EFFICIENT CLUSTERING USING ALTERNATING MINIMIZATION AND A PROJECTION-GRADIENT METHOD FOR DIMENSION REDUCTION

Cyprien Gilet[1], Marie Deprez[2], Pacal Barbry[2] ,Jean-Baptiste Caillau[3] and Michel Barlaud[1], Fellow,IEEE

[1] Université Côte d'Azur, CNRS, I3S
[2] Université Côte d'Azur, CNRS, IPMC
[3] Université Côte d'Azur, CNRS, Inria, LJAD

ABSTRACT

This paper deals with unsupervised clustering in high dimensional space. The problem is to estimate both labels and a sparse projection matrix of weights. To address this combinatorial non-convex problem maintaining a strict control on the sparsity of the matrix of weights, we propose an alternating minimization of the Frobenius norm criterion. We provide a new efficient algorithm named k-sparse which alternates k-means with projection-gradient minimization. The projection-gradient step is a method of splitting type, with exact projection on the ℓ^1 ball to promote sparsity. The convergence of the gradient-projection step is addressed, and a preliminary analysis of the alternating minimization is provided. Experiments on Single Cell RNA sequencing datasets show that our method significantly improves the results of spectral clustering, SIMLR, and Sparcl methods. The complexity of our method is linear with the number of samples (cells), so that the method scales up to large datasets.

1. RELATED WORKS

This paper deals with unsupervised clustering and removal of noisy features in high dimensional space. Clustering in high dimension using classical algorithms such as k-means [1, 2] suffers from the curse of dimensionality. As dimensions increase, vectors become indiscernible and the predictive power of the aforementioned methods is drastically reduced [3]. We advocate the use of sparsity promoting methods as they allow not only to perform feature selection (dimensionality reduction), but also to use efficient state-of-the-art algorithms from convex optimization. An early approach proposed in [4, 5] is to combine clustering and dimension reduction by means of *Linear Discriminant Analysis* (LDA). The heuristic used in [5] is based on alternating minimization, which consists in iteratively computing a projection subspace by LDA, using the labels y at the current iteration and then running k-means on the projection of the data onto the subspace. Departing from this work, the approach [6] proposes a convex relaxation in terms of a suitable semi-definite program (SDP). Another efficient method is spectral clustering where the main tools are graph Laplacian matrices [7, 8]. The approach [9] use a lagrangian lasso-type penalty to select the features and propose a sparse k-means method. A main issue is that optimizing the values of the Lagrangian parameter λ [9] is computationally expensive. All these methods [4, 5, 6, 9] require a k-means heuristic to retrieve the labels.

2. CONSTRAINED UNSUPERVISED CLASSIFICATION

2.1. General Framework

Let X be the (nonzero) $m \times d$ matrix made of m line samples $x_1, \ldots, x_m$ belonging to the d-dimensional space of features. Let $Y \in \{0,1\}^{m \times k}$ be the matrix of labels where $k \geqslant 2$ is the number of clusters. Note that we assume that this number is known; It is indeed the case for the applications we present in Section 3, while estimating k is in general a delicate matter out of the scope of this paper. Each line of Y has exactly one nonzero element equal to one, $y_{ij} = 1$ indicating that the sample x_i belongs to the j-th cluster. Let $W \in \mathbb{R}^{d \times \bar{d}}$ be the projection matrix, where the dimension in the projected space, $\bar{d}$, is understood to be much smaller than d. Let then μ be the $k \times \bar{d}$ matrix of centroids of the projected data, XW:

$$\mu(j,:) := \frac{1}{\sum_{i=1}^{m} y_{ij}} \sum_{i \text{ s.t. } y_{ij}=1} (XW)(i,:).$$

The j-th centroid is the model for all samples x_i belonging to the j-th cluster ($y_{ij} = 1$). The clustering criterion can be cast as the *Within-Cluster Sum of Squares* (WCSS) [10, 9] in the projected space

$$\min_{Y \in \{0,1\}^{m \times k}, W \in \mathbb{R}^{d \times \bar{d}}} \frac{1}{2} \|Y\mu - XW\|_F^2 \qquad (1)$$

where $\|.\|_F$ is the Frobenius norm induced by the Euclidean structure on $m \times \bar{d}$ matrices,

$$(A|B)_F := \mathrm{tr}(A^T B) = \mathrm{tr}(AB^T), \quad \|A\|_F := \sqrt{(A|A)_F}\,.$$

In contrast with the Lagrangian formulation, we want to have a direct control on the value of the ℓ^1 bound, so we

constrain W according to

$$\|W\|_1 \leqslant \eta \quad (\eta > 0), \tag{2}$$

where $\|.\|_1$ is the ℓ^1 norm of the vectorized $d \times \bar{d}$ matrix of weights:

$$\|W\|_1 := \|W(:)\|_1 = \sum_{i=1}^{d} \sum_{j=1}^{\bar{d}} |w_{ij}|.$$

The problem is to estimate labels Y together with the sparse projection matrix W. As Y and W are bounded, the set of constraints is compact and existence of minimizers holds.

Proposition 1 *The minimization of the norm (1), jointly in Y and W has a solution.*

To attack this difficult nonconvex problem, we propose an alternating (or Gauss-Seidel) scheme as in [4, 5, 9]. Another option would be to design a global convex relaxation to address the joint minimization in Y and W (see, *e.g.*, [6]) The first convex subproblem is to find the best projection from dimension d to dimension $\bar{d}$ for a given clustering.

Problem 1 *For a fixed clustering Y (and a given $\eta > 0$),*

$$\underset{W \in C}{\text{minimize}} \ \frac{1}{2}\|Y\mu - XW\|_F^2,$$

where $C := \{W \in \mathbb{R}^{d \times \bar{d}} : \|W\|_1 \leqslant \eta\}$.

Given the matrix of weights $W \in C$, the second subproblem is the standard k-means on the projected data.

Problem 2 *For a fixed projection matrix $W \in C$,*

$$\underset{Y \in \{0,1\}^{m \times k}}{\text{minimize}} \ \frac{1}{2}\|Y\mu - XW\|_F^2.$$

2.2. Exact gradient-projection splitting method

To solve Problem 1, we use a gradient-projection method. It belongs to the class of splitting methods [11, 12, 13, 14, 15]. It is designed to solve minimization problems of the form

$$\min_{W \in C} \ \varphi(W), \tag{3}$$

using separately the convexity properties of the function φ on one hand, and of the convex set C on the other. We use the following forward-backward scheme to generate a sequence of iterates:

$$V_n := W_n - \gamma_n \nabla\varphi(W_n), \tag{4}$$

$$W_{n+1} := P_C(V_n) + \varepsilon_n, \tag{5}$$

where P_C denotes the projection on the convex set C (a subset of some Euclidean space). Under standard assumptions on the sequence of gradient steps $(\gamma_n)_n$, and on the sequence of projection errors $(\varepsilon_n)_n$, convergence holds (see, *e.g.*, [16]).

Theorem 1 *Assume that (3) has a solution. Assume that φ is convex, differentiable, and that $\nabla\varphi$ is β-Lipschitz, $\beta > 0$. Assume finally that C is convex and that*

$$\sum_n |\varepsilon_n| < \infty, \quad \inf_n \gamma_n > 0, \quad \sup_n \gamma_n < 2/\beta.$$

Then the sequence of iterates of the forward-backward scheme (4-5) converges, whatever the initialization. If moreover $(\varepsilon_n)_n = 0$ (exact projections), there exists a rank N and a positive constant K such that, for $n \geqslant N$,

$$\varphi(W_n) - \inf_C \varphi \leqslant K/n. \tag{6}$$

In our case, $\nabla\varphi$ is Lipschitz since it is affine,

$$\nabla\varphi(W) = X^T(XW - Y\mu), \tag{7}$$

and we recall the estimation of its best Lipschitz constant.

Lemma 1 *Let A be a $d \times d$ real matrix, acting linearly on the set of $d \times k$ real matrices by left multiplication, $W \mapsto AW$. Then, its norm as a linear operator on this set endowed with the Frobenius norm is equal to its largest singular value, $\sigma_{\max}(A)$.*

Proof. The Frobenius norm is equal to the ℓ^2 norm of the vectorized matrix,

$$\|W\|_F = \| \begin{bmatrix} W^1 \\ \vdots \\ W^h \end{bmatrix} \|_2, \quad \|AW\|_F = \| \begin{bmatrix} AW^1 \\ \vdots \\ AW^h \end{bmatrix} \|_2, \tag{8}$$

where $W^1, \ldots, W^h$ denote the h column vectors of the $d \times h$ matrix W. Accordingly, the operator norm is equal to the largest singular value of the $kd \times kd$ block-diagonal matrix whose diagonal is made of k matrix A blocks. Such a matrix readily has the same largest singular value as A. □

As a byproduct of Theorem 1, we get

Corollary 1 *For any fixed step $\gamma \in (0, 2/\sigma_{\max}^2(X))$, the forward-backward scheme applied to the Problem 1 with an exact projection on ℓ^1 balls converges with a linear rate towards a solution, and the estimate (6) holds.*

Proof. The ℓ^1 ball being compact, existence holds. So does convergence, provided the condition of the step lengths is fulfilled. Now, according to the previous lemma, the best Lipschitz constant of the gradient of φ is $\sigma_{\max}(X^TX) = \sigma_{\max}^2(X)$, hence the result. □

2.3. Alternating minimization algorithm and convergence

The resulting alternating minimization procedure is summarized in Algorithm 1. Labels Y are for instance initialized by spectral clustering on X, while the k-means computation

relies on standard methods such as k-means++ [2]. We denote by $P_\eta^1(W)$ the (reshaped as a $d \times \bar{d}$ matrix) ℓ^1 projection of the vectorized matrix $W(:)$ which is computed by efficient methods [17, 18].

Algorithm 1 Alternating minimization clustering.

Input: $X, Y_0, \mu_0, W_0, L, N, k, \gamma, \eta$
$Y \leftarrow Y_0$
$\mu \leftarrow \mu_0$
$W \leftarrow W_0$
for $l = 0, \ldots, L$ **do**
  for $n = 0, \ldots, N$ **do**
    $V \leftarrow W - \gamma X^T(XW - Y\mu)$
    $W \leftarrow P_\eta^1(V)$
  end for
  $Y \leftarrow \texttt{kmeans}(XW, k)$
  $\mu \leftarrow \texttt{centroids}(Y, XW)$
end for
Output: Y, W

Similarly to the approaches advocated in [4, 5, 6, 9], our method involves non-convex k-means optimization for which convergence towards local minimizers only can be proved [19]. In practice, we use k-means++ with several replicates to improve each clustering step. We assume that the initial guess for labels Y and matrix of weights W is such that the associated k centroids are all different. We note for further research that there have been recent attempts to convexify k-means (see, *e.g.*, [20, 21]). As each step of the alternating minimization scheme decreases the norm in (1), which is nonnegative, the following readily holds.

Proposition 2 *The Frobenius norm $\|Y\mu - XW\|_F$ converges as the number of iterates L in Algorithm 1 goes to infinity.*

This property is illustrated in Fig. 2 on biological data. Further analysis of the convergence may build on recent results on proximal regularizations of the Gauss-Seidel alternating scheme for non convex problems [22, 23].

3. EXPERIMENTAL EVALUATION ON SINGLE CELL RNA-SEQ CLUSTERING

3.1. Single cell datasets

Our algorithm can be readily extended to multiclass clustering of high dimensional databases in computational biology (single cell clustering, mass-spectrometric data...). In this paper, we provide an experimental evaluation on Single-cell sequencing dataset [24]. We performed our algorithm on three public single-cell RNA-seq datasets: Klein [25], Zeisel [26], Usoskin [27].

The Klein scRNA-seq dataset [25] characterizes the transcriptome of 2,717 cells clustered in $k = 4$ classes from 10,322 genes. The Zeisel scRNA-seq dataset [26] collected 3,005 mouse cells clustered in $k = 9$ classes and 7364 genes from the primary somatosensory cortex (S1) and the hippocampal CA1 region. The Usoskin scRNA-seq dataset [27] collected 622 cells clustered in $k = 4$ classes and 9195 genes from the mouse dorsal root ganglion.

3.2. Experimental settings

The problem of estimating the number of clusters is out of the range of this study, and we refer to the popular GAP method [28]. We compare our method with spectral clustering [8], SIMLR (Single-cell Interpretation via Multikernel Learning) [29] [1] and Sparcl (Sparse k-means clustering) [9] using the *R* software package `Sparcl`.

In order to evaluate the results of each algorithms, we compare the labels obtained by each clustering method with the true labels to compute the clustering accuracy. We also report the popular *Adjusted Rank Index* (ARI) [30] and *Normalized Mutual Information* (NMI) criteria as well as silhouette coefficients [31]. Processing times are obtained on a computer using an i7 processor (2.5 Ghz).

Regarding our method and as illustrated in Fig. 1, the parameter η has a direct impact on the number of selected features. Here, we chose η such that it allows to achieve both the best silhouette coefficient and also to discard a large number of noisy features.

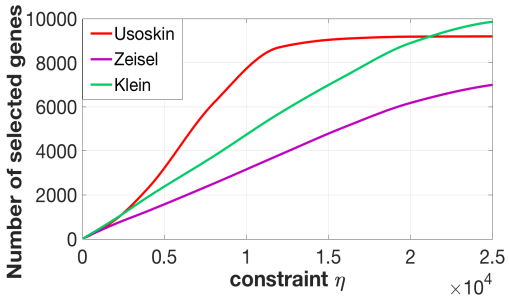

Fig. 1: Evolution of the number of selected genes versus the constraint is a smooth monotonous function. In our constrained approach, the parameter η is directly connected to the number of genes.

3.3. Evaluation and comparison between methods

We observe in Tables 1, 2 and 3 that our k-sparse algorithm significantly improves the results of Sparcl and SIMLR methods in terms of silhouette, accuracy, ARI and NMI.

Moreover, Fig. 3 illustrates the clustering results of each method in a 2D visualization using the `tsne` representation [32]. Each colored point represents a cell and misclassified cells are reported in black. We can observe that our k-sparse

[1] https://github.com/BatzoglouLabSU/SIMLR/tree/SIMLR/MATLAB

algorithm significantly improves visually the results of Sparcl and SIMLR methods. Note that SIMLR fails to discover one class on Usoskin.

Table 1: **Usoskin dataset** (4 clusters, 622 cells, 9,195 genes): Comparison between methods and with real labels. With $\eta = 5000$, k-sparse selected $3,095$ genes out of a total of a total of 9,195 and outperforms others methods in terms of accuracy by 15%.

Usoskin	Spectral	SIMLR	Sparcl	k-sparse
Silhouette	0.61	0.88	-	**0.95**
Accuracy (%)	60.13	76.37	57.24	**92.60**
ARI (%)	26.46	67.19	31.30	**87.42**
NMI	0.33	0.75	0.39	**0.85**
Time (s)	**0.91**	15.67	1,830	57.07

Table 2: **Klein dataset** (4 clusters, 2,717 cells, 10,322 genes): Comparison between methods and with real labels. For $\eta = 25000$, k-sparse selected $9,870$ genes out of a total of a total of 10,332 and has an accuracy close to 100%. SIMLR has similar performances (accuracy, ARI and NMI) than k-sparse (which is 5 times faster than SIMLR).

Klein	Spectral	SIMLR	Sparcl	k-sparse
Silhouette	0.73	0.95	-	**0.96**
Accuracy (%)	63.31	**99.12**	65.11	**99.12**
ARI (%)	38.91	**98.34**	45.11	**98.34**
NMI	0.54	**0.97**	0.56	**0.97**
Time (s)	**20.81**	511	30,384	97.10

Table 3: **Zeisel dataset** (9 clusters, 3,005 cells, 7,364 genes): Comparison between methods and with real labels. With $\eta = 12500$, k-sparse selected $3,981$ genes out of a total of a total of 7,364 and outperforms others methods in terms of accuracy by 16%. For this clustering K-sparse is 6 times faster than SIMLR.

Zeisel	Spectral	SIMLR	Sparcl	k-sparse
Silhouette	0.56	0.82	-	**0.83**
Accuracy (%)	59.30	71.85	65.23	**88.15**
ARI (%)	50.55	64.8	59.06	**84.17**
NMI	0.68	0.75	0.69	**0.81**
Time (s)	**23**	464	28,980	71.60

Fig. 2 illustrates that K-sparse converged within around $L = 10$ loops on each database. Regarding the computation time of each algorithm, Ksparse was particularly faster than SIMLR on the two largest datasets, as reported in Table 2 and Table 3. Finally, let us note that optimizing the values of the Lagrangian parameter using permutations (as in the Sparcl algorithm) is computationally more expensive than the projection onto the the ℓ_1 ball [17, 18] (as in our Ksparse approach). Our Matlab code is available at `github.com/cypgilet/Ksparse_Clustering`.

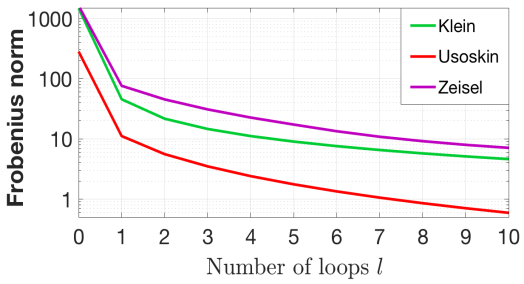

Fig. 2: Decay of the Frobenius norm for the three datasets versus the number of loops of the alternating minimization scheme emphasizes the fast and smooth convergence of our algorithm.

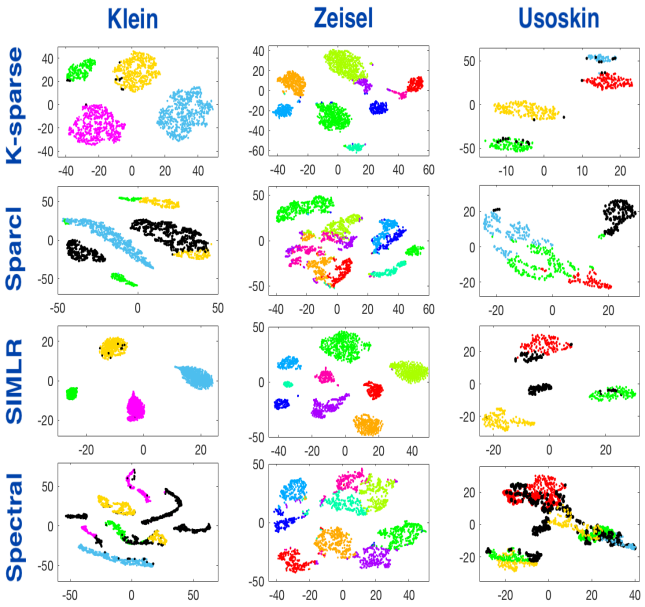

Fig. 3: Comparison of 2D visualization using `tsne` [32]. Each point represents a cell. Misclassified cells are reported in black.

4. CONCLUSION

In this paper, we focus on unsupervised clustering. We provide a new efficient algorithm based on alternating minimization introducing an ℓ^1 constraint in the gradient-projection step. This step, of splitting type, uses an exact projection on the ℓ^1 ball to promote sparsity, and is alternated with k-means. Convergence of the projection-gradient method is established, and each iterative step of our algorithm necessarily lowers the cost. Experiments on single-cell RNA-seq dataset in Section 3 demonstrate the efficiency of our method.

5. REFERENCES

[1] J.-B. McQueen, "Some methods for classification and analysis of multivariate observations," *Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1967.

[2] D. Arthur and S. Vassilvitski, "k-means++: The advantages of careful seeding," *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, 2007.

[3] M. Radovanovic, A. Nanopoulos, and M. Ivanovic, "Hubs in space : Popular nearest neighbors in high-dimensional data," *Journal of Machine Learning Research. 11: 2487?2531.*, 2010.

[4] F. de la Torre and T. Kanade, "Discriminative cluster analysis," *ICML 06 Proceedings of the 23rd international conference on Machine learning, Pittsburgh, Pennsylvania, USA*, 2006.

[5] C. Ding and T. Li, "Adaptive dimension reduction using discriminant analysis and k-means clustering," in *Proceedings of the 24th International Conference on Machine Learning*, ser. ICML '07. ACM, 2007, pp. 521–528.

[6] F. R. Bach and Z. Harchaoui, "Diffrac: a discriminative and flexible framework for clustering," in *Advances in Neural Information Processing Systems 20*, J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis, Eds. Curran Associates, Inc., 2008, pp. 49–56.

[7] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in *Advances in Neural Information Processing Systems 14*, T. G. Dietterich, S. Becker, and Z. Ghahramani, Eds. MIT Press, 2002, pp. 849–856.

[8] U. Von Luxburg, "A tutorial on spectral clustering," *Statistics and computing*, 2007.

[9] D. M. Witten and R. Tibshirani, "A framework for feature selection in clustering," *Journal of the American Statistical Association*, vol. 105, no. 490, pp. 713–726, 2010.

[10] S. Z. Selim and M. A. Ismail, "K-means-type algorithms: A generalized convergence theorem and characterization of local optimality," *IEEE Trans. Patt. An. Machine Intel.*, vol. PAMI-6, pp. 81–87, 1984.

[11] P. L. Combettes and J.-C. Pesquet, "Proximal splitting methods in signal processing," in *Fixed-point algorithms for inverse problems in science and engineering*. Springer, 2011, pp. 185–212.

[12] P.-L. Lions and B. Mercier, "Splitting algorithms for the sum of two nonlinear operators," *SIAM Journal on Numerical Analysis*, vol. 16, no. 6, pp. 964–979, 1979.

[13] S. Mosci, L. Rosasco, M. Santoro, A. Verri, and S. Villa, "Solving structured sparsity regularization with proximal methods," in *Machine Learning and Knowledge Discovery in Databases*. Springer, 2010, pp. 418–433.

[14] S. Sra, S. Nowozin, and S. J. Wright, *Optimization for Machine Learning*. MIT Press, 2012.

[15] M. Barlaud, W. Belhajali, P. L. Combettes, and L. Fillatre, "Classification and regression using an outer approximation projection-gradient method," *IEEE TRANSACTIONS ON SIGNAL PROCESSING*, vol. 65, no. 17, pp. 4635–4643, 2017.

[16] H. H. Bauschke and P. L. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Springer, 2011.

[17] L. Condat, "Fast projection onto the simplex and the l1 ball," *Mathematical Programming Series A*, vol. 158, no. 1, pp. 575–585, 2016.

[18] G. Perez, M. Barlaud, L. Fillatre, and J.-C. Régin, "A filtered bucket-clustering method for projection onto the simplex and the ℓ_1 ball," *Mathematical Programming*, May 2019.

[19] L. Bottou and Y. Bengio, "Convergence properties of the k-means algorithms," in *Advances in Neural Information Processing Systems 7*, G. Tesauro, D. S. Touretzky, and T. K. Leen, Eds. MIT Press, 1995, pp. 585–592.

[20] F. Bunea, C. Giraud, M. Royer, and N. Verzelen, "PECOK: A convex optimization approach to variable clustering," no. 1606.05100, 2016.

[21] L. Condat, "A convex approach to k-means clustering and image segmentation," *HAL*, no. 01504799, 2017.

[22] H. Attouch, J. Bolte, P. Redont, and A. Soubeyran, "Proximal alternating minimization and projection methods for nonconvex problems: an approach based on the kurdyka-lojasiewicz inequality," *Mathematics of Operations Research*, 2010.

[23] J. Bolte, S. Sabach, and M. Teboulle, "Proximal alternating linearized minimization for nonconvex and nonsmooth problems," *Mathematical Programming*, vol. 146, no. 1, pp. 459–494, Aug 2014.

[24] D. Evanko, "Method of the year 2013: Methods to sequence the dna and rna of single cells are poised to transform many areas of biology and medicine." *Nature Methods, Vol 11*, 2014.

[25] A. Klein, "Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells," *Cell*, 2015.

[26] A. Zeisel, "Cell types in the mouse cortex and hippocampus revealed by single-cell rna-seq," vol. 347, no. 6226, pp. 1138–1142, 2015.

[27] D. Usoskin, "Unbiased classification of sensory neuron types by large-scale single-cell rna sequencing," *Nature Neuroscience*, vol. 18, pp. 145–153, 2015.

[28] R. Tibshirani, G. Walther, and T. Hastie, "Estimating the number of clusters in a data set via the gap statistic." *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 411–423, 2001.

[29] B. Wang, J. Zhu, E. Pierson, D. Ramazzotti, and S. Batzoglou, "Visualization and analysis of single-cell rna-seq data by kernel-based similarity learning," *Nature methods*, no. 14, 2017.

[30] L. Hubert and P. Arabie, "Comparing partitions," *Journal of Classification*, vol. 2, no. 1, pp. 193–218, 1985.

[31] P. Rousseuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, November 1987.

[32] L. J. P. Van der Maaten and G. E. Hinton, "Visualizing high-dimensional data using t-sne," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.