini prima dell'allineamento. È stata poi calcolata, sia per il caso con quattro immagini che per il caso con tre, la matrice media delle immagini prima e dopo l'allineamento. Se le immagini non sono allineate le coordinate dei punti non sono veritiere. In questo caso quindi ogni punto della matrice media deriva dalla media di punti incorrelati e di conseguenza essa è di fatto una matrice casuale; al contrario, dopo l'allineamento, punti nelle stesse coordinate sono effettivamente punti corrispondenti e si calcola quindi la media di oggetti omogenei. Sono stati applicati algoritmi di *clustering* e si è valutata l'omogeneità dei *cluster* ottenuti tramite la *silhouette*. Si è deciso di non imporre a priori un numero specifico di *cluster* e sono quindi stati applicati due algoritmi basati sulle reti, l'algoritmo di Louvain e il *walktrap*. Per entrambi gli algoritmi si è osservata la *silhouette* media globale sia per l'immagine media delle quattro immagini dello stesso individuo che per l'immagine media delle tre immagini di individui diversi prima e dopo l'allineamento al variare dell'iperparametro che definisce il numero di vicini più vicini condivisi. In tutti i casi si nota che la *silhouette* media globale massima calcolata sulle matrici medie delle immagini allineate è superiore rispetto alla stessa calcolata sulla matrice media delle immagini non allineate. I risultati sono più evidenti nell'applicazione alla media di tre immagini di individui diversi, probabilmente perché le quattro immagini dello stesso individuo erano in origine già allineate.

Una volta appurato che l'allineamento rende le immagini effettivamente più simili, se ne è voluto investigare l'effetto da un punto di vista biologico. Seguendo quanto fatto nell'articolo di riferimento, si è inizialmente pensato di sfruttare l'informazione relativa agli strati. Un obiettivo dell'articolo di riferimento infatti è stato ottenere *cluster* che massimizzassero la concordanza con le etichette degli strati in maniera da poter sviluppare uno strumento che permettesse di ottenere la classificazione laminare senza dover etichettare manualmente tutti i punti. Non hanno però ottenuto risultati soddisfacenti,

infatti l'indice di Rand medio per la concordanza tra i *cluster* da loro ottenuti e gli strati manualmente assegnati è pari a 0.2. Dato che allineando due immagini una funge da riferimento, si sono prese le etichette degli strati dell'immagine di riferimento e sono state confrontate con i *cluster* ottenuti con i vari algoritmi. La concordanza, sempre misurata tramite indice di Rand aggiustato, è rimasta costante e confrontabile con i valori ottenuti da Maynard et al. (2021). Inoltre, è stata valutata la qualità della partizione ottenuta tramite il calcolo della *silhouette* ed è stato notato che molti punti hanno *silhouette* negativa. Si conclude quindi che il segnale esibito dai dati non rispecchia interamente l'appartenenza allo strato corticale e si è deciso di abbandonare questa strada.

Si è infine deciso di analizzare l'effetto dell'allineamento Procuste da un punto di vista inferenziale. L'esperimento ideale avrebbe richiesto di avere dati di diversi individui in diverse condizioni biologiche al fine di vedere l'effetto dell'allineamento su entrambe le variabili. Si vuole infatti che renda confrontabili le immagini di diversi individui, assorbendo quindi l'effetto di soggetto ma lasciando invariata la quota di variabilità dovuta alle differenze biologiche. Il *dataset* analizzato in questa tesi è però uno dei primi *dataset* ottenuti con la tecnologia *Visium* descritta nel Capitolo 1 e presenta solamente dati relativi a soggetti sani. Si è quindi pensato di condurre un esperimento nullo allineando otto immagini, quattro di un individuo e quattro di un altro, e applicare modelli di analisi differenziale. Dato che non sussistono differenze biologiche tra gli individui presenti in questo *dataset*, si può assumere che i geni differenzialmente espressi che emergono dall'analisi differenziale siano per la maggior parte falsi positivi. È stato adottato un approccio *pseudo-bulk* ed è stato applicato il modello *limma-trend*, confrontando i risultati ottenuti sulle immagini allineate con quelli derivanti dall'applicazione della stessa *pipeline* sulle immagini non allineate ma log-normalizzate con metodi standard. Si sono effettivamente osservate meno differenze tra i dati allineati. Inoltre, in questa applicazione, è stato per la prima volta applicato il modello ProMises Efficiente a matrici con diverso numero di colonne.

Nonostante i buoni risultati ottenuti, questo elaborato presenta alcune limitazioni. Innanzitutto, come già detto, sarebbe interessante ripetere l'analisi su dati di diversi individui in diverse condizioni biologiche. A questo scopo sarebbe inoltre interessante proporre un'estensione del modello ProMises che permetta maggiore variabilità. Ci sono due possibili strade per raggiungere questo obiettivo: definire una diversa matrice di riferimento per ogni sotto-popolazione, riformulando quindi il modello ProMises come una mistura, oppure ipotizzare diversa informazione a priori per ogni condizione biologica specificando diversi parametri di posizione.

Anche dal punto di vista della modellazione biologica si può pensare a dei miglioramenti. I modelli fino ad oggi proposti per l'analisi dei dati di espressione genica si basano infatti o sulla loro natura di conteggio o sulle loro caratteristiche a seguito di una normalizzazione campione-specifica. Sarebbe interessante studiare più approfonditamente le caratteristiche a seguito della rotazione Procuste e modificare i modelli già esistenti perché le rispettino al meglio. Si potrebbero per esempio applicare modelli ad effetti casuali, inserendo un effetto casuale per individuo.

Appendice A

Figure supplementari

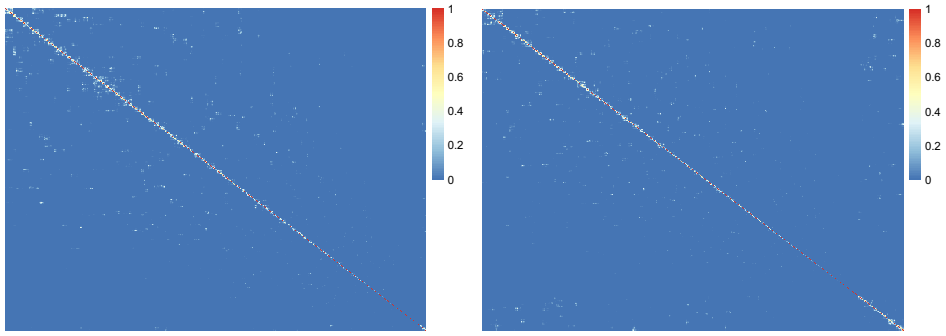


Figura A.1: Parametro di posizione F^* della distribuzione a priori degli R_i^* per l'allineamento delle quattro immagini dello stesso individuo (grafico a sinistra) e delle tre immagini di soggetti diversi (grafico a destra).

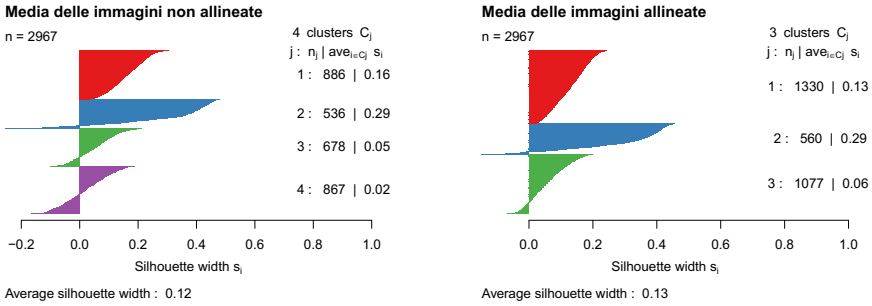


Figura A.2: Silhouette degli *spot* a seguito dell'applicazione dell'algoritmo di Louvain alla media delle quattro immagini dello stesso individuo prima e dopo l'allineamento.

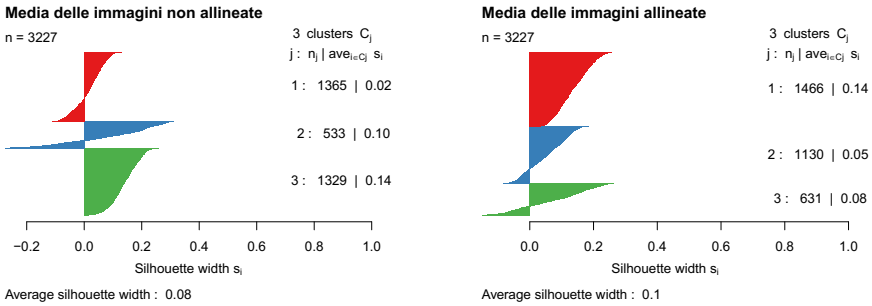


Figura A.3: Silhouette degli *spot* a seguito dell'applicazione dell'algoritmo di Louvain alla media delle tre immagini di individui diversi prima e dopo l'allineamento.

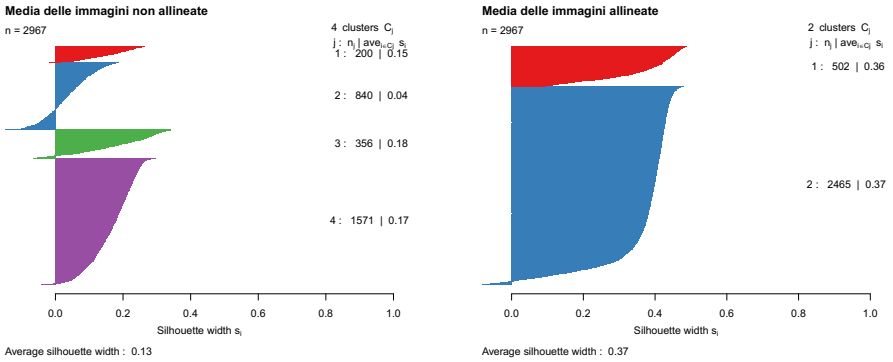


Figura A.4: Silhouette degli *spot* a seguito dell'applicazione dell'algoritmo walktrap alla media delle quattro immagini dello stesso individuo prima e dopo l'allineamento.

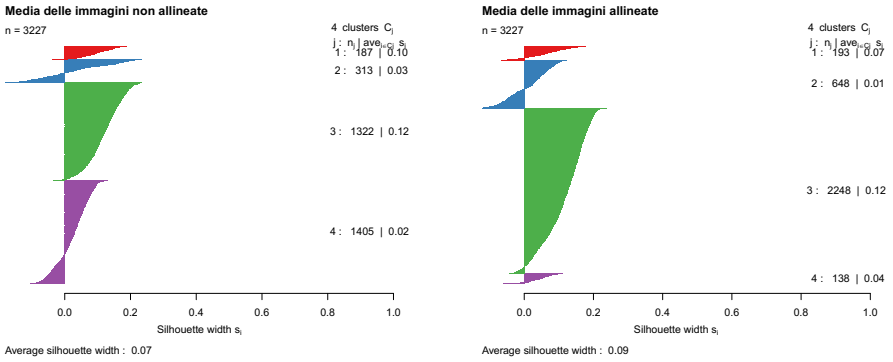


Figura A.5: Silhouette degli *spot* a seguito dell'applicazione dell'algoritmo walktrap alla media delle tre immagini di individui diversi prima e dopo l'allineamento.

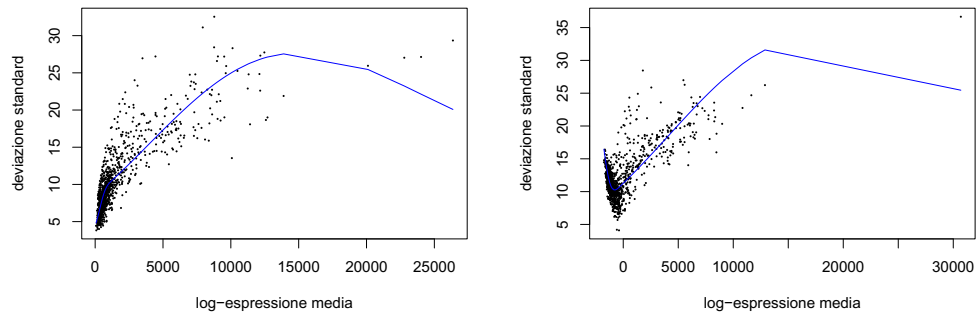


Figura A.6: Trend media-varianza stimato dal modello *limma-trend* applicato ai dati non allineati (grafico a sinistra) e ai dati allineati (grafico a destra).

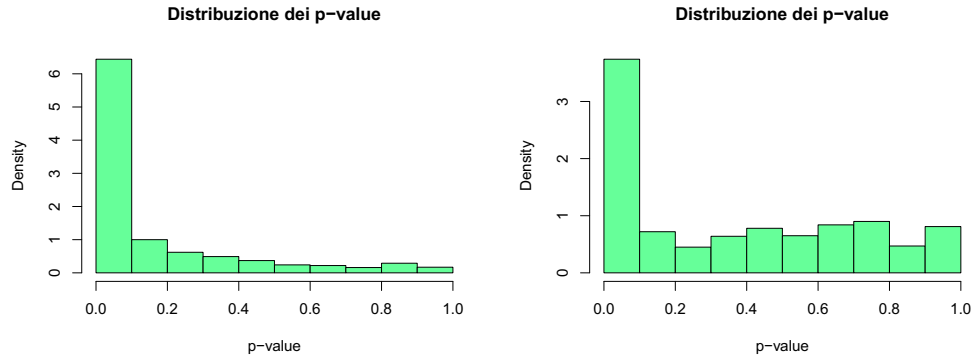


Figura A.7: Istogramma dei *p-value* ottenuti per i dati non allineati (grafico a sinistra) e per i dati allineati (grafico a destra).

Bibliografia

Andreella, Angela (2021). *Challenges in functional magnetic resonance imaging (fmri) data analysis: from functional alignment to selective inference*. Tesi di dottorato, Universitá degli Studi di Padova.

Andreella, Angela e Livio Finos (2022). "Procrustes analysis for high-dimensional data". *ArXiv preprint arXiv:2008.04631v4*.

Aran, Dvir, Agnieszka P. Looney, Leqian Liu, Esther Wu, Valerie Fong, Austin Hsu, Suzanna Chak, Ram P. Naikawadi, Paul J. Wolters, Adam R. Abate, Atul J. Butte e Mallar Bhattacharya (2019). "Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage". *Nat. immunol.* 20, 163–172.

Asp, Michaela, Joseph Bergenstråhle e Joakim Lundeberg (2020). "Spatially resolved transcriptomes—next generation tools for tissue exploration". *BioEssays* 42(10), 1900221.

Bai, Zhaojun, James Demmel, Jack Dongarra, Axel Ruhe e Henk van der Vorst (2000). *Templates for the solution of algebraic eigenvalue problems: a practical guide*. SIAM.

Bakken, Trygve E, Rebecca D Hodge, Jeremy A Miller, Zizhen Yao, Thuc Nghi Nguyen, Brian Aevermann, Eliza Barkan, Darren Bertagnolli, Tamara Casper, Nick Dee et al. (2018). "Single-nucleus and single-cell transcriptomes compared in matched cortical cell types". *Plos One* 13(12), e0209648.

Barndorff Nielsen, Ole (1973). *Exponential families and conditioning*. Tesi di dottorato, Universitá di Copenhagen.

Berge, Jos MF (1977). "Orthogonal Procrustes rotation for two or more matrices". *Psychometrika* 42(2), 267–276.

Blondel, Vincent D, Jean-Loup Guillaume, Renaud Lambiotte e Etienne Lefebvre (2008). "Fast unfolding of communities in large networks". *Journal of Statistical Mechanics: Theory and Experiment* 2008(10), P10008.

Crick, Francis (1970). "Central dogma of molecular biology". *Nature* 227, 561–563.

Darmanis, Spyros, Steven A Sloan, Ye Zhang, Martin Enge, Christine Caneda, Lawrence M Shuer, Melanie G Hayden Gephart, Ben A Barres e Stephen R Quake (2015). "A survey of human brain transcriptome diversity at the single cell level". *Proceedings of the national academy of sciences* 112(23), 7285–7290.

Downs, Thomas D (1972). "Orientation statistics". *Biometrika* 59(3), 665–676.

Goodall, Colin (1991). "Procrustes methods in the statistical analysis of shape". *Journal of the Royal Statistical Society: Series B (Methodological)* 53(2), 285–321.

Gower, John C (1975). "Generalized Procrustes Analysis". *Psychometrika* 40(1), 33–51.

Gower, John C, Garmt B Dijksterhuis et al. (2004). *Procrustes problems*. Vol. 30. Oxford University Press on Demand.

Green, B. (1952). "The orthogonal approximation of an oblique structure in factor analysis". *Psychometrika* 17, 429–440.

Green, Peter J e Kanti V Mardia (2006). "Bayesian alignment using hierarchical models, with applications in protein bioinformatics". *Biometrika* 93(2), 235–254.

Hartigan, John A e Manchek A Wong (1979). "Algorithm as 136: a k-means clustering algorithm". *Journal of the Royal Statistical Society. Series c (Applied Statistics)* 28(1), 100–108.

Hubert, Lawrence e Phipps Arabie (1985). "Comparing partitions". *Journal of Classification* 2(1), 193–218.

Jupp, Peter E e Kanti V Mardia (1979). "Maximum likelihood estimators for the matrix von mises-fisher and bingham distributions". *The Annals of Statistics* 7(3), 599–606.

Kaufman, Leonard e Peter J Rousseeuw (1990). *Partitioning around medoids (program pam)*. John Wiley & Sons, Ltd. Cap. 2, 68–125.

Law, Charity W, Yunshun Chen, Wei Shi e Gordon K Smyth (2014). "Voom: precision weights unlock linear model analysis tools for rna-seq read counts". *Genome Biology* 15(2), 1–17.

Love, Michael I., Wolfgang Huber e Simon Anders (2014). "Moderated estimation of fold change and dispersion for rna-seq data with deseq2". *Genome Biology* 15 (12), 550.

Maynard, Kristen R, Leonardo Collado-Torres, Lukas M Weber, Cedric Uytingco, Brianna K Barry, Stephen R Williams, Joseph L Catallini, Matthew N Tran, Zachary Besich, Madhavi Tippani et al. (2021). "Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex". *Nature neuroscience* 24(3), 425–436.

McCarthy, Davis J., Kieran R. Campbell, Aaron T. L. Lun e Quin F. Willis (2017). "Scater: pre-processing, quality control, normalisation and visualisation of single-cell RNA-seq data in R". *Bioinformatics* 33 (8), 1179–1186.

McCarthy, Davis J, Yunshun Chen e Gordon K Smyth (2012). "Differential expression analysis of multifactor rna-seq experiments with respect to biological variation". *Nucleic Acids Research* 40(10), 4288–4297.

Myronenko, Andriy e Xubo Song (2009). "On the closed-form solution of the rotation matrix arising in computer vision problems". *ArXiv preprint arXiv:0904.1613*.

Phipson, Belinda, Stanley Lee, Ian J Majewski, Warren S Alexander e Gordon K Smyth (2016). "Robust hyperparameter estimation protects against hypervariable genes and improves power to detect differential expression". *The Annals of Applied Statistics* 10(2), 946.

Pons, Pascal e Matthieu Latapy (2005). "Computing communities in large networks using random walks". *International symposium on computer and information sciences*. Springer, 284–293.

Risso, Davide e Michael Cole (2021). *Scrnaseq: collection of public single-cell rna-seq datasets*. R package version 2.6.1.

Ritchie, Matthew E, Belinda Phipson, Di Wu, Yifang Hu, Charity W Law, Wei Shi e Gordon K Smyth (2015). "limma powers differential expression analyses for RNA-sequencing and microarray studies". *Nucleic Acids Research* 43(7), e47.

Robinson, Mark D, Davis J McCarthy e Gordon K Smyth (2010). "Edger: a bioconductor package for differential expression analysis of digital gene expression data". *Bioinformatics* 26(1), 139–140.

Rousseeuw, Peter J (1987). "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis". *Journal of Computational and Applied Mathematics* 20, 53–65.

Schönemann, Peter H (1966). "A generalized solution of the orthogonal Procrustes problem". *Psychometrika* 31(1), 1–10.

Skene, Nathan G, Julien Bryois, Trygve E Bakken, Gerome Breen, James J Crowley, Héléna A Gaspar, Paola Giusti-Rodriguez, Rebecca D Hodge, Jeremy A Miller, Ana B Muñoz-Manchado et al. (2018). "Genetic identification of brain cell types underlying schizophrenia". *Nature Genetics* 50(6), 825–833.

Smyth, Gordon K (2004). "Linear models and empirical bayes methods for assessing differential expression in microarray experiments". *Statistical Applications in Genetics and Molecular Biology* 3(1).

Ståhl, Patrik L, Fredrik Salmén, Sanja Vickovic, Anna Lundmark, José Fernández Navarro, Jens Magnusson, Stefania Giacomello, Michaela Asp, Jakub O Westholm, Mikael Huss et al. (2016). "Visualization and analysis of gene expression in tissue sections by spatial transcriptomics". *Science* 353(6294), 78–82.

Sweet, Robert, Kenneth Fish e David Lewis (2010). "Mapping synaptic pathology within cerebral cortical circuits in subjects with schizophrenia". *Frontiers in Human Neuroscience* 4, 44.

Trendafilov, Nickolay T e Ross A Lippert (2002). "The multimode procrustes problem". *Linear Algebra and its Applications* 349(1-3), 245–264.

Van der Maaten, Laurens e Geoffrey Hinton (2008). "Visualizing data using t-sne." *Journal of Machine Learning Research* 9(11).

Velmeshev, Dmitry, Lucas Schirmer, Diane Jung, Maximilian Haeussler, Yonatan Perez, Simone Mayer, Aparna Bhaduri, Nitasha Goyal, David H. Rowitch e Arnold R. Kriegstein (2019). "Single-cell genomics identifies cell type–specific molecular changes in autism". *Science* 364(6441), 685–689.