Model-based Dimensionality Reduction for Single-cell RNA-seq using Generalized Bilinear Models

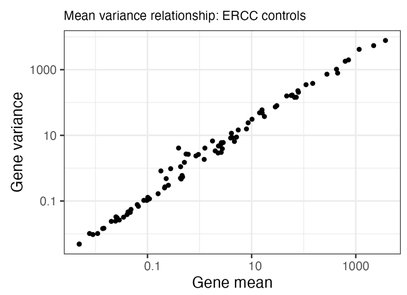
Phillip B. Nicol [1,2] Jeffrey W. Miller [1]

[1]Department of Biostatistics, Harvard University [2]Department of Data Science, Dana-Farber Cancer Institute

1. Introduction

Single-cell RNA-seq (scRNA-seq) is a revolutionary technology that measures gene expression at the level of individual cells. Because thousands of genes are measured for millions of cells, dimension reduction is a key first step in the analysis of these data. The observed data are unique molecular identifier (UMI) counts which can span several orders of magnitude. Consequently, principal component analysis (PCA) can not be directly applied due to extreme heterogeneity in the gene variances:

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



2. Shortcomings of transformation-based approaches

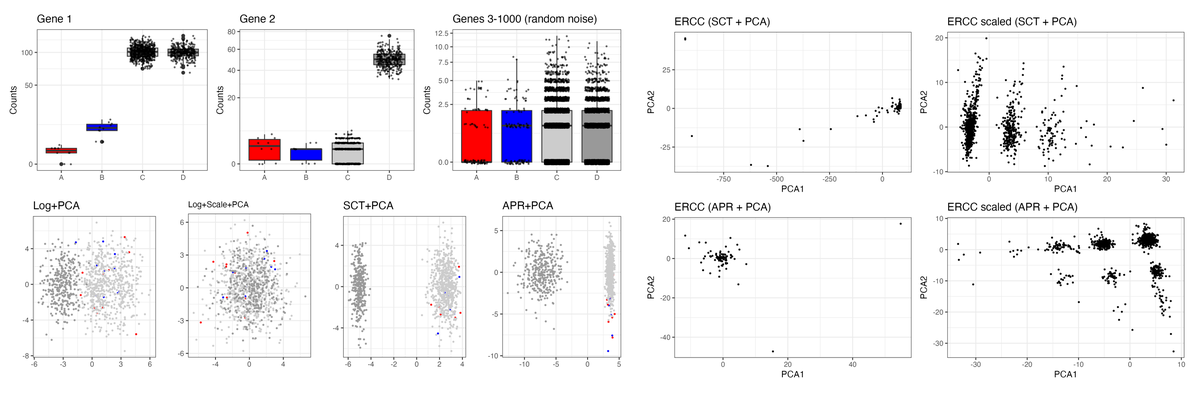
Let $Y \in \mathbb{R}^{I \times J}$ denote the count matrix (I genes, J cells). The standard approach is to transform Y to remove the unwanted variability prior to applying PCA. One of the most common trasformations is the *Pearson residual:*

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

scTransform (SCT) [1] obtains $\hat{\mu}_{ij}$ and $\hat{\theta}_i$ by fitting gene-wise GLMs whereas analytic Pearson residuals (APR) [2] use a closed-form for $\hat{\mu}_{ij}$ and $\hat{\theta}_i = 100$ for all i.

(a) Simulated data with four clusters: PCA on transformed count matrix is unable to identify rare cell types.

(b) ERCC technical control data with one cluster: Scaling each row (gene) causes PCA to introduce spurious clusters.

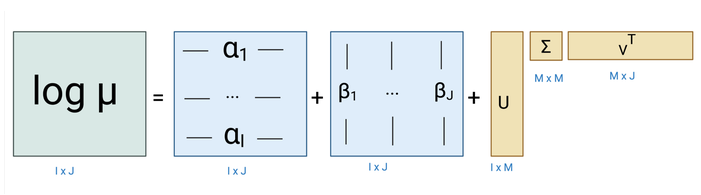


3. scGBM fits a Poisson bilinear model to the count matrix

scGBM estimates latent factors in log space while simultaneously modeling the count distribution of the data:

$$\begin{aligned} Y &\sim \text{Poisson}(\mu),\ \mu \in \mathbb{R}_+^{I \times J} \\ \log(\mu) &= \alpha \mathbf{1}_J^T + \mathbf{1}_I \beta^T + U\Sigma V^T \\ \Sigma &= \text{diag}(\sigma_1, \ldots, \sigma_M);\ \sigma_m \sim \text{Expo}(\tau) \end{aligned} \tag{2}$$

Similar to PCA, we enforce the factor loadings to be orthonormal, $U^\top U = I_M$.



4. Accurate estimation using iteratively reweighted singular value decompositions

Writing $X = U\Sigma V^\top$, we aim to maximize the penalized log-likelihood

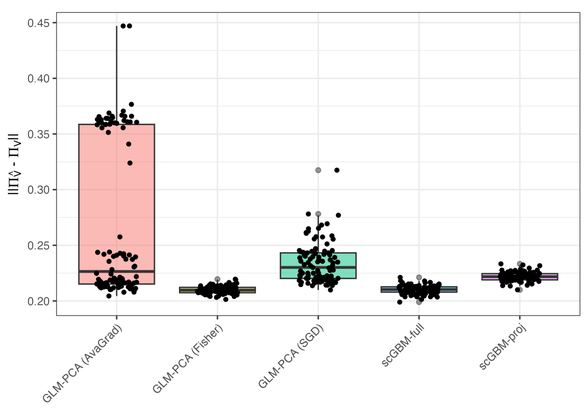
$$\ell(\alpha, \beta, X) = \sum_{ij} \left[Y_{ij}(\alpha_i + \beta_j + X_{ij}) - \exp(\alpha_i + \beta_j + X_{ij}) \right] - \tau ||X||_* \tag{3}$$

where $||\cdot||_*$ is the nuclear norm. Applying proximal gradient descent yields an update for the latent factors:

$$\hat{X}^{(t)} = \text{SVD}_{M,\tau}\left(\hat{X}^{(t-1)} + \gamma(Y - \hat{\mu}^{(t-1)}) \right) \tag{4}$$

where $\text{SVD}_{M,\tau}$ returns the rank-M truncated SVD of a matrix where the first M singular values have been soft-thresholded by τ. After this, the intercepts have a closed-form update.

We compare this method (**scGBM-full**) to an existing method GLM-PCA [4] on simulated data. We measure performance in estimating cell-embeddings by $||\Pi_V - \Pi_{\hat{V}}||_F$, where $\Pi_\cdot$ is a projection matrix.



scGBM outperforms the gradient-based methods in GLM-PCA and performs equally to Fisher scoring. However, scGBM-full attains a faster runtime of $O(IJM)$ per iteration compared to Fisher scoring which is quadratic in J.

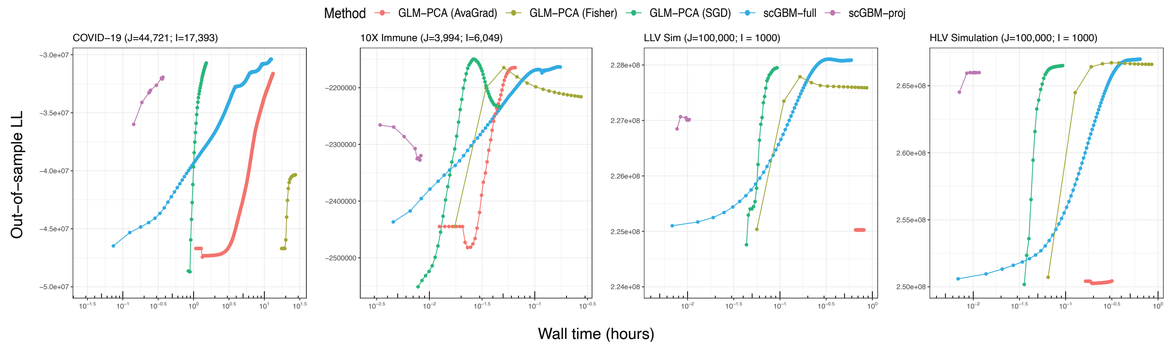
5. Fast estimation by subsampling cells

For very large datasets, we begin by estimating gene-specific parameters (U, α) using a much smaller subset of cells. Then the remaining cell-specific parameters $(\beta, V\Sigma)$ can be estimated by fitting J Poisson GLMs in parallel. For cell j, the linear predictor can be written as

$$\log(\mu_{ij}) = \hat{\alpha}_i + \beta_j + \sum_{m=1}^{M} (\sigma_m V_{jm}) \hat{U}_{im}. \tag{5}$$

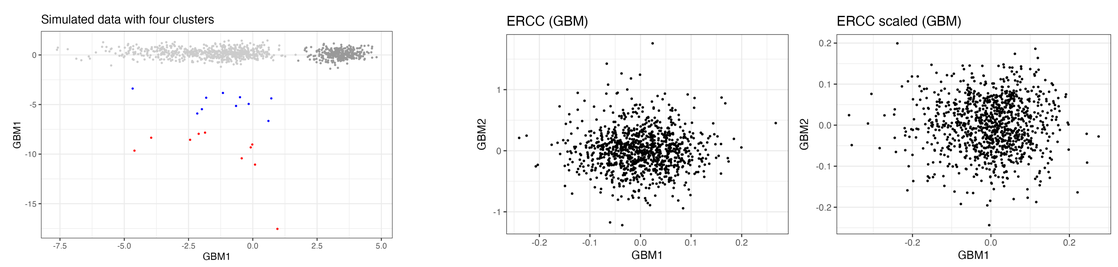
On simulated data, this method (**scGBM-proj**) is only slightly less accurate (see Section 4) but is orders of magnitude faster.

We assess out-of-sample performance by sampling $(Y^*)_{ij} \sim \text{Binomial}(Y_{ij}, 1/2)$ so that the matrices $Y^{**} := Y - Y^*$ and Y^* are iid assuming the data Y_{ij} are Poisson.



6. Simulation results

Applying scGBM to the simulated data from Section 2 demonstrates the improvement of directly modeling the counts and decomposing latent effects in log-space:



7. scGBM quantifies uncertainty in the low-dimensional embedding

To quantify uncertainty in V, we invert the diagonal blocks of the Fisher information for V, that is, the submatrices $F_1, \ldots, F_J \in \mathbb{R}^{M \times M}$ where

$$F_{j,m,m'} = -\mathbb{E}\left(\frac{\partial^2 \ell}{\partial V_{jm} V_{jm'}} \right). \tag{6}$$

The square root of the diagonals of F_j^{-1} give estimates for the standard errors of $\hat{V}_j$.

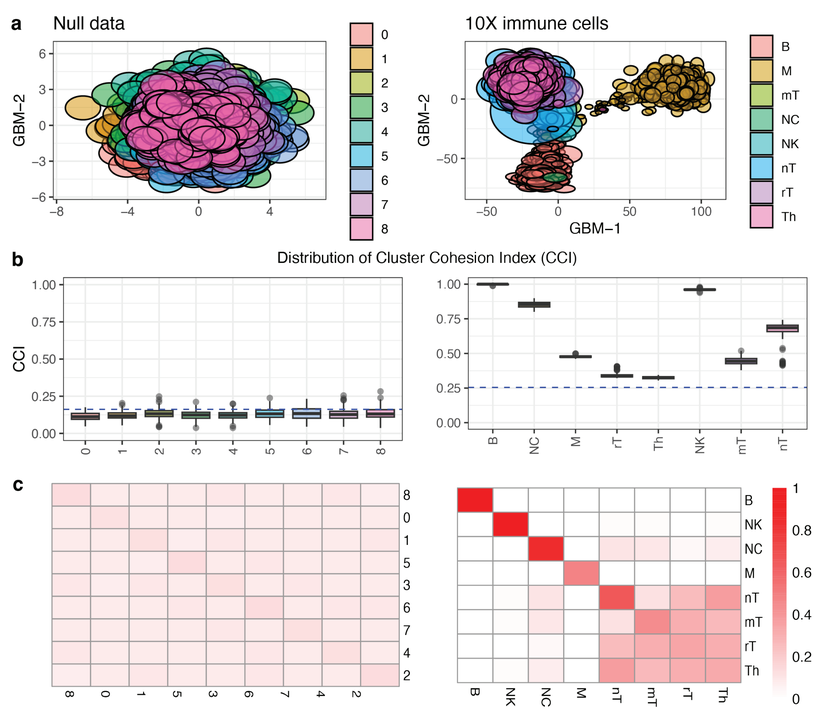
Cell embeddings (from PCA) are typically used to cluster the cells (to assign them to cell types). We defined a **cluster cohesion index** (CCI) that uses the scGBM uncertainty to measure the confidence interval overlap between cells in different clusters. To compute the CCIs, we start with an initial clustering $c_1, \ldots, c_J \in [K]$ and repeat the following steps $n = 100$ times:

1. Draw $\tilde{V}_{jm} \sim \mathcal{N}(\hat{V}_{jm}, \text{se}(\hat{V}_{jm})^2)$ for $j = 1, \ldots, J$, $m = 1, \ldots, M$.
2. Apply a clustering algorithm to the rows of $\tilde{V} = [\tilde{V}_{jm}]$ to obtain new cluster assignments $\tilde{c}_1, \ldots, \tilde{c}_J$.
3. For each pair of clusters $k, k' \in \{1, \ldots, K\}$, compute the fraction

$$f_{k,k'} = \frac{1}{|S_{k,k'}|} \sum_{(j,j') \in S_{k,k'}} \text{I}(\tilde{c}_j = \tilde{c}_{j'}) \tag{7}$$

where $S_{k,k'}$ is the set of pairs $j, j' \in \{1, \ldots, J\}$ such that $j \neq j'$, $c_j = k$, and $c_{j'} = k'$.

We demonstrate the use of the CCIs on (1) null data where $Y_{ij} \overset{\text{iid}}{\sim} \text{Pois}(1)$ and (2) 10X immune cell data consisting of 8 sorted cell types:



The low CCIs for the null data indicate that these clusters are likely the result of sampling variability.

8. 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):1–15, 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.

[3] Phillip B Nicol and Jeffrey W Miller. Model-based dimensionality reduction for single-cell rna-seq using generalized bilinear models. *bioRxiv*, 2023.

[4] F William Townes, Stephanie C Hicks, Martin J Aryee, and Rafael A Irizarry. Feature selection and dimension reduction for single-cell rna-seq based on a multinomial model. *Genome biology*, 20(1):1–16, 2019.

phillipnicol@g.harvard.edu