Hindawi
Journal of Applied Mathematics
Volume 2021, Article ID 6675829, 15 pages
https://doi.org/10.1155/2021/6675829

*Research Article*

# Coordinate Descent-Based Sparse Nonnegative Matrix Factorization for Robust Cancer-Class Discovery and Microarray Data Analysis

**Melisew Tefera Belachew**

*Department of Mathematics, Haramaya University, Ethiopia*

Correspondence should be addressed to Melisew Tefera Belachew; melisewt@gmail.com

Received 29 December 2020; Revised 3 August 2021; Accepted 31 August 2021; Published 23 October 2021

Academic Editor: Kannan Krithivasan



Determining the number of clusters in high-dimensional real-life datasets and interpreting the final outcome are among the challenging problems in data science. Discovering the number of classes in cancer and microarray data plays a vital role in the treatment and diagnosis of cancers and other related diseases. Nonnegative matrix factorization (NMF) plays a paramount role as an efficient data exploratory tool for extracting basis features inherent in massive data. Some algorithms which are based on incorporating sparsity constraints in the nonconvex NMF optimization problem are applied in the past for analyzing microarray datasets. However, to the best of our knowledge, none of these algorithms use block coordinate descent method which is known for providing closed form solutions. In this paper, we apply an algorithm developed based on columnwise partitioning and rank-one matrix approximation. We test this algorithm on two well-known cancer datasets: leukemia and multiple myeloma. The numerical results indicate that the proposed algorithm performs significantly better than related state-of-the-art methods. In particular, it is shown that this method is capable of robust clustering and discovering larger cancer classes in which the cluster splits are stable.

## 1. Introduction

Analyzing and interpreting microarray data which represent biological processes of cancers and other related diseases are among the big challenges in data science [1–3]. Moreover, treating cancer properly depends hugely on identifying specific therapies to different tumor types which is always a challenging task. For this reason, classifying cancer types properly and accurately plays a vital role in diagnosis and treatment of cancer and other related diseases [2, 3]. Numerous methods have been developed in the past to facilitate extracting and interpreting fundamental patterns of gene expressions hidden in microarray data. These methods have also been proved useful for classifying and clustering genes and samples that show similar patterns. One of these methods is hierarchical clustering (HC) which is widely used for analyzing and building hierarchy of clusters in high-dimensional data. For instance, Eisen et al. [4] applied HC to yeast and human microarray data to find out that it efficiently clusters genes of known similar patterns in the same category. HC has also been successfully used by Alizadeh et al. [5] to identify distinct types of diffuse large B-cell lymphoma among patients. The authors reported that molecular classification based on gene profiles together with HC helps to identify subtypes of cancer which play a vital role for undergoing clinical diagnosis. Perou et al. [6] experimentally proved that HC is very useful for classifying molecular portraits of breast tumors into subtypes distinguished by the differences that exist in the corresponding gene expression patterns. In spite of the abovementioned merits, HC has some disadvantages including failing to reveal clusters correctly, being very sensitive to the similarity measure used in the experiment, and being unable to depict local behavior [1, 3, 7, 8].

Other valuable and powerful approaches are types of artificial neural networks called self-organizing maps (SOMs). SOMs are dimensionality reduction techniques trained using unsupervised learning. SOMs can be used for

recognizing and classifying features in high-dimensional data. Tamayo et al. [9] applied SOMs for hematopoietic differentiation and reported that they are well suited for highlighting certain genes and pathways involved in differentiation therapy used in the treatment of acute promyelocytic leukemia. Golub et al. [2] reported that SOMs do not need any background information to effectively discover the difference between acute myeloid leukemia and acute lymphoblastic leukemia. On the other hand, SOMs fail to provide sparse parts-based localized features and since they are very sensitive to initial conditions, they may give different decompositions of the data to different initializations [1, 3, 8].

There are also various dimensionality reduction and matrix decomposition techniques developed for analyzing microarray data. For instance, Moloshok et al. [10] showed that Bayesian decomposition is capable of providing insights and identifying temporal patterns when applied to gene expression data obtained from yeast cell cycle experiments. Gasch and Eisen [11] used fuzzy $k$-means clustering to identify overlapping clusters of yeast genes and groups of functionally related and coregulated yeast genes. Alter et al. [12] have shown that singular value decomposition is suitable for processing and modeling genomewide expression patterns. The authors applied this decomposition method on elutriation yeast dataset and find out that it is capable of extracting gene expression patterns correlated with the original samples in the data. However, it reported that the above matrix decomposition methods have some drawbacks including being unable to capture full structures and local behaviors hidden in high-dimensional data [1, 7, 8]. Another efficient and well-known data exploratory tool is nonnegative matrix factorization (NMF). NMF is a linear dimensionality reduction technique famous for extracting basic and hidden features of high-dimensional data. The basic idea of NMF is to decompose a given nonnegative data matrix into two low-rank matrices: basis and weight matrix. Given a data matrix $Y \in \mathbb{R}_+^{m\times n}$ and a reduced rank $r$, the basic NMF problem based on the squared Frobenius norm can be stated as the following optimization problem:

$$\min_{U\in\mathbb{R}_+^{m\times r},\, V\in\mathbb{R}_+^{r\times n}} F(U,V) = \frac{1}{2}\|Y - UV\|_F^2. \tag{1}$$

NMF is found to be an essential tool for analyzing and interpreting gene expression patterns associated with microarray data. The choice and meaning of NMF matrices is application dependent: in analyzing microarray data, the columns of the data matrix $Y$ are samples of gene expression patterns and its rows represent genes. The columns and rows of the basis matrix $U$ stand for metagenes and genes, respectively, while the columns of the weight matrix $V$ are considered gene samples and its rows are referred as metagenes [1, 3, 8]. The interesting property of NMF is that it enables additive combination of parts to make a whole and is found very suitable for understanding the underlying structure in microarray data that represent various real-life phenomena and biological processes [1, 7, 8, 13–16]. Brunet et al. [1] successfully applied standard NMF algorithm based on Kullback-Leibler (KL) divergence by Lee and Seung [17] to extract meaningful information from different leukemia and tumor microarray data. The authors proved experimentally that NMF is a more powerful method for discovering cancer classes and molecular patterns when compared to HC and SOMs. Frigyesi and Hoglund [7] also used the divergence-based NMF algorithm to analyze some cancer and tumor data. Their experimental results showed that NMF facilitates the extraction of biologically relevant structure of microarray data and plays a vital role in understanding the properties of tumor and cancer-related diseases. Bocarelli et al. [13] witnessed that NMF is powerful in extracting biologically relevant genes about the symptoms and clinical conditions observed on patients suffering from multiple myeloma (MM) and monoclonal gammopathy of undetermined significance (MGUS). Carrasco et al. [14] modified KL-based NMF so that it enables to extract distinct genomic features and applied it to microarray data obtained from multiple myeloma patients. The results show that NMF is capable of providing useful information that helps to prepare specific drugs for certain patients. Kim and Tidor [15] applied NMF based on the squared Frobenius norm on yeast data and reported that NMF can predict functional relationships more accurately than conventional approaches. They have also showed that NMF is able to efficiently detect relationship among genes and functional subsystems at the molecular level.

It is possible to make standard NMF more appealing for the extraction of better and sparser localized features which are biologically more relevant. This can be done, for instance, by imposing sparsity constraints on the basic NMF problem. One way of obtaining sparse factors is by enforcing an $l_1$-norm-based sparsity constraint on the weight matrix $V$ of (1) as

$$\min_{U\in\mathbb{R}_+^{m\times r},\, V\in\mathbb{R}_+^{r\times n}} F(U,V) = \frac{1}{2}\|Y - UV\|_F^2 + \gamma\|V\|_1, \tag{2}$$

where $\gamma$ is a penalty parameter used to control the trade of between sparsity and reconstruction accuracy. Algorithms designed for treating different sparse NMF formulations including (2) are widely used for discovering cancer classes and analyzing gene expression microarray data. Gao and Church [3] formulated a sparse NMF (SNMF) problem by imposing a squared $l_2$-norm-based sparsity constraint on $V$. The authors designed an algorithm which has a multiplicative update rule component. The algorithm was applied on well-studied cancer and tumor datasets. They have concluded that their approach enables them to classify and discover cancer classes missed by standard NMF. Kim and Park [18] also considered Frobenius norm-based sparse NMF and used the concept of alternating nonnegative least squares to develop a new algorithm and applied it to analyze leukemia and tumor datasets. The numerical results reveal that their method has a better clustering performance and facilitates biological interpretation when compared to classical NMF. Kong et al. [19] used NMF with sparseness constraints

(NMFSC) by Hoyer [20] and reported that NMFSC performs more or less the same as SNMF when applied to leukemia and medulloblastoma datasets. Esposito et al. [21] considered a sparse NMF formulation based on KL divergence and designed a flexible multiplicative update rule to solve it. The authors successfully applied the method on 3D (a 3D microarray data is a 2D microarray data which evolves with time) microarray data obtained from multiple sclerosis patients. The numerical results show that their approach also helps to get a better understanding of the biological behavior of the disease.

Inspired by the aforementioned merits of sparse NMF formulation, we consider problem (2) and propose an algorithm that enables discovering larger cancer classes and facilitates the extraction of meaningful gene expression patterns. Instead of using multiplicative update rules, like most of the existing methods, we apply an algorithm which makes use of the block coordinate descent optimization method.

## 2. Methods

This section presents the proposed algorithm and the model selection strategy. The algorithm is obtained by extending the rank-one residue iteration (RRI) for NMF in [22]. A similar algorithm which is an extension of the hierarchical alternating least squares algorithm also exists in [23]. We were informed about this during the second review.

*2.1. Coordinate Descent Method for Sparse NMF.* In this section, we provide the derivation of the proposed algorithm by using the concept of the block coordinate descent (BCD) method which is popular for providing closed form solutions. This concept is interesting in the sense that it enables breaking down a given matrix-based nonconvex optimization problem into convex subproblems (based on blocks of columns) which are easier to solve [22, 24, 25].

Consider the following formulation of the sparse NMF problem:

$$\min_{U\in\mathbb{R}_+^{m\times r},\, V\in\mathbb{R}_+^{n\times r}} F(U,V)=\frac{1}{2}\|Y-UV^T\|_F^2+\gamma\|V\|_1. \quad (3)$$

Together with rank-one matrix factorization, BCD facilitates an optimal way of solving for the columns of $U$ and $V$ in an alternating fashion. We start by reformulating the sparse NMF problem (3) based on partitioning the columns of $U$ and $V$. The basic idea is to consider one column of $U$, $U_{:i}$, as an unknown and treat the rest columns of $U$ and the whole of $V$ as constants. We do this repeatedly until all the columns of $U$ and $V$ are approximated optimally. Using the above procedure leads to reformulating (3) in $U_{:i}$ as

$$\begin{aligned}\min_{U_{:i}\geq 0} F(U_{:i}) &= \frac{1}{2}\left\|\left(Y-\sum_{j\neq i}U_{:j}V_{:j}^T\right)-U_{:i}V_{:i}^T\right\|_F^2+\gamma\|V_{:i}\|_1\\ &= \frac{1}{2}\|R_j-U_{:i}V_{:i}^T\|_F^2+\gamma\|V_{:i}\|_1,\ i,j=1,2,\cdots,r,\end{aligned} \quad (4)$$

1: **Inputs**: datamatrix $Y\in\mathbb{R}_+^{m\times n}$ and rank $r$;
2: **Initialize and scale**$U\in\mathbb{R}_+^{m\times r}$ and $V\in\mathbb{R}_+^{n\times r}$;
3: Set sparsity parameter $\gamma$
4: **repeat**
5: **for**$i=1$: $r$**do**
6: $R_j=Y-\sum_{j\neq i}U_{:j}V_{:j}^T$;
7: Use (7) to update $U_{:i}$;
8: **if**$\|U_{:i}\|_2>0$ and $\|V_{:i}\|_2>0$**then**
9: $d=\sqrt{\|U_{:i}\|_2/\|V_{:i}\|_2}$, $U_{:i}=dU_{:i}$, and $V_{:i}=(1/d)V_{:i}$;
10: **end if**
11: Repeat steps 6 to 10 using (9) to update $V_{:i}$;
12: **end for**
13: **until** Stopping condition.

Algorithm 1: Pseudocode for CDSNMF

where $R_j=Y-\sum_{j\neq i}U_{:j}V_{:j}^T$. By the same token, we obtain the following subproblems corresponding to the columns of $V$:

$$\min_{V_{:i}\geq 0} F(V_{:i})=\frac{1}{2}\|R_j-U_{:i}V_{:i}^T\|_F^2+\gamma\|V_{:i}\|_1,\ i,j=1,2,\cdots,r. \quad (5)$$

Just like most optimization problems, the solutions of (4) and (5) are stationary points that satisfy the corresponding Karush-Kuhn-Tucker (KKT) optimality conditions. The KKT conditions for (4) are

$$U_{:i}\geq 0,\ \nabla_{U_{:i}}F(U_{:i})=-\left(R_j-U_{:i}V_{:i}^T\right)V_{:i}\geq 0,\ U_{:i}\#\nabla_{U_{:i}}F(U_{:i})=0, \quad (6)$$

where # denotes the elementwise product. The vectors that solve (4) are the ones that satisfy (6). They are given by

$$U_{:i}^*=\max\left(\frac{R_jV_{:i}}{\|V_{:i}\|_2^2},0\right). \quad (7)$$

Solutions for (5) are stationary points that satisfy the KKT conditions

$$\begin{aligned}V_{:i}\geq 0,\ \nabla_{V_{:i}}F(V_{:i}) &= -U_{:i}^T\left(R_j-U_{:i}V_{:i}^T\right)\\ &\quad+\gamma\mathbf{1}_{n\times 1}\geq 0,\ V_{:i}\#\nabla_{V_{:i}}F(V_{:i})=0.\end{aligned} \quad (8)$$

Such vectors are given by

$$V_{:i}^*=\max\left(\frac{R_j^TU_{:i}+\gamma\mathbf{1}_{n\times 1}}{\|U_{:i}\|_2^2},0\right). \quad (9)$$

The proposed algorithm which we call CDSNMF (coordinate descent for sparse NMF) uses (7) and (9) to compute the minimizers $U^*$ and $V^*$ of the objective function in (3). The main steps of CDSNMF are summarized in Algorithm 1.

*2.2. Model Selection.* In this paper, we use the model selection strategy described in [1]. One of the advantages of NMF algorithms is that given an $m$-by-$n$ microarray data, they are capable of grouping the $m$ samples into $r$ clusters,

where $r << \min(m, n)$ [1]. However, determining the number of clusters or rank $r$ which provide useful and meaningful interpretation is always a challenging task in data science. Moreover, NMF algorithms are sensitive to initializations and their stochastic nature poses a problem regarding converging to the same solution for different initial guesses. To overcome the issue of choosing $r$, Brunet et al. [1] developed a strategy by employing consensus clustering which is proved to facilitate the model selection process. For the sake of reducing the problem of convergence to the same point, the algorithm will be run several times for different initial points and a consensus matrix will be calculated by taking the average of several connectivity matrices which are of size $m$-by-$m$. The entries of a consensus matrix are valuable means of knowing whether related samples are clustered together. In the case of perfect clustering, its entries are either 0 or 1 and deviation from this optimal scenario indicates lack of stability in the clustering. Stability of clustering performance of an algorithm associated with rank $r$ can also be measured quantitatively by calculating the cophenetic correlation coefficient $\rho_r$ whose values range between 0 and 1. The fact of the matter is that $\rho_r$ closer to 1 indicates strong clustering performance (with $\rho_1 = 1$ meaning perfect clustering) whereas smaller values tell otherwise. In practice, one can study the behavior of $\rho_r$ by varying the value of $r$ and select the optimum value of the rank accordingly.

## 3. Numerical Results and Discussion

In this section, we conduct various experiments on leukemia and multiple myeloma datasets and discuss the results. Random initialization is used for all algorithms for its simplicity and easy implementation. As done in [1, 3, 19], the consensus matrices and cophenetic correlation coefficients were computed by averaging 50 connectivity matrices except for the new algorithm (CDSNMF) where averaging 10 of them is enough. We have compared CDSNMF with KL-NMF [1], SNMF [3], and NMFSC [19] which have employed the same model selection criteria. The tested algorithms involve some parameters: in CDSNMF, the sparsity parameter $\gamma = 5$ is used; in SNMF, $\lambda$ was set to 0.01 as suggested by the authors; and in NMFSC, the sparsity of factor $U$ was fixed to 0.5 whereas $V$ was left unconstrained as done by the authors. For each initialization (we did 50 of them), all algorithms were run until the maximum number of iteration reaches 1000. All experiments were conducted using MATLAB (R2019a) on a laptop Intel(R) Core(TM) i7-6500 U CPU @2.50 GHz 2.59 GHz 8 GB RAM.

*3.1. Leukemia Dataset.* The leukemia ALL-AML [2] gene expression dataset is a benchmark dataset widely used for cancer classification research and for comparing clustering performance of different algorithms. It contains 5000 genes and 38 tumor samples out of which 27 patients suffer from acute lymphoblastic leukemia (ALL) while 11 of them have acute myelogenous leukemia (AML). Since the id's of the columns (samples) of ALL-AML are very long to be used as axes labels, we prefer to use the corresponding indices (as given in the data) presented in Table 1. For instance, the fifth column has an index number of 5 while its id is ALL-9692-B-cell.

*3.2. Model Selection.* Figures 1(a)–1(e) depict the reordered consensus matrices for rank $r = 2, 3, 4, 5, 6$, respectively, and (f) shows the cophenetic correlation coefficient $\rho_r$ obtained by applying Algorithm 1 (CDSNMF) on the ALL-AML leukemia dataset. In all cases, i.e., for all the ranks, CDSNMF has produced very clear diagonal patterns as can be seen from this figure. The dark yellow colors in (a)–(e) correspond to a value of 1 and indicate that the gene samples are always clustered together based on their similarity whereas the dark blue ones are associated with a numerical value of 0 which means gene samples do not belong to the same cluster. The above qualitative results assure that the proposed algorithm is robust and cluster gene expression patterns without any dispersion whatsoever (except for $r = 6$ where there is a little dispersion). Moreover, we were able to assure what is seen in the reordered consensus matrices by using the quantitative measure called cophenetic correlation coefficient $\rho_r$ as shown in Figure 1(f). This measure also witnessed a perfect clustering $\rho_r = 1$ for all values of $r$ (except for $r = 6$ which has value $\rho_r = 0.9992$) which guaranteed the stability of the clustering performance of CDSNMF. Therefore, CDSNMF managed to give perfect consensus matrices with clear block diagonal patterns indicating that models for almost all the ranks are robust. Moreover, it can be observed that for almost all ranks the cluster splits are stable as per the model selection criteria and users can select a particular cluster according to their interest.

One can observe from Figure 2 that SNMF [3] was able to perfectly cluster the leukemia dataset in the case of $r = 2, 3, 4$ indicating robustness for these values. However, it is obvious that the result decreases a little when $r = 5$ and is not so good for $r = 6$ where there is a dispersion in the reordered consensus matrices. The value of $\rho_r$ is exactly one for $r = 2 - 4$ but drops significantly when $r$ increases from 5 to 6.

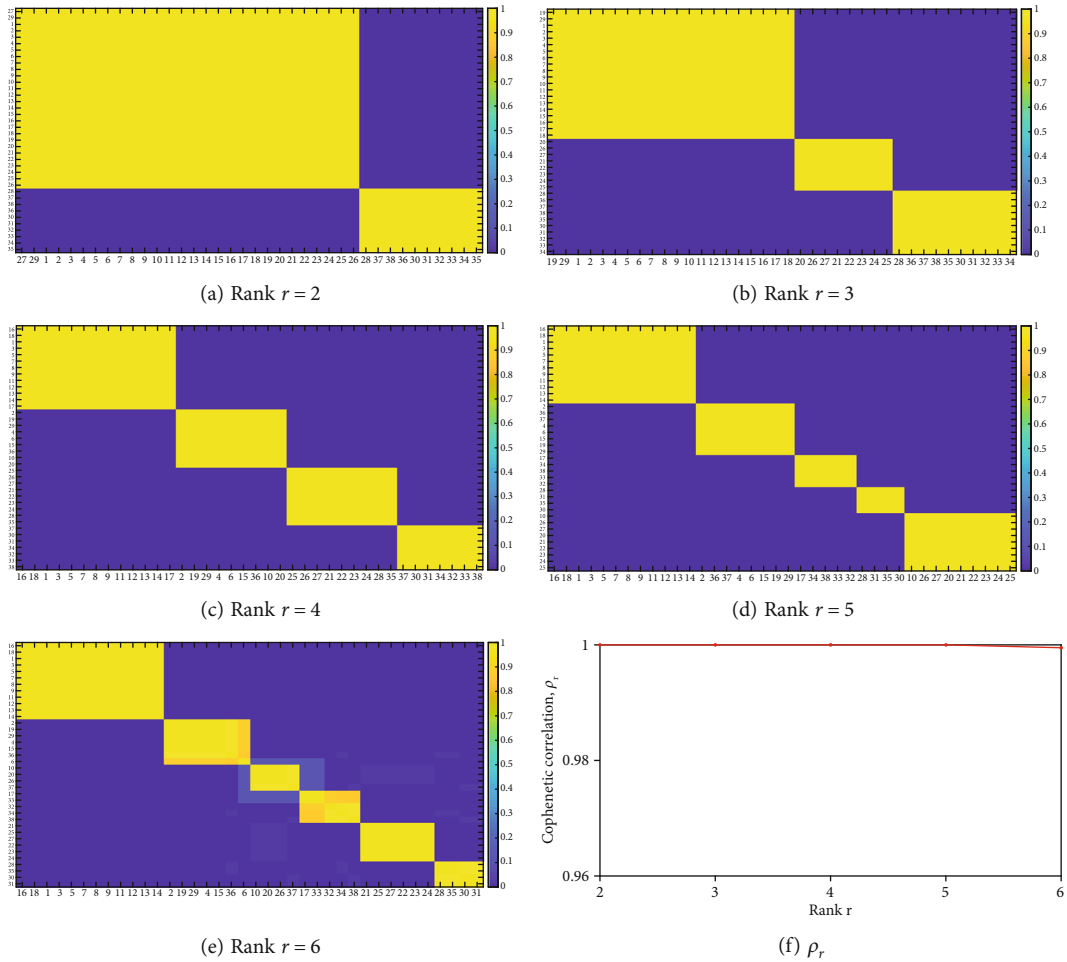
Figure 3 presents results obtained by applying NMFSC [19]. The consensus matrices show dispersion for all values of $r$. These results are replicated in the sense of $\rho_r$ indicating that the performance of NMFSC is not that good.

Figure 4 depicts results obtained by applying KL-based standard NMF [1]) to the leukemia dataset but the algorithm does not perform well in most cases. The reordered consensus matrices are clearly visible only in the case of $r = 2$, and dispersion of clustering increases as the rank $r$ increases from 2 to 6. The value of $\rho_r$ is also shown to decrease as $r$ increases indicating that the clustering performance of KL-NMF is not stable.

In conclusion, CDSNMF possesses a much higher clustering performance than SNMF, NMFSC, and KL-NMF. In fact, CDSNMF has shown a perfect clustering performance by providing very clear block diagonal patterns of reordered consensus matrices and a perfect score of 1 for the cophenetic correlation coefficient for all ranks except for $r = 6$ which is very close to 1 ($\rho_r = 0.9992$ to be exact, see Figure 1). CDSNMF assures that the samples are grouped

Table 1: Original indices of columns of Golub's ALL-AML leukemia data.

| Index | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| ALL | 19769-B-cell | 23953-B-cell | 28373-B-cell | 9335-B-cell | 9692-B-cell | 14749-B-cell | 17281-B-cell | 19183-B-cell | 20414-B-cell |
| Index | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 |
| ALL | 21302-B-cell | 549-B-cell | 17929-B-cell | 20185-B-cell | 11103-B-cell | 18239-B-cell | 5982-B-cell | 7092-B-cell | R11-B-cell |
| Index | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 |
| ALL | R23-B-cell | 16415-T-cell | 19881-T-cell | 9186-T-cell | 9723-T-cell | 17269-T-cell | 14402-T-cell | 17638-T-cell | 22474-T-cell |
| Index | 28 | 29 | 30 | 31 | 32 | 33 | 34 | 35 | 36 |
| AML | AML-12 | AML-13 | AML-14 | AML-16 | AML-20 | AML-1 | AML-2 | AML-3 | AML-5 |
| Index | 37 | 38 | | | | | | | |
| AML | AML-6 | AML-7 | | | | | | | |




(a) Rank $r = 2$ (b) Rank $r = 3$

(c) Rank $r = 4$ (d) Rank $r = 5$

(e) Rank $r = 6$ (f) $\rho_r$

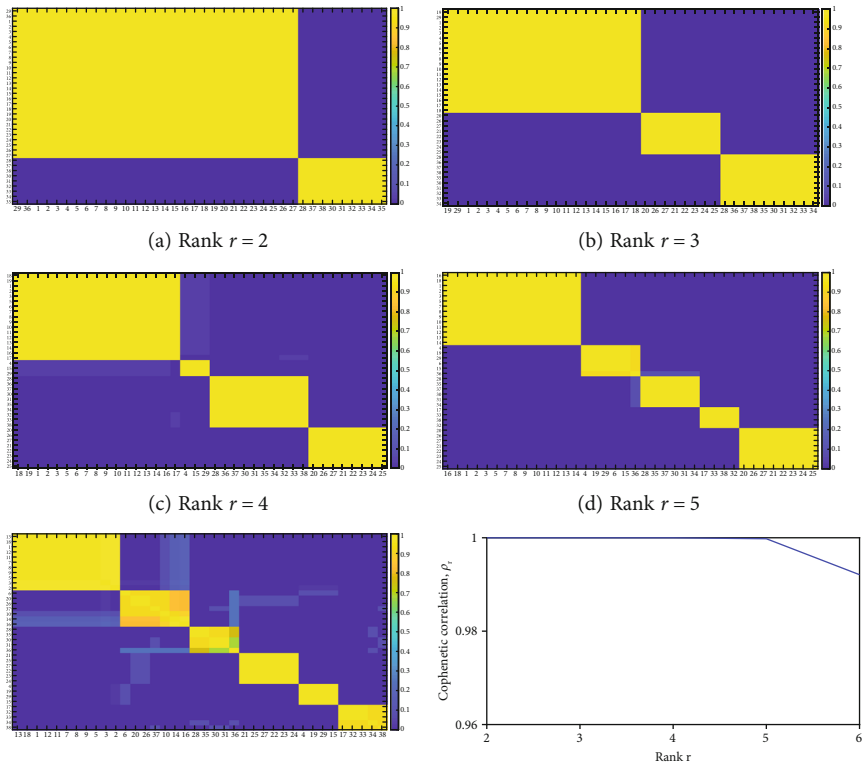
Figure 1: CDSNMF: (a–e) reordered consensus matrices and (f) cophenetic correlation coefficient $\rho_r$ for $r = 2$-6 using ALL-AML.

as per the similarity that exist in their metagene expression profiles and it makes sure that there is some regularities within the groups.

Next, we discuss how the groups of subjects (samples) change varying the algorithms and rank $r$. One can observe the following:

(1) $r = 2$: CDSNMF and SNMF provide two similar blocks (yellow squares) except that the samples are ordered a little differently. In addition, sample 36 (AML-5) belongs to the first group in SNMF whereas it is grouped in the second cluster in CDSNMF. KL-NMF also provided two clear diagonal blocks but the samples are grouped very differently. The clusters of NMFSC also do not share similarities with the other methods

(2) $r = 3$: here, it is interesting to observe that CDSNMF and SNMF provide exactly the same clusters. It is worth mentioning that the ordering of the samples

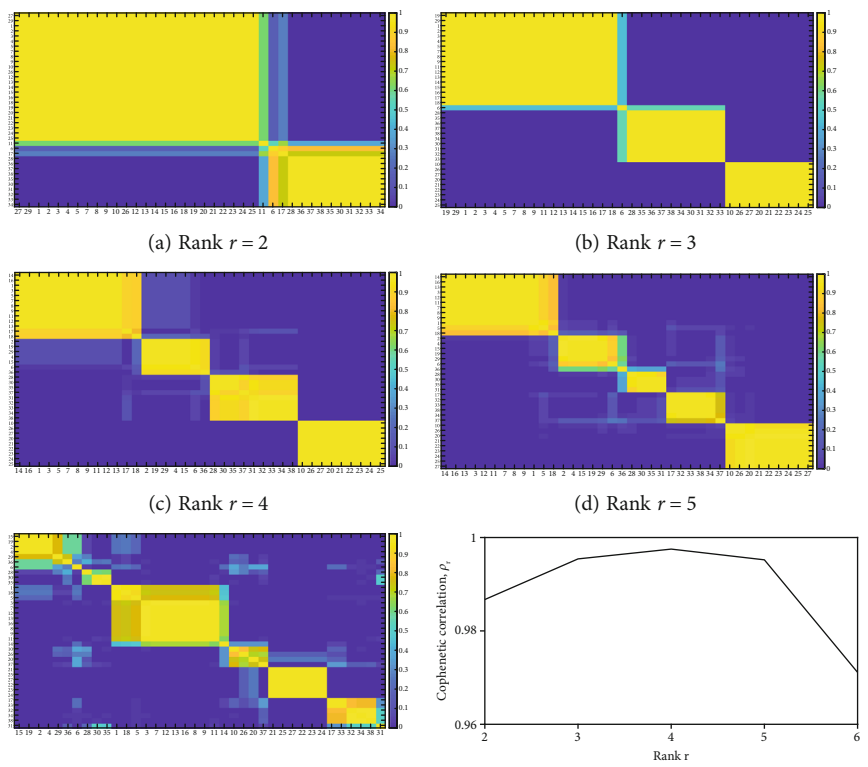



(a) Rank $r = 2$ (b) Rank $r = 3$

(c) Rank $r = 4$ (d) Rank $r = 5$

(e) Rank $r = 6$ (f) $\rho_r$

FIGURE 2: SNMF: (a–e) reordered consensus matrices and (f) cophenetic correlation coefficient $\rho_r$ for $r = 2$-6 using ALL-AML.

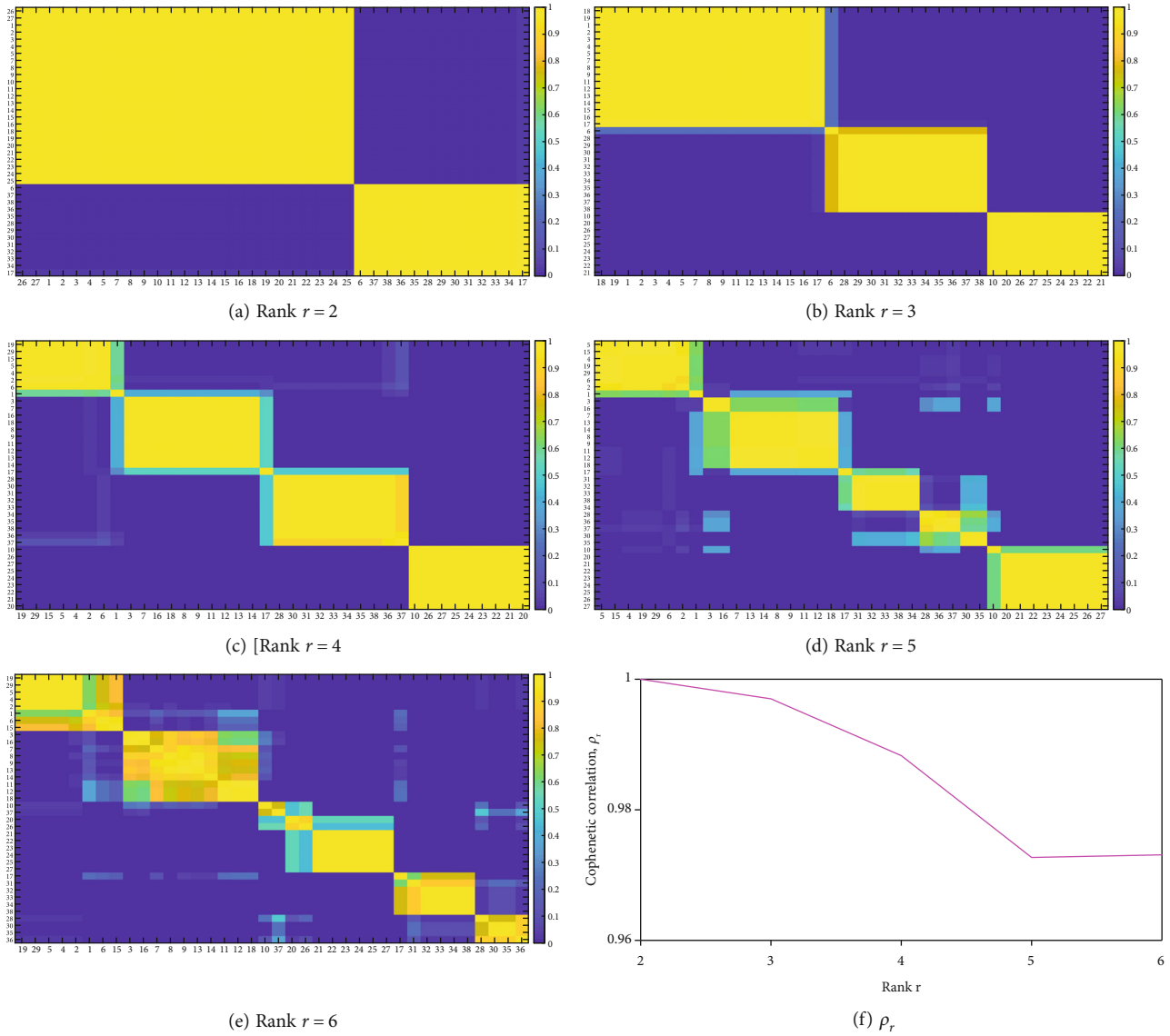



(a) Rank $r = 2$ (b) Rank $r = 3$

(c) Rank $r = 4$ (d) Rank $r = 5$

(e) Rank $r = 6$ (f) $\rho_r$

FIGURE 3: NMFSC: (a–e) reordered consensus matrices and (f) cophenetic correlation coefficient $\rho_r$ for $r = 2$-6 using ALL-AML.




(a) Rank $r = 2$ (b) Rank $r = 3$

(c) [Rank $r = 4$ (d) Rank $r = 5$

(e) Rank $r = 6$ (f) $\rho_r$

FIGURE 4: KL-NMF: (a–e) reordered consensus matrices obtained by averaging 50 connectivity matrices and (f) cophenetic correlation coefficient $\rho_r$ for $r = 2$-6 using ALL-AML.

in all the groups is identical. Once again, the groupings of NMFSC are completely different from the abovementioned methods. Unfortunately, the same is true for KL-NMF

(3) For the rest of the ranks, i.e, $r = 4$, $r = 5$, and $r = 6$, all algorithms cluster the samples of the data in a different manner.

*3.2.1. Optimal Rank, Clustering, and Metagene Analysis.* Figure 5(a) compares cophenetic correlation coefficient $\rho_r$ of the four algorithms by varying the value of $r$ from 2 to 9. This helps to determine the optimal rank as well as identify the best method. It is shown, in this figure, that CDSNMF performs better and $r = 5$ can be chosen as the optimal rank (of course $r = 6$ is also a nice choice).

Figure 5(b) depicts the reordered consensus matrix using CDSNMF and $r = 5$ (optimal rank). For this rank, CDSNMF provides five clusters which are the clear yellow blocks on the diagonal indicating that there are some regularities within the groups and the clustering is made as per the underlying similarities in the metagene expression profiles.

Figure 5(c) shows the stacked bar plot of the encoding matrix $V$ as provided by CDSNMF for $r = 5$. Each bar represents the sum of the composition of metagenes that make up the sample of the original data. This figure enables us to investigate the contribution of the metagenes to each sample and their distribution in the 5 different clusters. It is interesting to observe that all metagenes contribute to all clusters. There is a high expression of Metagene 1 in Cluster 4 (the least amount is a little over 30 percent in sample AML-3 while the highest percentage is a little over 70 in AML-14.

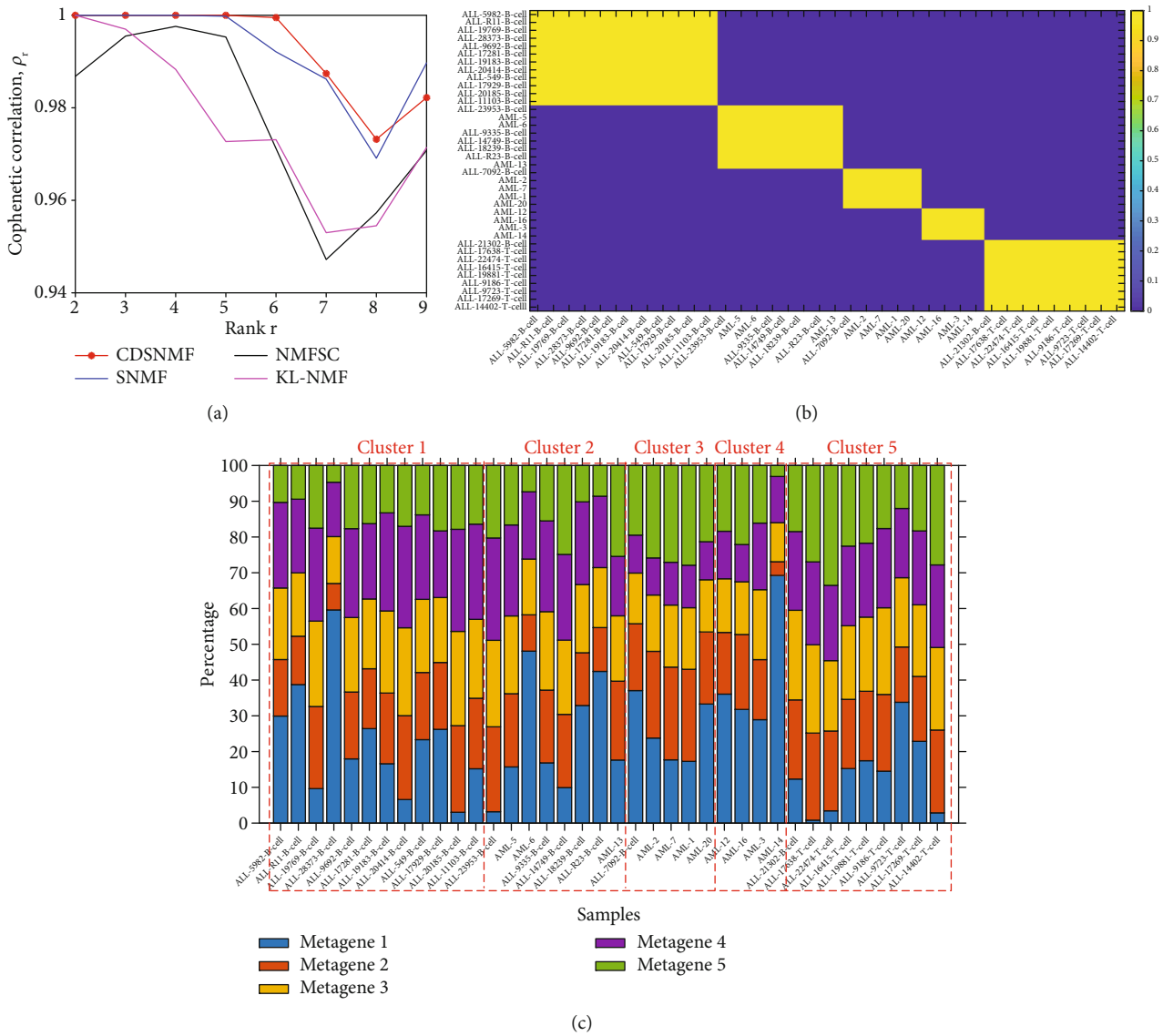



FIGURE 5: ALL-AML data: (a) cophenetic correlation coefficient by varying $r$ from 2 to 9. (b) Reordered consensus matrix for the optimal rank $r = 5$ obtained by using the winner algorithm CDSNMF. (c) Stacked bar plots of the encoding matrix $V$ provided by CDSNMF for the optimal rank $r = 5$.

On the other hand, its concentration is low in Cluster 5 (in particular in ALL-17638-T-cell and ALL-14402-T-cell). It is also very low in ALL-12085-B-cell of Cluster 1 and ALL-23953-B-cell of Cluster 2. Metagenes 2, 3, and 5 are seen to be fairly distributed in all clusters except Cluster 4. This cluster has a very low amount of Metagenes 2 and 5 in sample AML-14. One can also observe that Cluster 3 and Cluster 4 have relatively low expressions of Metagene 4.

Figure 6 depicts the first top 10 dominant genes that highly influence the basis matrix $U$ and the five metagenes extracted from ALL-AML by CDSNMF. This figure helps us to identify the genes that influence the obtained bases. The basis matrix $U$ is a gene-by-metagene matrix of size $5000 \times 5$. Rearranging the columns of $U$ by sorting the weights (entries) in decreasing order puts the dominant genes at the top, we can then use bar plots to investigate the effect of each gene. Figure 6(a) shows that gene M25079-s-at with 3.2 percent is the most dominant one in Metagene 1 followed by Z84721-cds2-at and X57351-s-at with a percentage of 1.68 and 0.93, respectively. One can also observe that the contribution of the genes in forming the five extracted metagenes is not far from one another (mostly less than 1 percent). For instance, the two 'most dominant' genes in Metagene 2 are M1147-at (0.675 percent) and X57351-s-at (0.669 percent), and the rest are also very close to one another. Next, let us discuss how the genes affect samples of the data. The contribution of Metagene 1 is very high in sample AML-14; this means the expressions of the genes M25079-s-at, Z84721-cds2-at, X57351-s-at, and X00274-at (taking only the first four) play a significant role in defining AML-14. On the other hand, the influence of the genes in Metagene 1 is very small in sample ALL-22474-T-cell

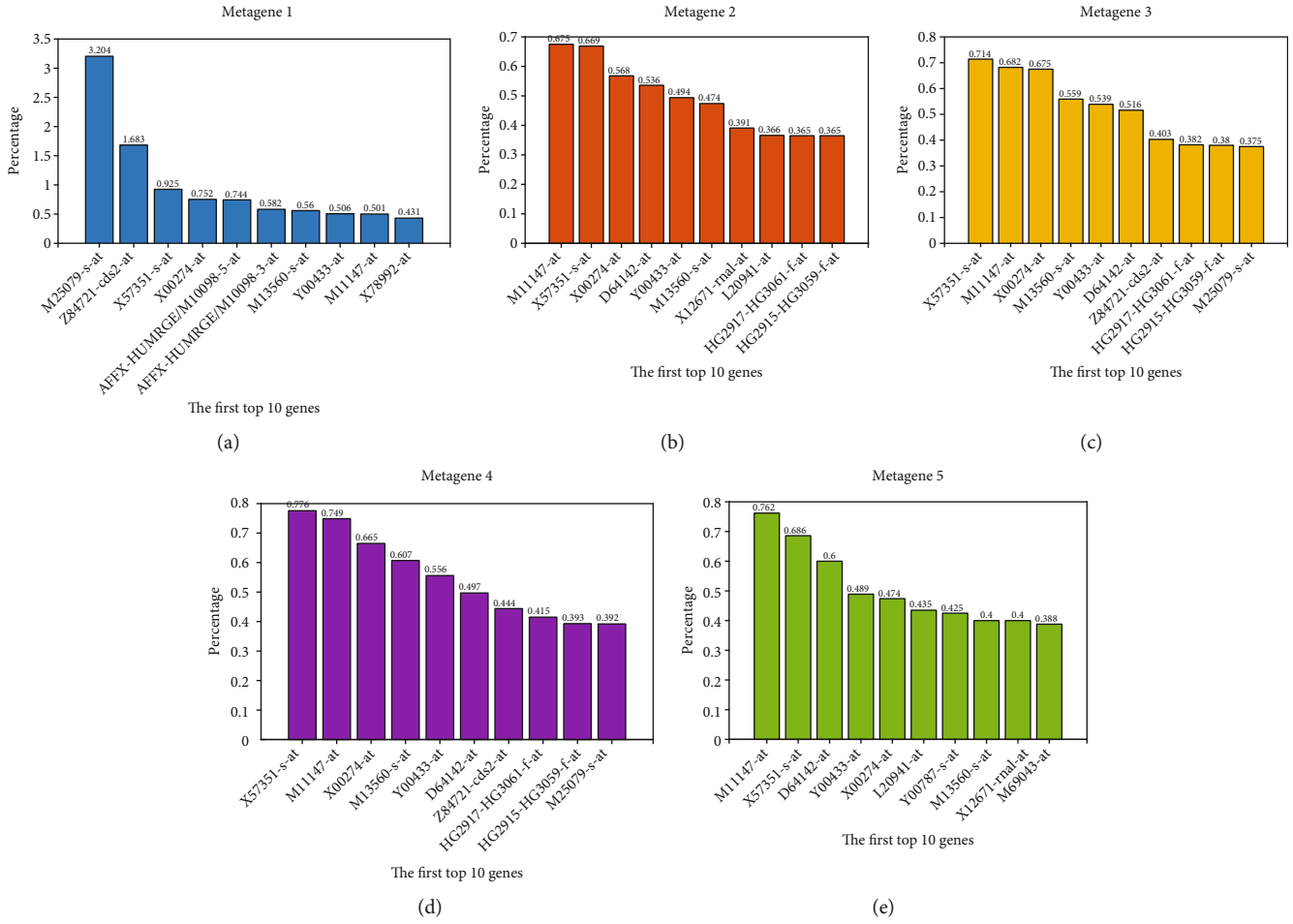


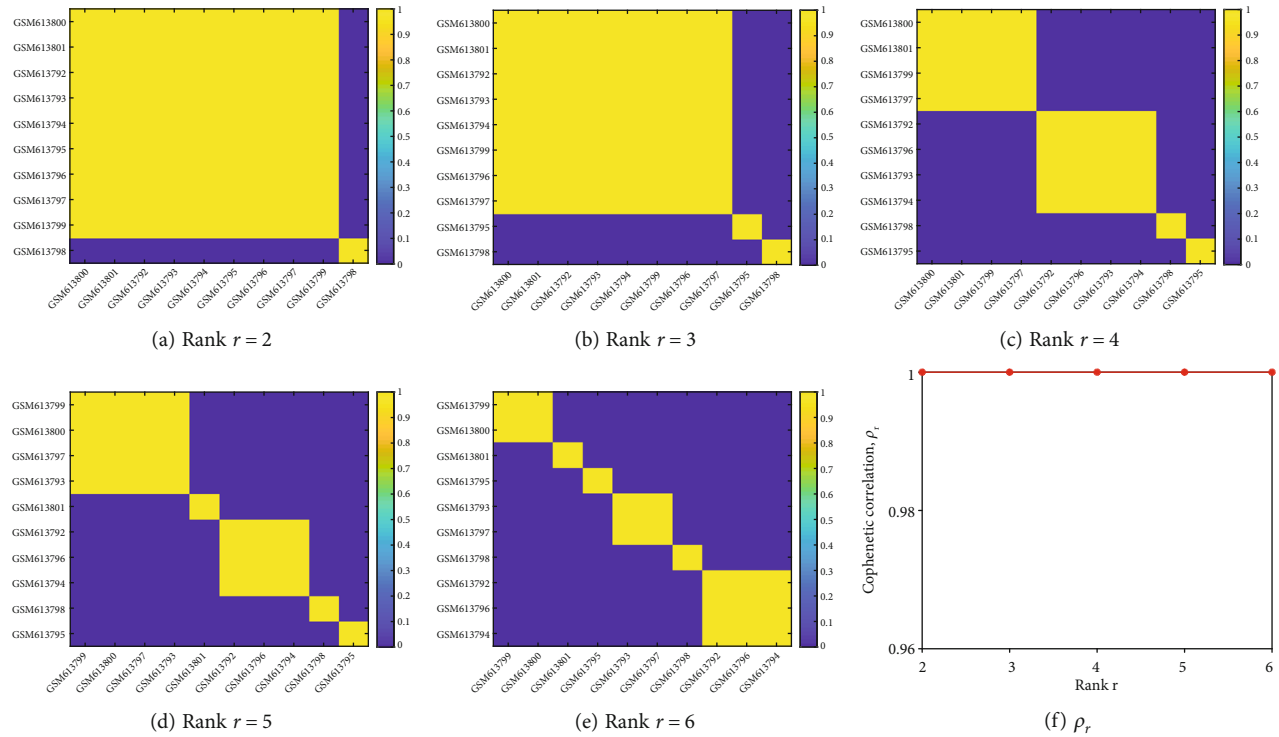

Figure 6: Composition of the five metagenes (provided by CDSNMF) of ALL-AML based on the first top 10 dominant original genes.

whereas the genes that define Metagene 5 (like M1147-at, X57351-s-at, and D64142-at) exist in large amount. We can proceed in a similar way to analyze the rest of the metagenes and the genes that define them.

*3.3. Multiple Myeloma (MM).* MM [13] is an incurable blood cancer that forms in a type of white blood cell called plasma cell. The disease can damage bones and main organs of a human body like the kidneys. It is characterized by the presence of a monoclonal component of plasma cells in the bone marrow. The disease also disrupts immune systems and red blood cell count. The dataset consists of 15464 genes and 10 samples.

*3.3.1. Model Selection.* CDSNMF has performed remarkably well in the MM dataset also and emerged as a clean winner once again. As shown in Figure 7, the block diagonal patterns of all the consensus matrices are clearly depicted and the clustering performance of CDSNMF is once again proved to be stable as it is accompanied by a perfect score of the cophenetic correlation coefficient $\rho_r$. This means that the models for the classes obtained by taking $r = 2$-6 are robust. As for the comparison, SNMF is found to be the second winner, since it has a robust performance for the ranks $r = 2, 3, 4$. But, as shown in Figure 8, the reordered consensus matrices provided by this algorithm for $r = 5, 6$ exhibit a bit of dispersion, more in the case of $r = 5$. The cophenetic correlation score of SNMF also drops when $r$ goes from 4 to 5 but goes up a little higher when $r = 6$ which indicates that its performance is not that stable. As depicted in Figure 9, NMFSC performs well for the first two values of $r$ but its performance decreases when $r$ increases from 3 to 4 and gets better when $r = 5$ but again becomes lower for $r = 6$. It can be observed that NMFSC is also not stable as its cophenetic correlation graph exhibits a clear zigzagging behavior. Figure 10 depicts the results obtained from the classical KL-NMF; one can see that this algorithm is not attractive as its performance zigzags here and there.

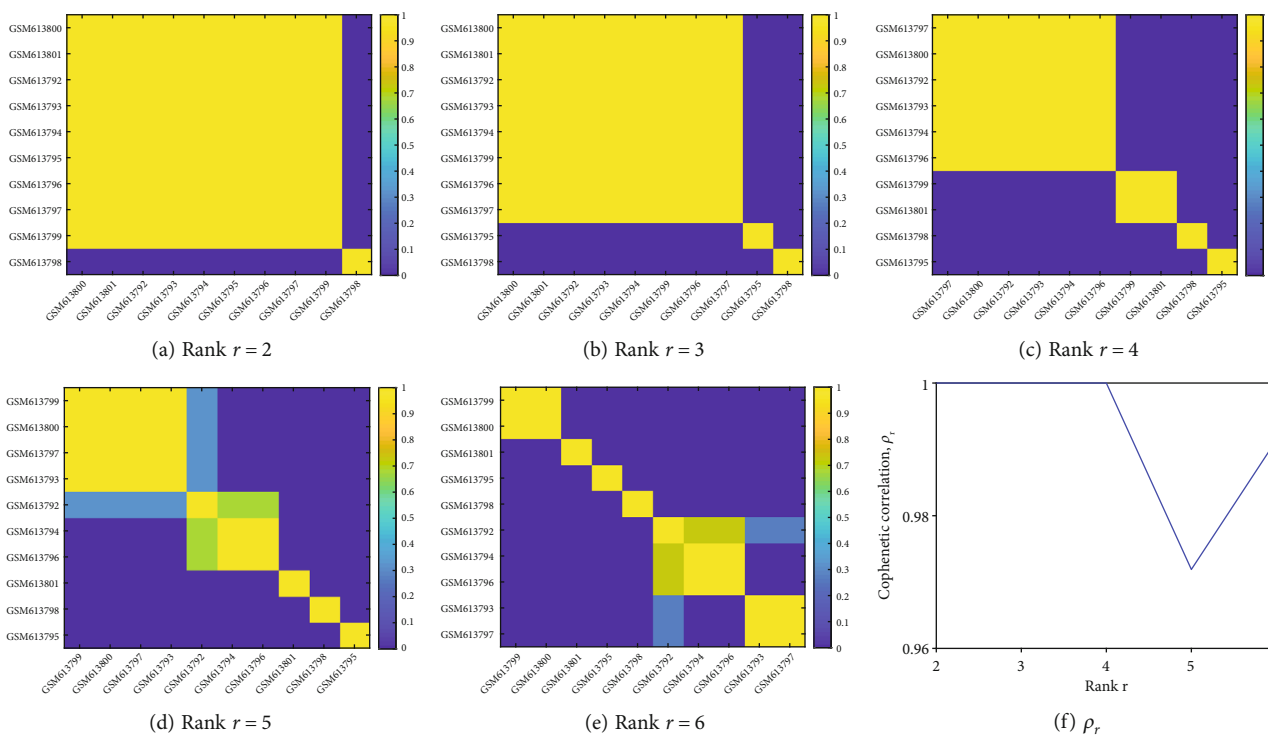
It is interesting to investigate how the tested algorithms cluster the samples for different values of $r$. For $r = 2$, all algorithms cluster samples in a very similar way. For $r = 3$, the clusters of samples provided by CDSNMF, SNMF, and NMFSC are identical. However, the only difference between KL-NMF and the other three is that the former places sample GSM613793 in Cluster 1 (instead of Cluster 2). For $r = 4$, only the last two clusters of the four algorithms are similar; they contain samples GSM613798 and GSM613795, respectively. For this particular rank, the only difference




(a) Rank $r = 2$ (b) Rank $r = 3$ (c) Rank $r = 4$

(d) Rank $r = 5$ (e) Rank $r = 6$ (f) $\rho_r$

Figure 7: CDSNMF: (a–e) reordered consensus matrices and (f) cophenetic correlation coefficient $\rho_r$ for $r = 2$-6 using MM.

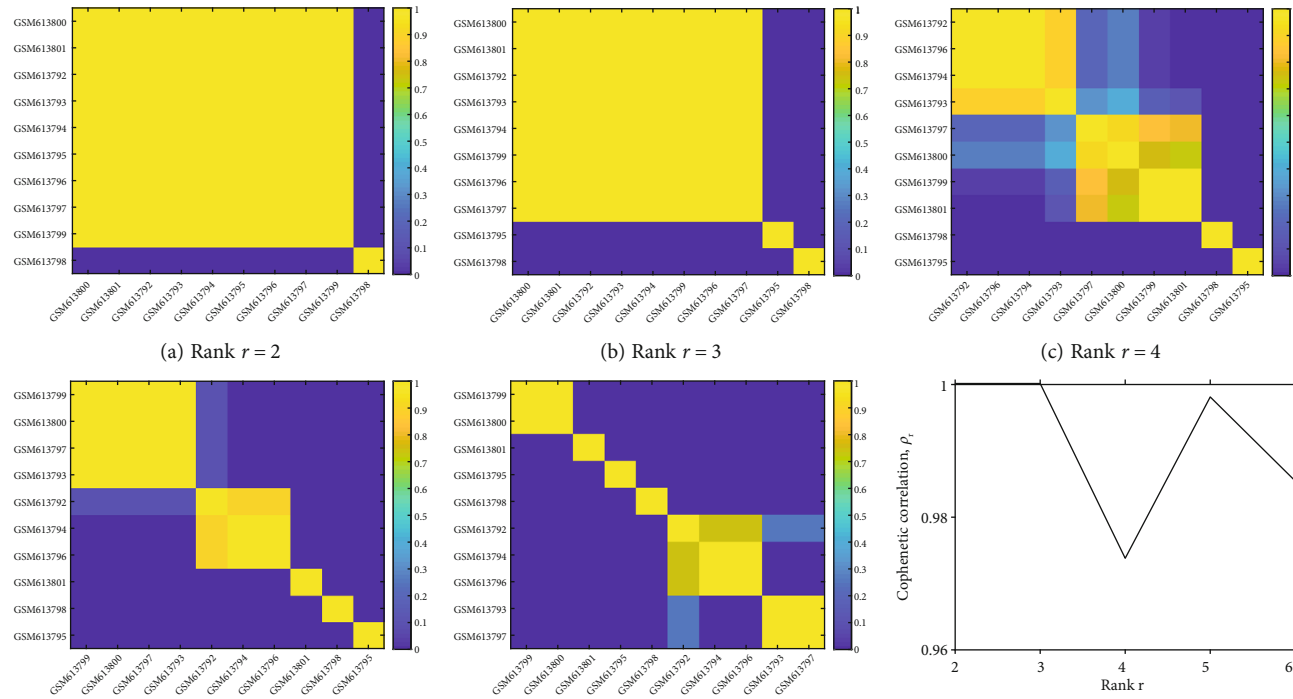



(a) Rank $r = 2$ (b) Rank $r = 3$ (c) Rank $r = 4$

(d) Rank $r = 5$ (e) Rank $r = 6$ (f) $\rho_r$

Figure 8: SNMF: (a–e) reordered consensus matrices and (f) cophenetic correlation coefficient $\rho_r$ for $r = 2$-6 using MM.

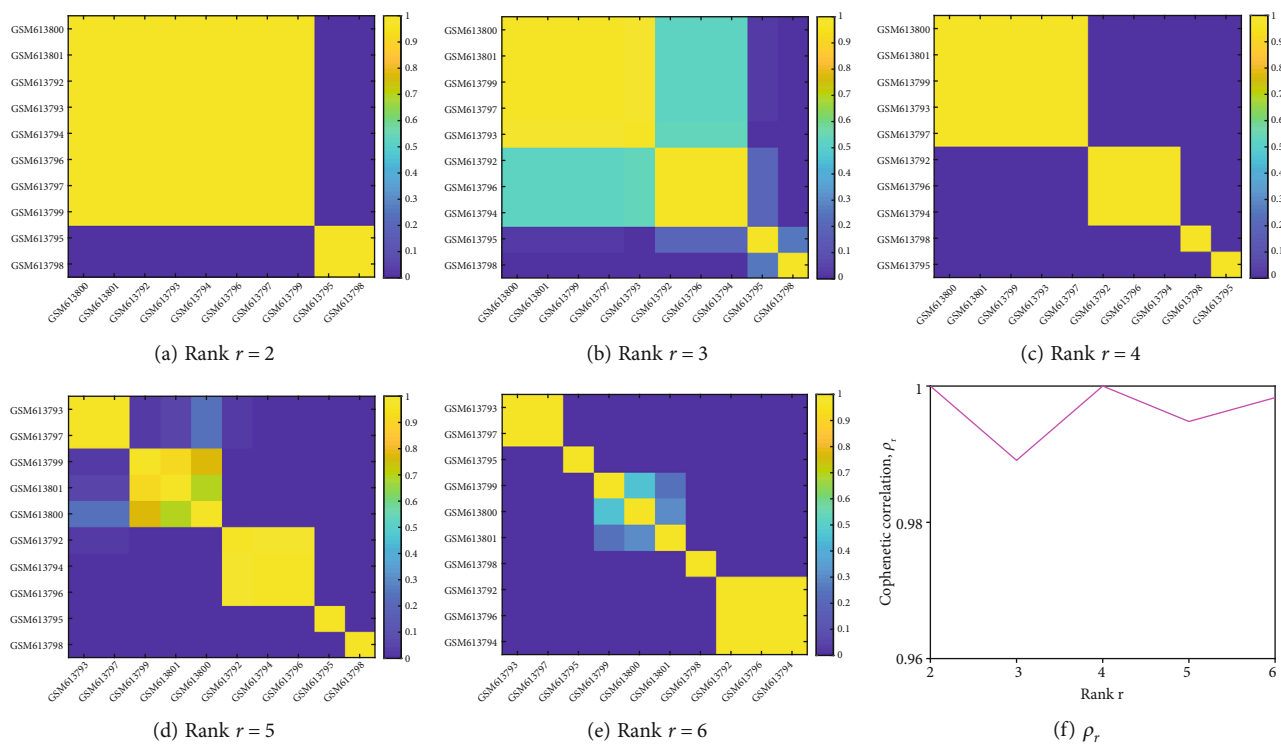
between CDSNMF and KL-NMF is that GSM613793 belongs to the first cluster in the former whereas it is placed in the second cluster by the later. It is not possible to draw a clear conclusion about SNMF and NMFSC regarding the first two clusters for $r = 4$. In the case when $r = 5$, Clusters 1, 4, and 5 of CDSNMF, SNMF, and




(a) Rank $r = 2$ (b) Rank $r = 3$ (c) Rank $r = 4$

(d) Rank $r = 5$ (e) Rank $r = 6$ (f) $\rho_r$

FIGURE 9: NMFSC: (a–e) reordered consensus matrices and (f) cophenetic correlation coefficient $\rho_r$ for $r = 2$-6 using MM.




(a) Rank $r = 2$ (b) Rank $r = 3$ (c) Rank $r = 4$

(d) Rank $r = 5$ (e) Rank $r = 6$ (f) $\rho_r$

FIGURE 10: KL-NMF: (a–e) reordered consensus matrices and (f) cophenetic correlation coefficient $\rho_r$ for $r = 2$-6 using MM.

NMFSC are exactly the same but the other clusters are totally different. However, KL-NMF clusters the samples very differently. For $r = 6$, the first three clusters of CDSNMF, SNMF, and NMFSC are identical. Once again, the clusters in KL-NMF do not share similarities with others.

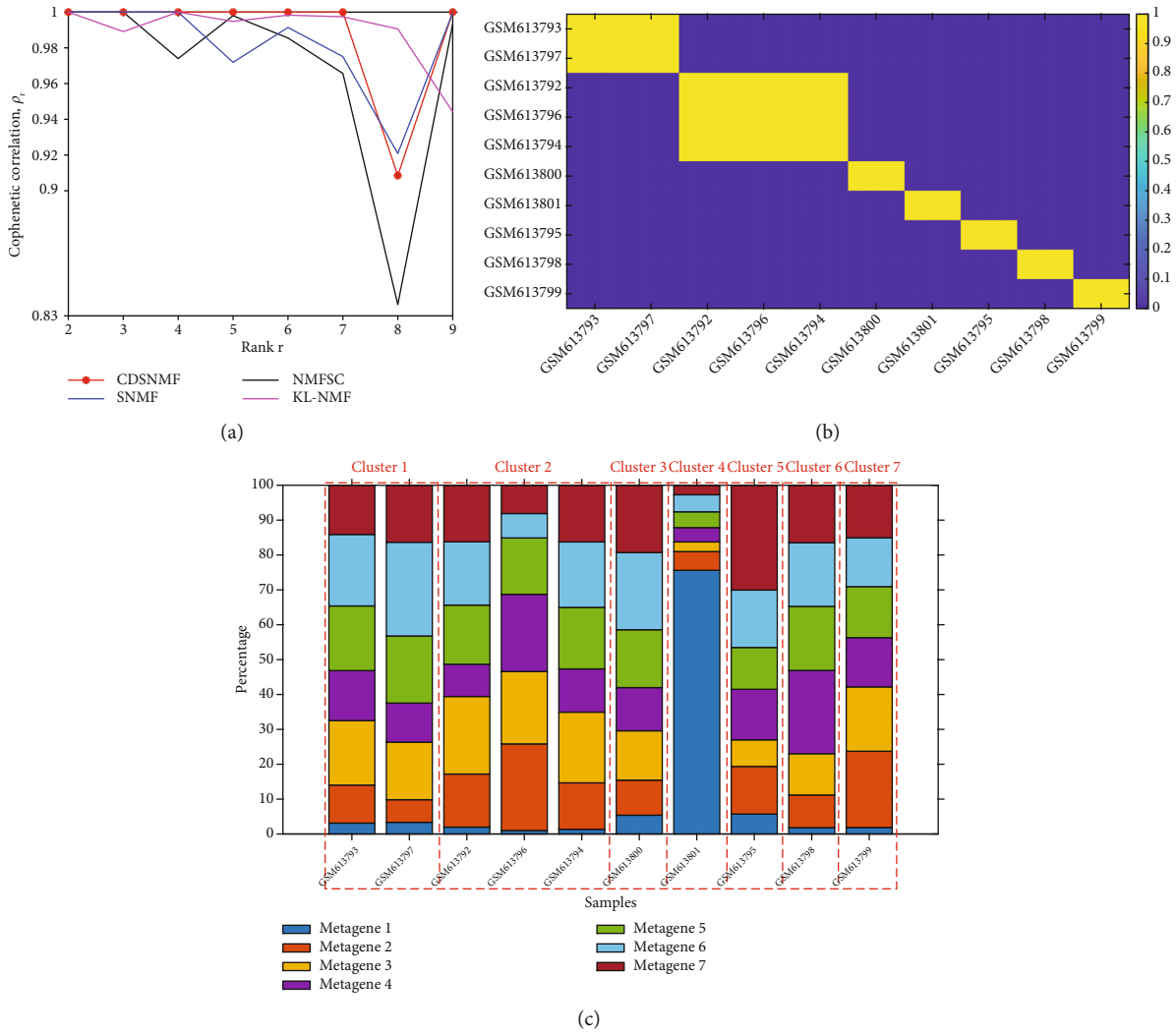



FIGURE 11: MM data: (a) cophenetic correlation coefficient by varying $r$ from 2 to 9. (b) Reordered consensus matrix for the optimal rank $r = 7$ using the winner algorithm CDSNMF. (c) Stacked bar plots of the encoding matrix $V$ provided by CDSNMF for the optimal rank $r = 7$.

*3.3.2. Optimal Rank, Clustering, and Metagene Analysis.* Figure 11(a) depicts the cophenetic correlation coefficient graphs obtained by varying the rank $r$ from 2 to 9. It can be seen that CDSNMF performs best and managed to find a larger class of clusters with an optimal rank of $r = 7$.

In Figure 11(b), the reordered consensus matrix for the optimal rank $r = 7$ is plotted using CDSNMF. For this rank, CDSNMF provides seven clear yellow diagonal blocks which assure the existence of regularities within the groups. It can also be concluded that the samples are clustered based on the underlying similarities in the metagene expression profiles.

Figure 11(c) presents the stacked bar plot of the encoding matrix $V$ as provided by CDSNMF for $r = 7$. As indicated in this figure, all the bars consist of proportions of the metagenes that define samples of the data. One can observe that there is a very low expression of Metagene 1 in all clusters except Cluster 4 (sample GSM613801). The expression of Metagenes 2 and 3 seem to be very high in Clusters 2. It is also good to notice that Cluster 5 (sample GSM613795) has the least amount of Metagene 3 but more of Metagene 7 as compared to the other samples. In most samples, the expression of Metagene 4 is more or less the same except in samples GSM613796 and GSM613798 which enjoy a little bit higher concentrations. Metagenes 5 and 6 are fairly be distributed in all the samples except Cluster 4 (sample GSM613801) whose share is relatively small.

Figure 12 depicts the first top 10 dominant genes that highly influence the basis matrix $U$ and the seven metagenes extracted from MM by CDSNMF. This figure helps us to identify the genes that influence the obtained bases. The basis matrix $U$ is a gene-by-metagene matrix of size 15464 $\times 7$. It is worth mentioning that, as presented in this figure, the role of the original genes of the data in defining each of the metagenes is very close to one another. This is justified by the fact that the contribution of each of the genes is less than one percent. For instance, the most dominant genes that play a vital role in defining the seven metagenes have percentages 0.832 (TOM1L2), 0.135 (ZBTB32), 0.104

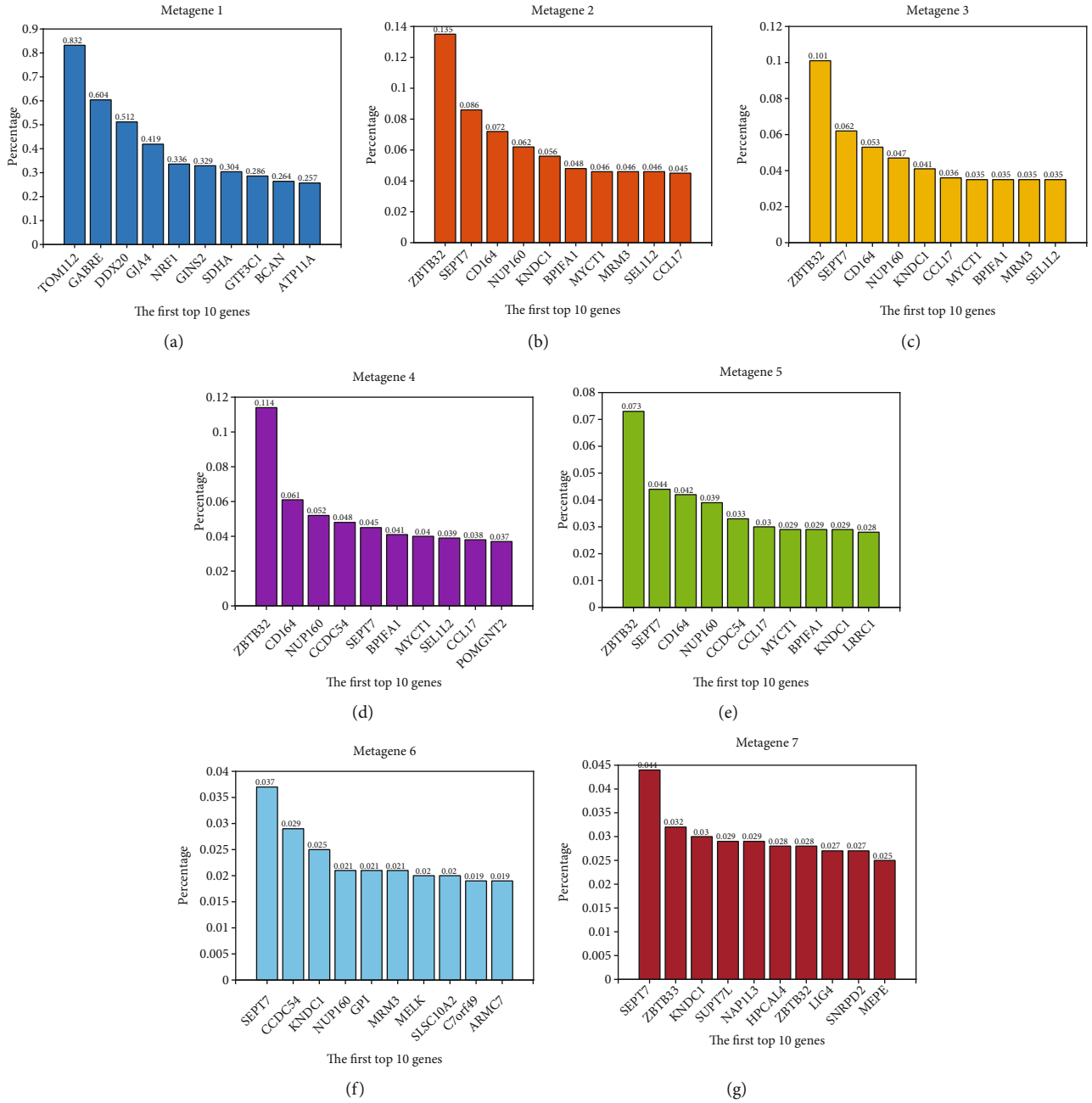



Figure 12: Composition of the seven metagenes (provided by CDSNMF) of MM based on the first top 10 original genes.

(ZBTB32), 0.114 (ZBTB32), 0.073 (ZBTB32), 0.037 (SEPT7), and 0.044 (SEPT7), respectively. The other thing worth mentioning here is that the heaviest genes that exist in Metagene 1, namely, TOM1L2, GABRE, DDX20, GJA4, and NRF1 (considering only the first five) highly influence how sample GSM613801 is made. On the other hand, the contribution of these genes is negligible in samples GSM613796 and extremely small in GSM613794. In addition, genes SEPT7, ZBTB33, KNDC1, SUPT7L, and NAP1L3 (only taking the tops 5 genes from Metagene 7) play a paramount role in sample GSM613795 whereas there is a very low concentration of them in sample GSM613801. Similar analysis can be made for the remaining genes and samples.

## 4. Conclusive Remarks

Nonnegative matrix factorization has become very popular for handling high-dimensional data. By its very nature, NMF facilitates the extraction and interpretation of real-life datasets including gene expression microarray datasets. In this paper, we extended the RRI algorithm which we abbreviate as CDSNMF to solve a nonconvex optimization problem posed in the form of sparse nonnegative matrix factorization. The algorithm uses block coordinate descent approach and is proved to work very well in practice. We have considered two cancer datasets, namely, ALL-AML and MM. The experimental results show that the new algorithm is capable of

discovering larger cancer classes and providing perfect consensus matrices whose block diagonal patterns are very clear. This signifies that the models for different rank factorizations are robust. Moreover, CDSNMF gives a perfect score of 1 (in almost all cases) for the cophenetic correlation coefficient indicating that the clustering performance of the algorithm is stable. In addition, the experimental results reveal that the new algorithm significantly outperforms other related state-of-the-art methods. The algorithm is also shown to be capable of identifying the dominant genes that influence the basis elements and extracted metagenes.

## Data Availability

Data are available at (1) doi:10.1126/science.286.5439.531, (2) https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE24990, and (3) https://translational-medicine.biomedcentral.com/articles/10.1186/s12967-018-1589-1.

## Conflicts of Interest

The author declares that there are no conflict of interests regarding the submission and publication of this work from any party.

## Acknowledgments

The author would like to pass his heartfelt thanks to Dr. Flavia Esposito at University of Bari (Italy) for providing the preprocessed versions of the datasets used in the paper.

## References

[1] J.-P. Brunet, P. Tamayo, T. R. Golub, and J. P. Mesirov, "Metagenes and molecular pattern discovery using matrix factorization," *Proceedings of the National Academy of Sciences*, vol. 101, no. 12, pp. 4164–4169, 2004.

[2] T. R. Golub, D. K. Slonim, P. Tamayo et al., "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," *Science*, vol. 286, no. 5439, pp. 531–537, 1999.

[3] Y. Gao and G. Church, "Improving molecular cancer class discovery through sparse non-negative matrix factorization," *Bioinformatics*, vol. 21, no. 21, pp. 3970–3975, 2005.

[4] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein, "Cluster analysis and display of genome-wide expression patterns," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 95, no. 25, pp. 14863–14868, 1998.

[5] A. A. Alizadeh, M. B. Eisen, R. E. Davis et al., "Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling," *Nature*, vol. 403, no. 6769, pp. 503–511, 2000.

[6] C. M. Perou, T. Sorlie, M. B. Eisen et al., "Molecular portraits of human breast tumours," *Nature*, vol. 406, no. 6797, pp. 747–752, 2000.

[7] A. Frigyesi and M. Hoglund, "Non-negative matrix factorization for the analysis of complex gene expression data: identification of clinically relevant tumor subtypes," *Cancer Informatics*, vol. 6, pp. 275–292, 2008.

[8] K. Devarajan, "Nonnegative matrix factorization: an analytical and interpretive tool in computational biology," *PLoS Computational Biology*, vol. 4, no. 7, article e1000029, 2008.

[9] P. Tamayo, D. Slonim, J. Mesirov et al., "Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 96, no. 6, pp. 2907–2912, 1999.

[10] T. D. Moloshok, R. R. Klevecz, J. D. Grant et al., "Application of Bayesian decomposition for analysing microarray data," *Bioinformatics*, vol. 18, no. 4, pp. 566–575, 2002.

[11] A. P. Gasch and M. B. Eisen, "Exploring the conditional coregulation of yeast gene expression through fuzzy k-means clustering," *Genome Biology*, vol. 3, no. 11, article research0059.1, 2002.

[12] O. Alter, P. O. Brown, and D. Botstein, "Singular value decomposition for genome-wide expression data processing and modeling," *PNAS*, vol. 97, no. 18, pp. 10101–10106, 2000.

[13] A. Boccarelli, F. Esposito, M. Coluccia, M. A. Frassanito, A. Vacca, and N. del Buono, "Improving knowledge on the activation of bone marrow fibroblasts in MGUS and MM disease through the automatic extraction of genes via a nonnegative matrix factorization approach on gene expression profiles," *BMC Translational Medicine*, vol. 16, no. 1, p. 217, 2018.

[14] D. R. Carrasco, G. Tonon, Y. Huang et al., "High-resolution genomic profiles define distinct clinico-pathogenetic subgroups of multiple myeloma patients," *Cancer Cell*, vol. 9, no. 4, pp. 313–325, 2006.

[15] P. M. Kim and B. Tidor, "Subsystem Identification through dimensionality reduction of large-scale gene expression data," *Genome Research*, vol. 13, no. 7, pp. 1706–1718, 2003.

[16] G. Casalino, M. Coluccia, M. L. Pati et al., "Intelligent microarray data analysis through non-negative matrix factorization to study human multiple myeloma cell lines," *Applied Sciences*, vol. 9, no. 24, p. 5552, 2019.

[17] D. Lee and S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[18] H. Kim and H. Park, "Sparse nonnegative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis," *Bioinformatics*, vol. 23, no. 12, pp. 1495–1502, 2007.

[19] X. Kong, C. Zheng, Y. Wu, and L. Shang, "Molecular cancer class discovery using non-negative matrix factorization with sparseness constraint," in *Advanced Intelligent Computing Theories and Applications with Aspects of Theoretical and Methodological Issues*, vol. 4681, pp. 792–802, Springer, 2007.

[20] P. O. Hoyer, "Nonnegative matrix factorization with sparseness constraints," *Journal of Machine Learning Research*, vol. 5, pp. 1457–1469, 2004.

[21] F. Esposito, N. Gillis, and N. Del Buono, "Orthogonal joint sparse NMF for microarray data analysis," *Mathematical Biology*, vol. 79, no. 1, pp. 223–247, 2019.

[22] N.-D. Ho, *Nonnegative matrix factorization algorithms and applications, [Ph.D. thesis]*, Universite Catholique de Louvain, 2008.

[23] A. Cichocki, R. Zdunek, A. H. Phan, and S. Amari, *Nonnegative Matrix and Tensor Factoriza-tions: Applications to Exploratory Multi-way Data Analysis and Blind Source Separation*, Wiley, UK, 2009.

[24] A. Cichocki, R. Zdunek, and S. Amari, "Hierarchical ALS algorithms for nonnegative matrix and 3D tensor factorization," in *Proceedings of 7th International Conference on Independent Component Analysis and Signal Separation*, pp. 169–176, Berlin, Heidelberg, 2007.

[25] M. T. Belachew, "Efficient algorithm for sparse symmetric nonnegative matrix factorization," *Pattern Recognition Letters*, vol. 125, pp. 735–741, 2019.