Clusterwise Survival Analysis of Lung Cancer Data

Srijan Chattopadhyay, Swapnaneel Bhattacharyya

Indian Statistical Institute, Kolkata

Introduction

- Spatial multiplex imaging allows for the simultaneous visualization and analysis of multiple biomarkers within a single tissue sample while preserving spatial context.
- Biopsies are taken from each patient, and Regions of Interest (ROIs) are identified from Tumor Micro Arrays (TMA), which are slides containing one or more patient tumors.
- For each ROI, the locations and types of each cell in the ROI are obtained.

Data Description

We have the **spatial multiplex images** for the lung cells of **153 patients** specifying the positions and cell types in an image with the patient's demographic information (Gender, Age, Stage, Adjuvant Therapy, MHCII Status).

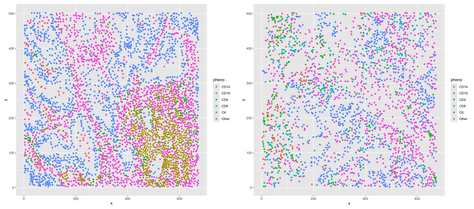


Figure 1. Visualization of the cell structure

Estimating Intensities

- We estimate spatial point pattern intensities for each cell type per patient using a kernel density estimator and we convert matrices to vectors.
- Combine vectors row-wise and perform PCA, retaining components explaining 90% of the variation.
- Resulting in a matrix of principal components for all cell types.

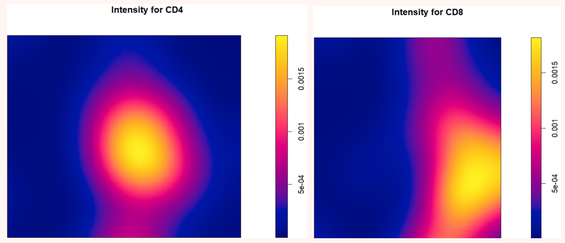


Figure 2. Visualizing the intensities

K functions and K cross functions

- For a spatial point pattern, $K-$function at a particular radius r is defined as

$$\widehat{K}(r) = \frac{|A|}{n(n-1)} \sum_{i=1}^{n} \sum_{i \neq j} \mathbf{1}\left(d\left(c_i, c_j\right) \leq r\right) e_{ij}$$

  where $|A|$ is the area of the image and $\{c_i\}_{i=1}^{n}$ are the position of the points.
- Another version of these functions are $K-$cross and permuted $K-$functions($\tilde{K}(r)$) which are used to extract spatial information embedded in a multiplex image.
- For each image we find **deciles** of $|\hat{K}(r) - \tilde{K}(r)|$ and use as spatial covariates.

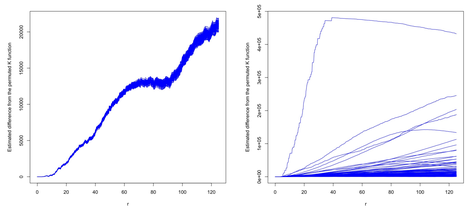
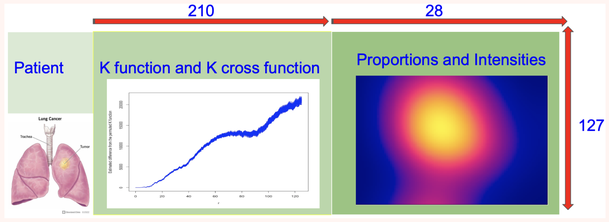


Figure 3. Visualization of $|K(r) - \tilde{K}(r)|$

The Final Dataset

- The principal components of the estimated intensities, deciles of the $K-$Functions, and the proportions of the cell types in each image are combined as the final dataset having **127 patients** and **238 spatial features**. Clustering is done based on this dataset after scaling columnwise.

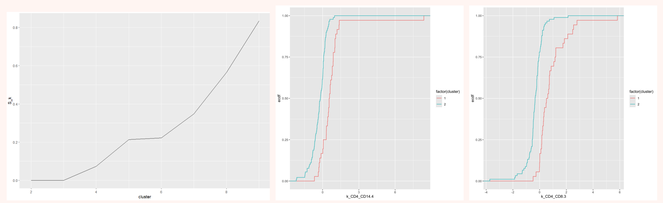


Clustering

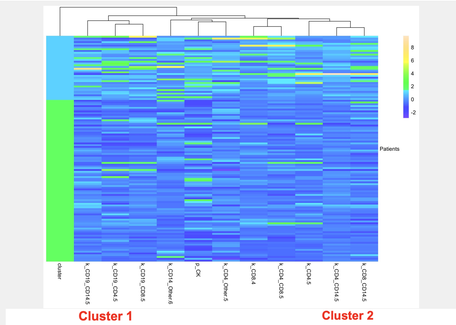
- In the combined dataset we use **Partitioning around Medoid (PAM)** algorithm for robust clustering of patients according to all the spatial features collected from the images.
- For a particular no of clusters(k) the PAM algorithm searches for k representative objects in a data set (k medoids) and then assigns each object to the closest medoid to create clusters aiming towards minimizing the intercluster variation.

Choosing the optimal no of Clusters(K)

- For a k, let $A_1, \cdots, A_k$ be the data matrices broken cluster-wise where a column represents a feature.
- For the rth feature, let p_r be the median of $p_{i,j}^r$ over $(i,j) : i,j \in \{1,2,\cdots,n\}, i \leq j$ where for each of $\binom{n+1}{2}$ pairs (i,j), $p_{i,j}^r$ is the p-value for the exact two-sample Kolmogorov-Smirnov test between the vectors $A_i[,r]$ and $A_j[,r]$.
- Let $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(238)}$ be the ordered values $p_1, \cdots, p_{238}$. (There are 238 features in total). Let $S_k = p_{(1)} + \cdots + p_{(24)}$ and choose the K which minimizes S_k.
- The optimal number of clusters turns out to be 2.



Feature Variation across Clusters



Cluster 1

55.56% in 1st stage, 19.44% in 2nd stage, 22.22% in 3rd stage and 2.78% in 4th stage

55.56% are male and 44.44% are female

91.67% didn't receive therapy

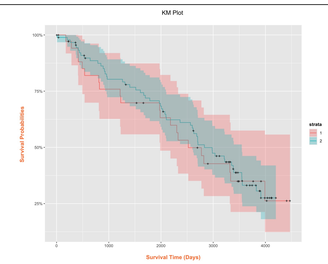
Cluster 2

63.74% in 1st stage, 21.98% in 2nd stage, 9.89% in 3rd stage, 4.39% in 4th stage

49.45% are male and 50.55% are female

83.52% didn't receive therapy

Clusterwise Survival Analysis



Conclusions

- Cluster 1 has more Cancer Cells than Cluster 2
- Cluster 1 has more interaction between macrophage and B Cells than Cluster 2
- Cluster 1 has more interaction between macrophage and helper T Cells than Cluster 2
- Cluster 1 has more interaction between macrophage and Other Cells than Cluster 2
- Cluster 1 has more clustered helper T cells, cytotoxic T cells than Cluster 2
- Cluster 1 has more interaction between cytotoxic T cells and macrophages than Cluster 2
- Cluster 2 people has more survival chance than Cluster 1

Limitations

- In this study we can not capture the joint spatial effect of three or more cell types.
- The Principal Components of the intensities are not very much interpretable.
- Choosing deciles in place of other quantiles can be justified more statistically by some sensitivity analysis.

Appendix - Abbreviations for Cell Types

CD14 (macrophages), CD4 (helper T cells), CD8 (cytotoxic T cells), CD19 (B cells), CK (Cancer Cells), and Other Cell Types