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

Phillip Nicol
Department of Biostatistics, Harvard University

NESS 2024 (Storrs, CT)

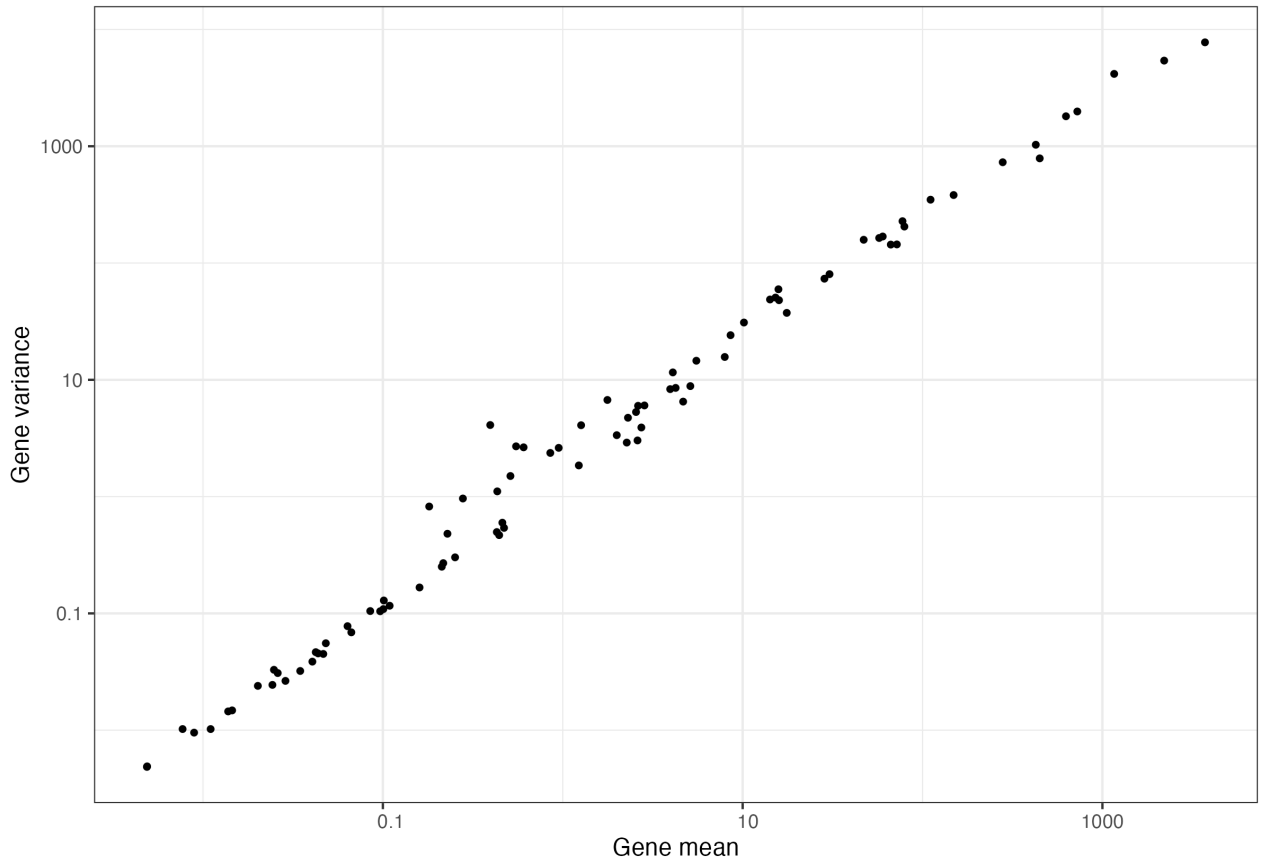
May 24, 2024

Introduction

- Single-cell RNA-seq is a revolutionary technology that allows gene expression to be quantified at the level of individual cells.
- Data comes in the form of $I \times J$ count matrix Y.
- For large datasets $I \approx 10^4$ and $J \approx 10^7$.
- Dimension reduction is a critical first step before downstream analysis (clustering, visualization, etc.)

	Cell 1	Cell 2	...	Cell J
Gene 1	1	0	.	0
Gene 2	14	11	.	3
...	.	.	.	.
Gene I	0	5	.	0

PCA cannot be directly applied due to heterogeneous variances



Standard approach is to transform the counts prior to PCA

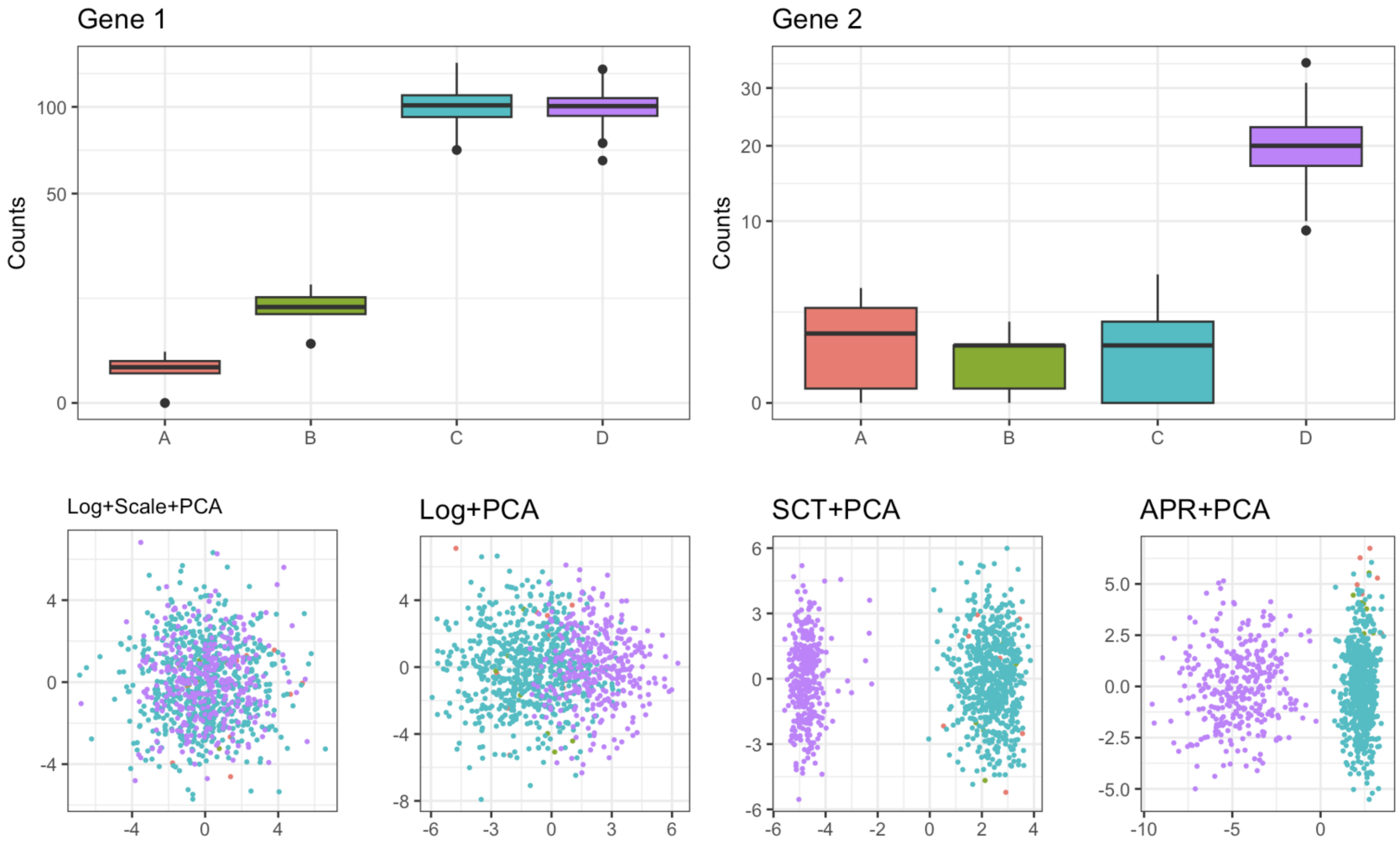
The counts are commonly pre-processed by computing the *Pearson residual*:

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

- scTransform (SCT) [1]: Estimate $\hat{\mu}$ and $\hat{\alpha}$ with a NB GLM.
- APR [2]: Fix $\hat{\alpha} = 0.01$ and use a closed-form approximation to $\hat{\mu}$.

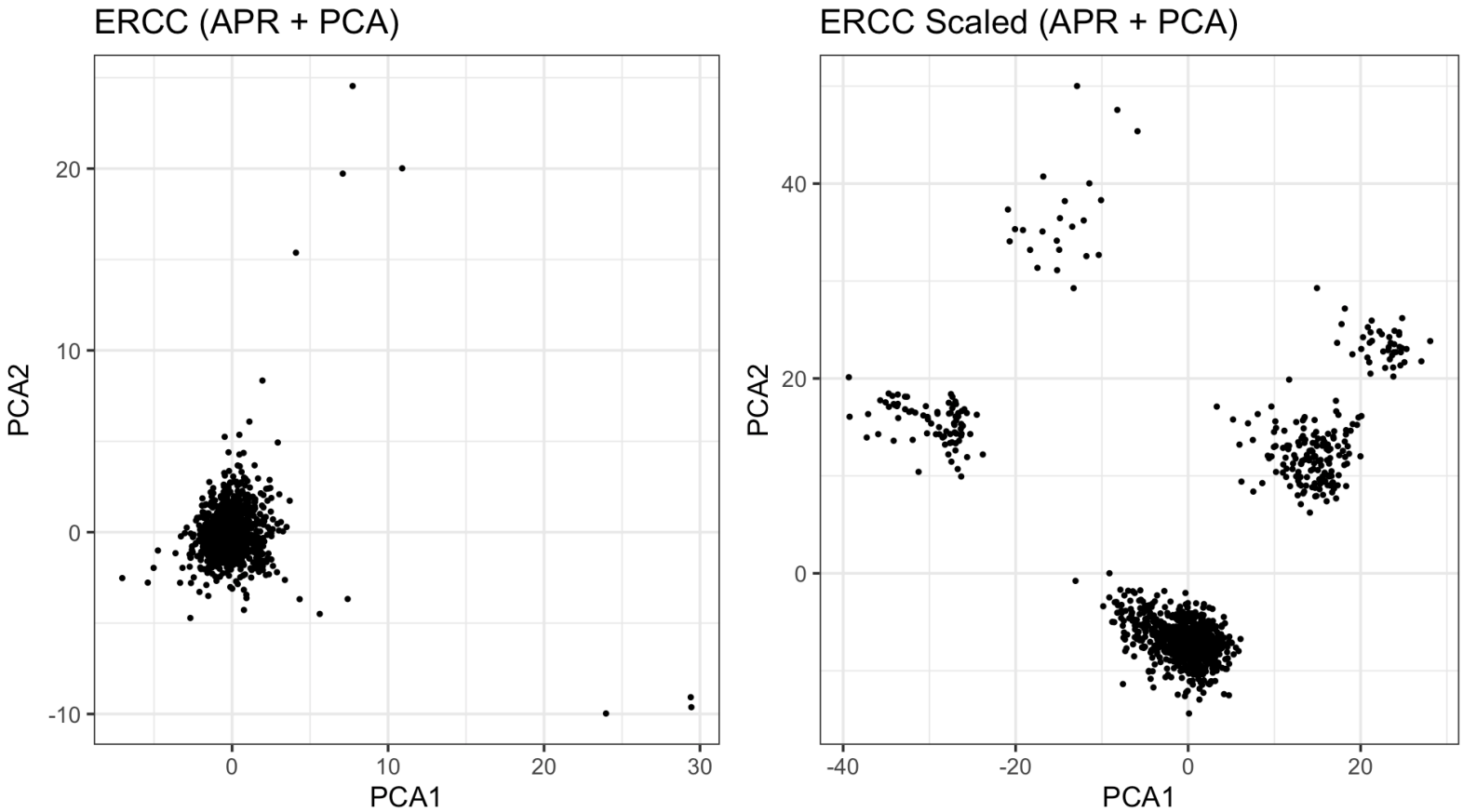
PCA on Pearson residuals can fail to capture rare cell types

If $\hat{\mu}_i$ is large then $Z_{ij} \approx Z_{ij'}$ even for very different counts.



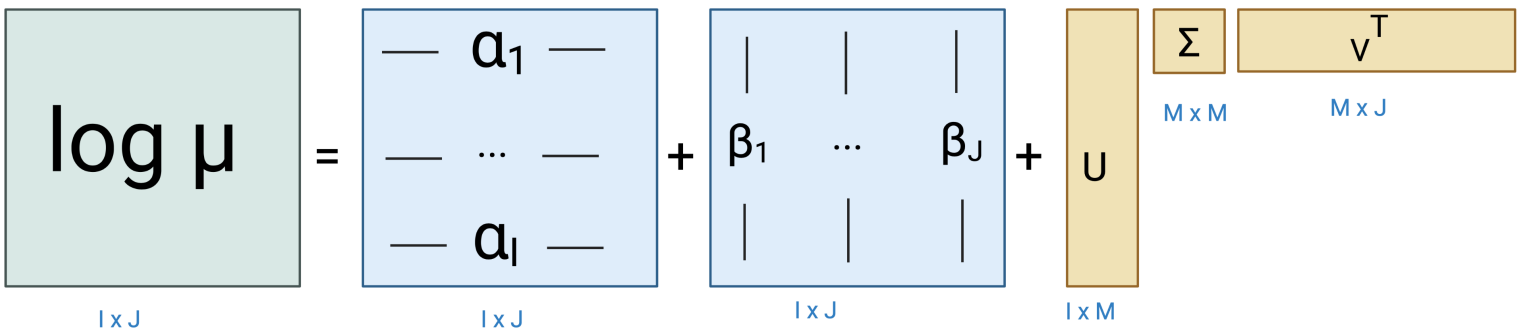
PCA on Pearson residuals is sensitive to changes in baseline expression

Starting with a dataset of technical replicates, scale each gene (row) by $\kappa_i \sim \text{Expo}(100)$:



scGBM performs model-based dimension reduction

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



$V \in \mathbb{R}^{J \times M}$ are the low-dimensional cell-embeddings

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:

$$\hat{X}^{(t+1)} = \text{SVD}_{M,1/a}\left(\hat{X}^{(t)} + \gamma(Y - \hat{\mu})\right) \tag{2}$$

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

Faster estimation using scGBM-proj

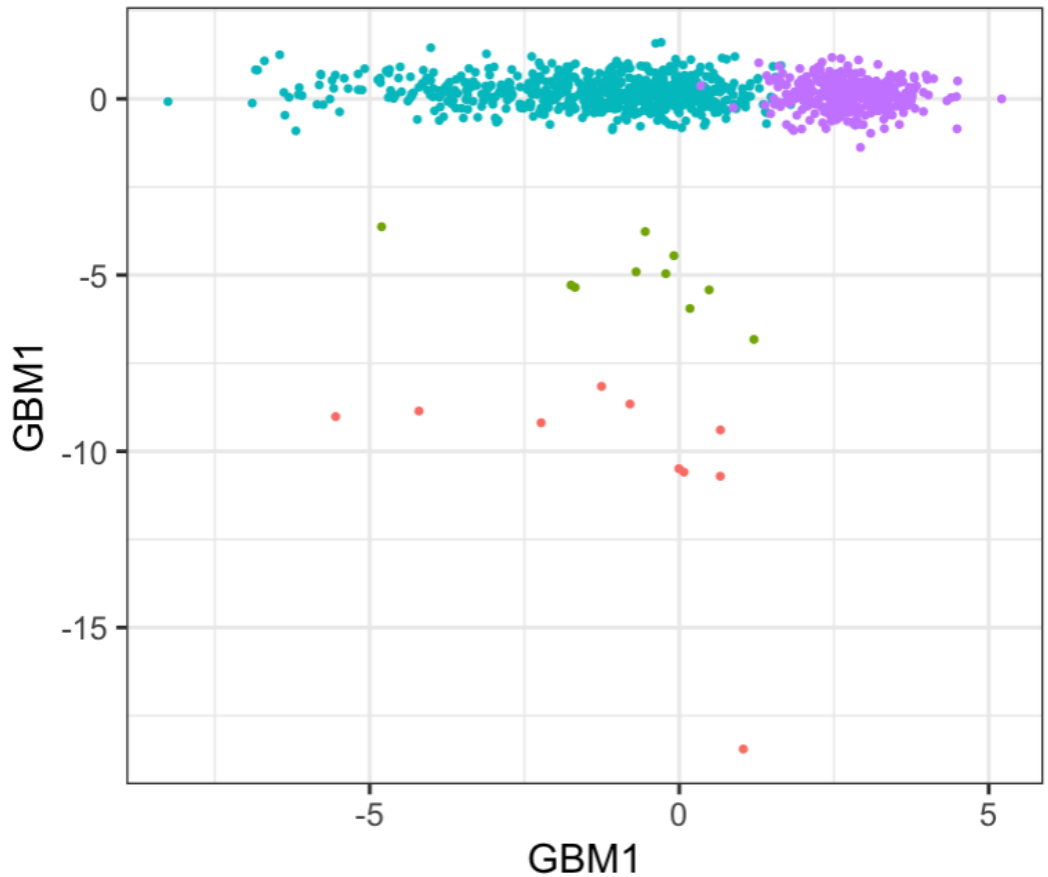
When J is very large, first estimate $\hat{\alpha}, \hat{U}$ using a smaller subset of cells.

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

Rare cell type simulation



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)

References

[1] Christoph Hafemeister and Rahul Satija. Normalization and variance stabilization of single-cell rna-seq data using regularized negative binomial regression. *Genome biology*, 20(1):296, 2019.

[2] Jan Lause, Philipp Berens, and Dmitry Kobak. Analytic pearson residuals for normalization of single-cell rna-seq umi data. *Genome biology*, 22:1–20, 2021.