Dropout-Aware Weighted NMF on Single-Cell RNA Sequencing Data

Jiaqi Zhang [1]

Abstract

Non-negative matrix factorization (NMF) aims to infer low-dimensional structure from high-dimensional genomic data to understand and visualize complex biological processes. In recent years, the rapid development of single-cell RNA sequencing (scRNA-seq) enables a single-cell level of gene profiling. Previous studies have applied NMF on scRNA-seq data to obtain a finer resolution of biological system understanding, but the high sparsity and the complicated structure of dropouts (i.e., zero values) in scRNA-seq data pose challenges to NMF optimization. To address this problem, in this study, we propose a dropout-aware weighted non-negative matrix factorization (`DA-WNMF`) to deal with the high sparsity based on Bayesian factorization. On experimental scRNA-seq data, `DA-WNMF` outperforms previous works when the data is not too sparse (e.g., has around 80% zeros) but fails for very-sparse data (i.e., has more than 90% zeros).

1. Introduction

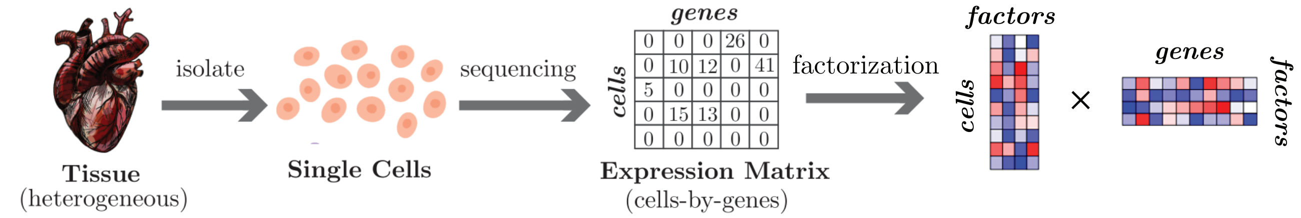
Non-negative matrix factorization (NMF) (Sra & Dhillon, 2005; Wang & Zhang, 2012) is a type of matrix factorization method that decomposes a matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ into two non-negative factor matrices $\mathbf{W} \in \mathbb{R}^{n \times k}_{\geqslant 0}$ and $\mathbf{H} \in \mathbb{R}^{k \times q}_{\geqslant 0}$ with $k << q$ such that $\mathbf{X} = \mathbf{WH}$. Previous works have used NMF to reveal low-dimensional structures from the high-dimensional genomic data and understand biological interactions. NMF has shown superior performance on cell clustering (Liu et al., 2008), gene interaction network inference (Ding et al., 2021), and differentially expressed genes detection (Jia et al., 2015; Akçay et al., 2022) on microarray and bulk RNA sequencing (RNA-seq) data.

In recent years, the rapid development of single-cell RNA sequencing (scRNA-seq) technique has enabled researchers to profile transcriptomes in individual cells and analyze biological processes in a finer resolution (Hwang et al., 2018; Chen et al., 2019). The scRNA-seq technique contains several steps. It first isolates tissues or a group of heterogeneous cells into individual cells. Then it will profile gene expression at the single-cell level and formulate expression into a cell-by-gene matrix (Fig. 1). The scRNA-seq datasets are large, generally consisting of expression for hundreds of thousands of cells and genes (Macosko et al., 2015; Picelli et al., 2014), providing richer information on gene expression and helping with data-driven population-based factorization. However, single-cell data opportunities also present new challenges for statistical analysis (Lähnemann et al., 2020). First, the scRNA-seq data can have very high sparsity, having more than 80% of zeros (Hicks et al., 2018; Qiu, 2020). It is notoriously difficult to deal with high sparsity. Second, the dropouts (i.e., zero values) have complex generation processes. The dropout can be a biological signal indicating the gene is indeed not expressed in some cells, or it is a technical artifact such that the gene has been expressed but the sequencing technique has lost its value due to technical reasons. This means we cannot treat all zeros as missing values like we generally do in other applications to deal with high sparsity.

Previous studies have proposed several NMF methods and applied them to scRNA-seq data. But they do not treat dropouts properly. These NMF variations can be largely divided into three groups. First, (Hoyer, 2002; Taslaman & Nilsson, 2012; Shen et al., 2014) proposed sparsity-regularized NMF models. They attached an ℓ_1 norm regularizer to obtain sparse factorizations. Second, (Cai et al., 2010; 2008; Xing et al., 2021; Xiao et al., 2018; Wang et al., 2019) utilized the geometric structure of data and proposed graph-regularized NMF models. The last group (Schmidt et al., 2009; Cemgil, 2009; Sherman et al., 2020; Durif et al., 2019; Sun et al., 2019) assumed observations or factor matrices follow certain distributions and learned the factorization by maximizing the likelihood. Although these methods successfully apply structural or statistical constraints in the optimization, when applied to scRNA-seq data, these models will break down since they treat all ob-

[1] Department of Computer Science, Brown University. Correspondence to: Jiaqi Zhang <jiaqi_zhang2@brown.edu>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022. Copyright 2022 by the author(s).



0	0	0	26	0
0	10	12	0	41
5	0	0	0	0
0	15	13	0	0
0	0	0	0	0

Figure 1. Illustration of NMF on scRNA-seq data. Given tissues or a group of heterogeneous cells, scRNA-seq first isolate them into individual cells. Then it will profile gene expression at the single-cell level and formulate expression into a cell-by-gene matrix, in which each element is a count value. The NMF will factorize expression matrix into a cell factor and gene factor matrices.

servations as informative values and ignore the existence of dropouts. Therefore, efficient and robust NMF models for scRNA-seq data are urgently needed in the field.

In this study, we propose a novel method called dropout-aware weighted non-negative matrix factorization (`DA-WNMF`) to deal with complex dropouts and high sparsity in scRNA-seq data. We introduce a weight matrix into the loss function to let factorization focus more on observations instead of missing values. Since dropouts contain both missing values and biological zeros, we estimate the probability that each expression value is missing. Specifically, we assume the normalized gene expression follows a Gamma-Normal mixture distribution, and we estimate distribution parameters by maximizing the likelihood. Then we compute the dropout rate for each expression value and construct the weight matrix, with a lower value denoting it is more likely to be missing value. After testing `DA-WNMF` on three experimental scRNA-seq datasets, we find that our `DA-WNMF` outperforms baseline methods concerning cell clustering when data is not too sparse, i.e., has around 20% non-zeros. But `DA-WNMF` also fails for very sparse data with only around 5% non-zeros. So how to extend NMF to scRNA-seq data is still an open problem.

2. Background

Notations: $\mathbf{X} \in \mathbf{Z}_{\geqslant 0}^{n \times p}$ denotes the scRNA-seq count matrix of n cells and p genes. Its corresponding factorization is the cell factor $\mathbf{W} \in \mathbf{R}_{\geqslant 0}^{n \times k}$ and gene factor $\mathbf{H} \in \mathbf{R}_{\geqslant 0}^{k \times p}$ with $k << p$ is the factor size. $\parallel \mathbf{X} \parallel_{\mathrm{F}} = \sqrt{\sum_{i,j} X_{ij}^2}$ and $\parallel \mathbf{X} \parallel_1 = \sqrt{\sum_{i,j} |X_{ij}|}$ are Frobenius norm and ℓ_1 norm respectively. $[n] = \{1, 2, 3, \cdots, n\}$.

2.1. Non-Negative Matrix Factorization (`NMF`)

`NMF` is a group of linear matrix factorization in which a matrix $\mathbf{X}$ is decomposed into linear multiplication of two non-negative matrices $\mathbf{W}$ and $\mathbf{H}$ such that

$$\mathbf{X} = \mathbf{WH}, \text{ with } W_{ij}, H_{ij} \geqslant 0 \text{ for } i \in [n] \text{ and } j \in [p]. \tag{1}$$

The conventional way to optimize factorization is minimizing the difference between $\mathbf{X}$ and reconstructed data $\mathbf{WH}$ via

$$\begin{aligned} \widehat{\mathbf{W}}, \widehat{\mathbf{H}} = \underset{\mathbf{W}, \mathbf{H}}{\operatorname{argmin}} \quad & \parallel \mathbf{X} - \mathbf{WH} \parallel_{\mathrm{F}}^2 \\ \text{s.t.} \quad & W_{ij}, H_{ij} \geqslant 0. \end{aligned} \tag{2}$$

2.2. Bayesian Inference of scRNA-seq Distribution

The conventional NMF formulation (Eq. 2) implicitly assumes the residual $\mathbf{X} - \mathbf{WH}$ should follow a Gaussian distribution. However, this is not the case for scRNA-seq data. To this end, previous works have proposed Bayesian NMF models, applying explicit data distribution assumptions. Specifically, this group of methods assumes observations

$$\mathbf{X} \sim \mathcal{D}_x(\mathbf{\Theta}) \tag{3}$$

where

$$\mathbf{\Theta} = f(\mathbf{WH}). \tag{4}$$

Here, f is a function that maps the multiplication $\mathbf{WH}$ to parameters of the observation distribution $\mathcal{D}_x$. Different works have different settings of $\mathcal{D}_x$. For example, (Durif et al., 2019) assumes $\mathcal{D}_x$ is a multivariate Poisson distribution.

2.3. Weighted Matrix Completion

In another line of work, matrix completion also faces the problem of missing values. The matrix completion is filling in missing entries for a partially observed matrix. The problem can be formulated into estimating a low-rank [1] matrix through

$$\widehat{\mathbf{M}} = \underset{\text{rank}(\mathbf{M}) = r}{\operatorname{argmin}} \quad \parallel \mathbf{X} - \mathbf{M} \parallel_{\mathrm{F}}^2 \tag{5}$$

given $\mathbf{X}$ with missing values. Eq. 5 assumes missing values are sampled in a uniformly random manner. However, in

[1] The low-rank assumption makes the problem well-posed.

many applications, the sampling strategy is not uniformly random, such that the conventional matrix completion methods are improper.

To solve this problem, (Negahban & Wainwright, 2012; Foucart et al., 2020; Cheng & Ge, 2018; Li et al., 2016) have proposed and analyzed the so-called weighted matrix completion. It revises the matrix completion problem by introducing a weight matrix $\mathbf{W} \in \mathbb{R}_{\geqslant 0}^{n \times p}$ as

$$\widehat{\mathbf{M}} = \underset{\text{rank}(\mathbf{M})=r}{\text{argmin}} \quad \| \mathbf{W} \circ (\mathbf{X} - \mathbf{M}) \|_{\text{F}}^2 \,. \tag{6}$$

Here, $\circ$ is the Hadamard (element-wise) product such that $[\mathbf{A} \circ \mathbf{B}]_{ij} = A_{ij} B_{ij}$. Real applications often use a binary weight matrix as

$$W_{ij}^{\text{Bi}} = \begin{cases} 1, & X_{ij} \neq 0, \\ 0, & X_{ij} = 0 \end{cases} \quad . \tag{7}$$

With the binary weight matrix, the weighted model focuses more on observations and avoids the misleading effect of missing values. In Sec. 3.2, we extend the weighted matrix completion to weighted NMF by defining a weight matrix considering special properties of scRNA-seq data.

3. Method

3.1. Previous NMF Models Break Down when Data Sparsity Grows

To validate that previous work will fail when data become too sparse, we tested five methods of multiple categories on simulated datasets.

Concretely, we simulate scRNA-seq data from a Gamma-Poisson distribution. Because it is equivalent to negative binomial distribution and better fit the real-world scRNA-seq data (Risso et al., 2018; Townes et al., 2019; Svensson, 2020). We first simulate block-diagonal cell and gene factor matrices with each block generated from $\mathbf{W}, \mathbf{H} \sim \text{Gamma}(1, 10)$. We simulate 500 cells, 100 genes, and 20 factors. The cell factor $\mathbf{W}$ has five diagonal blocks, and the gene factor $\mathbf{H}$ has four blocks. Then we can generate the count expression matrix with $\mathbf{X} \sim \text{Poisson}(\mathbf{\Theta})$ where $\mathbf{\Theta} = \mathbf{WH}$. Finally, we add dropouts by randomly setting expression values to zeros according to $\text{Binomial}(s)$ with s controls the sparsity level. Fig. 2 provides an illustration of the simulated data generation.

Here, we generate simulated datasets with the ratio of zeros of $\{0.0, 0.1, 0.2, \cdots, 0.9\}$. We test five methods: (1) conventional NMF, (2) L1-regularized NMF, (3) graph-regularized NMF, (4) `pCMF`, a Bayesian NMF based on the negative binomial distribution, and (5) the weighted NMF with binary weights To evaluate models, we cluster cells and genes from cell and gene factor matrices with KMeans algorithm. The number of clusters is set to the ground truth, which is 5 for cell clustering and 4 for gene clustering. Given clustering results, we compute the adjusted rand index (ARI) to evaluate clustering performance, as to evaluate the performance of factorizing observations into low-dimensional factors. The ARI value ranges from 0 to 1, with 1 denoting perfect clustering and 0 denoting random labeling.

Fig. 3 clearly shows that as the data becomes more and more sparse, `NMF`, `L1 NMF`, `Graph NMF`, and `pCMF` have worse performance. They break down when the dataset has more than 50% of zeros and have ARI lower than 0.3. On the contrary, the weighted model with binary weights, although its performance also decreases when the data sparsity increases, still outperform other methods. So the weighted NMF seems to be a promising way to deal with sparse scRNA-seq data. But the binary weight ignores the properties of dropout by assuming all observed zeros as missing values. So in the following, we proposed an approach to estimate the weight matrix considering the scRNA-seq data properties.

3.2. Proposed: Dropout-Aware Weighted Non-Negative Matrix Factorization (`DA-WNMF`)

As we mentioned, the scRNA-seq data dropouts contain two generation processes: biological zeros and technical missings. So we want to estimate a weight matrix giving zeros only to those technical missing values, since the biological zeros are informative. To do so, we adopt a similar idea as (Li & Li, 2018) and use the Gamma-Normal mixture distribution. Concretely, we use the normalized expression matrix. Given $\mathbf{X}$, we normalize it through two steps:

1. Normalize each cell into the same library size (i.e., the total number of counts for a cell) via

$$\mathbf{X}_{i:} = \frac{\mathbf{X}_{i:}}{\sum_k \mathbf{X}_{ik}} \cdot M \tag{8}$$

   where M is a scale factor and we set it as $M = 10^4$ in our experiments.

2. After this, we log-transform data with

$$\mathbf{X}_{ij} = \log_{10}^{(\mathbf{X}_{ij}+1)} \tag{9}$$

   to reduce the data variance.

This is a commonly-used normalization approach used in many scRNA-seq studies (Wolf et al., 2018). We assume the normalized expression for a gene j follows a Gamma-Normal mixture distribution with the density function as

$$f_j(x) = \lambda_j \text{Gamma}(x;\, \alpha_j, \beta_j) + (1 - \lambda_j) \text{Normal}(x;\, \mu_j, \sigma_j). \tag{10}$$

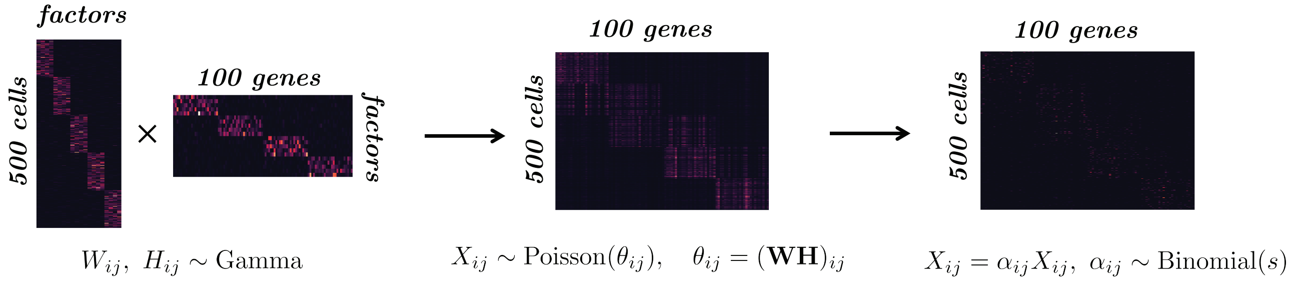

factors
500 cells
100 genes
factors
100 genes
500 cells
100 genes
500 cells
$W_{ij},\ H_{ij} \sim \text{Gamma}$
$X_{ij} \sim \text{Poisson}(\theta_{ij}),\quad \theta_{ij} = (\mathbf{WH})_{ij}$
$X_{ij} = \alpha_{ij} X_{ij},\ \alpha_{ij} \sim \text{Binomial}(s)$

Figure 2. Illustration of data simulation process. Generate cell and gene factor matrices from a Gamma distribution. Then the expression matrix is generated from a Poisson distribution where parameters depend on factor matrices. Finally, we add dropouts with the Binomial distribution.

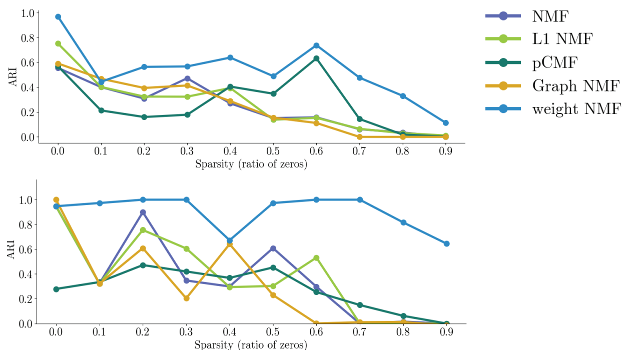


Figure 3. Compare five methods on simulated expression matrix.

The Gamma component represents technical missings, and the normal part denotes actual biological expressions. The parameter λ_j controls the trade-off between these two components.

We estimate five distribution parameters λ_j, α_j, β_j, μ_j, and σ_j in an expectation-maximization (EM) algorithm that maximizes the log-likelihood

$$\widehat{\lambda}_j, \widehat{\alpha}_j, \widehat{\beta}_j, \widehat{\mu}_j, \widehat{\sigma}_j = \operatorname{argmax} \sum_{j=1}^{p} f_j(x;\ \lambda_j, \alpha_j, \beta_j, \mu_j, \sigma_j). \tag{11}$$

Once parameters are inferred from data, we can compute a dropout rate d_{ij} for each expression value through

$$d_{ij} = 1 - \frac{\widehat{\lambda}_j \text{Gamma}(X_{ij};\ \widehat{\alpha}_j, \widehat{\beta}_j)}{\widehat{\lambda}_j \text{Gamma}(X_{ij};\ \widehat{\alpha}_j, \widehat{\beta}_j) + (1-\widehat{\lambda}_j)\text{Normal}(X_{ij};\ \widehat{\mu}_j, \widehat{\sigma}_j)}. \tag{12}$$

Intuitively, given a matrix element X_{ij}, this dropout rate d_{ij} evaluates the probability of the expression being modeled by the Gaussian component. So a high value indicates it is a true biological expression with high confidence, and a low dropout rate means it is more likely to be a technical miss. Finally, we can construct the **dropout-aware weight** matrix by thresholding small values as

$$W_{ij}^{\text{DA}} = \begin{cases} d_{ij}, & d_{ij} \geqslant 0.5, \\ 0, & d_{ij} < 0.5 \end{cases}. \tag{13}$$

Embedding this drop-out aware weight in weighted NMF model, we propose the **dropout-aware weighted non-negative matrix factorization** (`DA-WNMF`) model:

$$\begin{aligned} \widehat{\mathbf{W}}, \widehat{\mathbf{H}} = \operatorname*{argmin}_{\mathbf{W},\mathbf{H}} \quad & \| W^{\text{DA}} \circ (\mathbf{X} - \mathbf{WH}) \|_{\text{F}}^2 \\ \text{s.t.} \quad & W_{ij}, H_{ij} \geqslant 0. \end{aligned} \tag{14}$$

4. Experiment on Real scRNA-seq Data

We compared our `DA-WNMF` with five baselines on three experimental scRNA-seq datasets. These datasets cover multiple sequencing protocols and species, having a different number of cells, sparsity levels, and cell types. Detailed statistics of these datasets are listed in Table 1. We selected the top 500 highly variable genes for each dataset with the `scanpy` package (Wolf et al., 2018). We utilize the automatic differentiation in `PyTorch` apckage to solve Eq. 14.

Table 2 shows the comparison of cell clustering ARI values of `DA-WNMF` and five baseline models. First, on the Mouse Cortex dataset, the weighted NMF models, both binary weights and dropout-aware weight, outperform other baselines. Specifically, our `DA-WNMF` based on dropout-aware weight is better than the binary weight, having the best performance.

However, our `DA-WNMF` still fails on very-sparse data. The Mose Cortex dataset is the least sparse dataset in our experiments, having around 20% non-zeros. The sparsity levels of the other two datasets, Human PBMC and Quake Lung, are around 4% and 7%. Results on these two more sparse datasets indicate:

- Weighted NMF is better than others while binary weight and dropout-aware weight are comparable.

Table 1. Basic statistics of three experimental datasets.

Name	Species	Protocol	# Cells	Sparsity (ratio of non-zeros)	# Cell Types
Mouse Cortex	Mouse	Smart Seq2	643	21.55%	7
Human PBMC	Human	Drop Seq	6438	3.50%	9
Quake Lung	Mouse	Smart Seq2	1676	7.32%	11

- All methods are less than ideal since the clustering ARI value is less than 0.2. So for some very sparse data, our `DA-WNMF` is still not practical for biologists.

5. Discussion and Conclusion

In this study, we propose the dropout-aware weight to extend NMF to the single-cell analysis by considering the special properties of single-cell data. Although the model is less than ideal for very sparse data, it is still applicable in some cases. Some sequencing techniques generate less sparse datasets. Also, in pre-processing, some quality control steps can remove lowly-expressed cells and genes, such that they can reduce sparsity.

In future works, the dropout rate estimation can be improved. We now use the Gaussian component from the Gamma-Normal mixture to explain the actual expression (Eq. 10). However, in some single-cell data, especially those with very high sparsity, the actual expression part is not Gaussian, but more likely to be negative binomial (Risso et al., 2018; Townes et al., 2019; Svensson, 2020). So recent studies use zero-inflated negative-binomial distribution. It also assumes two zero generation processes, which might improve the weighted model performance.

Moreover, in scRNA-seq analyses, several pre-processing steps are proposed to improve estimations. For example, the normalization can reduce data variations (Stuart et al., 2019; Wolf et al., 2018; Choudhary & Satija, 2022; Borella et al., 2022). Moreover, scRNA-seq imputation methods are used to impute zero counts with values inferred from data (Andrews & Hemberg, 2018; Hou et al., 2020) to reduce data sparsity. So testing whether such pre-processing can improve the NMF model is another direction of improvement.

References

Akçay, S., Güven, E., Afzal, M., and Kazmi, I. Non-negative matrix factorization and differential expression analyses identify hub genes linked to progression and prognosis of glioblastoma multiforme. *Gene*, 824:146395, 2022.

Andrews, T. S. and Hemberg, M. False signals induced by single-cell imputation. *F1000Research*, 7, 2018.

Borella, M., Martello, G., Risso, D., and Romualdi, C. Psinorm: a scalable normalization for single-cell rna-seq data. *Bioinformatics*, 38(1):164–172, 2022.

Cai, D., He, X., Wu, X., and Han, J. Non-negative matrix factorization on manifold. In *2008 eighth IEEE international conference on data mining*, pp. 63–72. IEEE, 2008.

Cai, D., He, X., Han, J., and Huang, T. S. Graph regularized nonnegative matrix factorization for data representation. *IEEE transactions on pattern analysis and machine intelligence*, 33(8):1548–1560, 2010.

Cemgil, A. T. Bayesian inference for nonnegative matrix factorisation models. *Computational intelligence and neuroscience*, 2009, 2009.

Chen, G., Ning, B., and Shi, T. Single-cell rna-seq technologies and related computational data analysis. *Frontiers in genetics*, pp. 317, 2019.

Cheng, Y. and Ge, R. Non-convex matrix completion against a semi-random adversary. In *Conference On Learning Theory*, pp. 1362–1394. PMLR, 2018.

Choudhary, S. and Satija, R. Comparison and evaluation of statistical error models for scrna-seq. *Genome biology*, 23(1):1–20, 2022.

Ding, Q., Sun, Y., Shang, J., Li, F., Zhang, Y., and Liu, J.-X. Nmfna: A non-negative matrix factorization network analysis method for identifying modules and characteristic genes of pancreatic cancer. *Frontiers in genetics*, pp. 1115, 2021.

Durif, G., Modolo, L., Mold, J. E., Lambert-Lacroix, S., and Picard, F. Probabilistic count matrix factorization for single cell expression data analysis. *Bioinformatics*, 35 (20):4011–4019, 2019.

Foucart, S., Needell, D., Pathak, R., Plan, Y., and Wootters, M. Weighted matrix completion from non-random, non-uniform sampling patterns. *IEEE Transactions on Information Theory*, 67(2):1264–1290, 2020.

Table 2. Cell clustering results of six methods on three experimental datasets. Red bold number indicates the best result; blue underlined number indicates the second best.

Dataset	Cell Clustering ARI ↑ (**best**, second best)					
	NMF	L1 NMF	Graph NMF	pCMF	weighted NMF (binary)	DA-WNMF
Mouse Cortex	0.4	0.38	0.27	0.16	0.43	**0.51**
Human PBMC	0.06	0.01	0.06	0.08	**0.11**	0.10
Quake Lung	0.10	0.07	0.04	0.02	0.18	**0.19**

Hicks, S. C., Townes, F. W., Teng, M., and Irizarry, R. A. Missing data and technical variability in single-cell rna-sequencing experiments. *Biostatistics*, 19(4):562–578, 2018.

Hou, W., Ji, Z., Ji, H., and Hicks, S. C. A systematic evaluation of single-cell rna-sequencing imputation methods. *Genome biology*, 21(1):1–30, 2020.

Hoyer, P. O. Non-negative sparse coding. In *Proceedings of the 12th IEEE workshop on neural networks for signal processing*, pp. 557–565. IEEE, 2002.

Hwang, B., Lee, J. H., and Bang, D. Single-cell rna sequencing technologies and bioinformatics pipelines. *Experimental & molecular medicine*, 50(8):1–14, 2018.

Jia, Z., Zhang, X., Guan, N., Bo, X., Barnes, M. R., and Luo, Z. Gene ranking of rna-seq data via discriminant non-negative matrix factorization. *PloS one*, 10(9):e0137782, 2015.

Lähnemann, D., Köster, J., Szczurek, E., McCarthy, D. J., Hicks, S. C., Robinson, M. D., Vallejos, C. A., Campbell, K. R., Beerenwinkel, N., Mahfouz, A., et al. Eleven grand challenges in single-cell data science. *Genome biology*, 21(1):1–35, 2020.

Li, W. V. and Li, J. J. An accurate and robust imputation method scimpute for single-cell rna-seq data. *Nature communications*, 9(1):1–9, 2018.

Li, Y., Liang, Y., and Risteski, A. Recovery guarantee of weighted low-rank approximation via alternating minimization. In *International Conference on Machine Learning*, pp. 2358–2367. PMLR, 2016.

Liu, W., Yuan, K., and Ye, D. Reducing microarray data via nonnegative matrix factorization for visualization and clustering analysis. *Journal of biomedical informatics*, 41(4):602–606, 2008.

Macosko, E. Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A. R., Kamitaki, N., Martersteck, E. M., et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, 161(5):1202–1214, 2015.

Negahban, S. and Wainwright, M. J. Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. *The Journal of Machine Learning Research*, 13(1):1665–1697, 2012.

Picelli, S., Faridani, O. R., Björklund, Å. K., Winberg, G., Sagasser, S., and Sandberg, R. Full-length rna-seq from single cells using smart-seq2. *Nature protocols*, 9(1): 171–181, 2014.

Qiu, P. Embracing the dropouts in single-cell rna-seq analysis. *Nature communications*, 11(1):1–9, 2020.

Risso, D., Perraudeau, F., Gribkova, S., Dudoit, S., and Vert, J.-P. A general and flexible method for signal extraction from single-cell rna-seq data. *Nature communications*, 9 (1):1–17, 2018.

Schmidt, M. N., Winther, O., and Hansen, L. K. Bayesian non-negative matrix factorization. In *International Conference on Independent Component Analysis and Signal Separation*, pp. 540–547. Springer, 2009.

Shen, B., Liu, B.-D., Wang, Q., and Ji, R. Robust nonnegative matrix factorization via l 1 norm regularization by multiplicative updating rules. In *2014 IEEE International Conference on Image Processing (ICIP)*, pp. 5282–5286. IEEE, 2014.

Sherman, T. D., Gao, T., and Fertig, E. J. Cogaps 3: Bayesian non-negative matrix factorization for single-cell analysis with asynchronous updates and sparse data structures. *BMC bioinformatics*, 21(1):1–6, 2020.

Sra, S. and Dhillon, I. Generalized nonnegative matrix approximations with bregman divergences. *Advances in neural information processing systems*, 18, 2005.

Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck III, W. M., Hao, Y., Stoeckius, M., Smibert, P., and Satija, R. Comprehensive integration of single-cell data. *Cell*, 177(7):1888–1902, 2019.

Sun, S., Chen, Y., Liu, Y., and Shang, X. A fast and efficient count-based matrix factorization method for detecting cell types from single-cell rnaseq data. *BMC systems biology*, 13(2):1–8, 2019.

Svensson, V. Droplet scrna-seq is not zero-inflated. *Nature Biotechnology*, 38(2):147–150, 2020.

Taslaman, L. and Nilsson, B. A framework for regularized non-negative matrix factorization, with application to the analysis of gene expression data. *PloS one*, 7(11):e46331, 2012.

Townes, F. W., Hicks, S. C., Aryee, M. J., and Irizarry, R. A. Feature selection and dimension reduction for single-cell rna-seq based on a multinomial model. *Genome biology*, 20(1):1–16, 2019.

Wang, C.-Y., Liu, J.-X., Yu, N., and Zheng, C.-H. Sparse graph regularization non-negative matrix factorization based on huber loss model for cancer data analysis. *Frontiers in genetics*, 10:1054, 2019.

Wang, Y.-X. and Zhang, Y.-J. Nonnegative matrix factorization: A comprehensive review. *IEEE Transactions on knowledge and data engineering*, 25(6):1336–1353, 2012.

Wolf, F. A., Angerer, P., and Theis, F. J. Scanpy: large-scale single-cell gene expression data analysis. *Genome biology*, 19(1):1–5, 2018.

Xiao, Q., Luo, J., Liang, C., Cai, J., and Ding, P. A graph regularized non-negative matrix factorization method for identifying microrna-disease associations. *Bioinformatics*, 34(2):239–248, 2018.

Xing, Z., Ma, Y., Yang, X., and Nie, F. Graph regularized nonnegative matrix factorization with label discrimination for data clustering. *Neurocomputing*, 440:297–309, 2021.