Statistical Methods for Exploring the Arrangement of Myelin Basic Protein in Human Visual Cortex

Supervisor: Pratheepa Jeganathan

Student: Hainan Xu

1 Introduction

Spatial transcriptomics data have revolutionized the way of characterizing biological tissues at molecular levels. Benefited from the recent technological advancements in high-throughput sequencing and the development of cellular imaging techniques, we can obtain myriad spatial information on multiple aspects of gene expressions. However, few tools have been developed to analyze and summarize spatial point pattern data from multiple views. If we can use spatial variation methods to uncover the common patterns in the views, then we can integrate multiple views for downstream analysis. We propose to use first and second-order summaries of spatial point data to integrate multiple views.

Myelin is a layer of multi-protein structure that that insulates around axons Siu et al. (2015). Recent contributions of Murphy et al. (2020) indicate that myelin might be associated with adult cortical plasticity and aging. One of the major structural components of myelin is myelin basic protein (MBP), and it is expressed by the oligodendrocytes in the central nervous system. Analyzing the spatial arrangement of oligodendrocytes is important for us to uncover the spatial pattern of myelin basic protein, and thus help us understand the variation of cortical plasticity across one's lifespan.

To investigate the spatial pattern of the gene expressions of Myelin basic protein within human visual cortex samples, acquired from the Allen Institute of Brain Science, which consists of multiple specimens of visual cortex with different racial and age backgrounds. Each of the specimens is an ISH image that preserves the information of multiple gene expressions of the human primary visual cortex V1, which includes the cell information and the spatial coordinates of oligodendrocytes. A well-designed data infrastructure, SptialExperiment Righelli et al. (2022), can be used to integrate and analyze the molecular and spatial information.

Until now, a limited number of tools have been developed to integrate spatial statistics into multiple view analysis of spatial omics Dries et al. (2021). More tools and workflows need to be developed to store, retrieve and analyze the multi-view data. In this project, we propose spatial point pattern summaries to characterize multiple views of the human visual cortex (V1) and the spatial arrangement of points.

Typically, there are three types of spatial data: geostatistical data, lattice data, and point patterns Cressie (2015). There are two different kinds of point processes: the marked point process and the unmarked point process. An unmarked process considers only a collection of spatial coordinates as a random variable. A marked point process is referred to as a point process that has univariate or multivariate variables observed on points. Here we consider modelling the MBP and oligodendrocytes data by converting them into a point pattern derived from the point process under the random field assumption. At this stage, we focus on the unmarked point process. It's worth noticing that many features of the myelin basic protein expressed cells, such as cell size and shape, can be further investigated.

Through the model, we are able to know the distribution of the point pattern by considering first-order and second-order effects Holmes and Huber (2019) Schabenberger and Gotway (2017). After the image data is converted to a list of points coordinates, we are able to convert them to a point pattern and perform statistical point pattern analysis using R/spatstat package in R. The points in the point pattern

are represented by their Cartesian coordinates, and the data with their corresponding images can be saved into R/SpatialExperiment. The analysis methods include the G function, K function, L function, paired correlation function, etc Schabenberger and Gotway (2017).

The dataset is constituted of ISH images of 8 people of different sexes, races and ages, as shown in the following table. For each person, there are 3 to 4 ISH brain images that are stacked in order. Each ISH image includes spatial information and cell features of myelin basic protein, which is a key component of neuroplasticity. Sams (2021) suggests that oligodendrocytes have a reduced capacity to produce and maintain healthy myelin sheaths. Currently, there is limited information about how oligodendrocytes vary across the human lifespan. Recent research shows that there are drugs that boost myelin synthesis to promote a healthy aging brain Kremer et al. (9 01). Therefore, understanding the spatial pattern of myelin basic protein within the human visual cortex is helpful in understanding the dynamics of oligodendrocytes in the brain, and provides insights into the development of drugs that reduce the signs of aging. In ISH samples, the oligodendrocytes are characterized by having dark round, oval or irregularly shaped nuclei. It is not uncommon to see two or three oligodendrocytes colocate at similar positions, which is called doublet and triplet, respectively. Numerous oligodendrocytes exist within different layers of the human visual cortex, Siu and Murphy (2018), and thus the datasets can be extensive and challenging. In this report, we consider multiple views of the visual cortex of a single person.

Donor ID	number of slices	Age	Sex	Race
H07-0050	4	24	Male	White
H08-0010	3	42	Male	Black
H08-0001	4	40	Female	White
H08-0023	4	21	Female	Black

Table 1: Dataset information

To process the large-scale data and generate statistical summaries, we used the package spatstat in R. After that, the processed data information, spatial coordinates, and images can all be integrated into the S4 object, SpatialExperiment, using the package SpatialExperiment. Both packages are available using R/ Bioconductor. We use two different tools to extract the cell features and spatial information from Image data. The first one is the EBImage package in R, which provides tools to segment cells and extract various cell features, as well as generate analytical results for high-throughput cellular assays. (Holmes and Huber (2019)). Another tool is a software called Pipsqueak pro, which implements artificial intelligence engines to perform automatic cell detection for biomedical images. By first generating the training set by including the target cells as the region of interest(ROI) in a rectangular or oval frame, and setting up the background information, the semi-supervised custom AI model automatically frames the detected cells with a rectangular or oval-shaped region of interest, and return information table including spatial coordinates of ROI, cell count, ROI area(in pixel), background intensity, and statistics of image intensity. In this project, we use Pipsqueak to extract the spatial coordinates of target cells, use R/EBIimage to affirm the validity of the data, and implement R/spatstat and R/SpatialExperiment to perform the analysis.

In this project, we introduce the overall workflow and describe several statistical methods in section 2. Then, in section 3, we apply the methods to summarize the spatial arrangement of oligodendrocytes within the human visual cortex (V1) and interpret their results. We also introduce a threshold method to segment the different layers of the primary visual cortex. Finally, we discuss the possible extension of our work to integrate replicated multiple views of spatial point patterns.

2 Methods

In this project, we are concerned with the spatial arrangement of myelin basic protein within the human visual cortex. We present a workflow from extracting and storing spatial data, to analyzing and summarizing the spatial patterns. The procedures for data preprocessing are shown in 1. Starting from an ISH(In Situ

Hybridization) image of primary visual cortex specimen(V1), we mark out a rectangular observation window W that contains different layers of V1 through Pipsqueak. Next, as shown in fig1.C, we crop the W into 42 bins, and feed into the semi-supervised AI model in Pipsqueak to identify the cell features and spatial information. Finally, we integrate the datasets and convert them into a 2-dimensional point pattern object in R, which contains 3604 observations with mark size of each observation.

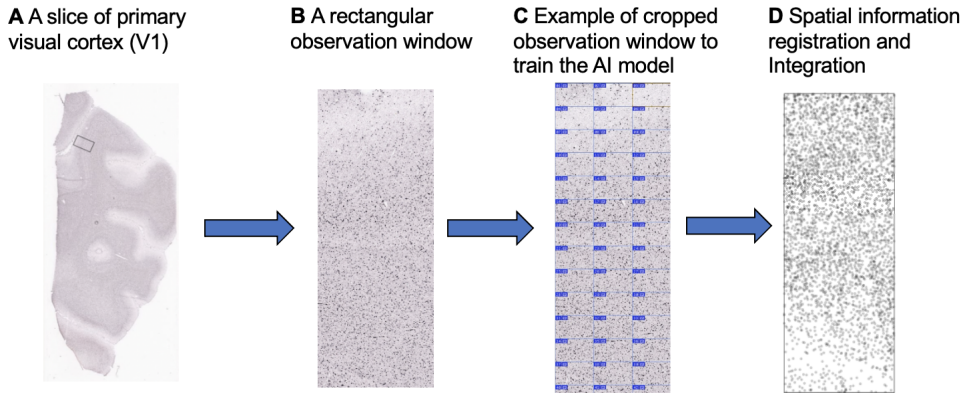

Figure 1: Data preprocessing

To have an overall understanding to the distribution of point pattern $Z(s)$, we first generate the scatter plot that displays the spatial information of each event. Then, based of the scatter plot, we plot histograms of marginal distributions of events along the x-axis and the y-axis.
Under the assumption of a random field, we consider the point pattern $\{Z(s) : s \in D \subset \mathbb{R}^2\}$ is a realization of the Poisson process. We estimate the first-order intensities and second-order intensities, compute the G functions, K functions and paired correlation functions to observe the arrangement and inter-event correlations of oligodendrocytes. We also implement the Monte Carlo method and generate simulation envelopes for each analysis function to guarantee the significance of the results.
After analyzing $Z(s)$ as a whole, we subset the observation window W into 4 non-overlapping windows based on the kernel density estimation using a threshold method. Then, we test for complete spatial randomness (CSR) for point patterns for each of the 4 subset windows to see if we can make any conclusion about the spatial patterns of each layer.

2.1 Quadrat count

The quadrat count method is a rectangular tessellation method of Z(s). To perform quadrat count, we subset the rectangular observation window into $r \times c$ blocks, then count the number of events that occurred within each block. We then perform a two-sided Chi-square test to check if the point pattern is a realization of the homogeneous Poisson process Schabenberger and Gotway (2017). We take the significance level at 0.05. If the computed p-value is smaller than 0.05, we reject the null hypothesis. It indicates that the number of events is not uniform within each block, and therefore the point pattern is not a realization of a homogeneous Poisson process. If p-value is higher than 0.05, we fail to reject the null hypothesis. It indicates that we do not have sufficient information to conclude the underlying point process is inhomogeneous.

2.2 First-order properties

The first order intensity of a point process $Z(s)$ measures the number of points within a unit disc centred at point **s**, which is defined as

$$\lambda(s) = \lim_{v(ds)\to 0} \frac{E[N(ds)]}{v(ds)},$$

where $N(ds)$ is a random variable that counts the number of events within the infinitesimal disc ds centred at s $\lambda(s)$ Schabenberger and Gotway (2017). If $\lambda(\mathbf{s}) = \lambda$ throughout the random domain D, the point process is a homogeneous process. Furthermore, if the point process is a Poisson process with constant first order intensity, it is a homogeneous Poisson process (HPP), and it preserves complete spatial randomness(CSR). In order to remove the bias caused by the edge effect, we define the applicable domain W as a 1500*1700 bounding rectangle. We apply edge correction for the points near the boundary of W, whose distance to the boundary of W compared is smaller than the disk radius ds. For these points, the intensity value at point s is estimated as

$$\lambda(s) = \sum_{i=1}^{n} K_h(s - s_i)e(s_i),$$

where k is the Gaussian smoothing kernel, $e(s_i)$ is an edge correction factor that takes the points that fall out of the W into account due to the application of the smoothing kernel.
In this analysis, we use the homogeneous Poisson process(HPP) as a null model to perform hypothesis testing to determine whether the observed point patterns have complete spatial randomness. If the null hypothesis is rejected, $Z(s)$ is not a realization of a homogeneous Poisson process. In that case, the point pattern $Z(s)$ may reveal some clustering or inhibition pattern.

2.3 Nearest neighbour distance distribution

After getting the intensity of the point pattern, we apply the nearest neighbour distance function to test for its departure of complete spatial randomness. The nearest neighbour distance function is a cumulative distribution function G that calculates the distance of a randomly chosen point s to its nearest other points. We calculate the empirical G and compare it with theoretical G derived from a Poisson point process, which is

$$\hat{G}(r) = \frac{1}{n} \sum_{i=1}^{n} I(h_i \leq h),$$

where h_i denotes the distance from event $\mathbf{s}_i$ to the nearest event, and let $I(h_i \leq h)$ be an indicator function which returns 1 if the condition is satisfied.
After getting the empirical $\hat{G}_0(r)$ function based on the observed point pattern Z(s), we then generated 200 Monte Carlo simulation that is conditioned on the number of events under the CSR hypothesis. The upper and lower simulations are given by

$$\hat{G}_l(r) = \min_{i=1,\ldots,200} \{\hat{G}_i(h)\}$$

and

$$\hat{G}_u(r) = \max_{i=1,\ldots,200} \{\hat{G}_i(h)\}.$$

If $\hat{G}_0(r)$ lies within the envelope bounded by the lower and upper simulations, we fail to reject the null hypothesis. $\hat{G}_0(r)$ is a homogeneous Poisson process. If $\hat{G}_0(r)$ falls outside of the envelope at some distance, $Z(s)$ may have some cluster or regular patterns.

2.4 Second-order properties

Second-order intensity takes account of the interaction between events.It gives the relationship of the expected number of the cross product of the event counts within an infinity small disks, which is given as

$$\lambda_2(s_i, s_j) = \lim_{|d\mathbf{S}_i|\to 0, |d\mathbf{S}_j|\to 0} \frac{E[N(d\mathbf{s}_i N(d\mathbf{s}_j)]}{|d\mathbf{s}_i||d\mathbf{s}_j|}.$$

We use Ripley's K function to analyze the second-order property of $Z(s)$, where K(h) is given as

$$K(h) = 2\pi\lambda^2 \int_o^h x\lambda_2(x)dx.$$

The Ripley's K function is a second-moment measure. $\lambda K(h)$ represents for the expected number of additional points within disk with radius h centred at a randomly picked event. To avoid the quadratic behaviour of K(h), we recommend applying a square root transformation on K, which is known as the L function:

$$L(h) = \sqrt{\frac{K(h)}{\pi}}.$$

In extension of $K(h)$, paired correlation function($g(h)$) is also a widely used method to study the dependence in point patterns. It has the following relationships with $K(h)$:

$$g(h) = \frac{1}{2n\pi}\frac{dK(h)}{dh}.$$

3 Results

3.1 Quadrat Count

In 2, each grey dot represents an event, and the radius corresponds to its mark size. A histogram of the event count of the x-axis lies on the top of the scatter plot. The height of the histogram represents the frequencies of events within the corresponding x range. There is no distinct variation in event count along the x-axis. On the right-hand side of the scatter plot, we plot the histogram of event count on the y-axis. The height of the histogram represents the frequencies of events within the corresponding x range. We can see the histogram shows a multi-modal distribution. Also, the cell count varies with the change of y. By inspection, we make a rough guess that the point process is not a homogeneous Poisson process. In addition, after checking duplicates, there are six pairs of points with the same spatial information but with different mark sizes. It is possible that each pair of points is doublet[1].

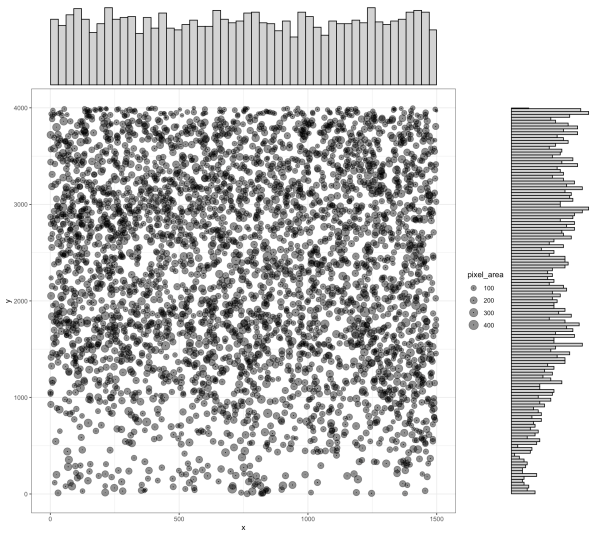

Figure 2: Scatter plot and histograms

[1]The doublet means a pair of oligodendrocytes colocate within a very small distance.

Next, we perform a quadrat count to test whether the process is homogeneous 3**A**. We divide the whole observation window into 36 blocks (3 columns and 13 rows), compute the event counts within each block, and perform a Chi-square test. The p-values is less than 0.05, indicating the intensity is non-constant. Therefore, we reject the null hypothesis. We have enough evidence to show that the point pattern is not a realization of homogeneous process.

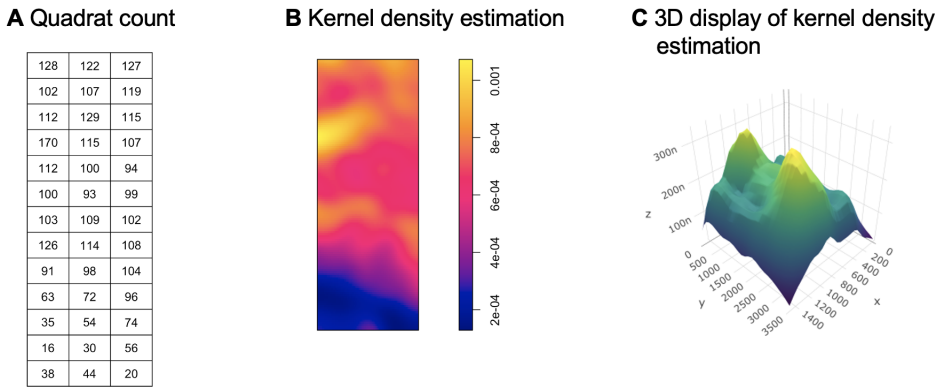

Figure 3: Visualizations of quadrat count and kernel density estimation

3.2 Intensity estimation

We compute first-order intensity with edge correction as shown in 3**B**. The kernel density plot demonstrates the variation of intensity within observation window W. Again, this strengthened our conclusion that the entire point process is not homogeneous. In addition, a trapezoidal area at the bottom of the observation window has a very low intensity. In the later section, we consider segmenting out the areas with similar intensities. We also generated a 3-D visualization of the kernel density plot using R/plotly.

3.3 G function

The observed G function is denoted as $\hat{G}_{obs}$) of the oligodendrocytes dataset shows the spatial arrangement of myelin basic protein within one slice of the visual cortex sample. The G_{theo} denotes the theoretical G function that derives from a complete spatial random point pattern within W. We plot the observed G function against the theoretical G function to see if there are differences between the oligodendrocytes' pattern and a complete random point pattern. To make the result significant, we performed 200 Monte-Carlo simulations of the random Poisson process with the same total number of oligodendrocytes and density as the dataset to generate a significance envelope. Any point pattern with $\hat{G}_{obs}$ lies outside of the significance envelope are considered non-random. 4**A** shows that $\hat{G}_{obs}$ lies outside of the envelope in many distances. Overall, Z(s) is not complete spatial random. 4**B** shows that $\hat{G}_{obs}$ is larger than G_{theo} for $r < 2$, indicating a possible clustering pattern. Some oligodendrocytes have the tendency to colocate within a very small distance. For $r > 2$, $\hat{G}_{obs}$ deviates from G_{theo} quickly. $\hat{G}_{obs} \leq G_{theo}$ is indicative of a tendency of inhibition. When $r > 25$, $\hat{G}_{obs}$ falls within the envelope. The oligodendrocytes tend to be independent from each other.

3.4 K function, L function and paired correlation function

In order to fully characterize the point pattern $Z(s)$, and to learn about inter-event dependency, we implement the K function. The K function estimates the additional number of points of a randomly picked reference point within a specific range r. 5**A** shows the estimated K for Z(s). $\hat{K}^{obs}$ and K^{theo} are plotted against each other. Similar to G functions, $\hat{K}^{obs}$ and K^{theo} crosses around r =2, indicating a strong inter-event

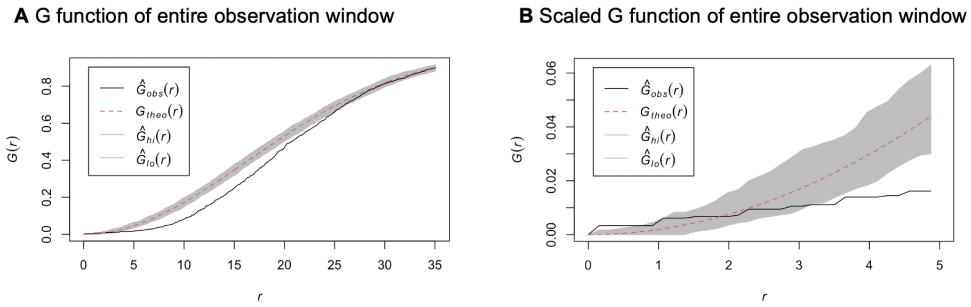

Figure 4: Visualization of G function

dependency for $r < 2$ and a weak inter-event dependency for $r > 2$.
With the increment of r for K function, the variance of K function can be very unstable, therefore we transform K into L(r) to get a better observation. The L function in 5**B** shows a similar pattern with K.
We also compute the paired correlation plot 6. $g_{pois}(r)$ is the theoretical paired correlation generated from the homogeneous Poisson process, which is a constant number 1. There is a strong paired correlation for distance around 1, and then it falls under $g_{pois}(r)$. This means the events cluster at a distances smaller than 2,and then have an inhibition pattern for distances smaller than 12. When the distance is greater than 2, the g(r) fluctuates around 1.

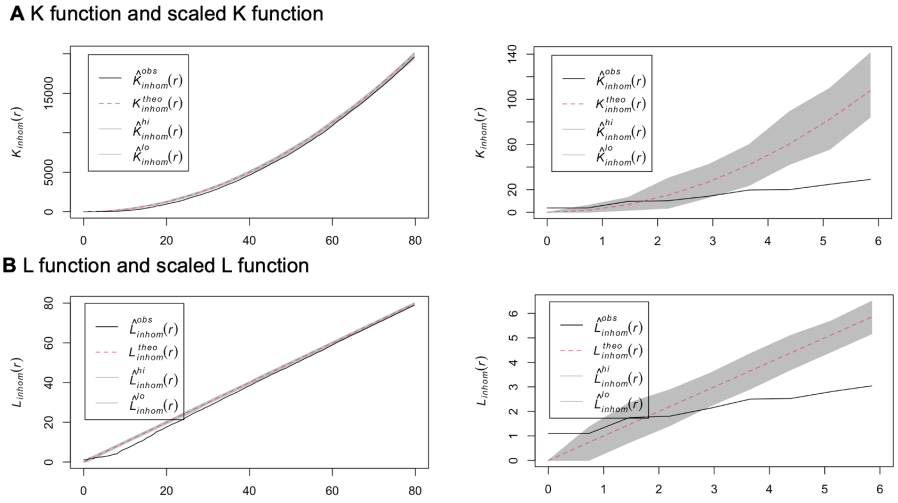

Figure 5: Visualization of K function and L function

3.5 Segmentation and test for CSR for segmented area

We know that the point pattern in the entire observation window is not complete spatial randomness. Given the fact that the primary visual cortex has a layered structure, we roughly segment out the entire observation window into four layers(regions) using the histogram generated from the kernel density estimation plot. Each region is denoted by a color. Region A in the histogram corresponds to the region with intensity under 3e-04, which roughly corresponds to the white matter layer; Region B denotes the region with intensity between 3e-04 and 6e-04; Region C denotes the region with intensity between 6e-04 and 6.75e-04; Region D in the histogram denotes the region with intensity greater than 6.75e-04. We test for complete spatial randomness for each of the four regions.

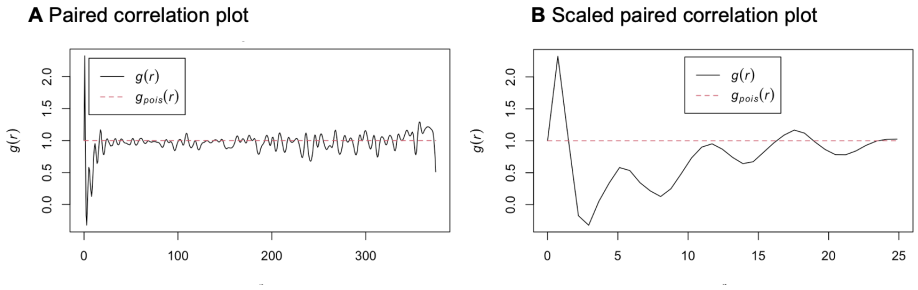

Figure 6: Visualization of paired correlation function

It turns out that region A is complete spatial random, indicating the white matter layer is a homogeneous Poisson process with constant λ, we can make a compact window for the other part of the point pattern.

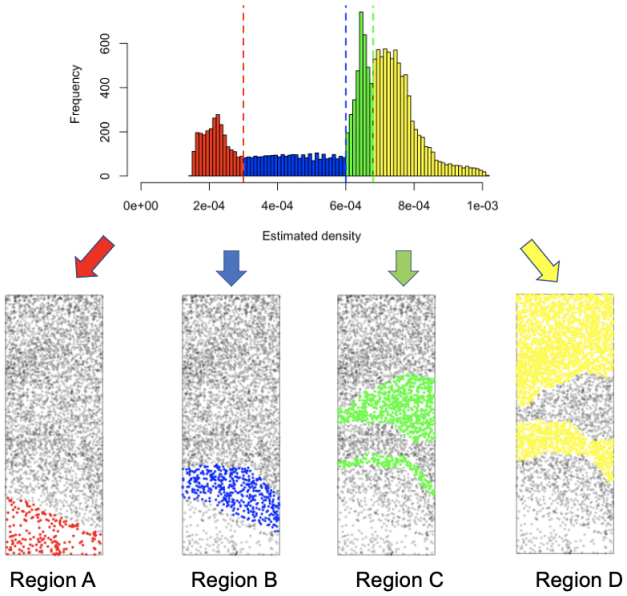

Figure 7: Segmentation

Region B is approximately CSR. To verify this, we compute the quadrat count within the compact window and perform a Chi-squared test of CSR with a significance level of 0.05. The p-value is approximately 0.023. Therefore, we fail to reject the null hypothesis that the point pattern is CSR.
From the G function, we know that the green part and yellow part are inhomogeneous Poisson processes. We may implement a more complex statistical model to simulate these two regions.
Besides, including the previous slice, we have four slices of human primary visual cortex specimen for Donor H07-0050. We perform a similar analysis for the slices as shown in the previous sections. And for each slice, region A is always CSR. For some of the slices, region B is also CSR. In addition, with different segmentation thresholds, there are CSR patterns among regions C and region D.

4 Discussion

In this project, we explored the spatial arrangement of myelin basic protein within human primary visual cortex using various statistical methods. We estimated the first-order intensity, computed and analyzed the nearest neighbour distribution, Ripley's K function, L function and paired correlation function. We talked about the different characteristics for these statistical methods. Then, we segment different layers of the human visual cortex from a data-driven perspective, which roughly corresponds to the actual biological layers. By testing for spatial randomness for the segmented regions, we were able to make a compact window only for the patterns that are not CSR, and thus improved the efficiency of computing. Meanwhile, the segmentation method is very sensitive to the threshold we chose. With the advent of more information on cell features, a better segmentation method can be developed using cluster analysis.

To ensure the accuracy of the data generated from the AI model and prevent artificial error when combining the datasets, we generated the dataset the second time using image processing Holmes and Huber (2019). By comparing the kernel density estimation plot generated from both datasets, we verified the correctness of the combined dataset.

We also tried to compute the Delaunay network for Z(s). However, the network only connects for region A of each specimen. This might be due to the excess of observations. Faster algorithms are needed for large-scale data.

Due to the length of the project time, only the data from donor H07-0050 were extracted and analyzed. In the future, with more data being processed and more features being extracted, it is possible to perform a longitudinal analysis by implementing the workflow proposed in this project.

5 Acknowledgement

I am particularly grateful to my supervisor Dr.Pratheepa Jeganathan for her detailed guidance. I also want to thank Ahad Daudi at the Visual Neuroscience Lab for training the AI model and providing the dataset. During this summer research period, I learnt about image processing and spatial statistics from scratch. We prepared a research schedule including reading materials on spatial analysis, learning about the usage of different R packages (SpatialExperiment, spatstat, EBImage), generating homogeneous and inhomogeneous point processes, reading literature reviews on advanced transcriptomics analysis, etc.

I have extensively used R, Github, Overleaf, Python, Atom and Zotero. My presentation skills were improved by doing presentations within the research group. In May, I attended the SSC annual conference and did a poster presentation. At the end of July, I participated Bioconductor conference and gave a lightning talk on my current research topic.

In the weekly meeting within the research group, I learned about high-performance computing. I also collaborated with students from other research groups and built connections with them.

References

Cressie, N. (2015). *Statistics for Spatial Data*. John Wiley & Sons. Google-Books-ID: MzN_BwAAQBAJ.

Dries, R., J. Chen, N. d. Rossi, M. M. Khan, A. Sistig, and G.-C. Yuan (2021). Advances in spatial transcriptomic data analysis. *31*(10), 1706–1718. Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab.

Holmes, S. and W. Huber (2019). *Modern Statistics for Modern Biology*. Cambridge University Press.

Kremer, D., R. Akkermann, P. Küry, and R. Dutta (2019-01). Current advancements in promoting remyelination in multiple sclerosis. *25*(1), 7–14.

Murphy, K. M., S. J. Mancini, K. V. Clayworth, K. Arbabi, and S. Beshara (2020). Experience-dependent changes in myelin basic protein expression in adult visual and somatosensory cortex. *14*.

Righelli, D., L. M. Weber, H. L. Crowell, B. Pardo, L. Collado-Torres, S. Ghazanfar, A. T. L. Lun, S. C. Hicks, and D. Risso (2022). SpatialExperiment: infrastructure for spatially-resolved transcriptomics data in r using bioconductor. *38*(11), 3128–3131.

Sams, E. (2021). Oligodendrocytes in the aging brain. *5*(3), NS20210008.

Schabenberger, O. and C. A. Gotway (2017). *Statistical Methods for Spatial Data Analysis: Texts in Statistical Science*. Chapman and Hall/CRC.

Siu, C. R., J. L. Balsor, D. G. Jones, and K. M. Murphy (2015). Classic and golli myelin basic protein have distinct developmental trajectories in human visual cortex. *9*.

Siu, C. R. and K. M. Murphy (2018). The development of human visual cortex and clinical implications. *10*, 25–36.