2011 IEEE International Workshop on Genomic Signal Processing and Statistics
December 4-6, 2011, San Antonio, Texas, USA

Clustering Gene Expression Data using Probabilistic Non-negative Matrix Factorization

Belhassen Bayar
Department of Electrical Engineering
Ecole Nationale d'Ingénieurs
de Tunis, Tunisia
Email: belhassen.bayar@gmail.com

Nidhal Bouaynaya
Department of Systems Engineering
University of Arkansas
at Little Rock, USA
Email: nxbouaynaya@ualr.edu

Roman Shterenberg
Department of Mathematics
University of Alabama
at Birmingham, USA
Email: shterenb@math.uab.edu

***Abstract*—Non-negative matrix factorization (NMF) has proven to be a useful decomposition for multivariate data. Specifically, NMF appears to have advantages over other clustering methods, such as hierarchical clustering, for identification of distinct molecular patterns in gene expression profiles. The NMF algorithm, however, is deterministic. In particular, it does not take into account the noisy nature of the measured genomic signals. In this paper, we extend the NMF algorithm to the probabilistic case, where the data is viewed as a stochastic process. We show that the probabilistic NMF can be viewed as a weighted regularized matrix factorization problem, and derive the corresponding update rules. Our simulation results show that the probabilistic non-negative matrix factorization (PNMF) algorithm is more accurate and more robust than its deterministic homologue in clustering cancer subtypes in a leukemia microarray dataset.**

I. Introduction

Numerous clustering methods have been proposed in the literature to identify distinct molecular patterns and extract relevant biological information from high-throughput gene expression data [1], [2], [3]. Common algorithms include hierarchical clustering (HC), which was applied in analyzing temporal expression patterns [2]. However, HC imposes a rigid tree structure on the data, and relies on an adhoc similarity measure to compute cluster association. Self-organizing maps (SOMs) provide another approach that imposes partial structure on the clusters. SOMs have been applied to hematopoietic differentiation by highlighting certain genes and pathways involved in differentiation therapy and used in the treatment of promyelocytic leukemia [3]. SOMs, however, yield different decompositions of the data depending on the initial geometry of the nodes. Recently, non-negative matrix factorization (NMF) was used to compress the thousands of genes in a genome in terms of a small number of metagenes. Samples can then be analyzed by summarizing their gene expression patterns in terms of expression patterns of the metagenes [1]. NMF appeared as an alluring clustering and classification technique as it does not impose any prior structure or knowledge on the data. Brunet et al. [1] experimentally showed that NMF leads to more accurate and more robust clustering than HC and SOM.

The NMF algorithm, however, is a deterministic technique. In particular, the algorithm does not take into account the noise in the data, and the effect of the data noise on the NMF in terms of convergence and robustness has not been investigated. On the other hand, gene expression profiles are known to be noisy, and therefore, must be processed and analyzed by systems that take into account the stochastic nature of the data. In this paper, we extend the NMF algorithm to the probabilistic case, where the data is described by its probability density function. We derive the probabilistic NMF update rules, and show that they converge to the desired factorization.

This paper is organized as follows: In Section II, we review the NMF algorithm and recall its update rules to minimize the squared error of the factorization. Section III introduces the PNMF algorithm and derives its corresponding update rules. Simulation results on a leukemia microarray dataset are presented in section IV, where we also conduct a performance comparison between the NMF and PNMF algorithms.

II. The NMF Algorithm

In this section, we present a concise overview of the NMF method. The NMF is a constrained matrix factorization method, where a non-negative matrix V is factorized into two non-negative matrices W and H, in the sense that 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, which do not impose the non-negativity constraint. Analytically, we have

$$V = WH + E, \tag{1}$$

where W and H are non-negative matrices, and E is a residual that can either be negative or positive. NMF is formulated as a constrained optimization problem, where a cost function is minimized. There are different types of non-negative matrix factorizations depending on the cost function used to measure the divergence between V and WH [4]. In this paper, we consider the squared error or Frobenius norm cost function. The factorization problem in the squared error version of NMF may be stated as: Given a non-negative matrix $V \in \mathbb{R}^{n\times m}$, find non-negative matrices $W \in \mathbb{R}^{n\times k}$ and $H \in \mathbb{R}^{k\times m}$ that minimize the function

$$F(W,H) = \|V - WH\|_F^2, \tag{2}$$

978-1-4673-0490-0/11/$26.00 ©2011 IEEE

Where $\|.\|_F$ denotes the Frobenius norm. Lee and Seung [4] showed that the following iterative update rules minimize the squared error in (2) subject to the non-negativity constraint,

$$H_{ij} \longleftarrow H_{ij}\frac{(W^TV)_{ij}}{(W^TWH)_{ij}} \tag{3}$$

$$W_{ij} \longleftarrow W_{ij}\frac{(VH^T)_{ij}}{(WHH^T)_{ij}} \tag{4}$$

In DNA microarrays analysis, the data matrix V consists of the expression levels of n genes in m samples. These samples could be time points, experiments or distinct tissues from different subjects. The goal is to find a small number of metagenes, each defined as a positive linear combination of the n genes. The gene expression patterns of the samples can then be studied through the expression patterns of the metagenes [1]. The $W = \{W_{ij}\}$ matrix represents a concatenation of k metagenes, which is also a basis of the genes vectors. Each entry w_{ij} is the coefficient of gene i in metagene j. The columns of the matrix $H = \{H_{ij}\}$ represent the metagene expression patterns of the corresponding sample. Each entry h_{ij} is the expression level of metagene i in sample j. In particular, the NMF algorithm allows us to group the m samples into k clusters, where $k < \min(n, m)$.

III. The Probabilistic NMF Algorithm

The NMF algorithm implicitly assumes that the data is deterministic. In particular, the convergence of the update rules in (3) and (4) and their robustness to noise in the data are unknown. In this section, we assume that the data, represented by the non-negative matrix V, is corrupted by Gaussian noise. Then, the data follows the conditional distribution,

$$p(V \mid W, H, \sigma^2) = \prod_{i=1}^{N}\prod_{j=1}^{M}[\mathcal{N}(V_{ij} \mid \mathbf{u}_i^T\mathbf{h}_j, \sigma^2)], \tag{5}$$

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 (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

$$p(U \mid \sigma_W^2) = \prod_{i=1}^{N}\mathcal{N}(\mathbf{u}_i \mid 0, \sigma_W^2 I) = p(W \mid \sigma_W^2) \tag{6}$$

$$p(H \mid \sigma_H^2) = \prod_{i=1}^{M}\mathcal{N}(\mathbf{h}_j \mid 0, \sigma_H^2 I) \tag{7}$$

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

$$\ln(p(W, H \mid V, \sigma^2, \sigma_H^2, \sigma_W^2)) = -\frac{1}{2\sigma^2}\sum_{i=1}^{N}\sum_{j=1}^{M}(V_{ij} - \mathbf{u}_i^T\mathbf{h}_j)^2$$
$$-\frac{1}{2\sigma_W^2}\sum_{i=1}^{N}\|\mathbf{u}_i\|^2 - \frac{1}{2\sigma_H^2}\sum_{j=1}^{M}\|\mathbf{h}_j\|^2 + C, \tag{8}$$

where C is a constant term depending only on the standard deviations σ, σ_W and σ_H. Maximizing (8) is equivalent to minimizing the following function

$$(W^*, H^*) = \underset{W,H\geq 0}{\arg\min}\, \|V - WH\|_F^2 + \lambda_W\|W\|_F^2 + \lambda_H\|H\|_F^2, \tag{9}$$

where $\lambda_W = \frac{\sigma^2}{\sigma_W^2}$ and $\lambda_H = \frac{\sigma^2}{\sigma_H^2}$. Observe that the PNMF formulation in (9) reduces to a weighted regularized matrix factorization problem. The following proposition provides the update rules for the PNMF constrained optimization problem.

Proposition 1. *Consider the function*

$$f(W, H) = \|V - WH\|_F^2 + \alpha\|W\|_F^2 + \beta\|H\|_F^2, \tag{10}$$

where α and β are non-negative real numbers. Then, f is non-increasing under the update rules

$$H_{ij} \longleftarrow H_{ij}\frac{(W^TV)_{ij}}{(W^TWH + \beta H)_{ij}} \tag{11}$$

$$W_{ij} \longleftarrow W_{ij}\frac{(VH^T)_{ij}}{(WHH^T + \alpha W)_{ij}}. \tag{12}$$

provided the initial values of the algorithm are chosen to be positive. Moreover, the function f is invariant under these updates if and only if W and H are at a stationary point.

The proof of the proposition is provided in the Appendix. Since the data matrix V is non-negative, the update rules in (11) and (12) lead to non-negative factors W and H as long as the initial values of the algorithm are chosen to be positive.

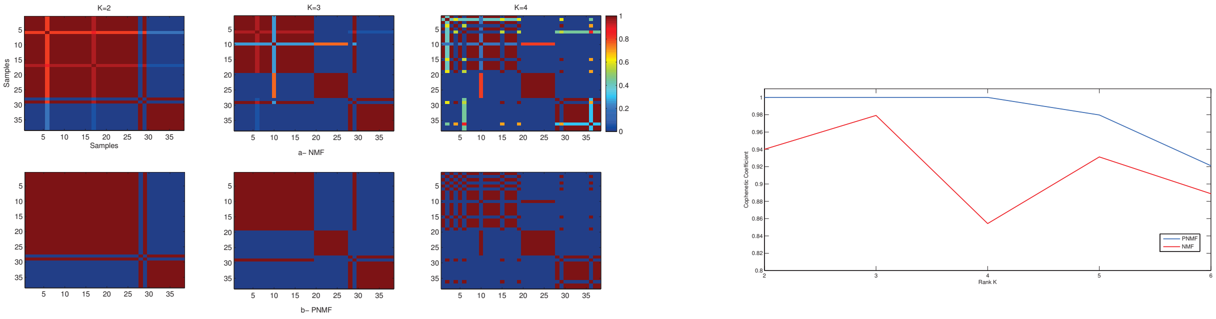
IV. Simulation Results

We illustrate the use of the probabilistic non-negative matrix factorization in elucidating cancer subtypes of leukemia. 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 [1]. We consider an ALL-AML dataset, which contains 5000 genes and 38 bone marrow samples (tissues from different patients for the considered genes) [1]. We would like to see how the proposed PNMF algorithm can discover these classes compared to the deterministic NMF algorithm.

A. Performance evaluation

In the non-negative factorization of the data matrix, the i^{th} gene belongs to the j^{th} cluster if the entry W_{ij} is the largest element in the considered row for the gene i, and the j^{th} sample belongs to the i^{th} cluster if the entry H_{ij} is the largest element in the considered column for the sample j. In this paper, we focus on sample clustering.

Convergence of the algorithm to a (local) minimum is assessed by the connectivity matrix C , whose entries refer to whether two samples belong to the same cluster by assigning 0 or 1 value. Specifically, $c_{ij} = 1$ if samples i and j belong to the same cluster, and $c_{ij} = 0$ otherwise. If C is almost invariant among runs, the objective function is considered to have reached a minimum. The number of clusters, k, is assessed quantitatively based on the average connectivity matrix, also



(a)

(b)

Fig. 1. (a) Top row: NMF Consensus matrices, bottom row: PNMF consensus matrices; (b) Cophenetic coefficient versus the rank k (NMF in red and PNMF in blue).

called consensus matrix [1]. The entries of the consensus matrix, $\bar{C}$ range from 0 to 1, reflecting the probability that samples i and j belong to the same cluster. Observe that perfect consensus translates into a matrix with all entries set to either 0 or 1. 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 would translate into a block-diagonal matrix with non-overlapping blocks of 1's along the diagonal, each block corresponding to a different cluster [1]. 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 [5].The measure is based on the correlation coefficient of the consensus matrix, ρ_k, also called the cophenetic correlation coefficient. Analytically, we have $\rho_k = \frac{1}{m^2}\sum_{ij} 4(c_{ij} - \frac{1}{2})^2$ [5]. 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 begins to fall (see Fig. 1(b)).

We apply the deterministic NMF and proposed PNMF algorithms to the leukemia dataset. Figure 1(a) shows the consensus matrices corresponding to $k = 2, 3, 4$ clusters for the NMF and PNMF algorithms. In this figure, the matrices are mapped using the gradient color in such a way 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' classification. Specifically, the classes, as distinguished by the PNMF algorithm, are better defined and the 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 NMF. Higher ranks k reveal further portioning of the samples, as shown in Fig. 1(a). The clear block diagonal patterns for $k = 2$ and 3 attest to the accuracy of the models with 2 and 3 classes, whereas a rank-4 factorization shows scattering from this structure. Rank $k = 4$ does not correspond to any biological significance. This observation is also reflected in the decreased value of the cophenetic correlation for rank 4, as shown in Fig. 1(b). It is also interesting to observe that the nested nature of the blocks for $k = 3$ corresponds to the known subdivision of the ALL samples into the T and B classes.

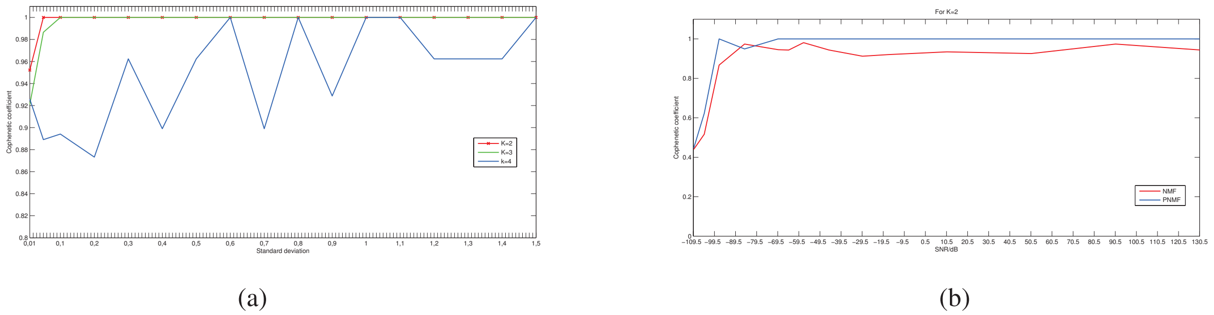
In order to assess the influence of the PNMF parameters, and especially the choice of σ on the algorithm, we study the performance of the PNMF for varying values of σ. Recall that, in the probabilistic model, σ measures the uncertainty in the data or the noise power in the gene expression measurements. 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 2(a) plots the cophenetic coefficient versus the standard deviation σ 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 converges to the NMF, which explains the drop in the cophenetic coefficient for values of σ near zero.

B. Robustness analysis

We assess 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,

$$V_{noisy} = V + \sigma_n R, \tag{13}$$

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 Fig. 2(b) the cophenetic coefficient versus the SNR, measured in dB, for both the NMF and PNMF algorithms. We observe that the PNMF algorithm leads to more robust clustering than the deterministic NMF for all SNR values. Moreover, the cophenetic coefficient of the PNMF algorithm stabilizes for SNR$\geq -97\ dB$ whereas it only stabilizes for SNR$\geq -85.5\ dB$ in the deterministic case.



(a)

(b)

Fig. 2. (a) Cophenetic coefficient versus standard deviation of the noise σ for $k = 2$ (red), 3 (green) and 4 (blue); (b) Cophenetic coefficient versus $SNR(dB)$ for $k = 2$ (NMF in red and PNMF in blue).

V. Conclusion

We presented an extension of the 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. The application to the leukemia dataset shows that the PNMF is able to recover biologically significant phenotypes and identify the known nested structure of leukemia classes. In addition, our simulation results showed that the PNMF leads to a more accurate and more robust clustering than its deterministic homologue, the NMF.

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

Proof of Proposition 1:

Definition 1. $G(h, h')$ *is an auxiliary function for* $F(h)$ *if* $G(h, h') \geq F(h)$ *and* $G(h, h) = F(h)$.

The following lemma in [4] shows the usefulness of the auxiliary function.

Lemma 1. *[4] if* G *is an auxiliary function, then* F *is nonincreasing under the update*

$$h^{(k+1)} = \underset{h}{argmin}\, G(h, h^{(k)}). \tag{14}$$

The following lemma provides an auxiliary function for the objective function F in (10).

Lemma 2. *Consider the diagonal matrix*

$$\Phi_{ij}(h^{(k)}) = \delta_{ij}(W^tWh^{(k)})_i/h_i^{(k)} + \beta\,\delta_{ij} = K_{ij}(h^{(k)}) + \beta\,\delta_{ij}, \tag{15}$$

where δ_{ij} *is the kronecker delta function. Then,*

$$\begin{aligned} G(h, h^{(k)}) = \quad & F(h^{(k)}) + (h - h^{(k)})^t \nabla F(h^{(k)}) + \\ & \frac{1}{2}(h - h^{(k)})^t \Phi(h^{(k)})(h - h^{(k)}) \end{aligned} \tag{16}$$

is an auxiliary function for $F(h) = \sum_i (v_i - \sum_j W_{ij}h_j)^2 + \alpha \|W\|_F^2 + \beta \sum_i \|h_i\|^2$.

Proof of Lemma 2: The fact that $G(h, h) = F(h)$ is obvious. Therefore, we need only to show that $G(h, h^{(k)}) \geq F(h)$. To do this, we compare

$$\begin{aligned} F(h) = \quad & F(h^{(k)}) + (h - h^{(k)})^T \nabla F(h^{(k)}) + \\ & \frac{1}{2}(h - h^{(k)})^T (W^TW + \beta I)(h - h^{(k)}) \end{aligned} \tag{17}$$

with Eq. (16) to find that $G(h, h^{(k)}) \geq F(h)$ is equivalent to

$$(h - h^{(k)})^T [K(h^{(k)}) - W^TW](h - h^{(k)}) \geq 0, \tag{18}$$

where $K(h^{(k)})$ is the diagonal matrix defined in Eq. (15) as $K_{ij}(h^{(k)}) = \delta_{ij}(W^tWh^{(k)})_i/h_i^{(k)}$. The proof of the semi-definiteness of the matrix in (18) is provided in [4]. ■

Replacing G in Eq. (14) by its expression in Eq. (16) results in the update rule

$$h^{(k+1)} = h^{(k)} - \Phi(h^{(k)})^{-1} \nabla F(h^{(k)}). \tag{19}$$

Since G is an auxiliary function of F, $F$is non-increasing under this update rule. Writing the components of Eq. (19), we obtain

$$h_i^{(k+1)} = h_i^{(k)} \frac{(W^Tv)_i}{(W^TWh^{(k)} + \beta h^{(k)})_i}. \tag{20}$$

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

References

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

[2] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein, "Cluster analysis and display of genome-wide expression patterns," *Genetics*, vol. 95, pp. 14 863–14 868, December 1998.

[3] P. Tamayo, D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, E. S. Lander, and T. R. Golub, "Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation," *Genetics*, vol. 96, pp. 2907–2912, March 1999.

[4] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," *Advances in Neural Information Processing Systems*, vol. 13, pp. 556–562, 2001.

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