Procrustes analysis for spatial transcriptomics data

Daniela Corbetta[1],

Angela Andreella[2], Livio Finos[1], Davide Risso[1]

[1]Department of Statistical Sciences, University of Padova
[2]Department of Economics, Ca' Foscari University of Venice

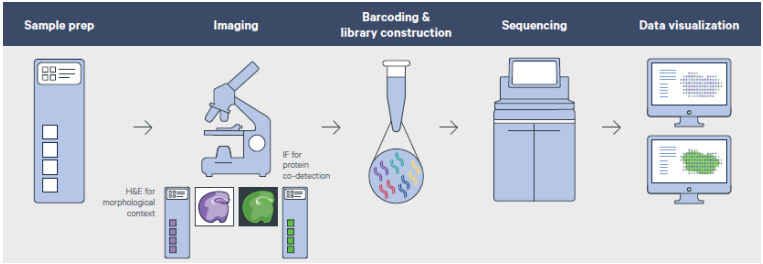
30 August, 2023

Introduction

- Genetic studies on the brain are fundamental for the study of neuropsychiatric disorders;
- The cerebral cortex has a **layered structure** divided into six layers plus a layer of white matter $\rightarrow$ alterations in gene expression have been found in specific cortical layers for certain pathologies $\rightarrow$ importance of spatial localization;
- Conventional DNA sequencing technologies do not provide the coordinates of cells $\rightarrow$ **spatial transcriptomics**.

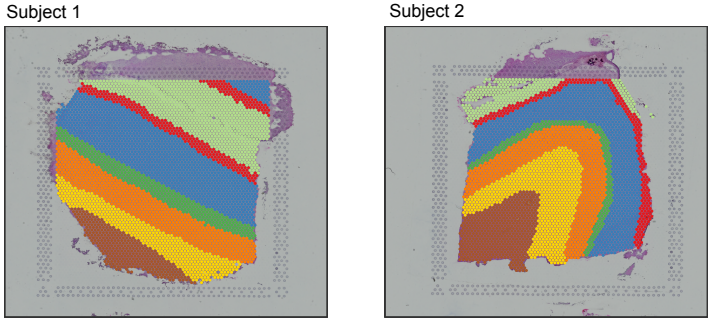
Spatial Transcriptomics

- Ståhl's approach (Ståhl et al., 2016): transcripts captured *in situ* and then sequenced *ex situ*.
- For each analyzed sample, two data matrices are obtained: one with gene expression counts and one with the two-dimensional coordinates of the sequenced spots.



Limitations

- Brains of different individuals are not functionally aligned;
- Variability of data from different individuals: **biological variability** (of interest) and **technical variability** due to misalignment (disturbance) $\rightarrow$ functional alignment algorithms.



Procrustes Analysis

- **Objective**: Match different matrices through similarities (rotations, reflections, translations, and scaling).
- Two matrices only, $X_1, X_2 \in \mathbb{R}^{n \times m}$:
  - Objective function: $R = \text{argmin}_R \|X_1 - X_2 R^\top\|_F^2$, subject to the constraint $R^\top R = I_m$.
  - **Closed-form solution**: $\hat{R} = UV^\top$, with $UDV^\top$ being the singular value decomposition of $X_1^\top X_2$.
- More than two matrices: iterative algorithms $\rightarrow$ **Efficient ProMises model** (Andreella & Finos, 2022)

Efficient ProMises Model

Let $X_1 \in \mathbb{R}^{n \times m_1}, \ldots, X_N \in \mathbb{R}^{n \times m_N}$ be the matrices to be aligned. The Efficient ProMises model assumes

$$X_i Q_i = (M + E_i) R_i^\top,$$

where

- $Q_i \in \mathbb{R}^{m_i \times n}$ is a semi-orthogonal transformation obtained from the thin SVD (Bai et al., 2000) of X_i,
- $E_i \sim \mathcal{MN}_{n,n}(0, \sigma^2 I_n, I_n)$,
- R_i is an orthogonal rotation/reflection parameter,
- $M \in \mathbb{R}^{n \times n}$ is the common space to which the matrices are aligned.

Comments on R_i

Let $X_i Q_i = X_i^*$.

- R_i follows a **von Mises-Fisher prior** distribution with a location parameter $F \in \mathbb{R}^{n \times n}$ and concentration parameter k:

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

- **conjugate distribution** for the matrix normal distribution
- The posterior distribution of R_i is still a **von Mises-Fisher** distribution with location parameter $X_i^{*\top} M + kF \rightarrow$ weighted average between the maximum likelihood estimator and the prior mode
- estimation of R_i: **maximum a posteriori** $\rightarrow \hat{R}_i = U_i V_i^\top$, with $U_i D_i V_i^\top$ being the SVD of $X_i^{*\top} M + kF$

Algorithm

1. Dimensionality reduction: $X_i^* = X_i Q_i \in \mathbb{R}^{n \times n}$, with Q_i semi-orthogonal transformation.
2. $\hat{M} = \sum_{i=1}^{N} X_i^* / N$
3. For $i = 1, \ldots, N$:
   1. SVD$(X_i^{*\top} \hat{M} + kF) = U_i D_i V_i^\top$
   2. $\hat{R}_i = U_i V_i^\top$
   3. $\hat{X}_i^* = X_i^* \hat{R}_i$
4. Update $\hat{M} = \sum_{i=1}^{N} \hat{X}_i^* / N$ and evaluate convergence. If convergence is reached, proceed to step 5; otherwise, go back to step 3.
5. Projection onto the original space: $\hat{X}_i = \hat{X}_i^* Q_i^\top$

Comments on the Model

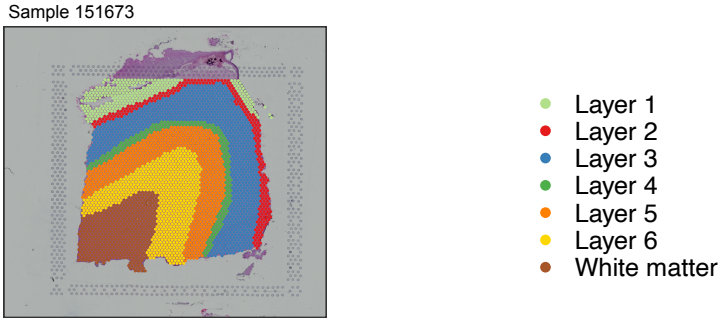
- The model leads to a **unique solution**, which is crucial in applications where matrices have a spatial interpretation.
- The parameter F allows to incorporate information regarding the **spatial organization** of the data.
- The model can be applied to matrices with different numbers of columns.

Application to real data

- **Objective:** show that the ProMises transformation absorbs the technical variability due to the misalignment.
- Comparison of results of differential analysis applied to aligned data and to log-normalized data with standard methods.

Data

- Sections of tissue from the dorsolateral prefrontal cortex of **three healthy** adult subjects (Maynard et al., 2021)
- Four images per subject $\rightarrow$ **12 images in total**
- Approximately **4000 spots** sequenced per image
- Annotation of the corresponding cortical layer for each spot.



Differential Analysis

Alignment of 8 images, 4 from one individual and 4 from another.
Pseudo-bulk approach with three cases:

- Single cluster
- One cluster per layer
- One cluster per cell type

Differential analysis using the **limma-trend** model (Law et al., 2014)

- Empirical Bayes linear model
- Response variable: log-normalized or rotated expression level
- Explanatory variable: subject

The Procrustes rotation is expected to absorb the variability due to misalignment, resulting in fewer differentially expressed genes.

Single Cluster

- Two-entry table showing the number of differentially expressed genes obtained from rotated and log-normalized images
- Type I error controlled with the **false discovery rate**, allowing for a 5% proportion of false positives

	Non aligned images		
Aligned images		$p < 0.05$	$p > 0.05$	Total
	$p < 0.05$	242	91	333
	$p > 0.05$	319	348	667
	Total	561	439	1000

One cluster per layer

- Number of differentially expressed genes between the two subjects in each layer.
- Fewer differences in the upper layers.
- Differences still present in the innermost layers and white matter.

Layer	Non aligned images	Aligned images
Layer 1	953	1
Layer 2	15	3
Layer 3	747	313
Layer 4	413	128
Layer 5	60	119
Layer 6	531	561
White matter	879	834

One cluster for each cell type

- Number of differentially expressed genes between the two subjects for each cell type.
- Differences still present for neurons and oligodendrocytes.

Cell type	Non aligned images	Aligned images
Astrocytes	512	0
Neurons	539	340
Endothelial cells	0	0
Microglia	0	0
Hybrid cells	286	0
OPC	0	0
Oligodendrocytes	982	695

Conclusions and further research directions

- ProMises transformation **absorbs a portion of the variability** attributed to misalignment. The number of differentially expressed genes is consistently lower when applying differential expression models to aligned data.
- The analyzed data are **real data**, and the actual differentially expressed genes cannot be determined. It would be necessary to conduct **simulation studies**.
- The model used for differential analysis was proposed to address the characteristics of log-normalized counts with a cell-specific normalization factor.
- It would be interesting to have data from individuals in **different biological conditions**.

References

Andreella, A. & Finos, L. (2022). Procrustes analysis for high-dimensional data. *psychometrika*, 87(4), 1422–1438.

Bai, Z., Demmel, J., Dongarra, J., Ruhe, A., & van der Vorst, H. (2000). *Templates for the solution of algebraic eigenvalue problems: a practical guide*. SIAM.

Law, C. W., Chen, Y., Shi, W., & Smyth, G. K. (2014). voom: Precision weights unlock linear model analysis tools for rna-seq read counts. *Genome biology*, 15(2), 1–17.

Maynard, K. R., Collado-Torres, L., Weber, L. M., Uytingco, C., Barry, B. K., Williams, S. R., Catallini, J. L., Tran, M. N., Besich, Z., Tippani, M., et al. (2021). Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex. *Nature neuroscience*, 24(3), 425–436.

Ståhl, P. L., Salmén, F., Vickovic, S., Lundmark, A., Navarro, J. F., Magnusson, J., Giacomello, S., Asp, M., Westholm, J. O., Huss, M., et al. (2016). Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science*, 353(6294), 78–82.