Dimension reduction for single-cell and spatial RNA-seq using generalized bilinear models

Phillip Nicol
Department of Biostatistics, Harvard University

NESS 2024 (Storrs, CT)

May 24, 2024

Introduction

PCA cannot be directly applied due to heterogeneous variances

Standard approach is to transform the counts prior to PCA

The counts are commonly pre-processed by computing the *Pearson residual*:

$$Z_{ij} := \frac{Y_{ij} - \hat{\mu}_i}{\sqrt{\hat{\mu}_i - \hat{\mu}_i^2/\hat{\alpha}_i}} \tag{1}$$

- scTransform (SCT) (REF): Estimate $\hat{\mu}$ and $\hat{\alpha}$ with a NB GLM.
- APR (REF): Fix $\hat{\alpha} = 100$ (XX) and use a closed-form approximation to $\hat{\mu}$.

PCA on Pearson residuals can fail to capture rare cell types

If baseline mean is large then $Z_{ij} \approx Z_{ij'}$ even for very different counts.
FIG XXX

ERCC scaling

scGBM simultaneously models the counts and reduces dimensionality

$$
\begin{aligned}
Y_{ij} &\sim \text{Pois}(\mu_{ij}) \\
\log(\mu_{ij}) &= \alpha_i + \beta_j + \sum_{m=1}^{M} \sigma_m u_{im} v_{jm} \qquad (2) \\
\sigma_m &\sim \text{Expo}(a)
\end{aligned}
$$

$V := [v_{im}] \in \mathbb{R}^{J \times M}$ is the (low-dimensional) cell embeddings.
FIG XX

Estimation with iteratively reweighted singular value decomposition

Define $\hat{X}^{(t)}$ to be the current estimate of $\hat{U}\hat{\Sigma}\hat{V}^{\top}$. The following update is used for the latent factors:

$$\hat{X}^{(t+1)} = \text{SVD}_{M,1/a}\left(\hat{X}^{(t)} + \gamma(Y - \hat{\mu})\right) \tag{3}$$

$\text{SVD}_{M,1/a}(\cdot)$ computes the rank M truncated SVD and then soft-thresholds the remaining singular values by $1/a$.

Faster estimation using scGBM-proj

When J is very large, first estimate $\hat{\alpha}, \hat{U}$ using a smaller subset of cells.

Then holding $\hat{\alpha}$ and $\hat{U}$ fixed, the parameters β and $V\Sigma$ can be estimated by fitting J GLMs in parallel.

By analogy to PCA, we call this the *projection method* (scGBM-proj)

scGBM is faster and more accurate than GLM-PCA

Single marker genes

ERCC Scaling

scGBM quantifies uncertainty in the low-dimensional embedding of cells

Cluster confidence index

Extending to spatial transcriptomics

Edge-aware spatial smoothing

Acknowledgements

- Jeff Miller (Harvard Biostatistics)
- Rafa Irizarry (DFCI Data Science)
- NIH Cancer Training Grant (REF)