Hybrid Clustering of single-cell gene-expression and cell spatial information via integrated NMF and k-means

Sooyoun Oh[1], Haesun Park[1,*], and Xiuwei Zhang[1,*]

[1] Georgia Institute of Technology, Atlanta GA 30332, USA
[*] Corresponding authors: hpark@cc.gatech.edu, xiuwei.zhang@gatech.edu

Abstract. Recent advances in single cell transcriptomics have allowed us to examine the identify of each single cell, thus have led to discovery of new cell types and provide a high resolution map of cell type composition in tissues. Technologies which can measure another type of data of a single cell in addition to the gene-expression data provide a more comprehensive picture of a cell, and meanwhile pose challenges for data integration tasks. We consider the spatial location of cells, which is an important feature of cells, combined with the cells' gene-expression profiles, to determine the cell type identity. We aim to jointly classify cells based on their locations relative to other cells in the system as well as their gene expression profiles. We have developed scHybridNMF (single-cell Hybrid Nonnegative Matrix Factorization), which performs cell type identification by incorporating single cell gene expression data with cell location data. We combined two classical methods, nonnegative matrix factorization with a k-means clustering scheme, to respectively represent high-dimensional gene expression data and low-dimensional location data together. Our method incorporates a novel cell location term to the gene expression clustering. We show that scHybridNMF can make use of the location data to improve cell type clustering. In particular, we show that under multiple scenarios, including that when the number of genes profiled is low, and when the location data is noisy, scHybridNMF outperforms the standalone algorithms NMF and k-means, and an existing method HMRF which also uses cell location and gene-expression data for cell type identification.

Keywords: Single-cell analysis · Multimodal clustering · Matrix low-rank approximations.

1 Introduction

Advances in single cell RNA-Sequencing (scRNA-Seq) technology have provided an unprecedented opportunity for researchers to study the identity and underlying mechanisms of single cells [16]. While scRNA-Seq data is a major type of data used to study single cells, it cannot fully determine the identity of a cell, which informs its cell type [15]. As such, it is important to consider other modalities such as chromatin accessibility [2], protein abundance [17], and spatial locations [19, 21] of single cells.

With the availability of these types of data, we have entered the era of multi-modality single-cell omics, and effective computational methods are crucial in integrating multi-modal data to learn a comprehensive picture of inter- and intra-cell processes [6, 20]. Spatial location data can provide important information on the cells' micro-environment and allow researchers to study cell-cell interactions [14]. This is because cells at nearby locations tend to form the same cell type – daughter cells tend to keep the same cell type and similar location as their mother cell.

Considering both the gene-expression and location data can therefore lead to more accurate cell type identification. Technologies that measure the location and gene-expression of the same set of cells often have to comprise on the number of genes measured [24]. Clustering cells using smaller gene-expression profiles can be inaccurate, so the cell location data can be used to improve the accuracy. However, reconciling single cell gene-expression and location data for cell type identification is challenging because different data types can have differing scale and distributions and exhibit different types and levels of noise [6].

We introduce a matrix low-rank approximation scheme, scHybridNMF (single-cell Hybrid NMF), to perform cell clustering by jointly processing 2-dimensional cell location and gene expression data. Previously, Zhu *et al* developed a HMRF (Hidden Markov Random Field) model and showed that the spatial location of cells can contribute to cell type identification [24]. We, however, use a matrix low-rank approximation scheme because of the ease of preserving data characteristics through constraints and optimization terms. Crafting a loss-based minimization objective that bakes in these data characteristics maximally utilizes this information to jointly-cluster cells. We combined nonnegative matrix factorization with a k-means clustering scheme to respectively represent high-dimensional gene expression data and low-dimensional location data together.

Such joint-clustering methods have been used in other contexts such as document clustering [4]. Additionally, promising NMF models have been developed for cell type identification for data ranging from just scRNA-Seq data to encompassing multiple modalities [5, 9, 12, 18, 22]. However, none of these methods incorporate cell locations. We compare our scHybridNMF model with both the standalone NMF and the k-means methods, as well as the HMRF method which uses spatial location information. We show that scHybridNMF are particularly advantageous in two application scenarios: to use when the number of genes with gene-expression data is small, or and when the location data is noisy.

2 Methods

Matrix low-rank approximations assume that a matrix can be well-approximated as a product of more concise matrices. Many clustering frameworks are designed as matrix low-rank approximation schemes because they can easily incorporate prior biological knowledge and

data constraints. We formulate our multimodal clustering algorithm as a combined matrix optimization scheme. This formulation is designed to guide the gene expression-based clustering of cells using cell location clusterings. As part of our design, we incorporate nonnegative matrix factorizations (NMF) and k-means clustering.

2.1 Review of NMF and K-Means Clustering

We incorporate cell location data and gene expression data using the intuition behind NMF and k-means clustering, respectively. We chose these methods because they can easily be formulated as matrix low-rank approximations, and creating an objective that incorporates both of these methods is intuitive. Additionally, the individual characteristics of each method strongly match the characteristics of the data.

K-means clustering is an unsupervised learning algorithm that clusters data points by comparing pairwise distances usually determined by the Euclidean distance metric. This metric naturally pairs with location-based data, as it quantifies the similarity between points by how physically close they are. The matrix formulation for a Euclidean distance-based k-means objective is below:

$$\min_{\substack{H_L \in \{0,1\}^{k\times n} \\ H_L^T \mathbf{1}_k = \mathbf{1}_n}} \|L - W_L H_L\|_F^2 , \tag{1}$$

where $\mathbf{1}_k$ and $\mathbf{1}_n$ are k-length and n-length vectors of all ones, respectively. The columns of W_L contain the cluster centroids, and the columns of H_L contain membership information for each data point. Since each data point must belong to one cluster, each column is all zero except for a one on the row corresponding to the cluster the point belongs to. Additionally, k-means clustering does not require any pre-processing on location data. Pre-processing input data may remove much of the underlying characteristics of the location data. As such, k-means clustering is a good fit for our two-dimensional location data because using the Euclidean metric to build clusters naturally follows from the underlying data representation.

NMF is a dimension reduction algorithm that computes two nonnegative factors of a specified low rank, whose product is designed to best approximate the nonnegative input matrix. Below is a typical formulation for NMF. The columns of W_A contain the cluster representatives, and the columns of H_A contain the cluster membership information for each data point.

$$\min_{\{W_A, H_A\} \geq 0} \|A - W_A H_A\|_F^2 . \tag{2}$$

Per its design, NMF is superior for clustering high dimensional data. Unlike k-means clustering, NMF produces soft clusters, which means that a data point can be represented as a linear combination of cluster representatives. As such, we chose NMF because it is a good fit for our high dimensional gene-expression data.

2.2 Multimodal Objective

Let $A \in \mathbb{R}^{m\times n}$ denote the gene-expression matrix and $L \in \mathbb{R}^{2\times n}$ denote the two-dimensional cell location coordinates, where m is the number of genes and n is the number of cells. We

use the following objective function for multimodal clustering:

$$\min_{\substack{\{W_A,H_A\}\geq 0\\ H_L\in\{0,1\}^{k\times n}\\ H_L^T\mathbf{1}_k=\mathbf{1}_n}} g(W_A,H_A)=\|A-W_AH_A\|_F^2+\alpha\,\|H_A-H_A\circ H_L\|_F^2\,, \tag{3}$$

where $\circ$ represents the element-wise product between two matrices and k is the number of clusters. This preserves the hard-clustering characteristic of k-means clustering on H_L and also the NMF quality that H_A must be nonnegative. Since we are comparing two matrices H_A and H_L, and H_A from NMF is not unique, we assume that the columns of W_A are of unit norm by normalizing the columns of the computed W_A each time, and modifying H_A accordingly.

The first term in Eqn. (3) represents the NMF objective as in Eqn. (2). The second term combines NMF and k-means clustering results by making the clustering results from NMF and k-means inform each other. Instead of forcing H_A and H_L to be similar overall, the second term forces H_A and H_L to be similar in terms of cluster membership discovered, i.e., we want the location of the largest element in each column of H_A and the location of the 1 element in the corresponding column of H_L to match as much as possible.

The main focus of this work is to use cell location information to aid the clustering of cells by gene expression. Because we are specifically adapting our gene clusters to incorporate location cluster information, our design seeks to align the cluster membership matrices found in both k-means and NMF while still considering the accuracy of the gene expression clustering. Because our method incorporates the predetermined location-based clusters, it would not make sense to adapt the cell location clustering in the consensus. That is why the k-means objective is not in Eqn. (3), but the resulting clustering membership matrix H_L is in Eqn. (3).

2.3 Proposed Algorithm

We devise scHybridNMF to optimize Eqn. (3) using a consensus clustering on the clusters determined by NMF on A and k-means on L. The steps in scHybridNMF are outlined below.

We first compute initial W_L, H_L by optimizing Eqn. (1) on L and initial W_A, H_A by running sparse NMF on A [10]. We use sparse NMF because it enforces the sparsity in the cluster membership matrix H_A, which allows for a better comparison against the hard-clustered H_L in term 3 of Eqn. (3). The formulation for sparse NMF is as follows:

$$\min_{\{W_A,H_A\}\geq 0} \|A-W_AH_A\|_F^2+\eta\,\|W_A\|_F^2+\beta\sum_{j=1}^{n}\|H_A(:,j)\|_1^2\,, \tag{4}$$

The crux of our algorithm is in the block coordinate descent for computing H_A and W_A. These two terms are computed via an alternating nonnegative least squares (ANLS) formulation. We isolate the terms that involve H_A and W_A in Eqn. (3) to formulate the inputs into ANLS.

To solve for H_A, we only need to combine the first and second terms in Eqn. (3). Given that the second term involves H_A twice, we reformulate the second term as follows:

$$\|H_A-H_A\circ H_L\|_F^2=\|H_A\circ\mathbf{1}_{k\times n}-H_A\circ H_L\|_F^2=\|H_A\circ C\|_F^2\,, \tag{5}$$

where $C = \mathbf{1}_{k\times n} - H_L$ and $\mathbf{1}_{k\times n}$ the $k \times n$ matrix of all ones. We can represent an element-wise product in a block-ANLS formulation by computing the formulation column-by-column. Therefore, the new update rule for the first and second terms of Eqn. (3) is as follows:

$$H_A(:, G_i) \leftarrow \underset{H_A(:,G_i)\geq 0}{\arg\min} \left\| \begin{pmatrix} W_A \\ \sqrt{\beta} * \text{diag}(\mathbf{1}_k - e_i) \end{pmatrix} H_A(:, G_i) - \begin{pmatrix} A(:, G_i) \\ \mathbf{0}_k \end{pmatrix} \right\|_F^2, \tag{6}$$

where $i \in \{1, \ldots, k\}$, $\mathbf{1}_k$ is a k-length vector of all ones, $\mathbf{0}_k$ is a k-length vector of all zeros, and e_i is the vector of all zeros, save for a one in position i. Each column in H_A is element-wise multiplied to each column in C in Eqn. (5), and since there are k different forms of C's columns, we group columns of H_A that share the same vector form in C to more efficiently compute H_A. Each group G_i is determined by whichever entry in a given column of C is zero, or not a one, and the column that defines G_i is $\mathbf{1}_k - e_i$. As such, $\bigcup_i G_i = \{1, \ldots, n\}$, and the pairwise intersection between any two G_i is empty. We use k different groups of columns in H_A to calculate Eqn. (6) because there are only k different forms the columns of C can take.

To solve for W_A, we only need to transpose the first term in Eqn. (3):

$$W_A \leftarrow \underset{W_A\geq 0}{\arg\min} \left\| (H_A)^T W_A^T - A^T \right\|_F^2, \tag{7}$$

As such, the overall algorithm is described in Algorithm 1. We used the projected gradient, as used in SymNMF, to be the stopping criterion of scHybridNMF [13].

Algorithm 1: scHybridNMF: I Algorithm to minimize Eqn. (3)

Input : gene expression matrix $A \in \mathbb{R}^{m\times n}$, cell location matrix $L \in \mathbb{R}^{2\times n}$, number of clusters k.

Compute W_L, H_L using Eqn. (1);
Compute W_A, H_A using Eqn. (4);
$C = \mathbf{1}_{k\times n} - H_L$;
while $\left\|\nabla^P g(W_A, H_A)\right\|_F > tol * \left\|\nabla^P g\left(W_A^{(0)}, H_A^{(0)}\right)\right\|_F$ **do**
  for $i = 1, \ldots, k$ **do**
    Compute G_i as the set of columns in C such that $\forall j \in G_i, C(:, j) = \mathbf{1}_k - e_i$;
    Compute $H_A(:, G_i)$ using Eqn. (6);
  end
  Compute W_A using Eqn. (7);
end

2.4 Convergence of Algorithm

We use a block coordinate descent (BCD) framework to optimize our objective function for clustering multimodal data. BCD solves subgroups of problems for each variable of interest, which iteratively minimizes the total objective function. Our objective aims to iteratively improve W_A and H_A, which defines a two block coordinate descent framework. These comprise the minimization version of the two-block Gauss-Seidel method, which assigns $H^{(j)}$ and $W^{(j)}$ values that minimize a shared objective function, Eqn. (3), one-at-a-time.

An important theorem regarding general block Gauss-Seidel methods states that if a continuously differentiable function over a set of closed convex sets is minimized by block

coordinate descent, every solution that uniquely minimizes the function in block coordinate descent is a stationary point [1]. This theorem has the additional property that the uniqueness of the minimum is not necessary for a two-block Gauss-Seidel nonlinear minimization scheme [8]. This has been used to show that a two-block formulation for solving Eqn. (2) via alternating least squares guarantees convergence [11].

Given the constrained nonlinear minimization objective in Eqn. (3), we can rewrite the block coordinate descent as two ANLS formulations, which follow from Eqns. (6) and (7):

$$H_A(:,G_i)^{(j)} \leftarrow \underset{H_A(:,G_i)\geq 0}{\arg\min} \left\| \begin{pmatrix} W_A^{(j-1)} \\ \sqrt{\beta} * \mathrm{diag}(\mathbf{1}_k - e_i) \end{pmatrix} H_A(:,G_i) - \begin{pmatrix} A(:,G_i) \\ \mathbf{0}_k \end{pmatrix} \right\|_F^2 \tag{8a}$$

$$W_A^{(j)} \leftarrow \underset{W_A\geq 0}{\arg\min} \left\| \left(H_A^{(j)}\right)^T W_A^T - A^T \right\|_F^2, \tag{8b}$$

Eqns. (8a) and (8b) are executed consecutively to solve for H_A and W_A. We consider $H_A^{(j)}$ to be one block calculation because the calculations for each individual group G_i are independent of each other. In other words, calculating G_i does not depend on the values calculated for G_h for all $h \neq i$. We are then able to apply this theorem because Eqns. (8a) and (8b) constitute a valid minimization scheme equivalent to minimizing Eqn. (3). As such, we get the following property, which guarantees the convergence of our algorithm:

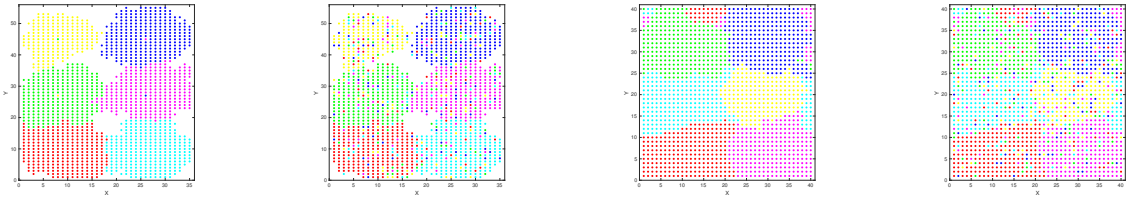
Theorem 1. *Every point* $\{W_A^{(j)}, H_A^{(j)}\}$ *calculated iteratively via Eqns. (8a) and (8b) is a stationary point of Eqn. (3).*

3 Experiments

We test the performance of scHybridNMF on both simulated and real data. We use SymSim [23] to simulate single cell gene-expression data where cells are from six cell types. Each dataset has 1600 cells and 600 genes. The number of genes is set to reflect the relatively low number of genes profiled in some spatially-resolved single cell gene-expression datasets.

We develop a new procedure to simulate the location data for the cells in a 2-d space such that cells belonging in the same cell type are closely located in the 2-d location space. This procedure mimics the cell division process in a tissue. First, in the 2-d space we choose a starting location for each cell type as the earliest cell for each cell type. Then for each cell type, a new cell is added in the following fashion: we randomly choose an existing cell of the same cell type as the parent cell of the new cell, and place the new cell next to the parent cell. If there is no available position next to the parent cell, then the new cell is located to a random empty position.

We consider different scenarios for the cell location data depending on how well the clusters are separated in the space. We denote clusters that are well separated as *w-separated*, and clusters that are not well separated as *n-separated*. For each of these scenarios, we generate location data with and without noise. In noisy data, cells from different cell types are mixed in the location space, and in data without noise, cells in the same cell type are all located together. We obtain the noisy location data from location data without noise by randomly choosing a percentage of cells and assigning them locations which are not in the main region of their original cell type. This is to more accurately emulate real-life data. Fig. 1 shows examples of these cases, where the case of n-separated with noise (as shown in Fig. 1d) is closest to real-life data.



(a) w-separated, no noise (b) w-separated, 20% noise (c) n-separated, no noise (d) n-separated, 20% noise

Fig. 1. An example of noise in location data. The data had $\sigma = 0.3$ and 20% noise in location. In each plot, there are six point colors that correspond to the colorized true cluster labels. (a) and (b) share the same ground truth w-separated cluster labels, and (c) and (d) share the same ground truth n-separated cluster labels.

SymSim has a parameter σ (Sigma) which adjusts the within-cluster heterogeneity. When σ increases, the clusters are less separable. In our experiments we test the performance of our algorithm with varying σ. The hypothesis is that when σ increases the data is more difficult for clustering algorithms using the gene-expression alone, and we should gain more improvement through integrating location data.

3.1 Results on Simulated Data

We use four different values for σ, $\sigma = \{0.3, 0.4, 0.5, 0.6\}$, in SymSim to generate single cell gene-expression data. For each parameter setting, 10 datesets are generated. To test on datasets where even less genes are measured, we randomly sample 30% and 50% of genes from the original gene-expression datasets with 600 genes. We then conduct the following experiments, which test different settings for gene-expression data and cell location data:

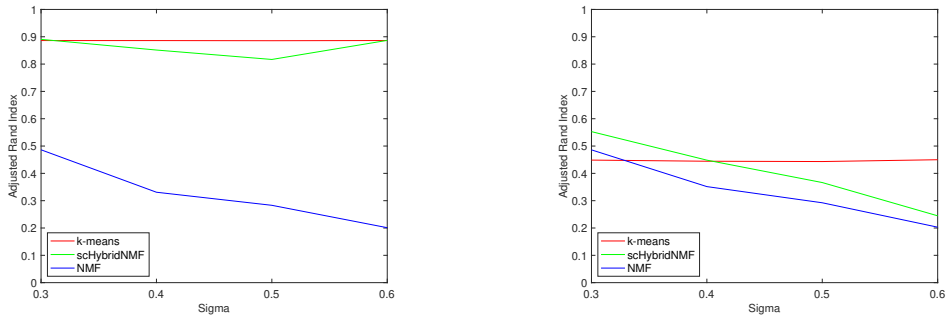
1. Gene-expression data that accounts for all genes; with location data with no noise.
2. Gene-expression data that accounts for all genes; with location data with noise.
3. Gene-expression data of random subsets of the total genes; with location data with noise.

The parameter β in our formulation Eqn. 4 has an impact on the results and we provide analytical forms of setting this parameter. For w-separated location data, we used $\eta = \beta = \frac{\sum_{(i,j)\in A} A(i,j)}{|A|}$, $\alpha = \frac{\|A\|_F^2}{\|H_A\|_F^2}$ and $tol = 10^{-3}$. For n-separated location data, we used $\eta = \beta = \frac{\sum_{(i,j)\in A} A(i,j)}{|A|}, \alpha = \frac{\|A - W_A H_A\|_F^2}{\|L - W_L H_L\|_F^2}$ and $tol = 10^{-3}$. We use the H_A, W_A, H_L, and W_L generated by steps 1 and 2 in Algorithm 1.

To evaluate the performance of scHybridNMF on our data, we calculated the adjusted Rand index (ARI) between the calculated clusters and the ground truth clusters for each set of experiments. In this context, ARI quantifies how similar two clusterings are to each other while correcting for chance. If the ARI of a clustering is very similar to the ground truth clustering, the ARI value should be close to 1. To ensure that there was an even comparison between NMF, k-means clustering, and scHybridNMF, we calculated the NMF and k-means clustering ARIs for the clusters that were used as steps 1 and 2 in Algorithm 1.

Experiment 1: A Motivating Example We start with the scenario where we have very informative location data, the location data without noise, to test whether the information in the location data can be transferred to improve the NMF clustering. Here we use the full

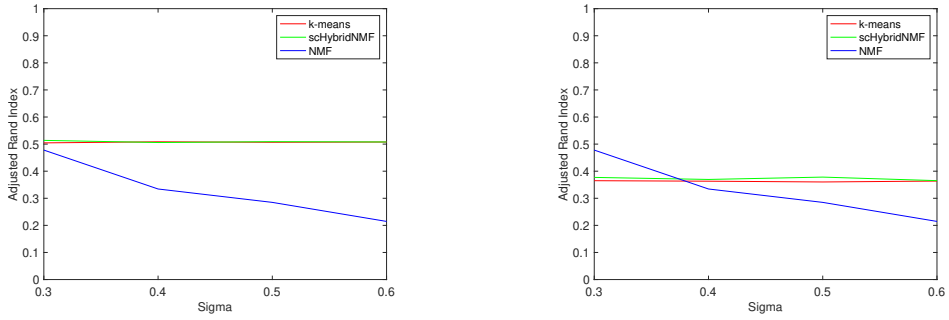
gene-expression data. The goal was to establish that our method beneficially incorporated cell location information with the gene expression clusters. For w-separated data, we calculated the average ARI over 10 location-gene expression pairs for each σ. For n-separated data, we calculated the average ARI over 100 location-gene expression pairs for each σ. We plotted the average values as a function of σ in Fig. 2.



(a) w-separated location data (b) n-separated location data

Fig. 2. Using all genes with location data without noise.

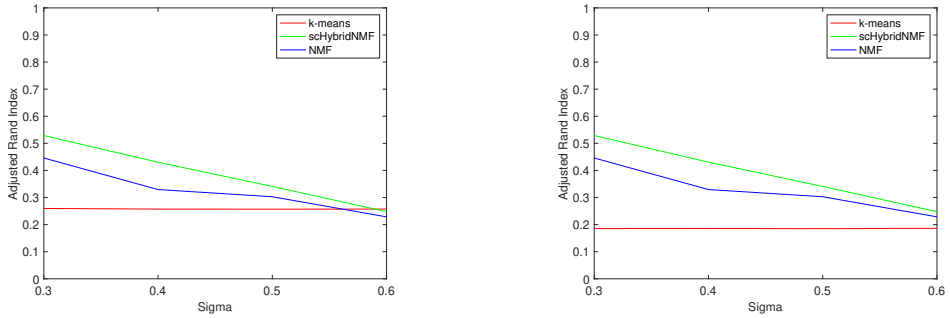
In Fig. 2 (a), since the location data is a very easy case for k-means clustering, we get high ARI by applying k-means alone on the location data. The scHybridNMF is able to elevate the performance of using NMF on only gene-expression data up to the level of the k-means performance. The performance of NMF suffers when σ increases, but scHybridNMF is not affected by the increase of σ thanks to the incorporation of location data. In Fig. 2 (b), where the location data is hard for k-means clustering, scHybridNMF now is more affected by σ, though still improves over NMF, and sometimes also improves over k-means, as now some information from the NMF clustering can be used to correct the wrong clustering of k-means results.



(a) 20% noise (b) 30% noise

Fig. 3. Using all genes with noisy data on w-separated data.

Experiment 2: All Genes, Noisy Location Data We now move to a more realistic setting where there is noise in the location data. We used location data with 20% and 30% noise. For each location data with no noise, we generate 10 noisy location datasets. For w-separated data, we calculated the average ARI over 100 location-gene expression pairs, for each σ and each noise percentage. For n-separated data, we calculated the average ARI over 1000 location-gene expression pairs, for each σ and each noise percentage. We plotted the average values as a function of σ in Figs. 3 and 4.

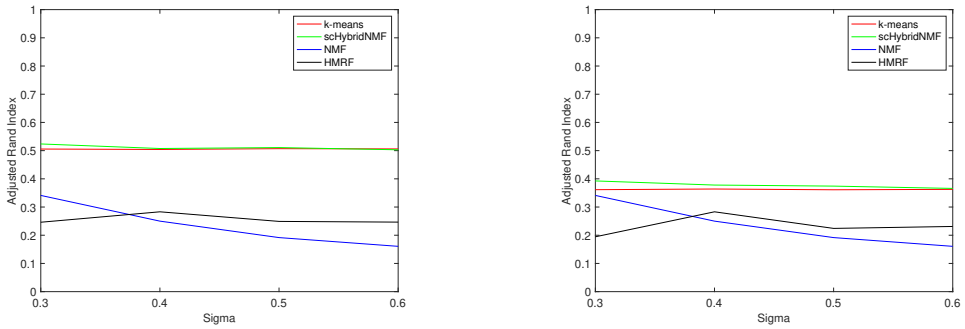


(a) 20% noise (b) 30% noise

Fig. 4. Using all genes with noisy data on n-separated data.

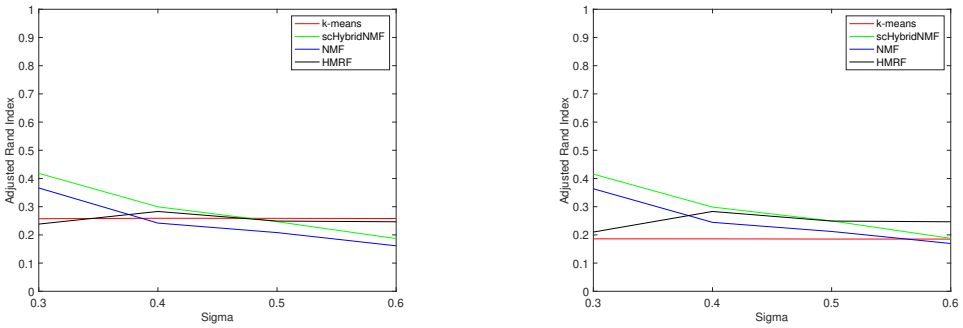
In Figs. 3 and 4, we observe that in both cases, our algorithm had higher ARI values than in using NMF alone. Increasing the amount of noise should decrease the performance of k-means clustering, which is evident from the plots. Even with the decreasing performance of the k-means clustering results, scHybridNMF improves tremendously over NMF. This is especially evident in Fig. 4, where scHybridNMF achieves a higher performance than both NMF and k-means, indicating that scHybridNMF is able to gather useful information from both standalone methods, and that it has high potential to be successful on real-world data.

Experiment 3: Sampled Genes, Noisy Location Data Finally, we investigate the scenario where we use noisy location data and even smaller number of genes (using 50% randomly sampled genes), which is the most challenging scenario. For w-separated data over each σ, we calculated 5 random gene samples over 10 location noise randomizations (for each noise percentage) for each of the 10 location-gene expression pairs. For n-separated data over each σ, we calculated 5 random gene samples over 10 location noise randomizations (for each noise percentage) for each of the 100 location-gene expression pairs. We plotted the average ARI values for sampling 50% of the genes as a function of σ in Figs. 5 and 6. In this experiment, we also run the existing HMRF [24] method which also performs cell clustering using both cell location and gene-expression data, on the same datasets.



(a) 20% noise (b) 30% noise

Fig. 5. Using 50% of the genes with 20% and 30% noise in w-separated data.

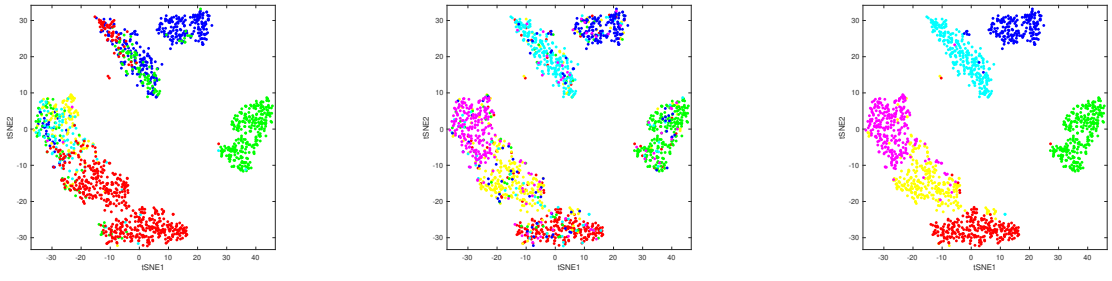


(a) 20% noise (b) 30% noise

Fig. 6. Using 50% of the genes with 20% and 30% noise in n-separated data.

The results show a clear distinction between scHybridNMF and NMF and k-means clustering. For the w-separated location data, the ARI values of scHybridNMF significantly exceeds those of NMF and HMRF clustering. In data with n-separated location data, scHybridNMF tends to outperform both k-means clustering and NMF, and ourperforms HMRF in a majority of cases.

These experiments show that scHybridNMF is robust to small datasets with noisy locations and a subset of the total number of genes. This sort of data is prevalent in the real world, and the fact that our algorithm performs the strongest relative to individually using NMF or k-means on this data indicates that it is likely to be successful for real data. Fig. 7 shows the tSNE plots, which visualize high-dimensional data, of the gene expression data clusters produced by NMF, scHybridNMF, and the ground truth labels. This shows that our method improves the performance of cell clustering of the gene-expression immensely.



(a) NMF (b) scHybridNMF (c) Ground truth

Fig. 7. An example of NMF vs scHybridNMF vs ground truth clusters for gene expression. The data had $\sigma = 0.3$, 30% of the genes sampled, 20% noise in location, and w-separated location data. In each plot, there are six point colors that correspond to the six cluster labels.

3.2 Results on Real Data

Having tested scHybridNMF extensively on simulated data, we now apply it to real data. We use the mouse brain gene-expression and location data from [3]. This data has been adapted from the seqFISH+ dataset [7], and it has been annotated with the locations of specific cells as well as their gene expression levels. To examine the regions that have varied genetic expressions, we isolated the 523 cortex cells and filtered the genes to keep those with

mean greater than 0.7 and correlation of variation greater than 1.2 measured across all cells. We then further sample only 20% of the genes. The analysis from [3] indicates that there may be 9 clusters, so we set the number of clusters $k = 9$. As in Fig. 8, we get 9 distinct clusters, which can correspond to the spatial domains in mouse brain, like those found in [3]. Each cluster contains cells both with similar gene-expression profiles and locate relatively closely in space. Further experiments or analysis can be performed to explore the biological meanings of the identified clusters.

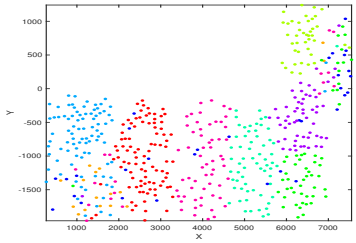


Fig. 8. The result of scHybridNMF on the mouse brain data with $k = 9$ in the location space. Each color corresponds to a cluster.

4 Conclusions and Discussions

In this paper, we present a hybrid clustering approach that can better identify cell types by incorporating the strengths of NMF and k-means clustering, which work well on the high-dimensional single cell gene-expression data and low-dimensional location data, respectively. We show that our hybrid framework, scHybridNMF, significantly improves over the clustering accuracy of using NMF alone on gene-expression data by integrating location information. This is particularly useful for the cases where NMF performance is affected by a low number of genes in the gene-expression data or high within-cluster heterogeneity. scHybridNMF also outperforms k-means clustering with only location data under realistic scenarios. Through combining two classical methods for clustering, NMF and k-means, scHybridNMF can exploit both the high and low dimensional data and achieve better performance than using either of the standalone methods, as well as an existing method HMRF.

The framework we use is flexible and can be extended to include more constraints and more types of data. For example, we can include gene-gene relationship data which represent potential gene-gene interaction in the data, and perform co-clustering of both cells and genes. The inferred gene clusters can be further used to study regulatory mechanisms in the cells and reconstruct gene regulatory networks.

5 Acknowledgements

We thank Grant Bruer and Ziqi Zhang for their editorial comments and discussions. This work was supported in part by the US National Science Foundation DBI-2019771. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of NSF.

References

1. Bertsekas, D.P., Hager, W., Mangasarian, O.: Nonlinear Programming. Athena Scientific Computing, Athena Scientific (1997)
2. Cusanovich, D.A., Daza, R., Adey, A., Pliner, H.A., Christiansen, L., Gunderson, K.L., Steemers, F.J., Trapnell, C., Shendure, J.: Multiplex single cell profiling of chromatin accessibility by combinatorial cellular indexing. Science **348**(6237), 910–914 (May 2015)
3. Dries, R., Zhu, Q., Eng, C.H.L., Sarkar, A., Bao, F., George, R.E., Pierson, N., Cai, L., Yuan, G.C.: Giotto, a pipeline for integrative analysis and visualization of single-cell spatial transcriptomic data (Jul 2019)
4. Du, R., Drake, B., Park, H.: Hybrid clustering based on content and connection structure using joint nonnegative matrix factorization. J. Global Optimiz. **74**(4), 861–877 (2019)
5. Duren, Z., Chen, X., Zamanighomi, M., Zeng, W., Satpathy, A.T., Chang, H.Y., Wang, Y., Wong, W.H.: Integrative analysis of single-cell genomics data by coupled nonnegative matrix factorizations. Proc. Natl. Acad. Sci. U. S. A. **115**(30), 7723–7728 (Jul 2018)
6. Efremova, M., Teichmann, S.A.: Computational methods for single-cell omics across modalities. Nat. Methods **17**(1), 14–17 (Jan 2020)
7. Eng, C.H.L., Lawson, M., Zhu, Q., Dries, R., Koulena, N., Takei, Y., Yun, J., Cronin, C., Karp, C., Yuan, G.C., Cai, L.: Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH. Nature **568**(7751), 235–239 (Apr 2019)
8. Grippo, L., Sciandrone, M.: On the convergence of the block nonlinear Gauss–Seidel method under convex constraints. Oper. Res. Lett. **26**(3), 127–136 (Apr 2000)
9. Jin, S., Zhang, L., Nie, Q.: scAI: an unsupervised approach for the integrative analysis of parallel single-cell transcriptomic and epigenomic profiles. Genome Biol. **21**(1), 25 (Feb 2020)
10. Kim, H., Park, H.: Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis. Bioinformatics **23**(12), 1495–1502 (Jun 2007)
11. Kim, J., He, Y., Park, H.: Algorithms for nonnegative matrix and tensor factorizations: a unified view based on block coordinate descent framework. J. Global Optimiz. **58**(2), 285–319 (Feb 2014)
12. Kotliar, D., Veres, A., Nagy, M.A., Tabrizi, S., Hodis, E., Melton, D.A., Sabeti, P.C.: Identifying gene expression programs of cell-type identity and cellular activity with single-cell RNA-Seq. Elife **8** (Jul 2019)
13. Kuang, D., Yun, S., Park, H.: SymNMF: nonnegative low-rank approximation of a similarity matrix for graph clustering. J. Global Optimiz. **62**(3), 545–574 (Jul 2015)
14. Mayr, U., Serra, D., Liberali, P.: Exploring single cells in space and time during tissue development, homeostasis and regeneration. Development **146**(12) (Jun 2019)
15. McKinley, K.L., Castillo-Azofeifa, D., Klein, O.D.: Tools and concepts for interrogating and defining cellular identity. Cell Stem Cell **26**(5), 632–656 (May 2020)
16. Morris, S.A.: The evolving concept of cell identity in the single cell era. Development **146**(12) (Jun 2019)
17. Peterson, V.M., Zhang, K.X., Kumar, N., Wong, J., Li, L., Wilson, D.C., Moore, R., McClanahan, T.K., Sadekova, S., Klappenbach, J.A.: Multiplexed quantification of proteins and transcripts in single cells. Nat. Biotechnol. **35**(10), 936–939 (Oct 2017)
18. Shao, C., Höfer, T.: Robust classification of single-cell transcriptome data by nonnegative matrix factorization. Bioinformatics **33**(2), 235–242 (Jan 2017)
19. Ståhl, P.L., Salmén, F., Vickovic, S., Lundmark, A., Navarro, J.F., Magnusson, J., Giacomello, S., Asp, M., Westholm, J.O., Huss, M., Mollbrink, A., Linnarsson, S., Codeluppi, S., Borg, Å., Pontén, F., Costea, P.I., Sahlén, P., Mulder, J., Bergmann, O., Lundeberg, J., Frisén, J.: Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. Science **353**(6294), 78–82 (Jul 2016)
20. Stuart, T., Satija, R.: Integrative single-cell analysis. Nat. Rev. Genet. (Jan 2019)
21. Wang, X., Allen, W.E., Wright, M.A., Sylwestrak, E.L., Samusik, N., Vesuna, S., Evans, K., Liu, C., Ramakrishnan, C., Liu, J., Nolan, G.P., Bava, F.A., Deisseroth, K.: Three-dimensional intact-tissue sequencing of single-cell transcriptional states. Science **361**(6400) (Jul 2018)
22. Welch, J.D., Kozareva, V., Ferreira, A., Vanderburg, C., Martin, C., Macosko, E.Z.: Single-Cell multi-omic integration compares and contrasts features of brain cell identity. Cell **177**(7), 1873–1887.e17 (Jun 2019)
23. Zhang, X., Xu, C., Yosef, N.: Simulating multiple faceted variability in single cell RNA sequencing. Nat. Commun. **10**(1), 2611 (Jun 2019)
24. Zhu, Q., Shah, S., Dries, R., Cai, L., Yuan, G.C.: Identification of spatially associated subpopulations by combining scRNAseq and sequential fluorescence in situ hybridization data. Nat. Biotechnol. (Oct 2018)