Co-clustering of Spatially Resolved Transcriptomic Data

Andrea Sottosanti & Davide Risso

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

Spatial transcriptomic experiments

The *10X-Visium* technology

- n: the number of genes, whose expression is measured in every spot;
- p: the number of spots;
- 1 spot usually contains more cells;
- the position of each spot on the grid is known.

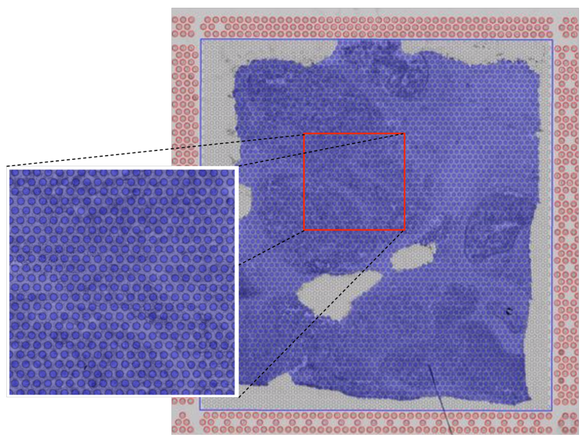

Figure 1. Breast tissue sample analysed with *10X-Visium*.

- The new *10X-Visium* transcriptomic protocol is a modern sequencing technology that allows scientists to achieve a full mapping of the cellular structure of a tissue sample in a relatively easy manner.
- The rise of such advanced technology has increased the interest for the so-called **spatially expressed (s.e.) genes [1]**.

Research issues

1. Determining the clustering of the areas of the tissue sample according to the spatial variation of the genes.
2. Testing if there exist clusters of genes which are *spatially expressed* only in some specific areas discovered from *i.)*.
3. Determining the highly variable genes in the areas discovered from *i.)* net of any spatial effect.

SpaRTaCo: a co-clustering model for spatially resolved data

- We assume the $n \times p$ data matrix $\mathbf{X}$ can be partitioned into $K \cdot R$ blocks, each of which representing a specific *co-cluster*:

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}^{11} & \dots & \mathbf{X}^{1R} \\ \vdots & \ddots & \vdots \\ \mathbf{X}^{K1} & \dots & \mathbf{X}^{KR} \end{bmatrix}, \quad \dim(\mathbf{X}^{kr}) = n_k \times p_r.$$

Definition *The block matrix $\mathbf{X}^{kr}$ is a set of n_k genes whose expression is measured in p_r spots. The spatial coordinates of the spots are contained in the matrix $\mathbf{S}^r$ of dimension $p_r \times 2$.*

- Let $\{\mathcal{Z}_i\}$ and $\{\mathcal{W}_j\}$ be the clustering variables of the genes and of the cells. The block $\mathbf{X}^{kr}$ is made by the the rows $\{i = 1, \dots, n : \mathcal{Z}_i = k\}$ and the columns $\{j = 1, \dots, p : \mathcal{W}_j = r\}$.
- We model the row i of block (k, r) as

$$\mathbf{x}_{i\cdot}^{kr} = \mu_{kr}\mathbf{1}_{p_r} + \sigma_{kr,i}\boldsymbol{\epsilon}_{i\cdot}^{kr}, \quad i = 1, \dots, n_k, \tag{1}$$

$$\sigma_{kr,i}^2 \sim \mathcal{IG}(\alpha_{kr}, \beta_{kr}), \qquad \boldsymbol{\epsilon}_{i\cdot}^{kr} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Delta}_{kr}), \tag{2}$$

where $\mu_{kr} \in \mathbb{R}$ is the co-cluster mean, $\sigma_{kr,i}^2$ is a gene-specific variance, $\boldsymbol{\epsilon}_{i\cdot}^{kr}$ is a Gaussian process with covariance matrix

$$\boldsymbol{\Delta}_{kr} = \underbrace{\tau_{kr}\mathbf{K}(\mathbf{S}^r; \boldsymbol{\phi}_r)}_{\textit{spatial effect}} + \underbrace{\xi_{kr}\mathbb{I}_{p_r}}_{\textit{nugget effect}},$$

where $\mathbf{K}(\cdot;\cdot)$ is a spatial covariance function parametrized by $\boldsymbol{\phi}_r$ and $\tau_{kr}, \xi_{kr} > 0$.

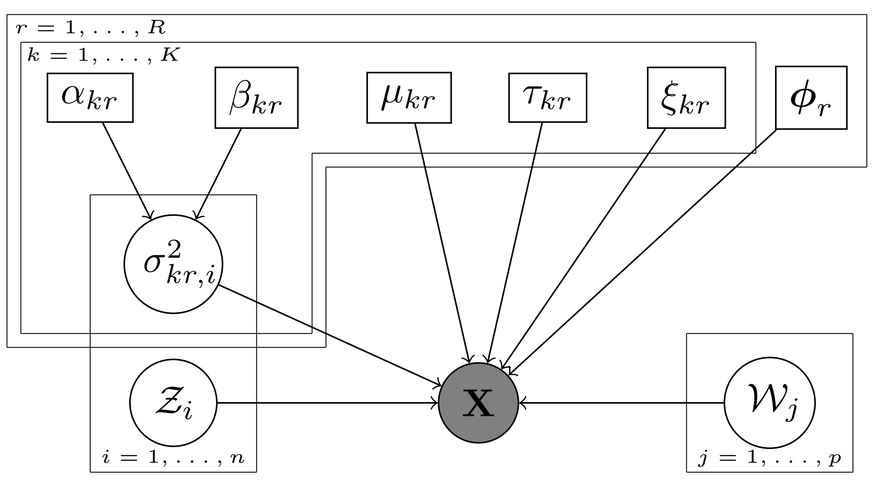

Figure 2. DAG of Model (1) - (2). Grey circle denotes the data, white circles are the unknown random variables, and white rectangles are the model parameters.

The log-likelihood function

- To make the model identifiable, we impose a constraint of the type $\tau_{kr} + \xi_{kr} = c$ for any k and r.
- The estimate of the model parameters are taken by maximizing the *classification log-likelihood*

$$\ell(\boldsymbol{\Theta}, \{\mathcal{Z}_i\}, \{\mathcal{W}_j\}) = \sum_{i=1}^{n}\sum_{k=1}^{K}\mathbb{1}(\mathcal{Z}_i = k)\left\{\sum_{r=1}^{R}\log p(\mathbf{x}_{i\cdot}^{r}; \boldsymbol{\theta}_{kr}, \boldsymbol{\phi}_r)\right\} \tag{3}$$

where $p(\cdot;\cdot)$ is the marginal model of $\mathbf{x}_i^{kr}$, with $\sigma_{kr,i}^2$ integrated out, $\mathbf{x}_{i\cdot}^{r}$ is the i-th row of the matrix $\mathbf{X}^{\cdot r}$, and $\boldsymbol{\theta}_{kr} = \{\mu_{kr}, \tau_{kr}, \alpha_{kr}, \beta_{kr}\}$.

- The number of clusters K and R and the spatial covariance function $\mathbf{K}(\cdot|\cdot)$ are selected using the ICL criterion. We consider here the exponential covariance function: $\exp\{-||\mathbf{s}_j - \mathbf{s}_{j'}||/\boldsymbol{\phi}_r\}$.

The CS-EM estimation algorithm

Our estimation algorithm iterates the following steps until convergence.

1. Given $\{\mathcal{Z}_i\}$ and $\{\mathcal{W}_j\}$, find $\hat{\boldsymbol{\Theta}}$ that maximizes (3).
2. Given $\{\mathcal{W}_j\}$ and $\hat{\boldsymbol{\Theta}}$, update the row clusters with a classification step (CEM algorithm) as in **[2]**.
3. Given $\{\mathcal{Z}_i\}$ and $\hat{\boldsymbol{\Theta}}$, propose a new column clustering configuration $\{\tilde{\mathcal{W}}_j\}$ as in **[3]**, and accept it with a Metropolis-Hastings move (SEM algorithm).

The Human Dorsolateral Prefrontal Cortex

- We consider the *human dorsolateral prefrontal cortex* (DLPFC) spatial transcriptomic dataset analysed with *10X-Visium* and contained in the `R` package `spatialLIBD`.
- We pre-process the dataset with the tools proposed by **[4]** and implemented in the `R` package `scry`; the final dataset analysed is made of **500 genes** measured in **3639 spots**.

- Genes in the second cluster ($n_2 = 129$) have an estimate of the spatial-nugget effect ratios τ_{2r}/ξ_{2r} substantially larger than those in the first ($n_1 = 371$), for every region $r = 1, \dots, 9$.
- The largest level of spatial expression is obtained in the first spot cluster ($p_1 = 243$).

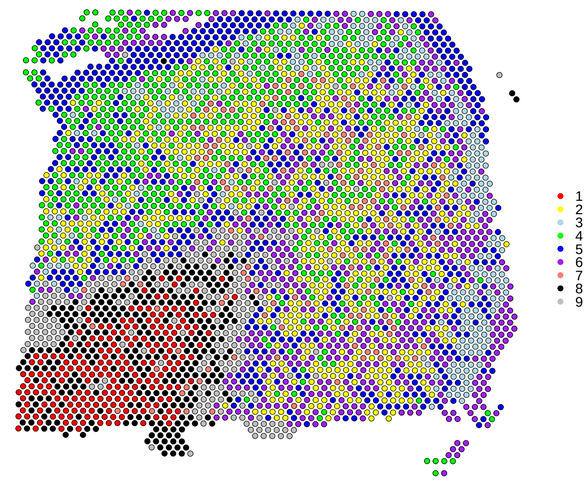

Figure 3. Example of a human DLPFC sample. The spots are coloured according to estimated clusters.

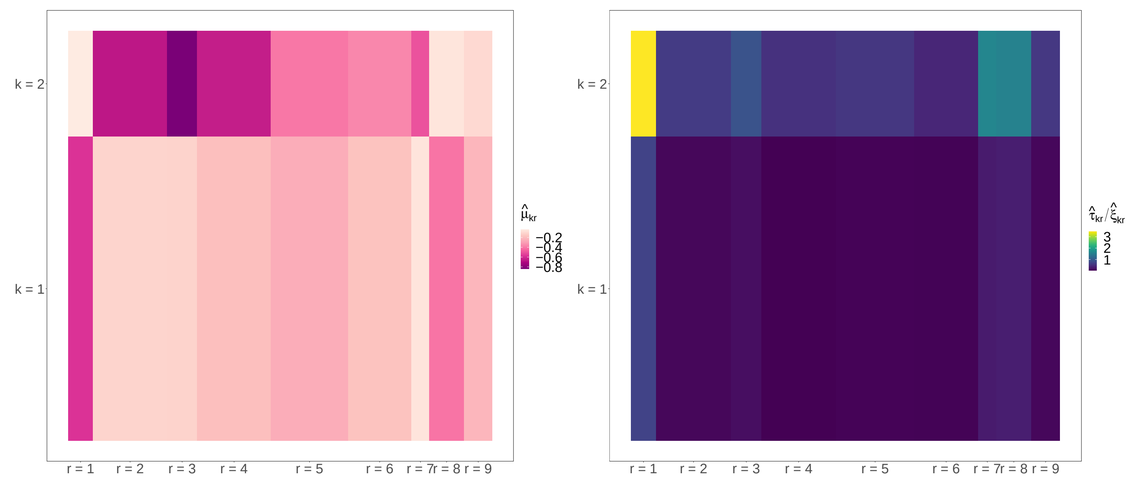

Figure 4. Left: Representation of $\mathbf{X}$ divided into the estimated blocks, coloured according to $\hat{\mu}_{kr}$ (left) and to $\hat{\tau}_{kr}/\hat{\xi}_{kr}$ (right).

- By exploiting the distribution of the gene variance $\sigma_i^2|\mathbf{X}, \{\mathcal{Z}_i\}_i, \{\mathcal{W}_j\}_j$ it is possible to determine the highly variable genes in every cluster of spots.
- For example, within the area $r = 1$ which corresponds to the white matter area, the gene `CERCAM` appears as highly variable, even though the pre-selection techniques of **[4]** did not consider it as highly variable (see the left panel of Figure 4).

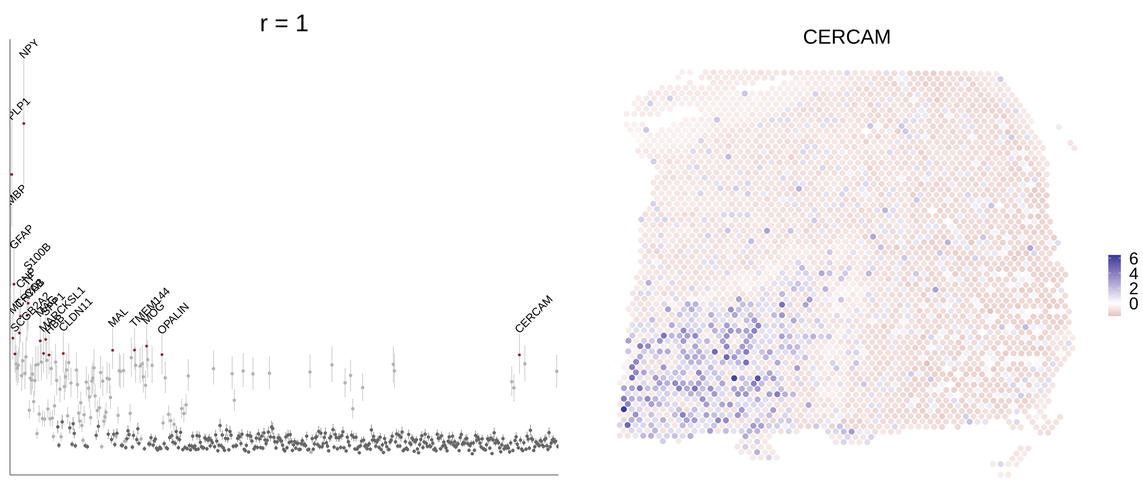

Figure 5. Left: Distribution of $\sigma_i^2|\text{data}$ in region $r = 1$. Right: Expression of gene `CERCAM` over the whole image.

References

1. Nobile, A., & Fearnside, A. T. (2007). Bayesian finite mixtures with an unknown number of components: the allocation sampler. *Statistics and Computing*, **17(2)**, 147–162.
2. Samè, A., Ambroise, C., & Govaert, G. (2007). An online classification EM algorithm based on the mixture model. *Statistics and Computing*, **17(3)**, 209–218.
3. Svensson, V., Teichmann, S. A., & Stegle, O. (2018). SpatialDE: identification of spatially variable genes. *Nature methods*, **15(5)**, 343–346.
4. Townes, F. W., Hicks, S. C., Aryee, M. J., & Irizarry, R. A. (2019). Feature selection and dimension reduction for single-cell RNA-Seq based on a multinomial model. *Genome biology*, **20(1)**, 1-16.

Contact information

andrea.sottosanti@unipd.it @ andreasottosanti.github.io A_Sottosanti