

Dirichlet Process Mixture Model of Negative Binomial distributions for clustering single-cell RNA-seq data in immunological studies

Master 2/Engineer school Internship

Background

Dirichlet process mixture models (DPMMs) are a Bayesian nonparametric approach to clustering in which the number of components is inferred from the data rather than pre-specified¹. This makes them well suited to gene expression data, particularly single-cell RNA-seq, where the number of cell types or states is usually unknown. Most existing DPMMs, however, use Gaussian components, which does not account for over-dispersed sequencing data. For mixture models in particular, the component variance plays a significant role in determining the number of components present², and therefore model misspecification can become an important issue.

In transcriptomics studies, count-based distributions have been used for differential expression³, normalization and variance stabilization⁴ as well as dimension reduction and feature selection⁵. On the other hand, they have received comparatively little attention as mixture kernels for nonparametric clustering. Apart from overdispersion, gene expression data sets also suffer from sparsity. Recent work shows zero-inflated Negative Binomial models perform better than standard multinomial or Negative Binomial models⁶.

Extending the collapsed VI framework our [vimixr](#)  package⁷ to such Negative Binomial kernels will likely provide both theoretical and computational edge over the state-of-the-art.

Objectives

1. Develop a theoretical framework of DPMM with Poisson-Gamma/Negative Binomial mixtures, followed by a VI-based computational implementation using .
2. Investigate the impact of different prior choices on the model sparsity and performance.
3. Apply this clustering approach to high-dimensional transcriptomics data in the context of immunology.

Required skills:

- Good knowledge in Bayesian Statistics
- Programming proficiency with  (R) and/or  (C++ appreciated)
- An interest for exploratory analysis of biomedical data
- English proficiency (both written and spoken)
- Scientific curiosity
- Enrolled in Master 2/Engineering school with a major in (Bio)-Statistics

Hosting laboratory:

SISTM team at Inserm U1219 *Bordeaux Population Health* & Centre Inria de l'Universit  de Bordeaux

Location:

Inserm U1219 *Bordeaux Population Health* research center

Universit  de Bordeaux

146, rue L o Saignat

33076 Bordeaux Cedex

Duration:

Internship of 5 to 6 month available starting from February 2027.

Compensation:

Intern gratification according to the official recommendations (15% of social security ceiling, i.e. around 625 /month).

Contact:

Send a detailed CV and a motivation letter to both **Boris Hejblum** [boris.hejblum@u-bordeaux.fr] & **Annesh Pal** [annesh.pal@u-bordeaux.fr]

Bibliography

1. Antoniak, C. E. Mixtures of dirichlet processes with applications to bayesian nonparametric problems. *Annals of Statistics* 1152–1174 (1974) doi:[10.1214/aos/1176342871](https://doi.org/10.1214/aos/1176342871).
2. Hejblum, B. P., Alkhassim, C., Gottardo, R., Caron, F. & Thi baut, R. [Sequential dirichlet process mixture of skew t-distributions for model-based clustering of flow cytometry data](#). *Annals of Applied Statistics* **13**, 638–660 (2019).
3. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Nature Precedings* 1–1 (2010) doi:[10.1186/gb-2010-11-10-r106](https://doi.org/10.1186/gb-2010-11-10-r106).
4. Hafemeister, C. & Satija, R. [Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression](#). *Genome biology* **20**, 296 (2019).

5. Townes, F. W., Hicks, S. C., Aryee, M. J. & Irizarry, R. A. [Feature selection and dimension reduction for single-cell RNA-seq based on a multinomial model.](#) *Genome biology* **20**, 295 (2019).
6. Wilson, T., Vo, D. H. & Thorne, T. [Identifying subpopulations of cells in single-cell transcriptomic data: A bayesian mixture modeling approach to zero inflation of counts.](#) *Journal of Computational Biology* **30**, 1059–1074 (2023).
7. Pal, A., Mimoun, A., Thiébaud, R. & Hejblum, B. P. Scalable dirichlet process mixture models with unknown concentration and adaptive covariance for high-dimensional clustering applied to leukemia transcriptomics. *arXiv Preprint* (2026) doi:[10.48550/arXiv.2601.21106](#).