CELL SUBCLASS IDENTIFICATION IN SINGLE-CELL RNA-SEQUENCING DATA USING ORTHOGONAL NONNEGATIVE MATRIX FACTORIZATION

Shuai Wang[†], *Peng Wu*[‡†], *Manqi Zhou*[†], *Tsung-Hui Chang*[†⋆] *and Song Wu*[‡]

[†]School of Science & Engineering, The Chinese University of Hong Kong, Shenzhen 518172, China
[⋆]Shenzhen Research Institute of Big Data, Shenzhen 518172, China
[‡]Institute of Urological Surgery of Shenzhen University, Shenzhen 518000, China

ABSTRACT

Identification of cell subclasses using single-cell RNA-Sequencing (scRNA-Seq) data is of paramount importance since it uncovers the hidden biological processes within the cell population. While the nonnegative matrix factorization (NMF) model has been reported to be effective in various unsupervised clustering tasks, it may still produce inappropriate results for some scRNA-Seq datasets with heterogeneous structures. In this paper, we propose the use of an orthogonally constrained NMF (ONMF) model for the subclass identification problem of scRNA-Seq datasets. The ONMF model in general can provide improved clustering performance, but is challenging to solve. We present a computationally efficient algorithm based on optimization techniques of variable splitting and alternating direction method of multipliers (ADMM). Through two scRNA-Seq datasets, we show that the proposed method can yield promising performance in identifying cell subclasses and detecting key genes over the existing methods. Moreover, the key genes identified by the proposed method are shown biologically significant via the gene set enrichment analysis.

***Index Terms*—** Cell subclass identification, single-cell RNA-Seq, orthogonal nonnegative matrix factorization, unsupervised clustering, gene extraction.

1. INTRODUCTION

Identification of cell subclasses and corresponding biological factors is of paramount importance for studying the hidden biological processes within a cell population. The subclasses detected may expose some previously undefined cell subtypes [1] or reveal the process of cell differentiation [2]. Meanwhile, the technique of single-cell RNA-Sequencing (scRNA-Seq) has been used recently to detect heterogeneity within the cell population at a high resolution [3]. Specifically, a group of single cells are clustered in an unsupervised fashion to elucidate cell subclasses and identify key genes. However, due to the ubiquitous noise in scRNA-Seq data [4], the task of unsupervised clustering on a group of seemingly similar cell samples remains challenging.

Numerous clustering methods have been applied to bioinformatics for classifying cells based on gene expression, including K-means, hierarchical clustering, principal component analysis (PCA) and nonnegative matrix factorization (NMF), to name a few. Among them, NMF recently gains significant attention in various aspects of computational biology [5], including molecular pattern discovery, class comparison and prediction, cross-platform and cross species analysis. More importantly, it has been shown to be a powerful tool to detect subclasses among cell samples due to its higher clustering accuracy and ability to extract key genes associated with each subclass [3, 6, 7, 8]. Despite of these successful examples, NMF may still fail in clustering some datasets with heterogeneous structures. Recently, it has been found that the orthogonally constrained NMF (ONMF) formulation [9, 10] is closely related to the K-means clustering and can provide improved clustering performance in various data mining tasks [11, 12, 13]. However, ONMF is less noticed in the literature of biological data analysis [14, 15], especially for scRNA-Seq datasets.

In this paper, we are interested in the use of ONMF for cell subclass identification and key gene extraction in scRNA-Seq datasets. Unfortunately, due to the orthogonality constraint, the ONMF formulation is much more challenging to solve than NMF. Existing algorithms for ONMF include the method by [9] which combines the multiplicative rule [16] and the penalty method in optimization, and augmented Lagrangian (AL) based methods [10, 17]. However, numerical experiences suggest that the non-parametric method in [9] cannot provide a good trade-off between the low-rank approximation accuracy and satisfaction of the orthogonality constraint. Thus the orthogonality constraint cannot be satisfied with high accuracy in general. The AL methods [10, 17] are less computationally efficient since they involve solving complicated subproblems.

To overcome these issues and inspired by the splitting of orthogonality constraint (SOC) method in [18], we present a new method based on the optimization techniques of variable splitting and alternating direction method of multipliers (VS-ADMM) [19] for the ONMF formulation. Specifically, in VS-ADMM, the updates of variables all have simple closed-form solutions and therefore are computationally efficient. Moreover, the VS-ADMM is amenable to yield near-orthogonal solutions, which is crucial for unsupervised clustering tasks. By considering two scRNA-Seq datasets, we examine the performance of the proposed VS-ADMM algorithm for cell subclasss identification and key gene extraction and perform comparison with the existing methods. Numerical results show that the proposed method yields superior performance over the existing methods. More importantly, the performed gene set enrichment analysis show that the key genes associated with each subclass as identified by the proposed method have significant biological meanings. In particular, for the bladder cancer dataset provided by the Institute of Urological Surgery of Shenzhen University, the extracted key genes are clearly responsible for known characteristics of cancer cells.

2. PROPOSED METHOD

2.1. NMF and ONMF Formulation

We consider a scRNA-Seq data set consisting of the expression levels of M genes of N cell samples. The data set is denoted by the non-negative matrix $\mathbf{X} \in \mathbb{R}_+^{M\times N}$. Typically, $M \gg N$ since the number of cells is usually small compared to the number of genes.

The NMF model has been found to be effective in numerous unsupervised learning tasks, including cell subclass identification for some scRNA-Seq datasets [3]. For the clustering task, the use of NMF is to find a low-dimensional representation of $\mathbf{X}$ via low-rank approximation, in order to improve both the computational and clustering performance. The NMF decomposes $\mathbf{X}$ into two nonnegative matrix factors $\mathbf{W} \in \mathbb{R}_+^{M\times K}$ and $\mathbf{H} \in \mathbb{R}_+^{K\times N}$ so that $\mathbf{X} \approx \mathbf{WH}$, where $K \ll \min\{N, M\}$ is the reduced dimension. Then the low-dimensional matrix $\mathbf{H}$ is used as an input to classical clustering methods such as K-means and hierarchical clustering. The matrix decomposition is implemented via solving the following optimization problem

$$\begin{aligned} \min_{\mathbf{W},\mathbf{H}} \ & \|\mathbf{X} - \mathbf{WH}\|_{\mathrm{F}}^2 \\ \text{s.t. } & \mathbf{W} \geq 0, \mathbf{H} \geq 0, \end{aligned} \tag{1}$$

where $\|\cdot\|_F$ denotes the Forbenius norm and $\mathbf{W} \geq 0$ ($\mathbf{H} \geq 0$) means that all elements of $\mathbf{W}$ ($\mathbf{H}$) are nonnegative [16].

In fact, it is known that the NMF model above has a strong connection with K-means [9]. Specifically, let K be the number of clusters, and let

$$[\mathbf{H}]_{k,n} = \begin{cases} \frac{1}{N_k} & \text{if cell } n \text{ belongs to cluster } k, \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

for all $k = 1, \ldots, K$, and $n = 1, \ldots, N$, where N_k is the number of cells assigned to cluster k. Then $\mathbf{X} \approx \mathbf{WH}$ can be interpreted as that each of the N cell samples is approximated by the kth column of $\mathbf{W} \triangleq [\boldsymbol{w}_1, \ldots, \boldsymbol{w}_K]$ if the cell is associated with cluster k. The column vector $\boldsymbol{w}_k$ is therefore the centroid of the kth cluster if $\boldsymbol{w}_k$ is the one that minimizes the average Euclidean distances of all cells associated with cluster k. Note that (2) is equivalent to

$$\mathbf{HH}^T = \mathbf{I}_K, \ [\mathbf{H}]_{k,n} \in \{0, \frac{1}{N_k}\}, \ \forall k, n, \tag{3}$$

where $\mathbf{I}_K$ is the $K \times K$ identity matrix. Thus the NMF model in (1) can be regarded as a *relaxed* formulation for K-means clustering since the NMF ignores the conditions that the rows of $\mathbf{H}$ are orthonormal and each (k, n)th element is either zero or $\frac{1}{N_k}$. However, the condition of orthogonality and set discreteness are difficult to deal with in general from an optimization point of view.

Therefore, the orthogonally constrained NMF (ONMF) model proposed in [9] is a tradeoff between the NMF model and the exact K-means clustering problem. Specifically, the ONMF problem is given by

$$\min_{\mathbf{W},\mathbf{H}} \ \|\mathbf{X} - \mathbf{WH}\|_{\mathrm{F}}^2 \tag{4a}$$

$$\text{s.t. } \mathbf{W} \geq 0, \mathbf{H} \geq 0, \tag{4b}$$

$$\mathbf{HH}^T = \mathbf{I}_K. \tag{4c}$$

It is worthy noting that the orthogonality constraint, together with the non-negativity constraint, enforces only one entry in each column of $\mathbf{H}$ to be nonzero. The nonzero entry thereby explicitly indicates the cluster index of the cell. As a result, an improved clustering performance is expected when comparing to the classical NMF model in (1). However, the orthogonality constraint (4c) makes the matrix decomposition problem even more difficult to handle.

For the ONMF model in (4), reference [9] proposed an algorithm by combining the multiplicative rule [16] and the penalty method. This algorithm is also applied to clustering problems in biomedical applications such as cancer cell clustering and integrative data analysis [15, 14]. Reference [18] proposed a splitting orthogonality constraint (SOC) method that relies on variable splitting and ADMM [19]. However, the SOC method neither considered non-negativity constraints nor applications of cell subclass identification in scRNA-Seq data. In the next subsection, we extend the idea of the SOC method to handle the ONMF model in (4).

2.2. Proposed VS-ADMM Algorithm

The ONMF model in (4) is challenging to solve because both the objective function and the orthogonality constraint are non-convex. By the fact that either projection onto the non-negative set or projection onto the orthogonality constraint is simple and has closed-form expression, we leverage the variable splitting technique and ADMM [19] to solve problem (4) in an efficient manner. Specifically, let us consider the following problem

$$\begin{aligned} \min_{\mathbf{W},\mathbf{H},\mathbf{S},\mathbf{P},\mathbf{Y}} \ & \|\mathbf{X} - \mathbf{WH}\|_{\mathrm{F}}^2, \\ \text{s.t. } & \mathbf{W} = \mathbf{S}, \mathbf{H} = \mathbf{P}, \mathbf{H} = \mathbf{Y}, \\ & \mathbf{S} \geq 0, \mathbf{P} \geq 0, \mathbf{YY}^T = \mathbf{I}_K. \end{aligned} \tag{5}$$

As seen, variables $(\mathbf{S}, \mathbf{P})$ are introduced for splitting the non-negative constraint from $\mathbf{W}$ and $\mathbf{H}$, respectively, and variable $\mathbf{Y}$ is used to split the orthogonality constraint from $\mathbf{H}$.

The second ingredient of the proposed method is ADMM. According to ADMM, we consider the (partial) augmented Lagrangian function of (5), which is given by

$$\begin{aligned} \mathcal{L}_{\mathrm{a}}(\mathbf{W}, \mathbf{H}, \mathbf{S}, \mathbf{P}, \mathbf{Y}, \mathbf{\Lambda}) \ = \ & \frac{1}{2}\|\mathbf{X} - \mathbf{WH}\|_{\mathrm{F}}^2 \\ + \mathrm{Tr}(\mathbf{\Lambda}_1^T(\mathbf{W} - \mathbf{S})) \ + \ & \frac{\rho_1}{2}\|\mathbf{W} - \mathbf{S}\|_{\mathrm{F}}^2 \\ + \mathrm{Tr}(\mathbf{\Lambda}_2^T(\mathbf{H} - \mathbf{P})) \ + \ & \frac{\rho_2}{2}\|\mathbf{H} - \mathbf{P}\|_{\mathrm{F}}^2 \\ + \mathrm{Tr}(\mathbf{\Lambda}_3^T(\mathbf{H} - \mathbf{Y})) \ + \ & \frac{\rho_3}{2}\|\mathbf{H} - \mathbf{Y}\|_{\mathrm{F}}^2, \end{aligned} \tag{6}$$

where $\mathbf{\Lambda} \triangleq (\mathbf{\Lambda}_1, \mathbf{\Lambda}_2, \mathbf{\Lambda}_3)$ in which $\mathbf{\Lambda}_1 \in \mathbb{R}^{M\times K}, \mathbf{\Lambda}_2 \in \mathbb{R}^{K\times N}$ and $\mathbf{\Lambda}_3 \in \mathbb{R}^{K\times N}$ are Lagrangian dual variables associated with linear equality constraints in (5), and $\rho_1, \rho_2, \rho_3 > 0$ are the penalty parameters for the augmented terms.

The ADMM iteratively minimizes the augumented Lagrangian function (6) with respect to the primal variables $(\mathbf{W}, \mathbf{H}, \mathbf{S}, \mathbf{P}, \mathbf{Y})$ in a Gauss-Seidel fashion, followed by updating the dual variables using gradient ascent. Specifically, at each iteration r, we sequentially perform the following updates

$$\mathbf{W}^{r+1} \leftarrow \arg\min_{\mathbf{W}} \mathcal{L}_{\mathrm{a}}(\mathbf{W}, \mathbf{H}^r, \mathbf{S}^r, \mathbf{P}^r, \mathbf{Y}^r, \mathbf{\Lambda}^r), \tag{7a}$$

$$\mathbf{H}^{r+1} \leftarrow \arg\min_{\mathbf{H}} \mathcal{L}_{\mathrm{a}}(\mathbf{W}^{r+1}, \mathbf{H}, \mathbf{S}^r, \mathbf{P}^r, \mathbf{Y}^r, \mathbf{\Lambda}^r), \tag{7b}$$

$$\mathbf{S}^{r+1} \leftarrow \arg\min_{\mathbf{S}\geq 0} \mathcal{L}_{\mathrm{a}}(\mathbf{W}^{r+1}, \mathbf{H}^{r+1}, \mathbf{S}, \mathbf{P}^r, \mathbf{Y}^r, \mathbf{\Lambda}^r), \tag{7c}$$

$$\mathbf{P}^{r+1} \leftarrow \arg\min_{\mathbf{P}\geq 0} \mathcal{L}_{\mathrm{a}}(\mathbf{W}^{r+1}, \mathbf{H}^{r+1}, \mathbf{S}^{r+1}, \mathbf{P}, \mathbf{Y}^r, \mathbf{\Lambda}^r), \tag{7d}$$

$$\mathbf{Y}^{r+1} \leftarrow \arg\min_{\mathbf{YY}^T=\mathbf{I}_K} \mathcal{L}_{\mathrm{a}}(\mathbf{W}^{r+1}, \mathbf{H}^{r+1}, \mathbf{S}^{r+1}, \mathbf{P}^{r+1}, \mathbf{Y}, \mathbf{\Lambda}^r), \tag{7e}$$

$$\mathbf{\Lambda}_1^{r+1} \leftarrow \mathbf{\Lambda}_1^r + \rho_1(\mathbf{W}^{r+1} - \mathbf{S}^{r+1}), \tag{7f}$$

$$\mathbf{\Lambda}_2^{r+1} \leftarrow \mathbf{\Lambda}_2^r + \rho_2(\mathbf{H}^{r+1} - \mathbf{P}^{r+1}), \tag{7g}$$

$$\mathbf{\Lambda}_3^{r+1} \leftarrow \mathbf{\Lambda}_3^r + \rho_3(\mathbf{H}^{r+1} - \mathbf{Y}^{r+1}). \tag{7h}$$

Interestingly, thanks to the variable splitting, all the subproblems in (7a) to (7e) admit simple closed-form solutions. Specifically, (7a)

and (7e) are unconstrained least squares problems with respect to $\mathbf{W}$ and $\mathbf{H}$. They have solutions as

$$\mathbf{W}^{r+1} \leftarrow (\mathbf{X}(\mathbf{H}^r)^T + \rho_1 \mathbf{S}^r - \mathbf{\Lambda}_1^r)[\mathbf{H}^r(\mathbf{H}^r)^T + \rho_1 \mathbf{I}_K]^{-1}, \quad (8)$$

$$\mathbf{H}^{r+1} \leftarrow (\mathbf{X}^T\mathbf{W}^{r+1} + \rho_2 \mathbf{P}^r - \mathbf{\Lambda}_2^r + \rho_3 \mathbf{Y}^r - \mathbf{\Lambda}_3^r) \times [(\mathbf{W}^{r+1})^T\mathbf{W}^{r+1} + (\rho_2 + \rho_3)\mathbf{I}_K]^{-1}. \quad (9)$$

It can be shown that subproblems (7c) and (7d) are equivalent to projecting $\mathbf{W}^{r+1} + \frac{1}{\rho_1}\mathbf{\Lambda}_1^r$ and $\mathbf{H}^{r+1} + \frac{1}{\rho_2}\mathbf{\Lambda}_2^r$ onto the non-negative set, which respectively have solutions as

$$\mathbf{S}^{r+1} \leftarrow \max(\mathbf{W}^{r+1} + \mathbf{\Lambda}_1^r/\rho_1, 0), \quad (10)$$

$$\mathbf{P}^{r+1} \leftarrow \max(\mathbf{H}^{r+1} + \mathbf{\Lambda}_2^r/\rho_2, 0), \quad (11)$$

where $\max(\cdot, 0)$ takes point-wise maximum value between the input matrix and zero. Subproblem (7e) involves projecting $\mathbf{H}^{r+1} + \frac{1}{\rho_3}\mathbf{\Lambda}_3^r$ onto the set of orthogonal matrices $\mathbf{Y}\mathbf{Y}^T = \mathbf{I}_K$. It is shown in [18] that this problem has a closed-from solution

$$\mathbf{Y}^{r+1} \leftarrow \mathbf{V}\mathbf{U}^T. \quad (12)$$

where $\mathbf{V} \in \mathbb{R}^{K\times K}$ and $\mathbf{U} \in \mathbb{R}^{N\times K}$ are the left and right singular vector matrices of $\mathbf{H}^{r+1} + \frac{1}{\rho_3}\mathbf{\Lambda}_3^r$, i.e., $\mathbf{H}^{r+1} + \frac{1}{\rho_3}\mathbf{\Lambda}_3^r = \mathbf{V}\mathbf{\Sigma}\mathbf{U}^T$ for some diagonal singular value matrices $\mathbf{\Sigma} \in \mathbb{R}^{K\times K}$.

As seen from (7) to (12), the proposed VS-ADMM algorithm is computationally efficient since each update has a close-form solution. In particular, the matrix inversion and singular value decomposition (SVD) required in (8), (9) and (12) involves matrices with dimension K and N which are much smaller than the number of genes in scRNA-Seq data. Another key advantage of the proposed VS-ADMM algorithm is that it allows a more flexible tradeoff between the feasibility of the orthogonality constraint and the low-rank approximation accuracy. This is in contrast to the non-parameter algorithm in [9] which does not allow such flexibility. Specifically, since variable $\mathbf{Y}^r$ satisfies the orthogonality constraint for all iterations in the proposed algorithm, the feasibility of the orthogonality constraint for variable $\mathbf{H}^r$ can be controlled by the proximity between $\mathbf{Y}^r$ and $\mathbf{H}^r$ through a proper choice of the penalty parameter ρ_3. This flexibility is important for clustering tasks since satisfying the orthogonality constraint better is usually more helpful for improving the clustering performance than achieving a lower low-rank approximation error, as we will demonstrate in Section 3.

2.3. Subclass Identification, Key Gene Extraction and Biological Significance Analysis

The outcome of ONMF, i.e., $\mathbf{H}$ and $\mathbf{W}$, can be further used for identifying subclasses of cells and finding key genes for each of the subclass. For subclass identification, while the low-dimensional matrix $\mathbf{H}$ obtained by ONMF is theoretically a subclass indicator matrix, one can improve the performance by further applying the K-means to $\mathbf{H}$. In particular, one may initialize the K-means with an initial subclass association obtained by assigning each cell n to subclass $\hat{k}$ if $\hat{k} = \arg\max_{k=1,\ldots,K} |[\mathbf{H}]_{k,n}|$, for all $n = 1, \ldots, N$. The number of subclasses K can be determined by the method of consensus clustering used in [6]. The key genes associated with one subclass should be these genes that are significant in their expression level only for cells in the subclass and weak for cells in other subclasses. Since the columns of $\mathbf{W}$ correspond to the centroids of the subclasses, the significance of each gene across subclasses can be identified based on $\mathbf{W}$. We adopt the scoring scheme in [7] which computes an entropy-related score for each row of $\mathbf{W}$. The genes are ranked according to the scores and those that are top ranked (e.g., the first 1000 ranked genes) are selected as the key genes for subclasses.

To understand the biological meanings of the subclasses and key genes, we further perform gene set enrichment analysis (GSEA) [20] using Kyoto Encyclopedia of Genes and Genomes (KEGG) [21] and REACTOME [22]. Both databases are widely used in biomedical research dealing with genomes, biological pathways, diseases and drugs.

Table 1. ScRNA-Seq datasets

Dataset	Samples	Genes	Clusters
1 Mouse Embryonic Fibroblasts	405	12117	5
2 Bladder Cancer	121	23048	4

Table 2. Clustering Performance of Different Methods for Dataset 1

	Purity	Rand Index	Sihouette
K-means	0.708	0.427	0.060
NMF (Euclidean) in [16]	0.731	0.483	0.538
NMF (KL) in [16]	0.742	0.489	0.616
DTPP in [9]	0.741	0.491	0.680
Proposed VS-ADMM	0.749	0.506	0.803

3. NUMERICAL RESULTS AND DISCUSSIONS

3.1. Datasets

To demonstrate the effectiveness of the proposed VS-ADMM algorithm, we consider two scRNA-Seq datasets as summarized in Table 1. Dataset 1 is publicly available from [23], and the data has clear labels for 5 subclasses. Dataset 2 is provided by the Institute of Urological Surgery of Shenzhen University. The dataset does not have prior subclass labels and is determined by the proposed algorithm as stated in Section 2.3.

3.2. Algorithm Convergence Performance

Let us first examine the convergence behavior of the proposed VS-ADMM algorithm and the comparison with the algorithm by Ding *et al.* (DTPP) in [9]. In Fig. 1(a) and Fig. 1(b), we present the normalized objective value with respect to the iteration number of the two algorithms applied to Dataset 1 and Dataset 2, respectively. In Fig. 1(c), we show the feasibility of the orthogonality constraint (measured by $\|\mathbf{H}\mathbf{H}^T - \mathbf{I}_K\|_{\mathrm{F}}^2/K^2$) achieved by the two algorithms when applied to the two datasets. The parameters in the experiment are set to $\rho_1 = 1 \times 10^{-4}$, $\rho_2 = 8.52 \times 10^{-3}$, $\rho_3 = 1 \times 10^{-1}$ for Dataset 1 and $\rho_1 = 2 \times 10^{-3}$, $\rho_2 = 3.5 \times 10^{3}$, $\rho_3 = 4.5 \times 10^{4}$ for Dataset 2. As one can see, for both datasets, while the proposed VS-ADMM algorithm achieves comparable or slightly higher objective values than the DTPP algorithm, the constraint feasibility achieved by the proposed algorithm is consistently lower. This implies that the proposed VS-ADMM gives better clustering performance, as we demonstrate next.

3.3. Subclass Identification Performance

Since Dataset 1 has labeled information about subclasses, we can examine the clustering performance of the proposed algorithm by comparing the obtained results with the labels. In addition to the proposed VS-ADMM and DTPP in [9], we also tested the K-means, and the classical NMF method with a Euclidean distance and KL divergence objective functions [16]. Three performance measures, namely Purity [24], Rand Index [25] and Average Silhouette Width [26], are evaluated by averaging over 10 trials each of which uses a randomly generated initial point for the algorithms. The results are displayed in Table 2. As seen, the proposed VS-ADMM algorithm

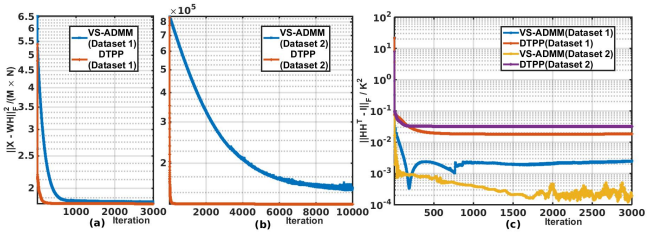

Fig. 1. The convergence of objective value and feasibility of orthogonality constraint of DTPP in [9] and proposed VS-ADMM algorithms applied to Datasets 1 and 2.

Table 3. Gene set enrichment analysis for key genes of Dataset 2 extracted by the proposed method.

Cluster	Biological Pathway	Genes	FDR q-value
1 (57 genes)	LIPID_TRANSPORTER_ ACTIVITY	3	1.89E-1
	ENDOMEMBRANE_ SYSTEM_ ORGANIZATION	4	2.07E-1
	CELLULAR_LIPID_ METABOLIC_PROCESS	5	2.07E-1
2 (24 genes)	NEGATIVE_ REGULATION_ OF_LIPASE_ACTIVITY	2	3.03E-2
	REGULATION_OF_ LIPASE_ACTIVITY	2	4.87E-1
3 (58 genes)	HISTONE_ DEMETHYLASE_ ACTIVITY	2	1.27E-1
	HISTONE_BINDING	3	1.27E-1
4 (861 genes)	IMMUNE_SYSTEM	33	1.85E-3
	CELL_CYCLE	19	3.46E-3
	MISMATCH_REPAIR	4	1.52E-2
	PPAR_SIGNALING_ PATHWAY	6	1.90E-2
	CELL_CYCLE_ CHECKPOINTS	8	1.90E-2
	REGULATION_OF_ ACTIN_CYTOSKELETON	10	3.99E-2

yields the best clustering performance over the other methods under test.

For Dataset 2, we illustrate the performance of subclass identification and key gene extraction by showing the heatmap of clustering results (the distance matrix based on the Pearson coefficient of entries of $\mathbf{H}$) and expression level of the top ranked 1000 genes in Fig. 2. Firstly, one can see from the heatmaps that, the proposed VS-ADMM provides a satisfactory result with four well-separated clusters, in contrast to the NMF and DTPP in [9]. Secondly, from the gene expression levels, one can observe that the proposed VS-ADMM gives a clear pattern for the key genes that are uniquely associated with each of the subclasses whereas the NMF and DTPP don't. Specifically, the 1000 key genes extracted by the VS-ADMM include 57 genes for Cluster 1, 24 genes for Cluster 2, 58 genes for Cluster 3 and 861 genes for Cluster 4.

3.4. Biological Significance Analysis of Extracted Key Genes

We validate the biological significance of the key genes of Dataset 2 by performing the gene set enrichment analysis (see Section 2.3). Table 3 presents the corresponding biological pathways or processes

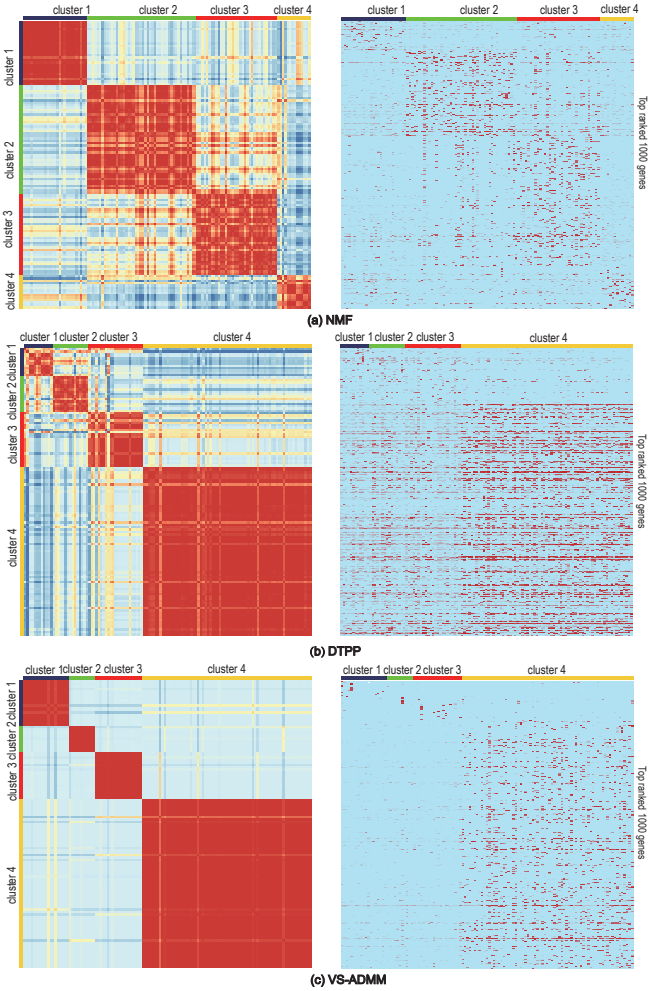

Fig. 2. The heatmaps of clustering results (Left) and expression level of the top ranked 1000 genes (Right) obtained by the NMF, ONMF using DTPP and ONMF using proposed VS-ADMM applied to Dataset 2.

for key genes of each cell subclass. Here, the analytical results are usually recognized as significant if the false discovery rate (FDR) q-value is less than 0.05 [27]. It is found that for Clusters 1, 2 and 3, the corresponding biological processes are not only less significant due to higher FDR q-values but also less relevant to cancer cells. In contrast, the key genes for Cluster 4 are enriched with pathways such as immune system, PPAR (Peroxisome proliferator-activated receptor) signal pathway, cell cycle and cytoskeleton, which are all important and commonly identified in cancer cells. Cell cycle and immune system pathways are usually found in cancer cells responsible for characteristics of immortalization and immune evasion of tumor. PPAR signal pathway related genes have already been found with high expression in muscle-invasive bladder cancer [28]. Also, changes in cytoskeleton, also known as a process called epithelial-mesenchymal transition, are essential for cancer cell invasion and metastasis [29]. These findings are consistent with the muscle-invasive features of bladder cancer cells of Dataset 2. In summary, the above results well demonstrate the heterogeneity of the dataset as well as the effectiveness of the proposed method by VS-ADMM for finding biologically significant cell subclasses and key genes for scRNA-Seq datasets.

4. REFERENCES

[1] F. Buettner, K. N. Natarajan, F. P. Casale, V. Proserpio, A. Scialdone, F. J. Theis, S. A. Teichmann, J. C. Marioni, and O. Stegle, "Computational analysis of cell-to-cell heterogeneity in single-cell RNA-Sequencing data reveals hidden subpopulations of cells," *Nature Biotechnology*, vol. 33, no. 2, pp. 155–160, Feb. 2015.

[2] L. Yan, M. Yang, H. Guo, L. Yang, J. Wu, R. Li, P. Liu, Y. Lian, X. Zheng, J. Yan, J. Huang, M. Li, X. Wu, L. Wen, K. Lao, R. Li, J. Qiao, and F. Tang, "Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells," *Nature Structural & Molecular Biology*, vol. 20, no. 9, pp. 1131–1139, Jul. 2013.

[3] X. Zhu, T. Ching, X. Pan, S. M. Weissman, and L. Garmire, "Detecting heterogeneity in single-cell RNA-Seq data by non-negative matrix factorization," *PeerJ*, vol. 5, no. e2888, pp. 1–20, Jan. 2017.

[4] V. Y. Kiselev, K. Kirschner, M. T. Schaub, T. Andrews, A. Yiu, T. Chandra, K. N. Natarajan, W. Reik, M. Barahona, A. R. Green, and M. Hemberg, "SC3: consensus clustering of single-cell RNA-Seq data," *Nature Methods*, vol. 14, no. 5, pp. 483–486, Mar. 2017.

[5] K. Devarajan, "Nonnegative matrix factorization: an analytical and interpretive tool in computational biology," *PLoS Computational Biology*, vol. 4, no. 7, pp. 1–12, Jul. 2008.

[6] J.-P. Brunet, P. Tamayo, T. R. Golub, and J. P. Mesirov, "Metagenes and molecular pattern discovery using matrix factorization," in *Proc. Natl. Acad. Sci. USA*, Mar, 2004, pp. 4164–4169.

[7] H. Kim and H. Park, "Spare non-negative matrix factorization via alternating non-negative-constrained least squares for microarray data analysis," *Bioinformatics*, vol. 23, no. 12, pp. 1495–1502, May 2007.

[8] C.-H. Zheng, D.-S. Huang, L. Zhang, and X.-Z. Kong, "Tumor clustering using nonnegative matrix factorization with gene selection," *IEEE. Transactions on Information Technology in Biomedicine*, vol. 13, no. 4, pp. 599–607, July. 2009.

[9] C. Ding, T. Li, W. Peng, and H. Park, "Orthogonal nonnegative matrix t-factorizations for clustering," in *Proc. KDD*, Philadelphia, PA, USA, Aug. 2006, pp. 20–23.

[10] F. Pompili, N. Gillis, P. A. Absil, and F. Glineur, "Two algorithms for orthogonal nonnegative matrix factorization with application to clustering," *Neurocomputing*, vol. 141, no. 2, pp. 15–25, Oct. 2014.

[11] J. Yoo and S. Choi, "Nonnegative matrix factorization with orthogonality constraints," *Journal of Computer Science and Engineering*, vol. 4, no. 2, pp. 97–109, May. 2010.

[12] H. Mansour, S. Rane, P. T. Boufounos, and A. Vetro, "Video querying via compact descriptors of visually salient objects," in *Proc. ICIP*, Paris, France, Oct. 2014, pp. 2789–2793.

[13] S. Choi, "Algorithms for orthogonal nonnegative matrix factorization," in *Proc. IJCNN*, Hong Kong, China, 2008, pp. 1828–1832.

[14] A. Mirzal, "Nonparametric orthogonal NMF and its application in cancer clustering," in *Proc. DaEng*, Kuala Lumpur, Malaysia, Dec. 2013, pp. 177–184.

[15] M. Strazar, M. Zitnik, B. Zupan, J. Ule, and T. Curk, "Orthogonal matrix factorization enables integrative analysis of multiple RNA binding proteins," *Bioinformatics*, vol. 32, no. 10, pp. 1527–1535, May. 2016.

[16] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Proc. NIPS*, Denver, CO, USA, Dec. 2000, pp. 556–562.

[17] W. Chen, H. Ji, and Y. You, "An augumented lagrangian method for L1-regularized optimization problems with orthogonality constraints," *SIAM Journal on Scientific Computing*, vol. 38, no. 4, pp. B570–592, Apri. 2016.

[18] R. Lai and S. Osher, "A splitting method for orthogonality constrained problems," *Journal of Scientific Computing*, vol. 58, no. 2, pp. 431–449, Feb. 2014.

[19] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends in Machine Learning*, vol. 3, no. 1, pp. 1–122, 2010.

[20] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov, "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles," in *Proc. Natl. Acad. Sci. USA*, 2005, pp. 15545–15550.

[21] M. Kanehisa and S. Goto, "KEGG: kyoto encyclopedia of genes and genomes," *Nucleic Acids Res.*, vol. 28, no. 1, pp. 27–30, Jan. 2000.

[22] I. Vastrik, P. D. Eustachio, E. Schmidt, G. Joshi-Tope, G. Gopinath, D. Croft, B. Bono, M. Gillespie, B. Jassal, S. Lewis, L. Matthews, G. Wu, E. Birney, and L. Stein, "Reactome: a knowledge base of biologic pathways and processes," *Genome Biology*, vol. 8, no. 3, pp. R39.1–13, Mar. 2007.

[23] B. Treutlein, Q. Y. Lee, J. G. Camp, M. Mall, W. Koh, S. A. M. Shariati, S. Sim, N. F. Neff, J. M. Skotheim, M. Wernig, and S. R. Quake, "Dissecting direct reprogramming from fibroblast to neuron using single-cell RNA-Seq," *Nature*, vol. 534, no. 7607, pp. 391–395, Jun. 2016.

[24] C. D. Manning, P. Raghavan, and H. Schutze, *Introduction to Information Retrieval*, Cambridge University Press, Cambridge, UK, 2008.

[25] W. M. Rand, "Objective criteria for the evaluation of clustering methods," *Journal of the American Statistical Association*, vol. 66, no. 336, pp. 846–850, Dec. 1971.

[26] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, no. 336, pp. 53–65, Nov. 1987.

[27] W. S. Noble, "How does multiple testing correction work?," *Nature Biotechnology*, vol. 27, no. 12, pp. 1135–1137, Dec. 2009.

[28] W. Choi, B. Czerniak, A. Ochoa, X. Su, A. Siefker-Radtke, C. Dinney, and D. J. McConkey, "Intrinsic basal and luminal subtypes of muscle invasive bladder cancer," *Nature Review Urology*, vol. 11, no. 7, pp. 400–410, Jun. 2014.

[29] M. Yilmaz and G. Christofori, "EMT, the cytoskeleton, and cancer cell invasion," *Cancer and Metastasis Reviews*, vol. 28, no. 1-2, pp. 15–33, Jun. 2009.