Dropout-Aware Weighted Non-negative Matrix Factorization on Single-Cell RNA Sequencing Data

Jiaqi Zhang

@CSCI 2952Q

Nov. 2022

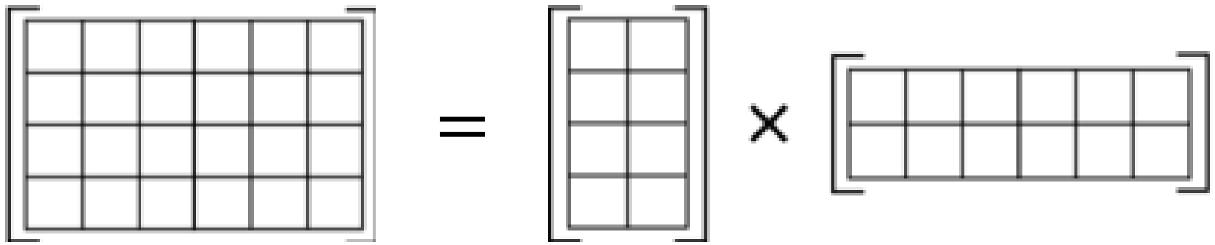
BROWN

CENTER FOR
Computational
Molecular Biology

Non-negative Matrix Factorization Helps Interpreting Complicated Systems

- **Non-negative matrix factorization (NMF)**: decompose a matrix into two factor matrices w/ non-neg. values

$\mathbf{X} \in \mathbb{R}_{+}^{m \times n}$ $\mathbf{W} \in \mathbb{R}_{+}^{m \times k}$ $\mathbf{H} \in \mathbb{R}_{+}^{k \times n}$

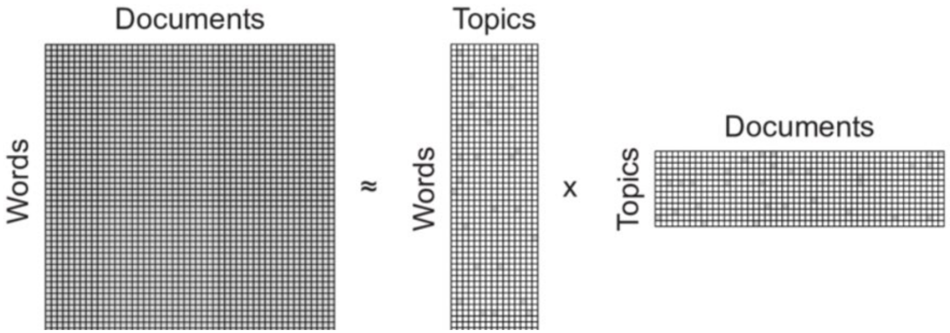


$$\mathbf{X} = \mathbf{WH} \quad \text{with} \quad W_{ij},\ H_{ij} \geqslant 0$$

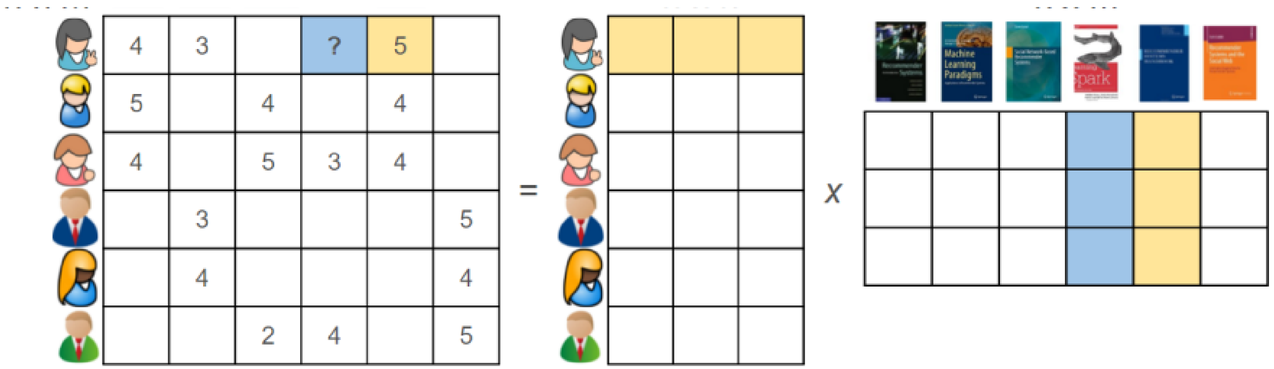
$$\operatorname{argmin}_{W_{ij}, H_{ij} \geqslant 0} \ \| \mathbf{X} - \mathbf{WH} \|_{\mathrm{F}}^{2}$$

- NMF is commonly used in various applications, including

topic modelling



recommender system



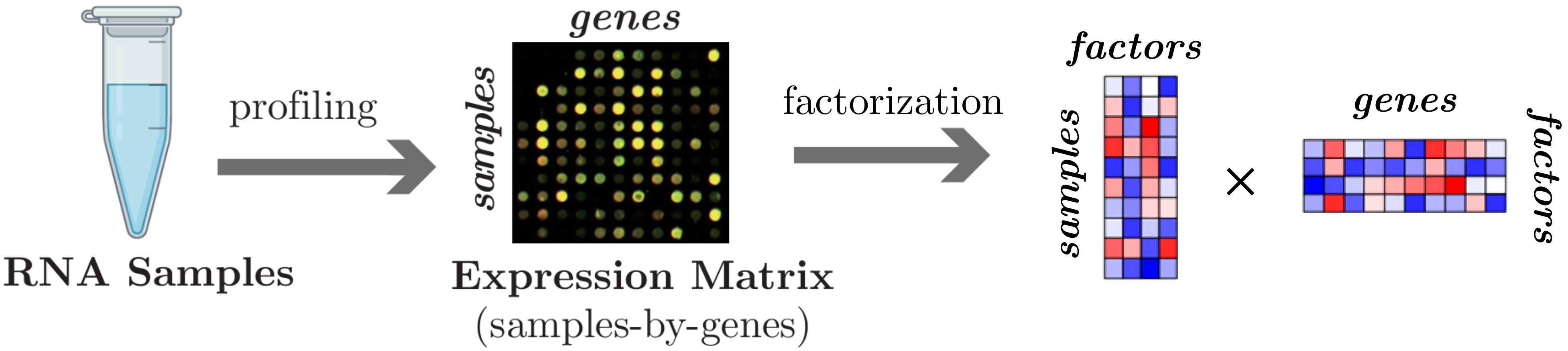
4	3		?	5	
5		4		4	
4		5	3	4	
	3				5
	4				4
		2	4		5

- NMF provides interpretation for relations between sample/features and latent factors

(Figures adopted from Wikipedia, Analytics Vidhya, and Medium)

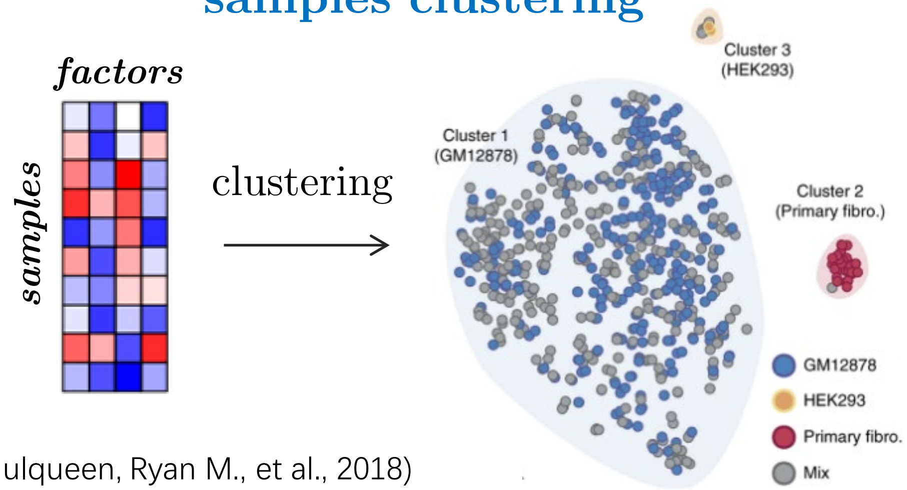
NMF Uncovers Knowledge from Bulk RNA-seq Data

- NMF is also popular in genomic data analyses
- NMF on bulk RNA sequencing (RNA-seq) data



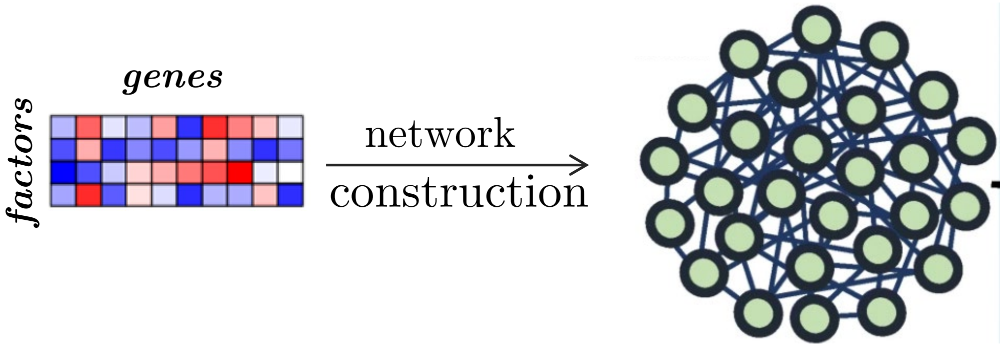
- NMF naturally fits RNA-seq data and has shown superior performances on:

samples clustering



(Mulqueen, Ryan M., et al., 2018)

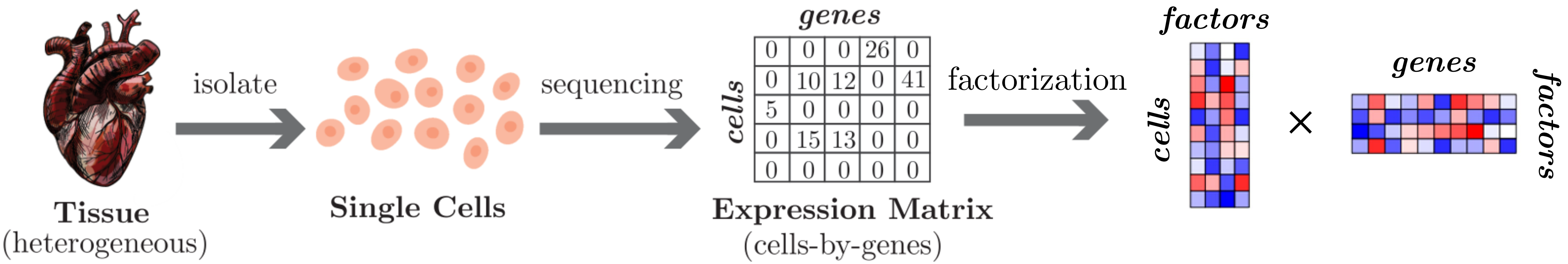
gene interaction network



(Xi, J., Wang, M., & Li, A., 2018)

Rapid Developments of Single-Cell RNA Sequencing Data Enables Finer Level of Analysis

- Development of single cell RNA sequencing (scRNA-seq) technique brings about single-cell level gene profiles



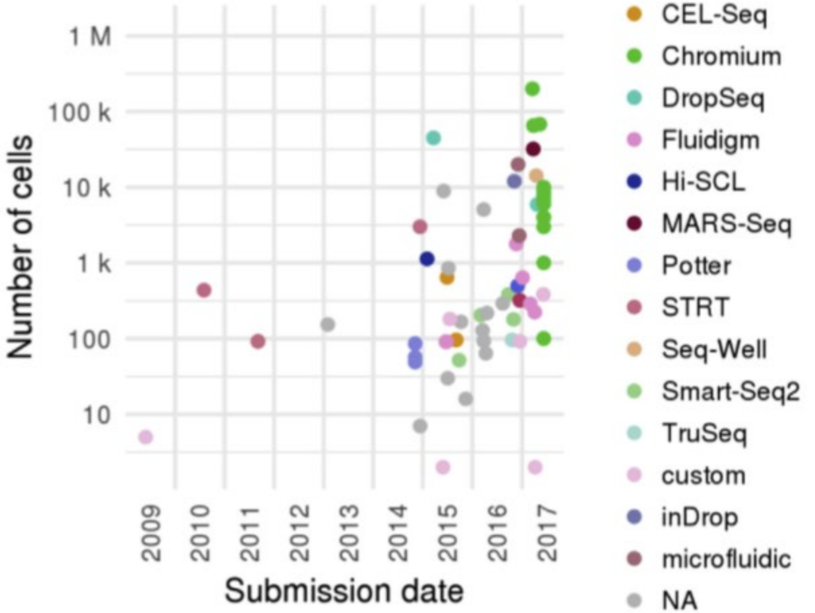
Tissue
(heterogeneous)

Single Cells

0	0	0	26	0
0	10	12	0	41
5	0	0	0	0
0	15	13	0	0
0	0	0	0	0

Expression Matrix
(cells-by-genes)

- scRNA-seq can provide data with 10K ~ 100K cells



factorization is unsupervised and data-driven

more cells offer more information of gene expression

should be helpful to understanding biological systems

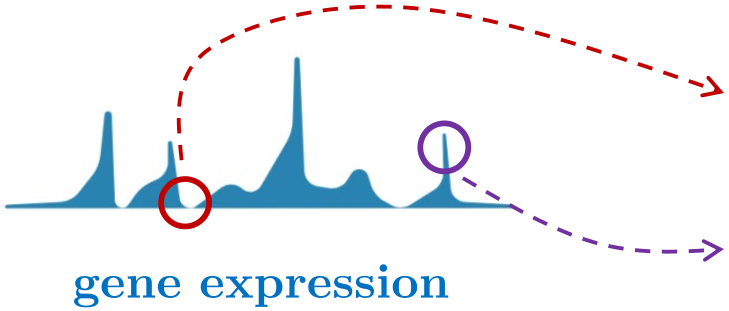
(Angerer, Philipp, et al. "Single cells make big data: New challenges and opportunities in transcriptomics." Current Opinion in Systems Biology 4 (2017): 85-91.)

Rapid Developments of Single-Cell RNA Sequencing Data Enables Finer Level of Analysis, But Also Bring Challenges

- scRNA-seq data properties pose challenges to the analysis
- High sparsity in scRNA-seq data

Dataset	Protocol	Sparsity (% of non-zeros)
Mouse Cortex	Smart Seq2	20.48%
Mouse Cortex	10x Genomics	7.58%
Human PBMC	Drop Seq	2.19%
Human PBMC	inDrops	1.92%

- Multiple sources of dropouts (i.e., zero values)



biological signals: gene is low-expressed at some cells

technical artifacts: missing data due to technical miss

Rapid Developments of Single-Cell RNA Sequencing Data Enables Finer Level of Analysis, But Also Bring Challenges (cont.)

- Previous studies proposed several NMF variations, but they didn't treat dropouts properly
  - L1-regularized NMF: $\| \mathbf{W} \|_1 + \| \mathbf{H} \|_1$ obtain unique factorizations with the sparsity constraint
  - Graph-regularized NMF: $\mathrm{tr}(\mathbf{H}^\top \mathbf{L} \mathbf{H})$ consider geometric structure of genes

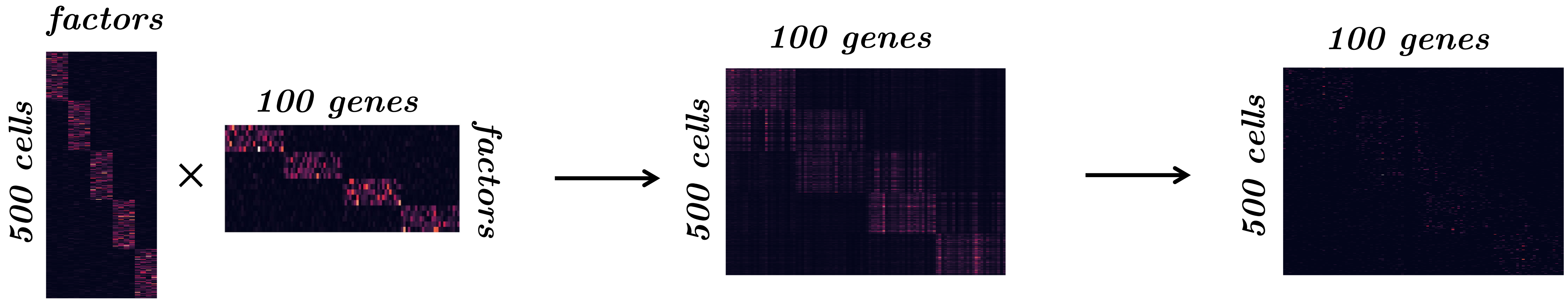
    L is the Laplacian matrix of the gene neighbor graph
  - Bayesian NMF: $\mathbf{X} \sim \mathcal{D}_x(\boldsymbol{\Theta})$ with $\boldsymbol{\Theta} = f(\mathbf{W}, \mathbf{H})$ assume more realistic distributions

    e.g., negative binomial (NB) distribution

- These NMF models break down for sparse scRNA-seq data, because they ignore missing values

Rapid Developments of Single-Cell RNA Sequencing Data Enables Finer Level of Analysis, But Also Bring Challenges (cont.)

- ***Validation***: test NMF methods on simulated data
  - Gamma-Poisson distribution ⟷ negative binomial distribution



$W_{ij},\ H_{ij} \sim \text{Gamma}$

$X_{ij} \sim \text{Poisson}(\theta_{ij}), \quad \theta_{ij} = (\mathbf{WH})_{ij}$

$X_{ij} = \alpha_{ij} X_{ij},\ \alpha_{ij} \sim \text{Binomial}(s)$

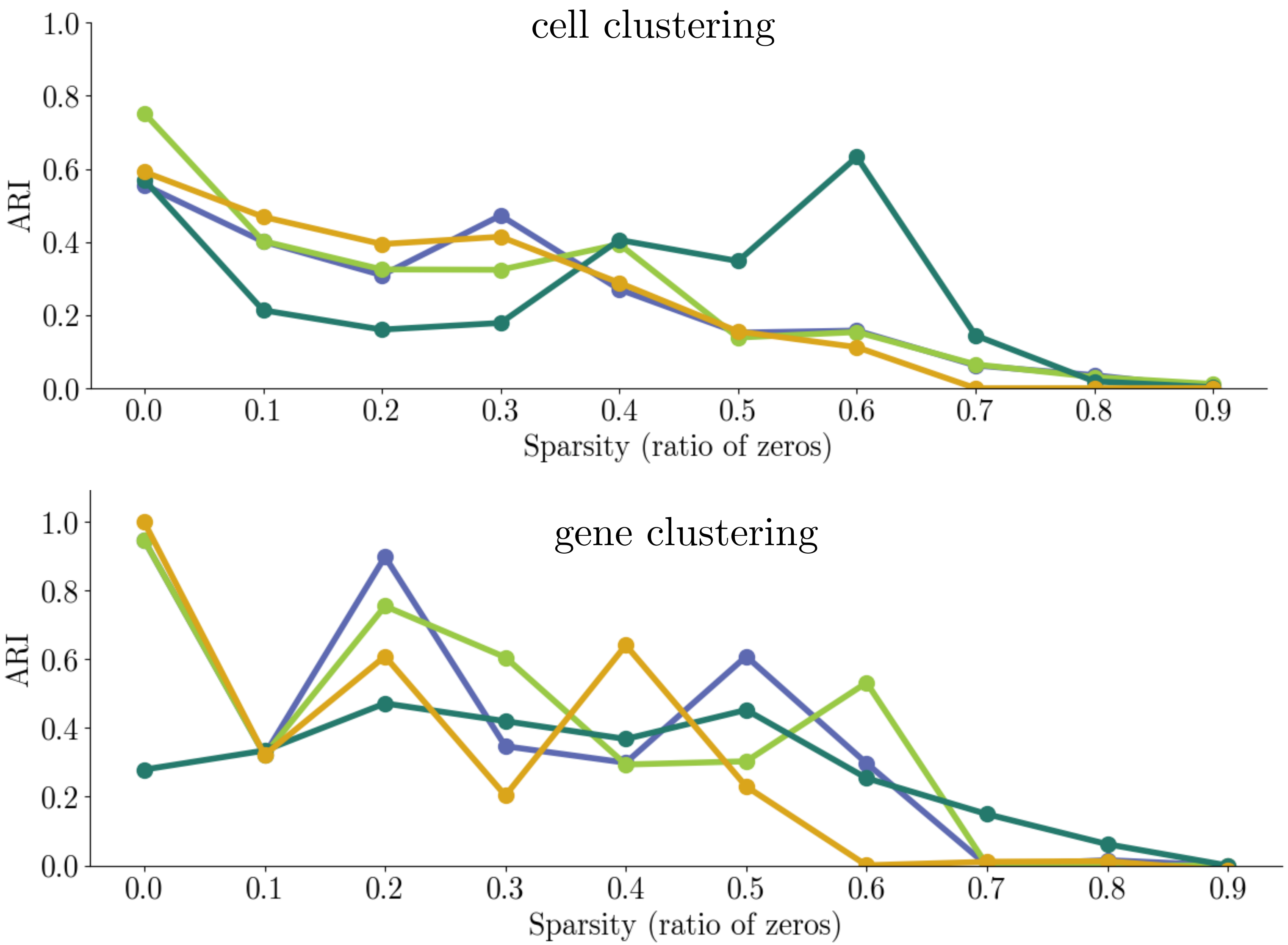
(5 cell clusters and 4 gene clusters)

  - Metric: adjusted rand index (ARI) ↑ on KMeans cell/gene clustering results

    ARI ∈ [0, 1], with 1 denoting perfect clustering and 0 indicating random labeling

Rapid Developments of Single-Cell RNA Sequencing Data Enables Finer Level of Analysis, But Also Bring Challenges (cont.)

●	NMF	(conventional NMF)
●	L1 NMF	(L1-regularized)
●	pCMF	(NB Bayesian NMF)
●	Graph NMF	(graph-regularized)



- Fail when data is too sparse, not applicable for scRNA-seq

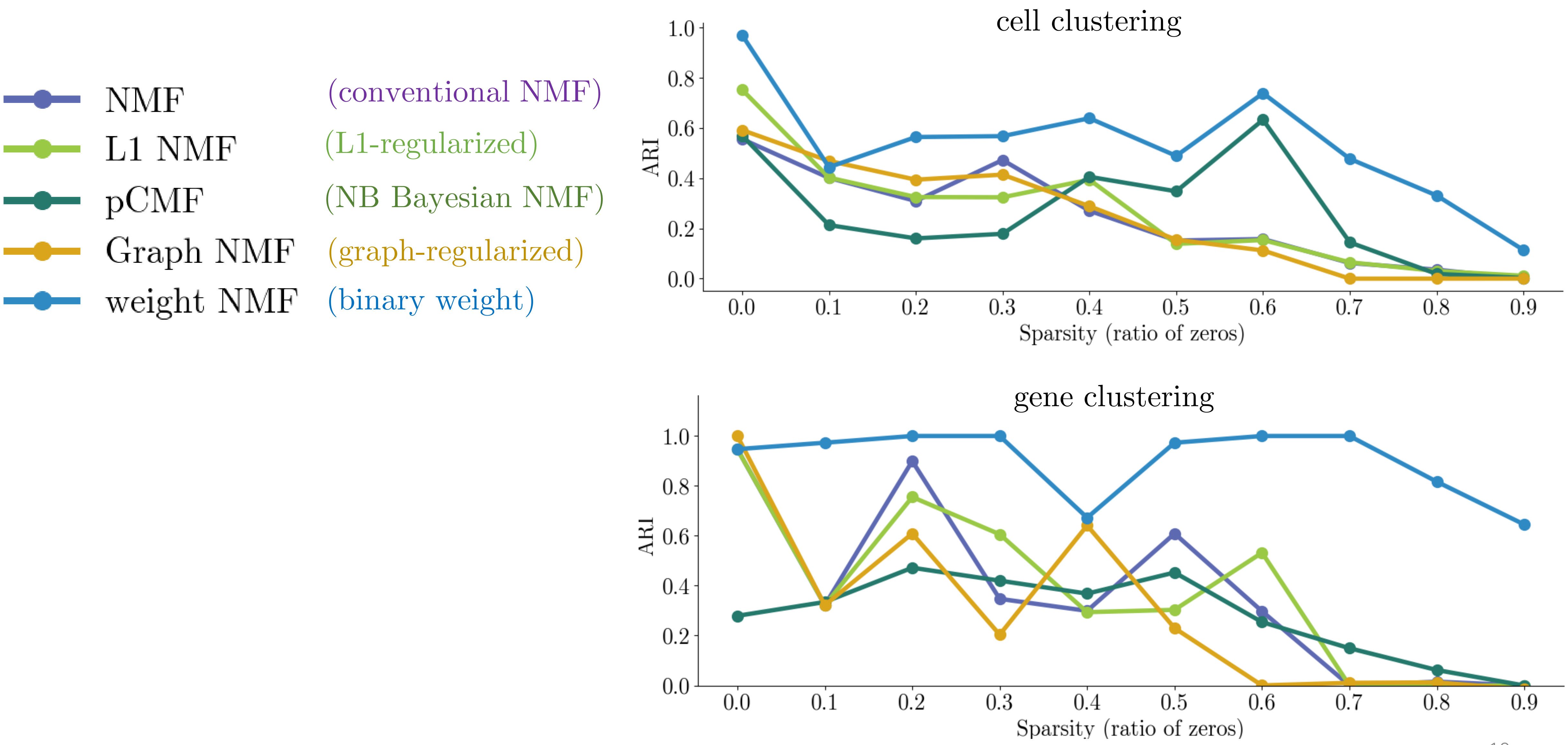
Weighted NMF is a Potential Solution Dealing with Sparsity

- Weighted NMF is proposed to deal with missing values
  - ○ Idea similar to weighted matrix completion $\operatorname{argmin}_{\mathbf{M}} \ \| \mathbf{W} \circ (\mathbf{X} - \mathbf{M}) \|_{\mathrm{F}}^2$ ($\circ$: element-wise multiply)
  - ○ weighted NMF $\operatorname{argmin}_{\mathbf{W},\mathbf{H}} \ \| \mathbf{W} \circ (\mathbf{X} - \mathbf{WH}) \|_{\mathrm{F}}^2$
  - ○ weight construction: **binary weight** $W_{ij} = \begin{cases} 1, & X_{ij} \neq 0 \\ 0, & X_{ij} = 0 \end{cases}$

    let the method only focus on observations (i.e., $W_{ij} = 1$)

Weighted NMF is a Potential Solution Dealing with Sparsity (cont.)

- Weighted NMF outperforms other methods on simulated data



Proposed: Dropout-Aware Weighted NMF for scRNA-seq Data

- Dropouts in scRNA-seq contains technical miss and biological zeros

- Estimate a weight matrix that gives zero to technical miss and non-zero to biological zero

- For each gene *j*, assume its normalized expression follows a Gamma-Normal mixture distribution with density

$$f_j(x) = \lambda_j \text{Gamma}(x;\ \alpha_j, \beta_j)\ +\ (1 - \lambda_j)\text{Normal}(x;\ \mu_j, \sigma_j)$$

technical fail

actual expression

- Estimate parameters with expectation-maximization (EM) algorithm with log-likelihood

$$\textstyle\sum_{j=1}^{n}\ f_j(x;\ \lambda_j, \alpha_j, \beta_j, \mu_j, \sigma_j)$$

- Estimate drop-out rate and construct dropout-aware weight matrix

$$d_{ij} = 1\ -\ \frac{\lambda_j \text{Gamma}(X_{ij};\ \alpha_j, \beta_j)}{\lambda_j \text{Gamma}(X_{ij};\ \alpha_j, \beta_j) + (1-\lambda_j)\text{Normal}(X_{ij};\ \mu_j, \sigma_j)} \quad \xrightarrow{\text{construct weight}} \quad W_{ij} = \begin{cases} d_{ij},\ \text{if } d_{ij} \geqslant 0.5 \\ 0,\ \text{if } d_{ij} < 0.5 \end{cases}$$

high value indicates it is true expression with high confidence

low value indicates it is more likely to be a technical miss

Dropout-Aware Weighted NMF Outperforms Baselines on Cell Clustering

- Experimental data: for each dataset, take top 500 highly variable genes

Dataset	**Protocol**	**# of Cells**	**Sparsity (% of non-zeros)**	**# of Cell Clusters**
Mouse Cortex	Smart Seq2	643	21.55%	7
Human PBMC	Drop Seq	6438	3.50%	9
Quake Lung	Smart Seq2	1676	7.32%	11

- Use weighted NMF + dropout-aware weight improve model performance

Dataset	**Cell Clustering ARI ↑ (best, second best)**					
	NMF	**L1 NMF**	**Graph NMF**	**pCMF**	**Weight NMF (binary)**	**Weight NMF (dropout-aware)**
Mouse Cortex	0.40	0.38	0.27	0.16	**0.43**	**0.51**

Dropout-Aware Weighted NMF Outperforms Baselines on Cell Clustering, But only for Less Sparse Data

- But the weighted NMF still fails on very-sparse data

Dataset	Protocol	# of Cells	Sparsity (% of non-zeros)	# of Cell Clusters
Mouse Cortex	Smart Seq2	643	21.55%	7
Human PBMC	Drop Seq	6438	3.50%	9
Quake Lung	Smart Seq2	1676	7.32%	11

Dataset	**Cell Clustering ARI ↑ (best, second best)**					
	NMF	**L1 NMF**	**Graph NMF**	**pCMF**	**Weight NMF (binary)**	**Weight NMF (dropout-aware)**
Mouse Cortex	0.40	0.38	0.27	0.16	**0.43**	**0.51**
Human PBMC	0.06	0.01	0.06	0.08	**0.11**	**0.10**
Quake Lung	0.10	0.07	0.04	0.02	**0.18**	**0.19**

Discussion & Summary

- The proposed dropout-aware weight extends NMF to single-cell analysis by considering special data properties

- Dropout-aware weighted NMF is still applicable in some cases
  - Some sequencing protocols provide less sparse data
  - Quality control in pre-processing can remove lowly-expressed cells/genes and reduce sparsity

- Better dropout rate estimation is required:
  - Gamma-Normal mixture: $f_j(x) = \lambda_j \text{Gamma}(x;\ \alpha_j, \beta_j)\ +\ (1-\lambda_j)\text{Normal}(x;\ \mu_j, \sigma_j)$
  - Zero-inflated negative binomial: $\Pr(x=0) = \lambda + (1-\lambda)\text{NB}(x=0; r, p)$

Thanks!