Contents

- 3 Different methods currently
- 1 NN (hard to interpret), 1 Bayesian (Cell-Segmentations), 1 density based (Segmentation-free)
- Use structure of SSAM but improve on methods in each section

1 Introduction

- Mention the larger project, what is this study going to be used for?
-
- Mention SSAM, Bayzor, Spage2Vec
-

2 Method and Theory

Should I mention SSAM? And then in each step say e.g. they use gaussian kernel with fixt bandwidth but we want an adaptive density estimation so we do...? Structure of the method:

1. Get genes from tissues as coordinates in 2-D (similar to GMM)
2. Pointwise density estimations
    - Tried one of the most known parameter free estimations (Kraskov et. al).
        - Cons: We don't have normalized density estimation or a density function. We use Rank. The rank only depends on the distance to the k:th nearest neighbor
        - Generalization: Use one of two estimates that make use of all neighbors $1, 2, ..., k$.

- Method 1: $\rho_1(i) = 1/\sum_{j=1}^{k} d_{ij}$, where d_{ij} is the euclidean distance from point i to point j and k is the number of nearest neighbors. We must fix k How do we decide on k?
- Method 2: $\rho_2(i) = d_i/vol(V)$ which is the stationary distribution in a similarity graph. Here $d_i = \sum_{j=1}^{n} \omega_{ij}$ where $\omega_{ij} = \frac{1}{d_{ij}^2+1}$ and $vol(V)$ is the sum of d_i. We only use ranks so $vol(V)$ does not contribute.

3. Some down sampling or local maximum finder How do we find the local maximums? Depends on normalization process.
    - Compared estimated mode and true mode from generated data. Check how well density estimators preserve spatial information of cluster modes I check how many neighbors away from the true cluster mode the estimated mode is.
    - Figure: Distribution of k from previous point shows peak close to 0 decaying quickly (good).
    - Further checks: Statistic from SASNE paper.
    - Method 1 seems to perform better (very slightly)
4. Check distribution of "True Mode" - "Estimated Mode", i.e. a plot around origo. Are the Estimated Modes biased? I.e. do they always lie to the left of the true mode. Or above? How about variance?
5. How does the gaussians generated look like? Does it look good? Bias? Variance?
6. Normalization of gene count in maximum
    - Standard is sctransform.
    - Cons: It makes a lot of assumption that do not necessarily hold. e.g. linearity between gene count and sequencing depth (might hold in sc-analysis but not in-situ since and same r in negative binomial for every cell (probably not true).
    - Assumption that each local maximum vector is the same as a cell. Can we really assume this?
    - Sequencing depth and gene count have correlation in sc-sequencing which needs to be normalized. For in-situ we do not do sequencing. Do we need to normalize for sequencing depth? Should we normalize for something else that affects gene count for in-situ samples instead?
    - We need another method for this
7. Clustering
    - Some density based clustering method with soft clustering

Notes:

- We have several types of genes. When doing density estimation and finding local maximums we bulk them all together. Afterwards they should be separate.

- Pros: Less issue with sparse areas which could be an issue when we do pointwise density estimations.
- Cons: Could this implement bias by favoring cells containing genes with high count? Does all cells have same total gene count? Probably not. Then maybe we miss some local max in favor or cells with high gene count.

3 Results

- Test entire pipeline on GMM generated data.
  - How many clusters can we find?
  - Try generate data from GMM but with different types of genes mimicing real cells. How well is the transcriptomics preserved by the model?
  - Test the method on benchmark data e.g. osmFish
  - Test on SciLifeLab data?
  - Interpret and explain why results are as they are!!

4 References

(sctransform): Hafemeister, C., Satija, R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. Genome Biol **20**, 296 (2019). https://doi.org/10.1186/s13059-019-1874-1

(Spage2Vec): Partel, G. & Wahlby, C. Spage2vec: Unsupervised representation of localized spatial gene expression signatures. FEBS J, **288**, 1859–187 (2021). https://doi.org/10.1111/febs.15572

(Bayzor): Petukhov, V., Xu, R.J., Soldatov, R.A. et al. Cell segmentation in imaging-based spatial transcriptomics. Nat Biotechnol (2021). https://doi.org/10.1038/s41587-021-01044-w

(SSAM): Park, J., Choi, W., Tiesmeyer, S. et al. Cell segmentation-free inference of cell types from in situ transcriptomics data. Nat Commun **12**, 3545 (2021). https://doi.org/10.1038/s41467-021-23807-4