CLUSTERING GENES SPATIAL EXPRESSION PROFILES WITH THE AID OF EXTERNAL BIOLOGICAL KNOWLEDGE

Andrea Sottosanti [1], Sara Agavni' Castiglioni [2], Stefania Pirrotta [3] , Enrica Calura[3], and Davide Risso [2]

[1] Department of Medicine, University of Padova,
(e-mail: andrea.sottosanti@unipd.it)

[2] Department of Statistical Sciences, University of Padova,

[3] Department of Biology, University of Padova

ABSTRACT: In the analysis of spatial transcriptomic experiments, the recently proposed SpaRTaCo model (Sottosanti & Risso, 2022) allows for the simultaneous clustering of genes and cells of a tissue sample, providing interesting insights abouve the underlying biological processes. In this work, we discuss how to integrate external knowledge such as manual cell-type annotations to inform gene clustering, with the by-product of substantially reducing the computational burden.

KEYWORDS: clustering, genomics, spatial statistics, spatial transcriptomics.

1 Introduction

Spatial transcriptomics is an innovative class of sequencing technologies, capable of providing the expression levels of thousands of genes in a tissue sample while retaining the spatial conformation of the analyzed tissue. With the aid of additional spatial information, researchers can better understand the complex biological processes that depend on the cellular organization of the tissue. New insights come from the discovery of *spatially expressed* (s.e.) genes, i.e., genes that exhibit specific patterns of variation in space (Svensson *et al.*, 2018).

Recently, we proposed SpaRTaCo (Sottosanti & Risso, 2022), a co-clustering model for spatial transcriptomic experiments, which has shown to be capable of determining s.e. genes active only in specific areas of a sample, providing insights that could not be achieved by competing methods in the literature. Clearly, it represents a useful tool for spatial transcriptomic data analysis; nevertheless, its estimation process is highly computationally demanding.

Here, we propose a modification of the original SpaRTaCo formulation that integrates external biological knowledge to speed up the computation. In

fact, spatial experiments often come with a manual annotation of the cellular composition of a sample made by a pathologist, providing a relevant source of information that can be integrated into the inferential process. Furthermore, we propose to estimate SpaRTaCo with a penalized maximum likelihood approach to prevent the model from capturing spurious spatial correlation, retaining relevant patterns only. We conclude with the analysis of a prostate cancer tissue sample analyzed with a recent spatial transcriptomic technology.

2 The semi-supervised SpaRTaCo with L_1 and L_2 penalizations

Let $\mathbf{X}$ be the $n \times p$ matrix of a spatial experiment having the expression of n genes measured over p spots, whose spatial locations are known. SpaRTaCo assumes the existence of K gene clusters and R spot clusters, inducing a partition of the experiment matrix into $K \times R$ blocks. Thus, the kr-th block has dimension $\dim(\mathbf{X}^{kr}) = n_k \times \ p_r$, and $\mathbf{X} = \left(\mathbf{X}^{kr}\right)$, with $k = 1, \dots, K$ and $r = 1, \dots, R$. The expression of the i-th gene with the kr-th block distributes as

$$\mathbf{x}_{i.}^{kr} | \sigma^2_{kr,i} \sim N_{p_r}\left(\mu_{kr}\mathbf{1}_{p_r}, \sigma^2_{kr,i}\mathbf{\Delta}_{kr}\right), \qquad \sigma^2_{kr,i} \sim I\mathcal{G}(\alpha_{kr}, \beta_{kr}) \tag{1}$$

where μ_{kr} is a mean parameter, $\mathbf{1}_{p_r}$ is a vector of ones of length p_r, $\sigma^2_{kr,i}$ is a gene-specific variance, and $\mathbf{\Delta}_{kr}$ is the covariance matrix of the spots with form

$$\mathbf{\Delta}_{kr} = \tau_{kr}\mathcal{K}(\mathbf{S}_r; \phi_r) + \xi_{kr}\mathbb{1}_{p_r}. \tag{2}$$

Notice that $\mathbf{\Delta}_{kr}$ is expressed as a linear combination of two matrix terms: the first is a kernel matrix with isotropic spatial covariance function $k(\cdot; \phi_r)$ that models the gene expression correlation across the spots of cluster r (with spatial coordinates $\mathbf{S}^r = (s_j)$), the second is an identity matrix of size p_r. The parameters τ_{kr} and ξ_{kr} quantify the amount of spatial variation and residual intra-block variability, respectively. Moreover, the quantity τ_{kr}/ξ_{kr} can be used to measure the amount of spatial variability compared to the residual variability, and for this reason it is called *spatial signal-to-noise ratio*. Last, the parameter $\sigma^2_{kr,i}$ in (1) is used to model the variance specific of gene i in the kr-th to account for the possible dependence across genes in the same cluster.

Even though SpaRTaCo is designed for clustering both rows (genes) and columns (cells) of $\mathbf{X}$, when a manual annotation of the tissue image is available, we can include it in the model in place of the column clustering labels to inform the inferential process. In addition, to improve the stability of parameter estimation, we can estimate the model with a penalized maximum likelihood approach. A *lasso* penalty on the parameters τ_{kr} discourages the

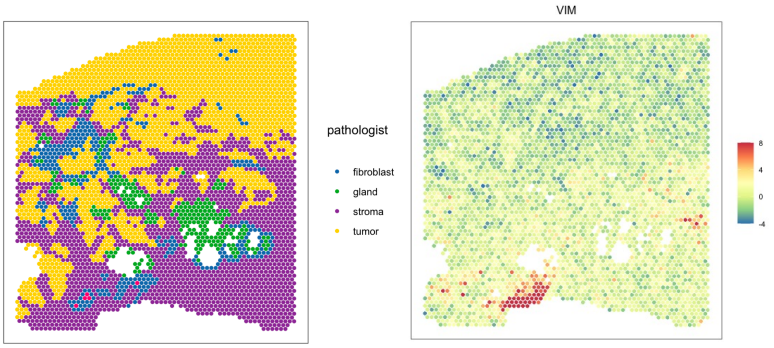

Figure 1: Left: human prostate tissue diagnosed with adenocarcinoma. Spot colours denote Dr. Esposito's annotation (red spots are not considered as they appear only 5 times). Right: spatial distribution of gene VIM.

model from capturing spurious correlation when no spatial effect is present, while a *ridge* penalty regularizes the mean parameters μ_{kr} since zero values do not have a clear biological meaning. The estimates of the model parameters $\Theta = \cup_r \{\cup_k (\mu_{kr}, \tau_{kr}, \alpha_{kr}, \beta_{kr}), \phi_r\}$ and of the clustering labels $\mathcal{Z}$ are obtained maximizing

$$\log \mathcal{L}(\Theta, \mathcal{Z}|\mathbf{X}, \mathcal{W}) - \lambda_\tau \sum_{k=1}^{K} \sum_{r=1}^{R} |\tau_{kr}| - \lambda_\mu \sum_{k=1}^{K} \sum_{r=1}^{R} \mu_{kr}^2, \tag{3}$$

where $\mathcal{W}$ is the vector containing the spot clustering labels that come with the data, $\log \mathcal{L}(\Theta, \mathcal{Z}|\mathbf{X}, \mathcal{W})$ is the classification log-likelihood, and λ_τ and λ_μ are the penalization terms associated to the τ and μ parameters, respectively. Simulation studies not reported here showed that $\lambda_\mu = 1.5$ and $\lambda_\tau = 0.3$ guarantee robust parameter estimates and prevent the model from capturing spurious spatial correlation. Notice that the parameters ξ_{kr} are not estimated, but are fixed a priori, for identifiability reasons. An exact solution to the maximization of (3) can be obtained using a classification EM algorithm.

3 Application to human prostate cancer data

We analyze a human prostate tissue diagnosed with adenocarcinoma processed with 10X-Visium platform (Righelli *et al.*, 2022). The slide was manually annotated by the pathologist Dr. Esposito (Veneto Oncology Institute, Italy), by analyzing microscope images that consider the cytoarchitecture of

the cells, i.e., the spatial organization and arrangement of cells within the tissue. Based on these characteristics, the tissue was divided into four macro categories: fibroblasts, glands, stroma, and tumour (Figure 1, left). After preliminary gene filtering and count normalization (Townes *et al.*, 2019), the final dataset had 1000 genes measured over 4366 locations (spots).

We estimated the semi-supervised SpaRTaCo using $K \in \{1,\dots,9\}$ and, after evaluating the *integrated complete log-likelihood* criterion and the clustering uncertainties (Sottosanti & Risso, 2022, Section 3.3 and 3.4), we selected the model with $K = 5$ gene clusters. The first two clusters that the model identifies have a substantial spatial variability in all tissue areas ($\hat{\tau}_{kr}/\hat{\xi}_{kr} > 1.5$, for $k = 1,2$ and $\forall r$) and particularly pronounced in the tumour area ($\hat{\tau}_{14}/\hat{\xi}_{14} = 7.12$, $\hat{\tau}_{24}/\hat{\xi}_{24} = 2.46$). In comparison, the remaining three gene clusters have moderate or absent spatial variability throughout the tissue and show substantial differences only at the mean level.

Thanks to gene-specific variance parameters $\sigma^2_{kr,i}$, we can provide a list of the most variable genes in every tissue area. As an example, the gene VIM appeared among the 20 most variable genes in the stromal region (Figure 1, right). VIM is a cancer growth promoter gene, and therefore, from this observed expression pattern, it can provide helpful information about the nature of the tumour and be the starting point for biological investigations. Alternative algorithms for selecting highly variable genes (e.g., Townes *et al.*, 2019) do not include VIM among the top 80 most informative genes, showing the importance of accounting for the spatial variability of the data in the analysis.

References

RIGHELLI, DARIO, WEBER, LUKAS M, CROWELL, HELENA L, & PARDO, BRENDA, ET AL. 2022. SpatialExperiment: infrastructure for spatially-resolved transcriptomics data in R using Bioconductor. *Bioinformatics*, **38**(11), 3128–3131.

SOTTOSANTI, ANDREA, & RISSO, DAVIDE. 2022. Co-clustering of Spatially Resolved Transcriptomic Data. *The Annals of Applied Statistics*. In press.

SVENSSON, V., TEICHMANN, S., & STEGLE, O. 2018. *SpatialDE: identification of spatially variable genes | Nature Methods*.

TOWNES, F. WILLIAM, HICKS, STEPHANIE C., ARYEE, MARTIN J., & IRIZARRY, RAFAEL A. 2019. Feature selection and dimension reduction for single-cell RNA-Seq based on a multinomial model. *Genome Biology*, **20**(1), 295.