Probabilistic Count Matrix Factorization for Single Cell Expression Data Analysis

Ghislain Durif*[1]

[1]CNRS / IMAG / Université de Montpellier – CNRS – France

Résumé

The development of high throughput single-cell technologies now allows the investigation of the genome-wide diversity of transcription at different scopes. First, the gene-to-gene variability (expression dynamics) can be quantified more accurately, thanks to the measurement of lowly-expressed genes. Second, the cell-to-cell variability is high, with a low proportion of cells expressing the same gene at the same time/level. Those emerging patterns appear to be very challenging from the statistical point of view, especially to represent and to provide a summarized view of single-cell expression data like single-cell RNA-seq data. Principal Component Analysis (PCA) is one of the most powerful framework to provide a suitable representation of high dimensional datasets, by searching for new axis catching the most variability in the data. Unfortunately, classical PCA is based on Euclidean distances and projections that work poorly in presence of over-dispersed counts that show zero-inflation (i.e. massive drop-outs or null values) such as single-cell RNA-seq data. We propose a probabilistic Count Matrix Factorization (pCMF) approach for single-cell expression data analysis. It relies on a sparse Gamma-Poisson factor model with a zero-inflated framework. This hierarchical model is inferred using a variational EM algorithm. We show how this probabilistic framework induces a geometry that is suitable for single-cell expression data, and produces a compression of the data that is very powerful for visualization or clustering purposes. Our method is competed to other standard representation methods like t-SNE or ZIFA (Zero-Inflated Factor Analysis), and we illustrate its performance for the representation of single-cell expression data. We especially focus on different publicly available single-cell RNA-seq datasets.

*Intervenant