Applied Mathematical Modelling 90 (2021) 875–888

Contents lists available at ScienceDirect

Applied Mathematical Modelling

journal homepage: www.elsevier.com/locate/apm

A kernel non-negative matrix factorization framework for single cell clustering

Hao Jiang[a,*], Ming Yi[b], Shihua Zhang[c]

[a] School of Mathematics, Renmin University of China, Beijing 100872, China
[b] School of Mathematics and Physics, China University of Geosciences, Wuhan, China
[c] Academy of Mathematics and Systems Science, CAS, Beijing 100190, China

ARTICLE INFO

Article history:
Received 27 February 2020
Revised 18 August 2020
Accepted 22 August 2020
Available online 26 September 2020

Keywords:
Single cell RNA-sequencing
Kernel non-negative matrix factorization
Cellular heterogeneity,

ABSTRACT

The emergence of single-cell RNA-sequencing is ideally placed to unravel cellular heterogeneity in biological systems, an extremely challenging problem in single cell RNA-sequencing studies. However, most current computational approaches lack the sensitivity to reliably detect nonlinear gene-gene relationships masked by dropout events. We proposed a kernel non-negative matrix factorization framework for detecting nonlinear relationships among genes, where the new kernel is developed using kernel tricks on cellular differentiability correlation. The newly constructed kernel not only provides a description on the gene-gene relationship, but also helps to build a new low-dimensional representation on the original data. Besides, we developed an efficient method for determining the optimal cluster number within each data set with the usage of Diffusion Maps. The proposed algorithm is further compared with representative algorithms: SC3 and several other state-of-the-art clustering methods, on several benchmark or real scRNA-Seq datasets using internal criteria (clustering number accuracy) and external criteria (Adjusted rand index and Normalized mutual information) to show effectiveness of our method.

1. Introduction

RNA-Seq measures the abundance of RNA transcripts, enabling the study of transcriptional structures, gene expression levels and so on [1]. In the past few years, bulk cell RNA-sequencing technology has become one of the most popular technologies for transcriptome profiling. However, it cannot address cellular heterogeneity problem for its inability to isolate individual cells from complex organisms and measure the corresponding transcriptional activity. Recent advances in single cell RNA-sequencing technology allow for quantification of intra-population heterogeneity across a range of cellular states at the level of single cell [2]. Consequently, scRNA-seq has gradually become the method of choice for understanding cellular heterogeneity in different complex biological conditions. One perspective is cell type identification [3–5]. In [3], hundreds of cells from mouse intestinal organoids were sequenced at the transcriptome level for characterization of rare cell types. In [4], authors proposed a method for detecting rare cell types from single-cell gene expression data. Tsoucas and Yuan [5] can successfully identify rare and common cell types at the same time with single cell RNA-sequencing data. Another perspective is cell development processes [6–9]. Wanderlust in [6] aligns single cells on a trajectory according to their developmental path. In [7], 124 individual cells including human oocytes and early embryos were sequenced using single cell RNA-sequencing

[*] Corresponding author.
E-mail addresses: jiangh@ruc.edu.cn (H. Jiang), yiming@cug.edu.cn (M. Yi), zsh@amss.ac.cn (S. Zhang).

https://doi.org/10.1016/j.apm.2020.08.065

technology to reveal the developmental heterogeneity. Chu et al. [8] used scRNA-seq technology to reveal novel regulators during stem cell differentiation. In [9], single-cell data was used to reconstruct developmental trajectories during zebrafish embryogenesis. Other perspectives include cancer evolution [10–12]. Gao et al. [10] studied evolution of copy number alterations from primary to circulating tumor cells. Brady et al. [11] reveals acquisition of malignant phenotypes after treatment with single cell analysis. Kim et al. [12] used single cell sequencing to reveal chemoresistance evolution in breast cancer. Tumor heterogeneity can also be revealed by single cell sequencing [13,14]. scRNA-seq data is structurally identical to that obtained from a bulk expression experiment, often has high dimension and highly noisy. In particular, scRNA-seq data is extremely sparse, having a lot of excess zeros. This is a typical phenomenon in scRNA-seq data where a gene is observed at a moderate expression level in some cell but undetected in other cells [15], making further analysis more difficult and challenging.

One way to deal with scRNA-seq data is to find an accurate representation for the high dimensional, noisy and sparse data. Several imputation approaches have been designed in the recent years, all of whom have various model assumptions. Bayesian approaches include BISCUIT, URSM and SAVER [16–18] where the former two methods were reported to be computational comprehensive. DrImpute [19] imputes zero values in scRNA-seq data by averaging strategy within the same cluster, scImpute [2] recovers gene expression profiles through non-negative least square model, and MAGIC [20] imputes zero values through Markov affinity graph. Taking into consideration on the high dimensionality in scRNA-seq data, researchers have also developed several dimension reduction methods. Identification of highly variable genes assumes that genes with high level of variance are caused by biological effects rather than mere technical noise [21]. Brennecke et al. [22] used a gamma generalized linear model to estimate the highly variable genes. BiSiCS [23] introduces Poisson-Gamma hierarchical models to identify highly variable genes. In [24], a simulation based framework was proposed to test the highly variable genes. M3Drop [25] implements two dropout-based feature selection methods with specific models for the null expectation. Another main approach for dealing with the curse of dimensionality is to project the high dimensional scRNA-seq data into a lower dimensional space. Pierson and Yau [26] applied zero-inflated factor analysis on the scRNA-seq data. tSNE is the most frequently used method for visualizing high dimensional datasets [2,27,28], however the drawback of which lies in the stochastic nature in producing different embeddings for the same dataset. As a nonliear projection method, diffusion maps preserve local and distant relationships between data points and thus are very suitable for large data sets [29].

Much effort has been devoted to find accurate representation of scRNA-seq data either in terms of imputation or dimension reduction, however most of the methods rely on different kinds of model assumptions thereby restricting the applications. Some recent work tried to seek an appropriate description on the relationship within the high dimensional, sparse and noisy scRNA-seq data. SNN-Cliq [30] defines the similarity between two data points using connectivity in the neighborhood, proposing a shared nearest neighborhood (SNN) similarity measure. However, the nearest neighbor identification in high dimensional data needs careful specification and the inherent graph structure in scRNA-seq data cannot always be satisfied. SIMLR [31] presents a multi-kernel learning framework to obtain the similarity between data points using a linear combination of gaussian kernel functions under different variance settings. It needs the predefinition of cluster number within the data set to start the learning process, and the usage of gaussian kernels in measuring data similarity may not always be suitable, sometimes can make the learning process stuck into overfitting. SC3 [32] combines principal component analysis and k-means clustering algorithm to construct a consensus similarity measure for the scRNA-seq data, proves to be a very robust method in identifying cellular heterogeneity. One of the main highlights in the method lies in the cluster-based similarity partitioning algorithm for consensus matrix construction, contributing to the stable and powerful performance. But in the preprocessing, principal component analysis as a linear dimension reduction method may not be appropriate for all the scRNA-seq data. Jiang et al. [33] proposed a new similarity construction method using cell-pair correlation, and the method shows robust power in cell type identification for small scRNA-seq data sets. The constructed similarity measure can be rewritten as a dot product and therefore can be regarded as a data transformation method. Results show that the transformation method gives a better representation on the original data and can help discover potential biomarkers.

In this paper, we focus on cell type identification problem and we will develop a kernel non-negative matrix factorization framework for learning similarity in the scRNA-seq data. The constructed kernel models the nonlinear relationship within the scRNA-seq data and we can get a low dimensional representation for the original data after non-negative matrix factorization on the constructed kernel matrix. Through the usage of cluster-based similarity partitioning algorithm, we finally construct a consensus matrix for measuring the relationships among the data points.

2. Methods

Assume that we have an expression matrix from scRNA-seq data denoted as $V = [v_1, v_2, \ldots, v_n] \in R^{p\times n}$, where n is the number os cells and p is the number of attributes used to represent a cell. In the following, we first give a brief introduction on non-negative matrix factorization and then we propose our kernel non-negative matrix factorization framework.

2.1. Preliminaries: Non-negative Matrix Factorization (NMF)

NMF [34] realizes non-negative factorization on V by finding two non-negative matrices: $W = [w_1, w_2, \ldots, w_k] \in R^{p\times k}$ and $H = [h_1, h_2, \ldots, h_n] \in R^{k\times n}$ such that

$$V \approx WH$$

where k is a specified parameter of reduced dimensionality, the columns of W represent the base vectors for transforming the original data V, and columns of H correspond to the coefficients describing each cell.

The solutions of W, H can be found through solving the following optimization problem

$$\min_{W\geq 0, H\geq 0} \sqrt{\sum_{i=1}^{p}\sum_{j=1}^{n}(V_{ij} - (WH)_{ij})^2} \tag{1}$$

where the objective function is actually the Frobenius norm to measure the error between the original matrix V and its low rank approximation WH.

However, NMF is essentially linear hence can not fully capture the nonlinear structures embedded in the scRNA-seq data. Kernel non-negative matrix factorization (KNMF) [35,36] as an nonlinear extension can be incorporated into scRNA-seq data analysis.

Assume that we have a nonlinear mapping ϕ from V to a feature space F of higher dimension or infinity dimension

$$\phi : v \in V \to \phi(v) \in F$$

Therefore, we have $\phi(V) = [\phi(v_1), \phi(v_2), \dots, \phi(v_n)]$. In a similar fashion to NMF, KNMF also tries to find two non-negative matrix W_ϕ, H_ϕ such that

$$\phi(V) \approx W_\phi H_\phi$$

When the nonlinear mapping ϕ is of infinity dimension, it is impractical to factorize $\phi(V)$ in a direct manner. But from $\phi(V) \approx W_\phi H_\phi$, we can derive the formula in the following:

$$\phi(V)^T\phi(V) \approx \phi(V)^T W_\phi H_\phi$$

The formula indicates that $K = \phi(V)^T\phi(V) \in R^{n\times n}$ is a valid kernel matrix and $\phi(V)$ is exactly the kernel induced nonlinear mapping. Therefore, $K = W_\phi H_\phi, W_\phi^K = \phi(V)^T W_\phi$ shares the same formalism with that of NMF, where W_ϕ^K are the bases for kernel matrix K and H_ϕ represents the coefficients matrix as dimension reduced representation for V. KNMF [35,36] naturally extends the NMF framework where it needs a careful description on the nonlinear mapping on the original data V, or alternatively constructing a delicate kernel for describing the relationship between different data points in scRNA-seq data set.

2.2. KDCorr: Kernel on Differentiability Correlation

There are a number of popular kernels:

- Linear Kernel.

  $$K(\mathbf{x}, \mathbf{x}') = \mathbf{x}^T\mathbf{x}',$$

  which is an inner product of $\mathbf{x}$ and $\mathbf{x}'$ in $\mathbf{R}^p$.
- Polynomial kernel

  $$K(\mathbf{x}, \mathbf{x}') = (\mathbf{x}^T\mathbf{x}' + 1)^d,$$
- Gaussian Radial Basis Function (RBF) kernel

  $$K(\mathbf{x}, \mathbf{x}') = \exp(-\frac{||\mathbf{x} - \mathbf{x}'||^2}{\sigma^2})$$

  where σ is parameter.

While scRNA-seq data is of high dimensionality, highly sparse and noisy, the above kernels may not be appropriate. In [31], authors proposed using a linear combination of RBF kernel functions under different variance settings, but the selection of variance is also a problem need to be addressed. Jiang et al. [33] proposed to use cell-pair correlation for modeling cellular relationships based on discrete version of differentiability. The method is computationally expensive when the involved number of cells become relatively large, say greater than 500. We here propose a kernel incorporating the advantages shown in [33], and is efficient for dealing with large scale data sets.

Clustering with *KDCorr-NMF*

In the first step, we define a continuous version of differentiability for each cell i in gene k

$$W_{i_k} = v_{ki} - \frac{(\sum_{l=1}^{n} v_{kl}) - v_{ki}}{n-1},$$

In the later step, we calculate the differentiability correlation among the given n cells in the data set

$$\text{DCorr}(i,j) = \frac{\sum_{k=1}^{p}\left(W_{i_k} - \overline{W_i}\right)\left(W_{j_k} - \overline{W_j}\right)}{\sqrt{\sum_{k=1}^{p}\left(W_{i_k} - \overline{W_i}\right)^2}\sqrt{\sum_{k=1}^{p}\left(W_{j_k} - \overline{W_j}\right)^2}}$$

where $\overline{W_k} = \sum_{l=1}^{p} W_{k_l}/p$.

The *KDCorr* kernel is then constructed as

$$\mathrm{KDCorr}(i,j)=\exp^{(\frac{1}{2}(\mathrm{Dcorr}(i,j)-1))}, i,j\in\{1,2,\dots,n\} \tag{2}$$

It can be easily proved that *KDCorr* is a valid kernel as by kernel trick [37].

Proof. We rewrite *KDCorr* in the following formulation.

$$\mathrm{KDCorr}(i,j)=\mathrm{e}^{-\frac{1}{2}}\exp^{\frac{1}{2}\mathrm{Dcorr}(i,j)}, i,j\in\{1,2,\dots,n\}$$

From Proposition 1 we have $\frac{1}{2}$Dcorr is a valid kernel, and we further get from Proposition 2 to know that $\mathrm{e}^{-\frac{1}{2}}\exp^{\frac{1}{2}\mathrm{Dcorr}(i,j)}$ is a valid kernel. □

We here list the related propositions without proof.

Proposition 1. *Let k_1 and k_2 be kernels over $X\times X, X\subset R^p, x,z\in X, a\in R^+, f(\cdot)$ a real-valued function on X, ϕ: $X\to R^N$ with k_3 a kernel over $R^N\times R^N$, and* ***B*** *a symmetric positive semi-definite $n\times n$ matrix. Then the following functions are kernels:*

$$\begin{cases}k(x,z)=k_1(x,z)+k_2(x,z)\\ k(x,z)=ak_1(x,z)\\ k(x,z)=k_1(x,z)k_2(x,z)\\ k(x,z)=f(x)f(z)\\ k(x,z)=k_2(\phi(x),\phi(z))\\ k(x,z)=x'\mathbf{B}z\end{cases}$$

Proposition 2. *Let k_1 be a kernel over $X\times X$, $X\subset R^p$, x, $z\in X$, $p(x)$ is a polynomial with positive coefficients. Then the following functions are also kernels. Then the following functions are kernels:*

$$\begin{cases}k(x,z)=p(k_1(x,z))\\ k(x,z)=\exp(k_1(x,z))\\ k(x,z)=\exp(-||x-z||^2/(2\sigma^2))\end{cases}$$

2.3. KDCorr for kernel non-negative matrix factorization framework in cellular heterogeneity analysis

KDCorr kernel evaluates the nonlinear relationship among cells in scRNA-seq data, suppose we have a factorization $K=W_\phi H_\phi$,where $W_\phi\in R^{n\times k}$, $H_\phi\in R^{k\times n}$, we can then use H_ϕ as the low-dimension representation for the original data and perform clustering to discover different cell types.

It should be noted that due to the non-convexity of $||K-W_\phi H_\phi||_F^2$ in both W_ϕ and H_ϕ, the numerical solutions of Eq. (1) may always converge to local minima. We here propose alternating least square algorithm to solve the minimization problem by considering a regularized optimization framework in the following:

$$\min_{W_\phi\geq0,H_\phi\geq0}L(W_\phi,H_\phi)=\min_{W_\phi\geq0,H_\phi\geq0}\sum_{i=1}^{n}\sum_{j=1}^{n}(K_{ij}-(w_i^\phi)^T h_j^\phi)^2+\lambda(||w_i^\phi||^2+||h_j^\phi||^2) \tag{3}$$

where $W_\phi=\begin{pmatrix}(w_1^\phi)^T\\ \vdots\\ (w_n^\phi)^T\end{pmatrix}$ and $H_\phi=\begin{pmatrix}(h_1^\phi) & \cdots & (h_n^\phi)\end{pmatrix}$.

Through differentiation on w_i^ϕ and h_j^ϕ, $i,j=1,2,\dots,n$, we can have

$$W_\phi^T=(H_\phi H_\phi^T+\lambda I)^{-1}H_\phi K^T$$

$$H_\phi=(W_\phi^T W_\phi+\lambda I)^{-1}W_\phi^T K$$

For a fixed low dimensionality k, the alternating least square algorithm involves computational complexity $O(2nk^3+n^2k^2)$. It does not depend on the attribute dimensionality p, hence can become more efficient.

The complete algorithm of the proposed method '*KDCorr-NMF*' is illustrated in the following Algorithm 1.

This algorithm is a consensus method incorporating non-negative matrix factorization on KDCorr kernel.

2.4. Optimal cluster number: diffusion maps with variance analysis

When the number of involved cells become large, the variance analysis based optimal cluster number is no longer efficient [33]. Here we propose to use diffusion maps [38] to construct a method for determining optimal cluster number C_{opt}

as it is very suitable for large data sets. The diffusion maps Ψ_t: $X \in R^p \to R^{s(\delta,t)}$ realize dimension reduction by embedding high-dimensional data into Euclidean space of $s(\delta, t)$ dimension in which the Euclidean distances best approximate the diffusion distances

$$||\Psi_t(v_i) - \Psi_t(v_j)|| \approx D_t(v_i, v_j)$$

where

$$D_t(v_i, v_j) = \left(\sum_{i=1}^{n} \lambda_l^{2t} (\psi_l(v_i) - \psi_l(v_j))^2 \right)^{\frac{1}{2}}$$

is the diffusion distance between v_i and v_j and

$$\Psi_t(v) = \begin{pmatrix} \lambda_1^t \psi_1(v) \\ \lambda_2^t \psi_2(v) \\ \vdots \\ \lambda_{s(\delta,t)}^t \psi_{s(\delta,t)}(v) \end{pmatrix}.$$

Here $\lambda_i, \psi_i, i = 1, 2 \ldots n$ are the descending sorted eigenvalues and eigenfunctions of diffusion matrix P,

$$s(\delta, t) = \max\{l \in \mathbf{N} \quad \text{such that} \quad |\lambda_l|^t > \delta|\lambda_1|^t\}$$

Hence diffusion maps can optimally preserves the intrinsic geometric structures of the data. We first evaluate the intrinsic dimension n_t embedded in the data set P_n with maximum likelihood estimation, and then perform diffusion maps with the denoted dimension n_t to get the transform data P_{n_t}. In the later stage, we introduce the same framework of variance analysis in [33] to obtain the final C_{opt}. We perform hierarchical clustering on P_{n_t} for different number of given clusters s, and check the changes of $R = \sum_{j=1}^{m} r_j$ as s increases, and compute the most frequent number determined among many C_m's we obtained, where m denotes the number of pre-selected genes.

3. Results

It is widely accepted that distinct cell types express different sets of genes, but in scRNA-seq data, notable technical noise in single-cell transcriptomes make cell type identification non-trivial. We hence test *KDCorr-NMF* for capability of cell type identification using a number of scRNA-seq datasets in a variety of cell types [39–41].

3.1. Comparison methods

The cell clustering results are presented in comparison with three widely used single cell clustering algorithms: SNN-Cliq [30] that can reflect cell types of origins in a highly accurate manner, SIMLR [31]: a stable multi-kernel learning framework for cell type identification and SC3 [32]: a robust and accurate consensus clustering method.

- *SNN-Cliq*: SNN-Cliq [30] is a graph-partitioning method. One of the major contributions in SNN-Cliq is the new similarity(SNN) construction based on primary similarity matrix using Euclidean distance. Local maximal quasi-cliques associated with each node in the subgraph induced by the node are extracted. Final clusters are constructed through merging quasi-cliques and assigning nodes to unique clusters. SNN-Cliq is able to automatically determine the number of clusters in the data, however the method usually gives an over-estimation on the cluster number involved in the data set. The computational complexity O(n^2) constrained the flexibility of the method in data sets with thousands of cells.
- *SIMLR*: SIMLR [31] mainly tries to learn cell-to-cell similarities S through the following optimization framework:

$$\min_{S,L,w} \sum_{i,j} D(v_i, v_j)S_{ij} + \beta||S||_F^2 + \gamma \text{tr}(L^T(I_n - S)L) + \rho \sum_l w_l \text{log} w_l$$

$$\text{s.t.} D(v_i, v_j) = \sum_l w_l D_l(v_i, v_j), \sum_l w_l = 1, w_l \geq 0$$

$$L^T L = I_c, \sum_j S_{ij} = 1, S_{ij} \geq 0, i, j \in 1, 2, \ldots, n$$

where v_i is the gene expression vector for cell i, $D(v_i, v_j)$ the linear combination of $D_l(v_i, v_j)$ is the distance between v_i and v_j, $||S||_F$ is the Frobenius norm of S, L is a low dimensional matrix enforcing low rank constraint on S. The general optimization problem is non-convex, and an alternating convex optimization method can be introduced to solve the tri-convex problem. The optimization process generates a c dimensional latent space to represent the high dimensional sc-RNAseq data, K-means clustering is then applied to obtain the final cluster assignment for each cell.
- *SC3*: SC3 [32] is a consensus clustering method which takes 5 elementary steps.

- Gene Filter.
  This step mainly removes rare genes that are not informative for clustering.
- Distance Calculations.
  This step calculates distances between different cells using Euclidean, Pearson and Spearman metrics where

$$\begin{cases} d_E(v_i, v_j) = ||v_i - v_j||_2 \\ d_P(v_i, v_j) = \frac{\sum_{k=1}^{p}(v_{ik}-\overline{v_i})(v_{jk}-\overline{v_j})}{\sqrt{\sum_{k=1}^{p}(v_{ik}-\overline{v_i})^2}\sqrt{\sum_{k=1}^{p}(v_{jk}-\overline{v_j})^2}} \\ d_S(v_i, v_j) = d_P(u_i, u_j) \end{cases}$$

  u_i and u_j are the transformed rank data for v_i and v_j respectively.
- Transformations.
  The distance matrices are further transformed by principal component analysis or graph Laplacian method.
- K-means clustering

  K-means clustering is conducted on the first d eigenvector of the transformed distance matrices $D_d = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}$. It is realized through iteratively solving optimization problem

$$\min_{\substack{c^{(1)}, c^{(2)}, \ldots, c^{(m)} \\ \mu_1, \mu_2, \ldots, \mu_k}} J(c^{(1)}, c^{(2)}, \ldots, c^{(m)}, \mu_1, \mu_2, \ldots, \mu_k)$$

$$= \min_{\substack{c^{(1)}, c^{(2)}, \ldots, c^{(m)} \\ \mu_1, \mu_2, \ldots, \mu_k}} \frac{1}{n} \sum_{i=1}^{n} ||Y_i - \mu_{c^{(i)}}||^2$$

  where $c^{(i)}$ is the index of clusters $(1, 2, \ldots, k)$ to which Y_i is currently assigned, μ_k is the cluster centroid k, $\mu_{c^{(i)}}$ is the cluster centroid of cluster to which Y_i has been assigned.
- Consensus Clustering
  A consensus matrix is constructed based on cluster-based similarity partitioning algorithm and the resulting consensus matrix is finally clustered with hierarchical clustering algorithm.

3.2. Evaluation metrics

The clustering results are evaluated with two broadly used measures: adjusted rand index (ARI) and normalized mutual information (NMI). The definitions of ARI and NMI are given below.

- *ARI* is a measure for the agreement between two partitions with different number of clusters. Suppose T_n is the true label information for a specific sc-RNAseq data set, L_n is the estimated label information based on clustering. It can be expressed as

$$\text{ARI} = \frac{\text{RI} - E(\text{RI})}{\max(\text{RI}) - E(\text{RI})}$$

  where RI can be computed as

$$\text{RI} = \frac{A + B}{C_n^2}$$

  A represents the number of pairs of objects in the same cluster in T_n and the same cluster in L_n, and B is the number of pairs that are neither in the same cluster in T_n nor in the same cluster in L_n.
- *NMI* is often used in clustering to measure the similarity between two clustering results. It is important measurement indicator that can objectively evaluate the accuracy of comparing a community division with a standard division. We further denote K_T, K_L as the number of clusters in T_n, L_n respectively. Entropy of T_n, L_n is defined as follows.

$$H(T_n) = -\sum_{i=1}^{K_T} P(i)\log(P(i)), H(L_n) = -\sum_{i=1}^{K_L} Q(i)\log(Q(i))$$

  where $P(i) = |T_n^i|/n, Q(i) = |L_n^i|/n$ represent the probability of cells in cluster i. T_n^i, L_n^i represent the indices of cells in cluster i, $|T_n^i|$, $|L_n^i|$ are the corresponding number of elements for T_n^i, L_n^i.
  NMI can be calculated as

$$\text{NMI}(T_n, L_n) = \frac{\text{MI}(T_n, L_n)}{\sqrt{H(T_n)H(L_n)}}$$

  where

$$\text{MI}(T_n, L_n) = \sum_{i=1}^{K_T}\sum_{j=1}^{K_L} P(i, j)\log(\frac{P(i, j)}{P(i)Q(j)}), P(i, j) = |T_n^i \bigcap L_n^j|/n$$

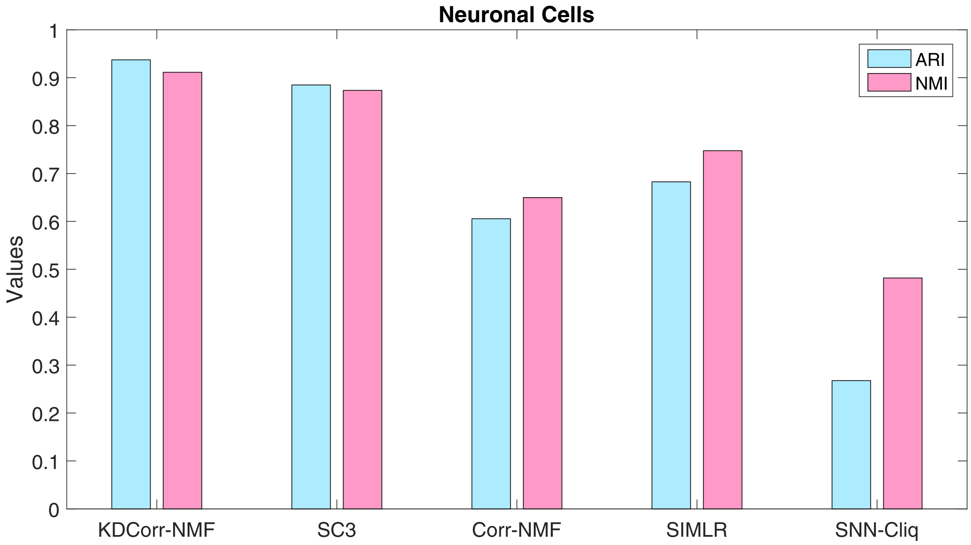

Fig. 1. Evaluation of clustering accuracy by ARI and NMI in neuronal cells.

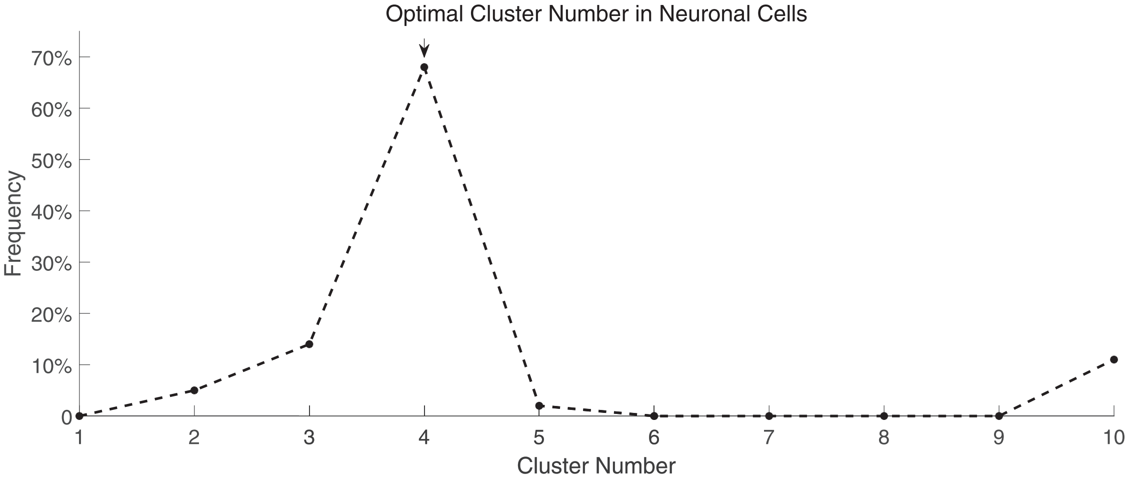

Fig. 2. Cluster number determination in neuronal cells by KDCorr-NMF.

3.3. *Performance on neuronal cells [39]*

This data set contains 622 neuronal cells with sensory subtypes: peptidergic nociceptors, nonpeptidergic nociceptors, neurofilament containing and tyrosine hydroxylase containing. It was obtained from mouse dorsal root ganglion, generated using Illumina Genome Analyzer IIx. The KNMF parameter k is set to be 16. For the hierarchical clustering with low dimensional representation generated by KNMF, the consensus parameter is set to be 20. The consensus$^+$ parameter is set to be 10. As shown in Fig. 1, the ARI value by **KDCorr-NMF** is 0.9373, NMI value is 0.9113. In comparison, the robust and accurate SC3 method achieves ARI value 0.8849, NMI value 0.8736, ranking the second best. SIMLR by similarity learning shows 0.6820 in ARI value, and 0.7476 in NMI value. SNN-Cliq finds 110 qausi-cliques, merged into 29 clusters, the final ARI value is 0.2677, with NMI value 0.4819.

It should be noted that we need to specify the number of clusters in advance in both SC3 and SIMLR, we all use the correct cluster number 4 as input for the two methods. While in *KDCorr-NMF*, we estimated the number of clusters using diffusion maps and variance analysis. We calculated the intrinsic dimension for neuronal cell data set to be 98, and conducted further variance analysis. Fig. 2 illustrates the determination of optimal cluster numbers for *KDCorr-NMF*, the final optimal cluster number is determined as the most frequently appeared one. We can see that diffusion maps with variance analysis helps determine the right cluster number, however, in SNN-Cliq, the cluster number estimated as 29 tends to be much larger than the real one.

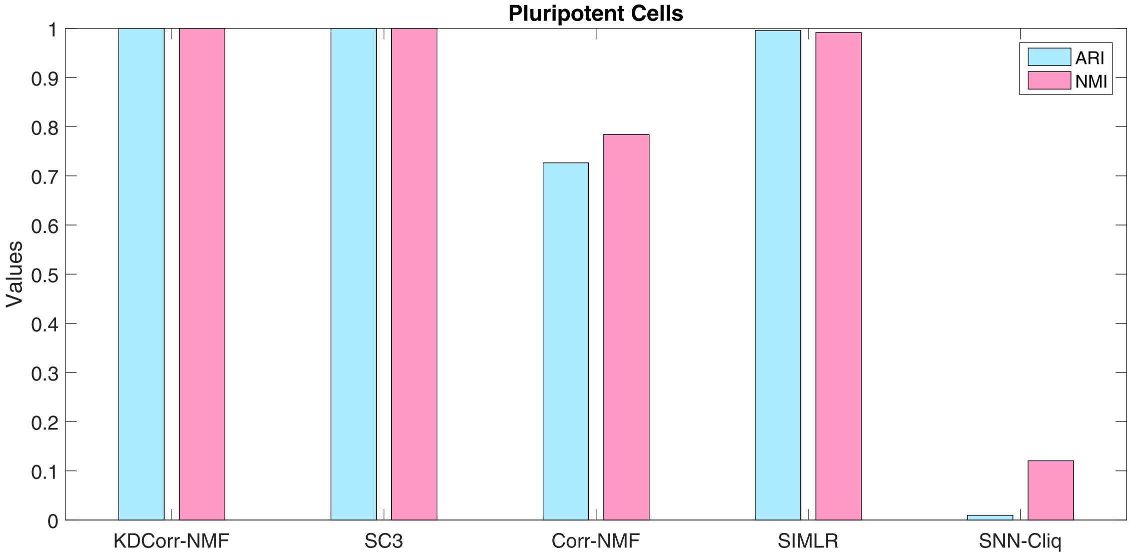

Fig. 3. Evaluation of clustering accuracy by ARI and NMI in pluripotent cells.

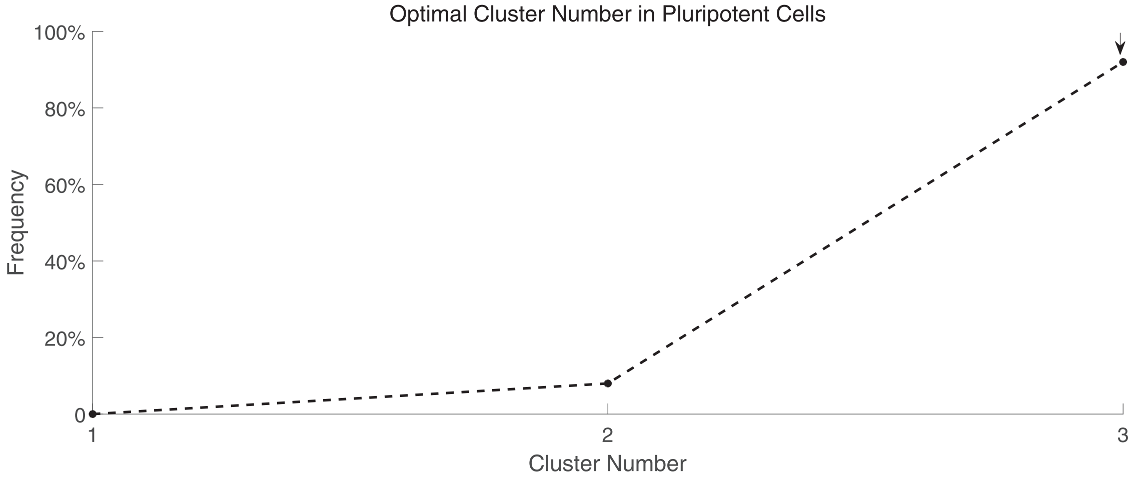

Fig. 4. Cluster number determination in pluripotent cells by KDCorr-NMF.

3.4. Performance on pluripotent cells [40]

This data set contains 704 mESCs involving 3 different culture conditions. It was obtained from a stem cell study considering the influence of culture conditions on pluripotent states of mESCs. The data set is available at ArrayExpress database (http://www.ebi.ac.uk/arrayexpress) under accession number E-MTAB-2600. The KNMF parameter k is set to be 9. For the hierarchical clustering with low dimensional representation generated by KNMF, the consensus parameter is set to be 20. The consensus$^+$ parameter is set to be 10. The ARI value by *KDCorr-NMF* is 1, NMI value is 1, showing in Fig. 3. In comparison, the robust and accurate SC3 method achieves ARI value 1, NMI value 1. Both *KDCorr-NMF* and SC3 show 100% in accuracy, showing that they can accurately distinguish mESCs in different culture conditions. SIMLR by similarity learning shows 0.9960 in ARI value, and 0.9915 in NMI value, which is almost 100% correct. SNN-Cliq finds 29 qausi-cliques, merged into 3 clusters, the final ARI value is 0.0096, with NMI value 0.1205. From the results we know that SNN-Cliq as a graph clustering algorithm may not be quite appropriate for this data set.

We calculate the intrinsic dimension estimated by maximum likelihood estimation for pluripotent cell data set as 64, and conduct variance analysis in a further step to estimate the optimal cluster number. Fig. 4 illustrates the determination of optimal cluster numbers for *KDCorr-NMF*, the final optimal cluster number is determined as the most frequently appeared one. It is very clear to show that the optimal cluster number is 3, occupying over 80% in the suggested optimal numbers. We can see that diffusion maps with variance analysis helps determine the right cluster number, It is interesting to see that

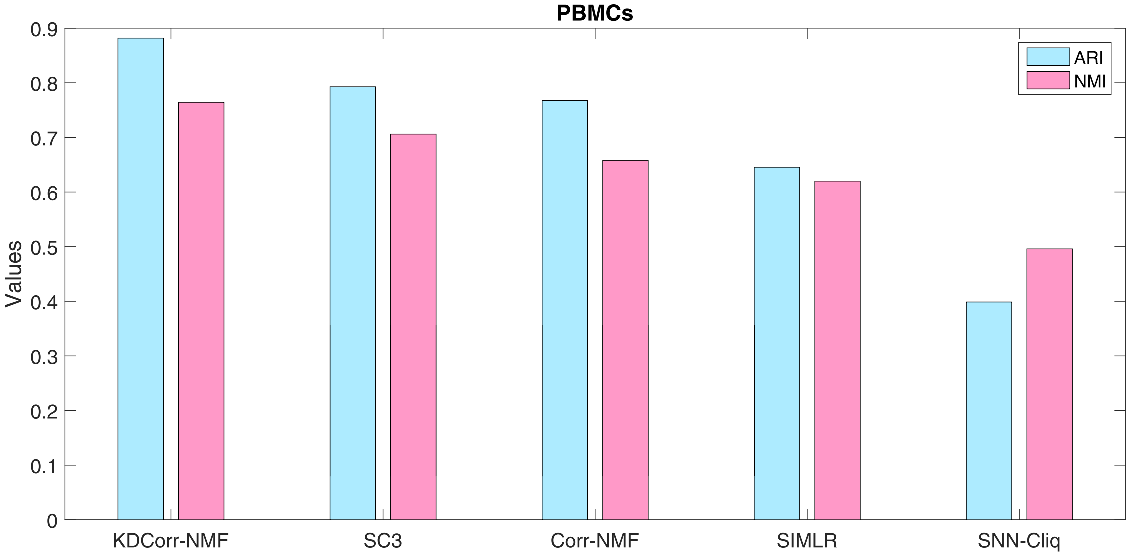

Fig. 5. Evaluation of clustering accuracy by ARI and NMI in PBMCs.

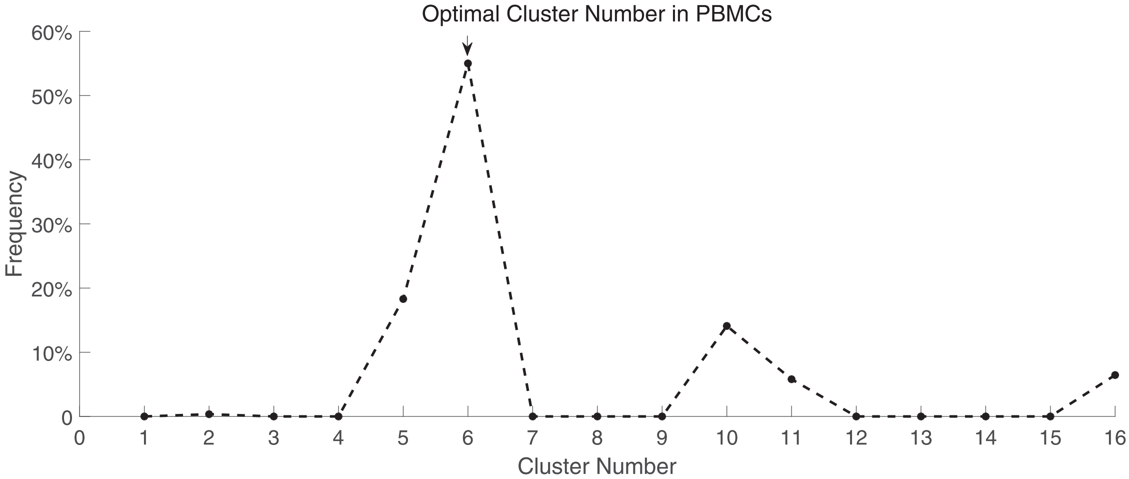

Fig. 6. Cluster number determination in PBMCs by KDCorr-NMF.

in SNN-Cliq, the cluster number estimated as 3 is also correct. However, the clustering performance in SNN-Cliq is quite unsatisfactory.

3.5. Performance on PBMCs [41]

We adopt an immune cell data set (https://www.nature.com/articles/nmeth.4179) using Seq-Well technology, that is more prone to drop out and noise errors. In the dataset, there are 4296 cells with 602 cells likely representing single-cell libraries of low complexity removed. In total, there are 3694 cells, containing 6 different cell types: B cell, NK cell, monocyte cell, CD4+ cell, CD8+ cell, Dendritic cell. The KNMF parameter k is set to be 36. For the hierarchical clustering with low dimensional representation generated by KNMF, the consensus parameter is set to be 20. The consensus$^+$ parameter is set to be 10. The ARI value by *KDCorr-NMF* is 0.8818, NMI value is 0.7643, showing in Fig. 5. In comparison, the robust and accurate SC3 method achieves ARI value 0.7927, NMI value 0.7061, inferior to *KDCorr-NMF*. SIMLR by similarity learning shows 0.6454 in ARI value, and 0.6200 in NMI value. SNN-Cliq finds 502 qausi-cliques, merged into 56 clusters, the final ARI value is 0.3987, with NMI value 0.4959. From the results we know that SNN-Cliq as a graph clustering algorithm may not be quite appropriate for this data set.

We calculate the intrinsic dimension estimated by maximum likelihood estimation for PBMC data set as 86, and conduct variance analysis in a further step to estimate the optimal cluster number. Fig. 6 illustrates the determination of optimal cluster numbers for *KDCorr-NMF*, the final optimal cluster number is determined as the most frequently appeared one. It is very clear to show that the optimal cluster number is 6, dominates the whole suggested cluster numbers. We can see that diffusion maps with variance analysis helps determine the right cluster number, In comparison in SNN-Cliq, the cluster number estimated as 56, providing an unreliable estimation on the cluster number. The clustering performance in SNN-Cliq is unsatisfactory as well.

4. Discussions

In the results part, we have included a comparison partner ‘Corr-NMF’ apart from SC3, SIMLR and SNN-Cliq. ‘Corr-NMF’ is different from *KDCorr-NMF* in the kernel construction part. Since we know that *KDCorr-NMF* as a kernel is induced by ‘DCorr’ where ‘DCorr’ itself is a valid kernel as well, therefore we also want to test if ‘Corr-NMF’ can be suitable for dealing with scRNA-seq data. However, it should be noted that ‘DCorr’ cannot always guarantee non-negativeness for all the entries. Based on the computational results, we can have the following discoveries.

- *KDCorr-NMF* is the best among all the comparison partners, achieving the best in both ARI and NMI values in all the considered data sets. In pluripotent cell data set, the data is relatively clean, *KDCorr-NMF* and SC3 show 100% in both ARI and NMI values, perfectly distinguish cells in different culture conditions, SIMLR also achieves 0.9980 in ARI value, ‘Corr-NMF’ yields 0.7265 in ARI value. SNN-Cliq only gets 0.0096 in ARI value, near 0. In neuronal cell data set with mediate noise, *KDCorr-NMF* can still show its power in cell type identification. SC3 ranked the second best and SIMLR is the third, achieving 0.6820 in ARI value. SNN-Cliq only shows 0.2677 in ARI value. In PBMCs data set with strong noise, *KDCorr-NMF* can show 0.8818 in ARI value, SC3 gets 0.7927 accordingly. ‘Corr-NMF’ is the third for this data set, superior to SIMLR who gets 0.6454 in ARI value. SNN-Cliq only shows 0.3987 in ARI value. ‘Corr-NMF’ is stable in a overall manner(see Supplementary file Table S1), but cannot compete with *KDCorr-NMF*.
- SNN-Cliq seems inappropriate for dealing with large scale scRNA-seq data sets. In neuronal cell data set and PBMCs data set, SNN-Cliq cannot give a relatively reasonable estimation on the cluster number, much larger than the truth. The evaluation measures are also not satisfactory. In particular for pluripotent cell data set, although SNN-Cliq provides accurate estimate on the cluster number within the data set, the ARI value and NMI value is almost 0.
- SC3 can demonstrate its ability in cell type identification. Although it cannot compete with *KDCorr-NMF*, it is the most robust and efficient method among the other competitors. There is no dominant method in SIMLR and ‘Corr-NMF’. In neuronal cell data set and Pluripotent data set, SIMLR is better than ‘Corr-NMF’ and in PMBCs data set ‘Corr-NMF’ is better than SIMLR.

KDCorr as a kernel may provide a relatively clear representation on the cellular relationships. Hence, we check if tSNE with the similarity matrix *KDCorr* may provide a proper low dimensional visualization on the highly noisy and sparse scRNA-seq data. Fig. 7 shows that tSNE with *KDCorr* can help preserve local structures better than tSNE with SIMLR. There are 6 type of cells involved, and cells with type 0,3,4,5 are clearly distinguished, leaving cells in type 1 and 2 blended in *KDCorr-NMF*. In SIMLR, we can see that cells in type 3 and 5 are perfectly isolated from other cells. But cells in type 0 are divided into two subclusters, and cells with type 1,3,4 are blended together. Similarly in the other two datasets (Please see supplementary file Figs. S1–S3 for more information), we can also see that *KDCorr* can provide clear representation on the cellular relationships.

Regarding the selection of k, we predetermine a feasible range of k values related to the number of determined optimal cluster number, k^2_{opt}. We test over a range of k values around k^2_{opt} to check if *KDCorr-NMF* is robust or not. We find that the consensus$^+$ framework helps to guarantee a robust and stable performance in clustering. (See Supplementary file Tables S1 and S2.) As shown in Fig. 8, in neuronal cells data, the performance of *KDCorr-NMF* is very stable when $k \in (16, 20)$. When k is chosen in a relatively small number, the performance is relatively less stable, but stable in a overall manner. And for pluripotent cells data, the performance is also very stable for $k \in (5, 13)$. For PBMC data, when $k \in (31, 40)$, the performance of *KDCorr-NMF* over multiple runs tends to be stable.

Algorithm 1 Framework of our algorithm ‘*KDCorr-NMF*’ based on kernel non-negative matrix factorization.

```
Input: The set of single cells, $P_n \in R^{n \times p}$;
Output: Cell types for the set of single cells, $L_n \in Z_+^{n \times 1}$;
  Evaluate continuous version of differentiability correlation for $i, j \in \{1, 2, \ldots, n\}$;
  Construct KDCorr kernel by Eq. (2);
  Determine Cluster Number $C_{\mathrm{opt}}$ based on Diffusion Maps and variance analysis;
  Consensus$^+$ Clustering with the above similarity measure and cluster number.
    • for certain $k$,perform KNMF on $KDCorr : K = W_{\phi_k}^K H_{\phi_k}$.
    • Consensus clustering on $H_{\phi_k}$ to get $L_{n_k} \in Z_+^{n \times 1}$
    • Consensus clustering on $L_{n_k} \in Z^{n \times 1}$ based on CSPA.
  Return: $L_n$;
```

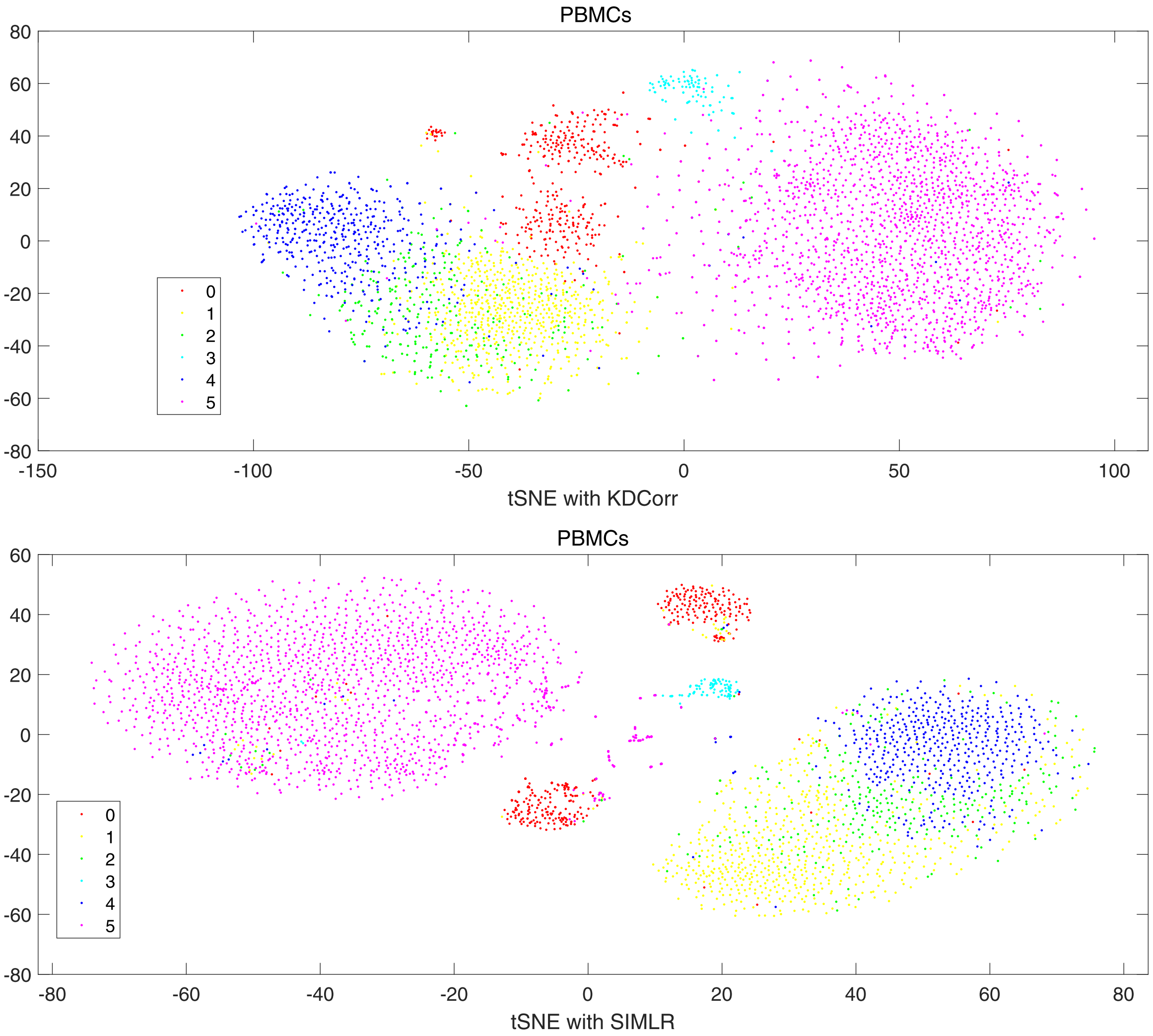

Fig. 7. tSNE in PBMCs by KDCorr-NMF and SIMLR.

Table 1
Performance of *KDCorr-NMF* framework for various datasets.

Neuronal	*k*	*k*=13	*k*=14	*k*=15	*k*=16	*k*=17	*k*=18	*k*=19	*k*=20	*k*=21	*k*=22
	ARI	0.9190	0.9272	0.9345	0.9459	0.9373	0.9373	0.9373	0.9373	0.9373	0.9373
Pluripotent	*k*	*k*=5	*k*=6	*k*=7	*k*=8	*k*=9	*k*=10	*k*=11	*k*=12	*k*=13	*k*=14
	ARI	1	1	1	1	1	1	1	1	1	1
PBMC	*k*	*k*=31	*k*=32	*k*=33	*k*=34	*k*=35	*k*=36	*k*=37	*k*=38	*k*=39	*k*=40
	ARI	0.8814	0.8783	0.8823	0.8893	0.8731	0.8818	0.8766	0.8798	0.8782	0.8799

Even when KDCorr-NMF is less stable for some certain *k*, we still can help identifying a near optimal clustering result through consensus$^+$ step. Table 1 shows the performance of *KDCorr-NMF* framework for various datasets with different *k* values. It is clear that consensus$^+$ helps guarantee a relatively stable and satisfactory result.

Apart from comparisons with state-of-the-art algorithms, we also test the robustness of our kernel non-negative framework in 3 different perspectives:

- Comparison in Kernel Construction step;
- Comparison with extension of non-negative matrix factorization taking into consideration on the geometric structure: Graph regularized nonnegative matrix factorization
- Comparison with symmetric non-negative matrix factorization with KDCorr kernel.

The detailed comparison results are attached in Supplementary files, Tables S2–S4 and Figs S4–S6. All results show that our kernel non-negative framework with KDCorr kernel turns out to be the best.

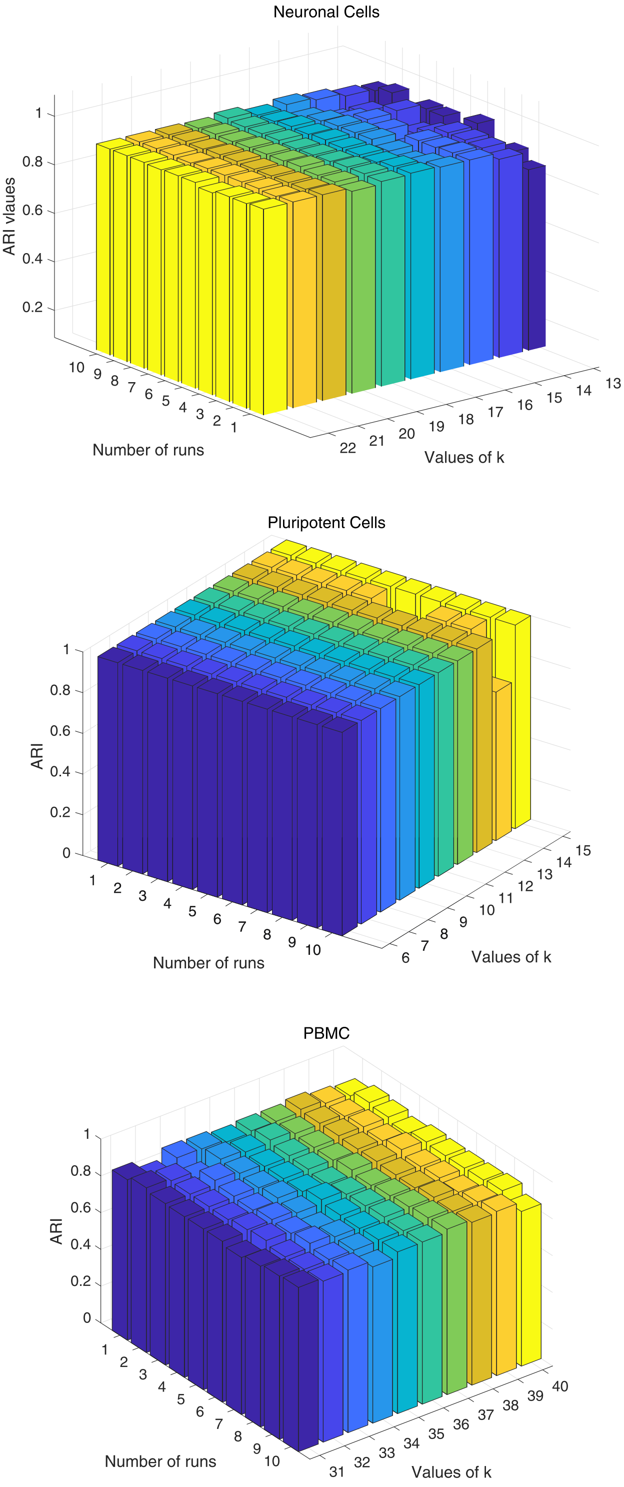

Fig. 8. Stability of *KDCorr-NMF*.

Through comparison with various state-of-the-art algorithms, we can summarize our findings here.

1. Our kernel non-negative matrix factorization framework proves to be effective in single cell clustering problems.
2. The kernel is a key component in kernel non-negative matrix factorization framework. We found that KDCorr kernel is superior to Dcorr kernel, the most widely used kernel: RBF kernel etc.
3. The most stable and robust clustering method to date: SC3, cannot compete with our method as well.
4. The consensus cluster-based similarity partitioning algorithm embedded in our kernel non-negative factorization framework helps to guarantee a robust clustering result.
5. The computational complexity for our method is $\mathrm{O}(n^2k^3 + 2nk^2)$, where n is the number of cells, and k is the reduced number of dimension after kernel non-negative factorization framework. It can be seen that when the number of cells becomes large, say greater than the number of attributes p, then the computational complexity is larger than that in non-negative factorization framework in the original single cell data matrix. And it would become computationally expensive in personal computers when n becomes hundreds of thousands. Therefore, our future endeavor will be devoted to developing proper methods for relatively large scale single-cell clustering problems (which satisfies $n < p$).

5. Conclusions

In this paper, we proposed kernel non-negative matrix factorization framework for single cell clustering. The KDCorr kernel provides a proper evaluation on cell similarity based on cell-pair differentiability correlation for nonlinear and noisy scRNA-seq data. And the consensus$^+$ framework provides a stable way for scRNA-seq data analysis. Experiments under different evaluation metrics support the effectiveness and feasibility of our proposed method. Besides, the cluster number determination method based on diffusion maps and variance analysis is a robust approach for estimating the intrinsic clusters embedded in the high-dimensional and noisy data.

Implementation and availability: The source code (Matlab) is available upon request.

Acknowledgments

This research is supported by National Natural Science Foundation of China Grant No. 11901575, 91730301 and 11675060.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.apm.2020.08.065.

References

[1] Z. Wang, M. Gerstein, M. Snyder, Rna-seq: a revolutionary tool for transcriptomics, Nat. Rev. Genet. 10 (2009) 57–63.

[2] V.L. Wei, J.J. Li, An accurate and robust imputation method scImpute for single-cell RNA-seq data, Nat. Commun. 9 (1) (2018) 997.

[3] D. Grün, A. Lyubimova, L. Kester, K. Wiebrands, O. Basak, N. Sasaki, H. Clevers, A. van Oudenaarden, Single-cell messenger RNA sequencing reveals rare intestinal cell types, Nature 525 (2015) 251–255.

[4] L. Jiang, H. Chen, L. Pinello, Y. Chen, GiniClust: detecting rare cell types from single-cell gene expression data with Gini index, Genome Biol. 17 (2016) 144.

[5] D. Tsoucas, G. Yuan, Giniclust2: a cluster-aware, weighted ensemble clustering method for cell-type detection., Genome Biol. 19 (2018) 58.

[6] S.C. Bendall, K.L. Davis, E. ad David Amir, M.D. Tadmor, E.F. Simonds, T.J. Chen, D.K. Shenfeld, G.P. Nolan, D. Pe'er, Single-cell trajectory detection uncovers progression and regulatory coordination in human B cell development., Cell 157 (3) (2014) 714–725.

[7] L. Yan, M. Yang, H. Guo, Y. Yang, J. Wu, R. Li, P. Liu, Y. Lian, X. Zheng, J. Yan, J. Huang, M. Li, X. Wu, L. Wen, K. Lao, R. Li, J. Qiao, F. Tang, Single-cell RNA-Seq profiling of human preimplantation 115 embryos and embryonic stem cells., Nat. Struct. Mol. Biol. 20 (2013) 1131–1139.

[8] L.F. Chu, N. Leng, J. Zhang, Z. Hou, D. Mamott, D.T. Vereide, J. Choi, C. Kendziorski, R. Stewart, J.A. Thomson, Single-cell RNA-seq reveals novel regulators of human embryonic stem cell differentiation to definitive endoderm, Genome Biol. 17 (1) (2016) 173.

[9] J.A. Farrell, Y. Wang, S.J. Riesenfeld, K. Shekhar, A. Regev, A.F. Schier, Single-cell reconstruction of developmental trajectories during zebrafish embryogenesis., Science 360 (2018) eaar3131.

[10] Y. Gao, X. Ni, H. Guo, Z. Su, N. Zhang, Single-cell sequencing deciphers a convergent evolution of copy number alterations from primary to circulating tumor cells, Genome Res. 27 (8) (2017) 1312.

[11] S.W. Brady, J.A. Mcquerry, Y. Qiao, S.R. Piccolo, G. Shrestha, D.F. Jenkins, R.M. Layer, B.S. Pedersen, R.H. Miller, A. Esch, Combating subclonal evolution of resistant cancer phenotypes, Nat. Commun. 8 (1) (2017) 1231.

[12] C. Kim, R. Gao, E. Sei, R. Brandt, J. Hartman, T. Hatschek, N. Crosetto, T. Foukakis, N.E. Navin, Chemoresistance evolution in triple-negative breast cancer delineated by single-cell sequencing, Cell (2018). S0092867418303659.

[13] K.M. Mann, J.Y. Newberg, M.A. Black, D.J. Jones, F. Amaya-Manzanares, L. Guzman-Rojas, T. Kodama, J.M. Ward, A.G. Rust, L. van der Weyden, Analyzing tumor heterogeneity and driver genes in single myeloid leukemia cells with SBCapSeq, Nat. Biotechnol. 34 (9) (2016) 962–972.

[14] S.F. Roerink, N. Sasaki, H. Lee-Six, M.D. Young, L.B. Alexandrov, S. Behjati, T.J. Mitchell, S. Grossmann, H. Lightfoot, D.A. Egan, Intra-tumour diversification in colorectal cancer at the single-cell level, Nature 556 (7702) (2018) 102.

[15] P.V. Kharchenko, L. Silberstein, D.T. Scadden, Bayesian approach to single-cell differential expression analysis, Nat. Methods 11 (7) (2014) 740.

[16] S. Prabhakaran, E. Azizi, A. Carr, D. Pe'Er, Dirichlet process mixture model for correcting technical variation in single-cell gene expression data, in: Proceedings of the International Conference on International Conference on Machine Learning, 2016.

[17] L. Zhu, L. Jing, B. Devlin, K. Roeder, A unified statistical framework for single cell and bulk rna sequencing data, Ann. Appl. Stat. 12 (1) (2017) 609–632.

[18] M. Huang, W. Jingshu, T. Eduardo, D. Hannah, S. Sydney, B. Roberto, M.J. I., R. Arjun, L. Mingyao, Z.N. R., Saver: gene expression recovery for single-cell rna sequencing, Nat. Methods 15 (2018) 539–542.

[19] W. Gong, I.Y. Kwak, P. Pota, N. Koyano-Nakagawa, D.J. Garry, Drimpute: imputing dropout events in single cell rna sequencing data, BMC Bioinform. 19 (1) (2018) 220.

[20] D.V. Dijk, R. Sharma, J. Nainys, K. Yim, P. Kathail, A.J. Carr, C. Burdziak, K.R. Moon, C.L. Chaffer, D. Pattabiraman, Recovering gene interactions from single-cell data using data diffusion, Cell 174 (3) (2018). S0092867418307244.

[21] R. Bacher, C. Kendziorski, Design and computational analysis of single-cell rna-sequencing experiments, Genome Biol. 17 (1) (2016) 63.
[22] P. Brennecke, S. Anders, J.K. Kim, A.A. Kolodziejczyk, X. Zhang, V. Proserpio, B. Baying, V. Benes, S.A. Teichmann, J.C. Marioni, Accounting for technical noise in single-cell rna-seq experiments, Nat. Methods 10 (11) (2013) 1093–1095.
[23] C.A. Vallejos, J.C. Marioni, S. Richardson, Basics: Bayesian analysis of single-cell sequencing data, PLoS Comput. Biol. 11 (6) (2015) e1004333.
[24] J.K. Kim, A.A. Kolodziejczyk, T. Ilicic, S.A. Teichmann, J.C. Marioni, Characterizing noise structure in single-cell rna-seq distinguishes genuine from technical stochastic allelic expression, Nat. Commun. 6 (2015) 8687.
[25] T.S. Andrews, M. Hemberg, M3Drop: Dropout-based feature selection for scRNASeq, Bioinformatics 35 (16) (2019) 2865–2867.
[26] E. Pierson, C. Yau, Zifa: Dimensionality reduction for zero-inflated single-cell gene expression analysis, Genome Biol. 16 (1) (2015) 241.
[27] E. Macosko, A. Basu, R. Satija, J. Nemesh, K. Shekhar, M. Goldman, I. Tirosh, A. Bialas, N. Kamitaki, E. Martersteck, Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets, Cell 161 (5) (2015) 1202–1214.
[28] M. Baron, A. Veres, S. Wolock, A. Faust, R. Gaujoux, A. Vetere, J.H. Ryu, B. Wagner, S. Shen-Orr, A. Klein, A single-cell transcriptomic map of the human and mouse pancreas reveals inter- and intra-cell population structure, Cell Syst. 3 (4) (2016) 346–360.e4.
[29] L. Haghverdi, M. Büttner, F.A. Wolf, F. Buettner, F.J. Theis, Diffusion pseudotime robustly reconstructs lineage branching, Nat. Methods 13 (10) (2016) 845.
[30] X. Chen, Z.C. Su, Identification of cell types from single-cell transcriptomes using a novel clustering method, Bioinformatics 31 (12) (2015) 1974–1980.
[31] B. Wang, J. Zhu, E. Pierson, D. Ramazzotti, S. Batzoglou, Visualization and analysis of single-cell RNA-Seq data by kernel-based similarity learning, Nat. Methods 14 (4) (2017) 414.
[32] V.Y. Kiselev, K. Kirschner, M.T. Schaub, T. Andrews, A. Yiu, T. Chandra, K.N. Natarajan, W. Reik, M. Barahona, A.R. Green, SC3: consensus clustering of single-cell RNA-Seq data, Nat. Methods 14 (5) (2017) 483–486.
[33] H. Jiang, L.L. Sohn, H.Y. Huang, L.N. Chen, Single cell clustering based on cell-pair differentiability correlation and variance analysis, Bioinformatics 34 (21) (2018) 3684.
[34] D.D. Lee, H.S. Seung, Learning the parts of objects by non-negative matrix factorization, Nature 401 (6755) (1999) 788–791.
[35] D. Zhang, Z.H. Zhou, S. Chen, (2006) Non-negative Matrix Factorization on Kernels. In: Yang Q., Webb G. (eds) PRICAI 2006: Trends in Artificial Intelligence. PRICAI 2006. Lecture Notes in Computer Science, vol 4099. Springer, Berlin, Heidelberg.
[36] B. Pan, J. Lai, W.S. Chen, Nonlinear nonnegative matrix factorization based on mercer kernel construction, Pattern Recognit. 44 (10) (2011) 2800–2810.
[37] J. Shawetaylor, N. Cristianini, Kernel Methods for Pattern Analysis: Pattern analysis, Cambridge University Press, 2004.
[38] R.R. Coifman, S. Lafon, Diffusion maps, Appl. Comput. Harmon. Anal. 21(1) 5–30.
[39] U. Dmitry, F. Alessandro, I. Saiful, A. Hind, L. Peter, L. Daohua, H.L. Jens, H. Jesper, K. Olga, P.V. Kharchenko, Unbiased classification of sensory neuron types by large-scale single-cell RNA sequencing, Nat. Neurosci. 18 (1) (2015) 145–153.
[40] A. Kolodziejczyk, J.K. Kim, J.H. Tsang, T. Ilicic, J. Henriksson, K. Natarajan, A. Tuck, X. Gao, M. Bühler, P. Liu, Single cell RNA-sequencing of pluripotent states unlocks modular transcriptional variation, Cell Stem Cell 17 (4) (2015) 471–485.
[41] T.M. Gierahn, M.H. Wadsworth, T.K. Hughes, B.D. Bryson, A. Butler, R. Satija, S. Fortune, J.C. Love, A.K. Shalek, Seq-well: portable, low-cost RNA sequencing of single cells at high throughput, Nat. Methods 14 (4) (2017) 395–398.