UNIVERSITY OF MICHIGAN

Clustering of the Lung Cancer Data

Big Data Summer Institute
University of Michigan

Swapnaneel Bhattacharyya
Srijan Chattopadhyay

Indian Statistical Institute, Kolkata
swapnanee@umich.edu\srijanch@umich.edu

Goals of this work

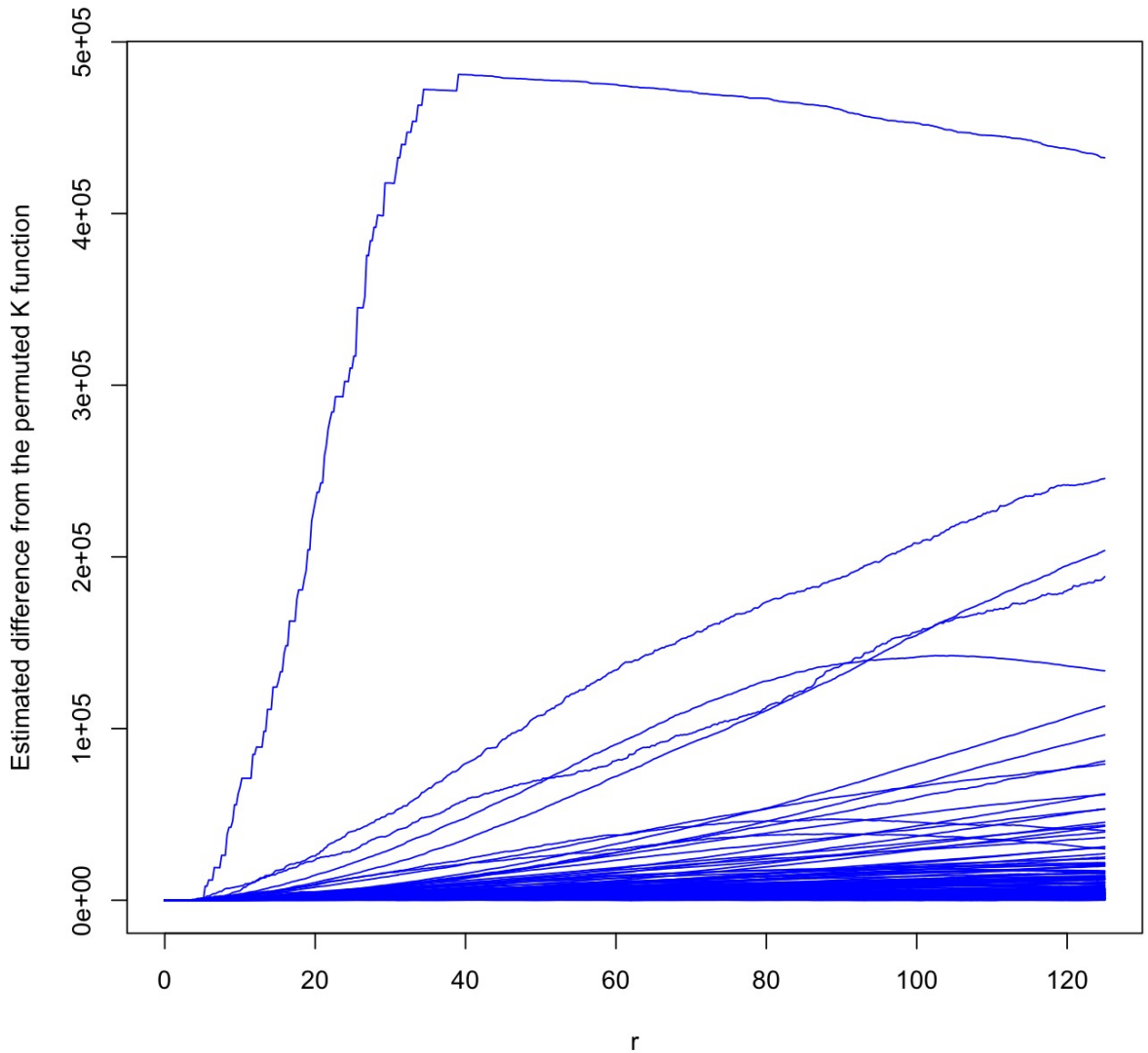
In this work,

- We obtain the clustering of the patients based on the spatial information present in the multiplex images.
- In the obtained clusters we find the clusterwise survival probabilities and plot the Kaplan–Meier curves.

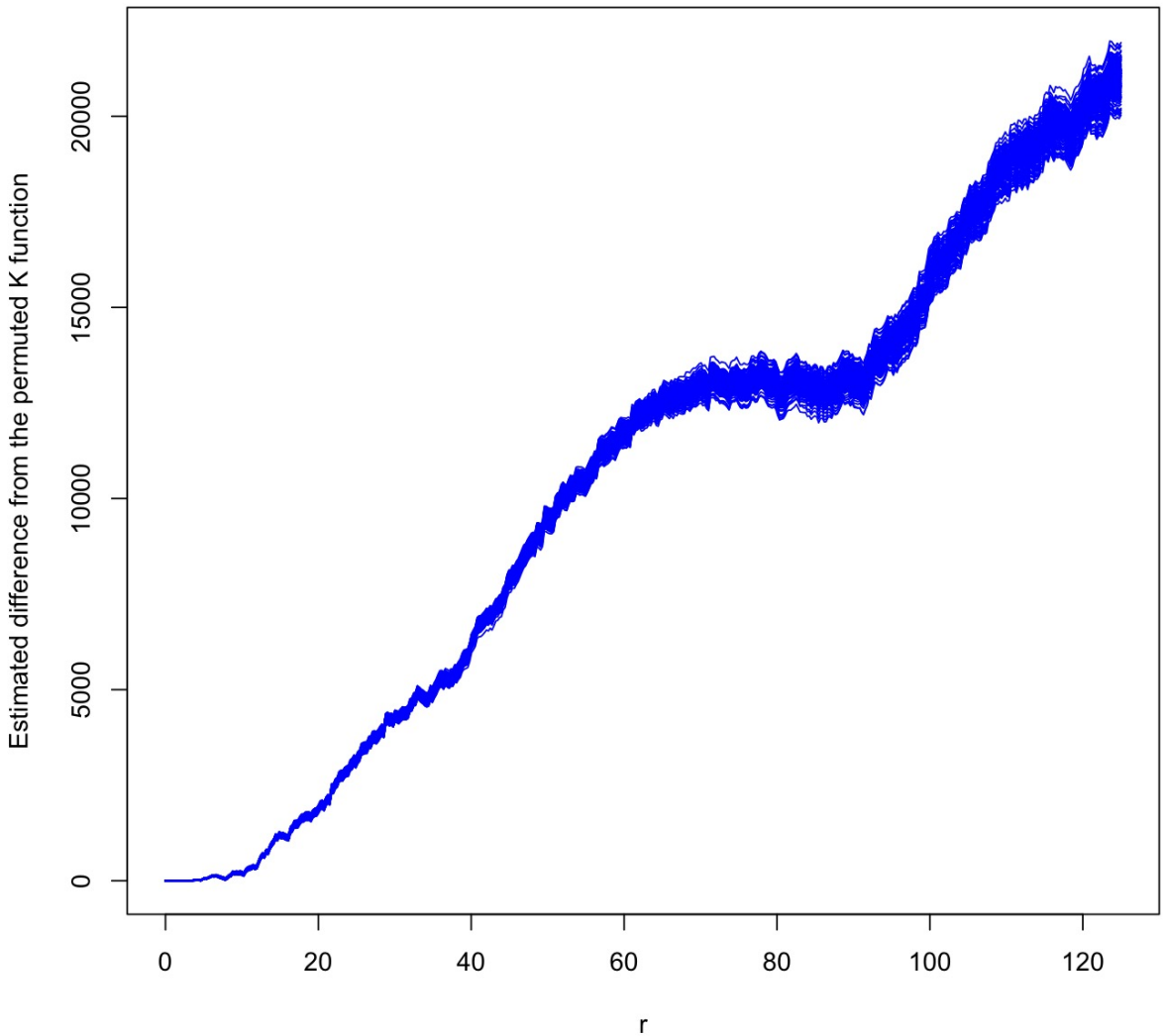
Clustering of Patients

- We first obtain the clustering of the patients based on the spatial information present in the multiplex images.
- For that, for each patient, we first choose the image with the highest number of points.
- Now from that image, we summarize the spatial information.
- For each patient, we find the **the deciles** of absolute difference of K**-function**$(K(r))$ and **Permuted K-function**$(\tilde{K}(r))$.
- We also **estimate** the **intensities of the point processes** generated by **each of the cell types** in the images of each patient.

Visualizing $|K(r) - \tilde{K}(r)|$



Visualizing $|K(r) - \tilde{K}(r)|$



Visualizing the Intensities

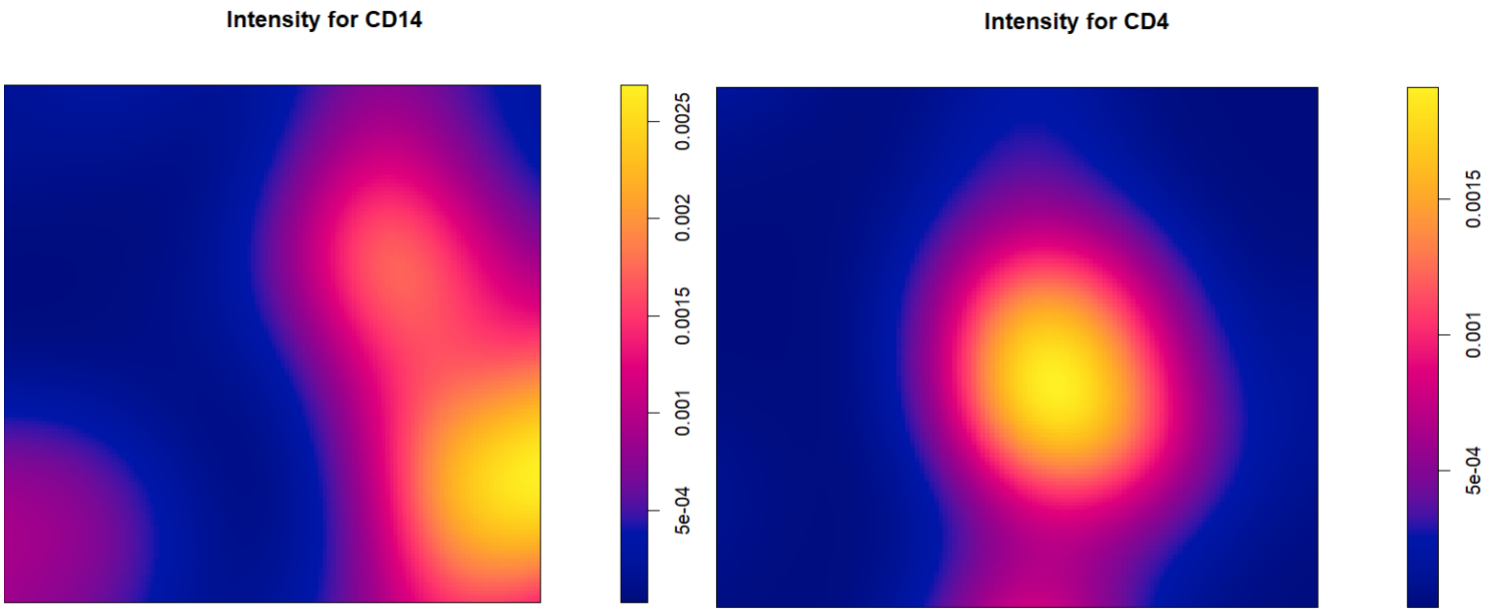


Figure: Intensities for CD14 (macrophages), CD4 (helper T cells)

Visualizing the Intensities

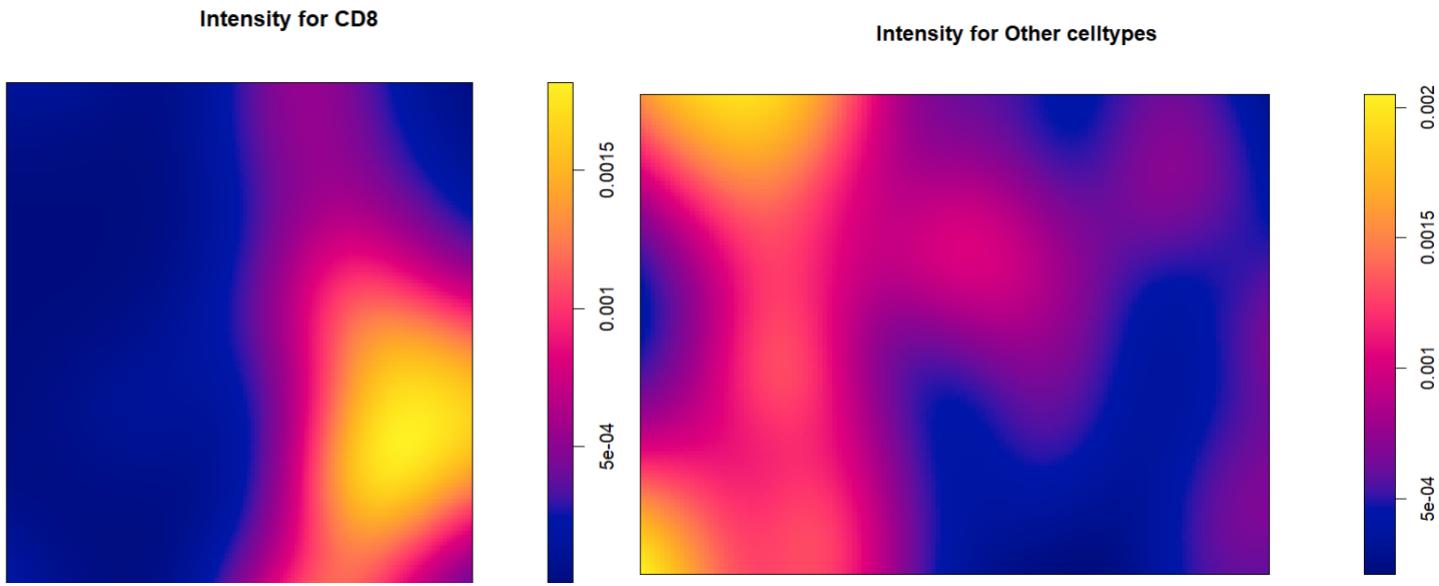


Figure: Intensities for CD8 (cytotoxic T cells) and Other Cell Types

Estimating Intensities

- For each patient, each of the cell types introduces a spatial point pattern.
- We estimate those intensities by the kernel density estimator ([1])

$$\hat{\lambda}(u) = \frac{\sum_{i=1}^{n} \kappa(u - x_i)}{\int_W \kappa(u - v)dv}$$

  where $\{x_i\}_{i=1}^{n}$ are point positions.
- That intensities are estimated in the form of a matrix, we make them vector.
- So for each patient, for a fixed cell type (say, B Cells) we have a vector of estimated intensities.
- Combine the vectors row-wise and perform PCA to reduce the number of columns of the data frame.

Estimating Intensities

- We consider the principal components explaining 90% of the variation.
- So for all of the cell types we have a matrix of principal components of the intensities.
- We take the $K-$function deciles and these estimated intensities as covariates.

Preparing the combined dataset

- We now combine the **deciles of $K-$functions, intensities, and the proportion of cell types** in a single dataset.
- So we have 238 spatial features in total.
- For some of the patients, we have missing values for some of these features. We remove those patients from our dataset.
- After removal, there are 127 patients in total.
- No significant correlations was found between almost all of the existing features.

A look at the final data

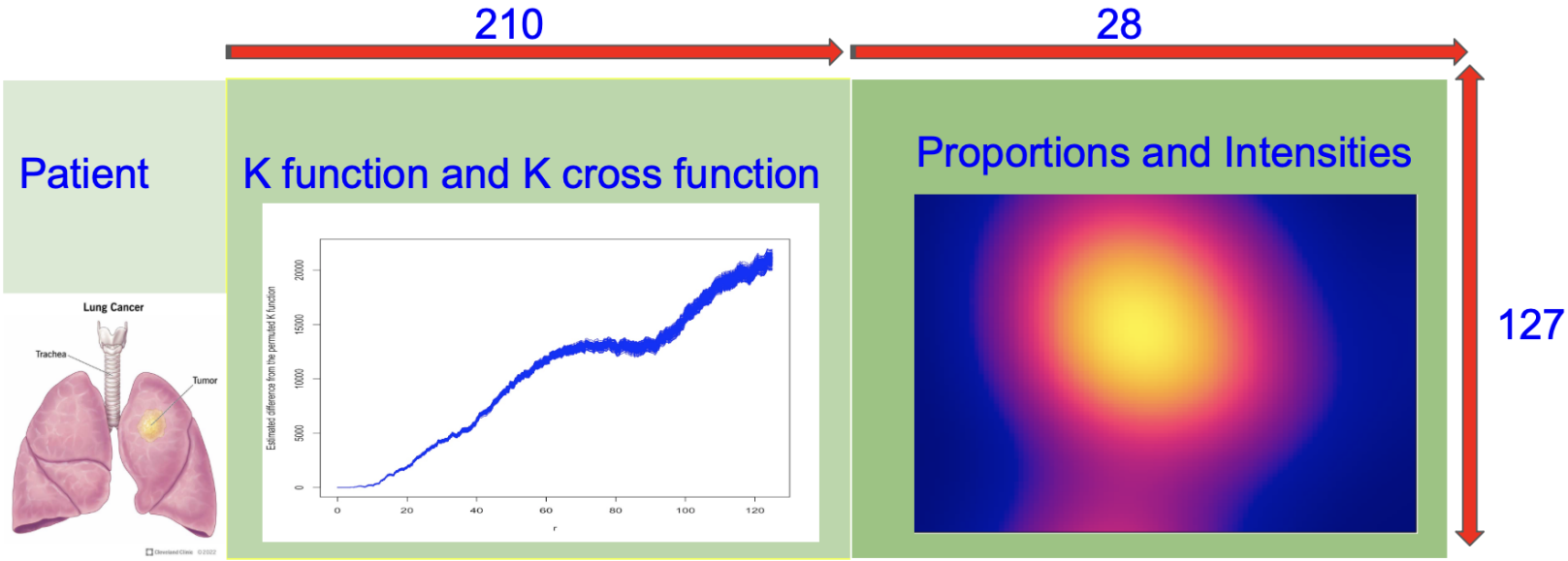
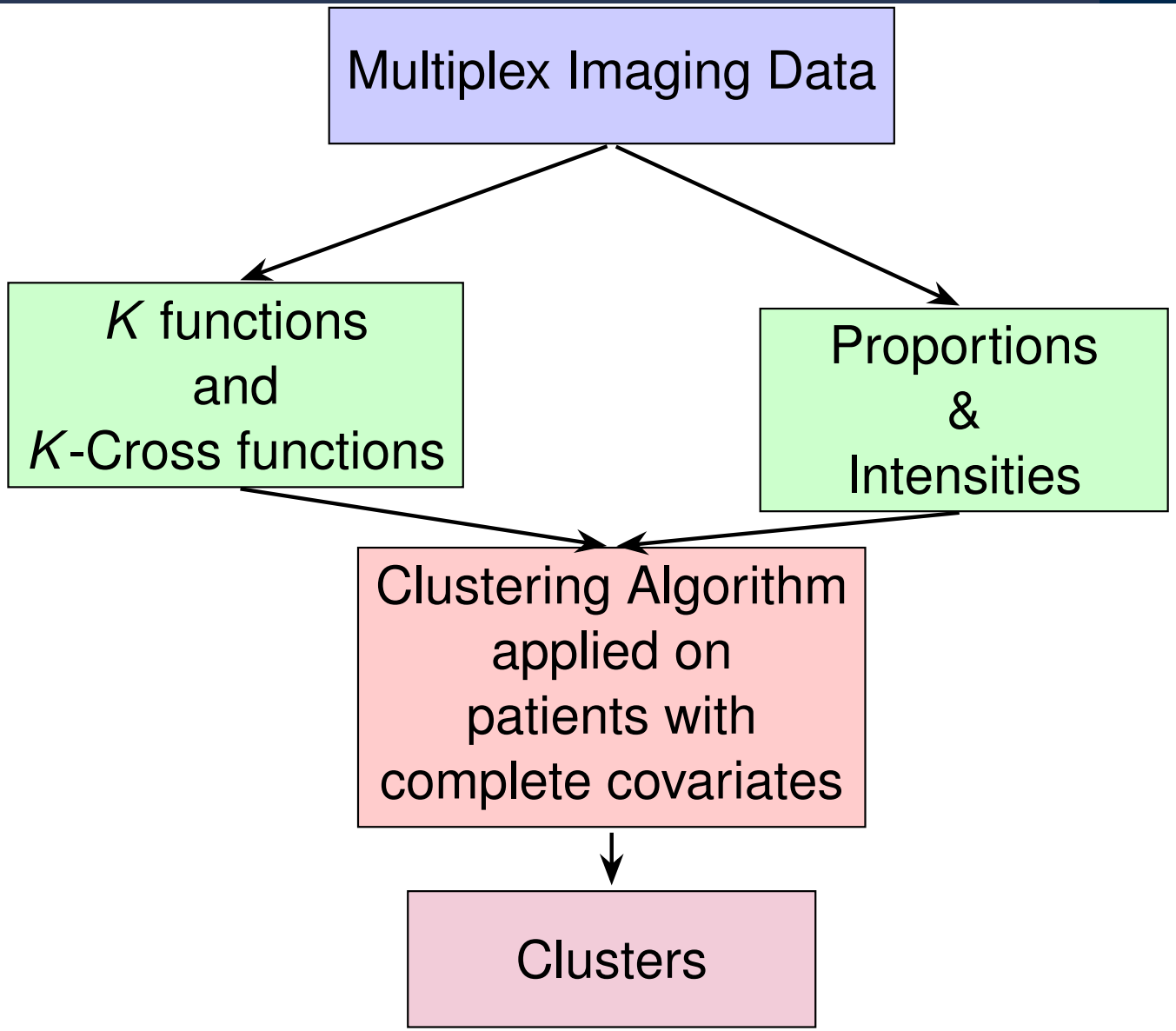


Figure: The Combined Dataset

At a Glance



Clustering of the patients

- In the combined dataset, we use this spatial information for clustering.
- We use **Partitioning around Medoid (PAM)** algorithm for clustering patients according to all the spatial features collected from the images.
- 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.
- It aims to minimize the sum of dissimilarities between the objects in a cluster and the center of the same cluster (medoid).
- It is known to be a robust version of the k-means algorithm as the median is more robust than the mean to outliers.

Choosing 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 $p^r_{i,j}$ over $(i,j) : i,j \in \{1,2,\cdots,n\}, i \leq j$ where for each of $\binom{n+1}{2}$ pairs (i,j), $p^r_{i,j}$ 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 $S_k = p_{(1)} + \cdots + p_{(24)}$, i.e. taking top 10% of the p values only.
- We choose the K which minimizes S_k.
- The optimal number of clusters turns out to be 2.

Optimal number of Clusters

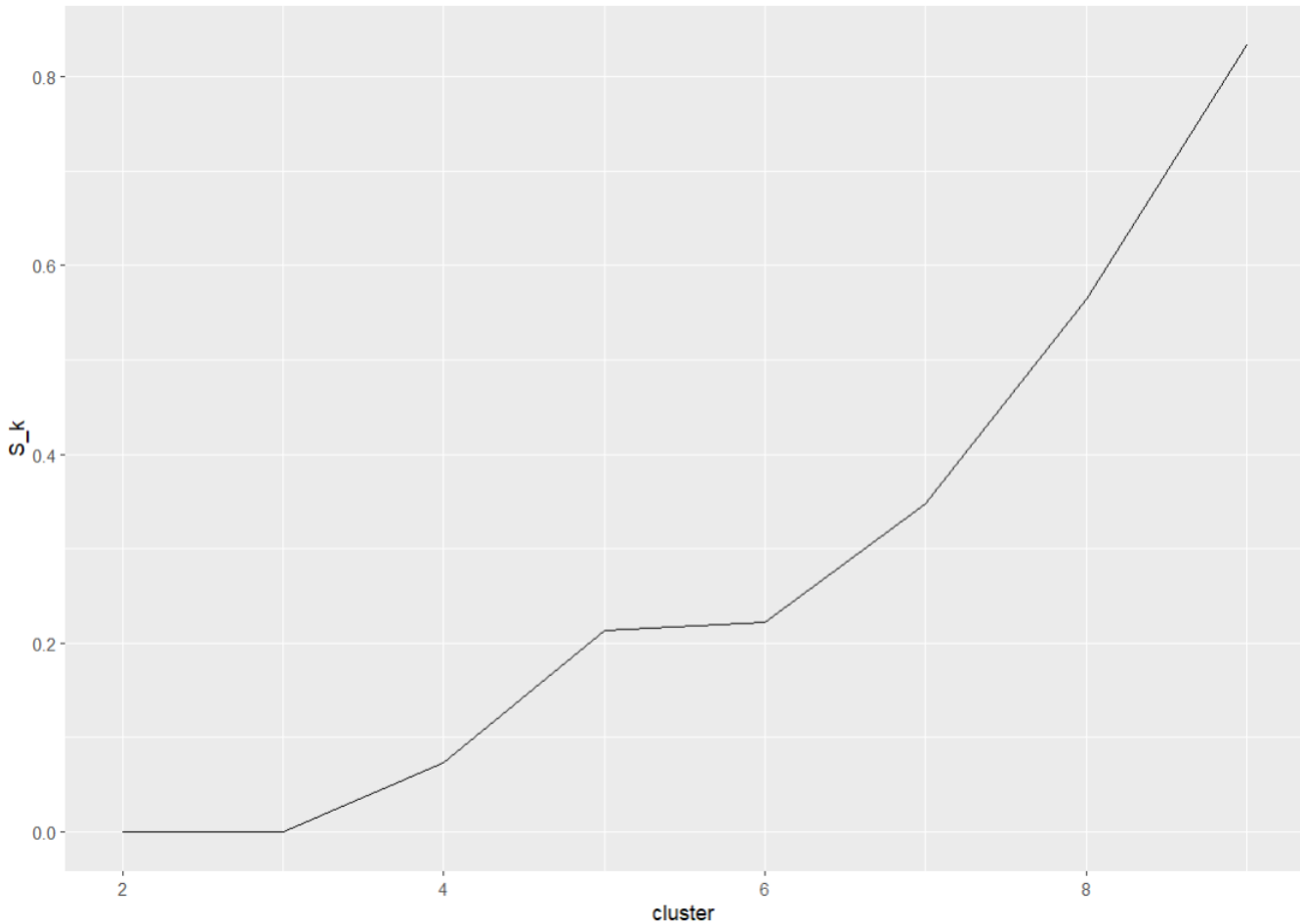


Figure: S_k vs k (k = Number of Clusters)

Clustering for $k = 2$

- There are 127 patients in the combined dataset.
- In the optimal clustering, the two clusters have 36 and 91 patients.
- In the 1st cluster 15 patients survived, while 21 couldn't. Whereas, in the 2nd cluster, 36 patients survived and 55 couldn't.

Heatmap of features across clusters

- We now find the features that have maximum variation across the two clusters.
- For each of the features, we perform a two-sided Kolmogorov Smirnov (KS) exact test between the set of values taken by the feature in two clusters.
- For the heatmap, we take the features having p-value < 0.05 with a multiple testing correction.

Heatmap

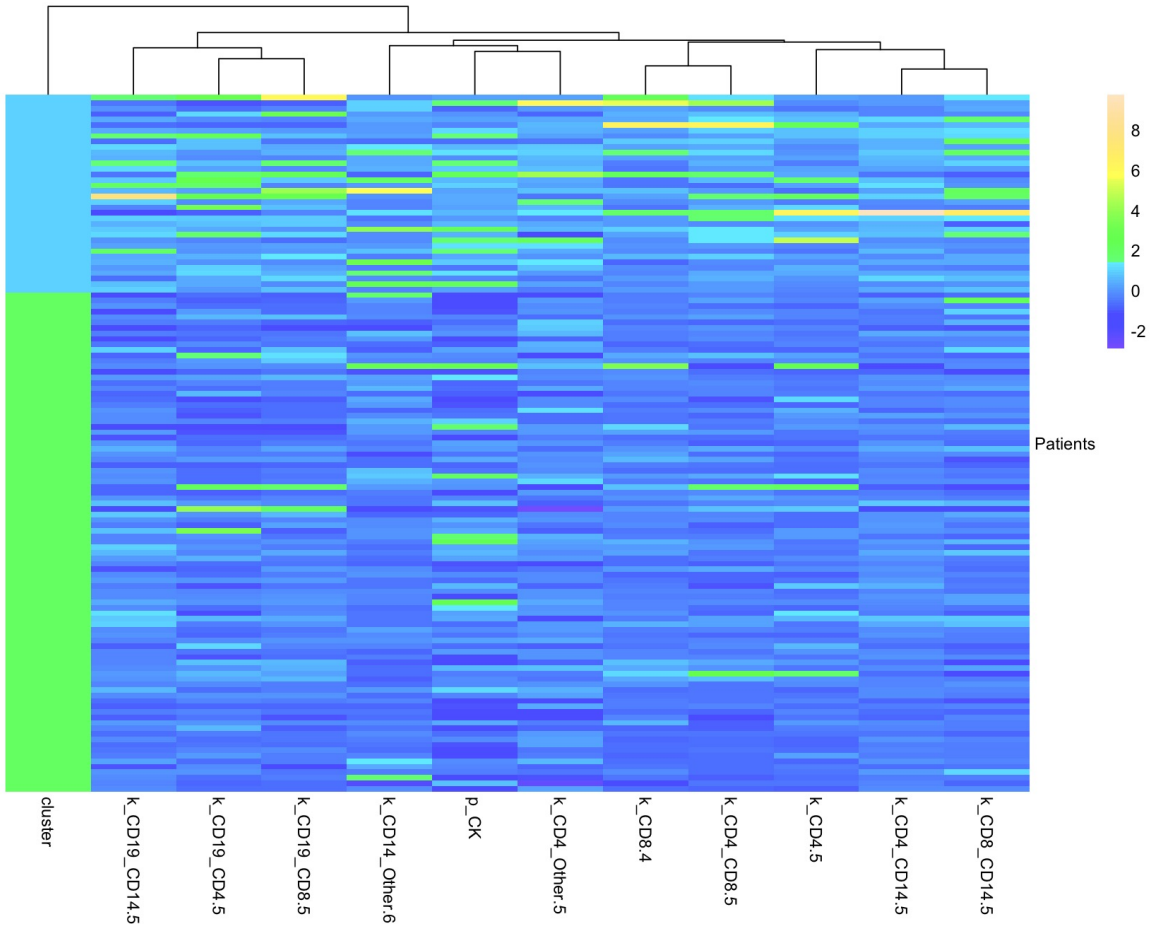
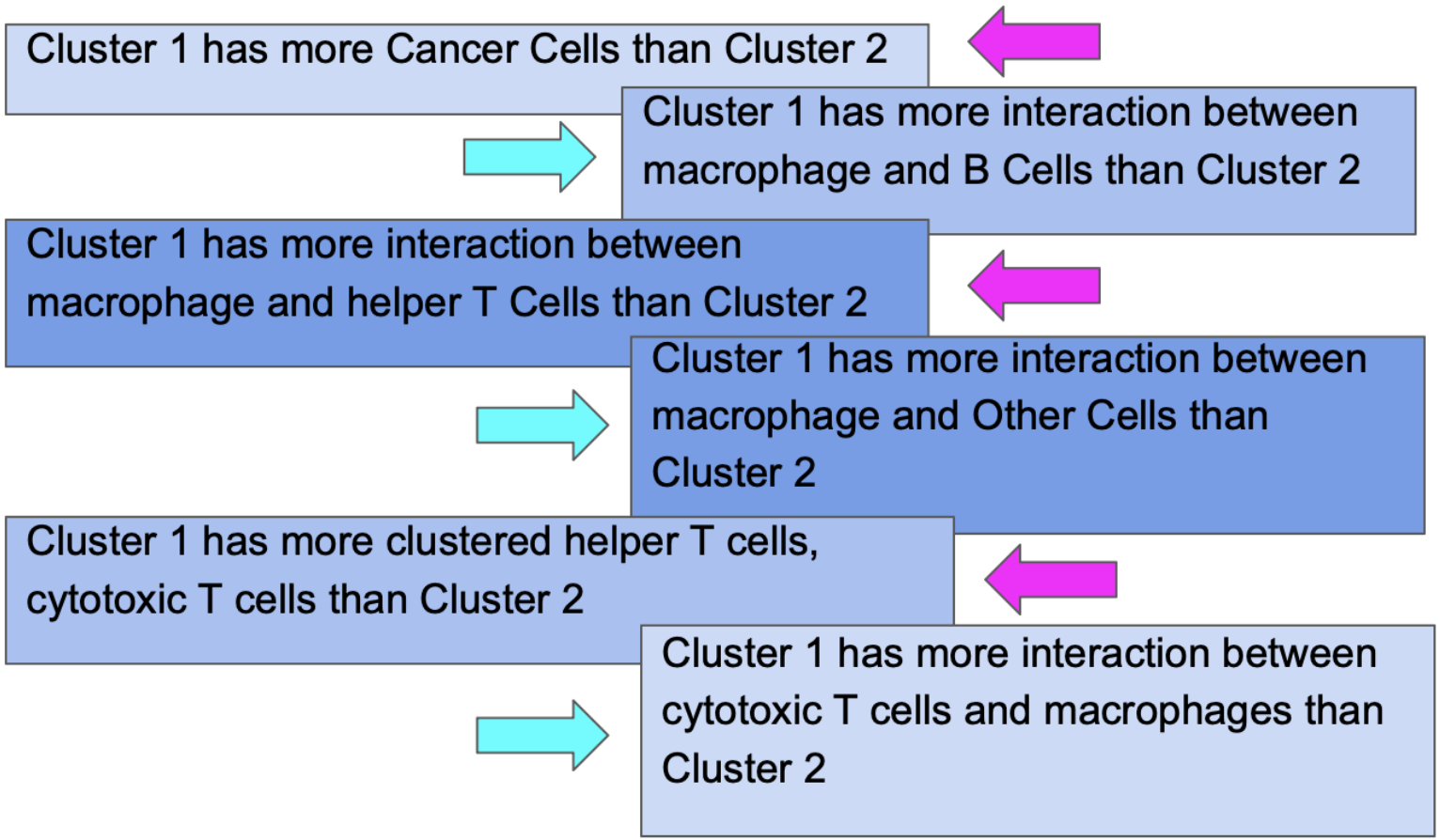
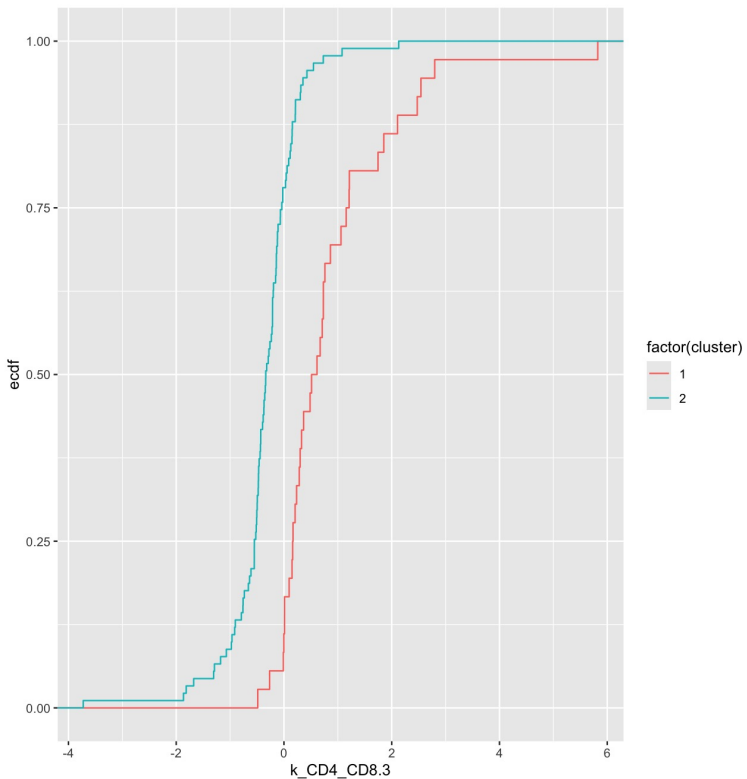


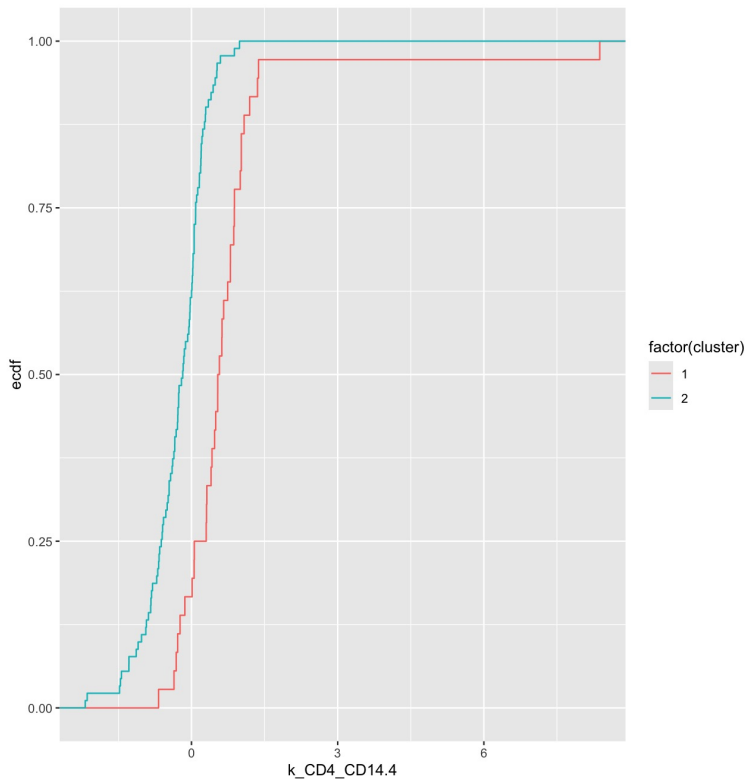
Figure: CD14 (macrophages), CD4 (helper T cells), CD8 (cytotoxic T cells), CD19 (B cells), CK (Cancer Cells) and Other Cell Types

At a glance

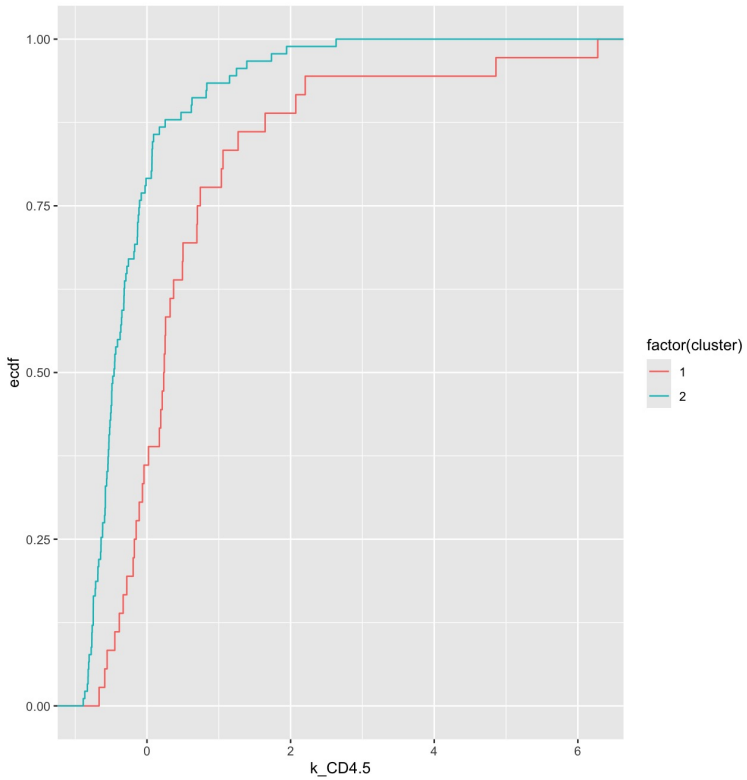


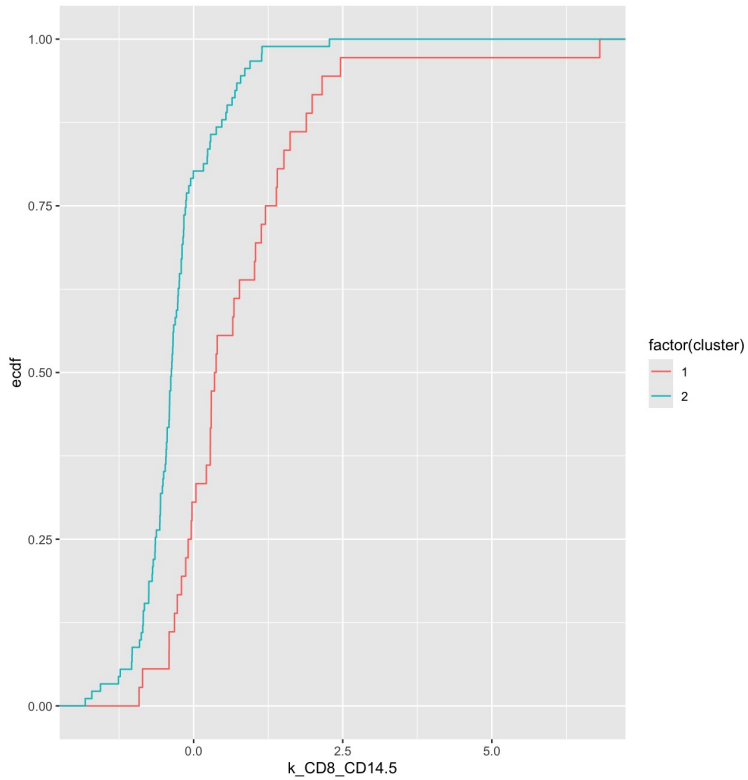
Estimated CDFs of the features across clusters





Estimated CDFs of the features across clusters

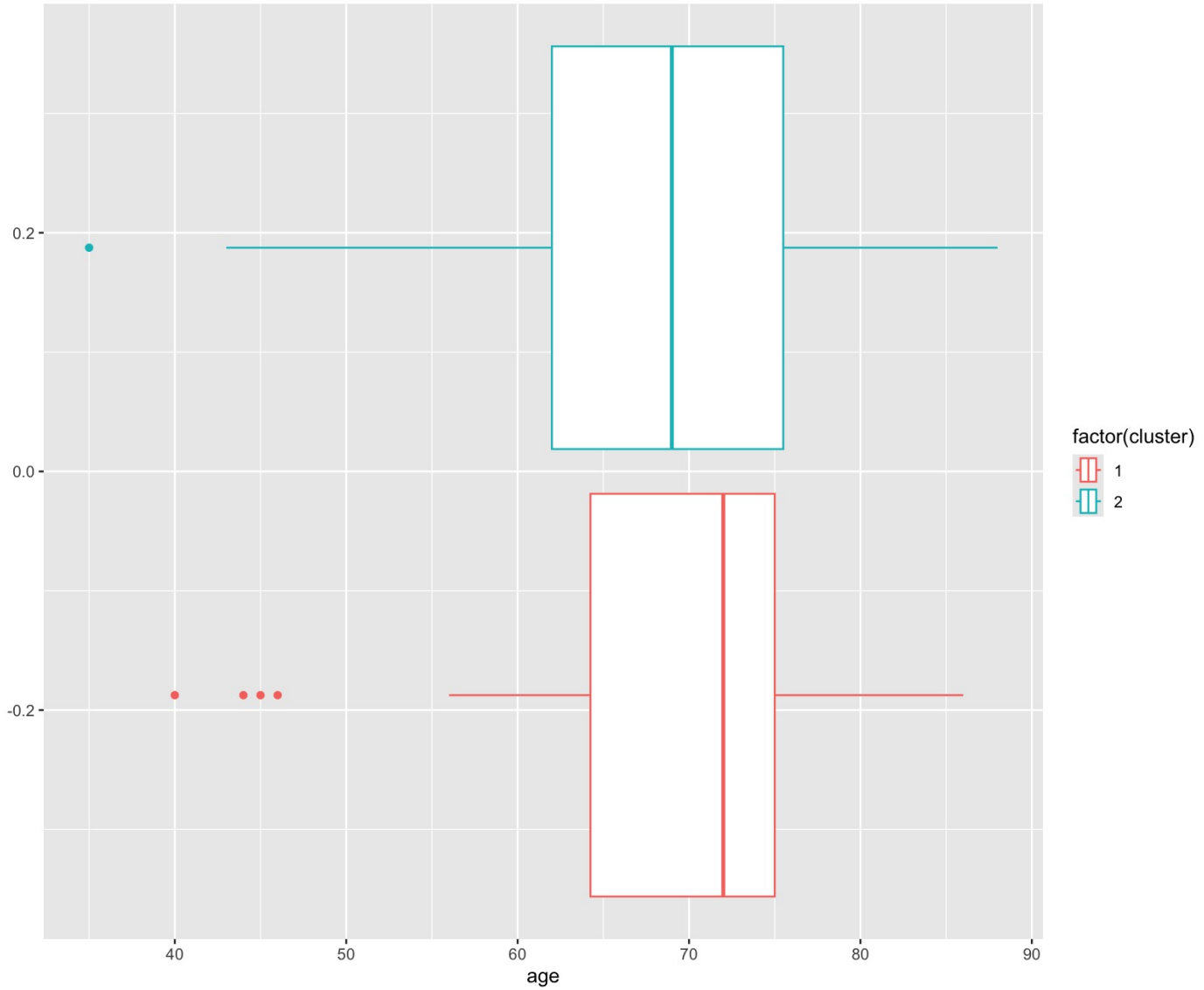




Clusterwise demographics

Cluster 1	Cluster 2
55.56% in 1st stage, 19.44% in 2nd stage, 22.22% in 3rd stage and 2.78% in 4th stage	63.74% in 1st stage, 21.98% in 2nd stage, 9.89% in 3rd stage, 4.39% in 4th stage
55.56% are male and 44.44% are female	49.45% are male and 50.55% are female
91.67% didn't receive therapy	83.52% didn't receive therapy

Comparison of Age



Survival Plot

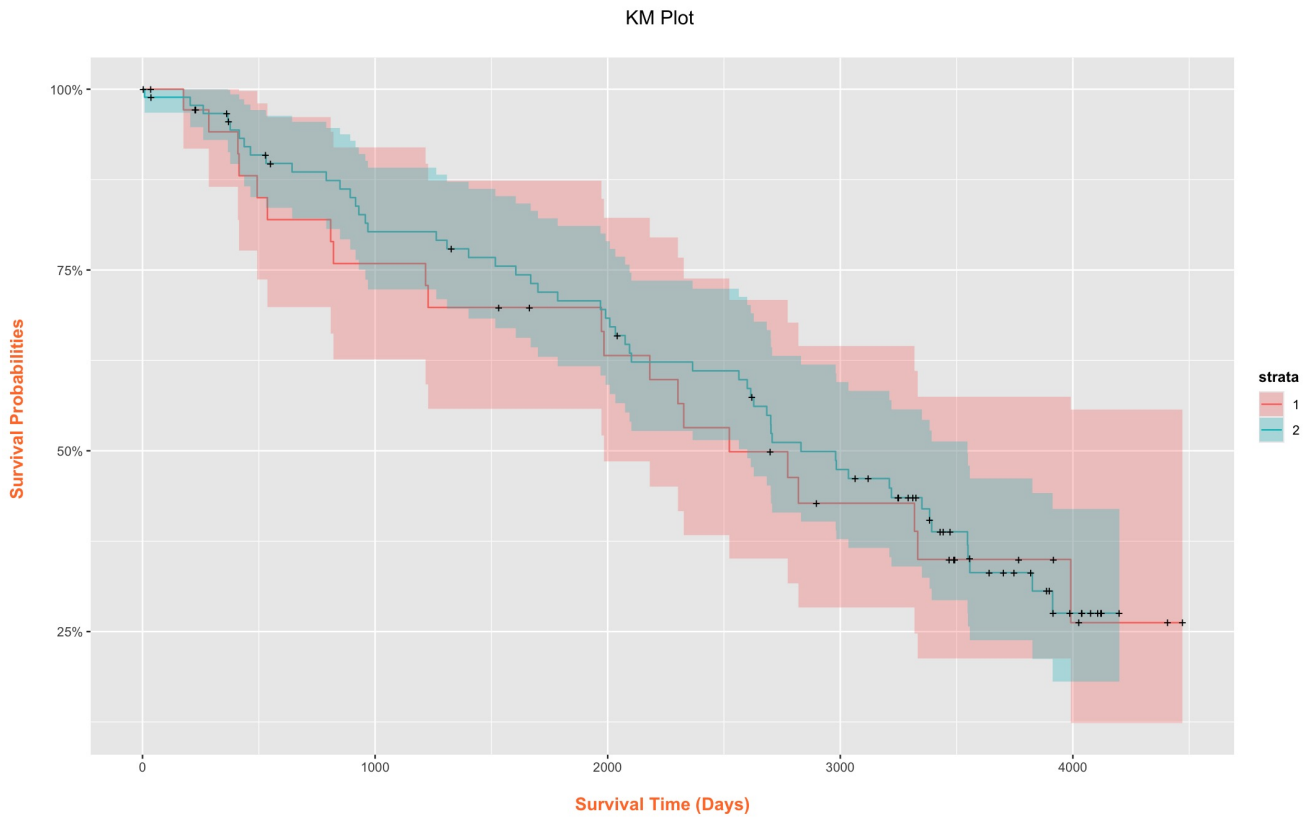
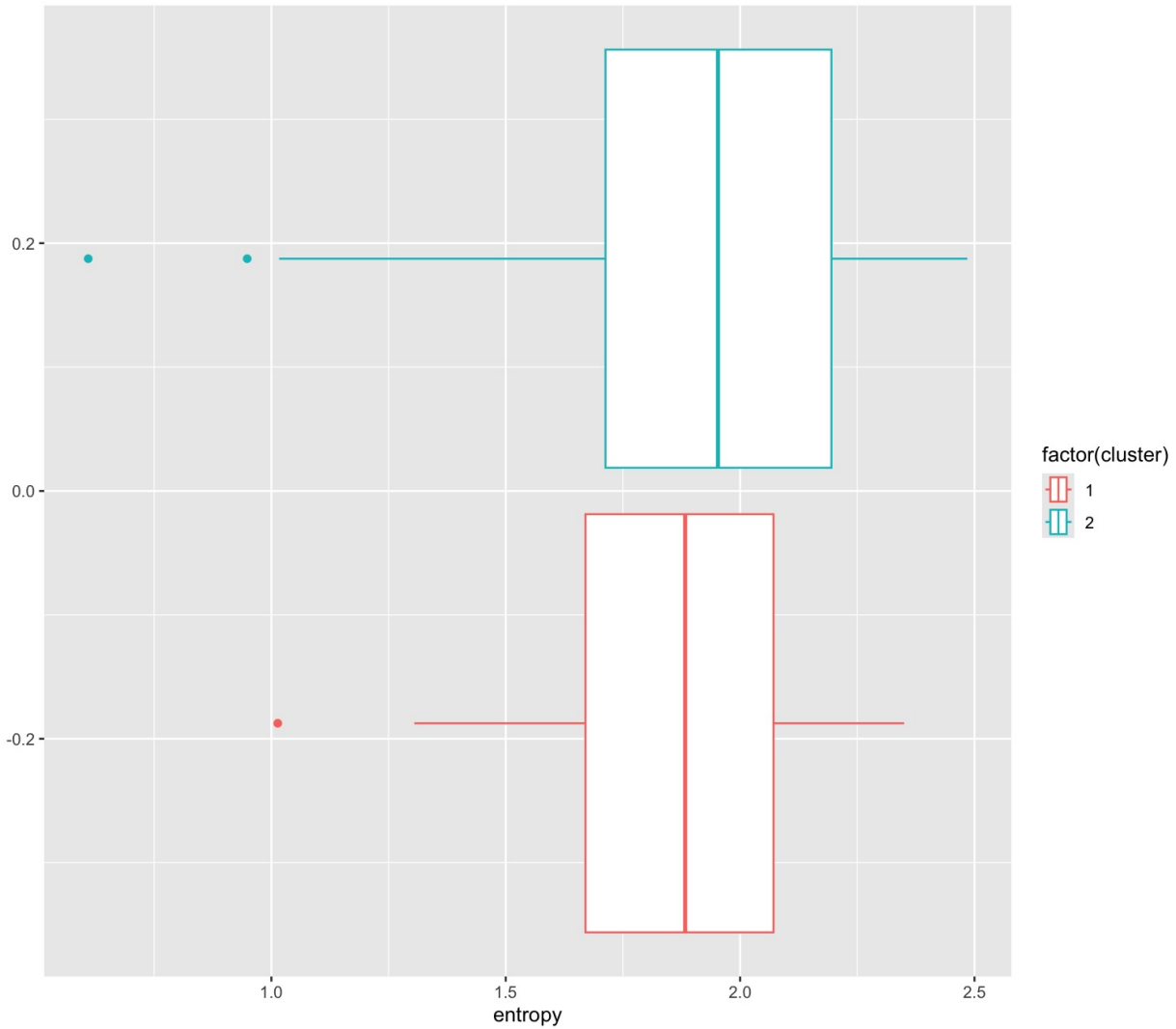


Figure: Clusterwise Survival Probabilities over Time

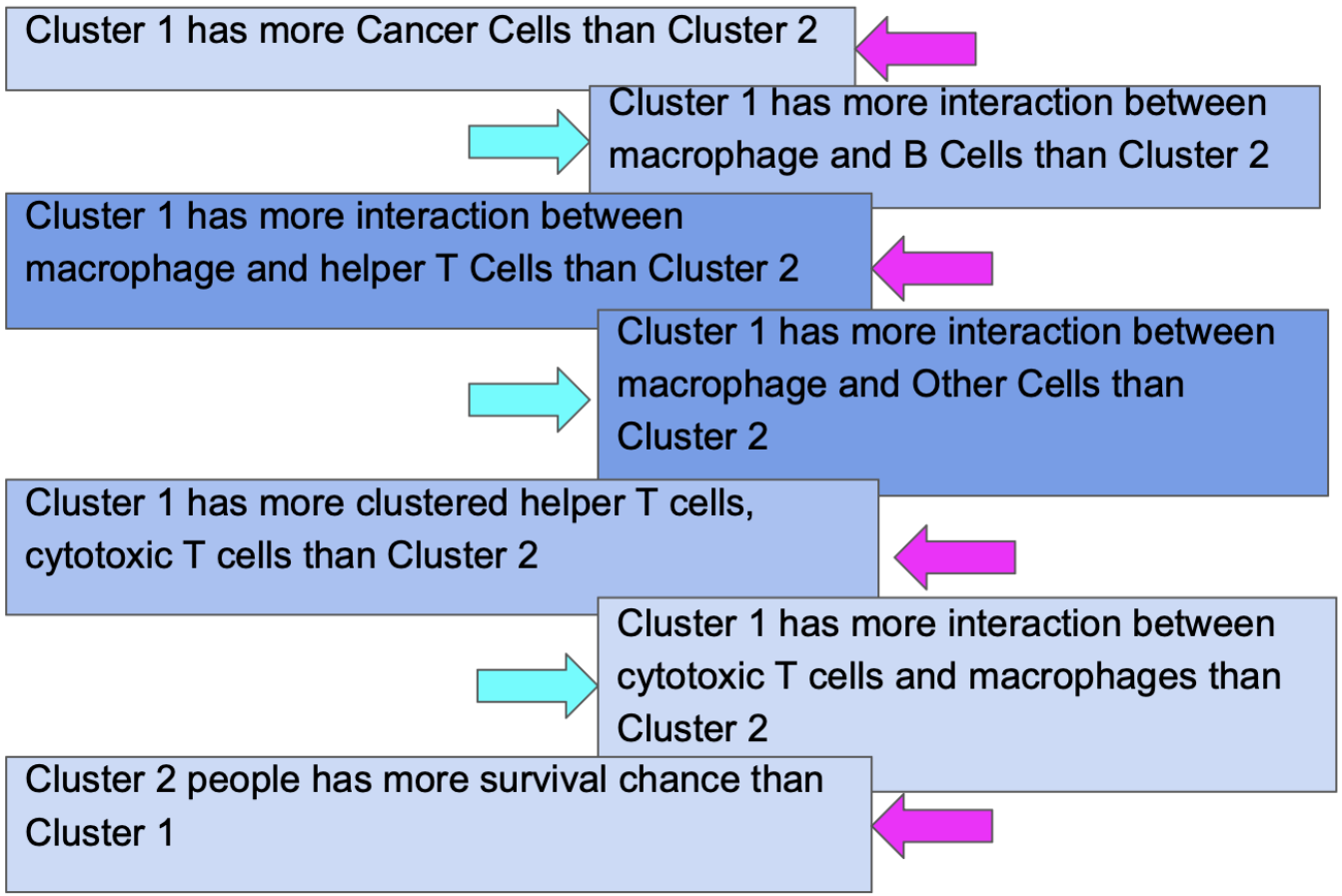
Comparing Cell type Proportions

- To compare the proportions of the cell types in images of two clusters we find the entropy of the cell-type proportion of each of the images.
- So for an image if the proportions of the 6 cell types are $p_1, \cdots, p_6$, we find $H = -\sum_{i=1}^{6} p_i \cdot \log_2 p_i$
- We compare the distribution of this H between two clusters.

Comparing Entropy - The Boxplot



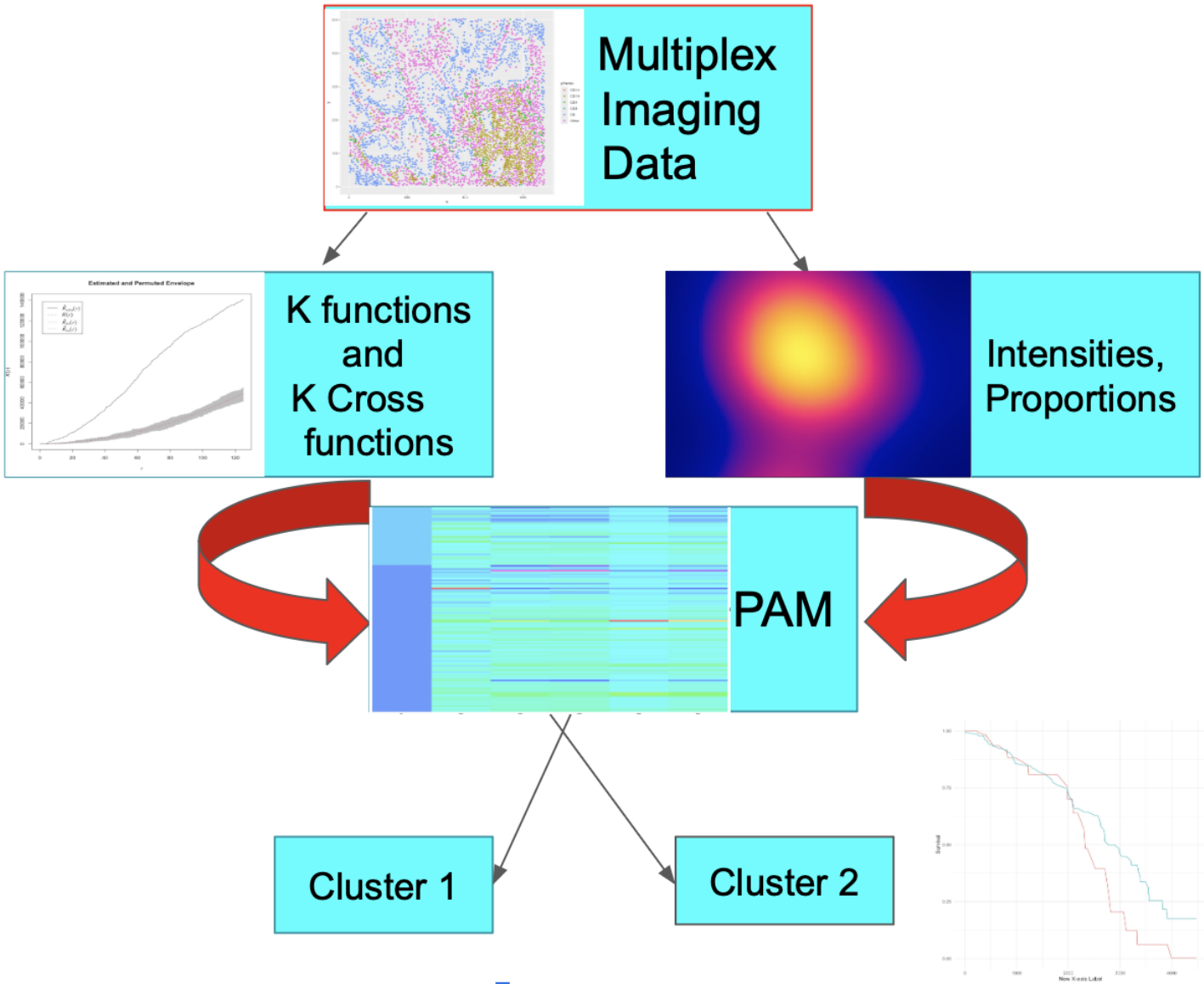
Conclusion



Limitations and Future Work

- In this study we extract the spatial information from the multiplex images using $K-$Functions and we can capture the joint effect of only two cell types (using $K-$Cross Function). So **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.
- The sample size (i.e. number of patients) is too small.
- A **Sensitivity analysis of chosing deciles** (instead of more detailed quantile levels) can be done to choose the most optimal cutpoints to capture maximum information.

Summary of the Whole Work



References

Masotti, Maria and Osher, Nathaniel and Eliason, Joel and Rao, Arvind and Baladandayuthapani, Veerabhadran

DIMPLE: An R package to quantify, visualize, and model spatial cellular interactions from multiplex imaging with distance matrices

Patterns, 4,12,2023,Elsevier.

Shariq Mohammed, Maria Masotti, Nathaniel Osher, Satwik Acharyya, Veerabhadran Baladandayuthapani

Statistical Modeling of Quantitative Medical Imaging Data

UNIVERSITY OF MICHIGAN

Thank You