A Semi-Parametric Probabilistic Method for Efficient Clustering of Single-Cell RNA-Seq Data

Taha Mohseni Ahooyi (1), Benjamin Stear (1), Shubha Vasisht (1), Erin Reichenberger (1), Yuanchao Zhang (1), Deanne M. Taylor (1, 2)

Children's Hospital of Philadelphia
Department of Biomedical & Health Informatics

1. Department of Biological & Health Informatics (DBHi), the Children's Hospital of Philadelphia
2. Perelman School of Medicine, University of Pennsylvania

1. Introduction

Single-cell RNA-seq (scRNA-seq) data analysis treats clustering as a key step towards studying cell type composition, differential expression profiling, marker gene selection, differentiation analysis and deconvolution [1]. Clustering applications typically used for the analysis of scRNA-seq data can inherit intrinsic limitations including undesirable computational complexity, inaccurate nonlinear behavior estimation, noise reduction and ambiguous definition of hard versus soft clusters [2].

We developed a probabilistic framework to systematically address some of these issues associated with conventional scRNA-seq clustering methods.

2. Method

The proposed method takes the following steps to cluster the data: **A)** raw expression data is fed into a feature extraction pipeline and principal components (PCs) are obtained. **B)** Through the kernel and copula methods, marginal and joint probability densities are estimated, respectively, according the following equations for PCs as input variables:

For every multivariate joint probability distribution there is a copula function $\mathrm{C}:[0,1]^d \to [0,1]$ such that

$$\Pr(X_1 \le x_1, \ldots, X_d \le x_d) = \mathrm{C}\left(F_{X_1}(x_1), \ldots, F_{X_d}(x_d)\right)$$

Using *C* we can estimate the joint density as:

$$\mathrm{c}(u_1, \ldots, u_d) = \frac{\partial^d \mathrm{C}(u_1, \ldots, u_d)}{\partial u_1 \ldots \partial u_d}$$

$$f_{\mathbf{X}}(x_1, \ldots, x_d) = \mathrm{c}\left(F_{X_1}(x_1), \ldots, F_{X_d}(x_d)\right) \prod_{i=1}^{d} f_{X_i}(x_i)$$

Here $\hat{C}$ and $\hat{F}_X$ are computed empirically. For $\hat{F}_X$ we utilize kernel density estimation with bandwidth *h*:

$$u = \hat{F}_{X,h}(x) = \frac{1}{n}\int_{-\infty}^{x} \frac{1}{h}\sum_{k=1}^{n} \mathrm{K}\left(\frac{t - \chi_k}{h}\right) dx$$

C) Pairwise density of PC_i and PC_j, $f(PC_i, PC_j)$ is heterogenized by log transformation. **D)** Regions of the heterogenized density distributed around local optima are segmented (watershed segmentation). **E)** Resulting clusters are assigned to each data-point index (barcodes). **F)** This process is repeated for different combinations of PCs. **G)** Barcodes are grouped based on a sequence of detected lower-dimensional clusters and visualized in a reduced dimension space such as UMAP.

3. Case Study Results

ScRNA-seq data was obtained from a recent study published on cell-type identification of nasal swabs from a cohort of COVID-19 patients [3] consisting of 32,588 cells and 18 detected clusters.. To examine our proposed method we used a set of PCs to detect clusters and re-group the barcodes according the steps in section 2 (**Figure 1**). In the presented results we used 5 PCs (PCs 1-5) for 32,588 cells. A total of 32 clusters were detected as opposed to 18 clusters. By design, the number of clusters is determined by the algorithm and is not explicitly provided by users.

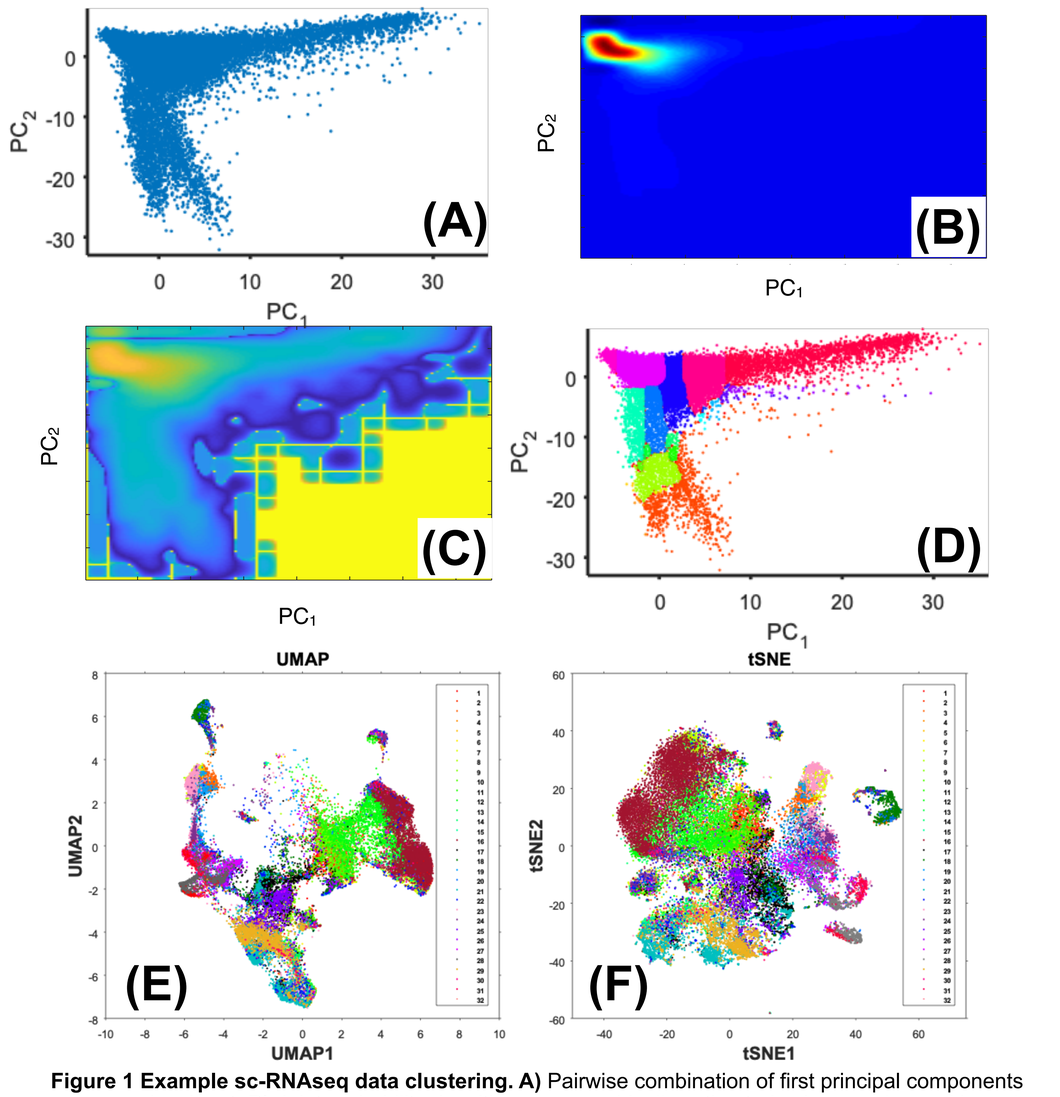

Figure 1 Example sc-RNAseq data clustering. A) Pairwise combination of first principal components were used as input. **B)** Joint probability densities were empirically estimated using copulas and kernel densities. **C)** Densities were heterogenized to make the local maxima detectable . **D)** Output of **C** was processed by segmenting the regions distributed around local peak densities and segment indices were assigned to 32,588 cells. A sequence of indices for a combination of 6 PCs were regrouped to assign each cell with A cluster number 1-32. **E)** UMAP and **F)** tSNE visualizations of clusters in color spectra.

4. Discussion

The proposed method differs from common scRNA-data clustering methods in that it defines the clusters as regions of feature space surrounding the local peak densities. To this end, the method uses an empirical copula to capture nonlinear relationships among the input variables. Overall, the method is fast and does not require pre-known number for clusters, instead cluster number is controlled by the number of input variables (here PCs). We are developing algorithms to measure the accuracy of the method as opposed to the ground truth. Correlation between the cluster labels and the labels provided in the reference study indicates maximum of 0.619. An informed cluster merging scheme is sought to enhance the method accuracy.

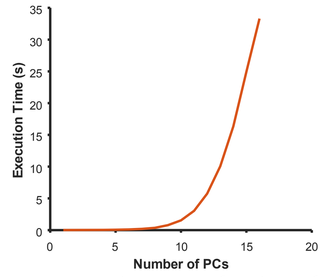

Figure 2: Convergence time as a function of number of PCs

5. Future Directions

We aim to combine different modules of this method as a standalone pipeline in R and Python also as an extension of the Scedar package published previously by Zhang et al [4]. Of our particular interest to automate and accelerate the performance by adopting optimized segmentation routines from available software packages (e.g. watershed function by Mathworks).

6. Acknowledgments

Authors sincerely appreciate all individuals at the CHOP DBHi center who provided us with technical insights over time as well as support by CHOP and national organizations..

7. Bibliography

[1] Menon, V. (**2018**). Clustering single cells: a review of approaches on high- and low-depth single-cell RNA-seq data. *Briefings in functional genomics*, *17*(4), 240-245.

[2] Kiselev, V. Y., Andrews, T. S., & Hemberg, M. (**2019**). Challenges in unsupervised clustering of single-cell RNA-seq data. *Nature Reviews Genetics*, *20*(5), 273-282.

[3] Ziegler, C. G., Miao, V. N., Owings, A. H., Navia, A. W., Tang, Y., Bromley, J. D., et al. & Ordovas-Montanes, J. (**2021**). Impaired local intrinsic immunity to SARS-CoV-2 infection in severe COVID-19. *Cell*, *184*(18), 4713-4733.

[4] Zhang, Y., Kim, M. S., Reichenberger, E. R., Stear, B., & Taylor, D. M. (**2020**). Scedar: a scalable Python package for single-cell RNA-seq exploratory data analysis. *PLoS computational* biology, 16(4), e1007794.