Seurat v3 HVG implementation

Adam Gayoso

March 2020

1 Introduction

This note describes the implementation of the Seurat v3 HVG method in Python. This arithmetic is required for handling sparse matrices.

Let μ_g and σ_g be the mean and regularized standard deviation per gene g as described in the Seurat v3 HVG method. Let X_{ng} be the UMI counts for cell n and gene g. N is the total number of cells.

This note is based on the implementation here `https://github.com/satijalab/seurat/blob/master/R/preprocessing.R` and here `https://github.com/satijalab/seurat/blob/master/src/data_manipulation.cpp`.

The variance of the gene after the variance stabilizing transformation is

$$\frac{1}{N-1}\sum_{i=1}^{N}(\frac{X_{ig}-\mu_g}{\sigma_g})^2 \tag{1}$$

This is due to the fact that after the transformation, each gene has mean 0. With some expansion,

$$\begin{aligned}\frac{1}{N-1}\sum_{i=1}^{N}(\frac{X_{ig}-\mu_g}{\sigma_g})^2 &= \frac{1}{N-1}\frac{1}{\sigma_g^2}\sum(X_{ig}^2-2X_{ig}\mu_g+\mu_g^2) &(2)\\ &= \frac{1}{N-1}\frac{1}{\sigma_g^2}N\mu^2+\frac{1}{N-1}\frac{1}{\sigma_g^2}\sum_{i=1}^{N}(X_{ig}^2-2X_{ig}\mu_g) &(3)\end{aligned}$$

Note that this equation is simple to compute with sparse matrices. Seurat v3 clips values in Equation 1 so that

$$\frac{1}{N-1}\sum_{i=1}^{N}(\min\{\frac{X_{ig}-\mu_g}{\sigma_g},\sqrt{N}\})^2 \tag{4}$$

This is equivalent to setting values that satisfy

$$\frac{X_i-\mu}{\sigma} > \sqrt{N} \tag{5}$$

to

$$X_i = \sigma\sqrt{N} + \mu. \tag{6}$$

This should be done before computing the sparse-friendly variance.