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[‡]

[†] The Chinese Univ. of Hong Kong, Shenzhen, [⋆]Shenzhen Research Institute of Big Data, [‡]Institute of Urological Surgery of Shenzhen Univ.

I. Introduction

Identification of cell subclasses using scRNA-Seq data is of paramount importance since it uncovers the hidden biological processes within the cell population. Our contributions are as follows:

- Propose the use of orthogonally constrained NMF (ONMF) model and a computationally efficient algorithm based on variable splitting and ADMM for subclass identification.
- Obtain promising results in identifying cell subclasses and detecting key (biologically meaningful) genes on real-world data.

II. Problem Formulation

Consider a scRNA-Seq dataset consisting of the expression levels of M genes of N cell samples, denoted by matrix $\mathsf{X} \in \mathbb{R}^{M\times N}$.

The NMF Model

$$\min_{\boldsymbol{W},\boldsymbol{H}} \ \|\boldsymbol{X}-\boldsymbol{W}\boldsymbol{H}\|_{\mathrm{F}}^2 \quad \text{s.t. } \boldsymbol{W}\geq 0, \boldsymbol{H}\geq 0, \tag{1}$$

where $\|\cdot\|_F$ denotes the Frobenius norm and $\boldsymbol{W}\geq 0$ ($\boldsymbol{H}\geq 0$) means that all elements of $\boldsymbol{W}$ ($\boldsymbol{H}$) are nonnegative.

- NMF has been shown to be powerful in detecting subclasses among cell samples.
- NMF may still fail in clustering some datasets with heterogeneous structures.

The ONMF Model

$$\min_{\boldsymbol{W},\boldsymbol{H}} \ \|\boldsymbol{X}-\boldsymbol{W}\boldsymbol{H}\|_{\mathrm{F}}^2 \tag{2}$$

$$\text{s.t. } \boldsymbol{W}\geq 0, \boldsymbol{H}\geq 0, \ \boldsymbol{H}\boldsymbol{H}^T=\boldsymbol{I}_K. \tag{3}$$

- Orthogonally constrained NMF (ONMF) is closedly related to K-means clustering.
- The orthogonality and non-negativity constraint enforces H to be a cluster indicator matrix.

$$[\boldsymbol{H}]_{k,n}=\begin{cases} c\neq 0 & \text{if cell } n \text{ belongs to cluster } k,\\ 0 & \text{otherwise}\end{cases}$$

- The ONMF problem is more challenging to solve with orthogonality constraint.

III. Proposed Algorithm

Variable Splitting

$$\min_{\boldsymbol{W},\boldsymbol{H},\boldsymbol{S},\boldsymbol{P},\boldsymbol{Y}} \ \|\boldsymbol{X}-\boldsymbol{W}\boldsymbol{H}\|_{\mathrm{F}}^2, \tag{4}$$

$$\text{s.t. } \boldsymbol{W}=\boldsymbol{S}, \boldsymbol{H}=\boldsymbol{P}, \boldsymbol{H}=\boldsymbol{Y}, \tag{5}$$

$$\boldsymbol{S}\geq 0, \boldsymbol{P}\geq 0, \boldsymbol{Y}\boldsymbol{Y}^T=\boldsymbol{I}_K. \tag{6}$$

New variables $(\boldsymbol{S},\boldsymbol{P},\boldsymbol{Y})$ split the non-negativity or orthogonality constraint from $(\boldsymbol{W},\boldsymbol{H})$.

ADMM Based Updates

$$\mathcal{L}_{\mathrm{a}}(\boldsymbol{W},\boldsymbol{H},\boldsymbol{S},\boldsymbol{P},\boldsymbol{Y},\Lambda)=\frac{1}{2}\|\boldsymbol{X}-\boldsymbol{W}\boldsymbol{H}\|_{\mathrm{F}}^2$$
$$+\mathrm{Tr}(\Lambda_1^T(\boldsymbol{W}-\boldsymbol{S})+\frac{\rho_1}{2}\|\boldsymbol{W}-\boldsymbol{S}\|_{\mathrm{F}}^2+\mathrm{Tr}(\Lambda_2^T(\boldsymbol{H}-\boldsymbol{P}))$$
$$+\frac{\rho_2}{2}\|\boldsymbol{H}-\boldsymbol{P}\|_{\mathrm{F}}^2+\mathrm{Tr}(\Lambda_3^T(\boldsymbol{H}-\boldsymbol{Y}))+\frac{\rho_3}{2}\|\boldsymbol{H}-\boldsymbol{Y}\|_{\mathrm{F}}^2,$$

At each iteration r, we update as follows:

$$\boldsymbol{W}^{r+1}\leftarrow \arg\min_{\boldsymbol{W}} \mathcal{L}_{\mathrm{a}}(\boldsymbol{W},\boldsymbol{H}^r,\boldsymbol{S}^r,\boldsymbol{P}^r,\boldsymbol{Y}^r,\Lambda^r),$$
$$\boldsymbol{H}^{r+1}\leftarrow \arg\min_{\boldsymbol{H}} \mathcal{L}_{\mathrm{a}}(\boldsymbol{W}^{r+1},\boldsymbol{H},\boldsymbol{S}^r,\boldsymbol{P}^r,\boldsymbol{Y}^r,\Lambda^r),$$
$$\boldsymbol{S}^{r+1}\leftarrow \arg\min_{\boldsymbol{S}\geq 0} \mathcal{L}_{\mathrm{a}}(\boldsymbol{W}^{r+1},\boldsymbol{H}^{r+1},\boldsymbol{S},\boldsymbol{P}^r,\boldsymbol{Y}^r,\Lambda^r),$$
$$\boldsymbol{P}^{r+1}\leftarrow \arg\min_{\boldsymbol{P}\geq 0} \mathcal{L}_{\mathrm{a}}(\boldsymbol{W}^{r+1},\boldsymbol{H}^{r+1},\boldsymbol{S}^{r+1},\boldsymbol{P},\boldsymbol{Y}^r,\Lambda^r),$$
$$\boldsymbol{Y}^{r+1}\leftarrow \arg\min_{\boldsymbol{Y}\boldsymbol{Y}^T=\boldsymbol{I}_K} \mathcal{L}_{\mathrm{a}}(\boldsymbol{W}^{r+1},\boldsymbol{H}^{r+1},\boldsymbol{S}^{r+1},\boldsymbol{P}^{r+1},\boldsymbol{Y},\Lambda^r),$$
$$\Lambda_1^{r+1}\leftarrow \Lambda_1^r+\rho_1(\boldsymbol{W}^{r+1}-\boldsymbol{S}^{r+1}),$$
$$\Lambda_2^{r+1}\leftarrow \Lambda_2^r+\rho_2(\boldsymbol{H}^{r+1}-\boldsymbol{P}^{r+1}),$$
$$\Lambda_3^{r+1}\leftarrow \Lambda_3^r+\rho_3(\boldsymbol{H}^{r+1}-\boldsymbol{Y}^{r+1}).$$

Each of above steps has closed-form solution.

$$\boldsymbol{W}^{r+1}\leftarrow(\boldsymbol{X}(\boldsymbol{H}^r)^T+\rho_1\boldsymbol{S}^r-\Lambda_1^r)[\boldsymbol{H}^r(\boldsymbol{H}^r)^T+\rho_1\boldsymbol{I}_K]^{-1}$$
$$\boldsymbol{H}^{r+1}\leftarrow(\boldsymbol{X}^T\boldsymbol{W}^{r+1}+\rho_2\boldsymbol{P}^r-\Lambda_2^r+\rho_3\boldsymbol{Y}^r-\Lambda_3^r)$$
$$\times[(\boldsymbol{W}^{r+1})^T\boldsymbol{W}^{r+1}+(\rho_2+\rho_3)\boldsymbol{I}_K]^{-1}.$$
$$\boldsymbol{S}^{r+1}\leftarrow\max(\boldsymbol{W}^{r+1}+\Lambda_1^r/\rho_1,0),$$
$$\boldsymbol{P}^{r+1}\leftarrow\max(\boldsymbol{H}^{r+1}+\Lambda_2^r/\rho_2,0),$$
$$\boldsymbol{Y}^{r+1}\leftarrow\boldsymbol{V}\boldsymbol{U}^T\ \ (SVD\ of\ (\boldsymbol{H}^{r+1}+\frac{1}{\rho_3}\Lambda_3^r)=\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{U}^T)$$

Subclass Identification & Key Gene Extraction

- Apply k-means on $\boldsymbol{H}$ with an initial subclass association from $\boldsymbol{H}$ to get clustering result.
- Adopt the scoring scheme [1] to obtain the significance of each gene from rows of $\boldsymbol{W}$ and key genes are those top ranked.

IV. Numerical Result I

Target Datasets

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

Figure 1: ScRNA-Seq datasets

Algorithm Convergence Performance

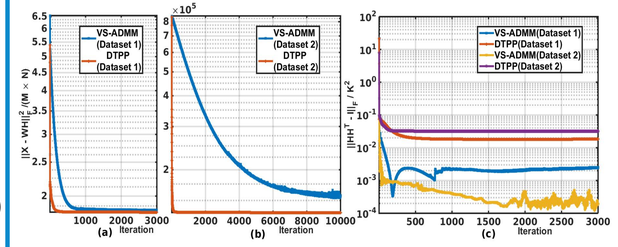

Figure 2: The convergence of objective value and feasibility of orthogonality constraint of DTPP and proposed VS-ADMM algorithms applied to Datasets 1 and 2.

Biological Significance Analysis

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

IV. Numerical Result II

Subclass Identification Performance

Table 1: Clustering Performance of Different Methods for Dataset 1

	Purity	Rand Index	Sihouette
K-means	0.708	0.427	0.060
NMF (Euclidean) [2]	0.731	0.483	0.538
NMF (KL) [2]	0.742	0.489	0.616
DTPP [3]	0.741	0.491	0.680
Proposed VS-ADMM	0.749	0.506	0.803

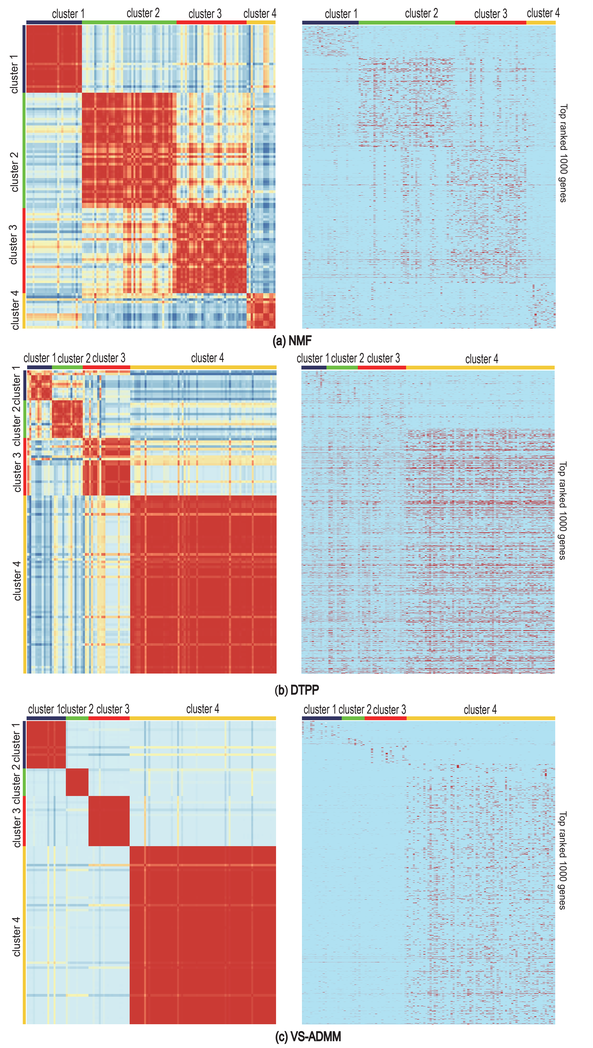

Figure 3: The heatmaps of clustering results (Left) and expression level of key genes (Right) obtained by the NMF, ONMF using DTPP and ONMF using proposed VS-ADMM applied to Dataset 2.

V. Reference

[1] H. Kim et al., "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.
[2] D. D. Lee et al., "Algorithms for non-negative matrix factorization," in Proc. NIPS, Denver, CO, USA, Dec. 2000, pp. 556-562.
[3] C. Ding et al., "Orthogonal nonnegative matrix t-factorizations for clustering," in Proc. KDD, Philadelphia, PA, USA, Aug. 2006, pp. 20-23.