# Supplementary file in A kernel non-negative matrix factorization framework for single cell clustering

Hao Jiang[a,1], Ming Yi[b,2], Shihua Zhang[c,3]

[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 55, Zhongguancun East Road Beijing 100190, China*

## 1. tSNE with KDCorr-NMF and SIMLR in different data sets

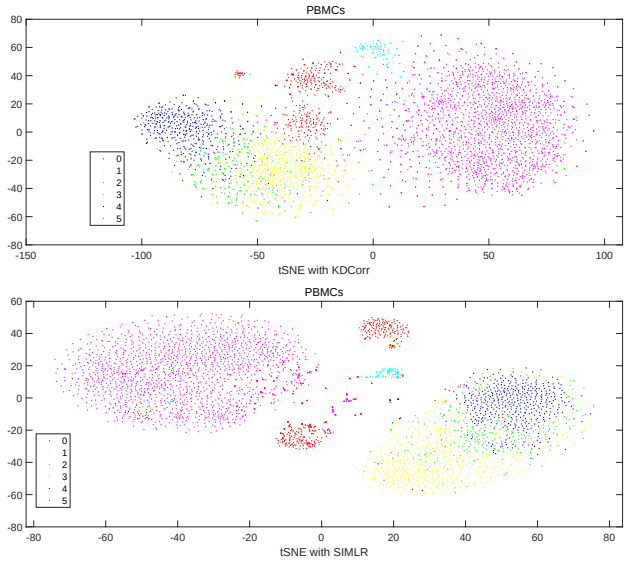


Figure S1: tSNE in PBMCs by KDCorr-NMF and SIMLR

[1]Email:jiangh@ruc.edu.cn
[2]Email:yiming@cug.edu.cn
[3]Email:zsh@amss.ac.cn

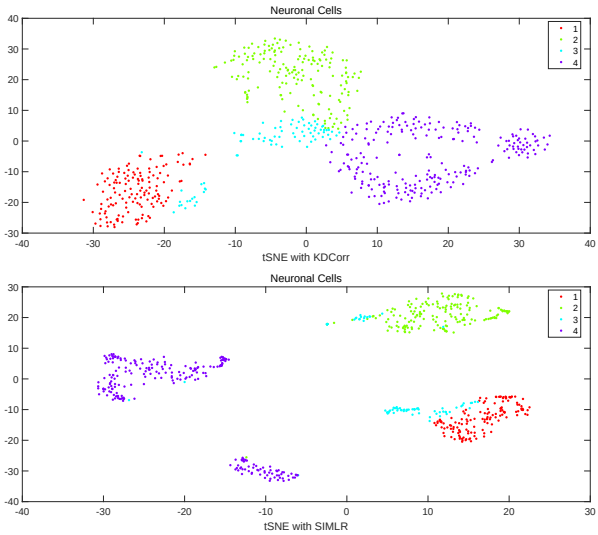


Figure S2: tSNE in Neuronal Cells by KDCorr-NMF and SIMLR

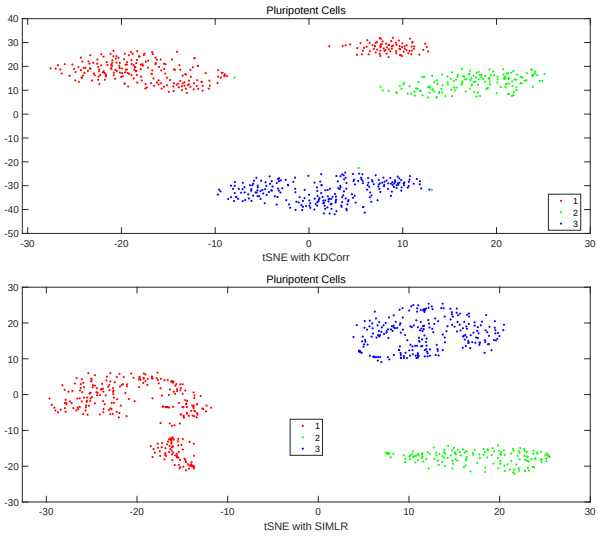


Figure S3: tSNE in Pluripotent Cells by KDCorr-NMF and SIMLR

## 2. Evaluation on different methods

### *2.1. Performance on Corr-NMF clustering*

Table S1: Performance of Corr-NMF for various datasets

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | k | k=13 | k=14 | k=15 | k=16 | k=17 | k=18 | k=19 | k=20 | k=21 | k=22 |
| Neuronal | ARI | 0.6249 | 0.6178 | 0.6216 | 0.6056 | 0.6032 | 0.6209 | 0.6182 | 0.6144 | 0.6182 | 0.6199 |
| | NMI | 0.6383 | 0.6353 | 0.6376 | 0.6498 | 0.6368 | 0.6487 | 0.6450 | 0.6380 | 0.6450 | 0.6446 |
| | k | k=5 | k=6 | k=7 | k=8 | k=9 | k=10 | k=11 | k=12 | k=13 | k=14 |
| Pluripotent | ARI | 0.7317 | 0.7265 | 0.7265 | 0.7265 | 0.7265 | 0.7265 | 0.7265 | 0.7265 | 0.7265 | 0.7265 |
| | NMI | 0.7870 | 0.7842 | 0.7842 | 0.7842 | 0.7842 | 0.7842 | 0.7842 | 0.7842 | 0.7842 | 0.7842 |
| | k | k=31 | k=32 | k=33 | k=34 | k=35 | k=36 | k=37 | k=38 | k=39 | k=40 |
| PBMC | ARI | 0.7634 | 0.7720 | 0.7626 | 0.7543 | 0.7043 | 0.7674 | 0.7688 | 0.7627 | 0.7317 | 0.7631 |
| | NMI | 0.6594 | 0.6660 | 0.6505 | 0.6732 | 0.6370 | 0.6580 | 0.6769 | 0.6586 | 0.6366 | 0.6736 |

### *2.2. Performance of KDCorr-NMF*

### *2.3.* **Performance for RBF-NMF framework**

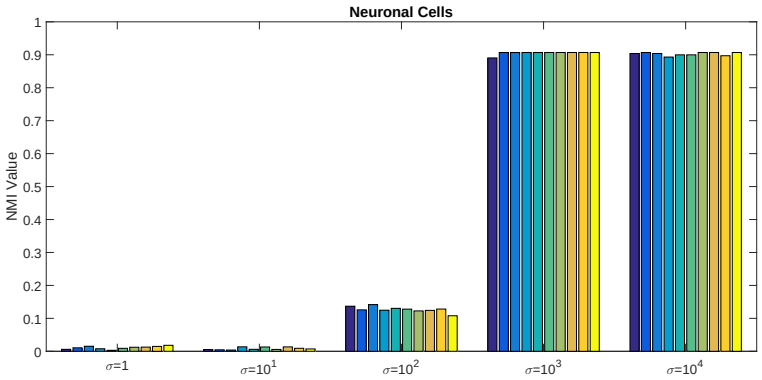


Figure S4: NMI Plot by RBF-NMF framework in Neuronal Cells

Table S2: Performance of KDCorr-NMF for various datasets

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | k | k=13 | k=14 | k=15 | k=16 | k=17 | k=18 | k=19 | k=20 | k=21 | k=22 |
| Neuronal | ARI | 0.9190 | 0.9272 | 0.9345 | 0.9459 | 0.9373 | 0.9373 | 0.9373 | 0.9373 | 0.9373 | 0.9373 |
| | NMI | 0.8928 | 0.9009 | 0.9083 | 0.9215 | 0.9113 | 0.9113 | 0.9113 | 0.9096 | 0.9113 | 0.9113 |
| | k | k=5 | k=6 | k=7 | k=8 | k=9 | k=10 | k=11 | k=12 | k=13 | k=14 |
| Pluripotent | ARI | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| | NMI | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| | k | k=31 | k=32 | k=33 | k=34 | k=35 | k=36 | k=37 | k=38 | k=39 | k=40 |
| PBMC | ARI | 0.8814 | 0.8783 | 0.8823 | 0.8893 | 0.8731 | 0.8818 | 0.8766 | 0.8798 | 0.8782 | 0.8799 |
| | NMI | 0.7624 | 0.7684 | 0.7607 | 0.7851 | 0.7652 | 0.7643 | 0.7673 | 0.7637 | 0.7703 | 0.7645 |

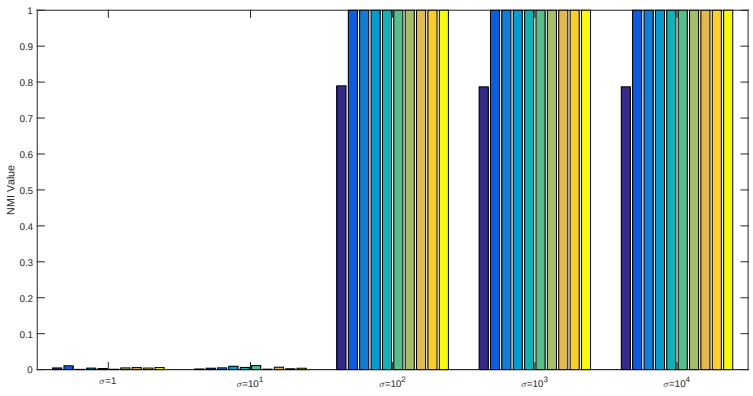


Figure S5: NMI Plot by RBF-NMF framework in Pluripotent Cells

## 3. Comparison in 4 different perspectives

Apart from comparisons with state-of-the-art algorithms, we also test the robustness of our kernel non-negative framework in 4 different perspectives.

- Comparison in Kernel Construction step.

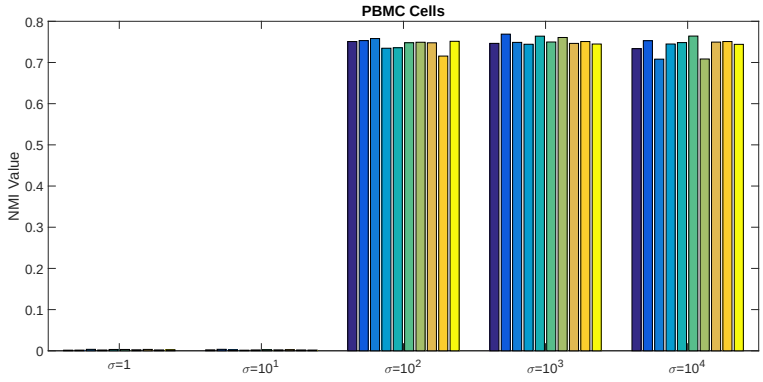


Figure S6: NMI Plot by RBF-NMF framework in PBMC Cells

We here tested the performance of kernel non-negative matrix factorization with the most widely used kernel: RBF kernel. RBF kernel as a kernel function is used in many different settings hence also can be used for kernel non-negative matrix factorization framework. However the selection of parameter $\sigma$ is also a problem need to be addressed. We show ARI results of RBF kernel non-negative matrix factorization framework under different selection of $\sigma$ parameters(The NMI results are attached in supplementary files Fig.S4-S6). In each figure, different bars correspond to different parameters of $k$. For neuronal cell data set, the $k$ parameter is from 13 to 22, and for pluripotent cell data set, the $k$ parameter is from 5 to 14; for pbmc data set, the $k$ parameter is from 31 to 40. We can have following discoveries.

- Different data sets fit for different $\sigma$ values. For example, in neuronal cells, RBF-NMF is unsatisfactory when $\sigma = 1, 10, 10^2$, and when $\sigma = 10^3, 10^4$, the performance tends to be robust and stable. In pluripotent cells, RBF-NMF is unsatisfactory when $\sigma = 1, 10$, and when $\sigma = 10^2, 10^3, 10^4$, the performance tends to be robust and stable.

– For a fixed value of $\sigma$ in each data set, the performance tends to be relatively stable for different values of $k$ parameters.

– The stability of performance differs in different datasets. In neuronal cell data set, the performance when $\sigma = 10^3$ and $\sigma = 10^4$ is similar with each other. Both of them are stable over different values of $k$ parameters. In pluripotent cell data set, the performance is similar when $\sigma = 10^2$, $\sigma = 10^3$ and $\sigma = 10^4$. Except for $k = 5$, RBF-NMF is not satisfactory over these $\sigma$ parameters. When $k$ ranges from 6 to 14, RBF-NMF can yield 100% in ARI values. In PBMC data set, the performance when $\sigma = 10^3$ is the most stable. However, when $\sigma = 10^4$, it is less stable instead.

Since $\sigma$ in RBF-NMF is a parameter that should be determined in the beginning stage, there is no good way to choose proper $\sigma$.

- Comparison with extension of non-negative matrix factorization taking into consideration on the geometric structure: Graph regularized non-negative matrix factorization.

  It solves the non-negative matrix factorization problem by incorporating a geometrically based regularizer. It constructs an affinity graph to encode the geometrical information and the optimization problem can then be described as

$$\min_{W\geq 0, H\geq 0} L(W,H) = \min_{W\geq 0, H\geq 0} \mathrm{tr}((V-WH)(V-WH)^T) + \lambda \mathrm{tr}(HL_V H^T) \tag{1}$$

  where $L_V$ is the graph laplacian. Using Lagrange multiplier method, the final update rules become

$$W_{ij} \leftarrow W_{ij} \frac{(VH^T)_{ij}}{(WHH^T)_{ij}}$$

$$H_{ij} \leftarrow H_{ij} \frac{(W^T V + \lambda H W_V^T)_{ij}}{(W^T WH + \lambda H D_V^T)_{ij}}$$

  From the performance of graph regularized non-negative matrix factorization framework in single cell clustering (in Table S3), we can see that the

Table S3: Performance of Graph regularized non-negative matrix factorization for various datasets

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | k | k=13 | k=14 | k=15 | k=16 | k=17 | k=18 | k=19 | k=20 | k=21 | k=22 |
| Neuronal | ARI | 0.5545 | 0.2570 | 0.5551 | 0.4974 | 0.5636 | 0.2871 | 0.3313 | 0.2463 | 0.6200 | 0.5622 |
| | NMI | 0.5782 | 0.3311 | 0.5625 | 0.5623 | 0.5926 | 0.3821 | 0.3782 | 0.3061 | 0.6221 | 0.6104 |
| | k | k=5 | k=6 | k=7 | k=8 | k=9 | k=10 | k=11 | k=12 | k=13 | k=14 |
| Pluripotent | ARI | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| | NMI | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| | k | k=31 | k=32 | k=33 | k=34 | k=35 | k=36 | k=37 | k=38 | k=39 | k=40 |
| PBMC | ARI | 0.7727 | 0.6929 | 0.7217 | 0.7166 | 0.7093 | 0.7251 | 0.7257 | 0.7124 | 0.7217 | 0.7094 |
| | NMI | 0.6575 | 0.5853 | 0.6200 | 0.5995 | 0.5944 | 0.6066 | 0.6036 | 0.5965 | 0.6069 | 0.5954 |

performance is not satisfactory except in pluripotent cell data set. And the performance in neuronal cell data set and PMBC data set is unstable when values of $k$ vary. The fluctuations of adjusted rand index and normalized mutual information in neuronal cell data set are obvious. Hence we can conclude that the affinity graph encoding the geometrical information in graph regularized non-negative matrix factorization framework may not always work in all the situations.

- Comparison with symmetric non-negative matrix factorization algorithm.

  Since KDCorr kernel is symmetric, symmetric non-negative matrix factorization can be applied. It is a special case of NMF, it aims to solve the following minimization problem

$$\min_{W_\phi \geq 0} L(W_\phi) = \min_{W_\phi \geq 0} ||K - W_\phi W_\phi^T||_F^2 \tag{2}$$

  The alternative non-negative least square algorithm is proposed to solve the alternative formulation of symmetric non-negative matrix factorization problem. The framework of the algorithm is in the following.

Table S4: Performance of symmetric non-negative matrix factorization for various datasets

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | k | k=13 | k=14 | k=15 | k=16 | k=17 | k=18 | k=19 | k=20 | k=21 | k=22 |
| Neuronal | ARI | 0.7507 | 0.5987 | 0.7676 | 0.9066 | 0.5581 | 0.8219 | 0.5472 | 0.8185 | 0.8364 | 0.8047 |
| | NMI | 0.7609 | 0.7183 | 0.7745 | 0.8666 | 0.6812 | 0.8000 | 0.6791 | 0.7679 | 0.8192 | 0.8006 |
| | k | k=5 | k=6 | k=7 | k=8 | k=9 | k=10 | k=11 | k=12 | k=13 | k=14 |
| Pluripotent | ARI | 0.7265 | 0.7265 | 0.7265 | 0.7265 | 0.7265 | 0.7265 | 0.7265 | 0.7239 | 0.7239 | 0.7240 |
| | NMI | 0.7842 | 0.7842 | 0.7842 | 0.7842 | 0.7842 | 0.7842 | 0.7842 | 0.7764 | 0.7764 | 0.7829 |
| | k | k=31 | k=32 | k=33 | k=34 | k=35 | k=36 | k=37 | k=38 | k=39 | k=40 |
| PBMC | ARI | 0.4725 | 0.7320 | 0.4624 | 0.4211 | 0.4542 | 0.7177 | 0.4354 | 0.4182 | 0.6974 | 0.4476 |
| | NMI | 0.6017 | 0.6519 | 0.5873 | 0.5835 | 0.5949 | 0.6474 | 0.5884 | 0.5796 | 0.6372 | 0.5869 |

**Algorithm 1** Framework of alternative non-negative least square algorithm for symmetric non-negative matrix factorization problem: $\min_{W,H\geq 0} ||K - WH^T||_F^2 + \lambda||W - H||_F^2$

**Input:** The number of data points n, number of clusters k, kernel matrix K, regularization parameter $\lambda > 0$, and tolerance parameter $0 < \mu \ll 1$;

**Output:** H;

repeat;

W $\leftarrow$ H;

Solve an NLS problem H $\leftarrow \text{argmin}_{H\geq 0}|| \begin{pmatrix} W \\ \sqrt{\lambda} I_k \end{pmatrix} H^T - \begin{pmatrix} K \\ \sqrt{\lambda} W^T \end{pmatrix} ||_F$;

until $|| \bigtriangledown^P g(W,H)||_F \leq \mu|| \bigtriangledown^P g(W^{(0)},H^{(0)})||_F$

**Return:** $H$;

We therefore test the performance of symmetric non-negative matrix factorization with KDCorr kernel over the considered datasets, see Table S4. It is obvious to see that symmetric non-negative matrix factorization with KDCorr kernel can not compete with KDCorr-NMF framework. And the

performance is unstable for different values of $k$ in symmetric non-negative matrix factorization with KDCorr kernel.

- Multiplicative update algorithm

  It is a fundamental algorithm for solving non-negative matrix factorization problems. Through differentiation on the objective function

$$\min_{W_\phi \geq 0, H_\phi \geq 0} L(W_\phi, H_\phi) = \min_{W_\phi \geq 0, H_\phi \geq 0} \sum_{i=1}^{n} \sum_{j=1}^{n} (K_{ij} - (w_i^\phi)^T h_j^\phi)^2$$

  with respect to $W_\phi, H_\phi$ and selecting appropriate update parameters we can have the update rules:

$$(H_\phi)_{ia} \leftarrow (H_\phi)_{ij} \frac{(W_\phi^T K)_{ia}}{(W_\phi^T W_\phi H_\phi + \delta)_{ia}}$$

$$(W_\phi)_{aj} \leftarrow (W_\phi)_{ij} \frac{(K H_\phi^T)_{aj}}{(W_\phi H_\phi H_\phi^T + \delta)_{aj}}$$

  where $\delta$ is a tuning parameter (almost 0) for avoiding division by zero.

The performance of multiplicative update algorithm is shown in Table S5.

Table S5: Performance of Multiplicative update algorithm with nonnegative matrix factorization 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 | -0.0011 | -0.0049 | -0.0051 | 0.0024 | 0.0012 | 0.0084 | -0.0036 | -0.0015 | -0.0016 | -0.0062 |
| | NMI | 0.0193 | 0.0036 | 0.0083 | 0.0027 | 0.0063 | 0.0105 | 0.0119 | 0.0026 | 0.0044 | 0.0176 |
| Pluripotent | k | k=5 | k=6 | k=7 | k=8 | k=9 | k=10 | k=11 | k=12 | k=13 | k=14 |
| | ARI | 0 | -0.0026 | 0.0012 | 0.0080 | 0.0024 | 0 | 0.0018 | 0.0043 | 0.0027 | 0.0028 |
| | NMI | 0.0003 | 0.0125 | 0.0091 | 0.0094 | 0.0081 | 0.0055 | 0.0081 | 0.0136 | 0.0099 | 0.0237 |
| PBMC | k | k=31 | k=32 | k=33 | k=34 | k=35 | k=36 | k=37 | k=38 | k=39 | k=40 |
| | ARI | 0.5502 | 0.5758 | 0.5187 | 0.5561 | 0.5376 | 0.5031 | 0.5696 | 0.5768 | 0.5710 | 0.5665 |
| | NMI | 0.4713 | 0.4908 | 0.4830 | 0.4847 | 0.4708 | 0.4803 | 0.4864 | 0.4856 | 0.5023 | 0.4823 |

From the performance of the multiplicative update algorithm with non-negative matrix factorization in TableS5, Among the 3 data sets, the performance in PBMC data is better than that in Neuronal and Pluripotent data. Overall, we can see that the performance is still not satisfactory. The possible reason may be that when the limit point of multiplicative update algorithm lies on the boundary of the feasible region, its stationarity cannot be determined. Hence the multiplicative update algorithm lacks sound optimization properties.