Exponential-family embedding with application to cell developmental trajectories for single-cell RNA-seq data

Kevin Z. Lin

University of Pennsylvania, Wharton Statistics Department, Philadelphia, PA

Jing Lei

Carnegie Mellon University, Statistics & Data Science Department, Pittsburgh, PA[*]

Kathryn Roeder

Carnegie Mellon University, Statistics & Data Science Department, Pittsburgh, PA[†]

Abstract

Scientists often embed cells into a lower-dimensional space when studying single-cell RNA-seq data for improved downstream analyses such as developmental trajectory analyses, but the statistical properties of such non-linear embedding methods are often not well understood. In this article, we develop the *eSVD* (exponential-family SVD), a non-linear embedding method for both cells and genes jointly with respect to a random dot product model using exponential-family distributions. Our estimator uses alternating minimization, which enables us to have a computationally-efficient method, prove the identifiability conditions and consistency of our method, and provide statistically-principled procedures to tune our method. All these qualities help advance the single-cell embedding literature, and we provide extensive simulations to demonstrate that the eSVD is competitive compared to other embedding methods.

[*] This work was supported by National Science Foundation (NSF) grants DMS-2015492, and DMS-1553884.

[†] This work was supported by National Institute of Mental Health (NIMH) grant R37MH057881 and R01MH123184 and and Simons Foundation Autism Research Initiative (SFARI) grants SF402281 and SF367561.

We apply the eSVD via Gaussian distributions where the standard deviations are proportional to the means to analyze a single-cell dataset of oligodendrocytes in mouse brains (Marques et al., 2016). Using the eSVD estimated embedding, we then investigate the cell developmental trajectories of the oligodendrocytes. While previous results are not able to distinguish the trajectories among the mature oligodendrocyte cell types, our diagnostics and results demonstrate there are two major developmental trajectories that diverge at mature oligodendrocytes.

Keywords: gene expression, oligodendrocytes, latent space models, matrix factorization, random dot product model

1 Introduction

Single-cell RNA-sequencing data give scientists an unprecedented opportunity to analyze the dynamics among individual cells based on their gene expressions, but many analyses require first embedding each cell into a lower-dimensional space as an important preprocessing step in order to make downstream methods more statistically or computationally tractable. For example, these low-dimensional embeddings can be used to visualize high-dimensional data, to control for batch effects, to cluster cells into cell types, to denoise or impute single-cell data, or to estimate trajectories to understand how cells develop over time; see Sun et al. (2019) for a comprehensive overview. Typically, these embeddings are computed from an $n \times p$ gene expression matrix, where each of the n rows and p columns represent a different cell and a different gene respectively. Two of the most common methods to compute these embeddings, uniform manifold approximation and projection (UMAP, McInnes et al. (2018)) and the singular value decomposition (SVD), have different weaknesses that we strive to remedy in this work. On one hand, UMAP produces flexible, non-linear embeddings that have seen widespread usage for visualization purposes (Becht et al., 2019). However, since UMAP does not yet have proven statistical properties such as consistency, there is a lack of consensus on how to tune this method, and methods that build on these UMAP embeddings inherit this statistical ambiguity; see Cao et al. (2019) and Bergen et al. (2020) for example. On the other hand, the SVD has been extensively studied in the statistical literature, but is often restrictive in practice since it yields only linear embeddings. In this article, we advance the literature by developing the eSVD (exponential-family SVD), a non-linear embedding method that retains desirable statistical properties. As the name suggests, the eSVD is a generalization of the SVD, and embeds each cell in a non-linear fashion into a lower-dimensional space with respect to any one-parameter exponential-family distribution, allowing the researcher to have much broader modeling flexibility. Methodologically, we design the eSVD such that it can be appropriately tuned using matrix-completion ideas. Theoretically, unlike similar work that also bridges this gap between the SVD and UMAP for single-cell applications such as Durif et al. (2017) and Risso et al. (2018), we leverage recent theoretical developments in the nonconvex optimization literature that formalize the statistical properties of the eSVD. With these insights, we use the eSVD to analyze single-cell data[1] in order to demonstrate better downstream analysis results.

To illustrate the importance of embeddings, we focus on analyzing oligodendrocytes – cells that enable rapid transmission of signals by producing myelin and providing metabolic support to neurons in the central nervous system. These cells are intriguing to study due to their constant development throughout a subject's lifetime, unlike many other cell types that mature at adulthood (Menn et al., 2006). As mentioned in Marques et al. (2016) and Cai and Xiao (2016), understanding how oligodendrocytes develop can lead to new insights into the cause of myelin disorders such as multiple sclerosis and Alzheimer's disease. We discuss the oligodendrocyte dataset and present a preliminary analysis in Section 2, where we provide various diagnostics demonstrating the shortcomings of the SVD. To better understand this phenomenon, we review the hierarchical model that the SVD implicitly assumes in Section 3. Specifically, suppose a hierarchical model where each cell and each gene has its own low-dimensional latent random vector. In the language of exponential-family distributions, this model assumes that the cell's expression of a particular gene is a one-parameter exponential-family random variable whose natural parameter is the inner product of the two corresponding latent vectors. By formulating this hierarchical model, we see that the SVD implicitly assumes a Gaussian distribution with constant variance. However, this assumption is often violated since the variance of each cell's gene expressions is observed to vary dramatically with their mean expression level (Love et al., 2014; Hicks et al.,

[1] We use the term "single-cell data" to refer to single-cell RNA-sequencing data specifically in the remainder of this article.

2017). Hence, as we will review later, there is a rich line of work that extends hierarchical models of this type to analyze single-cell data by replacing the Gaussian distribution with more appropriate exponential-family distributions (Pierson and Yau, 2015; Townes et al., 2017; Durif et al., 2017; Risso et al., 2018), of which this article continues.

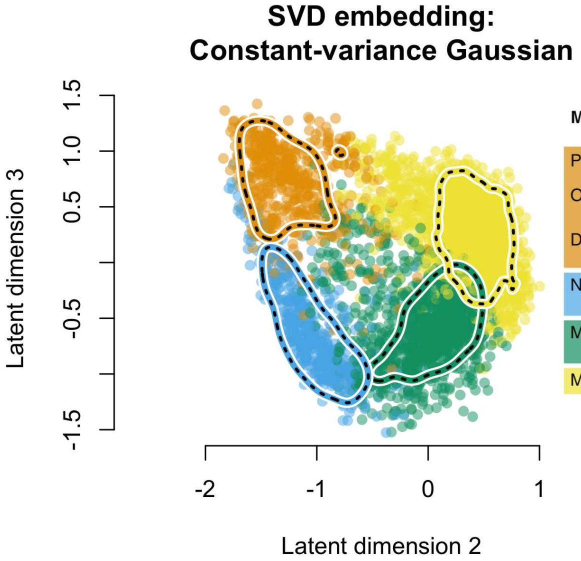
The aforementioned work often add additional nuances on top of the hierarchical model in order to model single-cell data better, but this often results in complicated estimators that become too intractable to statistically analyze. Hence, we design the eSVD in such a way that the posited statistical model retains the most important aspects common to the aforementioned work, while we can leverage recent theoretical developments to analyze the estimator's statistical properties. Specifically, the eSVD uses alternating minimization, a popular and computationally efficient approach used in the matrix factorization literature to solve the nonconvex optimization problem at hand, described in Section 4. We present the eSVD's statistical theory in Section 5, which builds upon the theoretical analyses of Zhao et al. (2015) and Lei (2018). These statistical properties include identifiability conditions and consistency, which ensure that researchers well-understand the estimated embedding and have a solid statistical foundation to build downstream analyses on top of. However, to ensure that the eSVD does not sacrifice too much modeling flexibility for theoretical tractability, we compare the eSVD to competing methods used to analyze single-cell data in Section 6 using synthetic data.

Finally, we return to our preliminary analysis of oligodendrocytes in Section 7, where we show that the eSVD embedding improves our analysis of cell developmental trajectories to match the latest scientific findings. These trajectories explain the heterogeneity among the oligodendrocytes by describing the smooth transition of gene expression among individual cells along a continuum, reflecting the cells' gradual transcriptional changes during development (Trapnell et al., 2014). Although early research suggest oligodendrocytes develop along a single trajectory (Kessaris et al., 2006), recent work suggest that oligodendrocytes could potentially branch out into various mature types (Marques et al., 2016; van Bruggen et al., 2017; Marques et al., 2018). Our improved analysis match these findings – we show the eSVD embedding estimates two distinct trajectories. We develop visualization tools to show our developmental trajectory findings, and conclude in Section 8 with practical extensions and theoretical questions left open for future work. While we focus on using the eSVD embedding to estimate cell developmental trajectories in this article, we emphasize that this embedding can be used for other applications highlighted earlier in this section, and provide additional analyses on another single-cell dataset in the appendix.

2 Preliminary analysis

We analyze a dataset of oligodendrocytes from mice brains collected by Marques et al. (2016) as a prototypical example to demonstrate shortcomings of the SVD embedding when applied to single-cell data. This dataset, henceforth called the Marques dataset, contains the gene expression of 5,069 oligodendrocytes that are clustered into thirteen cell sub-types using a biclustering algorithm (Zeisel et al., 2015) in Marques et al. (2016). These thirteen cell sub-types were later grouped into six major cell types and manually labeled based on cell-type specific marker genes (Zhang et al., 2014), shown in Figure 1.. We preprocess the data by selecting 983 highly informative genes, normalizing each cell by its library size (i.e., total counts across all genes), and $\log_2$-transforming each entry. These details are described in Appendix C. As suggested in the literature, it is common to $\log_2$-transform the gene expression matrix prior to using the SVD, since the $\log_2$-transformation can ideally stabilize the variance (Townes et al., 2017; Butler et al., 2018). However, as we see in the preliminary analysis in this section, this analysis strategy will result in modeling concerns that we wish to remedy in the rest of this article.

We review the SVD embedding, as it provides motivation for the eSVD in the next



Major cell type (Abbreviation)	Cell count	Number of sub-types
Pdgfra+ Precusor (Pdgfra+)	76	1
Oligodendrocyte precusor cells (OPC)	310	1
Differentiation-committed oligodendrocyte precursors (COP)	140	1
Newly formed oligodendrocytes (NFOL)	512	2
Myelin-forming oligodendrocytes (MFOL)	1283	2
Mature oligodendrocytes (MOL)	2748	6

Figure 1: The SVD embedding of the oligodendrocytes from Marques et al. (2016) after preprocessing including a $\log_2$-transformation, shown alongside a table summarizing the cell types. The six major cell types are listed in the table with the number of cells in each type, along with how they are differentiated into the thirteen different cell sub-types. The rows are organized from the "youngest" cell types to "most mature" cell types from top to bottom. The youngest three major cell types are colored orange. while the oldest three are colored blue, green and yellow respectively. The second and third latent dimensions are shown on the left, along with contours of the estimated densities to visualize high-density regions (one for each color of cells).

section. Let $A \in \mathbb{R}^{n\times p}$ represent the observed single-cell RNA-sequencing data matrix with rank m, where n is the number of cells and p is the number of genes. Here, loosely speaking, each entry A_{ij} measures how many instances of genetic material for gene j is observed for cell i after pre-processing. Let the SVD of A be denoted as $\widehat{U}\widehat{D}\widehat{V}^\top$ where $\widehat{U} \in \mathbb{R}^{n\times m}$ and $\widehat{V} \in \mathbb{R}^{p\times m}$ are both orthonormal matrices and $\widehat{D} \in \mathbb{R}^{m\times m}$ is a diagonal matrix. For a given latent dimensionality $k \leq m$, the SVD embedding for each cell $i \in \{1, \dots, n\}$ (denoted as

$\widehat{X}_i \in \mathbb{R}^k$) and each gene $j \in \{1, \ldots, p\}$ (denoted as $\widehat{Y}_j \in \mathbb{R}^k$) is defined as

$$\widehat{X}_i = \left(\frac{n}{p}\right)^{1/4} \cdot \left(\sqrt{\widehat{D}_{1,1}} \cdot \widehat{U}_{i,1}, \ldots, \sqrt{\widehat{D}_{k,k}} \cdot \widehat{U}_{i,k}\right), \quad \text{for } i \in \{1, \ldots, n\} \tag{2.1}$$

$$\widehat{Y}_j = \left(\frac{p}{n}\right)^{1/4} \cdot \left(\sqrt{\widehat{D}_{1,1}} \cdot \widehat{V}_{j,1}, \ldots, \sqrt{\widehat{D}_{k,k}} \cdot \widehat{V}_{j,k}\right), \quad \text{for } j \in \{1, \ldots, p\}. \tag{2.2}$$

We can see that the SVD embedding is a linear embedding since $\widehat{X}_i$ is the first k elements of the ith row in $(n/p)^{1/4} \cdot A\widehat{V}\widehat{D}^{-1/2}$. A scatterplot of the second and third latent dimensions of such an embedding is shown in Figure 1. Later in this article, we will show that this embedding implicitly assumes a constant-variance Gaussian distribution in Section 3, and show that this particular formulation handles identifiability concerns discussed in Section 4.

Now, we show that the SVD embedding (and its equivalent reparameterizations) does not model the data well, which could produce misleading results in downstream analyses. First, we visualize the quality of fit of the SVD embedding by purposefully omitting a small subset of randomly-selected entries in A and estimating the embedding as a matrix-completion problem. We can then assess the quality of fit of the embedding by comparing the values of these omitted entries in A to their predicted values. Figure 2 demonstrates this diagnostic, where the left plot shows the observed values in A that are not omitted (i.e., the "training set") verses their respective predicted values, while the right plot shows the observed values that are omitted (i.e., the "testing set") verses their respective predicted values. This missing-value diagnostic is commonly used both for assessing the quality of fit as well as for model selection (Li et al., 2020), and we will return to it in detail in Section 4. We see that while the embedding's performance on the testing set is more-or-less equivalent to that on the training set, the variability decreases as the gene expression increases. This is in opposition with the working model in the literature that suggests that larger gene expressions should be more variable than smaller ones (Witten, 2011; Risso et al., 2018). In fact, prior to taking the $\log_2$-transformation, Figure 3 demonstrates that the

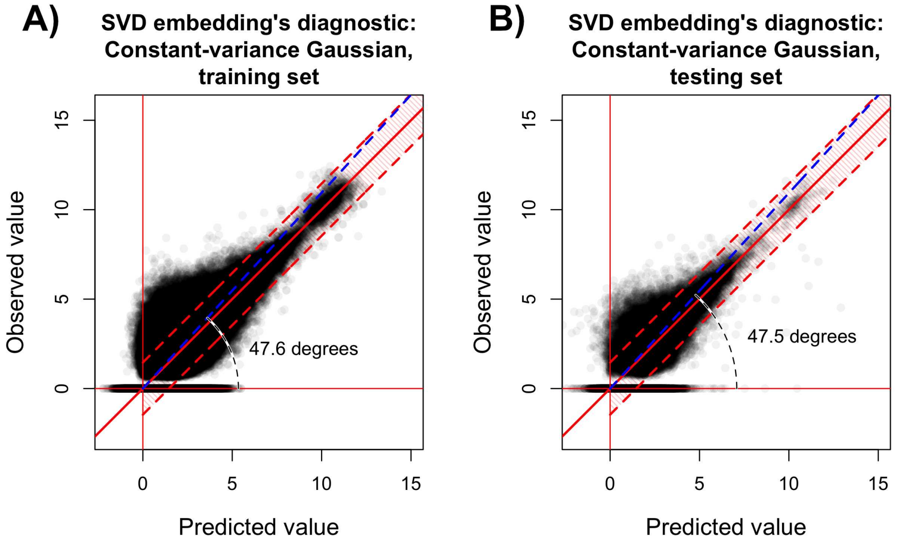


Figure 2: Diagnostic based on matrix completion to assess the fit using the SVD embedding for either the observed values that are not omitted (i.e., the “training set”) (A) or the observed values that are purposefully omitted (i.e., the “testing set”) (B), both verses their respective predicted values. This embedding is estimated using softImpute (Mazumder et al., 2010). The shaded red region is centered around the identity function (the ideal mean function) and marks the 10th to 90th quantiles of the constant-variance Gaussian model (based on the empirical variance) for different values of the predicted mean. The blue dotted line represents the principal angle between the observed values and their predicted value counterparts, where we mark its divergence from the identity function’s 45°. More details of this diagnostic is discussed in Section 4, while details of the fitting process using softImpute can be found in Appendix C.

variance in gene expression increases with its mean, reinforcing this model. Combined, all the diagnostics demonstrate that applying a $\log_2$-transform and using the SVD in conjunction seem to distort the properties of the data. This inspires us to develop a more appropriate embedding method that still retains desirable statistical properties. In the next section, we review the optimization problem that the SVD solves and see how it can be extended to one-parameter exponential families more generally, which will motivate our method, the eSVD.

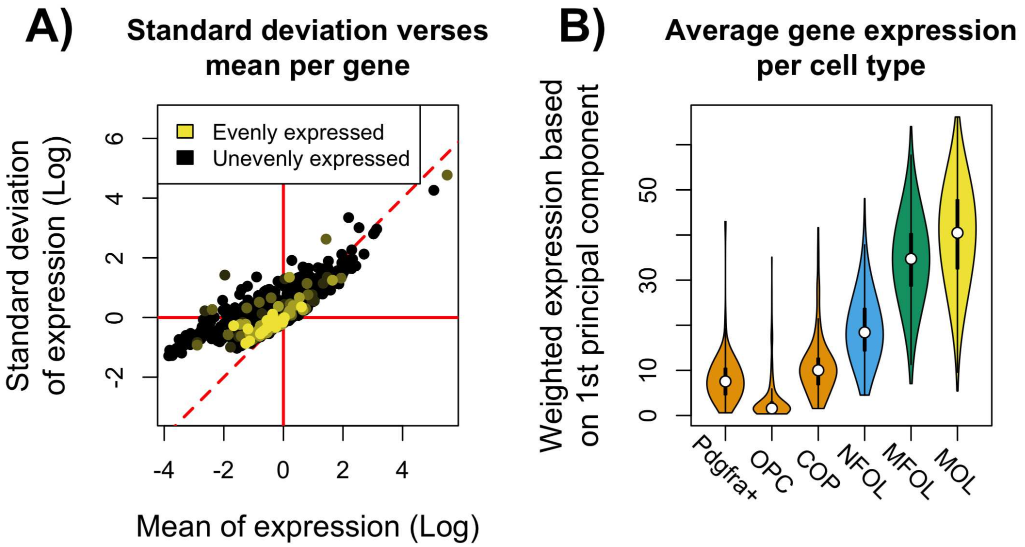


Figure 3: (A) The standard deviation of the expression verses the mean expression across all the cells, where each point represents one of the 983 genes in the preprocessed single-cell dataset. The color of each point depends on how evenly the gene is expressed among each of the six oligodendrocyte cell types show in Figure 1. The solid red horizontal and vertical lines and the dashed red line denoting the line $y = x$ are for visual reference. (B) Violin plot of the average expression of the genes reweighted according to the first principal component among the six oligodendrocyte cell types, using the color scheme in Figure 1. The statistics in both plots are computed prior to taking the $\log_2$-transformation, and is shown on the logarithm scale in Plot A purely for visualization purposes. More details about these plots are in Appendix C.

3 Statistical model and background

In this section, we explain the random dot product model that we investigate in this article, and its relation to other work.

3.1 Statistical model and estimation strategy

We model the entries of the single-cell dataset $A \in \mathbb{R}^{n\times p}$ as conditionally independent random variables drawn from a random dot product model – a latent hierarchical model commonly used in other work (Pierson and Yau, 2015; Townes et al., 2017; Durif et al., 2017; Risso et al., 2018). Specifically, for an appropriate one-parameter exponential-family distribution F parameterized by its natural parameter θ_{ij}, we impose model,

$$\begin{aligned} A_{ij} \sim F\big(\theta_{ij} = X_i^\top Y_j\big), &\quad \text{for } (i,j) \in \{1,\ldots,n\}\times\{1,\ldots,p\}, \\ \text{where} \quad X_1,\ldots,X_n \overset{i.i.d.}{\sim} G, &\quad \text{and} \quad Y_1,\ldots,Y_p \overset{i.i.d.}{\sim} H. \end{aligned} \tag{3.1}$$

where G and H represent two latent k-dimensional distributions, where k is much smaller than n or p. We assume all the latent random vectors X_i's and Y_j's are jointly independent, and the observed A_{ij}'s are independent conditioned on the X_i's and Y_j's. Let the density of the exponential-family distribution F be denoted as

$$p(A_{ij} \mid \theta_{ij}) = h(A_{ij}) \exp\big(T(A_{ij})^\top \eta(\theta_{ij}) - g(\theta_{ij})\big), \tag{3.2}$$

where $g(\cdot)$ is a known log-partition function for F with a domain $\mathcal{R}$, $\eta(\cdot)$ is a known natural parameter function, and $T(\cdot)$ is a known sufficient statistic function. For notational convenience, we denote $X \in \mathbb{R}^{n\times k}$ and $Y \in \mathbb{R}^{p\times k}$ as the matrices that collect all the latent vectors $X_1,\ldots,X_n$ and $Y_1,\ldots,Y_p$ row-wise, and denote $\Theta = XY^\top \in \mathbb{R}^{n\times p}$ as the rank-k natural parameter matrix that collects all elements θ_{ij}. Given the exponential-family form for F shown in (3.2), we need to impose the following assumption to ensure the inner products $X_i^\top Y_j$ for all $(i,j) \in \{1,\ldots,n\}\times\{1,\ldots,p\}$ yield valid natural parameters,

Assumption 3.1 (Bounded inner product)**.** *Let $\mathcal{R}$ denote the domain of the natural param-*

eters for the distribution F. Assume that for any $X_i \sim G$ and $Y_j \sim H$,

$$\mathbb{P}(X_i^\top Y_j \in \mathcal{R}) = 1, \quad \textit{almost surely.}$$

Given the above model, our goal is to estimate the latent random vectors $X_1, \ldots, X_n$ since these latent vectors represent the low-dimensional embedding of all n cells. For a given exponential-family distribution F, an intuitive strategy for estimating our desired embedding is to minimize the negative log-likelihood based on the observed data A over all possible vectors $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_p$. Specifically, plugging into the exponential-family form (3.2) into the model (3.1), we derive the loss function

$$\mathcal{L}_n(X, Y) = \frac{1}{np} \sum_{(i,j)} \Big[g(X_i^\top Y_j) - T(A_{ij})^\top \eta(X_i^\top Y_j) \Big], \tag{3.3}$$

with the constraints $X_i^\top Y_j \in \mathcal{R}$ for all pairs (i, j). The above loss function is nonconvex, but if F is the the constant-variance Gaussian distribution, this loss function is proportional to

$$\frac{1}{np} \sum_{(i,j)} (A_{ij} - X_i^\top Y_j)^2,$$

which is specifically what the SVD minimizes (Maezika, 2016). This particular model is convenient to use since the SVD provides a closed-form solution to the corresponding nonconvex optimization problem, and leads to the SVD embedding shown in (2.1).

3.2 Relation to other work modeling single-cell data

As we have discussed previously in Section 2, this constant-variance assumption is too restrictive to properly model single-cell data. Hence, many articles cited above replace the constant-variance Gaussian distribution with other exponential-family distributions for F

that allow the variance to increase with the mean. For example, Witten (2011) and Risso et al. (2018) consider the Poisson and negative binomial distribution specifically, after a suitable transformation of the natural parameters. Additionally, these models often add other random effects on top of the existing random dot product model (3.1) that influence the entries in A. For example, many methods like pCMF (Durif et al., 2017) allow researchers to incorporate *dropout* into the model – a characteristic of single-cell data where a substantial fraction of the gene expression for a cell is recorded as exactly 0 due to low amounts of RNA in the cell (Kharchenko et al., 2014). Other methods such as ZINB-WaVE (Risso et al., 2018) go further and allow covariate information such as gene length and cell size. Most recently, Lopez et al. (2018) use deep autoencoders to estimate the embedding. However, there is often a lack of theoretical analyses for the aforementioned estimators. This is because replacing the exponential-family distribution F with any distribution aside from the constant-variance Gaussian distribution leads to non-trivial nonconvex estimators that minimize the loss function shown in (3.3), which make traditional statistical techniques for analyzing these estimators unsuitable. Therefore, concerns such as identifiability are typically not addressed theoretically, leading to ambiguity on performance of the estimators of such models.

To resolve this theoretical ambiguity, we design the eSVD to estimate the embedding based on the random dot product model (3.1) for any exponential-family distribution F with no other random effects. In this way, we can tractably analyze the statistical properties of eSVD while retaining abundant flexibility to effectively model single-cell data.

3.3 Matrix factorization

To the best of our knowledge, the first statistical results for estimators that extended the SVD to generic exponential-family distributions by minimizing a loss function similar to

(3.3) come from Gunasekar et al. (2014) and Lafond (2015). There, the authors minimize the loss function over the natural parameter matrix Θ and add a trace penalization term to encourage the estimate to be low-rank. While this formulation yields a convex optimization problem, it requires solving a semidefinite program which can be computationally prohibitive for large datasets. This consideration has motivated researchers to investigate the statistical properties of estimators that minimize the loss function (3.3) directly as a non-convex optimization problem. Specifically, *alternating minimization* is a suitable candidate for this task, where each iteration alternates between optimizing either one of two low-rank matrices X and Y while treating the other fixed. This algorithmic strategy pre-dates the convex relaxation approach; see Collins et al. (2002), Jain et al. (2013), Udell et al. (2016), Landgraf and Lee (2019) and the references within for discussions and additional variants. From an algorithmic standpoint, our method is a direct continuation of such work. However, the statistical properties of such estimators have only recently been characterized rigorously. For example, to accommodate the constraints in Assumption 3.1, Wang et al. (2016), Yu et al. (2020) and Chi et al. (2019) adapt the theoretical framework to study slightly different estimators based on alternating projected gradient descent. In contrast, our work will build on techniques used in Zhao et al. (2015) and Balakrishnan et al. (2017) to retain our focus on alternating minimization.

However, all the aforementioned theoretical results do not directly apply the random dot product model (3.1), which contains an additional source of randomness induced by the hierarchical structure. In contrast, our theory is able to account for this additional source of randomness by drawing upon connections to the network literature. Specifically, the random dot product model (3.1) is similar to those used in latent position random graphs studied in the network literature (Hoff et al. (2002) and Athreya et al. (2017)). Hence, we draw inspiration from Lei (2018) on how to address these identifiability concerns and to develop proof techniques in this article.

There are many other embedding methods in the literature more broadly, such as non-negative matrix factorization, kernel PCA, and manifold-based embedding methods such as UMAP and Isomap. We defer a thorough discussion contrasting eSVD with such estimators to Appendix B.

4 Method: eSVD (Exponential-family SVD)

We describe the eSVD in this section, which is designed to be a general framework to minimize the loss function (3.3) for any choice of a one-parameter exponential-family distribution F. To keep the presentation clear, we describe some of the more nuanced implementation details in Appendix B. We also describe an important diagnostic to assess the quality of fit, as demonstrated in Section 2. This diagnostic can be also be used as a tuning procedure to select the most appropriate choice for F or nuisance parameters.

4.1 eSVD and its application to the curved Gaussian distribution

Similar to other nonconvex matrix factorization methods (Wang et al., 2016; Yu et al., 2020), our method requires an initial estimate of the rank-k matrix of natural parameters, $\widehat{\Theta}'$, where k is pre-determined. To achieve this, we use an initialization method based on Wang et al. (2016). To simplify the presentation here, we provide details in Appendix B.1. This initialization scheme performs a rank-k SVD based on transforming each entry of A via the inverse of the log-partition function $g(\cdot)$. Given this initial estimate, consider its SVD $\widehat{\Theta}' = UDV^\top$. To start the alternating minimization stage of our method, we set $\overline{Y}^{(0)} = V$.

After initialization, the eSVD then refines the estimate by performing alternating minimizations. Denoting a generic matrix and its SVD by $\widehat{\Theta} = UDV^\top$, let $\text{LeftSVD}(\widehat{\Theta}) = U$, the function that maps a matrix to its left singular vectors. Then, for iterations $t \in \{0, \ldots, T-1\}$,

$$
\begin{aligned}
X^{(t+1)} &= \operatorname*{argmin}_{X \in \mathbb{R}^{n \times k}} \mathcal{L}_n(X, \overline{Y}^{(t)}) \ : \ X_i^\top \overline{Y}_j^{(t)} \in \mathcal{R}, \quad \forall (i,j), && (4.1)\\
\overline{X}^{(t+1)} &= \sqrt{n} \cdot \text{LeftSVD}(X^{(t+1)}), && (4.2)\\
Y^{(t+1)} &= \operatorname*{argmin}_{Y \in \mathbb{R}^{p \times k}} \mathcal{L}_n(\overline{X}^{(t+1)}, Y) \ : \ (\overline{X}_i^{(t+1)})^\top Y_j \in \mathcal{R}, \quad \forall (i,j), && (4.3)\\
\overline{Y}^{(t+1)} &= \sqrt{p} \cdot \text{LeftSVD}(Y^{(t+1)}). && (4.4)
\end{aligned}
$$

After all T iterations, the eSVD outputs the final estimate after a reparameterization. That is, letting $\widehat{\Theta}^{(T)} = \overline{X}^{(T)}(Y^{(T)})^\top$ have a rank-k SVD of $\widehat{U}\widehat{D}\widehat{V}^\top$, the final estimates are

$$
\begin{aligned}
\widehat{X}_i &= \left(\frac{n}{p}\right)^{1/4} \cdot \left(\sqrt{\widehat{D}_{1,1}} \cdot \widehat{U}_{i,1}, \ldots, \sqrt{\widehat{D}_{k,k}} \cdot \widehat{U}_{i,k}\right), \quad i = 1, \ldots, n, && (4.5)\\
\widehat{Y}_j &= \left(\frac{p}{n}\right)^{1/4} \cdot \left(\sqrt{\widehat{D}_{1,1}} \cdot \widehat{V}_{j,1}, \ldots, \sqrt{\widehat{D}_{k,k}} \cdot \widehat{V}_{j,k}\right), \quad j = 1, \ldots, p. && (4.6)
\end{aligned}
$$

This is the same reparameterization used in (2.1) and (2.2).

Remarks about algorithmic design. We make a few remarks about the design of our algorithm. Optimizing over X and Y directly raises identifiability issues, since for any orthogonal matrix Q, $\mathcal{L}_n(XQ, YQ^\top) = \mathcal{L}_n(X, Y)$. To address this, Ge et al. (2017) append a penalty term

$$\frac{1}{8}\|X^\top X - Y^\top Y\|_F^2,$$

while Zhao et al. (2015) use the QR-decomposition between iterations, and we use the LeftSVD$(\cdot)$ operator. In practice, we found all three choices behave similarly. The factors $\sqrt{n}$ and $\sqrt{p}$ in (4.2) and (4.4) are included for theoretical reasons to ensure the spectrum of the Hessian is well-controlled and to ensure the values do not underflow if n or p are too large empirically. Also, the final reparameterizations in (4.5) and (4.6) are designed such that the sample second-moment matrices of $\widehat{X}$ and $\widehat{Y}$ are both equal and diagonal, i.e.,

$$\frac{1}{n}\widehat{X}^\top \widehat{X} = \frac{1}{p}\widehat{Y}^\top \widehat{Y},$$

which is important for our statistical analysis later on.

Lastly, to perform the constrained optimization (4.1) and (4.3), we use a first-order method called Frank-Wolfe (Jaggi, 2013), which we found more stable compared to using projected gradient descent. While there are theoretical guidelines for choosing step-sizes related to the convexity and smoothness for this method, we found these choices often led to poor empirical performance.

Example with the curved Gaussian distribution. To make the eSVD's workings more concrete, we demonstrate what the minimization in (4.1) would entail when we set F to be the *curved Gaussian* distribution. This will be useful later in this article when we use this distribution to analyze the oligodendrocytes. Specifically, we say the random variable A_{ij} follows a curved Gaussian distribution with a known parameter $\tau > 0$ if $A_{ij} \sim N(\mu_{ij}, \mu_{ij}^2/\tau^2)$, for an unknown mean parameter $\mu_{ij} > 0$ (Efron et al., 1978; Liu and Martin, 2020)[2]. This sets the standard deviation to be linearly proportional to the mean. Writing this distribution in exponential-family form (3.2) yields,

$$p(A_{ij} \mid \theta_{ij}) = \frac{\tau \exp(-\tau^2/2)}{\sqrt{2\pi}} \exp\Bigg(\begin{bmatrix} \tau^2 A_{ij} \\ -\tau^2 A_{ij}^2/2 \end{bmatrix}^\top \begin{bmatrix} \theta_{ij} \\ \theta_{ij}^2 \end{bmatrix} + \log(\theta_{ij}) \Bigg), \tag{4.7}$$

[2] We call it a curved Gaussian distribution since this distribution is a curved exponential-family distribution.

where the relation between the natural parameter θ_{ij} and the mean parameter μ_{ij} can be derived to be $\mu_{ij} = 1/\theta_{ij}$. Here, the domain of the natural parameters would be $\mathcal{R} = \mathbb{R}_+$, the positive half-line. After simple calculations, one can derive the negative log-likelihood and conclude that the minimization in (4.1) becomes

$$X^{(t+1)} = \operatorname*{argmin}_{X \in \mathbb{R}^{n \times k}} \frac{1}{np} \sum_{(i,j)} \Big[-\log\big(X_i^\top \overline{Y}_j^{(t)}\big) - \begin{bmatrix} \tau^2 A_{ij} \\ -\tau^2 A_{ij}^2/2 \end{bmatrix}^\top \begin{bmatrix} X_i^\top \overline{Y}_j^{(t)} \\ (X_i^\top \overline{Y}_j^{(t)})^2 \end{bmatrix} \Big].$$

Analogous calculations for other common distributions are shown in Appendix B.

The curved Gaussian distribution is relevant in practice since if $\tau \geq 2$, this distribution reflects the phenomenon that genes with larger expression also exhibit larger variance, while most of the distribution's mass is still positive. Additionally, in many instances, this distribution can capture more variability than the Poisson and negative binomial distributions, which can be beneficial when the single-cell data is intrinsically noisier. In general however, if the researcher wants to use the eSVD for an arbitrary one-parameter exponential-family distribution F, all she needs to pass into our implementation is the computation of the loss function (3.3), its gradients, and information about the domain $\mathcal{R}$.

4.2 Matrix-completion diagnostic and tuning procedure

We provide the following diagnostic to assess the embedding's quality of fit or to determine which choice of F is most appropriate for our data, which was used in Figure 2. Inspired by network cross-validation work such as Li et al. (2020), we use matrix completion to determine the quality of our model fit. As alluded to in Section 2, to do this, we omit a small percentage of the entries of A when estimating the embedding and compare these values to their predicted expected value counterparts. To compute this expected value, recall that for exponential-family distributions (3.2), $g'(\cdot)$ (the derivative of the log-partition function $g(\cdot)$) maps the natural parameter to the expected value. We note that we are able to adopt this matrix completion strategy to tune our method since our alternating minimization procedure can be adapted to handle missing values. In contrast, embedding methods like ZINB-WaVE (Risso et al., 2018) and PCMF (Durif et al., 2017) do not offer an analogous tuning procedure. This procedure is formalized below.

1. For bootstrap trials $b \in \{1, \ldots, B\}$:

    (a) Randomly sample m of the entries of A, denoted as $\mathcal{O} = \{(i_1, j_1), \ldots, (i_m, j_m)\}$, which will be omitted in the following estimation step. Here, m can be any small number, such as $\lceil 0.01 \cdot (np) \rceil$.

    (b) Estimate the latent vectors by $\widehat{X}$ and $\widehat{Y}$ according to Subsection 4.1 where the loss function (3.3) omits the entries in $\mathcal{O}$ and is parameterized based on the desired distribution of F.

    (c) Compute v_1, defined as the leading eigenvector of the matrix formed by the omitted observed values $A_{\mathcal{O}} = \{A_{i_1,j_1}, \ldots, A_{i_m,j_m}\}$ and their predicted expected value counterparts $g'(\widehat{X}\widehat{Y}^\top)_{\mathcal{O}} = \{g'(\widehat{X}_{i_1}^\top \widehat{Y}_{j_1}), \ldots, g'(\widehat{X}_{i_m}^\top \widehat{Y}_{j_m})\}$.

    (d) Compute model fit quality, $q^{(b)}$ defined as the angle between v_1 and the vector $(1, 1)$, representing the identity function.

2. Average the model fit qualities across all trials, $q^{(1)}, \ldots, q^{(B)}$.

Observe that we define the quality of fit $q^{(b)}$ above by how much the leading eigenvector v_1 deviates from 45°. This angle of v_1 is what we called the *principal angle* in Figure 2. Having an eigenvector's angle close to 45° means that on average, the predicted values correspond closely with the observed value. The advantage of this quality-of-fit's definition is that it is easily comparable even across different distributions F, unlike the negative log-likelihood or MSE.

While we advocate constructing plots such as Figure 2 to obtain a more holistic sense of how well the embedding fits the data in general, we can also use the above procedure to obtain an automated model selection method in the following way – if we try the above diagnostic for multiple distributions for F, the distribution that yields the smallest average of $q^{(1)}, \dots, q^{(B)}$ is deemed the most appropriate model for A. In this way, we can also use this diagnostic as a tuning procedure to select the dimensionality of the latent space k or nuisance parameters for exponential-family distributions such as τ in the curved Gaussian distribution (4.7) in a grid-search fashion. An additional variant of this tuning procedure is detailed in Appendix B.

5 Statistical theory

In this section, we prove the consistency for the eSVD when applied to the random dot product model (3.1), which is important to ensure that the eSVD is estimating a well-defined quantity. This result provides the needed statistical foundation for downstream tasks such as clustering and RNA velocity, as mentioned in Section 1. Additionally, our theory gives us better insights into the eSVD since our analysis also reveals the identifiability conditions that formalize to what degree the embedding can be estimated. While our theorems currently assume the correctly-specified setting (i.e., the eSVD's choice of F matches the true generating distribution and k is correctly-specified), we hope these theorems provide a roadmap for future work to prove analogous statements for broader settings or for more complex methods like ZINB-WaVE which currently do not exist.

We discuss the additional notation here. For a generic matrix A, let $\|A\|_F$ denote its Frobenius norm. For two sequences a_n and b_n and two random sequences A_n and B_n, let $a_n = O(b_n)$ and $A_n = O_P(B_n)$ denote that a_n/b_n or A_n/B_n is bounded for large enough n deterministically or in probability respectively. For all $i \in \{1, \dots, n\}$ and $j \in \{1, \dots, p\}$, let the the population second moment matrices of X_i and Y_j and the corresponding eigendecompositions be defined respectively as

$$\mathbb{E}[X_i X_i^\top] = C_X^* = \Phi^* \Lambda^* \Phi^{*\top}, \quad \text{and} \quad \mathbb{E}[Y_j Y_j^\top] = C_Y^* = \Psi^* \Gamma^* \Psi^{*\top}.$$

Our proposition below requires the following assumptions.

Assumption 5.1 (Sub-Gaussian distribution of latent vectors)**.** *Assume that X_i for all $i \in \{1, \dots, n\}$ are i.i.d. sub-Gaussian random vectors. That is, there exists a fixed constant D such that for any vector $v \in \mathbb{R}^k$ where $\|v\|_2 = 1$ and any integer $c \geq 1$, $(\mathbb{E}[|X_i^\top v|^c])^{1/c} \leq Dc^{1/2}$, and a similar assumption holds for Y_j for all $j \in \{1, \dots, p\}$ with also the same constant D.*

Assumption 5.2 (Second moment properties)**.** *First, assume the population second moment matrices C_X^* and C_Y^* are equal and are both diagonal matrices, where $(C_X^*)_{i,i} \geq (C_X^*)_{j,j}$ for any $1 \leq i < j \leq k$. Second, assume there exists positive numbers $c_1 \leq c_2$ and $1 < \alpha \leq \beta$ such that for all $\ell \in \{1, \dots, k\}$, the eigenvalues satisfy*

$$c_1 \ell^{-\alpha} \leq \lambda_\ell^* \leq c_2 \ell^{-\alpha}, \quad \textit{and} \quad \lambda_\ell^* - \lambda_{\ell+1}^* \geq c_1 \ell^{-\beta},$$

with the convention that $\lambda_{k+1}^* = 0$.

Both assumptions are common in work that study the spectrum associated with random dot product models (Lei, 2018). Assumption 5.1 assumes G and H are sub-Gaussian distributions, which enables sharp rates for estimating their second-moment matrices C_X^* and C_Y^* respectively (Vershynin, 2012). On the other hand, the second part of Assumption 5.2 enables our estimator to accurately estimate its eigenvalues and eigenvectors. Importantly however, the first part of Assumption 5.2 can be interpreted instead as an identifiability condition, which we formalize below.

Proposition 5.1. *Given two k-dimensional distributions G and H, each with at least two moments where the population second moment matrices are full rank, consider two independent random variables $X' \sim G$ and $Y' \sim H$. Then there exists a linear and invertible transformation R such that the population second moment matrices of $X = RX'$ and $Y = R^{-\top}Y'$ are the same, i.e.,*

$$\mathbb{E}[XX^\top] = \mathbb{E}[YY^\top].$$

Furthermore, both population second moment matrices of X and Y are diagonal matrices.

The proof is in Appendix I, which provides an explicit construction of the matrix R. Note that since R is invertible, we guarantee that the distribution of the inner product is preserved, i.e.,

$$\mathbb{P}\big((X')^\top Y' \leq t\big) = \mathbb{P}\big(X^\top Y \leq t\big), \quad \forall t \in \mathcal{R}.$$

Hence, the first part of Assumption 5.2 can be interpreted as an identifiability condition, since we can only estimate G and H only up to a linear transformation.

Proposition 5.2. *Assume the model in (3.1) where Assumptions 5.1 and 5.2 hold. If the estimator $\widehat{\Theta}$ satisfies $\|\Theta - \widehat{\Theta}\|_F \leq \epsilon$ conditioned on X and Y, and $k = o(\min\{n,p\})$, then up to sign,[3] eSVD achieves the rate after reparameterizations (4.5) and (4.6),*

$$\frac{1}{n}\|X - \widehat{X}\|_F^2 = O_P\Big(\max\Big\{\frac{k^{4\beta-\alpha+4}}{\min\{n,p\}}, \ \frac{k^{2\beta-\alpha+2}\max(\epsilon^2,\epsilon)}{np}\Big\}\Big). \tag{5.1}$$

Discussion of consistency. Assuming k is fixed, the above proposition states that the eSVD embedding is consistent as long as ϵ (the rate of convergence for the matrix of natural parameters, $\|\Theta - \widehat{\Theta}\|_F$) is faster than $O_P(\sqrt{np})$. We formalize this statement in Appendix D, where we add assumptions common to the literature such as strong convexity and smoothness

[3] We use "up to sign" similar to Fan et al. (2018), where each column of $\widehat{X}$ can be multiplied by ±1 since the SVD is not unique.

of the negative log-likelihood function associated with F. The details are deferred because these assumptions are technical to describe and detract from the main text. More generally, however, the above proposition addresses the additional source of randomness induced by the random dot product model mentioned in Section 3 that other theoretical investigations typically do not address. Therefore, *any* such estimator equipped with a rate for $\|\Theta - \widehat{\Theta}\|_F$ can be plugged into Proposition 5.2 directly. Additionally, note that a similar rate (5.1) holds for estimating Y, the embedding of the genes.

Application to curved Gaussian model. While Proposition 5.2 and the results in Appendix D apply for any generic one-parameter exponential-family distribution F satisfying certain conditions, we specifically apply these results to the curved Gaussian distribution (4.7) in Appendix E to demonstrate what the rates are for a particular exponential-family distribution. At a high level, we show that when n and p grow asymptotically at the same rate and k is fixed,

$$\frac{1}{n}\|X - \widehat{X}\|_F^2 = O_P\Big(\frac{\log^{1/2}(n)}{n^{1/2}}\Big).$$

6 Numerical study

In this section, we study the performance of the eSVD and other competitive methods based on synthetic data. Our setup for all the simulations in this section are as follows: based on model (3.1), we set the dimensionality to $k = 2$ and sample $X_1, \ldots, X_n$ i.i.d. uniformly from four connected linear segments (which we call the "trajectories") with additive Gaussian noise, as illustrated in Figure 4. These four segments loosely represent four cell types. We also sample $Y_1, \ldots, Y_p$ i.i.d. from a mixture of two Gaussians. These sampling procedures represent G and H respectively, up to identifiability conditions. We enforce $X_i^T Y_j \in \mathcal{R}$ for all pairs (i, j). The distribution family F varies among different simulations. We do not

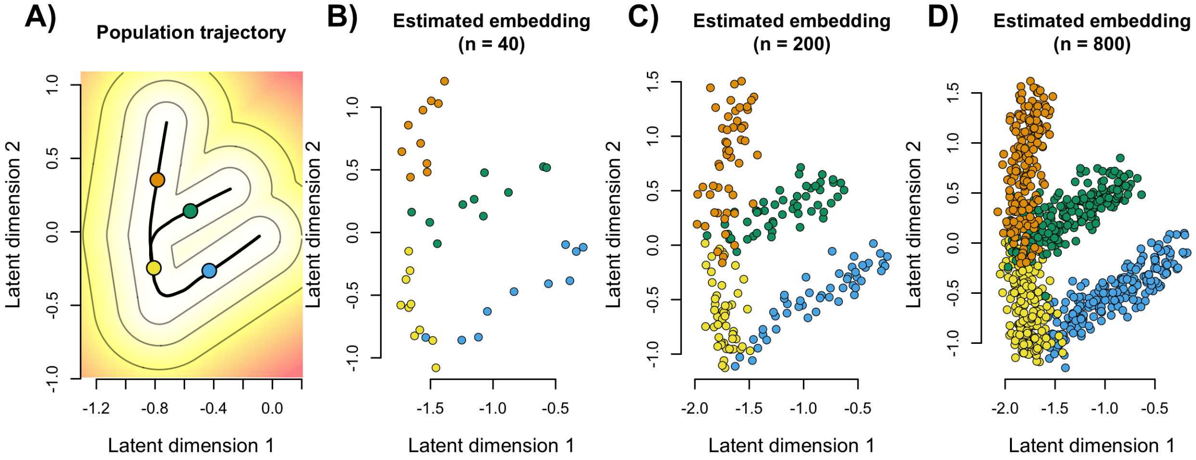


Figure 4: (A) The two-dimensional population density of G, visualized as a heat map with contour lines of the density along with the true "trajectories" (black lines). The mean vector for each of four cell types are labeled in a different color (blue, yellow, green, orange). (B to D) The estimated embedding $\widehat{X}_1, \ldots, \widehat{X}_n$ of the synthetically-generated A for varying levels of n, (i.e., number of cells or number of rows), colored by the true cell type, which are labels used only for visual reference and not used during estimation.

use R packages such as Splatter (Zappia et al., 2017) to generate our synthetic data because we want to have precise control over the true embedding. The full details of the simulation setups and usage of various estimators in this section are in Appendix F.

Consistency of the estimated embedding. In this first simulation suite, we demonstrate that the estimated embedding converges towards the true embedding. Specifically, we generate $A \in \mathbb{R}^{n \times p}$ where each entry A_{ij} is sampled independently from the negative binomial distribution with a natural parameter $\theta_{ij} = X_i^\top Y_j$ and $r = 50$, and fit the eSVD using the correctly specified model. Figure 4 is an illustration that demonstrates the asymptotic properties of the eSVD. Specifically, we see that the distribution of the embedding $\widehat{X}_1, \ldots, \widehat{X}_n$ approximates G as n increases. We provide details and additional results that verify the consistency of the eSVD's embedding in Appendix F.

Comparison of different embedding methods. In our second simulation suite, we demonstrate that the cells' latent positions estimated by the eSVD are more accurate in relation to one another than those estimated by other methods. Here, we fix $n = 200$ and $p = 400$. We compare the eSVD via the negative binomial distribution to nine other methods commonly used to embed single-cell data: ZINB-WaVE (Risso et al., 2018), pCMF (Durif et al., 2017), SVD, non-negative matrix factorization (NMF), independent component analysis (ICA), UMAP (Becht et al., 2019), t-SNE (Maaten and Hinton, 2008), Isomap (Tenenbaum et al., 2000) and diffusion map (Haghverdi et al., 2015). The first two methods and our tuning procedures are explained in Appendix F. Importantly, SVD implicitly assumes F is a constant-variance Gaussian distribution as mentioned in Section 3, while ZINB-WaVE and pCMF assume F is a negative binomial and Poisson distribution respectively.

We simulate data from a negative binomial model in this simulation, which is the distribution that is most commonly used to model sequencing data (Love et al., 2014). Specifically, we sample the observed count matrix A conditionally independent on $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_p$ where A_{ij}'s are sampled from a negative binomial distribution with natural parameter $X_i^\top Y_j$ and dispersion parameter $r = 50$. Then, when we estimate the embedding using eSVD, we use the tuning procedure mentioned in Section 4 to select the most appropriate value of the dispersion parameter r from the set $\{5, 50, 100\}$.

We find that on average across 100 trials, our method estimates the relative latent positions of each cell to be more accurate than other methods (Figure 5A). To define our notion of accuracy, consider each cell i and its Euclidean distance to all other $n - 1$ cells in the latent space in both the true and estimated embedding. We then compute the Kendall's tau correlation between these two vectors, which only relies on the ranks of the distances, and then average this value over all n cells. Hence, a high averaged Kendall's tau value suggests the latent positions of the n cells are well-estimated with respect to one another. We call this notion of accuracy the *relative embedding correlation.* We define our notion of accuracy in this way to ensure it is insensitive to arbitrary rotations or constant rescalings of the embedding. Figure 5B compares the different estimated embeddings to the true embedding as an illustration. Both the eSVD and ZINB-WaVE estimate embeddings where the four cell types are relatively in the correct configuration, and their accuracy are quite high. For the other methods, the high variability due to the overdispersion of the negative binomial seems to dramatically skew the embedding for certain cells. We defer additional simulations using other distributions F to Appendix F.

Investigation using misspecified models and the effect of k. In our third simulation suite, we empirically compare the eSVD to other methods when F is *misspecified*. We also demonstrate how different values of k can affect the quality of the embedding. While the theorems we developed in Section 5 do not currently handle such settings, these results help us understand how the eSVD performs in more realistic and more challenging scenarios. Due to space constraints, we defer these results to Appendix F, which show that the eSVD can roughly estimate the relative positions of the cells' embedding well compared to other methods despite the model misspecification. All-in-all, our takeaway message is that the eSVD's flexibility in choosing which exponential-family distribution F to use and the diagnostics provided in Section 4 allow our method to remain competitive among the ten methods.

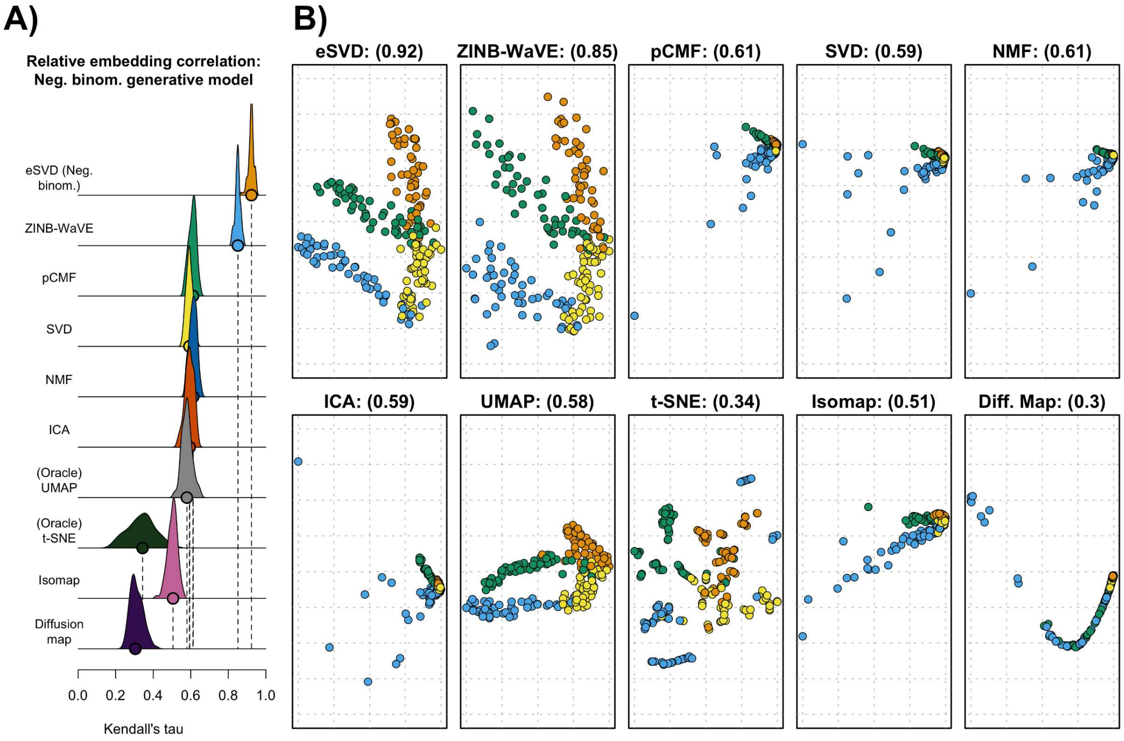


Figure 5: (A) The density plot of each embedding methods' accuracy (eSVD, ZINB-WaVE, pCMF, SVD, NMF, ICA, UMAP, t-SNE, Isomap and diffusion map), based on the relative embedding correlation. The circles along each method's x-axis denotes the median accuracy across the 100 trials. Here, both the data-generating distribution F and the eSVD use the negative binomial distribution. See Appendix F to see how we tuned the methods. (B) The ten estimated embedding, chosen among the trial with the median accuracy, noted in each plot's title in parenthesis. The coloring of the samples persists from Figure 4. The x- and y-axes represent the coordinate system estimated by the respective embedding methods, reflected in the dashed grids.

7 Single-cell analysis

We return to modeling the Marques dataset (Marques et al., 2016), as described in Section 2, to determine if the embedding based on the curved Gaussian distribution (4.7) is more appropriate than that based on the constant-variance Gaussian distribution, and if so, investigate how the embedding affects the downstream trajectory analysis. As alluded to in Section 2, the six major cell types in Figure 1 have a determined ordering, starting from Pdgfra+ precursors and ending with the mature oligodendrocytes. Our goal in this analysis is to estimate the trajectories among the cell sub-types constrained to this ordering. For example, in Marques et al. (2016), after embedding the cells into a latent space, the authors estimate one developmental trajectory connecting the first five major cell types starting from the Pdgfra+ precursors, but do not definitively conclude how the six mature oligodendrocyte cell sub-types differentiate. Instead, they relied on analyzing the percentage of different cell types across different brain regions to hypothesize that these six sub-types differentiate into multiple different trajectories.

7.1 Details of estimating cell developmental trajectories

We provide more details on how we estimate the developmental trajectories based on the low-dimensional embedding $\widehat{X}_1, \ldots, \widehat{X}_n$. As alluded in Section 1, these trajectories show how these different cell sub-types develop from one to another, assuming the latent vectors $\widehat{X}_i$ gradually change along the trajectories. Trajectory analyses are an important step in studying the cellular dynamics from single-cell data, as most single-cell technologies provide only a snapshot of all the cells. This is because most technologies destroy the cells during the sequencing step, which prevent longitudinal studies. In this article, we use Slingshot (Street et al., 2018) (with minor modifications) to estimate these cell developmental trajectories. Roughly speaking, Slingshot is a two-step algorithm that requires the latent vectors to already be clustered, where we treat each cell sub-type as a cluster. In the first stage, Slingshot estimates the number of trajectories and ordering of the cell sub-types based on minimizing the distances between cell sub-type centers via a minimum spanning tree. In the second stage, Slingshot fits variants of principal curves (Hastie and Stuetzle, 1989) that pass through the cell sub-type centers in the estimated ordering. These principal curves can be thought of as smooth curves that pass through high-density regions in the latent space. Throughout our analysis in this section, we apply Slingshot to the embedding using all latent dimensions, but only visualize the estimated trajectories with respect to the first three latent dimensions. More details about Slingshot and our modifications of it are given in Appendix G.

We briefly mention that the original study (Marques et al., 2016) uses the Monocle algorithm (Trapnell et al., 2014), to estimate the cell developmental trajectories. We use Slingshot instead as it is the current state-of-the-art method based on extensive benchmarking comparisons in Saelens et al. (2019).

7.2 Analysis using the constant-variance Gaussian distribution

Building on the analysis in Section 2, we perform a trajectory analysis using the SVD embedding shown in Figure 1 on the $\log_2$-transformed data, which assumes the constant-variance Gaussian model. Applying Slingshot directly to this embedding results in two trajectories, both heavily overlapping one another when visualized (Figure 6A). These results are similar to Marques et al. (2016) in two ways. First, the authors show that all cells develop from Pdgfra+ precursors to myelin-forming oligodendrocytes in the same way, which we estimate as well. Second, the authors do not definitively conclude if the mature oligodendrocytes diverge in their development. Our trajectories themselves also leave this ambiguity unresolved due to the heavy overlap between the two trajectories. However, we perform the following additional visual diagnostic to quantify if these two trajectories are well approximated by a single trajectory.

To formalize to what degree the different trajectories are the same, we use a bootstrap resampling procedure to construct a uniform uncertainty tube around each trajectory. These

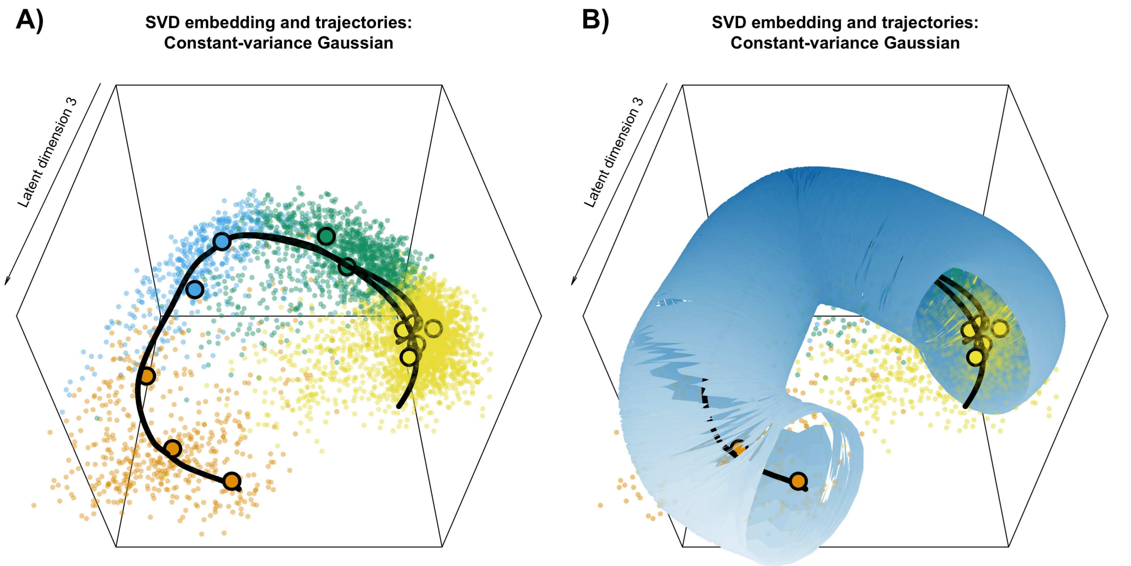


Figure 6: (A) Three-dimensional plot of the estimated latent positions via the SVD embedding with the two estimated cell developmental trajectories laid on top, corresponding to the data shown in the Figure 1. The thirteen bolded points correspond to the cluster centers of the thirteen cell sub-types, where the color scheme persists from Figure 1. (B) The uncertainty tube overlaid on top of Figure A.

tubes capture the variance of each estimated trajectory, and plotting these tubes is a useful descriptive tool. This is an important tool for our analysis because Slingshot is sensitive to small perturbations in the data due to its graph-based strategy to estimate the ordering of the cell sub-types. Specifically, small variations can dramatically change the number of estimated trajectories or ordering of cell sub-types within those trajectories. Hence, our procedure to construct these uniform uncertainty tubes first samples with replacement from all embedded cells within each of the thirteen cell sub-types. For each bootstrap sample, we apply Slingshot to estimate a new set of trajectories. We then compute the ℓ_2 distance between the new trajectories and the original trajectories. After applying this procedure multiple times, the 95% quantile of the ℓ_2 distances determines the uniform radius of the uncertainty tube, centered around the original trajectory. More details of this procedure are in Appendix G. Based on this construction, both trajectories lie in a single uncertainty tube (Figure 6B); hence, we conclude there is effectively one trajectory that connects all thirteen cell sub-types. This result explains why previous work such as Marques et al. (2016) had difficulty explaining how mature oligodendrocytes differentiate in their trajectory analysis.

7.3 Analysis using the curved Gaussian model

The above conclusions, however, rest on the questionable constant-variance Gaussian distributional assumption (see Figure 3). As we have seen in Figure 2, our matrix-completion diagnostic suggests that this assumption is not suitable for modeling the oligodendrocyte dataset at hand.

This finding motivates us to analyze the data using the eSVD to embed each cell with respect to the curved Gaussian distribution (4.7), and to re-examine the resulting diagnostics. Following suggestions from articles like Risso et al. (2018) and Durif et al. (2017), we no longer $\log_2$-transform the entries of A for our eSVD analysis, but rather model the counts in A directly after accounting for the library size. Based on our tuning procedure, the curved Gaussian distribution with $k = 5$ and $\tau = 2$ best fits the data, determined among a grid of candidate values. When we plot the resulting diagnostic for the eSVD in Figure 7, we obtain results that suggest a much better fit compared to that of the SVD. Specifically, the variance is appropriately increasing with the mean, unlike the trend shown in Figure 2. We conclude that the curved Gaussian model (without a $\log_2$-transformation) is more appropriate than the constant-variance Gaussian model (with a $\log_2$-transformation) for modeling our oligodendrocyte dataset.

We visualize the eSVD embedding in Figure 8 alongside its estimated trajectories and uncertainty tubes, and we see two distinct trajectories that differentiate among the mature

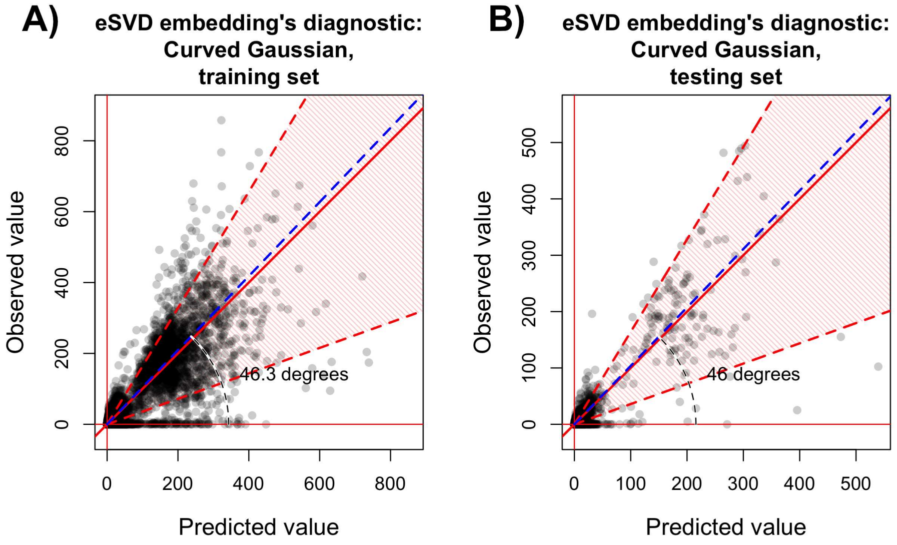


Figure 7: Diagnostic based on matrix completion to assess the fit using the eSVD embedding via the curved Gaussian model with $k = 5$ and $\tau = 2$ for either the observed values that are not omitted (i.e., the "training set") (A) or the observed values that are purposefully omitted (i.e., the "testing set") (B), both verses their respective predicted values. Both plots are comparable to those in Figure 2. Specifically, the 10th to 90th quantiles of the curved Gaussian model is marked by the shaded red region.

oligodendrocytes. Specifically, when we apply Slingshot to the eSVD embedding, we find that we still retain the conclusion that all cells from Pdgfra+ precursors to myelin-forming oligodendrocytes develop in the same way, similar to Marques et al. (2016). However, in contrast to that work, we are now able to observe substantial differentiation among the mature oligodendrocytes, with two distinct trajectories supported by the uncertainty tubes. Specifically, within this major cell type, only one of the six mature oligodendrocytes sub-types is shared between the two trajectories. Among the five remaining mature oligodendrocytes sub-types, three sub-types branch off in one trajectories while two sub-types branch into the other trajectory. This is in contrast with the analysis using the SVD embedding where all the estimated trajectories lay within one uncertainty tube (Figure 6B). We show ad-

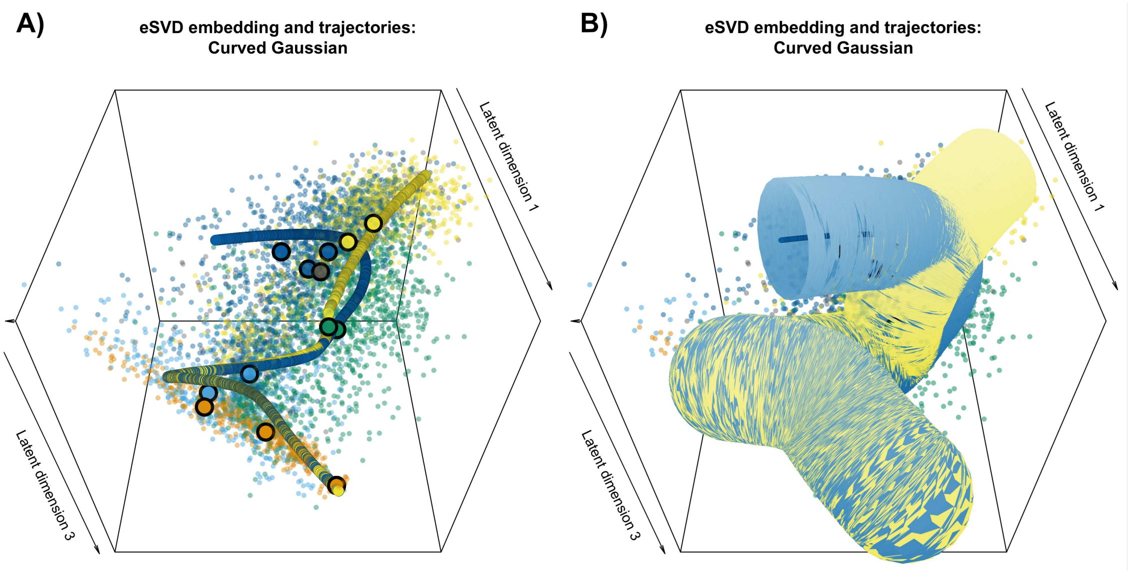


Figure 8: (A) Three-dimensional plot of the estimated latent positions via the eSVD embedding with the curved Gaussian distribution for $k = 5$ and $\tau = 2$ with the estimated cell developmental trajectory laid on top. The thirteen bolded points correspond to the cluster centers of the thirteen cell sub-types. The two estimated cell developmental trajectories are colored in yellow and blue. These correspond with the two of six mature oligodendrocytes cell sub-types unique to one trajectory (colored in yellow) and three mature oligodendrocytes cell sub-types unique to the other trajectory (colored in blue). The remaining mature oligodendrocyte cell sub-type is common to both trajectories, prior to the branching (colored in gray). The coloring of cells of other cell types persists from Figure 1. (B) The uncertainty tubes overlaid on top of Figure A. Both plots are comparable to Figure 6.

ditional plots corresponding to these results, as well as follow-up analyses and diagnostics of the oligodendrocytes using UMAP or ZINB-WaVE as well as plots based on the highly informative genes in Appendix H.

In summary, from the diagnostic (Figure 7), we conclude that the curved Gaussian distribution is more appropriate for the Marques data, and using this model we identify two distinct developmental trajectories (Figure 8). This is an improvement from the analysis in Marques et al. (2016) which suggested multiple trajectories, but was not able to directly verify this conjecture. Our comparison of the results obtained using the SVD versus the eSVD embeddings can help explain why previous scientific findings suggest that oligodendrocytes effectively follow a single developmental trajectory, while newer analyses based on more flexible statistical models (van Bruggen et al., 2017; Marques et al., 2018) suggest multiple trajectories.

We include an analysis of the single-cell dataset released in Zeisel et al. (2015) in Appendix H to demonstrate eSVD's performance in a different setting. There, the downstream task is to cluster cells rather than to infer developmental trajectories.

8 Discussion

In this article, we develop an estimator to non-linearly embed the cells in a single-cell RNA-sequencing dataset into a lower dimensional space with respect to a random dot product model where the inner product of two latent vectors is the natural parameter of a one-parameter exponential-family distribution F. This embedding method can greatly improve the estimation of cell developmental trajectories overall because it can handle distributions beyond the constant-variance Gaussian distribution, both in theory and practice. While the spirit of such embedding is not new, our contribution is two-fold. First, we develop the eSVD, an alternating minimizing estimator which is computationally efficient and also enables both a tuning procedure based on matrix completion and a theoretical investigation of its statistical properties such as identifiability and consistency. Second, we apply our estimator to analyze the oligodendrocytes in mouse brains, and our results coincide with recent scientific hypotheses (van Bruggen et al., 2017; Marques et al., 2018).

For future work, we plan to further the eSVD both in its modeling flexibility in practice as well as its theoretical properties, as we believe embeddings based on the random dot product model are appealing for single-cell analyses. Specifically, we plan to extend the eSVD to model the dropout effect, allow different nuisance parameters for each gene, or incorporate the library size directly into the statistical model directly. These trends occur in work such as Witten (2011), Pierson and Yau (2015), Townes et al. (2017) and Risso et al. (2018). While the models within these investigations are more flexible, we reiterate that their corresponding estimators were previously often believed to be too complicated to analyze from a theoretic perspective. Therefore, we are interested in studying how flexible our methods can be while still retaining tractability for theoretical analyses or how they perform in misspecified settings. Additionally, our current theory also does not address the trajectory estimation itself, which we think is another promising direction for theoretical investigation. On the methodological side, we plan to provide tuning procedures that are less computationally demanding compared to our current grid-search approach and to work on investigating other downstream applications of the eSVD embedding, such as cell clustering, batch correction, imputation, and RNA velocity (La Manno et al., 2018). Also, while this article focuses trajectory analysis based solely on the cells' latent vectors, other downstream applications such as gene clustering and finding marker genes could be developed based on the genes' latent vectors.

References

Athreya, A., Fishkind, D. E., Tang, M., Priebe, C. E., Park, Y., Vogelstein, J. T., Levin, K., Lyzinski, V., and Qin, Y. (2017). Statistical inference on random dot product graphs: A survey. *The Journal of Machine Learning Research*, 18(1):8393–8484.

Balakrishnan, S., Wainwright, M. J., Yu, B., et al. (2017). Statistical guarantees for the EM algorithm: From population to sample-based analysis. *The Annals of Statistics*, 45(1):77–120.

Becht, E., McInnes, L., Healy, J., Dutertre, C.-A., Kwok, I. W., Ng, L. G., Ginhoux, F., and

Newell, E. W. (2019). Dimensionality reduction for visualizing single-cell data using UMAP. *Nature biotechnology*, 37(1):38.

Bergen, V., Lange, M., Peidli, S., Wolf, F. A., and Theis, F. J. (2020). Generalizing RNA velocity to transient cell states through dynamical modeling. *Nature Biotechnology*, pages 1–7.

Butler, A., Hoffman, P., Smibert, P., Papalexi, E., and Satija, R. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature biotechnology*, 36(5):411–420.

Cai, Z. and Xiao, M. (2016). Oligodendrocytes and Alzheimer's disease. *International Journal of Neuroscience*, 126(2):97–104.

Cao, J., Spielmann, M., Qiu, X., Huang, X., Ibrahim, D. M., Hill, A. J., Zhang, F., Mundlos, S., Christiansen, L., Steemers, F. J., et al. (2019). The single-cell transcriptional landscape of mammalian organogenesis. *Nature*, 566(7745):496–502.

Chi, Y., Lu, Y. M., and Chen, Y. (2019). Nonconvex optimization meets low-rank matrix factorization: An overview. *IEEE Transactions on Signal Processing*, 67(20):5239–5269.

Collins, M., Dasgupta, S., and Schapire, R. E. (2002). A generalization of principal components analysis to the exponential family. In *Advances in neural information processing systems*, pages 617–624.

Durif, G., Modolo, L., Mold, J., Lambert-Lacroix, S., and Picard, F. (2017). Probabilistic count matrix factorization for single cell expression data analysis. In *Research in Computational Molecular Biology*, page 254. Springer.

Efron, B. et al. (1978). The geometry of exponential families. *The Annals of Statistics*, 6(2):362–376.

Fan, J., Wang, W., and Zhong, Y. (2018). An ℓ_∞ eigenvector perturbation bound and its application to robust covariance estimation. *Journal of Machine Learning Research*, 18(207):1–42.

Ge, R., Jin, C., and Zheng, Y. (2017). No spurious local minima in nonconvex low rank problems: A unified geometric analysis. pages 1233–1242.

Gunasekar, S., Ravikumar, P., and Ghosh, J. (2014). Exponential family matrix completion under structural constraints. In *International Conference on Machine Learning*, pages 1917–1925.

Haghverdi, L., Buettner, F., and Theis, F. J. (2015). Diffusion maps for high-dimensional single-cell analysis of differentiation data. *Bioinformatics*, 31(18):2989–2998.

Hastie, T. and Stuetzle, W. (1989). Principal curves. *Journal of the American Statistical Association*, 84(406):502–516.

Hicks, S. C., Townes, F. W., Teng, M., and Irizarry, R. A. (2017). Missing data and technical variability in single-cell RNA-sequencing experiments. *Biostatistics*.

Hoff, P. D., Raftery, A. E., and Handcock, M. S. (2002). Latent space approaches to social network analysis. *Journal of the american Statistical association*, 97(460):1090–1098.

Jaggi, M. (2013). Revisiting Frank-Wolfe: Projection-free sparse convex optimization. In *ICML (1)*, pages 427–435.

Jain, P., Netrapalli, P., and Sanghavi, S. (2013). Low-rank matrix completion using alternating minimization. In *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*, pages 665–674. ACM.

Kessaris, N., Fogarty, M., Iannarelli, P., Grist, M., Wegner, M., and Richardson, W. D. (2006). Competing waves of oligodendrocytes in the forebrain and postnatal elimination of an embryonic lineage. *Nature neuroscience*, 9(2):173.

Kharchenko, P. V., Silberstein, L., and Scadden, D. T. (2014). Bayesian approach to single-cell differential expression analysis. *Nature methods*, 11(7):740.

La Manno, G., Soldatov, R., Zeisel, A., Braun, E., Hochgerner, H., Petukhov, V., Lidschreiber, K., Kastriti, M. E., Lönnerberg, P., Furlan, A., et al. (2018). RNA velocity of single cells. *Nature*, 560(7719):494–498.

Lafond, J. (2015). Low rank matrix completion with exponential family noise. In *Conference on Learning Theory*, pages 1224–1243.

Landgraf, A. J. and Lee, Y. (2019). Generalized principal component analysis: Projection of saturated model parameters. *Technometrics*, pages 1–14.

Lei, J. (2018). Network representation using graph root distributions. *arXiv preprint arXiv:1802.09684*.

Li, T., Levina, E., and Zhu, J. (2020). Network cross-validation by edge sampling. *Biometrika*, 107(2):257–276.

Liu, C. and Martin, R. (2020). Inferential models and possibility measures. *arXiv preprint arXiv:2008.06874*.

Lopez, R., Regier, J., Cole, M. B., Jordan, M. I., and Yosef, N. (2018). Deep generative modeling for single-cell transcriptomics. *Nature methods*, 15(12):1053.

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology*, 15(12):550.

Maaten, L. v. d. and Hinton, G. (2008). Visualizing data using t-SNE. *Journal of machine learning research*, 9(Nov):2579–2605.

Maezika, M. (2016). The singular value decomposition and low rank approximation.

Marques, S., van Bruggen, D., Vanichkina, D. P., Floriddia, E. M., Munguba, H., Väremo, L., Giacomello, S., Falcão, A. M., Meijer, M., Björklund, Å. K., et al. (2018). Transcriptional convergence of oligodendrocyte lineage progenitors during development. *Developmental cell*, 46(4):504–517.

Marques, S., Zeisel, A., Codeluppi, S., van Bruggen, D., Falcão, A. M., Xiao, L., Li, H., Häring, M., Hochgerner, H., Romanov, R. A., et al. (2016). Oligodendrocyte heterogeneity in the mouse juvenile and adult central nervous system. *Science*, 352(6291):1326–1329.

Mazumder, R., Hastie, T., and Tibshirani, R. (2010). Spectral regularization algorithms for learning large incomplete matrices. *Journal of machine learning research*, 11(Aug):2287–2322.

McInnes, L., Healy, J., and Melville, J. (2018). UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Menn, B., Garcia-Verdugo, J. M., Yaschine, C., Gonzalez-Perez, O., Rowitch, D., and Alvarez-Buylla, A. (2006). Origin of oligodendrocytes in the subventricular zone of the adult brain. *Journal of Neuroscience*, 26(30):7907–7918.

Pierson, E. and Yau, C. (2015). ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis. *Genome biology*, 16(1):241.

Risso, D., Perraudeau, F., Gribkova, S., Dudoit, S., and Vert, J.-P. (2018). A general and flexible method for signal extraction from single-cell RNA-seq data. *Nature communications*, 9(1):284.

Saelens, W., Cannoodt, R., Todorov, H., and Saeys, Y. (2019). A comparison of single-cell trajectory inference methods. *Nature biotechnology*, 37(5):547.

Street, K., Risso, D., Fletcher, R. B., Das, D., Ngai, J., Yosef, N., Purdom, E., and Dudoit, S. (2018). Slingshot: Cell lineage and pseudotime inference for single-cell transcriptomics. *BMC genomics*, 19(1):477.

Sun, S., Zhu, J., Ma, Y., and Zhou, X. (2019). Accuracy, robustness and scalability of dimensionality reduction methods for single-cell RNA-seq analysis. *Genome biology*, 20(1):269.

Tenenbaum, J. B., De Silva, V., and Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323.

Townes, F. W., Hicks, S. C., Aryee, M. J., and Irizarry, R. A. (2017). Varying-censoring aware matrix factorization for single cell RNA-sequencing. *bioRxiv*, page 166736.

Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., Lennon, N. J., Livak, K. J., Mikkelsen, T. S., and Rinn, J. L. (2014). The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature biotechnology*, 32(4):381.

Udell, M., Horn, C., Zadeh, R., Boyd, S., et al. (2016). Generalized low rank models. *Foundations and Trends® in Machine Learning*, 9(1):1–118.

van Bruggen, D., Agirre, E., and Castelo-Branco, G. (2017). Single-cell transcriptomic analysis of oligodendrocyte lineage cells. *Current opinion in neurobiology*, 47:168–175.

Vershynin, R. (2012). How close is the sample covariance matrix to the actual covariance matrix? *Journal of Theoretical Probability*, 25(3):655–686.

Wang, L., Zhang, X., and Gu, Q. (2016). A unified computational and statistical framework for nonconvex low-rank matrix estimation. *arXiv preprint arXiv:1610.05275*.

Witten, D. M. (2011). Classification and clustering of sequencing data using a Poisson model. *The Annals of Applied Statistics*, 5(4):2493–2518.

Yu, M., Gupta, V., Kolar, M., et al. (2020). Recovery of simultaneous low rank and two-way sparse coefficient matrices, a nonconvex approach. *Electronic Journal of Statistics*, 14(1):413–457.

Zappia, L., Phipson, B., and Oshlack, A. (2017). Splatter: simulation of single-cell RNA sequencing data. *Genome biology*, 18(1):174.

Zeisel, A., Muñoz-Manchado, A. B., Codeluppi, S., Lönnerberg, P., La Manno, G., Juréus, A., Marques, S., Munguba, H., He, L., Betsholtz, C., et al. (2015). Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science*, 347(6226):1138–1142.

Zhang, Y., Chen, K., Sloan, S. A., Bennett, M. L., Scholze, A. R., O'Keeffe, S., Phatnani, H. P., Guarnieri, P., Caneda, C., Ruderisch, N., et al. (2014). An RNA-sequencing transcriptome and splicing database of glia, neurons, and vascular cells of the cerebral cortex. *Journal of Neuroscience*, 34(36):11929–11947.

Zhao, T., Wang, Z., and Liu, H. (2015). Nonconvex low rank matrix factorization via inexact first order oracle. *Advances in Neural Information Processing Systems.*