Journal of Bioinformatics and Computational Biology

Probabilistic Non-negative Matrix Factorization: Theory and Application to Microarray Data Analysis

Belhassen Bayar and Nidhal Bouaynaya

Department of Electrical and Computer Engineering, Rowan University,
201 Mullica Hill Road, Glassboro, New Jersey 08028
bayarb3@students.rowan.edu, bouaynaya@rowan.edu

Roman Shterenberg

Department of Mathematics, University of Alabama at Birmingham,
1300 University Blvd., Birmingham, AL 35294
shterenb@math.uab.edu

Non-negative matrix factorization (NMF) has proven to be a useful decomposition for multivariate data, where the non-negativity constraint is necessary to have a meaningful physical interpretation. NMF reduces the dimensionality of non-negative data by decomposing it into two smaller non-negative factors with physical interpretation for class discovery. The NMF algorithm, however, assumes a deterministic framework. In particular, the effect of the data noise on the stability of the factorization and the convergence of the algorithm are unknown. Collected data, on the other hand, is stochastic in nature due to measurement noise and sometimes inherent variability in the physical process. This paper presents new theoretical and applied developments to the problem of non-negative matrix factorization. First, we generalize the deterministic NMF algorithm to include a general class of update rules that converges towards an optimal non-negative factorization. Second, we extend the NMF framework to the probabilistic case (PNMF). We show that the Maximum A Posteriori estimate of the non-negative factors is the solution to a weighted regularized non-negative matrix factorization problem. We subsequently derive update rules that converge towards an optimal solution. Third, we apply the PNMF to cluster and classify DNA microarrays data. The proposed PNMF is shown to outperform the deterministic NMF and the sparse NMF algorithms in clustering stability and classification accuracy.

Keywords: Matrix Decomposition; Clustering Gene Expression Data; Tumors Classification.

1. Introduction

Extracting knowledge from experimental raw data and measurements is an important objective and challenge in signal processing. Often data collected is high dimensional and incorporates several inter-related variables, which are combinations of underlying latent components or factors. Approximate low-rank matrix factorizations play a fundamental role in extracting these latent components [9]. In many applications, signals to be analyzed are non-negative, e.g., pixel values in image

processing, price variables in economics and gene expression levels in computational biology. For such data, it is imperative to take the non-negativity constraint into account in order to obtain a meaningful physical interpretation. Classical decomposition tools, such as Principal Component Analysis (PCA), Singular Value Decomposition (SVD), Blind Source Separation (BSS) and related methods do not guarantee to maintain the non-negativity constraint. Non-negative matrix factorization (NMF) represents non-negative data in terms of lower-rank non-negative factors. NMF proved to be a powerful tool in many applications in biomedical data processing and analysis, such as muscle identification in the nervous system [31], classification of images [1], gene expression classification [7], biological process identification [20] and transcriptional regulatory network inference [24]. The appeal of NMF, compared to other clustering and classification methods, stems from the fact that it does not impose any prior structure or knowledge on the data. Brunet *et al.* successfully applied NMF to the classification of gene expression datasets [7] and showed that it leads to more accurate and more robust clustering than the Self-Organizing Maps (SOMs) and Hierarchical Clustering (HC). Analytically, the NMF method factors the original non-negative matrix V into two lower rank non-negative matrices, W and H such that $V = WH + E$, where E is the residual error. Lee and Seung [21] derived algorithms for estimating the optimal non-negative factors that minimize the Euclidean distance and the Kullback-Leibler divergence cost functions. Their algorithms, guaranteed to converge, are based on multiplicative update rules, and are a good compromise between speed and ease of implementation. In particular, the Euclidean distance NMF algorithm can be shown to reduce to the gradient descent algorithm for a specific choice of the step size [21]. Lee and Seung's NMF factorization algorithms have been widely adopted by the community [7,4,14,35].

The NMF method is, however, deterministic. That is, the algorithm does not take into account the measurement or observation noise in the data. On the other hand, data collected using electronic or biomedical devices, such as gene expression profiles, are known to be inherently noisy and therefore, must be processed and analyzed by systems that take into account the stochastic nature of the data. Furthermore, the effect of the data noise on the NMF method in terms of convergence and robustness has not been previously investigated. Thus, questions about the efficiency and robustness of the method in dealing with imperfect or noisy data are still unanswered.

In this paper, we extend the NMF framework and algorithms to the stochastic case, where the data is assumed to be drawn from a multinomial probability density function. We call the new framework Probabilistic NMF or PNMF. We show that the PNMF formulation reduces to a weighted regularized matrix factorization problem. We generalize and extend Lee and Seung's algorithm to the stochastic case; thus providing PNMF updates rules, which are guaranteed to converge to the optimal solution. The proposed PNMF algorithm is applied to cluster and classify gene expression datasets, and is compared to other NMF and non-NMF approaches including sparse NMF (SNMF) and SVM.

The paper is organized as follows: In Section 1.1, we discuss related work and clarify the similarities and differences between the proposed PNMF algorithm and other approaches to NMF present in the literature. In Section 2, we review the (deterministic) NMF formulation and extend Lee and Seung's NMF algorithm to include a general class of convergent update rules. In Section 3, we introduce the probabilistic NMF (PNMF) framework and derive its corresponding update rules. In Section 4, we present a data classification method based on the PNMF algorithm. Section 5 applies the proposed PNMF algorithm to cluster and classify gene expression profiles. The results are compared with the deterministic NMF, sparse NMF and SVM. Finally, a summary of the main contributions of the paper and concluding remarks are outlined in Section 6.

In this paper, scalars are denoted by lower case letters, e.g., n, m; vectors are denoted by bold lower case letters, e.g., $\boldsymbol{x}, \boldsymbol{y}$; and matrices are referred to by upper case letters, e.g., A, V. $\boldsymbol{x}_i$ denotes the i^{th} element of vector $\boldsymbol{x}$ and A_{ij} is the $(i,j)^{th}$ entry of matrix A. Throughout the paper, we provide references to known results and limit the presentation of proofs to new contributions. All proofs are presented in the Appendix section.

1.1. *Related work*

Several variants of the NMF algorithm have been proposed in the literature. An early form of NMF, called Probabilistic Latent Semantic Analysis (PLSA) [16], [17], [23], was used to cluster textual documents. The key idea is to map high-dimensional count vectors, such as the ones arising in text documents, to a lower dimensional representation in a so-called *latent semantic space*. PLSA has been shown to be equivalent to NMF factorization with Kullback-Leibler (KL) divergence, in the sense that they have the same objective function and any solution of PLSA is a solution of NMF with KL minimization [12].

Many variants of the NMF framework introduce additional constraints on the non-negative factor matrices W and H, such as sparsity and smoothness. Combining sparsity with non-negative matrix factorization is partly motivated by modeling neural information processing, where the goal is to find a decomposition in which the hidden components are sparse. Hoyer [18] combined sparse coding and non-negative matrix factorization into *non-negative sparse coding* (NNSC) to control the trade-off between sparseness and accuracy of the factorization. The sparsity constraint is imposed by constraining the l_1-norm. The NNSC algorithm resorts to setting the negative values of one of the factor matrices to zero. This procedure is not always guaranteed to converge to a stationary point. Kim and Park [19] solved the sparse NMF optimization problem via alternating non-negativity-constrained least squares. They applied sparse NMF to cancer class discovery and gene expression data analysis.

NMF has also been extended to consider a class of smoothness constraints on the optimization problem [25]. Enforcing smoothness on the factor matrices is desirable

in applications such as unmixing spectral reflectance data for space object identification and classification purposes [25]. However, the algorithm in [25] forces positive entries by setting negative values to zero and hence may suffer from convergence issues. Similarly, different penalty terms may be used depending upon the desired effects on the factorization. A unified model of constrained NMF, called versatile sparse matrix factorization (VSMF), has been proposed in [22]. The VSMF framework includes both l_1 and l_2-norms. The l_1-norm is used to induce sparsity and the l_2-norm is used to obtain smooth results. In particular, the standard NMF, sparse NMF [18], [19] and semi-NMF [11], where the non-negativity constraint is imposed on only one of the factors, can be seen as special cases of VSMF.

Another variant of the NMF framework is obtained by considering different distances or measures between the original data matrix and its non-negative factors [28], [33]. Sandler and Lindenbaum [28] proposed to factorize the data using the earth movers distance (EMD). The EMD NMF algorithm finds the local minimum by solving a sequence of linear programming problems. Though the algorithm has shown significant improvement in some applications, such as texture classification and face recognition, it is computationally very costly. To address this concern, the authors have proposed the wavelet-based approximation to the EMD distance, WEMD, and used it in place of EMD. They argued that the local minima of EMD and WEMD are generally collocated when using a gradient-based method. A similarity measure based on the correntropy, termed NMF MCC, has been proposed in [33]. The correntropy measure employs the Gaussian kernel to map the linear data space to a non-linear space. The optimization problem is solved using an expectation maximization based approach.

A collection of non-negative matrix factorization algorithms implemented for Matlab is available at `http://cogsys.imm.dtu.dk/toolbox/nmf/`. Except for PLSA, which was originally proposed as a statistical technique for text clustering, the presented NMF approaches do not explicitly assume a stochastic framework for the data. In other words, the data is assumed to be deterministic. In this work, we assume that the original data is a sample drawn from a multinomial distribution and derive the maximum a posteriori (MAP) estimates of the non-negative factors. The proposed NMF framework, termed Probabilistic NMF or PNMF, does not impose any additional constraints on the non-negative factors like SNMF or VSMF. Interestingly, however, the formulation of the MAP estimates reduces to a weighted regularized matrix factorization problem that resembles the formulations in constrained NMF approaches. The weighting parameters, however, have a different interpretation: they refer to signal to noise ratios rather than specific constraints.

2. Non-negative Matrix Factorization

The non-negative matrix factorization (NMF) is a constrained matrix factorization problem, where a non-negative matrix V is factorized into two non-negative

matrices W and H. Here, non-negativity refers to elementwise non-negativity, i.e., all elements of the factors W and H must be equal to or greater than zero. The non-negativity constraint makes NMF more difficult algorithmically than classical matrix factorization techniques, such as principal component analysis and singular value decomposition. Mathematically, the problem is formulated as follows: Given a non-negative matrix $V \in \mathbb{R}^{n\times m}$, find non-negative matrices $W \in \mathbb{R}^{n\times k}$ and $H \in \mathbb{R}^{k\times m}$ such that $V \approx WH$. The optimal factors minimize the squared error and are solutions to the following constrained optimization problem,

$$(W^*, H^*) = \arg\min_{W,H\geq 0} f(W,H) = \|V - WH\|_F^2, \tag{1}$$

where $\|.\|_F$ denotes the Frobenius norm and f is the squared Euclidean distance function between V and WH. The cost function f is convex with respect to either the elements of W or H, but not both. Alternating minimization of such a cost leads to the ALS (Alternating Least squares) algorithm [15], [32], [2], which can be described as follows:

(1) Initialize W randomly or by using any a priori knowledge.
(2) Estimate H as $H = (W^TW)^-W^TV$ with fixed W.
(3) Set all negative elements of H to zero or some small positive value.
(4) estimate W as $W = VH^T(HH^T)^-$ with fixed H.
(5) Set all negative elements of W to zero or some small positive value.

In this algorithm, A^- denotes the Moore-Penrose inverse of A. The ALS algorithm has been used extensively in the literature [15], [32], [2]. However, it is not guaranteed to converge to a global minimum nor even a stationary point. Moreover, it is often not sufficiently accurate, and it can be slow when the factor matrices are ill-conditioned or when the columns of these matrices are co-linear. Furthermore, the complexity of the ALS algorithm can be high for large-scale problems as it involves inverting a large matrix. Lee and Seung [21] proposed a multiplicative update rule, which is proven to converge to a stationary point, and does not suffer from the ALS drawbacks. In what follows, we present Lee and Seung's multiplicative update rule as a special case of a class of update rules, which converge towards a stationary point of the NMF problem.

Proposition 1. *The function $f(W,H) = \|V - WH\|_F^2$ is non-increasing under the update rules*

$$\begin{cases} \boldsymbol{h}^{k+1} = \boldsymbol{h}^k - K_h^{-1}(W^TW\boldsymbol{h}^k - W^T\boldsymbol{v}) \\ \\ \tilde{\boldsymbol{w}}^{k+1} = \tilde{\boldsymbol{w}}^k - K_w^{-1}(HH^T\tilde{\boldsymbol{w}}^k - H\tilde{\boldsymbol{v}}) \end{cases} \tag{2}$$

where $\tilde{\boldsymbol{w}}$ and $\tilde{\boldsymbol{v}}$ are the columns of W^T and V^T, respectively, and K_h and K_w satisfy the following conditions

a. *K_h and K_w are diagonal matrices with (strictly) positive elements for all vectors $\boldsymbol{h}$ and $\tilde{\boldsymbol{w}}$.*
b. *$K_h \boldsymbol{h}^k \geq W^T W \boldsymbol{h}^k$ and $K_w \tilde{\boldsymbol{w}}^k \geq H H^T \tilde{\boldsymbol{w}}^k$ where the inequality is elementwise.*
c. *The matrices $K_h - W^T W$ and $K_w - H H^T$ are positive semi-definite (p.s.d) for all $\boldsymbol{h}$ and $\tilde{\boldsymbol{w}}$.*

The function f is invariant under these update rules if and only if W and H are at a stationary point.

The following corollary presents a special choice of the matrices K_h and K_w, which leads to Lee and Seung's multiplicative rule for the NMF problem.

Corollary 1. *In Proposition 1, chose K_h and K_w as follows:*

$$(K_h)_{ij} = \delta_{ij} (W^T W \boldsymbol{h}^k)_i / \boldsymbol{h}_i^k, \tag{3}$$

$$(K_w)_{ij} = \delta_{ij} (H H^T \tilde{\boldsymbol{w}}^k)_i / \tilde{\boldsymbol{w}}_i^k, \tag{4}$$

Where $\boldsymbol{h}_i^k, \tilde{\boldsymbol{w}}_i^k$ are the i^{th} entries of the vectors $\boldsymbol{h}^k$ and $\tilde{\boldsymbol{w}}^k$, respectively, and δ_{ij} is the kronecker function, i.e., $\delta_{ij} = \begin{cases} 1, \text{ if } i = j \\ 0, \text{ otherwise.} \end{cases}$ This choice leads to the following update rule:

$$\begin{cases} H_{ij} \longleftarrow H_{ij} \frac{(W^T V)_{ij}}{(W^T W H)_{ij}} \\ \\ W_{ij} \longleftarrow W_{ij} \frac{(V H^T)_{ij}}{(W H H^T)_{ij}} \end{cases} \tag{5}$$

The function f is invariant under these updates if and only if W and H are at a stationary point.

Corollary 1 corresponds to the update rules proposed by Lee and Seung [21]. Proposition 1 presents a general class of update rules, which converge to a stationary point of the NMF problem. From the proof of the Proposition (detailed in the Appendix), it will be clear that conditions [a], [b] and [c] in Proposition 1 are only sufficient conditions for the update rules to converge towards a stationary point. That is, there may exist K_h and K_w that do not satisfy these conditions but that lead to update rules that converge towards a stationary point. The particular choice of K_h and K_w in Corollary 1 corresponds to the fastest convergent update rule among all matrices satisfying conditions [a]-[c] in Proposition 1. Observe also that since the data matrix V is non-negative, the update rule in (5) leads to non-negative factors W and H as long as the initial values of the algorithm are chosen to be non-negative.

3. Probabilistic Non-negative Matrix Factorization

3.1. *The PNMF framework*

In this section, we assume that the data, represented by the non-negative matrix V, is corrupted by additive white Gaussian noise. Then, the data follows the following

conditional distribution,

$$p(V \mid W, H, \sigma^2) = \prod_{i=1}^{N} \prod_{j=1}^{M} [\mathcal{N}(V_{ij} \mid \mathbf{u}_i^T \mathbf{h}_j, \sigma^2)], \tag{6}$$

where $\mathcal{N}(.|\mu, \sigma^2)$ is the probability density function of the Gaussian distribution with mean μ and standard deviation σ, $\mathbf{u}_i$ and $\mathbf{h}_j$ denote, respectively, the i^{th} column of the matrix $U = W^T$ (or the i^{th} row of W) and the j^{th} column of the matrix H. Zero mean Gaussian priors are imposed on $\mathbf{u}_i$ and $\mathbf{h}_j$ to control the model parameters. Specifically, we have

$$p(U \mid \sigma_W^2) = \prod_{i=1}^{N} \mathcal{N}(\mathbf{u}_i \mid 0, \sigma_W^2 I) = p(W \mid \sigma_W^2). \tag{7}$$

$$p(H \mid \sigma_H^2) = \prod_{i=1}^{M} \mathcal{N}(\mathbf{h}_j \mid 0, \sigma_H^2 I). \tag{8}$$

We estimate the factor matrices W and H using the maximum a posteriori (MAP) criterion. The logarithm of the posterior distribution is given by

$$\ln(p(W, H \mid V, \sigma^2, \sigma_H^2, \sigma_W^2)) = -\frac{1}{2\sigma^2} \sum_{i=1}^{N} \sum_{j=1}^{M} (V_{ij} - \mathbf{u}_i^T \mathbf{h}_j)^2$$

$$-\frac{1}{2\sigma_W^2} \sum_{i=1}^{N} \|\mathbf{u}_i\|^2 - \frac{1}{2\sigma_H^2} \sum_{j=1}^{M} \|\mathbf{h}_j\|^2 + C, \tag{9}$$

where C is a constant term depending only on the standard deviations σ, σ_W and σ_H. Maximizing (9) is equivalent to minimizing the following function

$$(W^*, H^*) = \underset{W,H \geq 0}{\arg\min} \|V - WH\|_F^2 + \lambda_W \|W\|_F^2 + \lambda_H \|H\|_F^2, \tag{10}$$

where $\lambda_W = \frac{\sigma^2}{\sigma_W^2}$ and $\lambda_H = \frac{\sigma^2}{\sigma_H^2}$. Observe that the PNMF formulation in (10) corresponds to a weighted regularized matrix factorization problem. Moreover, the PNMF reduces to the NMF for $\sigma = 0$. The following proposition provides the update rules for the PNMF constrained optimization problem.

Proposition 2. *The function*

$$f(W, H) = \|V - WH\|_F^2 + \alpha \|W\|_F^2 + \beta \|H\|_F^2 \tag{11}$$

is non-increasing under the update rules

$$\begin{cases} H_{ij} \longleftarrow H_{ij} \frac{(W^T V)_{ij}}{(W^T W H + \beta H)_{ij}} \\ W_{ij} \longleftarrow W_{ij} \frac{(V H^T)_{ij}}{(W H H^T + \alpha W)_{ij}} \end{cases} \tag{12}$$

The function f is invariant under these updates if and only if W and H are at a stationary point.

Observe that, since the data matrix V is non-negative, the update rules in (12) lead to non-negative factors W and H as long as the initial values of the algorithm are chosen to be non-negative.

4. PNMF-based Data Classification

In this section, we show how the PNMF output can be used to extract relevant features from the data for classification purposes. The main idea relies on the fact that metasamples extracted from the PNMF factorization contain the inherent structural information of the original data in the training set. Thus, each sample in a test set can be written as a sparse linear combination of the metasamples extracted from the training set. The classification task then reduces to computing the representation coefficients for each test sample based on a chosen discriminating function. The sparse representation approach has been shown to lead to more accurate and robust results [36]. The sparsity constraint is imposed through an l_1-regularization term [36]. Thus, a test sample may be represented in terms of few metasamples.

4.1. *Sparse Representation Approach*

We divide the data, represented by the $n \times m$ matrix V, into training and testing sets, where the number of classes k is assumed to be known. In Section 5, we describe a method to estimate the number of classes based on the PNMF clustering technique. The training data is ordered into a matrix A with n rows of genes and r columns of training samples with $r < m$. Thus, A is a sub-matrix of V used to recognize any new presented sample from the testing set. We arrange the matrix A in such a way to group samples which belong to the same class in the same sub-matrix A_i where $(1 \leq i \leq k)$. Then A can be written as $A = [A_1, A_2, ..., A_k]$ and each matrix A_i is a concatenation of r_i columns of the i^{th} class $A_i = [\boldsymbol{c}_{i,1}, \boldsymbol{c}_{i,2}, ..., \boldsymbol{c}_{i,r_i}]$

A test sample $\boldsymbol{y} \in \mathbb{R}^n$ that belongs to the i^{th} class can be written as the following linear combination of the A_i columns,

$$\boldsymbol{y} = \alpha_{i,1}\boldsymbol{c}_{i,1} + \alpha_{i,2}\boldsymbol{c}_{i,2} + ... + \alpha_{i,r_i}\boldsymbol{c}_{i,r_i}, \tag{13}$$

for some scalars $\alpha_{i,q} \in \mathbb{R}$, $1 \leq q \leq r_i$.
Equation (13) can be re-written as

$$\boldsymbol{y} = A\boldsymbol{x}, \tag{14}$$

where

$$\boldsymbol{x} = [0, ...0, \alpha_{i,1}, \alpha_{i,2}, ..., \alpha_{i,r_i}, 0..., 0]^T \in \mathbb{R}^r, \tag{15}$$

is the coefficient vector of the testing sample $\boldsymbol{y}$. $\boldsymbol{x}$ is a r_i-sparse vector whose nonzero entries are associated with the columns of the sub-matrix A_i, hence the name *sparse representation*. Therefore, predicting the class of test sample $\boldsymbol{y}$ reduces to estimating the vector $\boldsymbol{x}$ in Eq. (14).

We propose to find the sparsest least-squares estimate of the coefficient $\boldsymbol{x}$ as the solution to the following regularized least-squares problem [34]

$$\hat{\boldsymbol{x}} = \min_{\boldsymbol{x}} \{\|A\boldsymbol{x} - \boldsymbol{y}\|_2 + \lambda\|\boldsymbol{x}\|_1\}, \tag{16}$$

where $\|\boldsymbol{x}\|_1$ denotes the l_1-norm of vector $\boldsymbol{x}$, i.e., $\|\boldsymbol{x}\|_1 = \sum_i |\boldsymbol{x}_i|$, and λ is a positive scalar used to control the tradeoff between the sparsity of $\boldsymbol{x}$ and the accuracy of the reconstruction error. Donoho *et al.* showed that the l_1-norm approximates the l_0-norm, which counts the number of non-zero entries in a vector [13]. The l_0-norm problem, however, is NP hard, whereas the l_1-norm is convex. The optimization problem in (16) is therefore convex; thus, it admits a global solution, which can be efficiently computed using convex optimization solvers [10]. Actually, one can show that (16) is a Second-Order Cone Programming (SOCP) problem [6].

4.2. *PNMF-based classification*

The classifier's features are given by the metasamples computed by the PNMF algorithm. We first compute the PNMF factorization of each sub-matrix A_i as

$$A_i \sim W_i \times H_i, \tag{17}$$

where W_i and H_i are respectively $n \times k_i$ and $k_i \times r_i$ non-negative matrices. k_i refers to the number of metasamples needed to describe and summarize the i^{th} class. The value of k_i is experimentally determined and depends on the number of training samples r_i in each class and the total number of classes k. We subsequently concatenate all the W_i matrices to form the matrix $W = [W_1, W_2, .., W_k]$. Observe that the matrix W contains the metasamples of the entire training set. Therefore, a test sample $\boldsymbol{y}$ that belongs to the i^{th} class should approximately lie in the space spanned by the W_i columns.

The classification problem in (16) can therefore be re-written as

$$\hat{\boldsymbol{x}} = \min_{\boldsymbol{x}} \{\|W\boldsymbol{x} - \boldsymbol{y}\|_2 + \lambda\|\boldsymbol{x}\|_1\}, \tag{18}$$

Which can be easily solved using a SOCP solver [6].

PNMF-based classification algorithm The PNMF-based classification algorithm is summarized below.

Input: Gene expression data $V \in \mathbb{R}^{n\times m}$. It is assumed that V contains at least r labeled samples, which can be used in the learning or training process.

Step 1 Select the training samples $A \in \mathbb{R}^{n\times r}$ and the testing sample $\boldsymbol{y} \in \mathbb{R}^n$ from the original data V such that $\boldsymbol{y}$ is not a column of A.

Step 2 Reorder the training matrix $A = [A_1, A_2, ..., A_k]$ for k classes.

Step 3 Compute the matrix of features $W_i \in \mathbb{R}^{n \times k_i}$ from each sub matrix $A_i \in \mathbb{R}^{n \times r_i}$ Using the PNMF algorithm, $i = 1 : k$

Step 4 Solve the optimization problem in (18) for $W = [W_1, W_2, ..., W_k]$ using, for instance, the *cvx* environment in MATLAB. Let the solution $\boldsymbol{x} = [\boldsymbol{x}_1^T, \cdots, \boldsymbol{x}_k^T]^T$, where $\boldsymbol{x}_i \in \mathbb{R}^{k_i \times 1}$.

Step 5 Compute the residuals $e_i(\boldsymbol{y}) = \|\boldsymbol{y} - W\delta_i(\boldsymbol{x})\|_2$, $i = 1 : k$, where $\delta_i(\boldsymbol{x}) = [0, \cdots, 0, \boldsymbol{x}_i^T, 0, \cdots 0]^T$.

Step 6 Associate class($\boldsymbol{y}$)=$\arg\min_i e_i(\boldsymbol{y})$

---

5. Application to Gene Microarrays

We apply and compare the proposed PNMF-based clustering and classification algorithms with its homologue NMF-based clustering [7] and classification as well as the sparse-NMF classification method presented in [36]. We first describe the gene expression dataset used and present the clustering procedure.

5.1. *Data sets description*

One of the important challenges in DNA microarrays analysis is to group genes and experiments/samples according to their similarity in gene expression patterns. Microarrays simultaneously measure the expression levels of thousands of genes in a genome. The microarray data can be represented by a gene-expression matrix $V \in \mathbb{R}^{n \times m}$, where n is the number of genes and m is the number of samples that may represent distinct tissues, experiments, or time points. The m^{th} column of V represents the expression levels of all the genes in the m^{th} sample.

We consider seven different microarray data sets: leukemia [7], medulloblastoma [7], prostate [29], colon [3], breast-colon [8], lung [5] and brain [27]. The leukemia data set is considered a benchmark in cancer clustering and classification [7]. The distinction between acute myelogenous leukemia (AML) and acute lymphoblastic leukemia (ALL), as well as the division of ALL into T and B cell subtypes, is well known [7]. We consider an ALL-AML dataset, which contains 5000 genes and 38 bone marrow samples (tissues from different patients for the considered genes) [7]. The considered leukemia dataset contains 19 ALL-B, 8 ALL-T and 11 AML samples.

The medulloblastoma data set is a collection of 34 childhood brain tumors samples from different patients. Each patient is represented by 5893 genes. The pathogenesis of these brain tumors is not well understood. However, two known histological subclasses can be easily differentiated under the microscope, namely, classic (C) and desmoplastic (D) medulloblastoma tumors [7]. The medulloblastoma dataset contains 25 C and 9 D childhood brain tumors.

The prostate data [29] contains the gene expression patterns from 52 prostate

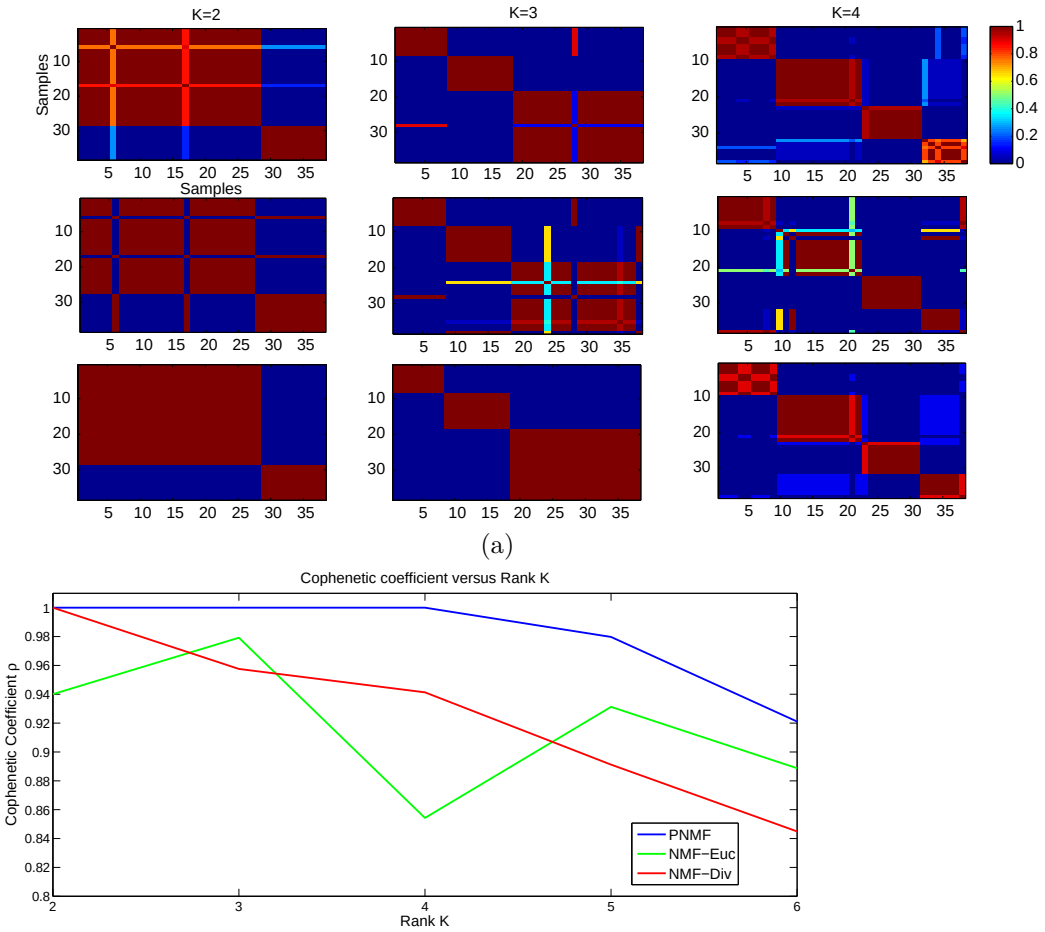


Fig. 1. Clustering results for the Leukemia dataset: (a) Consensus matrices: Top row NMF-Euc, Second row NMF-Div, bottom row: PNMF; (b) Cophenetic coefficient versus the rank k (NMF-Euc in green, NMF-Div in red and PNMF in blue).

tumors (PR) and 50 normal prostate specimens (N), which could be used to predict common clinical and pathological phenotypes relevant to the treatment of men diagnosed with this disease. The prostate dataset contains 102 samples across 339 genes.

The colondataset [3] is obtained from 40 tumors and 22 normal colon tissue samples across 2000 genes. The breast and colon data [8] contains tissues from 62 lymph node-negative breast tumors (B) and 42 Dukes' B colon tumors (C). The lung tumor data [5] contains 17 normal lung tissues (NL), 139 adenocarcinoma (AD), 6 small-cell

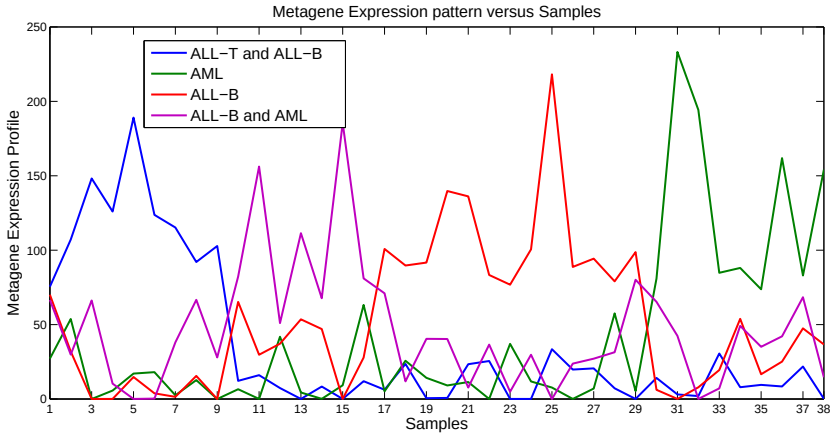


Fig. 2. Metagenes expression patterns versus the samples for $k = 4$ in the Leukemia dataset.

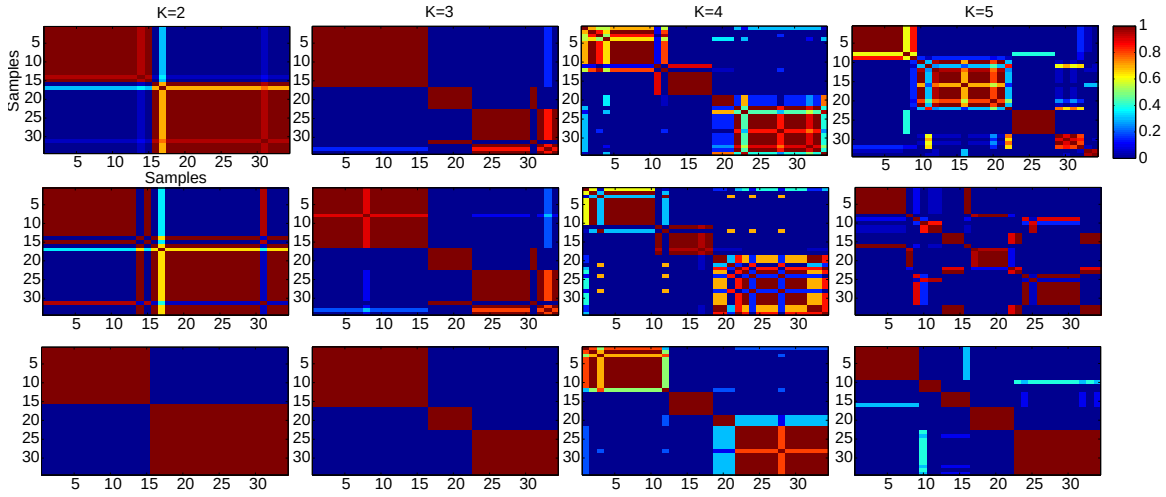
lung cancer (SCLC), 20 pulmonary carcinoids (COID) and 21 squamous cell lung carcinomas (SQ) samples across 12600 genes. The brain data [27] is the collection of embryonal tumors of the central nervous system. This data includes 10 medulloblastomas (MD), 10 malignant gliomas (Mglio), 10 atypical teratoid/rhabdoid tumors (Rhab), 4 normal tissues (Ncer) and 8 primitive neuroectodermal tumors (PNET). The brain samples are measured across 1379 genes.

5.2. *Gene expression data clustering*

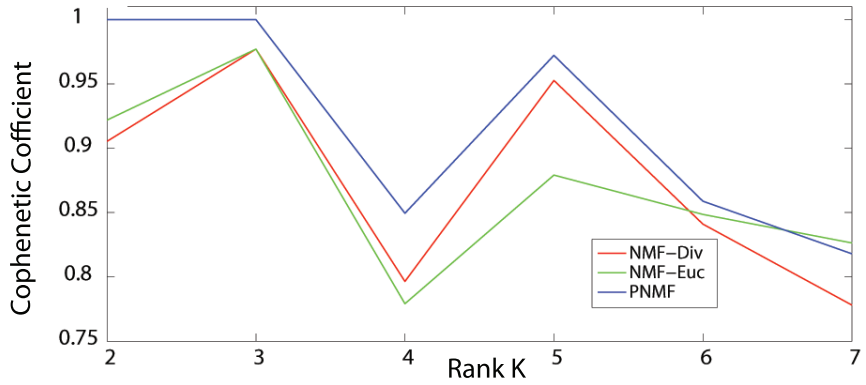
Applying the NMF framework to data obtained from gene expression profiles allows the grouping of genes as *metagenes* that capture latent structures in the observed data and provide significant insight into underlying biological processes and the mechanisms of disease. Typically, there are a few metagenes in the observed data that may monitor several thousands of genes. Thus, the redundancy in this application is very high, which is very profitable for NMF [9]. Assuming gene profiles can be grouped into j metagenes, V can be factored with NMF into the product of two non-negative matrices $W \in \mathbb{R}^{n\times j}$ and $H \in \mathbb{R}^{j\times m}$. Each column vector of W represents a metagene. In particular, w_{ij} denotes the contribution of the i^{th} genes into the j^{th} metagene, and h_{jm} is the expression level of the j^{th} metagene in the m^{th} sample.

5.2.1. *Clustering performance evaluation*

The position of the maximum value in each column vector of H indicates the index of the cluster to which the sample is assigned. Thus, there are j clusters of the samples. The stability of the clustering is tested by the so-called *connectivity matrix*



(a)



(b)

Fig. 3. Clustering results for the Medulloblastoma dataset: (a) Consensus matrices: Top row NMF-Euc, Second row NMF-Div, bottom row: PNMF; (b) Cophenetic coefficient versus the rank k (NMF-Euc in green, NMF-Div in red and PNMF in blue).

$C \in \mathbb{R}^{m \times m}$ [7], which is a binary matrix defined as $c_{ij} = 1$ if samples i and j belong to the same cluster, and $c_{ij} = 0$ otherwise. The connectivity matrix from each run of NMF is reordered to form a block diagonal matrix. After performing several runs, a *consensus matrix* is calculated by averaging all the connectivity matrices. The entries of the consensus matrix range between 0 and 1, and they can be interpreted as the probability that samples i and j belong to the same cluster. Moreover, if the entries of the consensus matrix were arranged so that samples belonging to the same cluster are adjacent to each other, perfect consensus matrix would translate into a block-diagonal matrix with non-overlapping blocks of 1's along the diagonal,

each block corresponding to a different cluster [7]. Thus, using the consensus matrix, we could cluster the samples and also assess the performance of the number of clusters k. A quantitative measure to evaluate the stability of the clustering associated with a cluster number k was proposed in [19]. The measure is based on the correlation coefficient of the consensus matrix, ρ_k, also called the cophenetic correlation coefficient. This coefficient measures how faithfully the consensus matrix represents the similarities and dissimilarities among observations. Analytically, we have $\rho_k = \frac{1}{m^2}\sum_{ij} 4(c_{ij} - \frac{1}{2})^2$ [19]. Observe that $0 \leq \rho_k \leq 1$, and a perfect consensus matrix (all entries equal to 0 or 1) would have $\rho_k = 1$. The optimal value of k is obtained when the magnitude of the cophenetic correlation coefficient starts declining.

5.2.2. *Clustering results*

Brunet *et al.* [7] showed that the (deterministic) NMF based on the divergence cost function performs better than the NMF based on the Euclidean cost function. The divergence cost function is defined as

$$(W^*, H^*) = \underset{W,H\geq 0}{\arg\min}\, g(W,H) = \sum_{i,j}(V_{ij}\log(\frac{V_{ij}}{(WH)_{ij}}) - V_{ij} + (WH)_{ij}) \tag{19}$$

The update rules for the divergence function are given by [21]

$$\begin{cases} H_{ij} \longleftarrow H_{ij}\frac{\sum_k (W_{ki}V_{kj})/(WH)_{kj}}{\sum_r W_{ri}} \\ \\ W_{ij} \longleftarrow W_{ij}\frac{\sum_k (H_{jk}V_{ik})/(WH)_{ik}}{\sum_r H_{jr}} \end{cases} \tag{20}$$

In this section, we compare the PNMF algorithm in (12) with both the Euclidean-based NMF in (5) and the divergence-based NMF in (19). We propose to cluster the leukemia and the medulloblastoma sample sets because the biological subclasses of these two datasets are known, and hence we can compare the performance of the algorithms with the ground truth. Figure 1(a) shows the consensus matrices corresponding to $k = 2, 3, 4$ clusters for the leukemia dataset. In this figure, the matrices are mapped using the gradient color so that dark blue corresponds to 0 and red to 1. We can observe the consensus matrix property that the samples' classes are laid in block-diagonal along the matrix. It is clear from this figure that the PNMF performs better than the NMF algorithm, in terms of samples' clustering. Specifically, the clusters, as identified by the PNMF algorithm, are better defined and the consensus matrices' entries are not overlapping and hence well clustered. In particular, PNMF with rank $k = 2$ correctly recovered the ALL-AML biological distinction with higher accuracy than the deterministic NMFs (based on the Euclidean and divergence costs). Consistent clusters are also observed for rank $k = 3$, which reveal further portioning of the samples when the ALL samples are classified as the B or T subclasses. In particular, the nested structure of the blocks

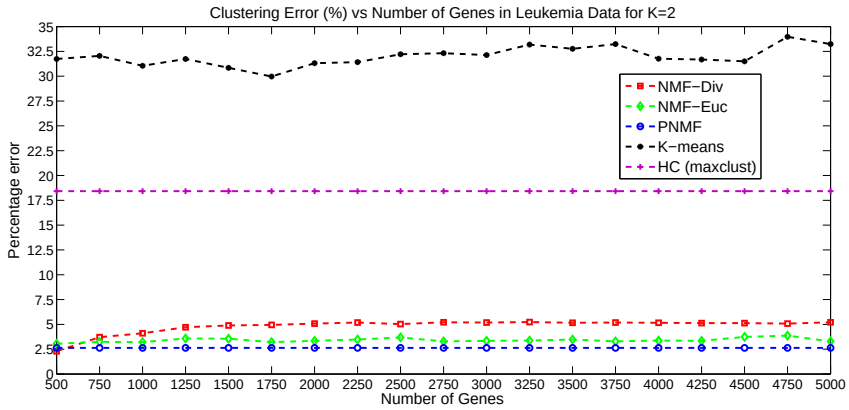


Fig. 4. Clustering Percentage Error versus *Nbr.* of genes (NMF-Euc in green, NMF-Div in red and PNMF in blue, K-means in black and Hierarchical Clustering in purple) in Leukemia dataset for $k = 2$.

for $k = 3$ corresponds to the known subdivision of the ALL samples into the T and B classes. Nested and partially overlapped clusters can be interpreted with the NMF approaches. Nested clusters reflect local properties of expression patterns, and overlapping is due to global properties of multiple biological processes (selected genes can participate in many processes) [9]. An increase in the number of clusters beyond 3 ($k = 4$) results in stronger dispersion in the consensus matrix. However, Fig. 1(b) shows that the value of the PNMF cophenetic correlation for rank 4 is equal to 1, whereas it drops sharply for both the Euclidean and divergence-based NMF algorithms. The Hierarchal Clustering (HC) method is also able to identify four clusters [7]. These clusters can be interpreted as subdividing the samples into sub-clusters that form separate patterns within the whole set of samples as follows: {(11 ALL-B), (7 ALL-B and 1 AML), (8 ALL-T and 1 ALL-B), (10 AML)}.

Figure 2 depicts the metagenes expression profiles (rows of H) versus the samples for the PNMF algorithm. We can visually recognize the different four patterns that PNMF and HC are able to identify.

Figure 3 shows the consensus matrices and the cophenetic coefficients of the medulloblastoma dataset for $k = 2, 3, 4, 5$. The NMF and PNMF algorithms are able to identify the two known histological subclasses: classic and desmoplastic. They also predict the existence of classes for $k = 3, 5$. This clustering also stands out because of the high values of the cophenetic coefficient for $k = 3, 5$ and the steep drop off for $k = 4, 6$. The sample assignments for $k = 2, 3$ and 5 display a nesting of putative medulloblastoma classes, similar to that seen in the leukemia dataset. From Fig. 3, we can see that the PNMF clustering is more robust, with respect to the consensus matrix and the cophenetic coefficient, than the NMF clustering. Furthermore, Brunet *et al.* [7] stated that the divergence-based NMF is able

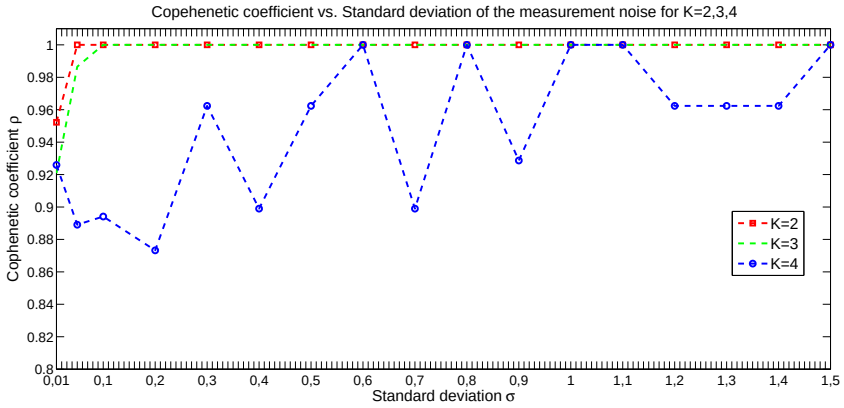


Fig. 5. The cophenetic coefficient versus the standard deviation of the measurement noise for $k = 2$ (red), 3 (green) and 4 (blue) in the Leukemia dataset.

to recognize subtypes that the Euclidian version cannot identify. We also reach a similar conclusion as shown in Fig. 3 for $k = 3, 5$, where the Euclidian-based NMF factorization shows scattering from these structures. However, the PNMF clustering performs even better than the divergence-based NMF as shown in Figs. 3(a) and 3(b).

To confirm our results we compare our proposed PNMF algorithm with the standard NMF algorithms, distance criterion-based Hierarchical Clustering (HC) and K-means. We plot in figure 4 the curve Error vs. Number of genes in the labeled Leukemia data set. We select genes with small profile variance using the Bioinformatics toolbox in MATLAB from 500 to 5000 genes and the experimental points are equally spaced. We run 100 Monte Carlo simulation then we take the average of the error. Our simulation results show that PNMF outperforms other clustering approaches.

5.2.3. *Robustness evaluation*

In this subsection, we assess the performance of the PNMF algorithm with respect to the model parameters, especially the choice of the noise power. Recall that, in the probabilistic model, σ measures the uncertainty in the data or the noise power in the gene expression measurements. We set the prior standard deviations $\sigma_W = \sigma_H = 0.01$, and compute the cophenetic coefficient for varying values of σ between 0.01 and 1.5. Figure 5 shows the cophenetic coefficient versus the standard deviation σ in the leukemia data set for ranks $k = 2, 3, 4$. We observe that the PNMF is stable to a choice of σ between 0.05 and 1.5 for the ranks $k = 2$ and 3, which correspond to biologically relevant classes. In particular, when σ tends to zero, the PNMF algorithm reduces to the classic NMF, which explains the drop in

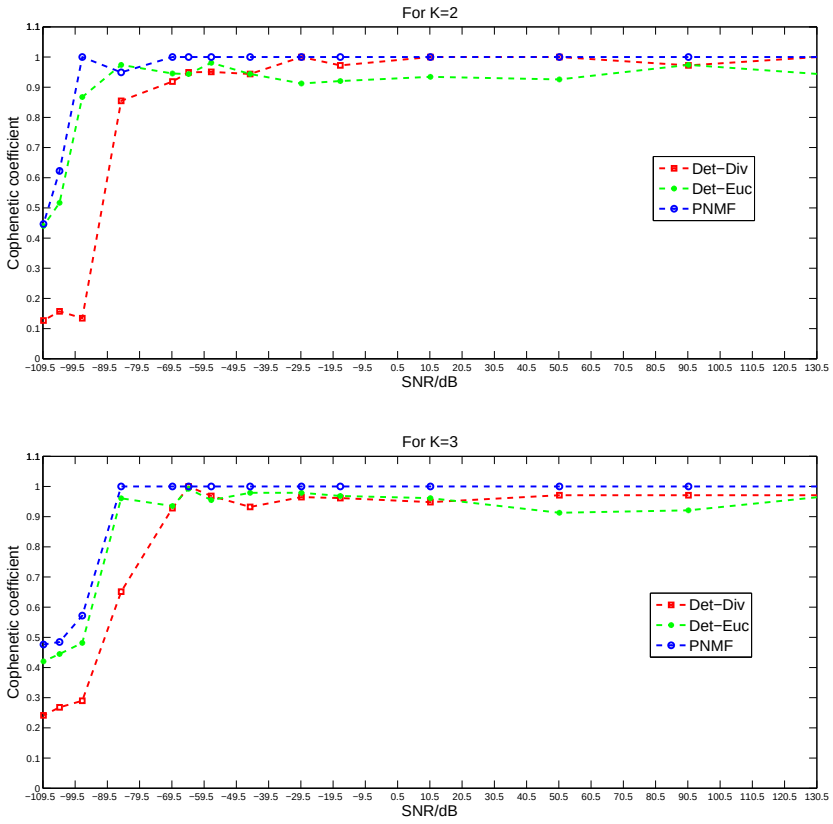


Fig. 6. Cophenetic versus SNR in dB (NMF-Euc in green, NMF-Div in red and PNMF in blue) in Leukemia dataset for $k = 2$ and $k = 3$.

the cophenetic coefficient for values of σ near zero.

We next study the robustness of the NMF and the proposed PNMF algorithms to the presence of noise in the data. To this end, we add white Gaussian noise, with varying power, to the leukemia dataset according to the following formula,

$$V_{noisy} = V + \sigma_n R, \tag{21}$$

where σ_n is the standard deviation of the noise, and R is a random matrix of the same size as the data matrix V, and whose entries are normally distributed with zero mean and unity variance. The signal to noise ratio (SNR) is, therefore, given by SNR $= \frac{P_V}{\sigma_n^2}$, where the signal power $P_V = \frac{1}{nm}\sum_i \sum_j v_{ij}^2 = \frac{1}{nm}\|V\|_F^2$. Since the cophenetic coefficient measures the stability of the clustering, we plot in Figures 6 and 7 the cophenetic coefficient versus the SNR, measured in dB, for both the Euclidean-based and divergence-based NMFs and PNMF algorithms using the leukemia and medulloblastoma data sets. We observe that the PNMF

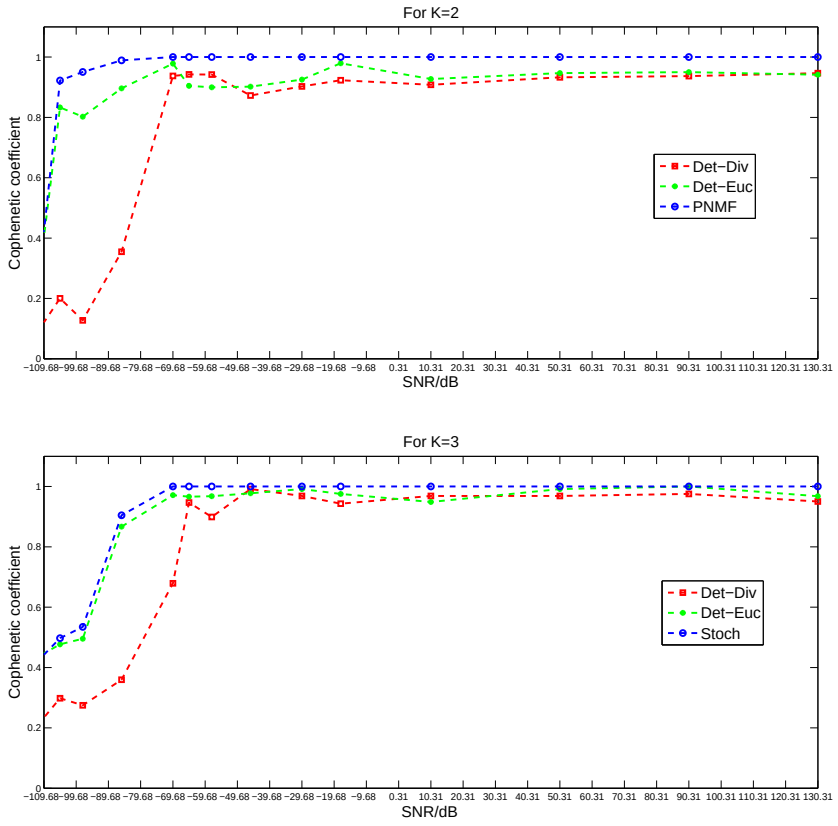


Fig. 7. Cophenetic versus SNR in dB (NMF-Euc in green, NMF-Div in red and PNMF in blue) in Medulloblastoma dataset for $k = 2$ and $k = 3$.

algorithm leads to more robust clustering than the deterministic NMF algorithms for all SNR values. Table 1 shows the minimum SNR values for which the cophenetic coefficient takes values higher or equal than 0.9. We say that the algorithm is "stable" for SNR values higher or equal than the minimum SNR. For the leukemia data, the Euclidean-based NMF and the divergence-based NMF algorithms stabilize respectively at $SNR = -93.5$ and $SNR = -73.5$ dB for $k = 2$, whereas the PNMF algorithm is stable at lower SNR values, $SNR = -99.5$ dB for $k = 2$. Similar results are obtained for the medulloblastoma dataset, where the NMF algorithms stabilize respectively as above at $SNR = -84.68$ and $SNR = -70.68$ dB, whereas the PNMF is stable at $SNR = -104.68$ dB. Thus, the PNMF algorithm is more stable than its deterministic homologue. Also, observe that the Euclidian-based NMF performs better than its divergence homologue for noisy data.

Table 1. Smallest SNR value for which the cophenetic coefficient is higher or equal than 0.9.

Datasets	$k = 2$			$k = 3$		
	NMF-Euc	NMF-Div	PNMF	NMF-Euc	NMF-Div	PNMF
Leukemia	−93.50	−73.50	**-99.50**	−87	−71	**-88.50**
Medulloblastoma	−84.68	−70.68	**-104.68**	−86	−65.50	**-86**

5.3. *NMF-based tumor classification*

Given that the proposed PNMF algorithm results in more stable clustering than its deterministic homologue, we expect that it will also lead to better feature extraction and classification. We classify the tumors in the seven gene expression datasets described in Section 5.1.

We assess the performance of the classification algorithm using the 10-fold cross-validation technique [36]. The number of metagenes k_i can be determined using the nested stratified 10-fold cross-validation. However, we follow the work in [36] and choose $k_i = 8$ if the number of samples in the i^{th} class $r_i > 8$. Otherwise we choose $k_i = r_i$. We selected the parameters α and β of PNMF in order to minimize the classification error in the training dataset based on a 10-fold cross-validation technique. The parameters of SNMF were selected using the same criterion and method, i.e. minimize the classification error in the training dataset. The classification results for the NMF, PNMF, SVM and SNMF [36] algorithms are summarized in Table 2. In particular, we compared the PNMF-based MSRC algorithm to the SVM algorithm which has been shown to outperform K-NN and neural network in tumor classification [30], [26]. In our experiment we use one-versus-rest SVM (OVRSVM) with Polynomial kernels approach which has been shown to be the best one [30]. The results can be obtained using the Gene Expression Model Selector (GEMS) publicly available online `http://www.gems-system.org/`. Observe that the PNMF-based classifier performs better than the other approaches for the considered data sets except for the prostate data where SVM achieves the highest classification accuracy. Moreover, the PNMF performs better than the SNMF for the prostate, lung and brain data sets. This is due to the high accuracy of the PNMF in feature extraction as compared to the SNMF algorithm, which is not guaranteed to converge to the optimal non-negative factorization [36].

6. Conclusion and Discussion

Studying and analyzing tumor profiles is a very relevant area in computational biology. Clinical applications include clustering and classification of gene expression profiles. In this work, we developed a new mathematical framework for clustering and classification based on the Probabilistic Non-negative Matrix Factorization

Table 2. Classification accuracy

Data sets	*Nbr.* of classes	NMF-Euc	NMF-Div	SNMF	SVM	PNMF
Prostate	2	85.29%	86.27%	88.24%	99%	**92.16%**
Medulloblastoma	2	85.29%	91.18%	94.12%	79.16%	**94.12%**
Colon	2	85.48%	88.71%	90.32%	89.04%	**90.32%**
Breast-Colon	2	98.08%	95.19%	98.08%	84.63%	**98.08%**
Leukemia	3	97.37%	97.37%	97.37%	95.50%	**97.37%**
Lung	5	92.61%	90.64%	93.60%	85.54%	**94.09%**
Brain	5	76.19%	78.57%	83.33%	77%	**85.71%**

(PNMF) method. We presented an extension of the deterministic NMF algorithm to the probabilistic case. The proposed PNMF algorithm takes into account the stochastic nature of the data due to the inherent presence of noise in the measurements as well as the internal biological variability. We subsequently casted the optimal non-negative probabilistic factorization as a weighted regularized matrix factorization problem. We derived updates rules and showed convergence towards the optimal non-negative factors. The derived update rules generalize Lee and Seung's multiplicative update rules for the NMF algorithm. We have also generalized Lee and Seung's algorithm to include a general class of update rules, which converge towards a stationary point of the (deterministic) NMF problem. We next derived a PNMF-based classifier, which relies on the PNMF factorization to extract features and classify the samples in the data. The PNMF-based clustering and classification algorithms were applied to seven microarray gene expression datasets. In particular, the PNMF-based clustering was able to identify biologically significant classes and subclasses of tumor samples in the leukemia and medulloblastoma datasets. Moreover, the PNMF clustering results were more stable and robust to data corrupted by noise than the classic (deterministic) NMF.

Thanks to its high stability, robustness to noise and convergence properties, the PNMF algorithm yielded better tumor classification results than the NMF and the Sparse NMF (SNMF) algorithms. The proposed PNMF framework and algorithm can be further applied to many other relevant applications in biomedical data processing and analysis, including muscle identification in the nervous system, image classification, and protein fold recognition.

Acknowledgment

This project is supported by Award Number R01GM096191 from the National Institute Of General Medical Sciences (NIH/NIGMS). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute Of General Medical Sciences or the National Institutes of Health.

Appendix

To prove the results in this paper, we need to define the notion of an auxiliary function.

Definition 1. $G(\boldsymbol{h}, \boldsymbol{h'})$ is an auxiliary function for $f(\boldsymbol{h})$ if $G(\boldsymbol{h}, \boldsymbol{h'}) \geq f(\boldsymbol{h})$ and $G(\boldsymbol{h}, \boldsymbol{h}) = f(\boldsymbol{h})$.

The following lemma in [21] shows the usefulness of the auxiliary function.

Lemma 1. [21] *if G is an auxiliary function, then f is nonincreasing under the update*

$$\boldsymbol{h}^{(k+1)} = \arg\min_{\boldsymbol{h}} G(\boldsymbol{h}, \boldsymbol{h}^{(k)}). \tag{A.1}$$

Proof of Proposition 1. We will prove the update rule for H. A similar reasoning would provide the update rule for W. Consider the two-variable matrix

$$\begin{aligned} G(\boldsymbol{h}, \boldsymbol{h}^{(k)}) = \; & f(\boldsymbol{h}^{(k)}) + (\boldsymbol{h} - \boldsymbol{h}^{(k)})^T \nabla f(\boldsymbol{h}^{(k)}) + \\ & \frac{1}{2}(\boldsymbol{h} - \boldsymbol{h}^{(k)})^T K_h(\boldsymbol{h}^{(k)})(\boldsymbol{h} - \boldsymbol{h}^{(k)}), \end{aligned} \tag{A.2}$$

where K_h is any function satisfying conditions [a]-[c] and $f(\boldsymbol{h}) = 1/2 \sum_i (\boldsymbol{v}_i - \sum_j W_{ij}\boldsymbol{h}_j)^2$. We show that G is an auxiliary function for f. It is straightforward to verify that $G(\boldsymbol{h}, \boldsymbol{h}) = f(\boldsymbol{h})$. We only need to show that $G(\boldsymbol{h}, \boldsymbol{h}^k) \geq f(\boldsymbol{h})$. To do this, we compare

$$\begin{aligned} f(\boldsymbol{h}) = \; & f(\boldsymbol{h}^{(k)}) + (\boldsymbol{h} - \boldsymbol{h}^{(k)})^T \nabla f(\boldsymbol{h}^{(k)}) + \\ & \frac{1}{2}(\boldsymbol{h} - \boldsymbol{h}^{(k)})^T (W^T W)(\boldsymbol{h} - \boldsymbol{h}^{(k)}) \end{aligned} \tag{A.3}$$

With Eq. (A.2) to find that $G(\boldsymbol{h}, \boldsymbol{h}^k) \geq f(\boldsymbol{h})$ is equivalent to

$$(\boldsymbol{h} - \boldsymbol{h}^{(k)})^T [K_h(\boldsymbol{h}^{(k)}) - W^T W](\boldsymbol{h} - \boldsymbol{h}^{(k)}) \geq 0, \tag{A.4}$$

From Condition [c], we have that $K_h - W^T W$ is positive semi-definite; thus, Eq. (A.4) is satisfied and $G(\boldsymbol{h}, \boldsymbol{h}^k) \geq f(\boldsymbol{h})$, proving that G is an auxiliary function of f. We next show that $\boldsymbol{h}$ is positive elementwise at every iteration k. From lemma 1, and taking the derivative of G with respect to $\boldsymbol{h}$, we obtain that

$$\begin{aligned} \boldsymbol{h}^{(k+1)} &= \boldsymbol{h}^{(k)} - K_h^{-1} \nabla f(\boldsymbol{h}^{(k)}) \\ &= \boldsymbol{h}^{(k)} - K_h^{-1}(W^T W \boldsymbol{h}^{(k)} - W^T \boldsymbol{v}) \\ &= [I - K_h^{-1} W^T W]\boldsymbol{h}^{(k)} + K_h^{-1} W^T \boldsymbol{v}, \end{aligned} \tag{A.5}$$

Let us assume that $\boldsymbol{h}^k$ is positive and show that $\boldsymbol{h}^{k+1}$ is also positive. From condition [a], K_h is diagonal and positive (elementwise). Therefore, K_h^{-1} is also diagonal and positive. Given that W and V are also positive, we have that $K_h^{-1} W^T \boldsymbol{v}$ is positive. From condition [b], we have that $[I - K_h^{-1} W^T W]\boldsymbol{h}^{(k)}$ is positive. Thus, $\boldsymbol{h}^{k+1}$ is

positive (elementwise). In particular, by choosing the initial point $\boldsymbol{h}^0$ positive, all iterations $\boldsymbol{h}^k$ are guaranteed to be positive.

This ends the proof of Proposition 1. Next, we show that Lee and Seung's choice of $(K_h)_{ij} = \delta_{ij}(W^T W \boldsymbol{h}^{(k)})_i / \boldsymbol{h}_i^{(k)}$ corresponds to the fastest convergent update rule among the class of matrices K_h that satisfy conditions [a]-[c].

From Eq. (A.5), we have

$$\begin{aligned} \|\boldsymbol{h}^{(k+1)} - \boldsymbol{h}^{(k)}\| &= \|K_h^{-1}(W^T W \boldsymbol{h}^{(k)} + W^T \boldsymbol{v})\| \\ &\leq \|K_h^{-1}\| \|W^T W \boldsymbol{h}^{(k)} + W^T \boldsymbol{v}\|. \end{aligned} \tag{A.6}$$

Thus, the smaller the norm of K_h (or the larger the norm of K_h^{-1}), the faster the convergence rate. From condition (b), we have that $K_h \boldsymbol{h}^k \geq W^T W \boldsymbol{h}^k$. Hence, the smallest choice of K_h corresponds to $(K_h)_{ij} = \delta_{ij}(W^T W \boldsymbol{h}^{(k)})_i / \boldsymbol{h}_i^{(k)}$. □

Proof of Proposition 2. The following lemma provides an auxiliary function for the objective function f in (11).

Lemma 2. *Consider the diagonal matrix*

$$\Phi_{ij}(\boldsymbol{h}^{(k)}) = \delta_{ij}(W^T W \boldsymbol{h}^{(k)})_i / \boldsymbol{h}_i^{(k)} + \beta. \tag{A.7}$$

We show that

$$\begin{aligned} G(\boldsymbol{h}, \boldsymbol{h}^{(k)}) = \; & f(\boldsymbol{h}^{(k)}) + (\boldsymbol{h} - \boldsymbol{h}^{(k)})^T \nabla f(\boldsymbol{h}^{(k)}) + \\ & \frac{1}{2}(\boldsymbol{h} - \boldsymbol{h}^{(k)})^T \Phi(\boldsymbol{h}^{(k)})(\boldsymbol{h} - \boldsymbol{h}^{(k)}) \end{aligned} \tag{A.8}$$

is an auxiliary function for $f(\boldsymbol{h}) = \sum_i (\boldsymbol{v}_i - \sum_j W_{ij}\boldsymbol{h}_j)^2 + \alpha \|W\|_F^2 + \beta \sum_i \|\boldsymbol{h}_i\|^2$.

The fact that $G(\boldsymbol{h}, \boldsymbol{h}) = f(\boldsymbol{h})$ is obvious. Therefore, we need only to show that $G(\boldsymbol{h}, \boldsymbol{h}^{(k)}) \geq f(\boldsymbol{h})$. To do this, we compare

$$\begin{aligned} f(\boldsymbol{h}) = \; & f(\boldsymbol{h}^{(k)}) + (\boldsymbol{h} - \boldsymbol{h}^{(k)})^T \nabla f(\boldsymbol{h}^{(k)}) + \\ & \frac{1}{2}(\boldsymbol{h} - \boldsymbol{h}^{(k)})^T (W^T W + \beta I)(\boldsymbol{h} - \boldsymbol{h}^{(k)}) \end{aligned} \tag{A.9}$$

with Eq. (A.8) to find that $G(\boldsymbol{h}, \boldsymbol{h}^{(k)}) \geq f(\boldsymbol{h})$ is equivalent to

$$(\boldsymbol{h} - \boldsymbol{h}^{(k)})^T [K(\boldsymbol{h}^{(k)}) - W^T W](\boldsymbol{h} - \boldsymbol{h}^{(k)}) \geq 0, \tag{A.10}$$

The proof of the semi-definiteness of the matrix in (A.10) is provided in [21]. Replacing G in Eq. (A.2) by its expression in Eq. (A.8) results in the update rule

$$\boldsymbol{h}^{(k+1)} = \boldsymbol{h}^{(k)} - \Phi(\boldsymbol{h}^{(k)})^{-1} \nabla f(\boldsymbol{h}^{(k)}). \tag{A.11}$$

Since G is an auxiliary function of f, f is non-increasing under this update rule. Writing the components of Eq. (A.11), we obtain

$$\boldsymbol{h}_i^{(k+1)} = \boldsymbol{h}_i^{(k)} \frac{(W^T \boldsymbol{v})_i}{(W^T W \boldsymbol{h}^{(k)} + \beta \boldsymbol{h}^{(k)})_i}. \tag{A.12}$$

Similarly, we can obtain the update rule for W. □

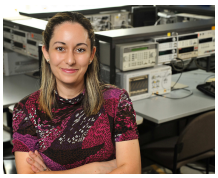
Proof of Corollary 1. Consider the diagonal matrices

$$(K_h)_{ij} = \delta_{ij}(W^T W H^k)_{ij}/H^k_{ij}. \tag{A.13}$$

$$(K_w)_{ij} = \delta_{ij}(W_k H H^T)_{ij}/W^k_{ij}. \tag{A.14}$$

It can be easily shown that K_h and K_w in Eqs. (A.13) and (A.14) satisfy conditions [a]-[c]. Corollary 1 follows directly from Proposition 1 by choosing K_h and K_w in proposition 1 as above. □

Belhassen Bayar Belhassen Bayar studied Mathematics and Physics at Institut Préparatoire aux Etudes d'ingénieurs de Tunis (IPEIT), Tunisia, in 2006-2008. He received the B.S. degree in Electrical Engineering from the Ecole Nationale d'Ingénieurs de Tunis (ENIT), Tunisia, in 2011. From 2011-2013, he was a research assistant at University of Arkansas at Little Rock (UALR), Little Rock. In 2013, he joined the Electrical and Computer Engineering Master program at Rowan University, New Jersey. Bayar won the Best Paper Award at the IEEE International Workshop on Genomic Signal Processing and Statistics in 2013. His main research interest is genomic signal processing.



Nidhal Bouaynaya Nidhal Bouaynaya received the B.S. degree in Electrical Engineering and Computer Science from the Ecole Nationale Supérieure de L'Electronique et de ses Applications (ENSEA), France, in 2002, the M.S. degree in Electrical and Computer Engineering from the Illinois Institute of Technology, Chicago, in 2002, the Diplôme d'Etudes Approfondies (DEA) in Signal and Image Processing from ENSEA, France, in 2003, the M.S. degree in Mathematics and the Ph.D. degree in Electrical and Computer Engineering from the University of Illinois at Chicago, in 2007. From 2007-2013, she was an Assistant then Associate Professor with the Department of Systems Engineering at the University of Arkansas at Little Rock. In fall 2013, she joined the Department of Electrical and Computer Engineering at Rowan University, where she is currently an Assistant Professor. Dr. Bouaynaya won the Best Student Paper Award in SPIE Visual Communication and Image Processing 2006 and the Best Paper Award at the IEEE International Workshop on Genomic Signal Processing and Statistics in 2013. She is serving as a Review Editor of Frontiers in Systems Biology. Her main research interests are in genomic signal processing, medical imaging, mathematical biology and dynamical systems.

Roman Shterenberg Roman Shterenberg received the B.S. degree in Physics in 1998, the M.S. and Ph.D. degrees in Mathematics in 2000 and 2003 from St. Petersburg State University, Russia. In 2005-2007, he was a Van Vleck Assistant Professor at the University of Wisconsin-Madison. In 2007, he joined the University of Alabama at Birmingham, where he is currently an Associate Professor in the Department of Mathematics. His research interests are in mathematical physics, spectral theory, inverse problems and mathematical biology.

References

1. Ahn JH, Kim SK, Oh JH, Choi S, Multiple nonnegative-matrix factorization of dynamic pet images, *Proceedings of Asian Conference on Computer Vision*, pp. 1009–1013, 2004.
2. Albright R, Cox J, Duling D, Langville AN, Meyer CD, *Algorithms, initializations, and convergence for the nonnegative matrix factorization*, *Tech Rep*, NCSU Technical Report Math 81706. http://meyer. math. ncsu. edu/Meyer/Abstracts/Publications. html, 2006.
3. Alon U, Barkai N, Notterman DA, Gish K, Ybarra S, Mack D, Levine AJ, Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays, *Proceedings of the National Academy of Sciences* **96**(12):6745–6750, 1999.
4. Berry M, Browne M, Langville A, Pauca P, Plemmon R, Algorithms and applications for approximate nonnegative matrix factorization, *Computational Statistics & Data Analysis* **52**(1):155–173, 2007.
5. Bhattacharjee A, Richards WG, Staunton J, Li C, Monti S, Vasa P, Ladd C, Beheshti J, Bueno R, Gillette M, Loda M, Weber G, Mark EJ, Lander ES, Wong W, Johnson BE, Golub TR, Sugarbaker DJ, Meyerson M, Classification of human lung carcinomas by mrna expression profiling reveals distinct adenocarcinoma subclasses., *Proceedings of the National Academy of Sciences* **98**(24):13790–13795, 2001.
6. Boyd SP, Vandenberghe L, *Convex optimization*, Cambridge university press, 2004.
7. Brunet JP, Tamayo P, Golub TR, PMesirov J, Metagenes and molecular pattern discovery using matrix factorization, *Proceedings of the National Academy of Sciences* **101**(12):4164–4169, 2004.
8. Chowdary D, Lathrop J, Skelton J, Curtin K, Briggs T, Zhang Y, Yu J, Wang Y, Mazumder A, Prognostic gene expression signatures can be measured in tissues collected in rnalater preservative., *The Journal of Molecular Diagnostics* **8**(1):31–39, 2006.
9. Cichocki A, Zdunek R, Phan AH, Amari SI, *Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-way Data Analysis and Blind Source Separation*, Wiley.com, 2009. ISBN 0470746661, 9780470746660.
10. CVX Research I, *CVX: Matlab Software for Disciplined Convex Programming, version 2.0*, `http://cvxr.com/cvx`, 2012.
11. Ding C, Li T, Jordan MI, Convex and semi-nonnegative matrix factorizations, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **32**:4555, 2010.
12. Dinga C, Lib T, Pengb W, On the equivalence between non-negative matrix factorization and probabilistic latent semantic indexing, *Computational Statistics and Data Analysis* **52**:39133927, 2008.

13. Donoho D, Compressed sensing, *IEEE Transactions on Information Theory* **52**:1289–1306, 2006.
14. Donoho D, Stodden V, When does non-negative matrix factorization give a correct decomposition into parts?, in *Advances in Neural Information Processing Systems 16*, eds., Thrun S, Saul L, Schölkopf B, MIT Press, Cambridge, MA, 2004.
15. Hancewicz TM, Wang JH, Discriminant image resolution: a novel multivariate image analysis method utilizing a spatial classification constraint in addition to bilinear nonnegativity, *Chemometrics and intelligent laboratory systems* **77**(1):18–31, 2005.
16. Hofmann T, Probabilistic latent semantic analysis, *Proceedings of the Annual Conference on Uncertainty in Artificial Intelligence*, Morgan Kaufmann, pp. 289–296, 1999.
17. Hofmann T, *Advances in Neural Information Processing Systems 12*. MIT Press, chap. Learning the Similarity of Documents : an information-geometric approach to document retrieval and categorization, pp. 914–920, 2000.
18. Hoyer PO, Non-negative matrix factorization with sparseness constraints, *The Journal of Machine Learning Research* **5**:1457–1469, 2004.
19. Kim H, Park H, Sparse non-negative matrix factorization via alternating non-negativity-constrained least squares for microarray data analysis, *Bioinformatics* **23**(12):1495–1502, 2007.
20. Kim PM, Tidor B, Subsystem identification through dimensionality reduction of large-scale gene expression data, *Genome research* **13**(7):1706–1718, 2003.
21. Lee DD, Seung HS, Algorithms for non-negative matrix factorization, *Proceedings of the Conference on Neural Information Processing Systems* pp. 556–562, 2001.
22. Li Y, Ngom A, Versatile sparse matrix factorization and its applications in high-dimensional biological data analysis, *Patten Recognition in Bioinformatics* **7986**:91–101, 2013.
23. Masseroli M, Chicco D, Pinoli P, Probabilistic latent semantic analysis for prediction of gene ontology annotations, *International Joint Conference on Neural Networks*, pp. 1 – 8, 2012.
24. Michael O, Elana F, Matrix factorization for transcriptional regulatory network inference, *Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), 2012 IEEE Symposium on*, IEEE, pp. 387–396, 2012.
25. Pauca VP, Piper J, Plemmons RJ, Nonnegative matrix factorization for spectral data analysis, *Linear Algebra and its Applications* **416**:29–54, 2006.
26. Pochet N, Smet FD, Suykens JAK, Moor BLRD, Systematic benchmarking of microarray data classification: assessing the role of non-linearity and dimensionality reduction, *Bioinformatics* **20**(17):3185–3195, 2004.
27. Pomeroy SL, Tamayo P, Gaasenbeek M, Sturla LM, Angelo M, McLaughlin ME, Kim JYH, Goumnerova LC, Black PM, Lau C, Allen JC, Zagzag D, Olson JM, Curran T, Wetmore C, Biegel JA, Poggio T, Mukherjee S, Rifkin R, Califano A, Stolovitzky G, Louis DN, Mesirov JP, Lander ES, Golub TR, Prediction of central nervous system embryonal tumour outcome based on gene expression., *Nature* **415**(6870):436–442, 2002.
28. Sandler R, Lindenbaum M, Nonnegative matrix factorization with earth mover's distance metric for image analysis **33**(8):1590–1602.
29. Singh D, Febbo PG, Ross K, Jackson DG, Manola J, Ladd C, Tamayo P, Renshaw AA, D'Amico AV, Richie JP, Lander ES, Loda M, Kantoff PW, Golub TR, Sellers WR, Gene expression correlates of clinical prostate cancer behavior, *Cancer Cell* **1**(2):203–209, 2002.
30. Statnikov AR, Aliferis CF, Tsamardinos I, Hardin DP, Levy S, A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer

diagnosis **21**(5):631–643.
31. Tresch MC, Cheung VCK, D'avella SAOA, Oviedo GT, Ting LH, Krouchev N, Kalaska JF, Drew T, Macpherson JM, Ivanenko YP, Lacquaniti F, Matrix factorization algorithms for the identification of muscle synergies: evaluation on simulated and experimental data sets, *Journal of Neurophysiology* **95**:2199–2212, 2006.
32. Wang JH, Hopke PK, Hancewicz TM, Zhang SL, Application of modified alternating least squares regression to spectroscopic image analysis, *Analytica Chimica Acta* , 2003.
33. Wang JJY, Wang X, Gao X, Non-negative matrix factorization by maximizing correntropy for cancer clustering, *BMC Bioinformatics* **1**(1):1590–1602, 2013.
34. Wright J, Yang AY, Ganesh A, Sastry SS, Ma Y, Robust face recognition via sparse representation, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **31**:210–227, 2009.
35. Xu W, Liu X, Gong Y, Document clustering based on non-negative matrix factorization, *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, ACM Press, pp. 267–273, 2003.
36. Zheng CH, Zhang L, Ng TY, Shiu SC, Huang DS, Metasample-based sparse representation for tumor classification, *IEEE Transactions on Computational Biology and Bioinformatics* **8**:1273–1282, 2011.