ASD/MF Project Notes

Ameet Rahane

March 3, 2019

1 Problem Set Up

1.1 What do we have?

We start with data from Wholebrain. What this looks like for one mouse is a matrix in $\mathbb{R}^{2,n}$, where n is the number of regions. This matrix consists of two vectors. They respectively stand for ipsilateral and contralateral hemispheres. Each vector contains cell counts for each region that we're looking at. Formally, we have

$$v \in \mathbb{R}^{2,n} \tag{1}$$

1.2 What do we want to find?

We want to find whether or not this mouse is male or female or if it has autism or not. Generalized, we want to find how the region counts are related to a given experimental condition. Since the outputs of these are yes or no questions, we can have it be an output. Thus, in essence, according to my understanding of this, at least, we want to find:

$$Tv = w, v \in \mathbb{R}^{2,n}, v \in \{0, 1\} \tag{2}$$

and

$$T : \mathbb{R}^{2,m} \to \{0, 1\} \tag{3}$$

That is, we're trying to approximate a transformation from the space of relative region counts to a yes or no output. As in, a general binary classification problem. A strong question does still lie in whether we want to approximate this function directly or not, or rather, do we want to exactly build a classifier to do this (no harm in trying right?).

2 Basic Exploratory Approaches

There are some approaches that data scientists generally take to look at the space of data given. We can start by simply building a correlation or covariance matrix. This will tell us the correlation between regions and probably, we could use it to find the correlations between hemispheres, as well.

Alternatively (or in conjunction), we could take a look at the space of data by principal component analysis or singular value decomposition. Either of these methods manipulate orthogonality and eigenvalue decomposition in order to find where the most variation is. If we choose to do something like this, we could probably use it as a dimensionality reduction technique, as well as just telling us something about the data. We can also use canonical correlaiton analysis, but this is very similar and would serve the same purpose.

3 More Advanced Approaches and Examples

We can use methods like LDA, random forests, or other binary classification techniques in order to directly find this classification map. The Cell paper did this as well.

3.1 Kim's (Cell) indirect, but more direct measure of which regions significantly matter

Let's call a region Y. The assumption they made is the following:

$$Y \sim -Binom(\mu) \tag{4}$$

μ here is the mean, but it's related to the experimental condition. Let's call this experimental condition X, in reality, this is the question we're trying to answer.

$$\mathbb{E}[Y] = \alpha + \beta X, \alpha, \beta \in \mathbb{R} \tag{5}$$

With this, we know that if β is statistically significant, then the region being tested is important to the experimental condition. We can use iterative linear least squares squares error to find the values of α, β with some confidence interval. They also use the Benjami-Hochberg procedure to reduce error, will need to look this up. Perhaps, if there are multiple regions here, we can find the joint distribution. Also, maybe this distribution becomes normal if it is significant? Not entirely sure about this.

3.2 Using KNN clustering to approximate this function

Changing distance function to be what we want.

4 Questions

I do still have a number of questions. In fact, I think this exercise has probably increased the number of questions that I have.

- How much do we care about the relationship between ipsilateral and contralateral?

  I know that we generally care about the projections, but do we need to look for the experimental relationship between the hemispheres?

- If so, is it important to first find its relationship or its joint distribution?
- Are we looking at a certain cell type? If so, what is this?
- What kind of labelling are we using?
- In the cell paper, why are they using a negative binomial distribution for the conditional distributions? I'm sure there is a reason, but it's not clear.

5 Takeaway and starting point

Overall, it seems like the Cell paper actually uses a similar technique to the one I suggested, although I found it a bit circular. They did quite a bit of analysis on the data itself, before trying to find a mapping. This is actually very useful, density maps are nice and we should try to do the same.

I think a good starting point would be to start with PCA or a similar dimensionality reduction technique. Then, we can run LDA or some other linear classifier. After that, we can measure the effects pseudo-directly by the probabilistic model, they outlined. Although, I do still have a few questions about why they chose the distributions that they did.

Beyond that, we should do some region analytics too. That is, density maps and maybe trying to figure out the underlying probablity density distribution of the region data.

I'm not entirely sure how to incorporate ipsilateral vs. contralateral in this whole thing. Presumably, if we try to learn a map, we should see which matters more. Beyond that, the direct probabilistic modeling will give us a sense of which contributes more. We should treat ipsi-regions and contra-regions separately in the probabilistic sense, which would then give us the ability to compare how they look (lots of plots).