Procrustes analysis for high dimensional data: `alignProMises`

Daniela Corbetta, Angela Andreella, Livio Finos, Davide Risso

daniela.corbetta@studenti.unipd.it

Introduction

- **Spatial transcriptomics** data provide both **genomic and spatial information**
- **The structure of the brains differs between subjects** → Brains of different subjects cannot be compared since they are not aligned

Aim of the analysis: rotate brains of different subjects to absorb the unwanted variability caused by the misalignment

Alignment methods based on Procrustes theory: a statistical shape analysis that aligns matrices using **similarity** transformations

- **2 matrices → explicit solution**: $\hat{X}_1 = X_1 \hat{R}$, where $\hat{R} = UV^\top$ and U and V derive from the **SVD** of $(X_1^\top X_2)$
- **More than two matrices → iterative algorithms**: Andreella and Finos (2022) proposed the **ProMises model** and the **Efficient ProMises model**

Data

Let $X_i \in \mathbb{R}^{n \times m}$, where $i = 1, \ldots, N$ represents the sample, m is the total number of spots and n is the total number of genes (Maynard *et al.*, 2021):

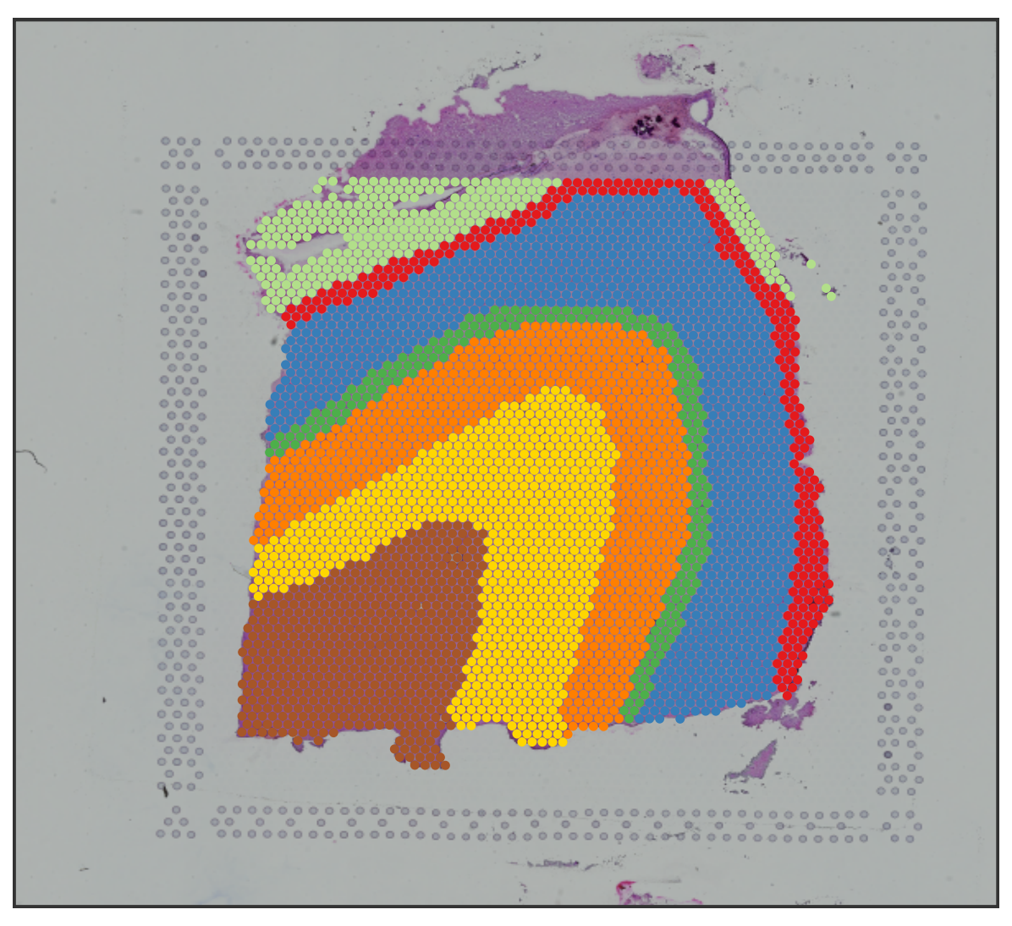
- 3 subjects, 4 samples per subject: 12 samples ($i = 1, \ldots, 12$)
- 4000 spots per image (m=4000)
- 1000 genes (n=1000)

Genomic counts

	Spot 1	Spot 2	…	Spot m
Gene 1	y_{11}	y_{21}	…	y_{m1}
Gene 2	y_{12}	y_{22}	…	y_{m2}
⋮	⋮	⋮	⋱	⋮
Gene n	y_{1n}	y_{2n}	…	y_{mn}

Coordinates data

	Spot 1	Spot 2	…	Spot m
coord x	x_1	x_2	…	x_m
coord y	y_1	y_2	…	y_m



ProMises model

Every X_i is the **rotation of a common reference matrix plus an error term**:

$$\boxed{X_i = (M + E_i) R_i^\top \quad \text{subject to} \quad R_i R_i^\top = R_i^\top R_i = I_v}.$$

- $E_i \sim \mathcal{MN}_{n,m}(0, \sigma^2 I_n, I_m)$.
- M is the **common mean** matrix with dimension $n \times m$.
- R_i is the **orthogonal rotation parameter**. It has **von Mises-Fisher prior distribution** with location parameter Q and concentration parameter k → coniugate prior for the matrix Normal distribution.

The MAP estimate for R_i is $\hat{R}_i = U_i V_i^\top$, where U_i and V_i derive from the SVD of $X_i^\top M + kQ$.

Limitation: high computational load

Efficient ProMises model: same logic as ProMises model but with a preliminary **dimension reduction** step: $X_i \rightarrow X_i^* \in \mathbb{R}^{n \times n} \rightarrow$ suitable for **matrices with different dimensions**.

Package overview

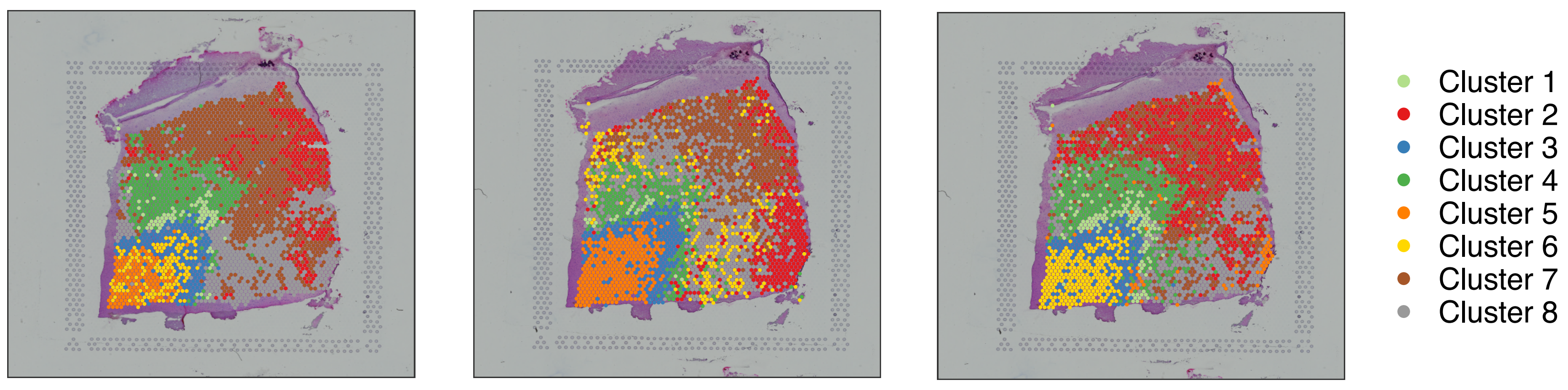
4 main functions:

- `GPASub`: performs functional alignment of a matrix by the ProMises model with known reference matrix M;
- `ProMisesModel`: performs the functional alignment using the ProMises model with unknown reference matrix M. If there are only two matrices, one is rotated with the explicit solution;
- `EfficientProMises`: performs the functional alignment using the Efficient ProMises model decomposing the mean matrix to obtain the light matrices X_i^*: $X_i^* = X_i T^\top$ where $T^\top$ derives from the light-SVD of $\hat{M} = \sum_{i=1}^{N} X_i / N$;
- `EfficientProMisesSubj`: performs the functional alignment using the Efficient ProMises model decomposing the single X_i to obtain the light matrices X_i^*: $X_i^* = X_i T_i^\top$ where $T_i^\top$ derives from the light-SVD of X_i.

1: Two matrices

Explicit solution → `ProMisesModel` function

```
>out = alignProMises::ProMisesModel(data)
```

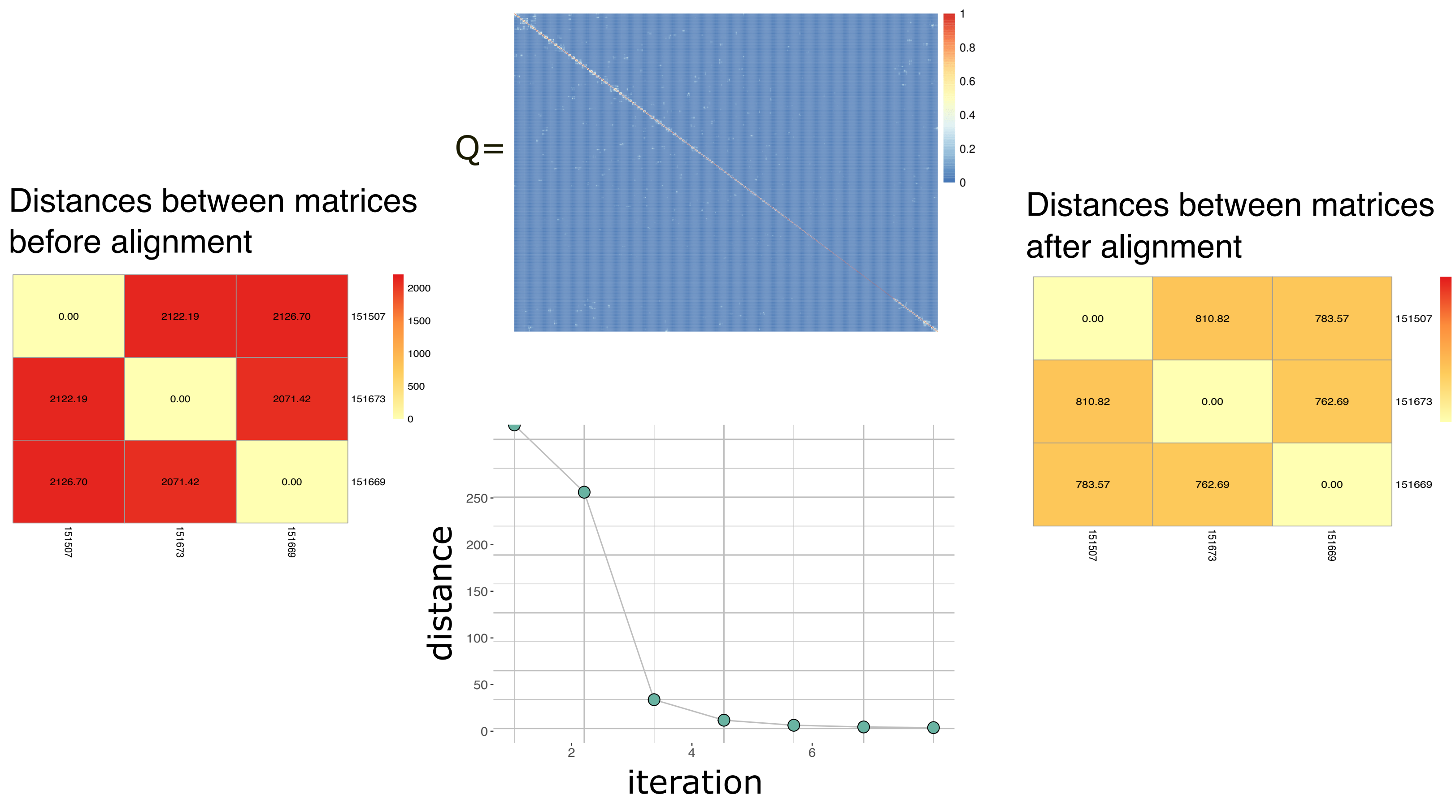


X_2: reference image — X_1 before alignment — X_1 after alignment

2: Three matrices with same dimensions

Efficient ProMises model → `EfficientProMises` function

```
>out <- alignProMises::EfficientProMises(data, t =
1, maxIt = 100, Q=Q, k=1, scaling = T, centered = F)
```



3: Eight matrices with different dimensions

Efficient ProMises model → `EfficientProMisesSubj` function

```
>out <- alignProMises::EfficientProMisesSubj(data, t = 1, maxIt =
100, Q=Q, k=1, scaling = T, centered = F)
```

Two sources of variability:

- **biological variance:** true differences among subjects
- **technical variance** due to the lack of alignment

All subjects in our dataset are **healthy** → differences are mostly **false positives** generated by the technical variability → the Efficient ProMises model absorbs this variability

Number of different expressed genes among two subjects in each layer

Layer	Raw 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

References

[1] Andreella, A., Finos, L. (2022) "Procrustes analysis for high-dimensional data." Psychometrika, 1-17.;

[2] Maynard, K. R. *et al.* (2021). "Transcriptome- scale spatial gene expression in the human dorsolateral prefrontal cortex". Nature neuroscience 24(3), 425–436.