An Advanced NMF-Based Approach for Single Cell Data Clustering

Peng Zhao
School of Computer Science and Engineering
University of Electronic Science and Technology of China
Chengdu, China
zhaop211@gmail.com

Yongpan Sheng
School of Big Data and Software Engineering
Chongqing University
Chongqing, China
shengyp2011@gmail.com

Xiaohui Zhan*
School of Basic Medicine
Chongqing Medical University
Chongqing, China
xhzhan@cqmu.edu.cn

***Abstract*—Single-cell RNA sequencing (scRNA-seq) provides transcriptomic profiling for individual cells, allowing researchers to study the heterogeneity of tissues, recognize rare cell identities and discover new cellular subtypes. Clustering analysis is usually used to predict cell class assignments and infer cell identities. However, The performance of existing single-cell clustering methods is extremely sensitive to the presence of noise data and outliers. Nevertheless, there is still no consensus on the best performing method. To address this issue, we utilize an advanced NMF for scRNA-seq data clustering based on soft self-paced learning (S3NMF). We will gradually add cells from simple to complex to our model until the model converges. In this way, the influence of noisy data and outliers can be significantly reduced. The proposed method achieves the best performance on both simulation data and real scRNA-seq data.**

***Keywords*—single cell, clustering, sequence data, NMF**

I. Introduction

In recent years, advances in single cell RNA sequencing (scRNA-seq) have promoted the study of computational methods for analyzing transcriptome data from single cells. scRNA-seq can be regarded a powerful new approach for studying the transcriptomes of cell lines, tissues, tumors, and diseased states. Since the information about sequential cells is only partial, cluster analysis is usually used to discover cell subtypes or to distinguish and better characterize known cell subtypes [1].

Unlike bulk RNA-seq data, single-cell RNA-seq data are more sparse and have a high dropout, which make clustering very challenging. Recently, several methods and tools have been developed for single-cell RNA-seq clustering. [2] used non-negative matrix factorization to incorporate information from a larger annotated dataset and then applies transfer learning to perform the clustering. CIDR (Clusteing through Imputation and Dimensionlity Reduction) performs data imputation before clustering a PCA-reduced (Principal Component Analysis) representation using hierarchical clustering [3]. SOUP (Semisoft Clustering with Pure cells) handles both pure and transitional cells and uses the expression similarity matrix to compute soft cluster memberships [4]. [5] proposed, Isomap, a parallelized dimensionality reduction method. and [6] presented a mathod to improved protein function prediction.

Most of the available clustering techniques are inevitably unable to deal with the noise and outliers of data well. A new machine learning framework called curriculum learning (CL) [7] has gotten a lot of interest as a solution to these issues. The concept stems from the reality that people learn better when they begin with simple knowledge and work their way up to more complicated knowledge. To formalize this strategy in machine learning, [8] proposed self-paced learning (SPL) to achieve the curriculum designing's aim by including an SPL regularization term in the objective function. For example, SPL first trains the clustering model on "simple" examples before progressively adding "difficult" instances to cluster. SPL has been shown to be capable of avoiding the noise and outliers, so SPL has superior generalization ability [8], [9], [10], [11]. Ren et al. employed SPL to overcome the non-convex problem caused by feature corruption approaches [12]. As a result, SPL is frequently used to improve solutions for non-convex problems with many local minima.

Due to the non-convexity of the NMF model, it is simple for these approaches to produce unsatisfactory local solutions when noise and outliers are present. This research offers a soft self-step learning-based NMF-based single-cell RNA-seq data clustering approach (S3NMF) to address this problem. The stepwise integration of single cells into the NMF process, from basic to complex, takes full advantage of the benefits of SPL and has been found to assist model clustering avoid the effects of noise and outliers.

In summary, we highlight our contributions of this paper in the following:

- We utilize a novel NMF-based method for single cell RNA-seq data clustering in the framework of soft self-paced learning. The outliers and noise from scRNA-seq

∗: Corresponding author.

978-1-6654-8219-6/22/$31.00 ©2022 IEEE

data are addressed.

- We conduct experiments on both simulation and real scRNA-seq datasets, and compare our result against several baselines. Experimental results verify the effectiveness of our proposed method.

II. PRELIMINARIES

A. Nonnegative Matrix Factorization

After conducting the data pre-processing, the scRNA-seq data can be represented as a matrix $\mathbf{E} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n] \in \mathbb{R}^{m\times n}$, where n is the number of cells and m is the genes. $\mathbf{E}_{ij}$ denotes the gene expression of gene i in the single cell j. NMF [13] aims to find two nonnegative matrices $\mathbf{U} \in \mathbb{R}^{m\times k}$ and $\mathbf{V} \in \mathbb{R}^{n\times k}$ which minimize the following objective function

$$\begin{aligned} J &= \|\mathbf{E} - \mathbf{U}\mathbf{V}^T\|_F^2 \\ \text{s.t.} \quad & \mathbf{U} \geq 0, \mathbf{V} \geq 0, \end{aligned} \tag{1}$$

where $\|\cdot\|_F$ is Frobenius norm (F-norm). Lee et al. [13] proposed an algorithm of alternating iteration U and V to optimize the objective function.

In the clustering setting of NMF [14], $\mathbf{V} \in \mathbb{R}^{n\times k}$ is the cluster assignment matrix where k is the number of clusters. In reality, we have $k \ll n$ and $k \ll m$.

Noting that the i-th row of $\mathbf{V}$ can be regarded as the low-dimensional representation of i-th single cell with respect to the new basis $\mathbf{U}$.

B. Self-paced learning

Before introducing our innovative work, we briefly introduce the original self-paced learning (SPL) framework as bridging knowledge.

The goal of SPL is to learn a weight variable $\mathbf{w} = [w_1, \ldots, w_n]^T$ as well as the model parameter $\boldsymbol{\theta}$. The original objective function of SPL is as follows [8]:

$$\min_{\boldsymbol{\theta},\mathbf{w}} J(\boldsymbol{\theta}, \mathbf{w}; \lambda) = \sum_{i=1}^{n} w_i l_i + f(\lambda, \mathbf{w}), \tag{2}$$

where l_i denotes the reconstruction error which is computed by a loss function, and $\boldsymbol{\theta}$ is the model parameter of the loss function. λ denotes a regular term coefficient. When $\mathbf{w}$ is fixed, Eq. (2) denotes traditional machine learning problem. While $\boldsymbol{\theta}$ is fixed, the solution of $\mathbf{w}$ depends on the definition of $f(\lambda, \mathbf{w})$ and the range of values of $\mathbf{w}$. [8] lets $\mathbf{w} \in \{0, 1\}$ and defines $f(\lambda, \mathbf{w})$ as

$$f(\lambda, \mathbf{w}) = -\lambda \sum_{i=1}^{n} w_i, \tag{3}$$

and the optimal $\mathbf{w}^*$ can be calculated by

$$w_i^* = \begin{cases} 1, & \text{if } l_i < \lambda \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

Since w_i $(i = 1, \ldots, n)$ is either 1 or 0, the strategy mentioned above can be treated as hard weighting. Here, $\lambda > 0$ is initially tuned to a small value such that the data points with small loss values can be selected to train. With the increasing of λ, more and more data points will be selected until all data points are chosen.

III. METHODS

A. S3NMF model

To the best of our knowledge, We have greatly improved this model to processing single-cell RNA-seq clustering data. Based on the SPL regularization term presented in Eq. (3), The objective function is as follows:

$$\begin{aligned} \min_{\mathbf{U},\mathbf{V},\mathbf{w}} & \|diag(\mathbf{w})(\mathbf{E} - \mathbf{U}\mathbf{V}^T)\|_F + f(\lambda, \mathbf{w}) \\ \text{s.t.} \quad & \mathbf{U} \geq 0, \mathbf{V} \geq 0, \mathbf{w} \in [0, 1]^n, \end{aligned} \tag{5}$$

where w_i is the i-th diagonal element in diagonal matrix $diag(\mathbf{w})$. One of the simple regular functions $f(\lambda, \mathbf{w})$ is shown in Eq. (3). The core idea of S3NMF proposed in this paper is gradually select cells for decomposition from simple to complex. This strategy is similar to the successful application in the field of face image [15].

B. Optimization

We considered an iterative updating algorithm to solve the optimization problem of S3NMF. Specifically, the objective function will be optimized with regard to one variable while fixing other variables.

*a) **Step 1**:* Fix $\mathbf{U}$ and $\mathbf{V}$, update $\mathbf{w}$.

With the fixed parameters $\mathbf{U}$ and $\mathbf{V}$. The weight matrix $diag(\mathbf{w})$ is updated by

$$\mathbf{w}^* = \arg\min \sum_{i=1}^{n} w_i l_i + f(\lambda, \mathbf{w}), \tag{6}$$

where $l_i = \|\mathbf{e}_i - \mathbf{U}\mathbf{v}_i\|_2$.

We can observe from Eq. (2) that SPL chooses data points based on their loss values and a global λ.

To alleviate this, we consider assigning weights and gradually selecting data points from easy to complex. We choose to utilize a novel SPL regularization term.

The regularization term is defined as

$$f(\lambda, \mathbf{w}) = -\sum_{i=1}^{n} \zeta \ln(w_i + \zeta/\lambda), \tag{7}$$

and the optimal $\mathbf{w}^*$ is computed by

$$w_i^* = \begin{cases} 1, & \text{if } l_i \leq \zeta\lambda/(\zeta + \lambda) \\ 0, & \text{if } l_i \geq \lambda \\ \zeta/l_i - \zeta/\lambda, & \text{otherwise.} \end{cases} \tag{8}$$

It is clear that Eq. (8) is soft weighting strategy. And we set $\zeta = 0.5 \times \lambda$ for simplicity in our experiments. According to [16], Eq. (8) is also called mixture weighting.

*b) **Step 2**:* Fix $\mathbf{w}$, update $\mathbf{U}$ and $\mathbf{V}$.

When we fix $\mathbf{w}$, $f(\lambda, \mathbf{w})$ in Eq. (5) is a constant. Thus we can update the model parameters $\mathbf{U}$ and $\mathbf{V}$ iteratively as follows: $(\mathbf{a})$ Fix $\mathbf{V}$, update $\mathbf{U}$. $(\mathbf{b})$ Fix $\mathbf{U}$, update $\mathbf{V}$. The problem corresponds to the weighted NMF problem [17] and related algorithms can be employed for solving it.

Until now, we have all the update rules done. We optimize the model in an iterative way. That is, steps 1 and 2 are iteratively repeated until the model convergence. For detailed algorithms, please refer to [15]. We increase λ to select more cells to the factorization process. Specifically, we initialize λ such that more than half cells are picked in the first iteration. In the following iteration, λ is increased such that 10% more instances can be added. As a consequence, λ is automatically determined. And the model repeats until all the data points are chosen. Finally, Kmeans clustering is applied to the matrix V after iteration and the clustering results of scRNA-seq data are obtained. The clustering results will be evaluated and analyzed on the experiment section.

IV. Experiments

A. Evaluation metrics

All clustering results are measured by Acc (Clustering Accuracy), NMI, and Purity [18]. These cluster evaluation indicators will be briefly introduced as follows. Acc is used for discovering the one-to-one relationship between classes and clusters along with measure the extent to which each cluster contained data instances from the corresponding class. NMI (Normalized Mutual information) measures the amount of information obtained about one partition through observing the other partition, ignoring the permutations. Purity is applied to measure the extent to which each cluster contained data instances from primarily one class. Purity is quite simple to calculate. The larger the value of these evaluation indexes is, the better the clustering performance is.

B. Datasets

We will show the performance of the proposed method on both simulated data and the two real single-cell RNA-seq datasets. The statistical information of all datasets used in this study are shown in Tab I. We generated simulated data to evaluate the cluster performance by a tool named Splatter [19]. In the simulation data, the cells contained two different clusters of 100 cells each, each containing 10,000 genes. In the following chapters, we will introduce how to screen some genes for the calculation of cell clustering. The pre-processed read count matrix is treated as the input for our model and the others compared algorithms. Raw scRAN-seq read count data are pre-processed by the Python package Scanpy [20].

C. Baselines

To evaluate the performance of the proposed S3NMF, we compared it with several closely related nonnegative matrix factorization methods for scRNA-seq clustering:

TABLE I
A SUMMARY OF THE SCRNA-SEQ DATASETS USED IN THIS STUDY

datasets	# clusters	# cells	# genes
simulated data [19]	2	200	10000
human data [21]	2	171	887
mouse cortex [21]	2	2383	910

- K-means [22]. This is a simple iterative method to partition the given dataset into a cell-specified number of clusters, k.
- NMF [13]. This is a low-rank matrix approximation method for finding two low-rank non-negative matrices whose product provides a good approximation to the original non-negative matrix.
- ONMF [23]. This is a NMF-based method with orthogonal transformation.
- $l_{2,1}$-NMF [24]. This is a NMF-based method which exploits $l_{2,1}$-norm.

V. Experimental Results

A. Parameter settings

As described in section III, Cells are gradually added to the factorization model by increasing the λ. In the first iteration, we set λ such that 60% of the cells are selected. Then λ is increased such that 10% more samples is added in every following iteration. The matrix factorization component defaults to the number of true cell clusters. For simulation data, we selected 2000 highly variables genes (HVGs) [25] as inputs to our model.

B. Results for simulated data

To test the effectiveness of S3NMF we used in single-cell sequencing data, we now test it on simulated data. The results show that this method is superior to some algorithms proposed before and achieves the best performance. The detail of cluster result are show in Tab. II.

TABLE II
CLUSTERING RESULT ON SIMULATED SCRNA-SEQ DATA

Datasets	Acc	NMI	Purity
K-means	0.5135 ± 0.0147	0.0075 ± 0.0086	0.5135 ± 0.0147
NMF	0.5940 ± 0.0588	0.0373 ± 0.0411	0.5940 ± 0.0588
ONMF	0.5310 ± 0.0218	0.0122 ± 0.0113	0.5310 ± 0.0218
l_{21}**-NMF**	0.6025 ± 0.0466	0.0587 ± 0.0432	0.6025 ± 0.0466
S3NMF	$\mathbf{0.6953 \pm 0.0637}$	$\mathbf{0.1336 \pm 0.0757}$	$\mathbf{0.6953 \pm 0.0637}$

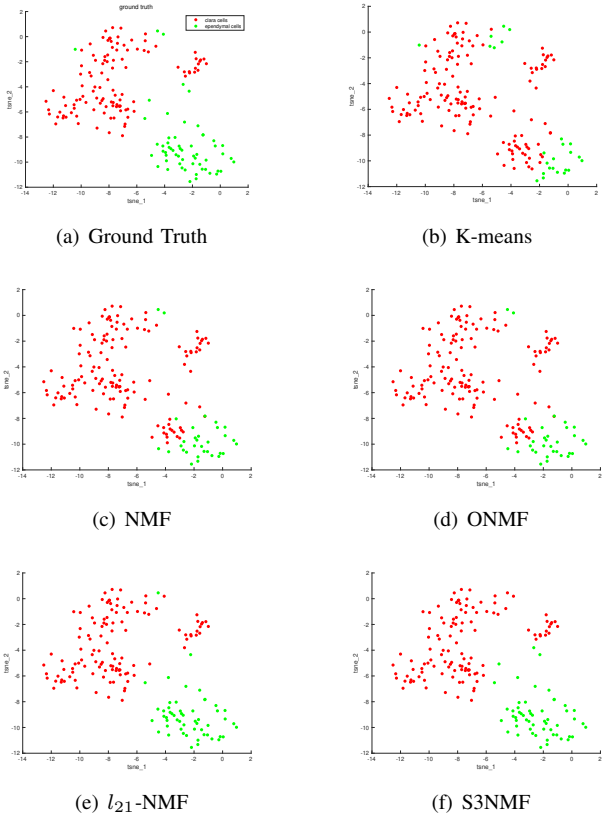
C. Example 1: clustering pulmonary alveolar type II, clara and ependymal cells of human scRNA-seq data

In the real dataset experiment, we first test on data containing 113 clara cells and 58 ependymal cells in the human scRNA-seq data. We used the true cell labels as a benchmark for evaluating the performance of the clustering methods. The detail of cluster result are show in Tab. III. In Fig. 1 shows t-sne for pulmonary alveolar type II, clara and ependymal cells of human scRNA-seq data cluster results. It can be seen from

the experimental results that the performance of S3NMF is better than other comparison methods in all three evaluation indexs, which is Acc, NMI, Purity. The NMI of S3NMF is almost 10% higher than that of the l_{21} based NMF.

TABLE III
CLUSTERING RESULT ON HUMAN SCRNA-SEQ DATA

Datasets	Acc	NMI	Purity
K-means	0.7317 ± 0.1241	0.3976 ± 0.1674	0.8671 ± 0.0712
NMF	0.7679 ± 0.1291	0.4583 ± 0.2438	0.8585 ± 0.0916
ONMF	0.8862 ± 0.0972	0.6805 ± 0.1884	0.9448 ± 0.0760
l_{21}**-NMF**	0.9376 ± 0.0102	0.7332 ± 0.0424	0.9768 ± 0.0073
S3NMF	$\mathbf{0.9616 \pm 0.0056}$	$\mathbf{0.8505 \pm 0.0121}$	$\mathbf{0.9825 \pm 0.0333}$



(a) Ground Truth (b) K-means (c) NMF (d) ONMF (e) l_{21}-NMF (f) S3NMF

Fig. 1. t-sne for pulmonary alveolar type II, clara and ependymal cells of human scRNA-seq data cluster results. (a) t-sne for ground truth.clara cells is shown in red, ependymal cells in blue. $(b) - (f)$ t-sne for S3NMF and The comparison algorithm of this study. The red dots represent Clara cells and the green dots represent Ependymal cells.

D. Example 2: clustering L4 and L2/3 cells in mouse cortex scRNA-seq data

To fully examine the validity of S3NMF on different single-cell data, we tested the algorithm on mouse cortex scRNA-seq data. The dataset contains two types of cell lines (1401 L4 and 982 L2/3 IT scRNA-seq cells, Where 'L4' and 'L2/3' stand for excitatory neurons in different neocortical layers; IT is the abbreviation of intratelencephalic neuron). We used the provided cell type labels as a benchmark for evaluating the performance of the clustering methods. The experimental results show that the clustering accuracy of this method is very high in the easy to distinguish single cell type clustering problem, and even higher than other advanced comparison algorithms.

TABLE IV
CLUSTERING RESULT ON MOUSE CORTEX SCRNA-SEQ DATA

Datasets	Acc	NMI	Purity
K-means	0.8552 ± 0.0448	0.6311 ± 0.0433	0.9948 ± 0.0013
NMF	0.8615 ± 0.0453	0.6406 ± 0.0432	0.9960 ± 0.0008
ONMF	0.9069 ± 0.0157	0.6820 ± 0.0140	0.9966 ± 0.0000
l_{21}**-NMF**	0.9074 ± 0.0155	0.6755 ± 0.0111	0.9933 ± 0.0000
S3NMF	$\mathbf{0.9201 \pm 0.0419}$	$\mathbf{0.7039 \pm 0.0383}$	$\mathbf{0.9960 \pm 0.0005}$

VI. CONCLUSION

In this study, a new sample selection strategy, self-paced learning, is introduced for scRNA-seq data clustering. Cells are grouped into clustered samples from easy to hard based on the loss of initialization. The cells were added into the clustering algorithm from easy to difficult, so as to avoid the influence of noise and outliers on the algorithm effectively, and avoid the local optimal dilemma of other algorithms easily. Experimental results show that S3NMF is effective on scRNA-seq data clustering and the best performance are obtained on simulation data and two real single cell transcriptome datasets. Sometimes single-cell data vary widely in the number of cells each category contains. We will improve the algorithm to adapt to the impact of unbalanced data on the clustering results in the future work. At the same time, as the scale of sequencing data expands, we will also focus on how to improve the algorithm to adapt to big data.

Acknowledgments. This work was supported by National Natural Science Foundation of China (No. 32100530) and National Key Research and Development Project (No. 2018YFB2101200).

REFERENCES

[1] Sui Huang. Non-genetic heterogeneity of cells in development: more than just noise. *Development*, 136(23):3853–3862, 2009.
[2] Bettina Mieth, James RF Hockley, Nico Görnitz, Marina M-C Vidovic, Klaus-Robert Müller, Alex Gutteridge, and Daniel Ziemek. Using transfer learning from prior reference knowledge to improve the clustering of single-cell rna-seq data. *Scientific reports*, 9(1):1–14, 2019.
[3] Peijie Lin, Michael Troup, and Joshua WK Ho. Cidr: Ultrafast and accurate clustering through imputation for single-cell rna-seq data. *Genome biology*, 18(1):1–11, 2017.
[4] Lingxue Zhu, Jing Lei, Lambertus Klei, Bernie Devlin, and Kathryn Roeder. Semisoft clustering of single-cell data. *Proceedings of the National Academy of Sciences*, 116(2):466–471, 2019.
[5] Arpita Joshi and Nurit Haspel. Clustering of protein conformations using parallelized dimensionality reduction. *Journal of Advances in Information Technology*, 2019.
[6] Haneen Altartouri and Tobias Glasmachers. Improved protein function prediction by combining clustering with ensemble classification,". *Journal of Advances in Information Technology (JAIT)*, 2021.
[7] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th International Conference on Machine Learning*, pages 41–48, 2009.
[8] M. Pawan Kumar, Benjamin Packer, and Daphne Koller. Self-paced learning for latent variable models. In *Advances in Neural Information Processing Systems*, pages 1189–1197, 2010.

[9] M. Pawan Kumar, Haithem Turki, Dan Preston, and Daphne Koller. Learning specific-class segmentation from diverse data. In *Proceedings of IEEE International Conference on Computer Vision*, pages 1800–1807, 2011.

[10] Lu Jiang, Deyu Meng, Qian Zhao, Shiguang Shan, and Alexander G. Hauptmann. Self-paced curriculum learning. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence*, pages 2694–2900, 2015.

[11] Kevin Tang, Vignesh Ramanathan, Fei-Fei Li, and Daphne Koller. Shifting weights: Adapting object detectors from image to video. In *Advances in Neural Information Processing Systems*, pages 647–655, 2012.

[12] Yazhou Ren, Peng Zhao, Zenglin Xu, and Dezhong Yao. Balanced self-paced learning with feature corruption. In *Proceedings of the International Joint Conference on Neural Networks*, pages 2064–2071, 2017.

[13] Daniel D Lee and H Sebastian Seung. Algorithms for non-negative matrix factorization. In *Advances in Neural Information Processing Systems*, pages 556–562, 2001.

[14] V Paul Pauca, Farial Shahnaz, Michael W Berry, and Robert J Plemmons. Text mining using non-negative matrix factorizations. In *SIAM International Conference on Data Mining*, pages 452–456, 2004.

[15] Xiangxiang Zhu and Zhuosheng Zhang. Improved self-paced learning framework for nonnegative matrix factorization. *Pattern Recognition Letters*, 97:1–7, 2017.

[16] Lu Jiang, Deyu Meng, Teruko Mitamura, and Alexander G. Hauptmann. Easy samples first: Self-paced reranking for zero-example multimedia search. In *Proceedings of the 22nd ACM International Conference on Multimedia*, pages 547–556. ACM, 2014.

[17] Yong-Deok Kim and Seungjin Choi. Weighted nonnegative matrix factorization. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1541–1544, 2009.

[18] Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of classification*, 2(1):193–218, 1985.

[19] Luke Zappia, Belinda Phipson, and Alicia Oshlack. Splatter: simulation of single-cell rna sequencing data. *Genome biology*, 18(1):1–15, 2017.

[20] F Alexander Wolf, Philipp Angerer, and Fabian J Theis. Scanpy: large-scale single-cell gene expression data analysis. *Genome biology*, 19(1):1–5, 2018.

[21] Oscar Franzén, Li-Ming Gan, and Johan LM Björkegren. Panglaodb: a web server for exploration of mouse and human single-cell rna sequencing data. *Database*, 2019:1–9, 2019.

[22] J. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297. University of California Press, 1967.

[23] Chris Ding, Tao Li, Wei Peng, and Haesun Park. Orthogonal nonnegative matrix t-factorizations for clustering. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 126–135, 2006.

[24] Deguang Kong, Chris Ding, and Heng Huang. Robust nonnegative matrix factorization using l21-norm. In *International on Conference on Information and Knowledge Management*, pages 673–682, 2011.

[25] Shun H Yip, Pak Chung Sham, and Junwen Wang. Evaluation of tools for highly variable gene discovery from single-cell rna-seq data. *Briefings in bioinformatics*, 20(4):1583–1589, 2019.