Supplementary materials of "SinNLRR: a robust subspace clustering method for cell type detection by nonnegative and low rank representation"

Ruiqing Zheng[1], Min Li[1, *], Zhenlan Liang[1], Fang-Xiang Wu[1,2] , Yi Pan[1,3] and Jianxin Wang[1]

[1] School of Computer Science and engineering, Central South University, Changsha 410083, China and [2] Division of Biomedical Engineering, University of Saskatchewan, Saskatoon SKS7N5A9, Canada, [3] Department of Computer Science, Georgia State University, Atlanta, GA 30302-4110, USA

A. The details of ADMM for solving NLRR

According to the Equation 3 in the main context, the augmented Lagrangian formulation is as follows:

$$\mathcal{L}_{\frac{1}{\gamma}}(C,J,Y) = \frac{1}{2}||X - XC||_F^2 + \lambda\left||J|\right|_* + Y^T(J - C) + \frac{1}{2\gamma}\left||J - C|\right|^2 \qquad (S1)$$

where *X* is the single cell RNA-seq matrix, *C* is the target similarity matrix, *J* is an auxiliary matrix, *Y* is the dual variable or Lagrange multiplier and γ is a user-defined parameter. The ADMM update the one of matrix *C*, *J*, *Y* by fixing others. So the *K*-th iteration of updating is as follows:

$$C^{k+1} = \mathrm{argmin}_C\ \mathcal{L}_{\frac{1}{\gamma}}(C, J^k, Y^k) \qquad (S2)$$

$$J^{k+1} = \mathrm{argmin}_J\ \mathcal{L}_{\frac{1}{\gamma}}(C^{k+1}, J, Y^k) \qquad (S3)$$

$$Y^{k+1} = \ Y^k + \frac{1}{\gamma}(C^{k+1} - J^{k+1}) \qquad (S4)$$

The Y^{k+1} could be calculated based on the Y^k, C^{k+1} and J^{k+1}. The optimal C^{k+1} and J^{k+1} could be derived as following rules:

$$\begin{aligned} C^{k+1} &= \mathrm{argmin}_C\ \mathcal{L}_{\frac{1}{\gamma}}(C, J^k, Y^k) \qquad (S5) \\ &= \mathrm{argmin}_C \left\{\frac{1}{2}\|X - XC\|^2 + \lambda\|J^k\|_* + Y^{k^T}(J^k - C) + \frac{1}{2\gamma}\|J^k - C\|^2\right\} \\ &= \mathrm{argmin}_C \left\{X^T(XC - X) - Y^k + \frac{1}{\gamma}(C - J)\right\} \\ &= \left(X^TX + \frac{1}{\gamma}\right)^{-1}(X^TX + Y^k + \frac{1}{\gamma}J^k) \end{aligned}$$

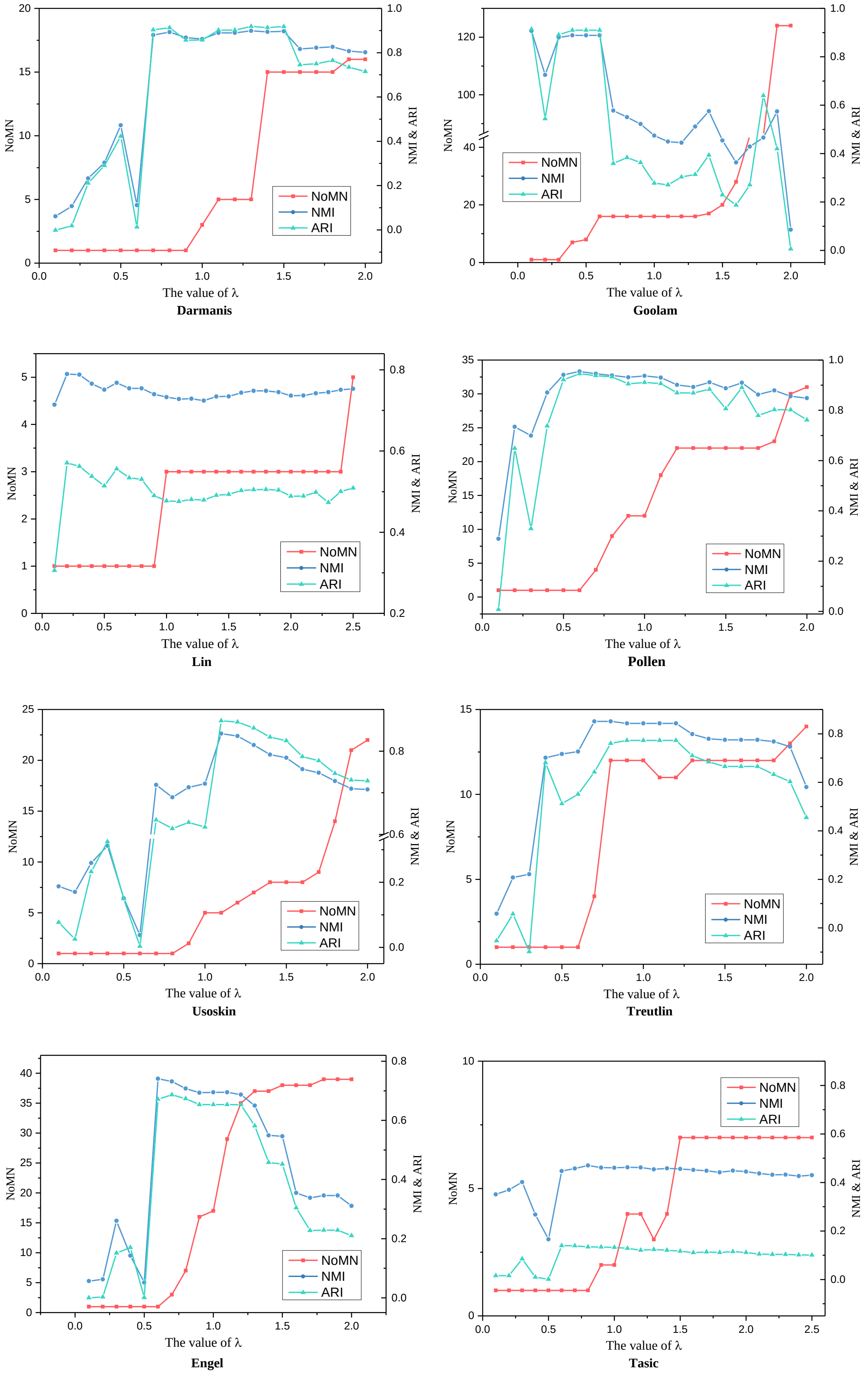
$$\begin{aligned} J^{k+1} &= \mathrm{argmin}_J\ \mathcal{L}_{\frac{1}{\gamma}}(C^{k+1}, J, Y^k) \qquad (S6) \\ &= \mathrm{argmin}_J \left\{\lambda\left||J|\right|_* + \frac{1}{2\gamma}\left||J - C^{k+1}|\right|^2\right\} \\ &= Soft_{\lambda,\gamma}(C^{k+1} - \gamma Y^k) \end{aligned}$$

where $Soft_{\lambda,\gamma}(*)$ is a soft-thresholding operator (Cai *et al.*, 2010), and $Soft_{\lambda,\gamma}(A) = UD_{\lambda,\gamma}(\Sigma)V^T, A = U\Sigma V^T$, $D_{\lambda,\gamma}(\Sigma) = \max(\sigma_{ii} - \lambda * \gamma, 0)$ and σ_{ii} is the diagonal elements of Σ.

B. Analysis of parameters

There are two main parameters, λ and γ, in SinNLRR. We select the proper λ by analyzing the locality of similarity matrix and the number of minimal neighbors (NoMN), which is described in Section 2.2 of main context. The corresponding NoMN, NMI and

ARI with the increment of λ are shown in Fig. S1. In most case, the tipping points of NoMN (larger than 3 for small scale datasets and larger than 1 for datasets with the number of cells larger than 1000) achieve better performance on NMI and ARI.



Darmanis

Goolam

Lin

Pollen

Usoskin

Treutlin

Engel

Tasic

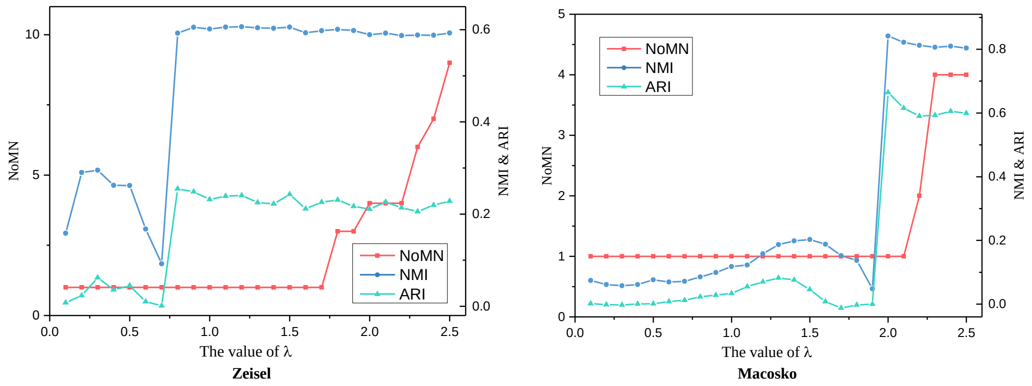


Fig. S1. The corresponding NoMN, NMI and ARI with different values of parameter λ **on ten datasets.** The left Y axis is the value of NoMN, while right Y axis denotes the value of NMI and ARI.

We also another parameter γ (we use 1/γ in the matlab code) and evaluate the 1/γ from 0.2 to 2.0 with increment of 0.2. The corresponding NMI and ARI are show in Fig. S2. The results on the ten datasets show that SinNLRR is not so sensitive to the selection of 1/γ with the proper λ.

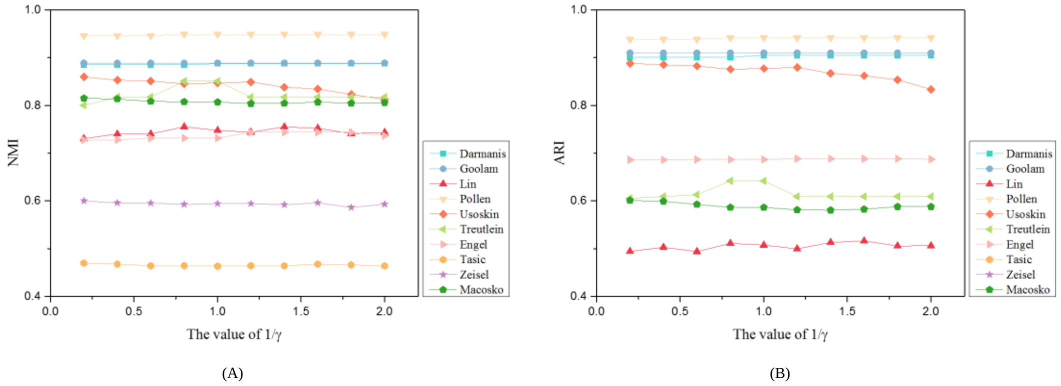


Fig. S2. The corresponding (A) NMI and (B) ARI with different values of parameter γ **on ten datasets.**

C. The estimated cluster number K

Given the normalized Laplacian matrix *L* of learned similarity matrix *C*, we apply *eigengap* (Von Luxburg, 2007) to determine the number of cluster *k* by maximizing the eigenvalues gap $|\lambda_k - \lambda_{k-1}|$, where $\lambda_1 < \lambda_2 \ldots < \lambda_n$ is the eigenvalues of the Laplacian matrix *L*. We run similar methods in SIMLR and MPSSC. Besides, the methods SNN-Cliq and Corr also provided the methods to estimate the number of clusters. The estimated cluster numbers in the ten datasets is shown in Table S1. Corr is not applied on Tasic, Zeisel and Macosko, because it is time-consuming for big datasets. SinNLRR obtains the number of clusters same with true cell types in two datasets (Pollen and Engel) and closest with the true cell types in five datasets (Darmanis, Goolam, Treutlein, Zeisel and Macosko). Although the cluster numbers estimated by these methods are not so consistent with the true cell types, SinNLRR achieves the better performance overall.

Table S1. The estimation of cluster number by five methods.

Dataset	Cell types	SIMLR	MPSSC	SNN-Cliq	Corr	SinNLRR
Darmanis (Darmanis et al., 2015)	8	13	18	17	6	9
Goolam (Goolam et al., 2016)	5	11	15	17	3	4
Lin (Lin et al., 2017)	16	11	26	73	7	10
Pollen (Pollen et al., 2014)	11	19	21	11	3	11
Usoskin (Usoskin, et al., 2015)	4	4	2	29	2	5
Treutlein (Treutlein et al., 2014)	5	10	15	14	3	3
Engel (Engel et al., 2016)	4	3	2	13	2	4
Tasic (Tasic et al., 2016)	48	7	7	36	-	22
Zeisel (Zeisel et al., 2016)	48	10	3	383	-	11
Macosko (Macosko et al., 2015)	39	6	11	548	-	26

C. The analysis of computational complex

In the procedure of obtaining similarity matrix, some operations, such as X^TX and $\left(X^TX+\frac{1}{\gamma}\right)^{-1}$, need to be computed only once. The main computational complexity is brought by the soft threshold function, which actually uses the SVD. The complexity is $O(n^2m+d_\lambda kn^3)$, where *n* is the number of cells, m is the number of genes, d_λ is the number of loops to obtain optimal λ based on NoMN and k is the number of updates in ADMM. In addition, the computational complexity of spectral clustering is $O(n^3)$. The complete computational complexity of SinNLRR is $O(n^2m+d_\lambda kn^3)$. To compare the computational time of these methods, we select four datesets which have different number of cells. The computational times are tested on the same sever (CPU 2.4GHz, 40 cores, 128G). Table S2 summarizes the running times in four datasets. SinNLRR achieves the comparable running times in small datasets while is more time-consuming in large scale datasets. The result shows SinNLRR is not feasible to solve a large number of cells because of the optimal λ searching and ADMM.

Table S2. The computational times (*seconds*) of SC, Corr, SIMLR, NMF, MPSSC and SinNLRR on four datasets. "-" denotes the running time is longer than 24 hours.

Dataset		**SC**	**SNN-Cliq**	**Corr**	**SIMLR**	**NMF**	**MPSSC**	**SinNLRR**
Goolam	(*size=124*)	4.100	1.1532	233.780	2.600	1995.244	3.398	4.570
Usoskin	(*size=622*)	10.499	20.1131	6181.127	8.053	1321.200	15.308	39.721
Tasic	(*size=1727*)	14.004	4886.403	-	62.142	3802.086	289.895	294.281
Macosko	(*size=6418*)	76.159	32063.506	-	1161.528	4178.579	6322.816	18121.312

References

Cai, J. F. *et al*. (2010) A singular value thresholding algorithm for matrix completion[J]. *SIAM Journal on Optimization*, **20**, 1956-1982.

Von-Luxburg, U. (2007). A tutorial on spectral clustering[J]. *Statistics and computing*, **17**, 395-416.

Darmanis, S. *et al*. (2015) A survey of human brain transcriptome diversity at the single cell level[J]. *Proceedings of the National Academy of Sciences*, **112**, 7285-7290.

Goolam, M. *et al*. (2016) Heterogeneity in Oct4 and Sox2 targets biases cell fate in 4-cell mouse embryos[J]. *Cell*, **165**, 61-74.

Lin, C. *et al*. (2017). Using neural networks for reducing the dimensions of single-cell RNA-Seq data[J]. *Nucleic acids research*, **45**, e156-e156.

Pollen, A. A. *et al*. (2014) Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex[J]. *Nature biotechnology*, **32**, 1053.

Usoskin, D. *et al*. (2015). Unbiased classification of sensory neuron types by large-scale single-cell RNA sequencing[J]. *Nature neuroscience*, **18**, 145.

Treutlein, B. *et al.* (2014). Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq[J]. *Nature*, **509**, 371.

Engel, I. *et al*. (2016) Innate-like functions of natural killer T cell subsets result from highly divergent gene programs[J]. *Nature immunology*, **17**, 728.

Tasic, B. *et al.* (2016). Adult mouse cortical cell taxonomy revealed by single cell transcriptomics[J]. *Nature neuroscience*, **19**, 335.

Zeisel, A. *et al.* (2015). Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq[J]. *Science*, **347**, 1138-1142.

Macosko, E. Z. *et al.* (2015). Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets[J]. *Cell*, **161**, 1202-1214.