Dimension Reduction and Dropout Imputation for Single-Cell RNA Sequencing Data Using Constrained Robust Nonnegative Matrix Factorization

Shuqin Zhang

Fudan University
E-mail: zhangs@fudan.edu.cn

Abstract: Single cell RNA-sequencing (scRNA-seq) technology is a powerful tool to analyze the whole transcriptome at single cell level, and it has been receiving more and more attention in recent years. Dimension reduction and clustering are the basic steps in scRNA-seq data analysis, and they are seriously affected by the dropout phenomenon, which is an important characteristic of scRNA-seq data. The cells with lower sequencing depth will tend to have more dropouts compared with the deeper sequenced cells. Moreover, scRNA-seq data is count-based and thus nonnegative.

In this paper, we propose a model for simultaneously implementing dropout imputation and dimension reduction of scRNA-seq data under the nonnegative matrix factorization (NMF) framework.

The dropouts modeled as a nonnegative sparse matrix are added to the observed data matrix, which is approximated by NMF. A weighted L_1 penalty taking into account the dependence of the dropouts on the sequencing depth in each single cell is imposed to ensure the sparsity pattern. The computational efficient method is developed to solve the formulated optimization problem. Experiments on both synthetic data and real data show that the dimension reduction can give more robust clustering results compared with the existing methods and the dropout imputation helps improve the differential expression analysis.