Supplementary materials of "Spectral clustering based on learning similarity matrix"

Seyoung Park and Hongyu Zhao

Department of Biostatistics, School of Public Health,

Yale University, New Haven, CT, 06511, USA

November 17, 2017

These supplementary materials include additional ADMM procedures, theoretical details of convergence of the algorithm, and additional figures.

A Algorithm

A.1 ADMM for Step 2

It is difficult to compute the optimizer directly by minimizing (9) because the penalty function is not easily handled with the constraints. Following Lu *et al.* (2016a), we reparameterize (9) by introducing a new parameter $Q := P$. Then, the minimization of (9) is equivalent to the following constrained

optimization problem: for $\bar{S} := \sum_l w_l^{i+1} \bar{G}^{(l)}$,

$$\min_{P,Q} c\|P\|_F^2 - \langle \bar{S}, P\rangle + \lambda \|P\|_{1,\bar{P}} \quad \text{s.t.} \;\; P = Q, \quad Q \in \mathrm{CH}(n, C). \tag{S1}$$

By the augmented Lagrangian method, we solve

$$\min_{P,Q,\Gamma} c\|P\|_F^2 - \langle \bar{S}, P\rangle + \lambda \|P\|_{1,\bar{P}} + \langle \Gamma, P - Q\rangle + \frac{\eta}{2}\|P - Q\|^2 \text{ s.t. } Q \in \mathrm{CH}(n, C), \tag{S2}$$

where the dual variables Γ_{jk} are the Lagrangian multipliers and $\eta > 0$ is the penalty parameter. Let $F(P, Q, \Gamma)$ be the objective function in (S2). We solve for the minimizer iteratively using the following steps at the i^{th} iteration until it converges.

$$\begin{aligned}
&\textbf{Update } P\textbf{:} \quad P^{i+1} = \operatorname*{argmin}_{P} F(P, Q^i, \Gamma^i) \\
&\textbf{Update } Q\textbf{:} \quad Q^{i+1} = \operatorname*{argmin}_{Q; Q \in \mathrm{CH}(n,C)} F(P^{i+1}, Q, \Gamma^i) \\
&\textbf{Update } \Gamma\textbf{:} \quad \Gamma^{i+1} = \Gamma^i + \eta(P^{i+1} - Q^{i+1}).
\end{aligned}$$

Since (9) is convex, the iterates of ADMM converge to an optimal point with the convergence rate $O(1/K)$ (He and Yuan, 2012; Monteiro and Svaiter, 2013; Lu *et al.*, 2016b), where K is the number of iterations. The following are details in updating P, Q, and Γ.

Details in updating P:

Let $t_{jk} = \bar{S}_{jk} - \Gamma_{jk} + \eta Q_{jk}$. Then, by KKT condition,

$$\begin{aligned}
P_{jk} &= (t_{jk} - \lambda \bar{p}_{jk})(2c + \eta)^{-1} \quad \text{if} \quad t_{jk} > \lambda \bar{p}_{jk} \\
P_{jk} &= (t_{jk} + \lambda \bar{p}_{jk})(2c + \eta)^{-1} \quad \text{if} \quad t_{jk} < -\lambda \bar{p}_{jk} \\
P_{jk} &= 0 \quad \text{if} \quad |t_{jk}| \le \lambda \bar{p}_{jk}.
\end{aligned}$$

Details in updating Q:

Minimizing $F(P^{i+1}, Q, \Gamma^i)$ with respect to Q is equivalent to solve

$$\min_{Q} \|Q - P^{i+1} - \Gamma/\eta\|_F^2 \quad \text{s.t.} \;\; Q \in CH(n, C).$$

Let $T_3 = P^{i+1} + \Gamma/\eta$. Let $B = \frac{T_3 + T_3^T}{2}$ and $B = U\,\mathrm{diag}(u)U^T$ be the spectral decomposition of B, where $u \in \mathbb{R}^n$ and $\mathrm{diag}(u)$ is the n by n diagonal matrix with $(\mathrm{diag}(u))_{ii} = u_i$. Then, by using Theorem 2 in Lu *et al.* (2016a), $P = U\,\mathrm{diag}(x^*)U^T$, where x^* is the solution to

$$\min_{x} \|x - u\|^2, \text{ s.t. } 0 \le x \le 1, \quad 1^T x = C,$$

which can be efficiently solved using existing algorithm.

Stopping Criterion

We use a termination criterion that $P^{i+1} - P^i$, $Q^{i+1} - Q^i$, and $\Gamma^{i+1} - \Gamma^i$ are nearly zero, i.e.,

$$\|P^{i+1} - P^i\|_F^2 + \|Q^{i+1} - Q^i\|_F^2 + \|\Gamma^{i+1} - \Gamma^i\|_F^2 \le \epsilon$$

for some small $\epsilon > 0$. In our implementation, we set $\epsilon = 0.005$.

A.2 ADMM for Step 3

It is difficult to compute the optimizer directly by minimizing (6) in the main paper because the penalty functions are not separable in $X_{i,\cdot}$. We introduce a new set of parameters $\Theta_{jk} = X_{j,\cdot} - X_{k,\cdot}$ for $1 \le j < k \le n$. The minimization of (6) in the main paper is equivalent to the constrained

optimization problem

$$\begin{aligned}&\min_{X,\Theta} \|X-\hat{P}\|_F^2+\mu\sum_{j<k}\frac{\|\Theta_{j,k}\|_2}{\|\hat{P}_{j,\cdot}-\hat{P}_{k,\cdot}\|_2}\\&\text{s.t. } \Theta_{jk}=X_{j,\cdot}-X_{k,\cdot},\ X\in \mathrm{CH}(n,C),\end{aligned}\tag{S3}$$

where $\Theta=\{\Theta_{jk}:\ j<k\}$. By the augmented Lagrangian, we solve

$$\begin{aligned}&\min_{X,\Theta,\gamma} \|X-\hat{P}\|_F^2+\mu\sum_{j<k}\frac{\|\Theta_{jk}\|_2}{\|\hat{P}_{j,\cdot}-\hat{P}_{k,\cdot}\|_2}+\\&\sum_{j<k}\langle\gamma_{jk},X_{j,\cdot}-X_{k,\cdot}-\Theta_{jk}\rangle+\frac{\eta}{2}\sum_{j<k}\|X_{j,\cdot}-X_{k,\cdot}-\Theta_{jk}\|^2\\&\text{s.t. }\ X\in \mathrm{CH}(n,C),\end{aligned}\tag{S4}$$

where the dual variables γ_{jk} are the Lagrangian multipliers and $\eta>0$ is the penalty parameter. Since (4) is convex, ADMM guarantees convergences of the iterates.

Let $F(X,\Theta,\gamma)$ be the objective function in (S4). We iteratively solve for the minimizer at the i^{th} iteration until it converges:

$$\begin{aligned}&\textbf{Update } X\textbf{:}\quad X^{i+1}=\operatorname*{argmin}_{X\in\mathrm{CH}(n,C)} F(X,\Theta^i,\gamma^i)\\&\textbf{Update } \Theta\textbf{:}\quad \Theta^{i+1}=\operatorname*{argmin}_{\Theta} F(X^{i+1},\Theta,\gamma^i)\\&\textbf{Update } \gamma_{jk}\textbf{:}\quad \gamma_{jk}^{i+1}=\gamma_{jk}^i+\eta(X_{j,\cdot}^{i+1}-X_{k,\cdot}^{i+1}-\Theta_{jk}^{i+1})\end{aligned}$$

Details in updating X:

Let $\Delta\in\mathbb{R}^{(n^2-n)/2\times n}$ such that $\Delta_{jk,\cdot}=e_j-e_k$ for $1\le j<k\le n$. It holds that $\Delta^T\Delta=nI_n-1_{n\times n}$, where $1_{n\times n}$ is the n by n matrix with all entries one and e_j is the 1 by n vector with the jth component equal to one and all other components equal to zero. Let $T_4=(\eta\Delta^T\Delta+2I)^{-1}(2\hat{P}+\eta\Delta^T\Theta^i-\Delta^T\gamma^i)$. Let $T_5=(T_4+T_4^T)/2$ and $U\,\mathrm{diag}(v)U^T$ be the spectral decomposition of

T_5, where $v \in \mathbb{R}^n$. Then, by using Theorem 2 in Lu *et al.* (2016a), $X^{i+1} = U\,\mathrm{diag}(\lambda^*)U^T$, where λ^* is the solution to

$$\min_{\lambda} \|\lambda - v\|^2, \text{ s.t. } 0 \leq \lambda \leq 1, \quad 1^T\lambda = C.$$

Details in updating Θ:

Let $d\hat{P}_{j,k} := \|\hat{P}_{j,\cdot} - \hat{P}_{k,\cdot}\|_2$ for $j < k$, and $t_{jk} = \frac{\gamma^i_{jk}}{\mu} + \frac{\eta}{\mu}(X^{i+1}_{j,\cdot} - X^{i+1}_{k,\cdot})$. By KKT condition and simple calculation, we can update such that if $\|t_{jk}\| \leq 1/d\hat{P}_{j,k}$, then $\Theta_{jk} = 0$, else if $\|t_{jk}\| > 1/d\hat{P}_{j,k}$,

$$\Theta_{jk} = \frac{\mu(\|t_{jk}\| - 1/d\hat{P}_{j,k})}{\eta\|t_{jk}\|} t_{jk}.$$

Stopping Criteria

We use a termination criterion that $X^{i+1} - X^i$, $\Theta^{i+1} - \Theta^i$, and $\gamma^{i+1} - \gamma^i$ are nearly zero, i.e.,

$$\|P^{i+1} - P^i\|_F^2 + \|\Theta^{i+1} - \Theta^i\|_F^2 + \|\gamma^{i+1} - \gamma^i\|_F^2 \leq \epsilon$$

for some small $\epsilon > 0$. In our implementation, we set $\epsilon = 0.005$.

Since (6) in the main paper is convex, the solution by ADMM is optimal with the convergence rate $O(1/K)$ (He and Yuan, 2012; Monteiro and Svaiter, 2013; Lu *et al.*, 2016b), where K is the number of iterations.

B Proof of Proposition

Proposition S1. *Let $G(P, W)$ be the objective function of* (5) *in the main paper. Then the iterates (P^i, W^i) converge to a global minimum point of G*

with

$$G(P^{i-1}, W^{i-1}) - G(P^i, W^i) \geq \frac{\rho}{4}\|W^i - W^{i-1}\|_F^2,$$

where the objective value $G(P^i, W^i)$ is monotonically decreasing.

Proof of Proposition S1. Fix $i \geq 1$. At the ith iteration of the iterative algorithm presented in (8) and (9) in the main paper, we first consider (8). Note that solving (8) is equivalent to solve

$$\min_W \langle \sum_l w_l \bar{G}^{(l)}, -P^i \rangle + \rho \sum_l w_l \log w_l \quad \text{s.t.} \sum_l w_l = 1,\ w_l \geq 0. \tag{S5}$$

We can easily show that optimizer w_l's of (S6) satisfy $w_l \geq 0$ for each l without the constraint $w_l \geq 0$. Hence, by the Lagrangian method, the above is equivalent to solve

$$\min_W \langle \sum_l w_l \bar{G}^{(l)}, -P^i \rangle + \rho \sum_l w_l \log w_l + \lambda(1 - \sum_l w_l) \tag{S6}$$

for some $\lambda > 0$. Let $H_1(W) = \langle \sum_l w_l \bar{G}^{(l)}, -P^i \rangle + \rho \sum_l w_l \log w_l$ and $H_2(W) = H_1(W) + \lambda(1 - \sum_l w_l)$. Then

$$\frac{\partial H_2(W)}{\partial w_l} = \langle \bar{G}^{(l)}, -P^i \rangle + \rho(1 + \log w_l) - \lambda, \quad \frac{\partial^2 H_2(W)}{\partial w_l \partial w_k} = 1_{\{l=k\}} \cdot \rho / w_l.$$

Since $W^{i+1} = \{w_l^{i+1}\}_l$ is solution to (S6), we have

$$\begin{aligned} H_2(W^i) &\geq H_2(W^{i+1}) + \frac{\rho}{2}(W^{i+1} - W^i)^T \operatorname{diag}(W^{i+1})^{-1}(W^{i+1} - W^i) \\ &\geq H_2(W^{i+1}) + \frac{\rho}{2}\|W^{i+1} - W^i\|^2. \end{aligned}$$

Since $H_1(W^i) = H_2(W^i)$ and $H_1(W^{i+1}) = H_2(W^{i+1})$, we have

$$H_1(W^i) - H_1(W^{i+1}) \geq \frac{\rho}{2}\|W^{i+1} - W^i\|^2,$$

hence for the G, which is the objective function in (5) of the main paper,

$$G(P^i, W^i) - G(P^i, W^{i+1}) \geq \frac{\rho}{2}\|W^{i+1} - W^i\|^2. \tag{S7}$$

Now consider the ADMM algorithm solving (9) in the main paper, i.e., at the ith iteration,

$$P^{i+1} = \underset{P:\ P \in \mathrm{CH}(n,C)}{\operatorname{argmin}} G(P, W^{i+1}). \tag{S8}$$

Since (S8) is convex in P, the iterates $\{P_j^{i+1}\}_{j\geq 1}$ of the ADMM converge to an optimal point of (9) (Boyd *et al.*, 2011; He and Yuan, 2012), i.e., $P_j^{i+1} \to P_*^{i+1}$ as $j \to \infty$, where P_*^{i+1} is the optimal point of (S8). Since $P^{i+1} = P_j^{i+1}$ for large enough j, we have

$$G(P^i, W^{i+1}) - G(P^{i+1}, W^{i+1}) \geq\ -\frac{\rho}{4}\|W^{i+1} - W^i\|^2. \tag{S9}$$

Combining (S7) and (S9), we have

$$G(P^i, W^i) - G(P^{i+1}, W^{i+1}) \geq \frac{\rho}{4}\|W^{i+1} - W^i\|^2,$$

that is, the objective value $G(P^i, W^i)$ is monotonically decreasing until convergence.

Since both (8) and (9) in the main paper are strictly convex due to $c > 0$, (8) and (9) have unique global minimizers, respectively. By Theorem 4.1 of Tseng (2001), the convergence of the proposed biconvex algorithm is achieved. □

C Time complexity of the algorithm

The computational complexity of the algorithm is $O(Kn^3)$, where n is the number of data points and K is the number of iterations. In the experiments,

K is less than 20. Note that the traditional spectral clustering algorithm has the complexity $O(n^3)$. The proposed algorithm is still fast for single-cell data, since n is relatively small compared with the number of variables (genes).

Most of the simulations and scRNA-seq applications were implemented on an Apple MacBook Pro (2.7 GHz, 8 GB of memory) using the MATLAB 2016b. However, certain computational or memory-intensive steps (e.g. larger-scale data sets) were run on the computing cluster (6 CPUs, 800 GB of memory).

D Evaluation metrics

We use the following three performance metrics to evaluate the consistency between the obtained clustering and the true labels: Normalized Mutual Information (NMI) (Strehl and Ghosh, 2003), Purity (Wagner and Wagner, 2007a), and Adjusted Rand Index (ARI) (Wagner and Wagner, 2007b). Given two clustering results U and V on a set of n data points with C_U and C_V clusters, respectively, the mutual information NMI is defined as

$$\mathrm{NMI}(U,V) = \frac{\sum_{p=1}^{C_U}\sum_{q=1}^{C_V}|U_p \cap V_q| \log \frac{n|U_p \cap V_q|}{|U_p| \times |V_q|}}{\max\left(-\sum_{p=1}^{C_U}|U_p| \log \frac{|U_p|}{n}, -\sum_{q=1}^{C_V}|V_q| \log \frac{|V_q|}{n}\right)},$$

where the numerator is the mutual information between U and V, and the denominator represents the entropy of the clustering U and V. For Purity, each identified cluster is assigned to the one which is most frequent in the cluster, and then the accuracy of this assignment is computed by counting the number of correctly assigned samples divided by the number n:

$$\mathrm{Purity}(U,V) = \frac{\sum_p \max_q |U_p \cap V_q|}{n}.$$

The ARI depends on the following four quantities: a_{uv}, the number of objects in a pair that are placed in the same group in U and V; a_u, the number of objects in a pair that are placed in the same group in U but in different groups in V; a_v, the number of objects in a pair that are placed in the same group in V but in different groups in U; a, the number of objects in a pair that are placed in the different group in U and V. The ARI is

$$\mathrm{ARI}(U,V) = \frac{\binom{n}{2}(a_{uv}+a) - [(a_{uv}+a_u)(a_{uv}+a_v)+(a_v+a)(a_u+a)]}{\binom{n}{2}^2 - [(a_{uv}+a_u)(a_{uv}+a_v)+(a_v+a)(a_u+a)]}.$$

Note that the NMI and Purity take on values between 0 and 1, but ARI can yield negative values. These metrics measure the concordance of two clustering results such that higher value refers to higher concordance with true labels.

E Simulation models

E.1 First simulation model

We generate the simulated data as follows:

(A1): Generate C points in the 2-dimensional latent space to create a circle, each of which is considered to be the center of one cluster:

$$O_l = d \times (\cos(2l\pi/C), \sin(2l\pi/C)) \quad \text{for } l = 1, \cdots, C.$$

The n_l points are generated by adding independent noises to the center O_l:

$$\widetilde{Z}_i^{(l)} \;=\; O_l^T + (Z_{l,i}^{(1)}, Z_{l,i}^{(2)})^T \quad \text{for } i = 1, \cdots, n_l, \tag{S10}$$

where $Z_{l,i}^{(1)}$ and $Z_{l,i}^{(2)}$ are i.i.d. $N(\mu, \sigma_l^2)$ random variables. Let $\widetilde{Z}_l = [\widetilde{Z}_1^{(l)}, \cdots, \widetilde{Z}_{n_l}^{(l)}]^T \in \mathbb{R}^{n_l \times 2}$, and $\widetilde{Z} = [\widetilde{Z}_1^T, \cdots, \widetilde{Z}_C^T]^T \in \mathbb{R}^{n \times 2}$, where $n = \sum_l n_l$.

(A2): Project the data $\widetilde{Z}$ to a p-dimensional space through a projection matrix $P \in \mathbb{R}^{2 \times p}$, where $P_{ij} \sim \text{unif}(0, 1)$ for $1 \leq j \leq q$ and $P_{ij} = 0$ for $j > q$. The $X = \widetilde{Z}P \in \mathbb{R}^{n \times p}$ represents gene expression data, where p corresponds to the number of genes.

(A3): Simulate a noisy gene expression matrix X' by adding independent Gaussian noise: $\quad X'_{ij} = X_{ij} + e_{ij}, \quad$ where $e_{ij} \sim N(0, \sigma^2)$.

(A4): Each entry X'_{ij} is independently observed with probability $1-\exp(-\gamma X'_{ij})$: we observe $Y \in \mathbb{R}^{n \times p}$, where

$$Y_{ij} = \begin{cases} X'_{ij} & \text{with probability } 1 - \exp(-\gamma X'_{ij}) \\ 0 & \text{with probability } \exp(-\gamma X'_{ij}). \end{cases}$$

We use $\gamma \in \{0.01, 0.006\}$. In (A1), we construct five clusters such that the centers of five clusters form a circle. One example of points in the latent space generated in (A1) is shown in Figure S1. In (A4), we introduce dropout events following Pierson and Yau (2015) and Wang *et al.* (2017). We fix $n = 250$, $p = 500$, $q = 50$, $C = 5$, $d = 10$, $\sigma^2 = 1$, and $\sigma_l = 1$. Note that d in (A1) controls distances between samples in different clusters and σ_l controls a density of samples within cluster. The σ^2 in (A3) represents the extrinsic noise from the environment outside the cell.

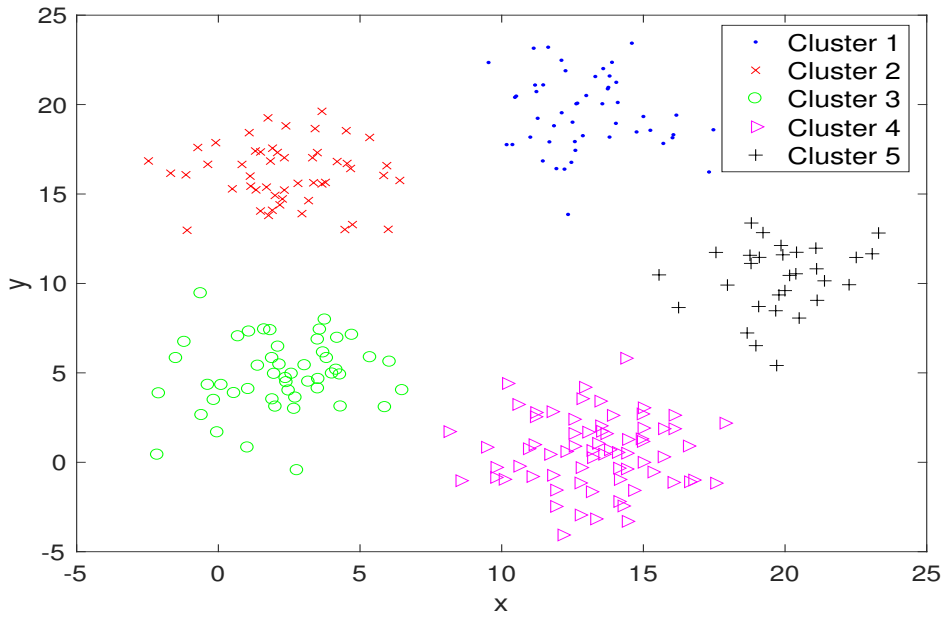

Figure S1: Simulated latent data set $\widetilde{Z}$ generated in (A1) with 5 clusters in the 2-D. We observe that the centers of the five clusters form a circle.

E.2 Second simulation model

The second simulation model are based on sparse Gaussian mixture model. We generate the simulated data as follows:

(A1): Generate elements of $\tilde{B} \in \mathbb{R}^{C\times q}$ as left singular matrix of i.i.d. standard gaussian random matrix. We get $B \in \mathbb{R}^{C\times p}$ as $B = [\sigma\tilde{B}, 0_{C\times(p-q)}]$.

(A2): Generate the cluster label $z_i \in [C]$ of the ith sample by random assignment to one group. Then generate membership matrix $Z \in \mathbb{R}^{n\times C}$ with $Z_{ij} = 1(z_i = j)$.

(A3): Generate data matrix $X = ZB + W$, where W is standard Gaussian noise matrix.

(A4): Each entry X_{ij} is independently observed with probability $1-\exp(-\gamma X'_{ij})$, and we observe $Y \in \mathbb{R}^{n\times p}$.

The (A1)-(A2) are based on sparse Gaussian mixture model, where σ controls the signal to noise ratio. To distinguish different cell types, it is likely that only a few genes are informative, and non-informative and highly noisy genes can increase the difficulty of identifying cell types. Under this context, in the simulation models, we only use q of the p attributes to distinguish the clustering labels. We set $n = 500$ $p = 1000$, $q = 50$, $C = 10$, $\sigma = 5$, and $\gamma \in \{0.6, 0.1\}$.

F Real data

We collected nine scRNA-seq data sets representing several types of dynamic processes such as cell differentiation, cell cycle, and response upon external stimulus. Each scRNA-seq data contains cells for which the labels were known a priori or validated in the respective studies. The characteristics of the nine data sets are summarized as follows:

- Pollen: 249 single cells from 11 populations using microfluidics, including neural cells and blood cells. The 11 clusters in the data set were from different sources (CRL-2338, CRL-2339, K562, BJ, HL60, hiPSC,

Keratinocyte, Fetal cortex (GW21+3), Fetal cortex (GW21), Fetal cortex (GW16), and NPC) that are expected to show robust differences in gene expression. Data were pre-filtered to exclude genes where more than 90% of cells had zero measurements and include only single cells with greater than 500,000 reads ($n = 249$).

- Buettner: Embryonic stem cells under different cell cycle stages. Buettner *et al.* (2015) assayed the transcriptional profile of 182 ESCs that had been staged for cell-cycle phase (G1, S, and G2M) based on sorting of the Hoechst 33342-stained cell area of a flow cytometry (FACS) distribution. The cells were sorted for three stages of the cell cycle, and they were validated using gold-standard Hoechst staining. The data have been deposited at ArrayExpress: E-MTAB-2805.

- Ting: Single Cell RNA-sequencing of Pancreatic Circulating Tumor Cells. We downloaded the data from GEO (GSE51372), which contains 5 subtypes from Single-cell transcriptomes from MEFs, the NB508 pancreatic cancer cell line, normal WBCs, bulk primary tumors diluted to 10 or 100 pg of RNA, and classical CTC.

- Treutlein: This data set contains single cell RNA-seq expression data for 80 lung epithelial cells at E18.5 together with the five putative cell type; AT1, AT2, Clara, BP, and ciliated. We downloaded the data from https://www.nature.com/articles/nature13173. We considered data with selected genes with 959 highest loadings in the first four PCA coefficients following the similar approach of Treutlein *et al.* (2014).

- Deng (Deng *et al.*, 2014): The Deng data set consists of transcriptomes for individual cells isolated from mouse embryos at different preimplantation stages. The data set consists of 135 cells and 19,703 genes, where cells belong to zygote, early 2-cell-stage, mid 2-cell-stage, late 2-cell-stage, 4-cell-stage, 8-cell-stage, and 16-cell-stage. We downloaded the processed data from GEO (GSE45719).

- Ginhoux (Schlitzer *et al.*, 2015): This data set contains the expression values of 15,752 genes for 251 dendritic cell progenitors in one of following three cellular states: Monocyte and Dendritic cell Progenitors (MDPs), Common Dendritic cell Progenitors (CDPs), and Pre-Dendritic Cells (PreDCs). The data set contains 59 MDPs, 96 CDPs, and 96 PreDCs. We downloaded the processed data from GEO (GSE60783).

- Tasic (Tasic *et al.*, 2016): Tasic *et al.* (2016) identified 49 transcriptomic cell types, including 23 GABAergic, 19 glutamatergic and 7 non-neuronal types. To identify cell types, they applied two parallel and iterative approaches for dimensionality reduction and clustering, iterative principal component analysis (PCA) and iterative weighted gene coexpression network analysis (WGCNA), and validated the cluster membership from each approach using a non-deterministic machine learning method (random forest). We downloaded the processed data from GEO (GSE71585).

- Zeisel (Zeisel *et al.*, 2015): Zeisel *et al.* (2015) have used large-scale single-cell RNA sequencing to classify cells in the mouse somatosensory

cortex and hippocampal CA1 region. 3,005 Cells from the mouse cortex and hippocampus collected. Zeisel *et al.* (2015) found 47 molecularly distinct subclasses identified by hierarchical biclustering and validated by gene markers.

- Macosko (Macosko *et al.*, 2015): Mouse retina cells with 39 subtypes. This data set is obtained by droplet-based high-throughput technique. The data set consists of 44,808 cells. The 39 cell types were identified via PCA and density-based clustering, and they were validated by differential gene expression. We filtered out cells with less than 1,200 genes (yielding 6,418 cells) for clustering analysis. We downloaded the data from GEO (GSE63473).

G Additional figures

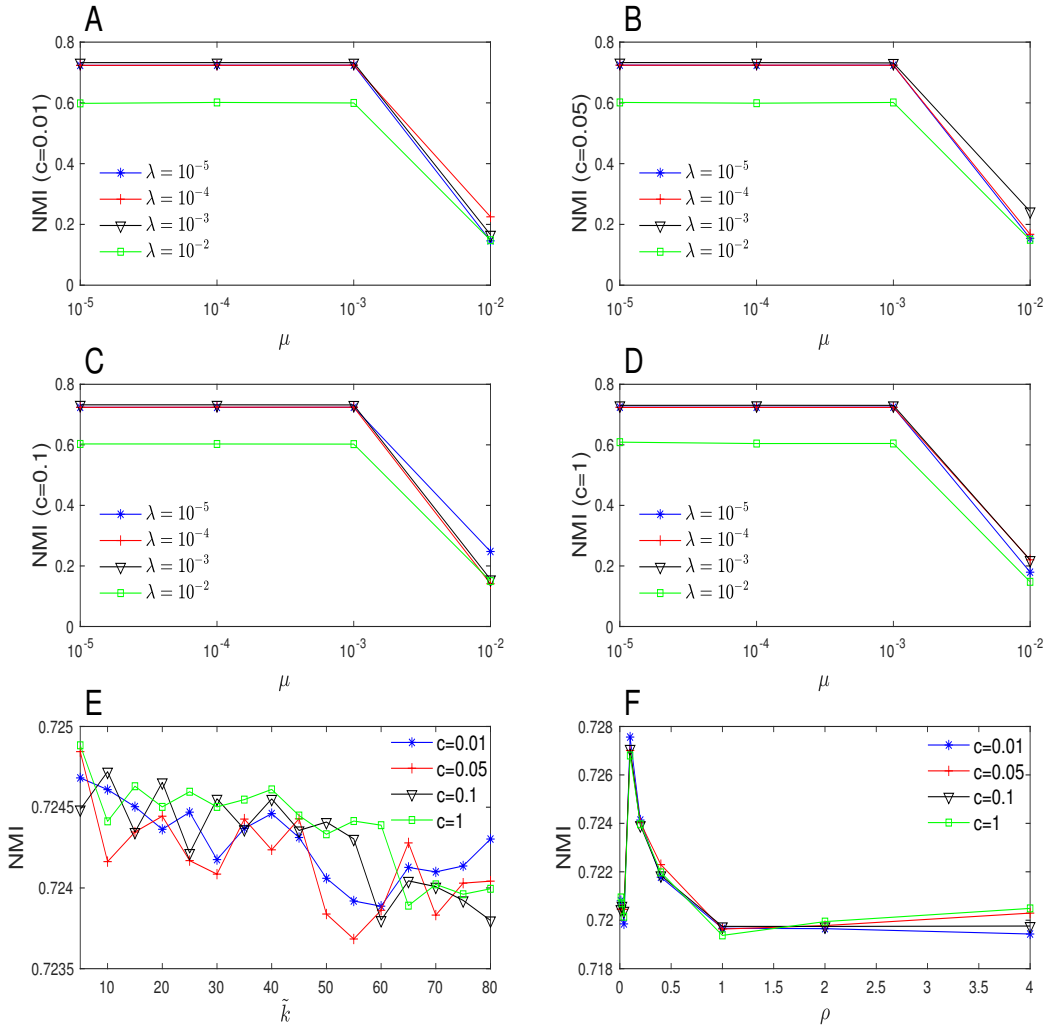

Figure S2: The effects of parameters with $c = 0.01(A)$, $0.05(B)$, $0.1(C)$, and $1.0(D)$ on the clustering results of the simulated data set with $d = 10$, $\sigma_k = 3$, and $\gamma = 0.2$ when $\tilde{k} = 10$. The (E) considers the case when $c \in \{0.01, 0.1, 0.5, 1.0\}$, $\tilde{k} \in \{5, 10, \cdots, 80\}$, and $\lambda = \mu = 0.0001$. The (F) considers when $c \in \{0.01, 0.05, 0.1, 1.0\}$ and $\rho \in [0.01, 4]$. We consider $(n, p, q, C) = (250, 500, 20, 5)$, $n_k = 50$, and $(d, \mu, \sigma_k, \sigma) = (10, 10, 3, 0.5)$.

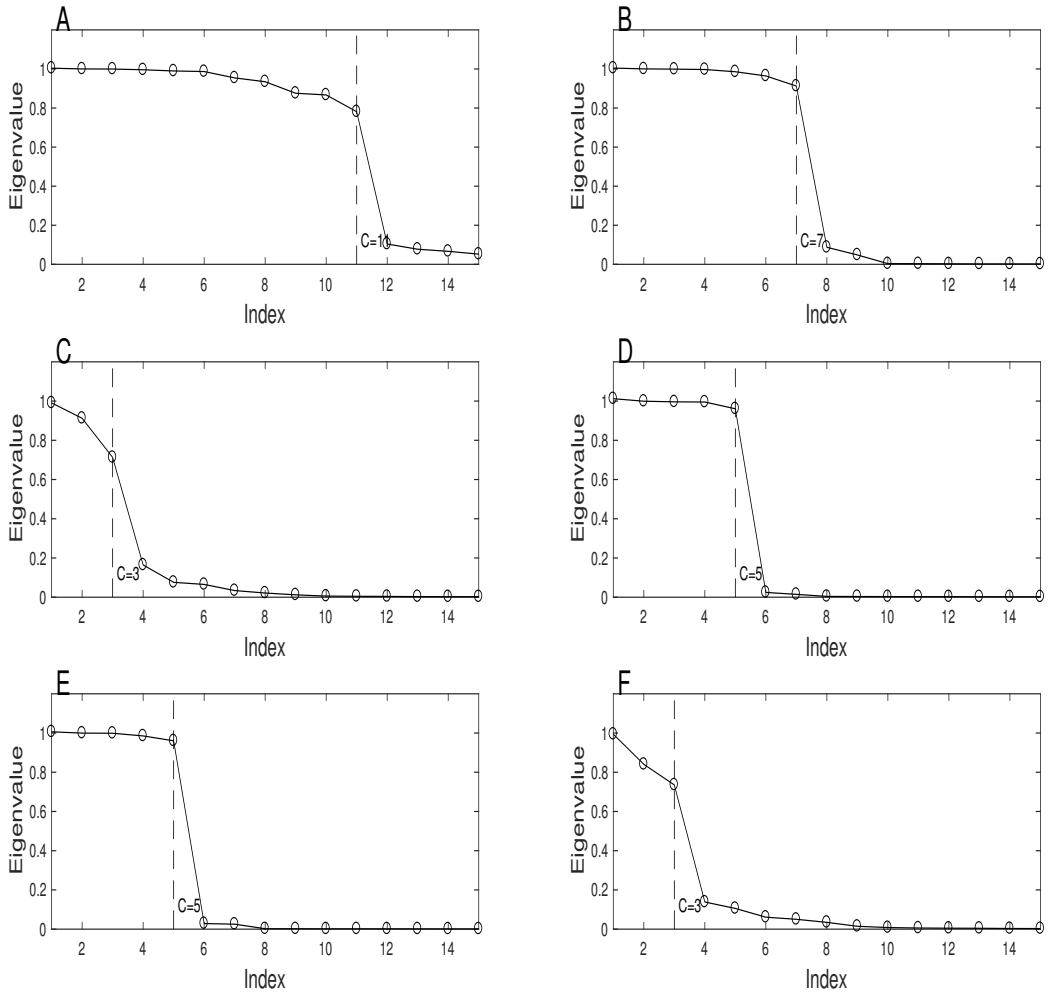

Figure S3: The inference of the number of clusters with the proposed method on the six real single-cell data sets; Pollen (Pollen *et al.*, 2014)(A), Deng (Deng *et al.*, 2014)(B), Ginhoux (Schlitzer *et al.*, 2015)(C), Ting (Ting *et al.*, 2014)(D), Treutlein (Treutlein *et al.*, 2014)(E), and Buettner (Buettner *et al.*, 2015)(F).

Figure S4: Evaluation of the six clustering methods by Purity (A) and ARI (B) for the six small-scale data sets.

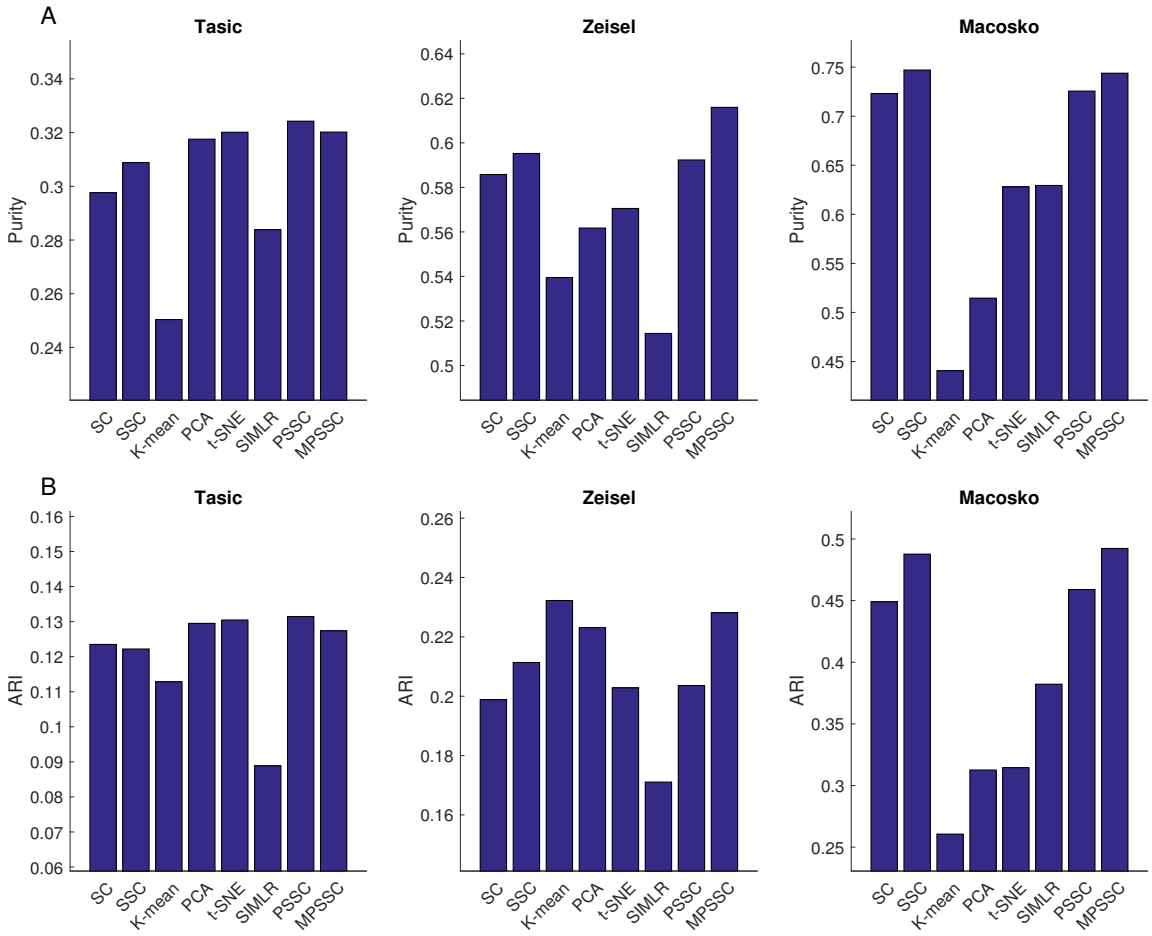

Figure S5: Evaluation of the six clustering methods by Purity (A) and ARI (B) for the three large-scale data sets.

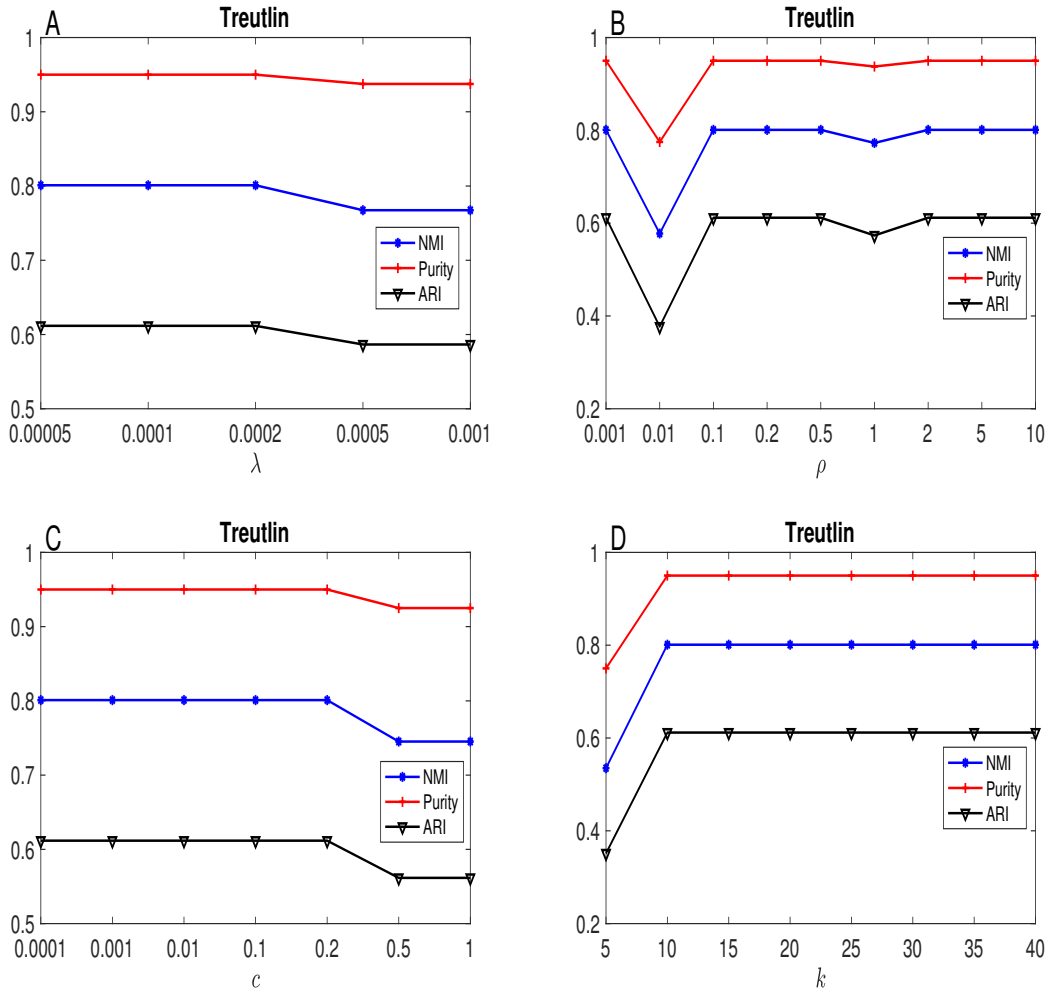

Figure S6: Sensitivity analysis of the proposed method on the Treutlein data set (Treutlein *et al.*, 2014). In (A), we set $\rho = 0.2, c = 0.1, k = 10$. In (B), we set $\lambda = 0.0001, c = 0.1, k = 10$. In (C), we use $\lambda = 0.0001, \rho = 0.2, k = 10$. In (D), we fix $\lambda = 0.0001, \rho = 0.2, c = 0.1, k = 10$.

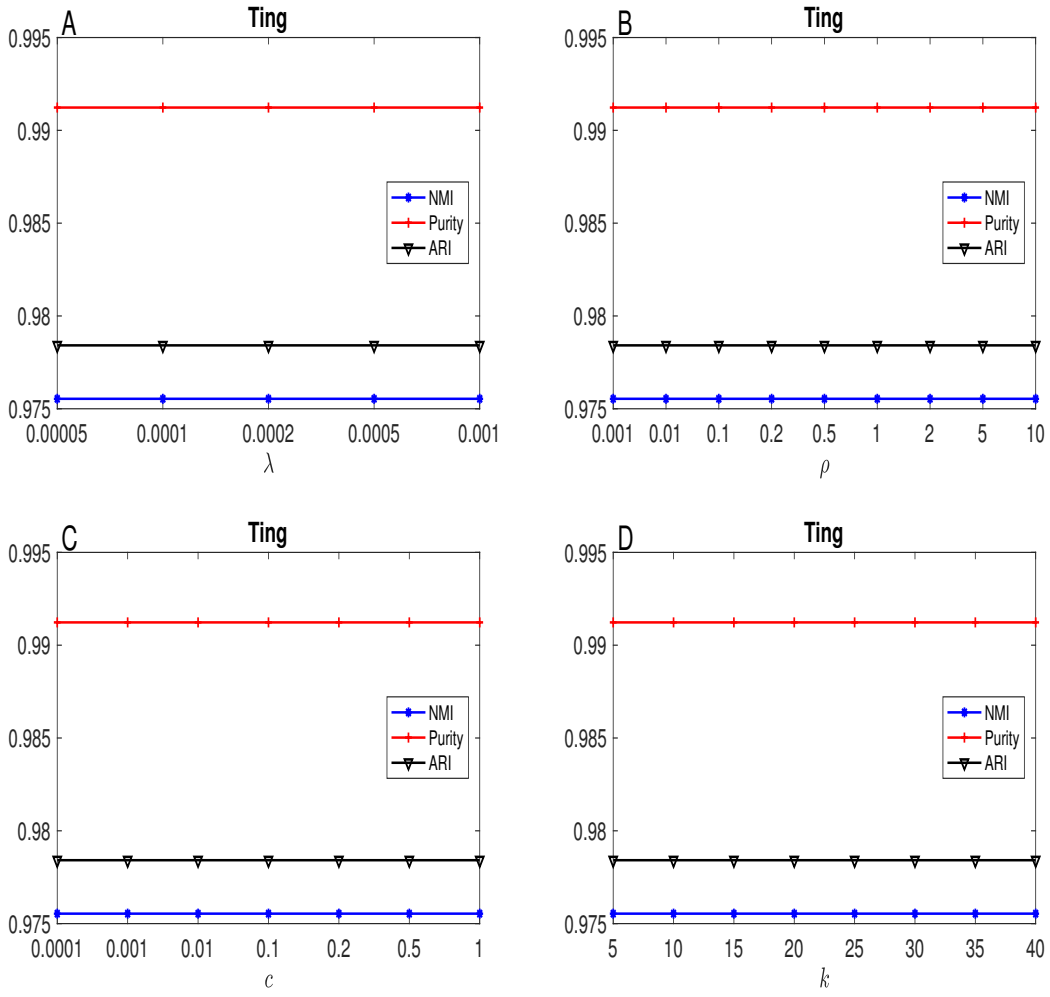

Figure S7: Sensitivity analysis of the proposed method on the Ting data set (Ting *et al.*, 2014). In (A), we set $\rho = 0.2, c = 0.1, k = 10$. In (B), we set $\lambda = 0.0001, c = 0.1, k = 10$. In (C), we use $\lambda = 0.0001, \rho = 0.2, k = 10$. In (D), we fix $\lambda = 0.0001, \rho = 0.2, c = 0.1, k = 10$.

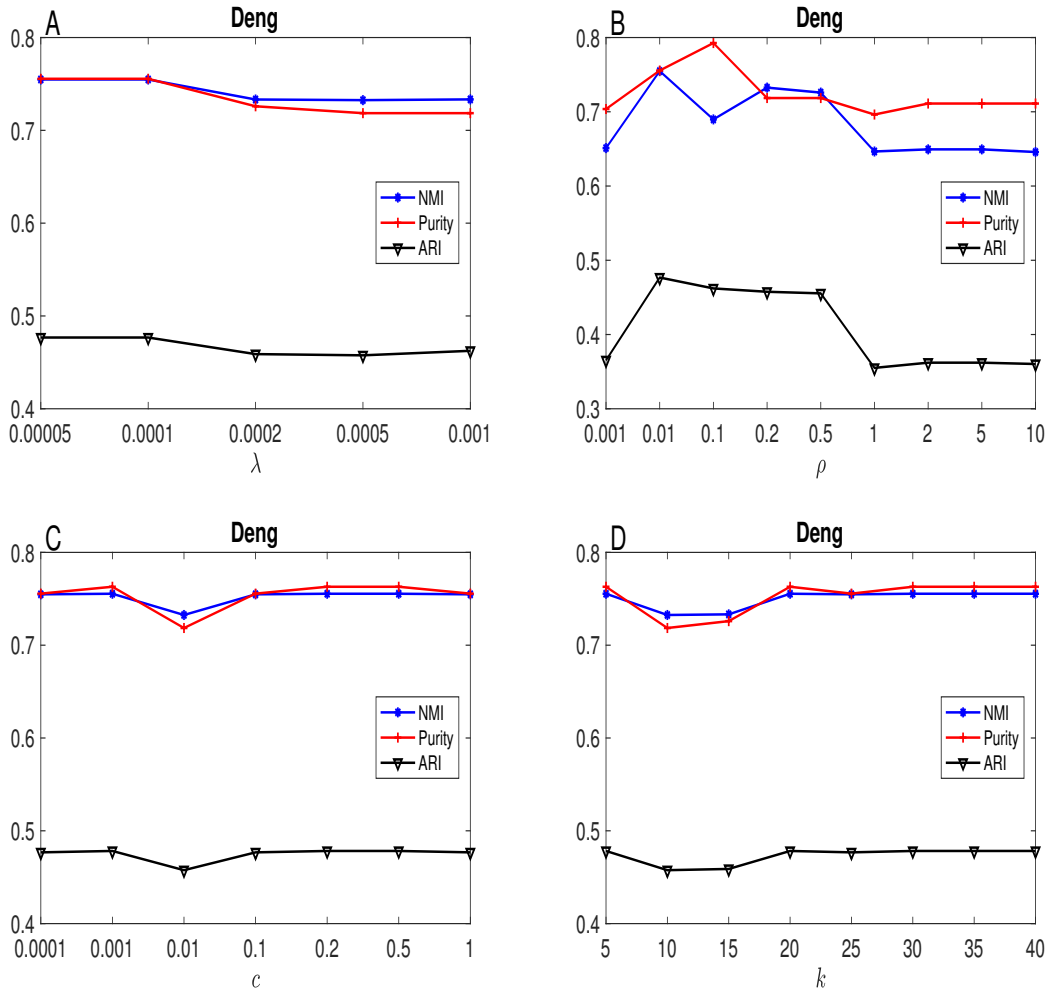

Figure S8: Sensitivity analysis of the proposed method on the Deng data set (Deng *et al.*, 2014). In (A), we set $\rho = 0.2, c = 0.1, k = 10$. In (B), we set $\lambda = 0.0001, c = 0.1, k = 10$. In (C), we use $\lambda = 0.0001, \rho = 0.2, k = 10$. In (D), we fix $\lambda = 0.0001, \rho = 0.2, c = 0.1, k = 10$.

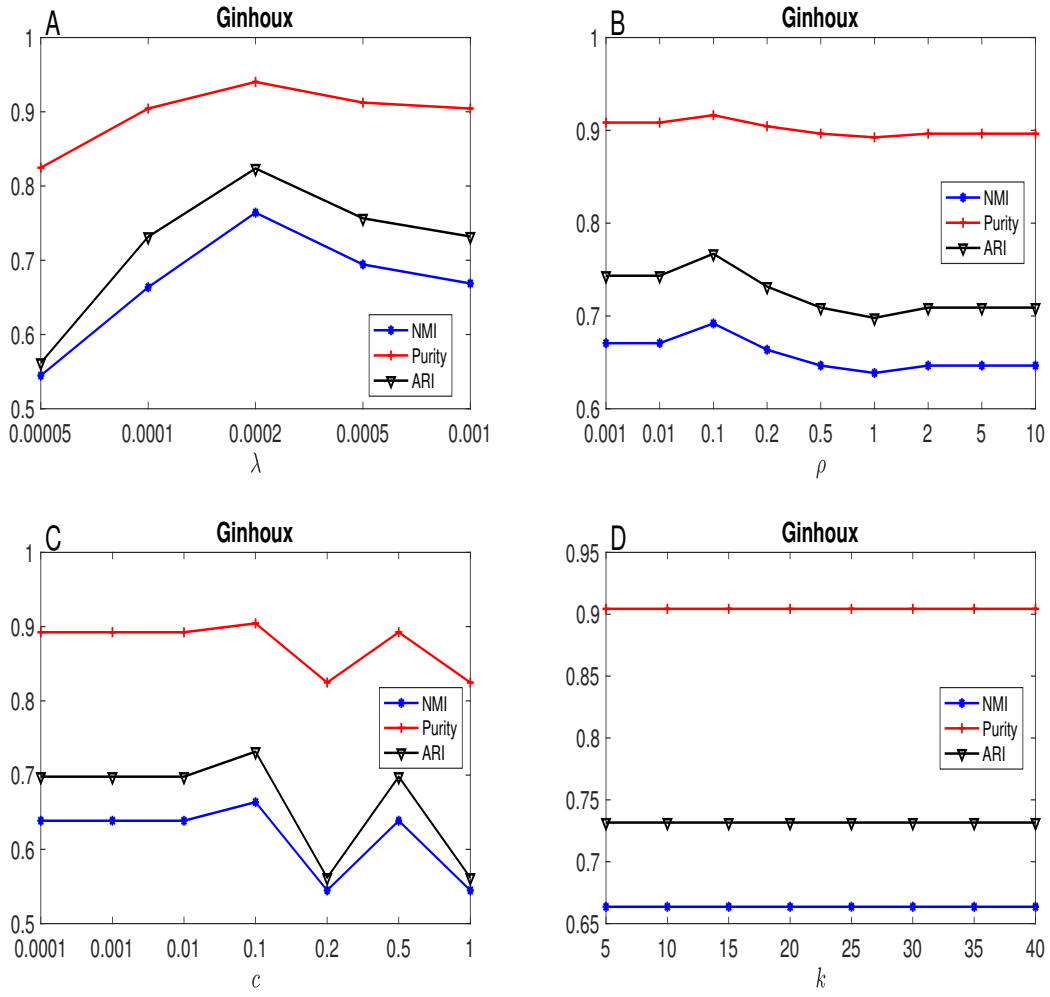

Figure S9: Sensitivity analysis of the proposed method on the Ginhoux data set (Schlitzer *et al.*, 2015). In (A), we set $\rho = 0.2, c = 0.1, k = 10$. In (B), we set $\lambda = 0.0001, c = 0.1, k = 10$. In (C), we use $\lambda = 0.0001, \rho = 0.2, k = 10$. In (D), we fix $\lambda = 0.0001, \rho = 0.2, c = 0.1, k = 10$.

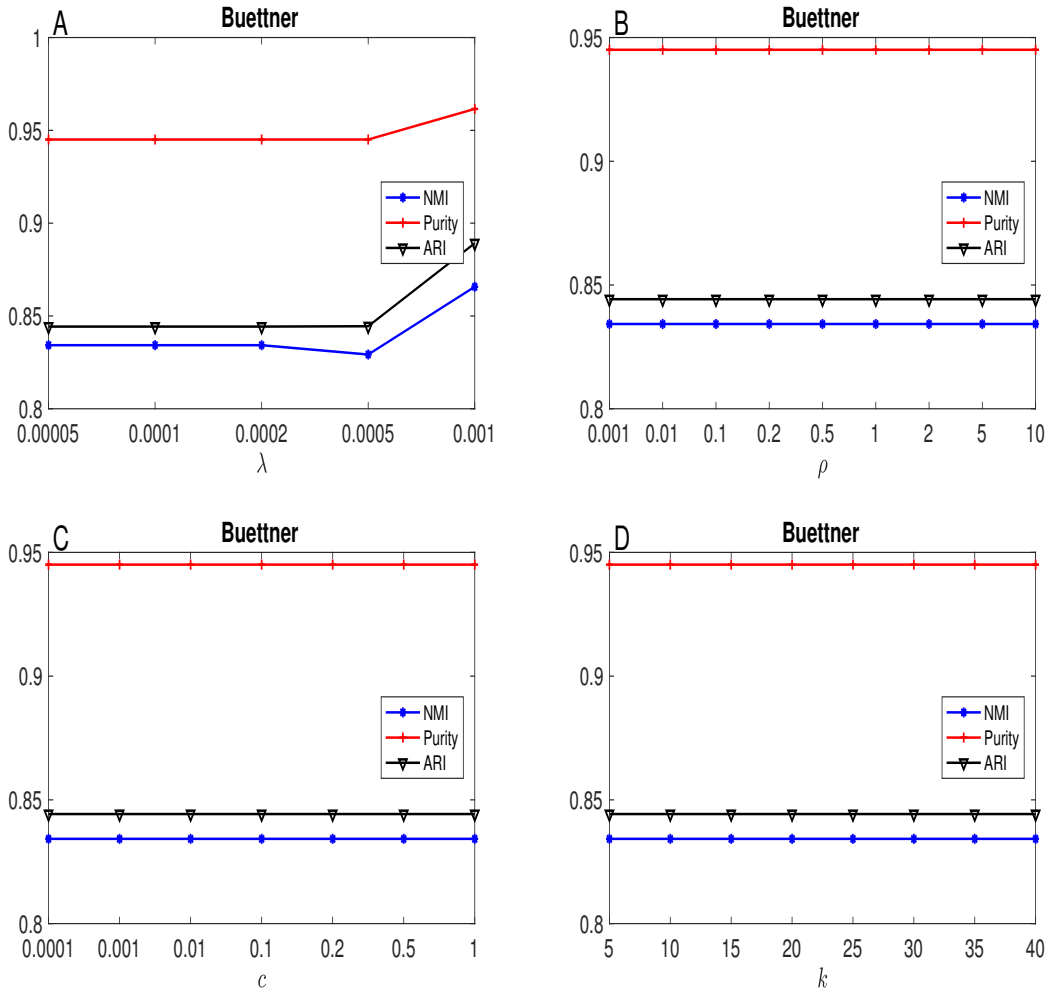

Figure S10: Sensitivity analysis of the proposed method on the Buettner data set (Buettner *et al.*, 2015). In (A), we set $\rho = 0.2, c = 0.1, k = 10$. In (B), we set $\lambda = 0.0001, c = 0.1, k = 10$. In (C), we use $\lambda = 0.0001, \rho = 0.2, k = 10$. In (D), we fix $\lambda = 0.0001, \rho = 0.2, c = 0.1, k = 10$.

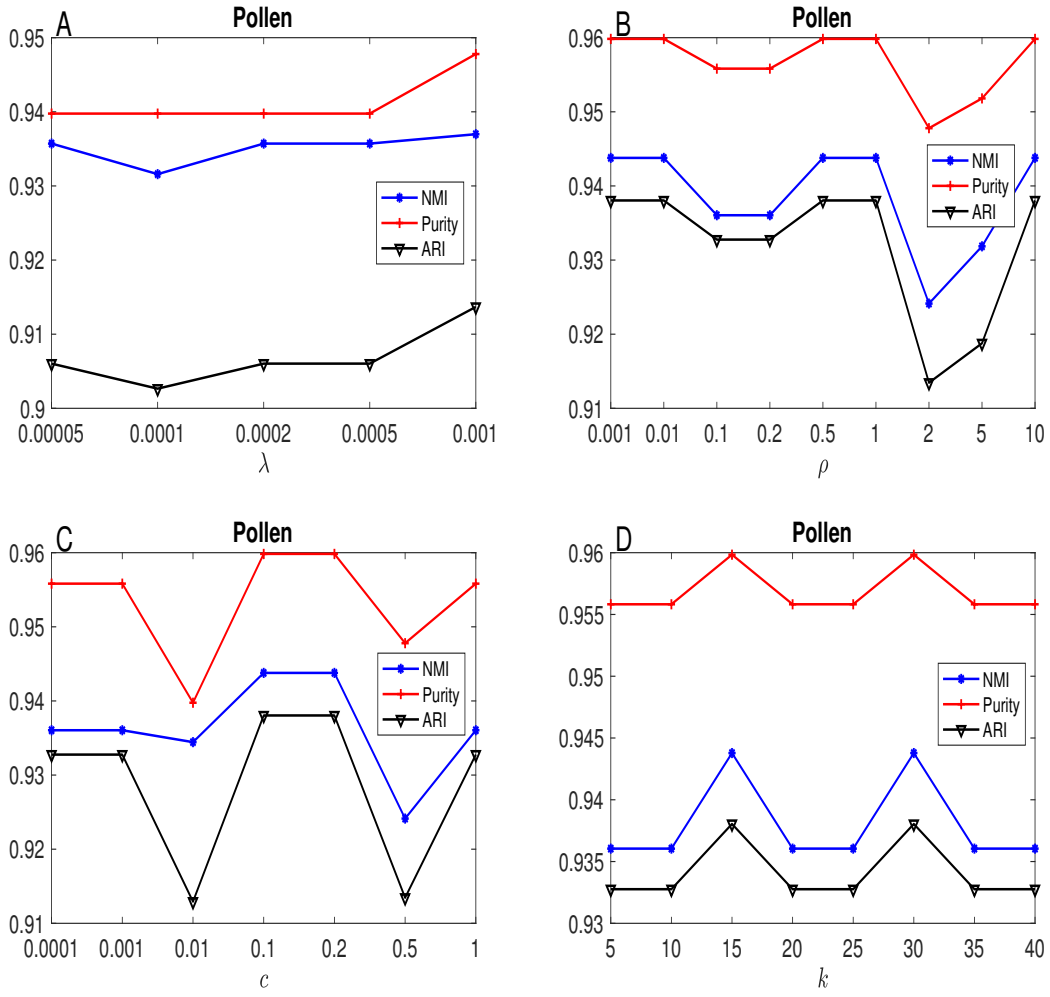

Figure S11: Sensitivity analysis of the proposed method on the Pollen data set (Pollen *et al.*, 2014). In (A), we set $\rho = 0.2, c = 0.1, k = 10$. In (B), we set $\lambda = 0.0001, c = 0.1, k = 10$. In (C), we use $\lambda = 0.0001, \rho = 0.2, k = 10$. In (D), we fix $\lambda = 0.0001, \rho = 0.2, c = 0.1, k = 10$.

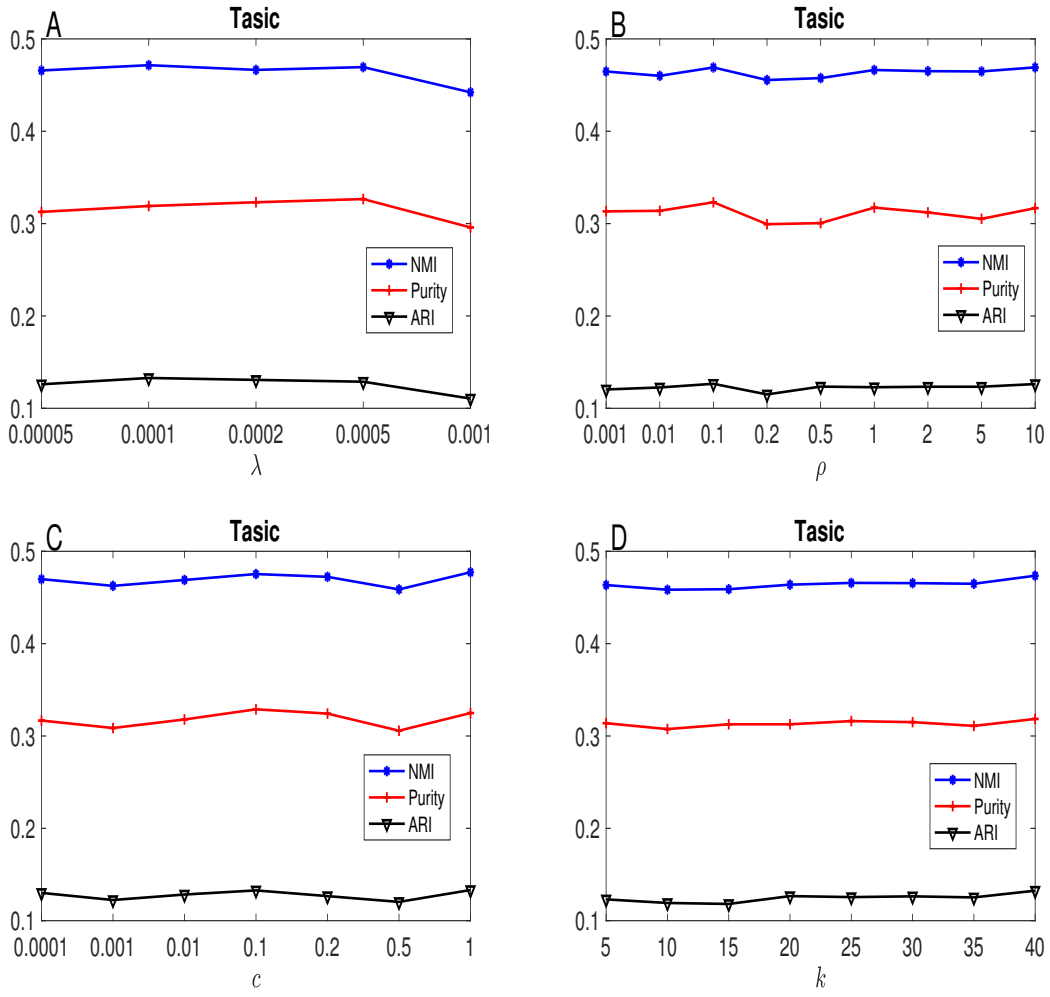

Figure S12: Sensitivity analysis of the proposed method on the Tasic data set (Tasic *et al.*, 2016). In (A), we set $\rho = 0.2, c = 0.1, k = 10$. In (B), we set $\lambda = 0.0001, c = 0.1, k = 10$. In (C), we use $\lambda = 0.0001, \rho = 0.2, k = 10$. In (D), we fix $\lambda = 0.0001, \rho = 0.2, c = 0.1, k = 10$.

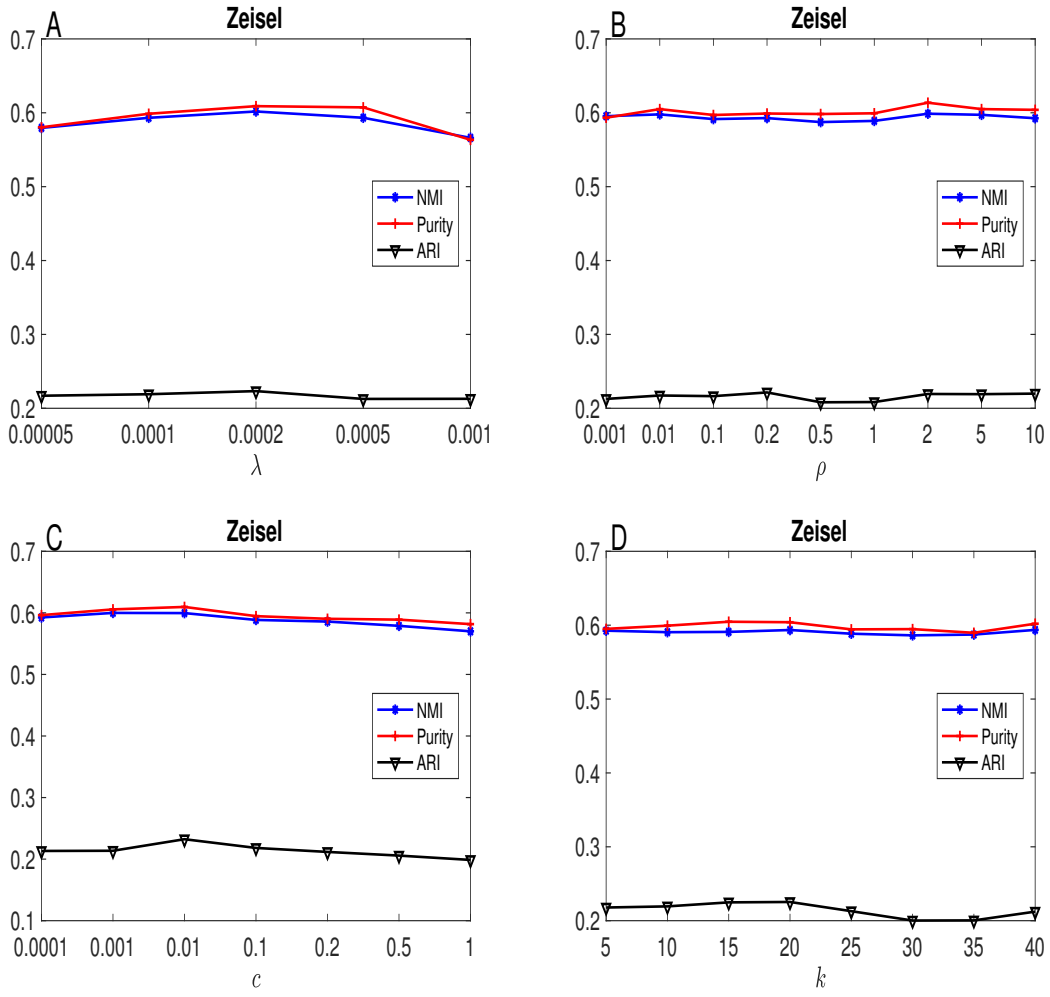

Figure S13: Sensitivity analysis of the proposed method on the Zeisel data set (Zeisel *et al.*, 2015). In (A), we set $\rho = 0.2, c = 0.1, k = 10$. In (B), we set $\lambda = 0.0001, c = 0.1, k = 10$. In (C), we use $\lambda = 0.0001, \rho = 0.2, k = 10$. In (D), we fix $\lambda = 0.0001, \rho = 0.2, c = 0.1, k = 10$.

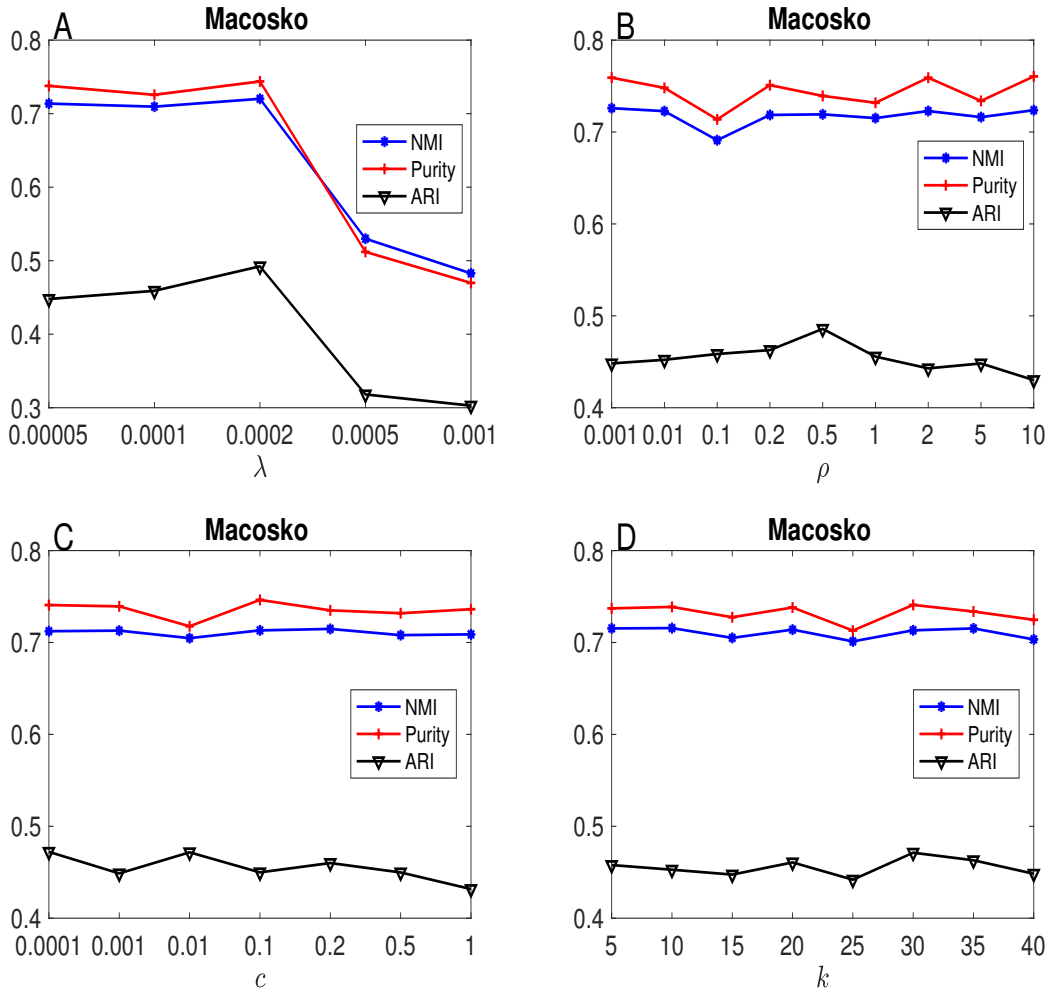

Figure S14: Sensitivity analysis of the proposed method on the Macosko data set (Macosko *et al.*, 2015). In (A), we set $\rho = 0.2, c = 0.1, k = 10$. In (B), we set $\lambda = 0.0001, c = 0.1, k = 10$. In (C), we use $\lambda = 0.0001, \rho = 0.2, k = 10$. In (D), we fix $\lambda = 0.0001, \rho = 0.2, c = 0.1, k = 10$.

References

Boyd, S., Parikh, N., Chu, E., Peleato, B., and Eckstein, J. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, **3**(1), 1–122.

Buettner, F. *et al.* (2015). Computational analysis of cell-to-cell heterogeneity in single-cell rna-sequencing data reveals hidden subpopulations of cells. *Nature Biotechnology*, **33**(2), 155–160.

Deng, Q. *et al.* (2014). Single-cell rna-seq reveals dynamic, random monoallelic gene expression in mammalian cells. *Science*, **343**, 193–196.

He, B. and Yuan, X. (2012). On the $o(1/n)$ convergence rate of the douglas-rachford alternating direction method. *SIAM Journal on Numerical Analysis*, **50**(2), 700–709.

Lu, C. *et al.* (2016a). Convex sparse spectral clustering: single-view to multi-view. *IEEE TRANSACTIONS ON IMAGE PROCESSING*, **25**(6), 2833–2843.

Lu, C. *et al.* (2016b). Fast proximal linearized alternating direction method of multiplier with parallel splitting. *Proc. AAAI Conf. Artif. Intell.*

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

Monteiro, R. and Svaiter, B. (2013). Iteration-complexity of block-decomposition algorithms and the alternating direction method of multipliers. *SIAM Journal on Optimization*, **23**(1), 475–507.

Pierson, E. and Yau, C. (2015). Zifa: Dimensionality reduction for zero-inflated single-cell gene expression analysis. *Genome Biology*, **16**(241), 1–10.

Pollen, A. *et al.* (2014). Low-coverage single-cell mrna sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. *Nature Biotechnology*, **32**, 1053–1058.

Schlitzer, A. *et al.* (2015). Identification of cdc1- and cdc2-committed dc progenitors reveals early lineage priming at the common dc progenitor stage in the bone marrow. *Nature Immunology*, **16**(7), 718–728.

Strehl, A. and Ghosh, J. (2003). Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research*, **3**, 583–617.

Tasic, B. *et al.* (2016). Adult mouse cortical cell taxonomy revealed by single cell transcriptomics. *Nat. Neurosci.*, **19**, 335–346.

Ting, D. T. *et al.* (2014). Single-cell rna sequencing identifies extracellular matrix gene expression by pancreatic circulating tumor cells. *Cell Reports*, **8**, 1905–1918.

Treutlein, B. *et al.* (2014). Reconstructing lineage hierarchies of the distal lung epithelium using single-cell rna-seq. *Nature Letter*, **509**, 371–375.

Tseng, P. (2001). Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal Optimization Theory and Applications*, **109**(3), 475–494.

Wagner, S. and Wagner, D. (2007a). Comparing clusterings- an overview. *Universit at Karlsruhe, Technical Report*.

Wagner, S. and Wagner, D. (2007b). Comparing clusterings: an overview. *Universit at Karl- sruhe, Fakult at fu r Informatik Karlsruhe*.

Wang, B. *et al.* (2017). Visualization and analysis of single-cell rna-seq data by kernel-based similarity learning. *Nature Methods*, **14**, 414–416.

Zeisel, A. *et al.* (2015). Brain structure. cell types in the mouse cortex and hippocampus revealed by single-cell rna-seq. *Science*, **347**, 1138–1142.