CPSC 545 Project Suggestions

Partners and topics For the final project, you will work in groups of ≤ 2 (it must be 1 in the case of the 4th direction below). You may consider the following directions: 1) developing new statistical models or algorithms and applying them to publicly available data, especially single-cell RNA-seq data, single-cell multi-omics data, or data generated by spatially-resolved technologies, 2) applying existing statistical models to new datasets, 3) reproducing results from journal papers and extending or improving previous analyses, or 4) writing a technical review on the topic of interest (with both biological and computational problems) similar to a short review paper.

Projects Components Submit a 1-page project proposal (excluding references, due at **11:59 PM, Monday, Oct. 21th, 2024.**) outlining the background, literature review, project commitments, deliverables, data to be used, models to be developed or used, and project milestones. The goal of the proposal is to ensure the project is feasible and to receive feedback from the instructors.

The final project deliverables are: 1) a 12-page final report (excluding supplementary materials and references), 2) code submission via public GitHub repository or Canvas 3) 15-minute presentation by all members of the group during the final week, and 4) self and peer evaluations of group contributions.

The final report should be structured like a research paper, including the abstract, introduction, results, discussion (focusing on the insights learned and limitations), and method details. In the presentation, please explain the project's motivation, background, methods, and results. The peer evaluations will be done during the presentation. Please submit your final report by **11:59 PM, Wednesday, December 11th, 2024.**

Please use the course latex template for the proposal and the report.

Potential Projects

Listed below are possible project topics, but by no means is this a complete list. One way to proceed is to find a research article on your topic of interest, which will give you some perspective and suggest open problems. Feel free to propose topics if you would like to work on something else, e.g., topics related to your (honor) thesis work or motivated by other interests, if relevant to this class.

Modelling Single-cell RNA-Seq Data Distributions Modeling single-cell RNA-seq data is a challenging task because of cell heterogeneity, sparsity, and technical artifacts. Researchers have proposed (zero-inflated) Poisson, negative binomial, multinomial distributions, etc for modelling scRNA-seq data, and numerous ways for parameter estimation. You may analyze which distributions and ways of parameter estimations best explain scRNA-seq data. There are lots of relevant papers [3, 7, 8, 11, 12, 14].

Single-cell RNA-Seq Data Analysis Pipeline A 'standard' single-cell RNA-Seq processing pipeline has many steps [5] (`https://satijalab.org/seurat/articles/pbmc3k_tutorial.html`). At the end, cells are clustered based on similarities in their types or states. Many clustering tools can be used, e.g., based on mixture models, density-based clustering, or methods based on graph-partition. However, we typically don't know if a cell is 'optimally' assigned to the respective cluster. You may investigate the uncertainty in cell cluster assignment [1, 6, 10].

Single-cell Studies of Human Diseases Single-cell RNA-Seq has numerous applications in disease studies and high translational values. For example, scientists have used scRNA-seq to study cancer [16], Inflammatory Bowel Disease [9], and neurodegenerative diseases [17], just to name a few [13]. You may

pick a study of interest, reproduce the results, and extend the original analyses by resolving the potential pitfalls.

Gene Programs An important problem in scRNA-seq data analysis is to extract interpretable 'gene programs' [18] using methods such as principal component analysis, non-negative matrix factorizations, topic modelling, and etc. You may use the existing tools [2, 4, 15, 19] to analyze some publicly available data, compare the strengths and weaknesses of these tools, and report your discoveries.

References

[1] I. Benhar, J. Ding, W. Yan, I. E. Whitney, A. Jacobi, M. Sud, G. Burgin, K. Shekhar, N. M. Tran, C. Wang, et al. Temporal single-cell atlas of non-neuronal retinal cells reveals dynamic, coordinated multicellular responses to central nervous system injury. *Nature Immunology*, pages 1–14, 2023.

[2] P. Carbonetto, A. Sarkar, Z. Wang, and M. Stephens. Non-negative matrix factorization algorithms greatly improve topic model fits. *arXiv preprint arXiv:2105.13440*, 2021.

[3] K. Choi, Y. Chen, D. A. Skelly, and G. A. Churchill. Bayesian model selection reveals biological origins of zero inflation in single-cell transcriptomics. *Genome biology*, 21:1–16, 2020.

[4] Z. J. DeBruine, K. Melcher, and T. J. Triche Jr. Fast and robust non-negative matrix factorization for single-cell experiments. *bioRxiv*, pages 2021–09, 2021.

[5] J. Ding, X. Adiconis, S. K. Simmons, M. S. Kowalczyk, C. C. Hession, N. D. Marjanovic, T. K. Hughes, M. H. Wadsworth, T. Burks, L. T. Nguyen, et al. Systematic comparison of single-cell and single-nucleus rna-sequencing methods. *Nature biotechnology*, 38(6):737–746, 2020.

[6] J. Goldberger, G. E. Hinton, S. Roweis, and R. R. Salakhutdinov. Neighbourhood components analysis. *Advances in neural information processing systems*, 17, 2004.

[7] R. Jiang, T. Sun, D. Song, and J. J. Li. Statistics or biology: the zero-inflation controversy about scrna-seq data. *Genome biology*, 23(1):1–24, 2022.

[8] T. H. Kim, X. Zhou, and M. Chen. Demystifying "drop-outs" in single-cell umi data. *Genome biology*, 21(1):196, 2020.

[9] L. Kong, V. Pokatayev, A. Lefkovith, G. T. Carter, E. A. Creasey, C. Krishna, S. Subramanian, B. Kochar, O. Ashenberg, H. Lau, et al. The landscape of immune dysregulation in crohn's disease revealed through single-cell transcriptomic profiling in the ileum and colon. *Immunity*, 2023.

[10] B. Liu, C. Li, Z. Li, D. Wang, X. Ren, and Z. Zhang. An entropy-based metric for assessing the purity of single cell populations. *Nature communications*, 11(1):3155, 2020.

[11] Y. Pan, J. T. Landis, R. Moorad, D. Wu, J. Marron, and D. P. Dittmer. The poisson distribution model fits umi-based single-cell rna-sequencing data. 2023.

[12] E. Pierson and C. Yau. Zifa: Dimensionality reduction for zero-inflated single-cell gene expression analysis. *Genome biology*, 16(1):1–10, 2015.

[13] J. E. Rood, A. Maartens, A. Hupalowska, S. A. Teichmann, and A. Regev. Impact of the human cell atlas on medicine. *Nature Medicine*, pages 1–11, 2022.

[14] V. Svensson. Droplet scrna-seq is not zero-inflated. *Nature Biotechnology*, 38(2):147–150, 2020.

[15] F. W. Townes, S. C. Hicks, M. J. Aryee, and R. A. Irizarry. Feature selection and dimension reduction for single-cell rna-seq based on a multinomial model. *Genome biology*, 20:1–16, 2019.

[16] I. Vázquez-García, F. Uhlitz, N. Ceglia, J. L. Lim, M. Wu, N. Mohibullah, J. Niyazov, A. E. B. Ruiz, K. M. Boehm, V. Bojilova, et al. Ovarian cancer mutational processes drive site-specific immune evasion. *Nature*, pages 1–9, 2022.

[17] P. Wang, L. Yao, M. Luo, W. Zhou, X. Jin, Z. Xu, S. Yan, Y. Li, C. Xu, R. Cheng, et al. Single-cell transcriptome and tcr profiling reveal activated and expanded t cell populations in parkinson's disease. *Cell Discovery*, 7(1):52, 2021.

[18] H. Xu, J. Ding, C. B. Porter, A. Wallrapp, M. Tabaka, S. Ma, S. Fu, X. Guo, S. J. Riesenfeld, C. Su, et al. Transcriptional atlas of intestinal immune cells reveals that neuropeptide α-cgrp modulates group 2 innate lymphoid cell responses. *Immunity*, 51(4):696–708, 2019.

[19] Y. Zhao, H. Cai, Z. Zhang, J. Tang, and Y. Li. Learning interpretable cellular and gene signature embeddings from single-cell transcriptomic data. *Nature communications*, 12(1):5261, 2021.