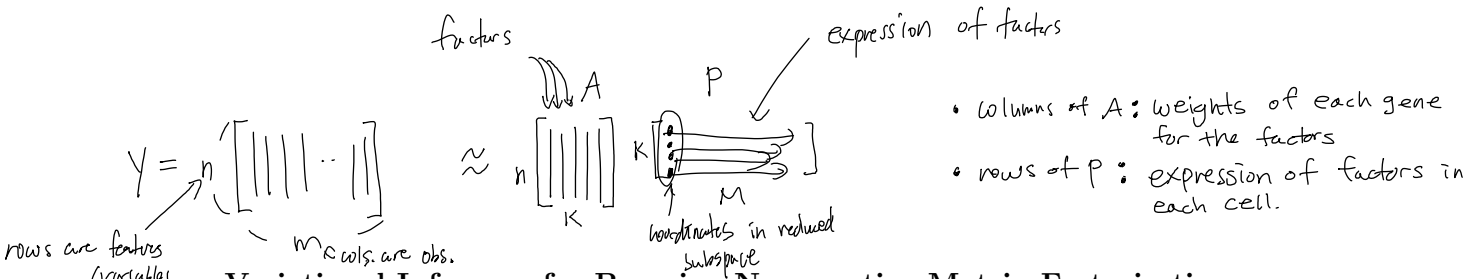


Variational Inference for Bayesian Nonnegative Matrix Factorisation

Gyu Hwan Park

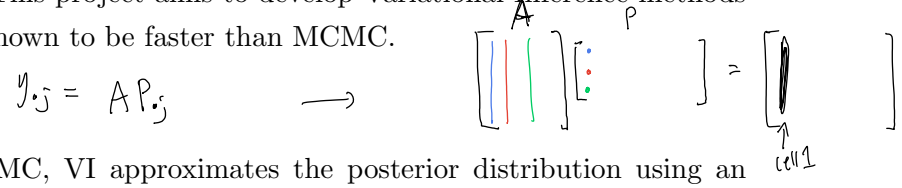
Motivation

Nonnegative matrix factorisation (NMF) is one of dimension reduction techniques, and it has been used to identify the molecular features (factors) associated with biological processes from single-cell RNA-seq data [1]. Given a data matrix $Y \in \mathbb{R}^{N \times M}$ of single cell data with N genes and M cells, NMF factorises it into two lower dimensional matrices, $A \in \mathbb{R}_+^{N \times K}$ and $P \in \mathbb{R}_+^{K \times M}$, where $K \ll N$ and M, as follows:

$$Y \approx AP + \varepsilon$$

$$\mathbb{E}[Y] = AP.$$

Here, the elements in the A and P matrices are assumed to be greater than or equal to zero. The columns of A represent the relative weights of each gene for the factors, and the corresponding rows of P represents the relative expression of the factors in each cell [2]. For Bayesian inference of NMF, Markov chain Monte Carlo methods (MCMC) have been developed [1][2]. MCMC achieves high accuracy yet is computationally intensive. Thus, it limits the wide application of NMF for large single-cell datasets. This project aims to develop Variational Inference methods (VI) for Bayesian NMF which are known to be faster than MCMC.



Proposed Methods

As an alternative approach to MCMC, VI approximates the posterior distribution using an optimisation approach. Specifically, given a family of probability distributions, VI aims to find a member which is most similar to the posterior distribution. VI is known to be more computationally efficient than MCMC, at the cost of loss in accuracy. The aims and tentative timeline of the project are as follows:

- Review and derive VI algorithm for Bayesian inference of NMF. (1.5 weeks)
- Implement the algorithm. (1.5 week)
- Compare the performance of VI and MCMC methods [1][2] by using simulated datasets and real single-cell RNA-seq data. (1.5 weeks)
- Create an R package implementing the VI for Bayesian inference of NMF. (1.5 week)

References

1. Stein-O'Brien et al. 2019. Decomposing Cell Identity for Transfer Learning across Cellular Measurements, Platforms, Tissues, and Species. Cell Systems.
2. Sherman et al. 2020. CoGAPS 3: Bayesian non-negative matrix factorization for single-cell analysis with asynchronous updates and sparse data structures. BMC Bioinformatics.