Supplementary material for Orthogonal Parametric Non-negative Matrix Tri-Factorization with α-Divergence for Co-clustering

Saeid Hoseinipour, Mina Aminghafari*, Adel Mohammadpour

Department of Mathematics and Computer Science, Amirkabir University of Technology (Tehran Polytechnic), Iran.

1. Introduction

In this supplementary material, we present a couple of complementary elements which were omitted in the main paper due to limitation the length of paper and redundancy of them. First, we present synthetic datasets for Bernoulli, Poisson, and Truncated Gaussian, and then implement OPNMTF on these. Second, we show more visualization of some datasets.

2. Synthetic data

We propose to use a Latent Block Model (LBM) (Govaert and Nadif, 2013) for generating a non-negative data matrix. This generative model considers $\boldsymbol{\mathcal{F}}$ and $\boldsymbol{\mathcal{G}}$ as the sets of possible labels $\mathbf{F}$ (rows) and $\mathbf{G}$ (columns), also $\varphi(X_{ij};\delta_{kh})$ denoted distribution of the values x_{ij} for co-cluster (blocks) kh. The total probability density function is as follows

$$f(\mathbf{X};\boldsymbol{\theta}) = \sum_{(\mathbf{F},\mathbf{G})\in\boldsymbol{\mathcal{F}}\times\boldsymbol{\mathcal{G}}} \prod_{i,k} \pi_k^{F_{ik}} \prod_{j,h} \rho_h^{G_{jh}} \prod_{i,j,k,h} \varphi(X_{ij};\delta_{kh}), \tag{1}$$

*Corresponding author
Email address: aminghafari@aut.ac.ir (Mina Aminghafari)

where $\boldsymbol{\theta} = (\boldsymbol{\pi}, \boldsymbol{\rho}, \boldsymbol{\delta})$, $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_g)$ and $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_s)$, also $\boldsymbol{\delta} = (\delta_{11}, \ldots, \delta_{gs})$ are all parameters co-clusters that are defined according to the probability density function.

- **Bernoulli LBM**: The values x_{ij} are distributed according to a Bernoulli distribution $\mathcal{B}(\gamma_{kh})$ with $\gamma_{kh} \in [0, 1]$, therefore

$$\varphi(x_{ij}; \gamma_{kh}) = \gamma_{kh}^{x_{ij}} (1 - \gamma_{kh})^{1-x_{ij}}. \tag{2}$$

- **Poisson LBM**: The values x_{ij} are distributed according to a Poisson distribution $\mathcal{P}(\omega_i \nu_j \alpha_{kh})$ where ω_i and ν_j are the effects of row i and column j, respectively. Also, α_{kh} is the effect of co-cluster kh for x_{ij} that belong to the k-th row cluster and h-th column cluster and

$$\varphi(x_{ij}; \alpha_{kh}) = \frac{e^{-\omega_i \nu_j \alpha_{kh}} (\omega_i \nu_j \alpha_{kh})^{x_{ij}}}{x_{ij}!}. \tag{3}$$

- **Truncated Gaussian LBM**: The values x_{ij} belong to co-cluster kh are distributed according to a Truncated Gaussian distribution $\mathcal{TG}(\mu_{kh}, \sigma_{kh})$ with $a \leq x_{ij} \leq b$, mean μ_{kh}, and variance σ_{kh}, therefore

$$\varphi(x_{ij}; \mu_{kh}, \sigma_{kh}, a, b) = \frac{1}{\sigma_{kh}} \frac{\phi\left(\frac{x_{ij} - \mu_{kh}}{\sigma_{kh}}\right)}{\Phi\left(\frac{b - \mu_{kh}}{\sigma_{kh}}\right) - \Phi\left(\frac{a - \mu_{kh}}{\sigma_{kh}}\right)}, \tag{4}$$

  where ϕ and Φ are standard Gaussian distribution and cumulative distribution functions, respectively.

Our motivation is to implement the introduced algorithms on the data matrix with a known distribution of X_{ij}. Due to having row and column true labels, we use a two-way accuracy evaluation criterion. Algorithm 1

Algorithm 1 Simulating data of the Latent Block Model

Input: $n, m, g, s, \boldsymbol{\pi}, \boldsymbol{\rho}$

Simulate $\mathbf{F}$ according to a Multinomial distribution with parameters $\boldsymbol{\pi} = (1, \pi_1, \ldots, \pi_g)$.

Simulate $\mathbf{G}$ according to a Multinomial distribution with parameters $\boldsymbol{\rho} = (1, \rho_1, \ldots, \rho_s)$.

Simulate each co-cluster $\mathbf{X}_{kh}$ according to Bernoulli (2), Poisson (3), and Truncated Gaussian (4) distributions.

Permutation all $\mathbf{X}_{kh}$.

Output: Data matrix $\mathbf{X}$ size of $n \times m$

present process simulation data of LBM, also NMTFcoclust as repository GitHub is available.

2.0.1. Setting

Table 1 shows the parameters that we need to generate a non-negative data matrix. Figure 1 illustrates three types of simulation data matrices. Figures 2 and 3 shows row and column clustering comparison with difference tuple parameter $(\lambda, \mu) \in \{(10^{-9}, 10^{-9}), (10^{-9}, 10^{+9}), (10^{+9}, 10^{-9}), (10^{+9}, 10^{+9})\}$ and $\alpha \in \{0, 0.5, 1\}$. We conducted experiments on synthetic and real-world datasets to validate the effectiveness of our proposed algorithm.

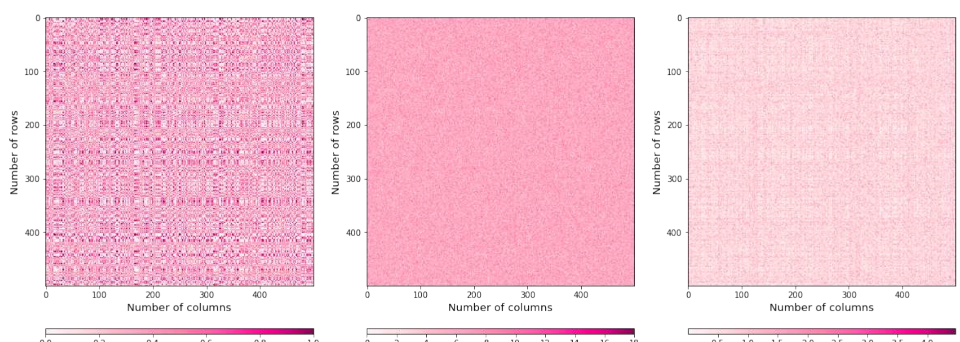
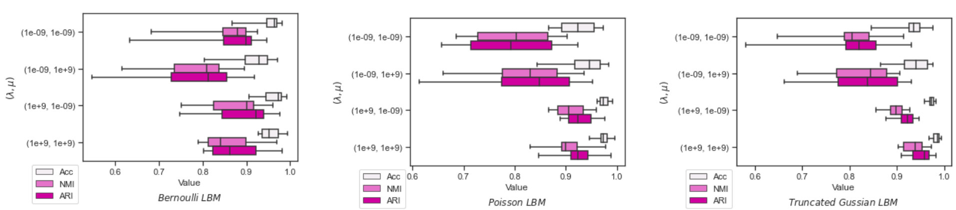


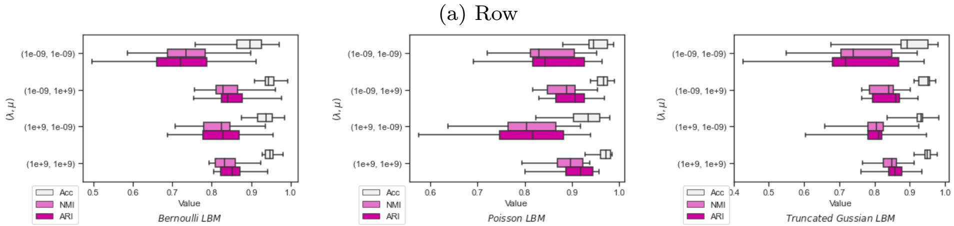
Figure 1: (Left) Data matrix **Bernoulli LBM**. (Middle) Data matrix **Poisson LBM**. (Right) Data matrix **Truncated Gaussian LBM**.

Table 1: Parameters required to generate synthetic data from LBM.

Inputs and Parameters	Setting
(n, m)	(500, 500)
(g, s)	(3,3)
LBM	$\boldsymbol{\pi} = (1/3, 1/3, 1/3)$ $\boldsymbol{\rho} = (1/3, 1/3, 1/3)$
Bernoulli LBM	$\boldsymbol{\gamma} = \begin{pmatrix} 0.9 & 0.01 & 0.01 \\ 0.01 & 0.9 & 0.01 \\ 0.01 & 0.01 & 0.9 \end{pmatrix}$
Poisson LBM	$\boldsymbol{\alpha} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$ $\boldsymbol{\omega} = (2, 2, 6)$ $\boldsymbol{\nu} = (7, 2, 1)$
Truncated Gaussian LBM	$\boldsymbol{\mu} = \begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix}, \boldsymbol{\sigma}^2 = \begin{pmatrix} 10 & 1 & 1 \\ 1 & 100 & 1 \\ 1 & 1 & 1000 \end{pmatrix}$

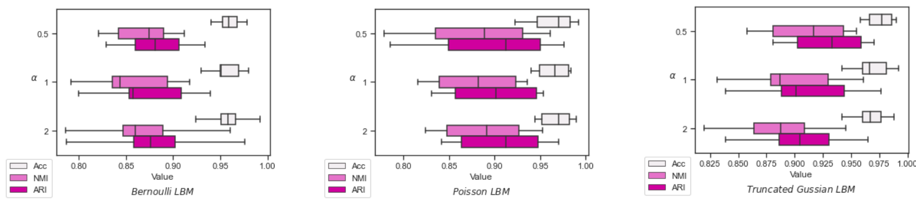


(a) Row

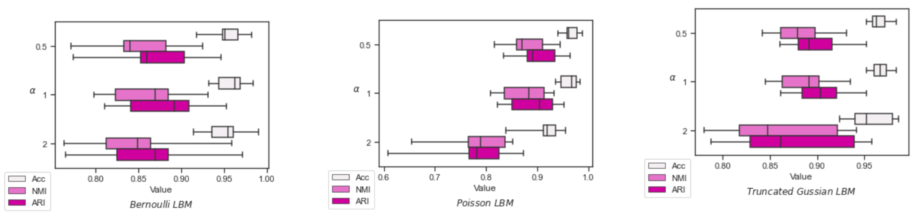


(b) Column

Figure 2: Results on three synthetic data (**Bernoulli**, **Poisson**, **Truncated Gaussian**) according to OPNMTF with $\alpha = 0.5$ and $(\lambda, \mu) \in \{(10^{-9}, 10^{-9}), (10^{-9}, 10^{+9}), (10^{+9}, 10^{-9}), (10^{+9}, 10^{+9})\}$.



(a) Row



(b) Column

Figure 3: Box plots of measures (Acc, NMI, ARI) for three synthetic data (**Bernoulli**, **Poisson**, **Truncated Gaussian**) based on 100 iterations of OPNMTF with $\lambda = 0.03$, $\mu = 1$, and $\alpha \in \{0, 0.5, 1\}$.

3. Visualization

We show a triple reorganization plot based on OPNMTF on **WebACE** dataset in Figure 4, also Figures 5 and 6 show different numbers of column clusters for **CSTR** and **Classic3**, respectively. We focus on **Classic3** with PCA in 7, as well as the co-clustering reorganization plot near by word clouds 3×3 in Figure 8.

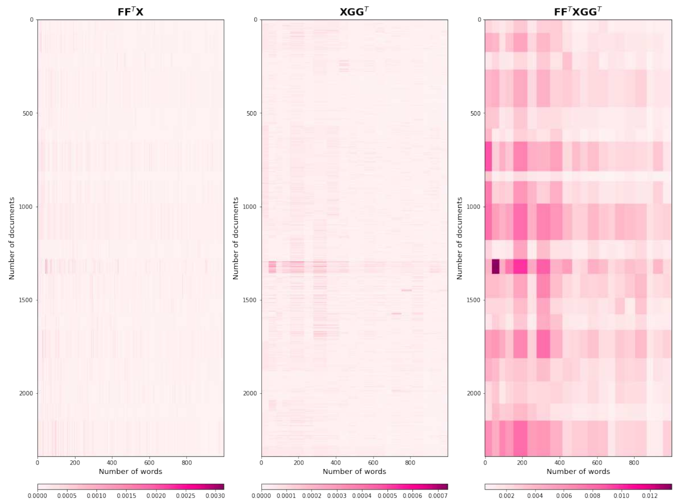


Figure 4: Result by OPNMTF with $\alpha = 1$, $\lambda = 0.03$, $\mu = 1$, $g = 20$, and $s = 10$ on **WebACE**, (Left) Clustering row data matrix ($\mathbf{FF}^\top\mathbf{X}$). (Middle) Clustering column data matrix ($\mathbf{XGG}^\top$). (Right) Co-clustering reorganization data matrix ($\mathbf{FF}^\top\mathbf{XGG}^\top$).

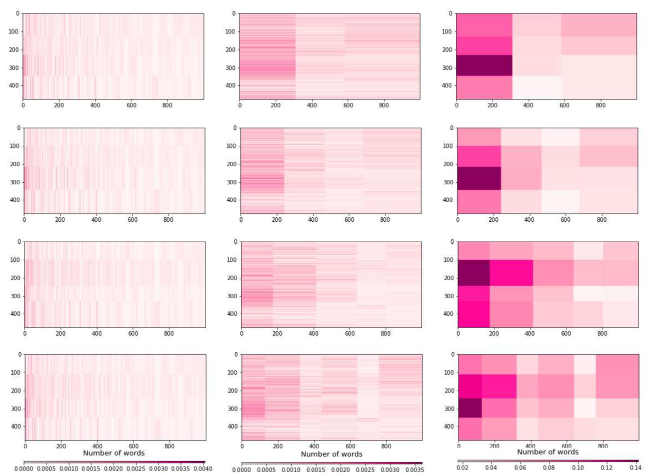


Figure 5: Result by OPNMTF with $\alpha = 0.4$, $\lambda = 0.4$, $\mu = 0.6$, and $g = 4$ on **CSTR**, (Left) Clustering rows data matrix ($\mathbf{FF}^\top\mathbf{X}$). (Middle) Clustering columns data matrix ($\mathbf{XGG}^\top$) with $s = 3, 4, 5, 6$. (Right) Co-clustering reorganization data matrix, ($\mathbf{FF}^\top\mathbf{XGG}^\top$).

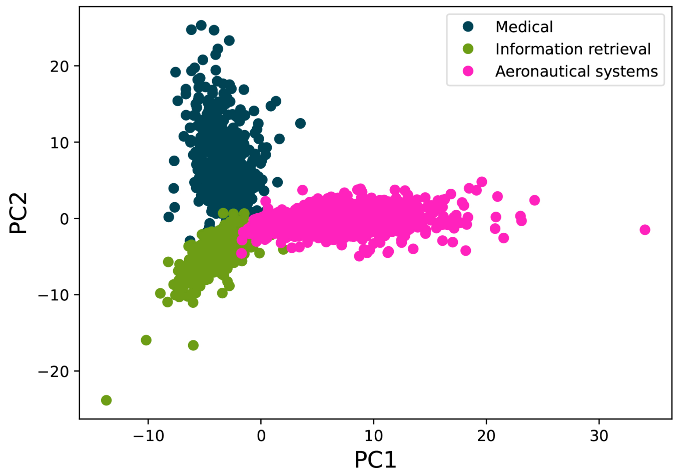


Figure 7: Principal Components Analysis (PCA) on **Classic3** for two components.

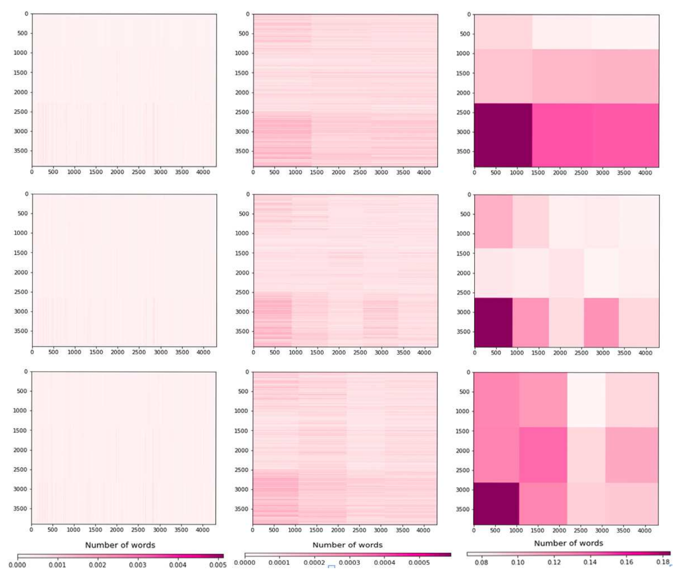


Figure 6: Result by OPNMTF with $\alpha = 0.4$, $\lambda = 0.03$, $\mu = 1$, and $g = 3$ on **Classic3**, (Left) Clustering row data matrix ($\mathbf{FF}^T\mathbf{X}$). (Middle) Clustering column data matrix ($\mathbf{XGG}^T$) with $s = 3, 4, 5$. (Right) Co-clustering reorganization data matrix ($\mathbf{FF}^T\mathbf{XGG}^T$).

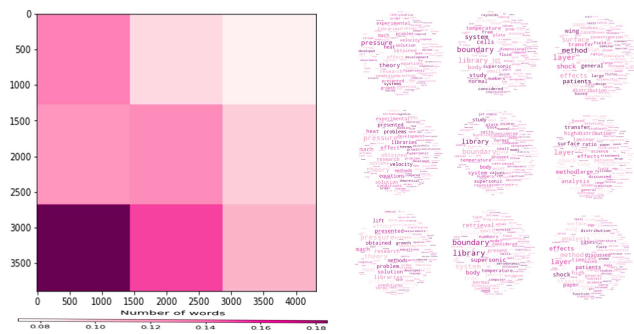


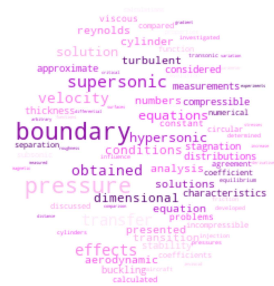
Figure 8: (left) Co-clustering reorganization **Classic3** data matrix according to the output of the OPNMTF$_{\alpha=0.5}$ (right) Top-20 words in each document-word co-cluster discovered by OPNMTF$_{\alpha=0.5}$ with $\lambda = 0.01$ and $\mu = 0.01$ for **Classic3** dataset.

In the original paper, we presented an analysis of **Classic3** with $s = 3$. Now, we want to consider $s = 2$ and implement OPNMTF for more analysis about polysemous words. Based on Table 2 as summarization matrix $\mathbf{S}$, the top 100 word clouds display in Figure 9, and the top 20 words are represented in Tables 3 and 4.

Table 2: The summarization matrix $\mathbf{S}$ in the middle of the three factor decomposition $\mathbf{FSG}^T$ determined by OPNMTF$_{\alpha=1}$ with $\lambda = 0.03$, and $\mu = 1$. Two word-clusters ($s = 2$), and three document-clusters ($g = 3$) are denoted by WC1-WC2, and DC1-DC3. Each of DC1-DC3 is associated with one of document-clusters with maximum value (boldface number).

	WC1	WC2
DC1: Medical	**0.153989**	0.126851
DC2: Information retrieval	**0.156837**	0.142681
DC3: Aeronautical systems	**0.238269**	0.181370

(a) WC1 with total frequency **0.549095** for all words.



(b) DC3-WC2 with frequency **0.181370** for all words.

Figure 9: Word clouds of the top 100 words for matrix **S** have been shown in Table 2 for **Classic3** dataset. The bigger word has more frequency.

Table 3: Top 20 words in WC1 are summarized when OPNMTF is applied. (total frequency is 0.549095)

WC1: Medical+Information retrieval+Aeronautical systems
library, **layer**, **theory**, **method**, **system**, **mach**, **shock**, **heat**
study, **body**, **effect**, **surface**, **wing**, **temperature**, **high**, **time**
large, **cases**, **problem**, **case**

Table 4: Top 20 words in WC2 are summarized when OPNMTF is applied.

WC2: Aeronautical systems
boundary, **pressure**, **supersonic**, **transfer**, **effects**, **obtained**, **velocity**, **solution**
hypersonic, **equations**, **conditions**, **dimensional**, **analysis**, **numbers**, **presented**, **reynolds**
turbulent, **solutions**, **aerodynamic**, **cylinder**

References

Govaert, G., Nadif, M., (2013). Co-clustering: models, algorithms and applications. John Wiley & Sons. https://books.google.com/books?hl=en&lr=&id=ESdMEAAAQBAJ&oi=fnd&pg=PA1&ots=UZa52Wdj-9&sig=mPswdHuxk-wociB2gh-FiTMHpUY.