



THSLRR: A Low-Rank Subspace Clustering Method Based on Tired Random Walk Similarity and Hypergraph Regularization Constraints

Tian-Jing Qiao, Na-Na Zhang, Jin-Xing Liu, Jun-Liang Shang, Cui-Na Jiao, and Juan Wang^(✉)

School of Computer Science, Qufu Normal University, Rizhao 276826, China
wangjuansdu@163.com

Abstract. Single-cell RNA sequencing (scRNA-seq) technology furnishes us with a certainly forceful tool for exploring biological mechanisms from the perspective of single-cell. By clustering scRNA-seq data, different types of cells can be effectively distinguished, which is helpful for disease treatment and the discovery of new cell types. Nevertheless, the existing clustering methods still cannot achieve satisfactory results attributed to the complexity of high-dimensional noisy scRNA-seq data. Therefore, we propose a clustering method called Hypergraph regularization sparse low-rank representation with similarity constraint based on tired random walk (THSLRR). Specifically, the sparse low-rank model rebuilds spatial information from a suite of high-dimensional subspaces by mapping data into subspaces, and removes superfluous information and errors in scRNA-seq data. The hypergraph regularization explores the higher-order manifold structure embedded in the scRNA-seq data. Meanwhile, the similarity constraint based on tired random walk can farther upgrade the learning ability and interpretability of the model. Then, the learned similarity matrix could be for spectral clustering, visualization and identification of marker genes. Compared with other advanced methods, the clustering results of the THSLRR method are more robust and accurate.

Keywords: scRNA-seq · Single-cell type identification · Hypergraph regularization · Similarity constraint

1 Introduction

In the past few years, advances in single-cell RNA sequencing (scRNA-seq) technology have provided a new window of opportunity to learn about biological mechanisms at the single-cell level, and guide scientists in exploring gene expression profiles at the single-cell level [1, 2]. By mining and analyzing scRNA-seq data, we can research cell heterogeneity and identify subgroups. The identification of cell types from scRNA-seq data facilitates the extraction of meaningful biological information, as a matter of unsupervised clustering. With the clustering model, cells that are highly similar will be

grouped into the same cluster. Because of biological factors and technical limitations, however, scRNA-seq data tend to be high-dimensional, sparse and noisy. Consequently, classical clustering methods like K-means [3] and Spectral Clustering (SC) [4] are no longer suitable for scRNA-seq data, and reliable clustering cannot always be used for downstream analysis.

At present, in order to iron out the difficulties existing in scRNA-seq data clustering research, scholars have put forward numerous clustering methods. For instance, through the in-depth research of shared nearest neighbors, Xu and Su came up with a quasi-cluster-based clustering method (SNN-Cliq), which shows greater superiority in clustering high-dimensional single-cell data [5]. Based on the profound study of multi-kernal learning, Wang et al. proposed the SIMLR method, working out dimensionality reduction as well as clustering of data [6]. Park et al. proposed the MPSSC method, in which the SC framework is modified by adding sparse structure constraint, and the similarity matrix is constructed by using multiple double random affinity matrices [7]. Jiang et al. took into account paired cell differentiability correlation and variance, then proposed the Corr model [8].

At the same time, researchers have also proposed a number of subspace clustering methods and proved that the similarity obtained by the subspace clustering method based on low-rank representation (LRR) is more robust than the pairwise similarity involved in the methods mentioned above [9, 10]. For example, Liu et al. proposed the LatLRR method, integrating feature extraction and subspace learning into a unified framework to better cope with severely corrupted observation data [11]. Zheng et al. presented the SinNLRR method, a low-rank based clustering method, that fully exploits the global information of the data by imposing low-rank and non-negative constraints on the similarity matrix [10]. In order to explore the local information of the data, Zhang et al. proposed the SCCLRR method based on SinNLRR with the addition of local feature descriptions to capture both global and local information of the data [9]. Zheng et al. proposed the AdaptiveSSC method based on subspace learning to figure out the matters of noise and high dimensionality in single-cell data, achieving improved performance on multiple experimental data sets [12].

In this paper, we propose a single-cell clustering method called Hypergraph regularization sparse low-rank representation with similarity constraint based on tired random walk (THSLRR), which aims to capture the global structure and local information of scRNA-seq data simultaneously in subspace learning. Concretely, on the basis of the sparse LRR model, the hypergraph regularization based on manifold learning is introduced to mine the complex high-order relationship in scRNA-seq data. At the same time, the similarity constraint based on tired random walk (TRW) further improves the learning ability of model. The final sparse low-rank symmetric matrix Z^* obtained by THSLRR is further operated to learn the affinity matrix H , then H is used for single-cell spectral clustering, t-distributed stochastic neighbor embedding (t-SNE) [13] visual analysis of cells and genes prioritization. Figure 1 illustrates the specific process and applications of THSLRR.

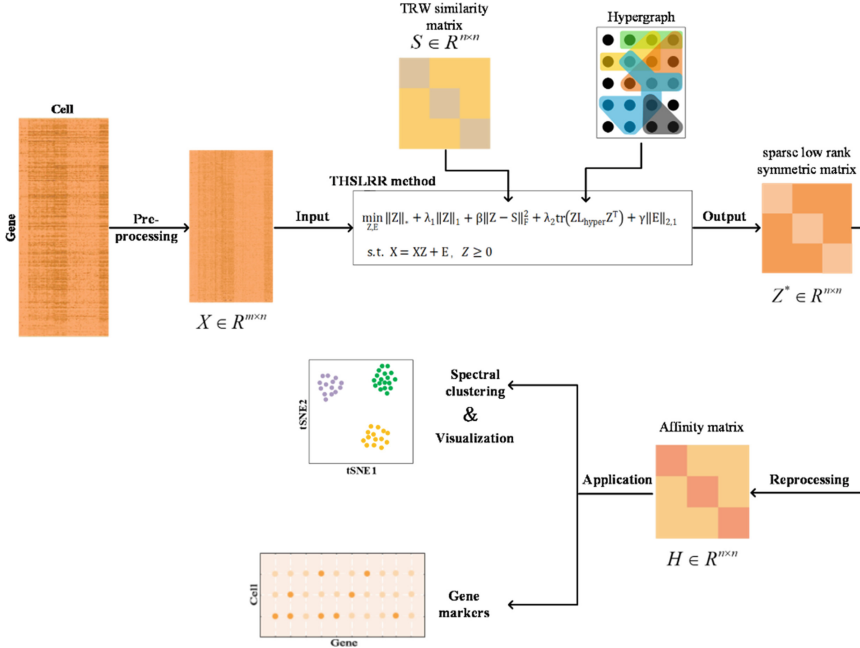


Fig. 1. The framework of THSLRR for scRNA-seq data analysis.

2 Method

2.1 Sparse Low-Rank Representation

The LRR model is a progressive subspace clustering method, which is widely used in data mining, machine learning and other fields. Finding the lowest rank representation of data on the basis of the given data dictionary is the central objective of LRR [14]. Given the scRNA-seq data matrix $X = [X_1, X_2, \dots, X_n] \in R^{m \times n}$, where m represents the number of genes and n is the number of cells, its LRR formula is expressed as follows:

$$\min_{Z,E} \|Z\|_* + \gamma \|E\|_{2,1} \text{ s.t. } X = XZ + E. \tag{1}$$

There, $\|*\|_*$ represents the kernel norm of the matrix, $\|*\|_{2,1}$ is the $l_{2,1}$ norm. E is the error item and Z is the coefficient matrix that demands to be optimized to achieve the lowest rank. $\gamma > 0$ is the parameter to coordinate the influence of errors.

The sparse representation model obtains the sparse coefficient matrix that unravels the close relationship between the data points, what is equivalent to solving the following optimization problem:

$$\min_Z \|Z\|_1 \text{ s.t. } X = XZ, \tag{2}$$

where $\| * \|_1$ is the l_1 norm. We further combine sparse and low-rank constraints for the extraction of salient features and noise removal to obtain the sparse LRR of the matrix, as follows:

$$\min_{Z,E} \|Z\|_* + \lambda \|Z\|_1 + \gamma \|E\|_{2,1} \text{ s.t. } X = XZ + E. \quad (3)$$

Here, λ and γ are regularization parameters.

2.2 Hypergraph Regularization

Extracting local information from high-dimensional sparse noisy data is also a problem worth considering. Therefore, we exploit the hypergraph to encode higher-order geometric relationships among multiple sample points, which can more fully extract the underlying local information of scRNA-seq data.

For a given hypergraph $G = (V, E, W)$, $V = \{v_1, v_2, \dots, v_n\}$ is the collection of vertices, $E = \{e_1, e_2, \dots, e_r\}$ is the collection of hyperedges, W is the hyperedge weight matrix. The incidence matrix R of the hypergraph G is calculated as follows:

$$R(v, e) = \begin{cases} 1 & \text{if } v \in e \\ 0 & \text{others} \end{cases} \quad (4)$$

The weight $w(e_i)$ of hyperedge e_i is obtained by the following formula:

$$w(e_i) = \sum_{\{v_i, v_j\} \in e_i} \exp \frac{-\|v_i - v_j\|_2^2}{\delta^2}, \quad (5)$$

where $\delta = \sum_{\{v_i, v_j\} \in e_i} \|v_i - v_j\|_2^2 / k$, and k represents the number of nearest neighbors of each vertex. The degree $d(v)$ of vertex v is as follows:

$$d(v) = \sum_{e \in E} w(e) R(v, e). \quad (6)$$

The degree $g(e)$ of hyperedge e is as follows:

$$g(e) = \sum_{v \in V} R(v, e). \quad (7)$$

Then, we obtain the non-normalized hypergraph Laplacian matrix L_{hyper} , as shown below:

$$L_{hyper} = D_v - RW_H(D_H)^{-1}R^T. \quad (8)$$

where vertex degree matrix D_v , hyperedge degree matrix D_H and hyperedge weight matrix W_H are diagonal matrices, and the elements on the diagonal are $d(v)$, $g(e)$ and $w(e)$ respectively.

Under certain conditions of the mapping, z_i and z_j are the mapping representations of the original data points x_i and x_j under the new basis, then the target formula of the hypergraph regularization constraint is as follows:

$$\begin{aligned} \min_Z \frac{1}{2} \sum_{e \in E} \sum_{(i,j) \in e} \frac{w(e)}{g(e)} \|z_i - z_j\|^2 &= \min_Z \text{tr} \left(Z \left(D_v - RW_H (D_H)^{-1} R^T \right) Z^T \right) \\ &= \min_Z \text{tr} \left(Z L_{\text{hyper}} Z^T \right) \end{aligned} \quad (9)$$

2.3 Tired Random Walk

The TRW model was proposed in [15] and proved to be a practical measurement of nonlinear manifold [16]. Therefore, the similarity constraint can not only improve the learning ability of the model for the overall geometric information of the data, but also ensure the symmetry of the similarity matrix, so that the model has better interpretability.

For an undirected weight graph with n vertexes, the transition probability matrix of the random walk is $P = D^{-1}W$, W represents the affinity matrix of the graph, D represents the diagonal matrix with $D_{ii} = \sum_{j=1}^n W_{ij}$. According to [17], the cumulative transition probability matrix is $P_{TRW} = \sum_{s=0}^{\infty} (\tau P)^s$ for all vertices, where $\tau \in (0, 1)$ and the eigenvalue of P is at $[0, 1]$, so the TRW matrix is as follows:

$$P_{TRW} = \sum_{s=0}^{\infty} (\tau P)^s = (1 - \tau P)^{-1}. \quad (10)$$

In order to weaken the effect of errors existing in the primary samples and ensure that the paired sample points have consistent correlation weights, we further symmetrize P_{TRW} to obtain final TRW similarity matrix $S \in R^{n \times n}$ as follows:

$$S(x_i, x_j) = \frac{(P_{TRW})_{ij} + (P_{TRW})_{ji}}{2}. \quad (11)$$

2.4 Objective Function of THSLRR

THSLRR learns the expression matrix $Z \in R^{n \times n}$ from the scRNA-seq data matrix $X = [X_1, X_2, \dots, X_n] \in R^{m \times n}$ with m genes and n cells by the following objective function (12):

$$\begin{aligned} \min_{Z, E} \|Z\|_* + \lambda_1 \|Z\|_1 + \lambda_2 \text{tr}(Z L_{\text{hyper}} Z^T) + \beta \|Z - S\|_F^2 + \gamma \|E\|_{2,1} \\ \text{s.t. } X = XZ + E, Z \geq 0, \end{aligned} \quad (12)$$

where Z is the coefficient matrix to be optimized, $L_{\text{hyper}} \in R^{n \times n}$ is the hypergraph Laplacian matrix, $S \in R^{n \times n}$ is the symmetric cell similarity matrix generated by TRW, $E \in R^{m \times n}$ represents the errors term, $\|*\|_F$ is the Frobenius norm of the matrix, λ_1 , λ_2 , β and γ are the penalty parameters.

2.5 Optimization Process and Spectral Clustering of THSLRR Method

The objective function that has multiple constraints of THSLRR is a convex optimization problem. In order to effectively work out the problem (12), we adopt the Linearized Adaptive Direction Method with Adaptive Penalty (LADMAP) [18].

Initially, to separate the objective function (12) by using an auxiliary variable J , and then obtain formula (13):

$$\begin{aligned} \min_{Z, E, J} \|Z\|_* + \lambda_1 \|J\|_1 + \lambda_2 \text{tr}(ZL_{\text{hyper}}Z^T) + \beta \|Z - S\|_F^2 + \gamma \|E\|_{2,1} \\ \text{s.t. } X = XZ + E, Z = J, Z \geq 0. \end{aligned} \quad (13)$$

Then, the augmented lagrangian multiplier method is introduced to eliminate the linear constraints existing in (13). Therefore, we get the following formula:

$$\begin{aligned} L(Z, E, J, Y_1, Y_2) = \|Z\|_* + \lambda_1 \|J\|_1 + \lambda_2 \text{tr}(ZL_{\text{hyper}}Z^T) + \beta \|Z - S\|_F^2 + \gamma \|E\|_{2,1} \\ + \langle Y_1, X - XZ - E \rangle + \langle Y_2, Z - J \rangle \\ + \frac{\mu}{2} \left(\|X - XZ - E\|_F^2 + \|Z - J\|_F^2 \right). \end{aligned} \quad (14)$$

Here, μ is a penalty parameter, Y_1 and Y_2 are lagrangian multipliers.

Finally, the optimization problem is ironed out by updating one of the variables by turn while fixing the other variables. Therefore, the update rules of Z , E , and J are as follows:

$$Z_{k+1} = \theta \frac{1}{\eta\mu} \left(Z_k - \frac{\nabla_Z q(Z_k)}{\eta} \right). \quad (15)$$

$$E_{k+1}(i, :) = \begin{cases} \frac{\|p_i\| - \frac{\gamma}{\mu_k}}{\|p_i\|} p_i, & \frac{\gamma}{\mu_k} < \|p_i\|. \\ 0, & \text{otherwise} \end{cases} \quad (16)$$

$$J_{k+1} = \max \left\{ \theta \frac{\lambda}{\mu_k} \left(Z_{k+1} + Y_2^k / \mu_k \right), 0 \right\}. \quad (17)$$

The sparse low-rank symmetric matrix Z^* is obtained with our THSLRR method, and the elements on both sides of the main diagonal of the matrix Z^* correspond to the similarity weights of the data sample points. Inspired by [19], we use the main direction angle information of matrix Z^* to learn the affinity matrix H . Finally, we use learned matrix H as the input of SC method to obtain the clustering results.

3 Results and Discussion

3.1 Evaluation Measurements

In the experiment, two commonly used indicators are used to assess the effectiveness of THSLRR, namely adjusted rand index (ARI) [20] and normalized mutual information (NMI) [21]. The value of ARI belongs to $[-1, 1]$ while the value of NMI is $[0, 1]$.

Given the real cluster label $T = \{T_1, T_2, \dots, T_K\}$ and the predicted cluster label $Y = \{Y_1, Y_2, \dots, Y_K\}$ of n sample points. The formula of ARI is as follows:

$$ARI(T, Y) = \frac{\binom{n}{2}(a_{ty} + a) - [(a_{\Sigma} + a_t)(a_{ty} + a_y) + (a_t + a)(a_y + a)]}{\binom{n}{2} - [(a_{ty} + a_t)(a_{ty} + a_y) + (a_t + a)(a_y + a)]}. \quad (18)$$

Here, a_{ty} denotes the number of data points put in the same class, whereas a_t denotes the number of data points in the same class T but separate Y classes. a_y represents the number of data point pairs that are in the same cluster in Y but not in the same cluster in T , whereas a is the number of data point pairs that are neither in the same cluster of Y nor in the same cluster of T .

NMI is defined as follows:

$$NMI(T, Y) = \frac{\sum_{t \in T} \sum_{y \in Y} p(t, y) \ln \left(\frac{p(t, y)}{p(t)p(y)} \right)}{\sqrt{H(T) \cdot H(Y)}}, \quad (19)$$

Here, $H(T)$ and $H(Y)$ represent the information entropy of the tags T and Y , respectively. $p(t)$ and $p(y)$ are the marginal distribution of t and y , $p(t, y)$ represents the joint distribution function of t and y .

3.2 scRNA-seq Datasets

In this paper, nine different scRNA-seq datasets were used to do the relevant experimental analysis. The datasets involved in the experiment include Treutlein [22], Ting [23], Pollen [24], Deng [25], Goolam [26], Kolod [27], mECS, Engel4 [28] and Darmanis [29]. The detailed information of the nine scRNA-seq data sets are shown in Table 1.

3.3 Parameters Setting

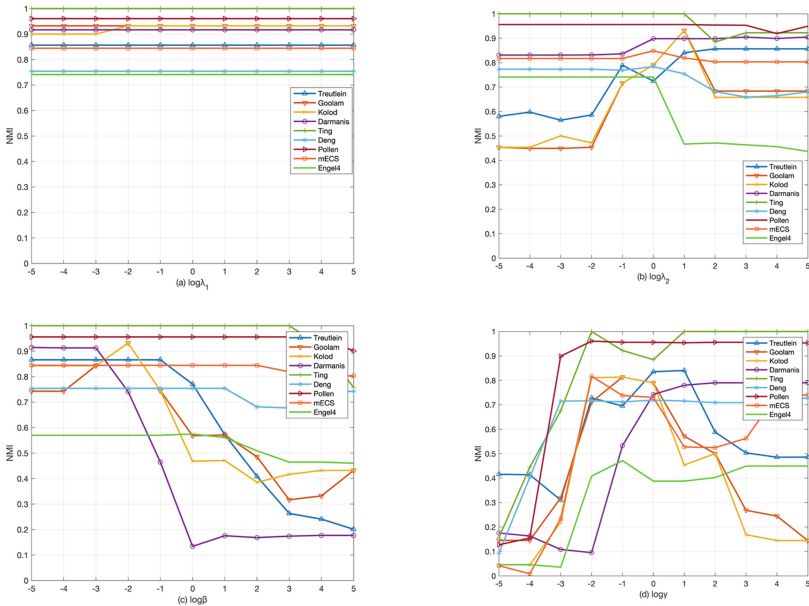
In this part, we specifically discuss the influence of different parameters with regard to the effectiveness of THSLRR method. We make use of the grid search method to determine

Table 1. The scRNA-seq data sets used in experiments.

Data set	Cells	Genes	Cell type	Species
Treutlein	80	959	5	Homo sapiens
Ting	114	14405	5	Mus musculus
Deng	135	12548	7	Mus musculus
Pollen	249	14805	11	Homo sapiens
Goolam	124	40315	5	Mus musculus
Kolod	704	10685	3	Mus musculus
mECS	182	8989	3	Mus musculus
Engel4	203	23337	4	Homo sapiens
Darmanis	420	22085	8	Homo sapiens

Table 2. The optimal values of four parameters for scRNA-seq data sets.

Data sets	λ_1	λ_2	β	γ
Treutlein	$10^{0.5}$	10^1	10^{-1}	10^1
Ting	10^1	10^1	10^{-4}	10^1
Deng	$10^{0.7}$	$10^{1.1}$	10^{-2}	$10^{-2.3}$
Pollen	10^0	10^1	10^3	$10^{-1.5}$
Goolam	$10^{0.1}$	10^1	10^{-2}	$10^{-1.1}$
Kolod	$10^{0.5}$	10^1	10^{-2}	$10^{-1.1}$
mECS	$10^{0.9}$	10^2	10^2	$10^{-2.2}$
Engel4	$10^{0.8}$	10^{-1}	$10^{2.6}$	$10^{-1.2}$
Darmanis	$10^{0.2}$	10^1	$10^{-1.9}$	$10^{-0.2}$

**Fig. 2.** Sensitivity of different parameters to clustering performance of nine scRNA-seq data sets. (a) λ_1 varying. (b) λ_2 varying. (c) β varying. (d) γ varying.

the optimal combination of parameters. The four parameters change in separate intervals $[10^{-5}, 10^5]$, and when one of the parameters changes, the other parameters are fixed, and then we get Fig. 2. In Fig. 2, the clustering results are insensitive to different λ_1 , while λ_2 , β and γ have a greater impact on the model performance. Fortunately, within a certain range, we can choose the appropriate combination of parameters to achieve the optimal clustering result. Therefore, we obtain the optimal parameters of different datasets, as shown in Table 2.

3.4 Comparative Analysis of Clustering

We conduct experiments on nine scRNA-seq data sets recounted in Table 1 to discuss the clustering performance of THSLRR. t-SNE, K-means, SIMLR, SC, Corr, MPSSC and SinNLRR are selected as comparison methods. In order to ensure the fairness and objectivity of the comparison, we furnish the real number of classes to THSLRR as well as the other seven methods, and their parameters are all set to the optimal parameters. The comparison results are shown in Fig. 3 and Table 3.

By observing Fig. 3 and Table 3, we can draw the following conclusions:

- 1) In Fig. 3(a), the median ARI for comparison methods in all datasets is below 0.7, while the median value of THSLRR is greater than 0.9. Furthermore, it is the flattest compared to the box plots of the other seven methods, indicating that the performance of THSLRR is more stable. Similar results can be found in Fig. 3(b).
- 2) In Table 3, SinNLRR outperforms SIMLR, MPSSC, and Corr on most datasets, and the average ARI for SinNLRR is approximately 11%, 6% and 20% higher, respectively. THSLRR exceeds SIMLR, MPSSC and Corr on all datasets except mECS, and outperforms SIMLR, MPSSC and Corr in terms of average ARI by about 27%, 22% and 36% respectively. As can be seen, the low-rank based clustering methods SinNLRR and THSLRR achieve satisfactory clustering results on most of the data sets, indicating the critical contribution of global information to improve the clustering performance once again. In contrast, SIMLR, MPSSC and Corr only take into consideration the local information between samples, their clustering performance is not as impressive as SinNLRR and THSLRR on most of the datasets.
- 3) It can also be seen from Table 3 that THSLRR exceeds the SinNLRR method by about 16% in ARI score. There are two main factors. First, the THSLRR method utilizes the hypergraph regularization to thoroughly mine the complex high-order relationships of scRNA-seq data, while sinNLRR simply considers the overall information of the data. Secondly, the similarity based on TRW captures the global manifold structure information of the data and improves the learning ability of the model.

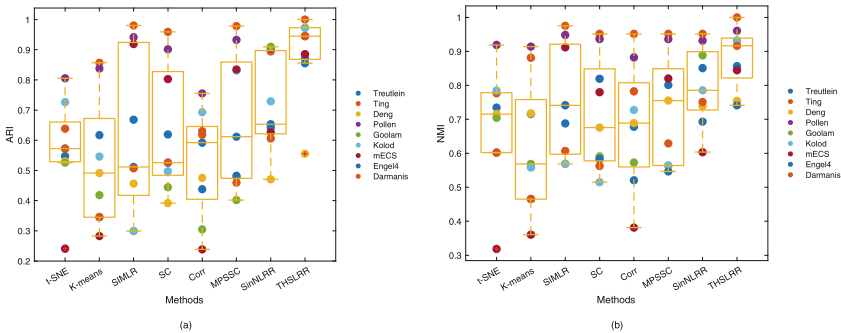


Fig. 3. Clustering results of eight clustering methods on nine scRNA-seq data sets. (a) ARI. (b) NMI

In conclusion, THSLRR achieves the best results on most data sets. Moreover, the average ARI and NMI of THSLRR increase by approximately 12% and 22% compared with comparison methods. Therefore, the THSLRR method is rational and it has certain advantages in cell type identification.

Table 3. The clustering performance on the scRNA-seq data

Method	ARI							
	t-SNE	K-means	SIMLR	SC	Corr	MPSSC	SinNLRR	THSLRR
Treutlein	0.5473	0.6172	0.5114	0.6191	0.5919	0.6117	0.6419	0.8722
Ting	0.6384	0.8567	0.9803	0.9592	0.6302	0.9784	0.8943	1.0000
Deng	0.5301	0.4914	0.4565	0.3917	0.4753	0.4783	0.4706	0.5553
Pollen	0.8055	0.8378	0.9415	0.9013	0.7553	0.9328	0.9051	0.9448
Goolam	0.5255	0.4182	0.2991	0.4445	0.3046	0.402	0.9097	0.9727
Kolod	0.7265	0.5462	0.2991	0.4974	0.6928	0.8306	0.7291	0.9727
mECS	0.2408	0.2824	0.9186	0.8028	0.2385	0.8347	0.6263	0.8857
Engel4	0.5725	0.3453	0.6682	0.5258	0.4377	0.4821	0.6533	0.8554
Darmanis	0.5725	0.3453	0.5069	0.5258	0.6183	0.4593	0.6057	0.9452
average	0.5732	0.5494	0.6202	0.6427	0.5269	0.6678	0.7288	0.8893

3.5 Visualize Cells Using t-SNE

According to [6], we make use of the improved t-SNE to map the learned matrix H to the two-dimensional space to observe the structure representation performance of THSLRR method. We only analyze the visualization results for the Ting and Darmanis datasets because of space limitations.

As shown in Fig. 4(a), THSLRR does not distinguish class 1 from class 4 on the Treutlein data, but the boundaries among other types of cells are more obvious. SinNLRR does not distinguish the three cell types 1, 3 and 4, the boundary between classes 2 and 5 is also very blurred. The distribution of t-SNE, SIMLR and MPSSC cells are also scattered. In Fig. 4(b), the result of t-SNE is the worst, SIMLR divides cells belonging to the same class into two clusters, SinNLRR and MPSSC fail to separate the two types of cells and THSLRR can correctly separate five cell types. All methods do not show promising results on the Pollen and Darmanis datasets in Fig. 4(c) and Fig. 4(d), while THSLRR performed best overall because almost all cells belonging to the same cluster are segregated into the same group and the boundaries between clusters were relatively clear.

3.6 Gene Markers Prioritization

In this section, the affinity matrix H learned from THSLRR is used to prioritize genes. First, the bootstrap Laplacian score that is proposed in [6] is used for identifying gene

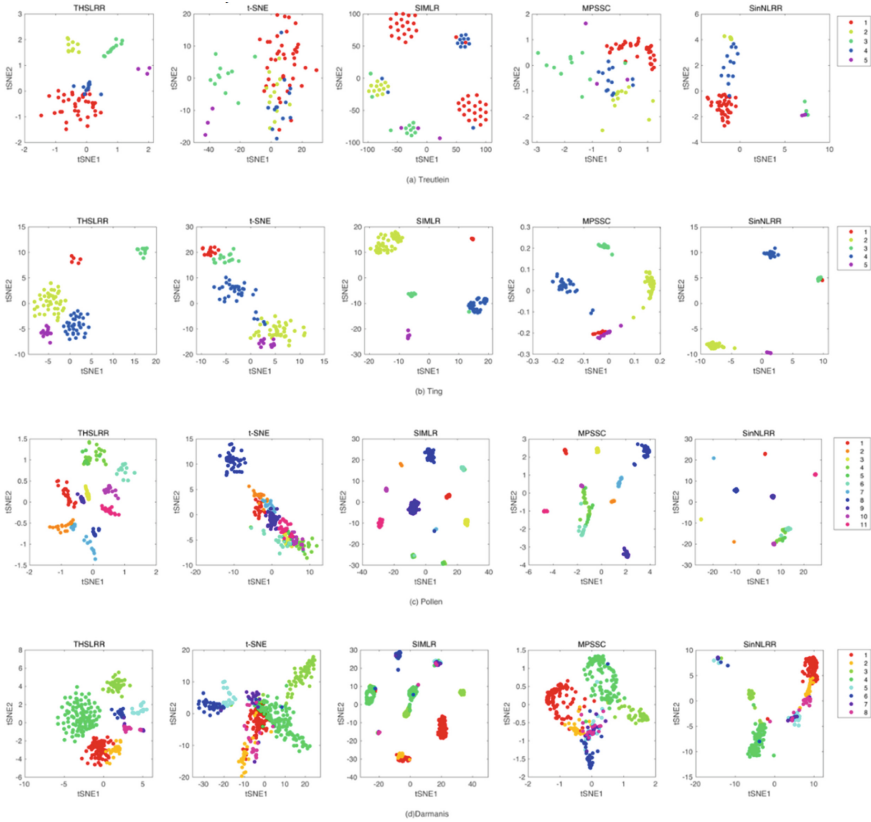


Fig. 4. Visualization results of the cells on (a) Treutlein, (b) Ting, (c) Pollen, and (d) Darmanis datasets.

markers on the matrix H . Then, the genes are placed in descending order in the light of their importance in distinguishing cell subpopulations. Finally, the top ten genes are selected for visual analysis. We use Engel4 and Darmanis data sets for gene markers analysis.

On Darmanis and Engel4 data sets, we select the top 10 gene markers as shown in Fig. 5(a) and Fig. 5(b) respectively. The color of the ring indicates the mean expression level of the gene, and the darker the color, the higher the average expression level of the gene. The size of the ring means the percentage of gene expression in the cell.

Figure 5(a) shows the top ten genes of Darmanis data set. The genes *SLC1A3*, *SLC1A2*, *SPARCL1* and *AQP4* have a high level of expression in astrocytes, and they play an essential part in early development of astrocytes. In fetal quiescent, *SOX4*, *SOX11*, *TUBA1A* and *MAP1B* have a high level of expression and have been proven to be marker genes with specific roles [30–33]. *MAP1B* in neurons is also highly expressed. *PLP12* and *CLDND1* with high expression in oligodendrocytes can be regarded as gene markers of oligodendrocytes [34]. In the Engel4 data, as shown in Fig. 5(b), Engel et al. have been confirmed for *Serp1b1a*, *Tmsb10*, *Hmgb2* and *Malta1* [28]. The remaining genes have also been selected as marker genes in related literature [35, 36].

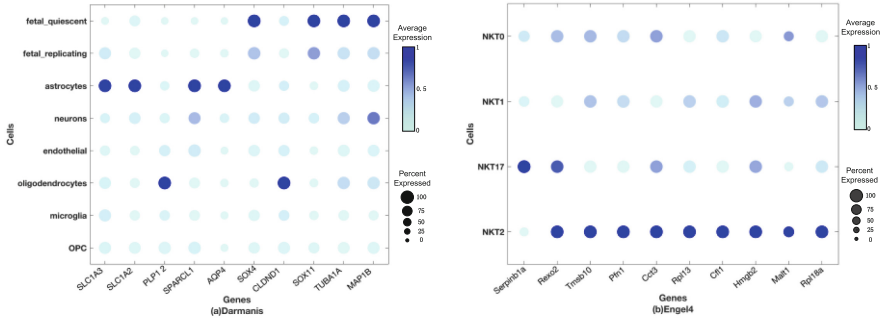


Fig. 5. The top ten gene markers. (a) Darmanis data set. (b) Engel4 data set.

4 Conclusion

In this paper, we propose a clustering method based on subspace learning, named THSLRR. There are mainly two differences where our method differs from other subspace clustering methods. The first aspect is the introduction of hypergraph regularization, which is used to encode higher-order geometric relationships among data and to mine the internal information of data. Compared with other subspace clustering methods, the complex relationships of data can be extracted by our method. Another aspect is the similarity constraint based on TRW, it can mine the global nonlinear manifold structure information of the data and improve the clustering performance and the interpretability of the model. Comparative experiments prove the effectiveness of the THSLRR method. Moreover, the THSLRR method can also provide guidance for data mining as well as be employed in other related domains.

Now, we would like to discuss the limitations of our model. Primarily, although the optimal combination of parameters can be searched by the grid search method, it would be helpful if the optimal parameters could be determined automatically based on some strategy. Second, we use the single similarity criterion in our model, which may not be comprehensive for capturing similarity information from the data. So we can try to use measurement fusion to capture more accurate prior information in the next work.

Funding. This work was supported in part by the National Science Foundation of China under Grant Nos. 62172253 , 61972226 and 62172254.

References

1. Kalisky, T., Quake, S.R.: Single-cell genomics. *Nat. Methods* **8**(4), 311–314 (2011). <https://doi.org/10.1038/nmeth0411-311>
2. Pelkmans, L.: Using cell-to-cell variability—a new era in molecular biology. *Science* **336**(6080), 425 (2012). <https://doi.org/10.1126/science.1222161>
3. Forgy, E.W.: Cluster analysis of multivariate data: efficiency versus interpretability of classifications. *Biometrics* **21**(3) (1965). <https://doi.org/10.1080/00207239208710779>
4. Luxburg, U.V.: A tutorial on spectral clustering. *Stat. Comput.* **17**(4), 395–416 (2004). <https://doi.org/10.1007/s11222-007-9033-z>

5. Xu, C., Su, Z.: Identification of cell types from single-cell transcriptomes using a novel clustering method. *Bioinformatics* **12**, 1974–1980 (2015). <https://doi.org/10.1093/bioinformatics/btv088>
6. Wang, B., Zhu, J., Pierson, E., Ramazzotti, D., Batzoglou, S.: Visualization and analysis of single-cell RNA-seq data by kernel-based similarity learning. *Nat. Methods* **14**(4), 414 (2017). <https://doi.org/10.1038/nmeth.4207>
7. Park, S., Zhao, H., Birol, I.: Spectral clustering based on learning similarity matrix. *Bioinformatics* **34**(12) (2018). <https://doi.org/10.1093/bioinformatics/bty050>
8. Jiang, H., Sohn, L.L., Huang, H., Chen, L.: Single cell clustering based on cell-pair differentiability correlation and variance analysis. *Bioinform. (Oxf. Engl.)* **21**, 3684 (2018). <https://doi.org/10.1093/bioinformatics/bty390>
9. Zhang, W., Li, Y., Zou, X.: SCCLRR: a robust computational method for accurate clustering single cell RNA-seq data. *IEEE J. Biomed. Health Inform.* **25**(1), 247–256 (2020). <https://doi.org/10.1109/JBHI.2020.2991172>
10. Zheng, R., Li, M., Liang, Z., Wu, F.X., Pan, Y., Wang, J.: SinNLRR: a robust subspace clustering method for cell type detection by nonnegative and low rank representation. *Bioinformatics* (2019). <https://doi.org/10.1093/bioinformatics/btz139>
11. Liu, G., Yan, S.: Latent Low-Rank Representation for subspace segmentation and feature extraction. *IEEE* (2012). <https://doi.org/10.1109/ICCV.2011.6126422>
12. Zheng, R., Liang, Z., Chen, X., Tian, Y., Cao, C., Li, M.: An adaptive sparse subspace clustering for cell type identification. *Front. Genet.* **11**, 407 (2020). <https://doi.org/10.3389/fgene.2020.00407>
13. Van der Maaten, L., Hinton, G.: Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9**(11), 2579–2605 (2008)
14. Liu, G., Lin, Z., Yan, S., Sun, J., Yu, Y., Ma, Y.: Robust recovery of subspace structures by low-rank representation. *IEEE Trans. Pattern Anal. Mach. Intell.* **35**(1), 171–184 (2012). <https://doi.org/10.1109/TPAMI.2012.88>
15. Tu, E., Cao, L., Yang, J., Kasabov, N.: A novel graph-based k-means for nonlinear manifold clustering and representative selection. *Neurocomputing* **143**, 109–122 (2014). <https://doi.org/10.1016/j.neucom.2014.05.067>
16. Wang, H., Wu, J., Yuan, S., Chen, J.: On characterizing scale effect of Chinese mutual funds via text mining. *Signal Process.* **124**, 266–278 (2016). <https://doi.org/10.1016/j.sigpro.2015.05.018>
17. Et, A., Yz, B., Lin, Z.C., Jie, Y.D., Nk, E.: A graph-based semi-supervised k nearest-neighbor method for nonlinear manifold distributed data classification. *Inf. Sci.* **367–368**, 673–688 (2016). <https://doi.org/10.1016/j.ins.2016.07.016>
18. Lin, Z., Liu, R., Su, Z.: Linearized alternating direction method with adaptive penalty for low-rank representation. In: *Advances in Neural Information Processing Systems*, pp. 612–620 (2011). <https://doi.org/10.48550/arXiv.1109.0367>
19. Chen, J., Mao, H., Sang, Y., Yi, Z.: Subspace clustering using a symmetric low-rank representation. *Knowl. Based Syst.* **127**, 46–57 (2017). <https://doi.org/10.1016/j.knosys.2017.02.031>
20. Meilă, M.: Comparing clusterings—an information based distance. *J. Multivar. Anal.* **98**(5), 873–895 (2007). <https://doi.org/10.1016/j.jmva.2006.11.013>
21. Strehl, A., Ghosh, J.: Cluster ensembles - a knowledge reuse framework for combining multiple partitions. *J. Mach. Learn. Res.* **3**(3), 583–617 (2002). <https://doi.org/10.1162/153244303321897735>
22. Treutlein, B., et al.: Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. *Nature* (2014). <https://doi.org/10.1038/nature13173>

23. Ting, D.T., et al.: Single-cell RNA sequencing identifies extracellular matrix gene expression by pancreatic circulating tumor cells. *Cell Rep.* **8**(6), 1905–1918 (2014). <https://doi.org/10.1016/j.celrep.2014.08.029>
24. Pollen, A.A., et al.: Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. *Nat. Biotechnol.* **32**(10), 1053–1058 (2014). <https://doi.org/10.1038/nbt.2967>
25. De Ng, Q., Ramskld, D., Reinius, B., Sandberg, R.: Single-Cell RNA-Seq reveals dynamic, random monoallelic gene expression in mammalian cells. *Science* **343** (2014). <https://doi.org/10.1126/science.1245316>
26. Goolam, M., et al.: Heterogeneity in Oct4 and Sox2 targets biases cell fate in 4-cell mouse embryos. *Cell* **165**(1), 61–74 (2016). <https://doi.org/10.1016/j.cell.2016.01.047>
27. Kolodziejczyk, A.A., et al.: Single cell RNA-sequencing of pluripotent states unlocks modular transcriptional variation. *Cell Stem Cell* **17**(4), 471–485 (2015). <https://doi.org/10.1016/j.stem.2015.09.011>
28. Engel, I., et al.: Innate-like functions of natural killer T cell subsets result from highly divergent gene programs. *Nat. Immunol.* (2016). <https://doi.org/10.1038/ni.3437>
29. Darmanis, S., et al.: A survey of human brain transcriptome diversity at the single cell level. *Proc. Natl. Acad. Sci. U. S. A.* **112**(23), 7285–7290 (2015). <https://doi.org/10.1073/pnas.1507125112>
30. Takemura, R., Okabe, S., Umeyama, T., Kanai, Y., Hirokawa, N.: Increased microtubule stability and alpha tubulin acetylation in cells transfected with microtubule-associated proteins MAP1B, MAP2 or tau. *J. Cell Sci.* **103**(Pt 4), 953–964 (1993). <https://doi.org/10.1083/jcb.119.6.1721>
31. Uwanogho, D., et al.: Embryonic expression of the chicken Sox2, Sox3 and Sox11 genes suggests an interactive role in neuronal development. *Mech. Dev.* **49**(1–2), 23–36 (1995). [https://doi.org/10.1016/0925-4773\(94\)00299-3](https://doi.org/10.1016/0925-4773(94)00299-3)
32. Medina, P.P., et al.: The SRY-HMG box gene, SOX4, is a target of gene amplification at chromosome 6p in lung cancer. *Huma. Mol. Genet.* **18**(7), 1343 (2009). <https://doi.org/10.1093/hmg/ddp034>
33. Cushion, T.D., et al.: Overlapping cortical malformations and mutations in TUBB2B and TUBA1A. *Brain A J. Neurol.* **2**, 536–548 (2013). <https://doi.org/10.1093/brain/aws338>
34. Numasawa-Kuroiwa, Y., et al.: Involvement of ER stress in dysmyelination of pelizaeus-merzbacher disease with PLP1 missense mutations shown by iPSC-derived oligodendrocytes. *Stem Cell Rep.* **2**(5), 648–661 (2014). <https://doi.org/10.1016/j.stemcr.2014.03.007>
35. Yu, N., Liu, J.X., Gao, Y.L., Zheng, C.H., Shang, J., Cai, H.: CNLLRR: a novel low-rank representation method for single-cell RNA-seq data analysis. *Hum. Genomics* (2019). <https://doi.org/10.1101/818062>
36. Jiao, C.-N., Liu, J.-X., Wang, J., Shang, J., Zheng, C.-H.: Visualization and analysis of single cell RNA-seq data by maximizing correntropy based non-negative low rank representation. *IEEE J. Biomed. Health Inform.* (2021). <https://doi.org/10.1109/JBHI.2021.3110766>