

# Non-Negative Factorization for Clustering of Microarray Data

L. Morgos

## Lucian Morgos

Dept. of Electronics and Telecommunications  
Faculty of Electrical Engineering and Information Technology  
University of Oradea  
Romania, 419987 Oradea, Universitatii, 1  
lmorgos@uoradea.ro

**Abstract:** Typically, gene expression data are formed by thousands of genes associated to tens or hundreds of samples. Gene expression data comprise relevant (discriminant) information as well as irrelevant information often interpreted as noise. The irrelevant information usually affects the efficiency of discovering and grouping meaningful latent information correlated to biological significance, process closely related to data clustering. Class discovery through clustering may help in identifying latent features that reflect molecular signatures, ultimately leading to class forming. One solution for improving the class discovery efficiency is provided by data dimensionality reduction, where data is decomposed into lower dimensional factors, so that those factors approximate original data.

**Keywords:** computational intelligence, microarray data analysis, clustering, recognition.

## 1 Introduction

When analyzing high dimensional data the analysis process often becomes prohibitive and computational intractable if the data is analyzed in its receiving form. Fortunately, in many cases, data comprises redundant information, irrelevant or insignificant entries which can be discarded without significantly affecting the final result or decision, process known as *dimensionality reduction*. Discovering latent and meaningful low-dimensional factors from high-dimensional data is of great importance in many research fields. Basically, dimensionality reduction may be accomplished by either feature selection or data transformation. The latter is a factorization process that decomposes data into meaningful factors so that their product approximates the initial data as good as possible. By factorization, some initial information is lost; however, the factorization of gene expression should be performed so that the factors provide a biologically meaningful decomposition of the data, which is a major goal for this particular task. The cluster statistics should match the known sample labels when statistics are available from the full data.

Let us consider our data containing  $m$  samples of  $n$  - dimensional vectors stored into a  $n \times m$  matrix  $\mathbf{V}$ , where the vector elements indexed as  $i = 1 \dots n$  correspond to the receiving information values. Dimensionality reduction refers to transforming the  $n$  - dimensional data into a lower  $r$  - dimensional data. This further relates to decomposing the matrix  $\mathbf{V}$  into factors  $\mathbf{W}$  of dimension  $n \times r$  and  $\mathbf{H}$  of dimension  $r \times m$ , where  $r < (n, m)$  is often named the factorization rank, so that their product approximates  $\mathbf{V}$ , i.e.  $\mathbf{WH} \cong \tilde{\mathbf{V}} \approx \mathbf{V}$ . For analyzing the receiving information the  $n$  - dimensional space is replaced by the lower  $r$  - dimensional space where the new information resides. Typically,  $\mathbf{W}$  represents the *basis* factors and  $\mathbf{H}$  the *encoding* (also know as the decomposition *coefficients*). Thus, each  $j = 1 \dots m$  information embedded into an  $n$  - dimensional column vector  $\mathbf{v}_j$  can be represented as linear (or nonlinear) combination of the basis matrix  $\mathbf{W}$  and the corresponding low dimensional coefficients vector  $h_j$ , more precisely,  $\mathbf{v}_j = \mathbf{W}h_j$ . When the factorization process is applied to gene expression data from a set of

microarray experiments, we typically deal with gene expression profile for a single gene across many samples or experimental conditions. More precisely, the matrix  $\mathbf{V}$  comprises in its rows the expression levels of genes, while the columns correspond to samples. Usually,  $n$  is of order of thousands while  $m$  is often less than 100. In this context NMF seeks for finding a small number of metagenes. Therefore, each column of  $\mathbf{W}$  represents a metagene and each column of  $\mathbf{H}$  denotes the expression profile.

## 2 Approach

One of the frequently used dimensionality reduction approaches is given by singular values decomposition (SVD). SVD has been proposed [1] to reduce the gene expression space where the columns of  $\mathbf{W}$  form orthonormal basis called eigenarrays, while  $\mathbf{H}$  forms eigengenes. Sorting the genes according to the descending values of eigengenes gives a global picture of the dynamics of the gene expression, in which individual genes and arrays appear to be clustered into groups of similar functions or cellular states. A robust SVD variant [2] has been further proposed to cope with outliers or missing values often found in gene expression. Other approaches refer to novel methods for analyzing SVD for obtaining gene groups along with a confidence measure [3], or SVD based biclustering through incorporating simultaneous normalization of rows and columns [4].

A relatively recent approach for data factorization is provided by nonnegative matrix factorization, where both factors are constrained to have nonnegative entries. The work of Lee and Seung published in *Nature* [5] can be considered as a starting point from where the interest in this approach almost exploded due to the algorithm's simplicity and its compliance with physiological principles. Ever since, new NMF variants have been developed in the attempt to improve its convergence, numerical stability, processing speed or to extend the approach to temporal domain. Original NMF has been applied to gene clustering [6] and, recently a sparse NMF approach has been developed [7], that appears to yield better clustering performance compared to the standard NMF.

Although the original NMF has been successfully applied to gene expression clustering, its application was rather limited to three data sets. We have extended the previous work of [6] by employing three more expression data sets in order to validate the efficiency of this decomposition approach. In addition, two more NMF variants were applied for a systematic comparison. The extension also includes the use of  $k$ -means clustering strategy along with the expression intensity provided by the rows of  $\mathbf{H}$ .

Moreover, the suitability of NMF approaches was investigated for the recognition task. Finally, we rank the algorithms according to their clustering and recognition performance as well as with respect to their stability, as defined later.

## 3 Methods

### 3.1 Non-negative matrix factorization

In the NMF formulation the difference between the initial data matrix and the product of its decomposition factors can be expressed in different ways leading to various cost functions  $f_{NMF}$  which quantify the decomposition quality. One frequently cost function is described by the Kullback-Leibler ( $KL$ ) divergence based cost function:

$$f_{NMF}(\mathbf{v}, \mathbf{w}, \mathbf{h}) \triangleq \sum_{i,j} \left( v_{ij} \ln \frac{v_{ij}}{\sum_l w_{il} h_{lj}} + \sum_l w_{il} h_{lj} - v_{ij} \right), \quad (1)$$

for any  $i = 1 \dots n$ ,  $j = 1 \dots m$ , and  $l = 1 \dots r$ .

One solution for minimizing the above cost functions is given by adopting an Expectation - Maximization (EM) like strategy [5] for iteratively updating the factors. Starting with random values for  $\mathbf{W}$  and  $\mathbf{H}$ , at each iteration, the following updating rules are used to guarantee the minimization (towards a local minimum) of the *KL*-based cost function:

$$h_{lj} \leftarrow h_{lj} \frac{\sum_i w_{li} \frac{v_{ij}}{\sum_i w_{li} h_{lj}}}{\sum_i w_{il}} \quad (2)$$

$$w_{il} \leftarrow w_{il} \frac{\sum_j h_{jl} \frac{v_{ij}}{\sum_k w_{ik} h_{lj}}}{\sum_j h_{lj}} \quad (3)$$

### 3.2 Local non-negative matrix factorization

To enhance the decomposition sparseness, Li et al [8] have developed the Local Non-negative Matrix Factorization (LNMF) algorithm, imposing more constraints to the KL cost function to get more localized basis factors. The associated cost function is then given by:

$$h_{lj} \leftarrow \sqrt{h_{lj} \frac{\sum_i v_{ij} \frac{w_{il}}{\sum_l w_{ik} h_{lj}}}{\sum_i v_{ij} \frac{w_{il}}{\sum_l w_{ik} h_{lj}}}} \quad (4)$$

$$w_{il} \leftarrow \frac{w_{il} \sum_j v_{ij} \frac{h_{lj}}{\sum_i w_{li} h_{lj}}}{\sum_j h_{lj}} \quad (5)$$

However, we should note that, here, although sparseness issue refers to the basis factors, this approach does not guarantee a high degree of sparseness for the expression profiles. As we shall see in the experimental section, the coefficient sparseness indeed seems to positively correlate with the accuracy performance.

### 3.3 Polynomial non-negative matrix factorization

Unlike NMF, the polynomial NMF [9] [10] was developed as alternative to the standard NMF for decomposing data in some nonlinear fashion. PNMF assumes that the input data  $\mathbf{V} \in \mathcal{V} \subseteq \mathbb{R}^{n \times m}$  are transformed to a higher dimensional space  $\mathcal{F} \subseteq \mathbb{R}^{l \times m}$ ,  $l \gg n$ . By denoting the set of the transformed input data with  $\mathbf{F} = [\phi(\mathbf{v}_1), \phi(\mathbf{v}_2), \dots, \phi(\mathbf{v}_m)]$ , with the  $l$  - dimensional vector expressed as  $\phi(\mathbf{v}_j) = [\phi(\mathbf{v})_1, \phi(\mathbf{v})_2, \dots, \phi(\mathbf{v})_s, \dots, \phi(\mathbf{v})_l]^T \in \mathcal{F}$ , a matrix  $\mathbf{Y} = [\phi(\mathbf{w}_1), \phi(\mathbf{w}_2), \dots, \phi(\mathbf{w}_r)]$ ,  $\mathbf{Y} \in \mathcal{F}$ , that approximates the transformed data set, such that  $r < n$ , can be found. Therefore, each vector  $\phi(\mathbf{v})$  can be written as a linear combination as  $\phi(\mathbf{v}) \approx \mathbf{Y}\mathbf{h}$ . Defining a kernel function  $\kappa$  that satisfies  $\kappa(\mathbf{v}\mathbf{w}) \triangleq \kappa(\mathbf{v}, \mathbf{w}) = \langle \phi(\mathbf{v}), \phi(\mathbf{w}) \rangle$ , for all  $\mathbf{v}, \mathbf{w} \in \mathcal{V}$ , where  $\phi$  is a mapping from  $\mathcal{V}$  to an (inner product) feature space  $\mathcal{F}$ ,  $\phi : \mathbf{v} \rightarrow \phi(\mathbf{v}) \in \mathcal{F}$ , the following squared Euclidean distance based cost function was proposed [9]:

$$f_{PNMF}(\mathbf{v}, \mathbf{w}, \mathbf{h}) = \frac{1}{2} \|\phi(\mathbf{v}_j) - \mathbf{Y}\mathbf{h}_j\|^2 \quad (6)$$

$$= \mathbf{k}(\mathbf{v}, \mathbf{v}) - 2\mathbf{k}(\mathbf{v}, \mathbf{w}_l)\mathbf{h} + \mathbf{h}^T \mathbf{k}(\mathbf{w}_o, \mathbf{w}_l)\mathbf{h},$$

The cost function  $f_{PNMF}$  should be minimized subject to  $h_l, w_{il} \geq 0$ , and  $\sum_{i=1}^n w_{il} = 1$ ,  $l = 1 \dots r$ .

## 4 Results

The experiments were conducted with respect to the following aspects: (i) *clustering error* defined as the number of misclustered samples expressed in percentage, and (ii) *model stability* defined through the standard deviation of the clustering error for several NMF runs, and (iii) *recognition error* defined as the percent of misclassified samples from the test set.

### 4.1 Data set description

We have applied NMF, LNM, and PNMF to six publicly available microarray data sets: *leukemia* [11], *medulloblastoma* [12], *colon* [13], *DLBCL* [14], *SRBCT* [15], and *prostate* [16] data set, respectively. The first two data sets are exactly the ones used by [6]. We should also note that the original *DLBCL* data set initially contained 7126 genes. By setting an intensity thresholds at 20 - 16000 units, then filtering out genes with  $\max/\min \leq 3$  or  $(\max - \min) \leq 100$ , the final number of genes was 6285. The filtering procedure was similar to the one described in [17]. For the *prostate*, the conditions for gene selection was a threshold at 100 - 16000 units, with  $\max/\min \leq 5$  or  $(\max - \min) \leq 50$ , conducting from a total of initial 12600 to 5966 genes.

### 4.2 Clustering

Starting from random values for both factors  $\mathbf{W}$  and  $\mathbf{H}$ , NMF iteratively modifies them until the algorithm converges to a local minimum. More precisely, the running stops when the difference between  $\mathbf{V}$  and the product of its decomposition factors reaches a minimum imposed threshold. Due to its stochastic nature, the algorithm does not always converge to the same solution with different initial random values for the factors. Therefore, NMF should run several times. In our experiments we ran NMF 30 times, and the minimum (*Min.*) and average (*Av.*) clustering error along with its standard deviation (*Std.*) are reported. Standard deviation is a good indicator for the model stability. A small value reveals better stability, i.e. the algorithm leads to approximately the same performance for multiple runs.

Once NMF converged and found its final decomposition factors, it is the  $\mathbf{H}$  that contains clustering information. There are two approaches to assign sample labels. One approach suggested by [6] refers to the label assignment according only to the relative values in each column of  $\mathbf{H}$ , i.e. the class labels are determined by the index of the highest metagene expression. When the number of rows of  $\mathbf{H}$  equals the number of known clusters we count the number of maximum values for each column of  $\mathbf{H}$  with respect to the known class range. Clustering errors are simply the misassignments, i.e. the mismatch between the index and associated value. Table 1 depicts the clustering errors, where the table column denoted by  $H$  indicates the first clustering criteria. It may happen that this procedure does not lead to the correct number of clustering for a particular run. Even worse, whenever this approach fails to discover the correct number of clusters for any of those 30 runs, we report this case as *N/A* in the table.

The first clustering approach only applies when the number of rows of  $\mathbf{H}$  (viewed as dimension of  $\mathbf{H}$ , or the decomposition *rank*)  $r$  equals the number of known clusters. The second clustering approach deals with varying ranks. The NMF algorithms were run for  $r = \{2, 3, 4, 5, 6, 7, 8\}$ . Each  $r$ -dimensional column vector of  $\mathbf{H}$  was clustered with the help of  $K$ -means clustering strategy. We have also carried out experiments with SVD, as baseline. The SVD method always fails to discover the correct number of clusters for the first clustering approach for all data sets. To emphasize the importance of dimensionality reduction task, experiments with the initial data dimension were also conducted and shown in Table 1 corresponding to the row termed "No processing". Indeed, the high dimension technically halves the performance of the  $K$ -means clustering.

Table 1: Clustering results for the NMF methods. Clustering results for SVD as dimensionality reduction approach is also tabulated. For comparison purpose, the clustering errors are also shown when K means is applied for the initial data dimension. The first two lowest errors are emphasized in bold.

Method	Approach	Clust. Err. (%)	Data					
			Leukemia	Medulloblastoma	Colon	DLBCL	SRBCT	Prostate
No processing	K means	Min.	13.15	14.70	20.96	28.57	34.93	36.27
		Min.	<b>5.26</b>	11.76	<b>11.29</b>	<b>18.18</b>	N/A	<b>18.62</b>
		Av.	23.97	34.11	29.26	27.27	N/A	38.67
SVD	K means	Std.	17.42	9.61	9.27	6.94	N/A	10.86
		Min.	2.63	38.23	25.80	28.57	N/A	N/A
		Av.	4.94	39.64	26.25	34.41	N/A	N/A
NMF	H	Std.	1.15	1.49	0.73	4.76	N/A	N/A
		Min.	<b>0</b>	<b>5.88</b>	<b>11.29</b>	<b>12.98</b>	<b>21.68</b>	24.50
		Av.	9.29	8.08	25.00	23.11	26.80	40.68
	K means	Std.	10.75	6.23	9.96	5.60	8.65	6.69
		Min.	10.52	38.23	32.25	29.87	N/A	N/A
		Av.	17.57	38.23	34.38	32.20	N/A	N/A
LNMF	H	Std.	2.24	0	2.45	0.99	N/A	N/A
		Min.	2.63	35.29	46.77	27.27	33.73	N/A
		Av.	12.65	35.29	46.77	29.71	50.96	N/A
	K means	Std.	9.90	0	0	2.06	4.50	N/A
		Min.	2.63	41.17	33.87	28.57	N/A	N/A
		Av.	2.63	41.17	33.87	28.57	N/A	N/A
PNMF	H	Std.	0	0	0	0	N/A	N/A
		Min.	<b>0</b>	11.76	14.51	23.37	32.53	<b>16.66</b>
		Av.	15.41	29.41	21.58	30.99	43.01	33.23
	K means	Std.	9.05	9.03	4.72	6.00	5.90	7.98

As far as the clustering ability is concerned, NMF seems to yield the lowest clustering error for all data sets but *prostate*. Comparing the two clustering approaches, *K* - means clearly outperforms the first approach for all data sets and all the decomposition methods. Moreover, except the *prostate* case for LNMF, this clustering approach always finds the correct number of classes for several NMF runs. For the *leukemia* data set, *K* - means conducted to perfect clustering (zero clustering error) for both NMF and LNMF. Brunet et al. [6] as well as other authors, including Kim and Park [7] reported that two ALL samples were consistently misclassified by most methods, including NMF. The same tendency has been noticed by us when applying the first clustering strategy but not with the second one. However, we should stress that this data set is probably an easy one compared to the others, in terms of method's clustering ability.

It is worth mentioning that, amongst all data sets, the *prostate* data seems to be the most difficult one to be correctly clustered. This indicates hard overlaps within samples from both tumor and non-tumor classes. Surprisingly, although PNMf performs modest on the other data sets, in this case it had conducted to the lowest clustering error. This behaviors could be due to the nonlinear nature of the PNMf algorithm that enables to find discriminant non-linear biological structures for this data set. For the PNMf algorithm the best result corresponds to the polynomial degree of 2.

Analyzing the model stability, LNMF followed by the PNMf algorithm are the most stable models and tend to make the same errors for various numbers of algorithm runs, as indicated by low or even zero standard deviation value. However, the LNMF was outperformed by all other methods.

The stability of the algorithm was also measured through the cophenetic correlation coefficient [6] computed for various decomposition ranks and displayed in Figure 1. All algorithms tend to be instable as rank increases. However, the behavior differs from one data set to another.

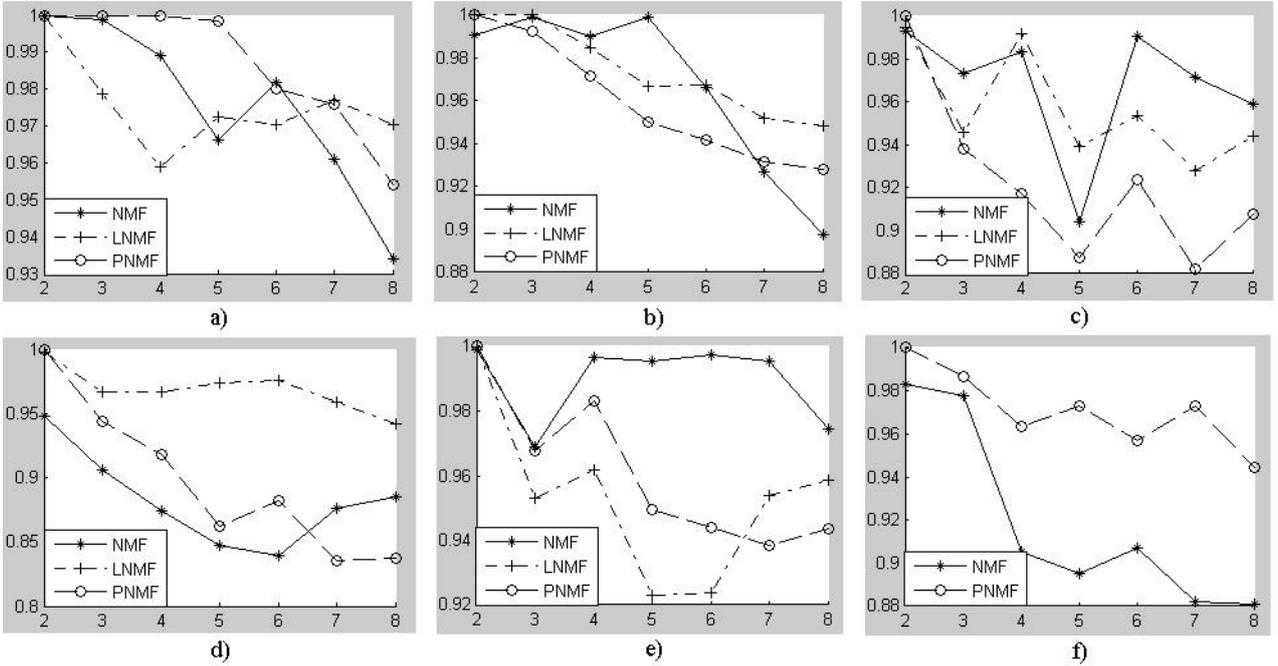


Figure 1: The cophenetic correlation coefficient versus decomposition rank for the *leukemia*, *medulloblastoma*, *colon*, *DLBCL*, *SRBCT*, and *prostate* data set, respectively.

### 4.3 Sample recognition

The second major experimental task involves the recognition of samples previously misclassified by the clustering  $K$ -means clustering approach. The correctly assigned samples were all gathered to form a training set  $\mathbf{V}_{tr}$ , while a test set  $\mathbf{V}_{te}$  comprises all wrongly clustered samples associated to the minimum clustering error. The NMF methods were only applied to the training set. Thus, there is a big difference between clustering and recognition tasks in terms of training. While in the latter case the NMF uses all samples to extract relevant information, for the recognition task, NMF employs only a subset of this statistics, more precisely, those associated to the correctly clustered samples. Let us denote the final decomposition factors corresponding to the training set by  $\mathbf{W}_{tr}$  and  $\mathbf{H}_{tr}$ , respectively. To form a training feature vector, the training data were projected into the inverse of  $\mathbf{W}_{tr}$ , i.e.,  $\mathbf{F}_{tr} = \mathbf{W}_{tr}^{-1} * \mathbf{V}_{tr}$ . Similarly, the test feature vectors are formed as  $\mathbf{F}_{te} = \mathbf{W}_{tr}^{-1} * \mathbf{V}_{te}$ . We should note that, unlike the clustering procedure where the clustering results depend on  $\mathbf{H}$ , the sample recognition is solely affected by  $\mathbf{W}$ . By projecting the data into the inverse of  $\mathbf{W}$ ,  $\mathbf{F}_{tr}$  comprises mixed positive and negative values. To keep the non-negativity constraints we have also ran experiments by projecting the data into the  $\mathbf{W}$ . However, the classification results were much lower in this case.

A simple maximum correlation classifier (MCC) relying on the minimum Euclidean distance between vectors was chosen as distance measurement. The distance from any  $\mathbf{h}_{te}$  to any  $\mathbf{h}_{tr}$  is expressed as  $\|\mathbf{h}_{te} - \mathbf{h}_{tr}\|^2 = -2g_c(\mathbf{h}_{te} + \mathbf{h}_{tr})^T \mathbf{h}_{te}$ , where  $g_c(\mathbf{h}_{te}) = \mathbf{h}_{tr}^T \mathbf{h}_{te} - 1/2\|\mathbf{h}_{tr}\|^2$  is a linear discriminant function of  $\mathbf{h}_{te}$ . By computing  $Q$  linear discriminant functions, a test sample is assigned to the label according to the

$$MCC = \operatorname{argmax}_c \{g_c(\mathbf{h}_{te})\} \quad (7)$$

In other words, the label of the training feature vector corresponding to the maximum correlation is associated to the test feature vector (test sample). The recognition results are tabulated

Table 2: Sample recognition results with MCC.

Method	Data					
	Leukemia	Medulloblastoma	Colon	DLBCL	SRBCT	Prostate
NMF	-	50	57.14	40	<b>55.55</b>	<b>28</b>
LNMF	0	66.66	86.20	71.42	64.28	-
PNMF	-	<b>25</b>	<b>44.44</b>	<b>33.33</b>	66.66	47.05

in Table 2. A recognition error of 100% indicates no improvement over the misclustered samples. LNMF wrongly clustered only one *leukemia* sample that have been finally correctly recognized by the recognition procedure. Four misclustered *medulloblastoma* samples are provided by the PNMf, while three of them were also further correctly classified by the recognition procedure. An impressive improvement was obtained in the case of NMF for the *prostate* data set. While the clustering strategy led to 25 mislabelled samples out of 102, the recognition procedure had further reduced those samples to 7.

## 5 Conclusions

Microarray analysis typically involves tens of samples where each sample is formed of thousands of values (genes). This recasts into a high-dimensional matrix. The analysis of such matrix is associated in the computer science community with the small size problem negatively affecting both the clustering and recognition accuracy. Wisely reducing data dimension is one solution to this issue. This paper addressed data dimensionality reduction applied to clustering and recognition tasks by three non-negative decomposition variants, NMF, LNM, and PNMf. The novelty of this work resides in the following: (a) we have carried out a systematic comparison in terms of clustering of these NMF algorithms involving six gene expression data sets publicly available; (b) we have shown that, unlike the clustering approach where the sample assignment depends only on the relative (maximum) values in each column of  $\mathbf{H}$ , strategy that failed to discover the correct class for some certain methods and data sets, by employing a  $K$  - means clustering approach the correct class discovery is always guaranteed for sufficient number of runs.

Concluding, the experiments indicate superior performance for the standard NMF over the other two, leading to the lowest clustering errors. The standard NMF was closely followed by SVD and PNMf. The nonlinear NMF approach (PNMF) outperforms NMF for only one data set, i.e. *prostate*.

## Bibliography

- [1] Alter, O. et al.; Singular value decomposition for genome-wide expression data processing and modeling, *Proc. Nat. Acad. Sci. USA*, 97: 10101-10106, 2000.
- [2] Liu L. et al; Robust singular value decomposition of microarray data, *Proc. Nat. Acad. Sci. USA*, 100: 13167-13172, 2003.
- [3] Wall, M.E. et al; SVDMAN singular value decomposition analysis of microarray data, *Bioinformatics*, 17: 566-568, 2001.
- [4] Kluger, Y. et al; Spectral biclustering of microarray data: Co-clustering genes and conditions, *Genome Research*, 13: 703-716, 2003.

- 
- [5] Lee, D.D.; and Seung, H.S.; Learning the parts of the objects by non-negative matrix factorization, *Nature*, 401: 788-791, 1999.
- [6] Brunet, J.P. et al; Metagenes and molecular pattern discovery using matrix factorization, *Proc. Nat. Acad. Sci. USA*, 101: 4164-4169, 2004.
- [7] Kim, H.; and Park, H. (2007); Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis, *Bioinformatics*, 23: 1495-1502, 2007.
- [8] Li, S.Z. et al; Learning spatially localized, parts-based representation, *Proc. Int. Conf. Computer Vision and Pattern Recognition*, 207-212, 2001.
- [9] Buciu, I. et al; Non-negative matrix factorization in polynomial feature space, *IEEE Trans. on Neural Networks*, 19: 1090-1100, 2008.
- [10] Buciu, I., Non-negative Matrix Factorization, A New Tool for Feature Extraction: Theory and Applications, *Int J Comput Commun*, ISSN 1841-9836, Vol. 3, Supplement: Suppl. S, 3(S): 67-74, 2008.
- [11] Golub, T.R. et al; Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring, *Science*, 286:531-537, 1999.
- [12] Pomeroy, S. L. et al; Prediction of central nervous system embryonal tumour outcome based on gene expression, *Nature*, 415:436-442, 2002.
- [13] Alon, U. et al; Broad patterns of gene expression revealed by clustering of tumor and normal colon tissues probed by oligonucleotide arrays, *Proc. Nat. Acad. Sci. USA*, 96: 6745-6750, 1999.
- [14] Shipp, M.A. et al; Diffuse large B-cell lymphoma outcome prediction by gene expression profiling and supervised machine learning, *Nature Medicine*, 8: 68-74, 2002.
- [15] Khan, J. et al; Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks, *Nature Medicine*, 7: 673-679, 2001.
- [16] Singh, D. et al; Gene expression correlates of clinical prostate cancer behavior, *Cancer Cell*, 1: 203-209, 2002.
- [17] Yang, K. et al; A stable gene selection in microarray data analysis, *BMC Bioinformatics*, 7: 228, 2006.