BIOINFORMATICS ORIGINAL PAPER Vol. 23 no. 12 2007, pages 1495–1502
doi:10.1093/bioinformatics/btm134

Gene expression

Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis

Hyunsoo Kim* and Haesun Park*

College of Computing, Georgia Institute of Technology, 266 Ferst Drive, Atlanta, GA 30332, USA

Received on November 2, 2006; revised on February 19, 2007; accepted on April 1, 2007

Advance Access publication May 5, 2007

Associate Editor: David Rocke

ABSTRACT

Motivation: Many practical pattern recognition problems require non-negativity constraints. For example, pixels in digital images and chemical concentrations in bioinformatics are non-negative. Sparse non-negative matrix factorizations (NMFs) are useful when the degree of sparseness in the non-negative basis matrix or the non-negative coefficient matrix in an NMF needs to be controlled in approximating high-dimensional data in a lower dimensional space.

Results: In this article, we introduce a novel formulation of sparse NMF and show how the new formulation leads to a convergent sparse NMF algorithm via alternating non-negativity-constrained least squares. We apply our sparse NMF algorithm to cancer-class discovery and gene expression data analysis and offer biological analysis of the results obtained. Our experimental results illustrate that the proposed sparse NMF algorithm often achieves better clustering performance with shorter computing time compared to other existing NMF algorithms.

Availability: The software is available as supplementary material.

Contact: hskim@cc.gatech.edu, hpark@acc.gatech.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

In many data-mining problems, dimension reduction is imperative for efficient manipulation of massive quantity of high-dimensional data. The subspace method has demonstrated its success in numerous pattern recognition tasks including efficient classification (Kim *et al.*, 2005), clustering (Ding *et al.*, 2002) and fast search (Berry *et al.*, 1999) . There are two general approaches for reducing dimensionality, i.e. feature extraction and feature selection. Feature extraction is transforming the existing features into a lower dimensional space, while feature selection is selecting a subset of the existing features without a transformation. For feature extraction, principal component analysis (PCA), linear discriminant analysis (LDA) and non-negative matrix factorization (NMF) have been widely used. Many practical pattern recognition problems require non-negativity constraints. For example, pixels in digital images and chemical concentrations in bioinformatics are non-negative. NMF is a useful technique in approximating these high-dimensional data.

Given a non-negative matrix A of size $m \times n$, where each column of A corresponds to a data point in the m-dimensional space, and a positive integer $k < \min\{m, n\}$, NMF finds two non-negative matrices $W \in \mathbb{R}^{m \times k}$ and $H \in \mathbb{R}^{k \times n}$ so that

$$A \approx WH. \tag{1}$$

A solution to the NMF problem can be obtained by solving the following optimization problem:

$$\min_{W,H} f(W,H) \equiv \frac{1}{2}\|A - WH\|_F^2, \ s.t. \ W, H \geq 0, \tag{2}$$

where $W \in \mathbb{R}^{m \times k}$ is a basis matrix, $H \in \mathbb{R}^{k \times n}$ is a coefficient matrix, $\|\cdot\|_F$ is the Frobenius norm and $W, H \geq 0$ means that all elements of W and H are non-negative. Due to $k < m$, dimension reduction is achieved and a lower dimensional representation of A in a k-dimensional space is given by H.

Since NMF may give us direct interpretation due to non-subtractive combinations of non-negative basis vectors, it has recently received much attention and it has been applied to many interesting problems including text data mining (Chagoyen *et al.*, 2006; Lee and Seung, 1999; Pauca *et al.*, 2004) gene expression data analysis (Brunet *et al.*, 2004; Carmona-Saez *et al.*, 2006; Gao and Church, 2005; Kim and Tidor, 2003; Maher *et al.*, 2006), microarray comparative genomic hybridization (aCGH) data (Carrasco *et al.*, 2006) and functional characterization of gene lists (Pehkonen *et al.*, 2005). One of the interesting properties of NMF is that it often generates sparse basis vectors that allow us to discover parts-based basis vectors. NMF generated holistic basis images instead of parts-based basis images for a facial image dataset in the results presented in Li *et al.* (2001) and Hoyer (2004). Several approaches (Dueck *et al.*, 2005; Hoyer, 2004; Pascual-Montano *et al.*, 2006; Pauca *et al.*, 2006) have been proposed to explicitly control the degree of sparseness of W and H. However, if strong sparsity constraints are imposed on the basis matrix W, some useful information for gene selection in microarray data analysis may be lost. Some other approaches (Gao and Church, 2005; Pauca *et al.*, 2004) imposed sparsity

*To whom correspondence should be addressed.

constraints on only the coefficient matrix H, but there may be mathematical difficulties related to convergence.

In this article, we introduce a sparse NMF algorithm that can control the degree of sparseness in the coefficient matrix H via alternating non-negativity-constrained least squares, and apply it to microarray data analysis. The rest of this article is organized as follows. We give brief overviews on various NMF algorithms in Section 2.1. In Section 2.2, we introduce sparse NMF formulations and algorithms based on alternating non-negativity-constrained least squares and discuss their convergence properties. In Section 3, we describe stopping criteria, and present experimental results and biological analysis illustrating properties of the proposed sparse NMF. Summary is given in Section 4.

2 METHODS

2.1 Review of NMF algorithms

Lee and Seung (1999, 2000) suggested NMF algorithms based on multiplicative update rules of W and H. The distance $\|A - WH\|_F$ is non-increasing under the following multiplicative update rules:

$$H_{qj} \leftarrow H_{qj} \frac{(W^T A)_{qj}}{((W^T W)H)_{qj}},$$

for $1 \le q \le k$ and $1 \le j \le n$,

$$W_{iq} \leftarrow W_{iq} \frac{(AH^T)_{iq}}{(W(HH^T))_{iq}},$$

for $1 \le i \le m$ and $1 \le q \le k$. The divergence $D(A, WH) = \sum_{i=1}^{m} \sum_{j=1}^{n} (A_{ij} \ln(A_{ij}/(WH)_{ij}) - A_{ij} + (WH)_{ij})$ is non-increasing under the different multiplicative update rules (Lee and Seung, 2000). Gonzales and Zhang (2005) claimed that these non-increasing properties of multiplicative update rules may not imply the convergence to a stationary point within realistic amount of runtime for problems of meaningful sizes. More detailed review on NMF algorithms can be found in Berry *et al.* (2006).

Hoyer (2004) devised a sparse NMF algorithm based on the projected gradient descent method (SNMF/PGD) in order to constrain NMF to find solution with desired sparseness in W and H. To impose sparseness constraints on only one matrix W or H, this algorithm uses a multiplicative update rule for updating the counter matrix, which suffers from slow convergence. Puscual-Montano *et al.* (2006) claimed that non-smooth NMF (*ns*NMF) outperformed previous sparse NMF variants for their synthetic and real datasets. The *ns*NMF (Puscual-Montano *et al.*, 2006) is also based on multiplicative update rules.

Pauca *et al.* (2006) proposed a constrained NMF (CNMF) formulation,

$$\min_{W,H} \{\|A - WH\|_F^2 + \alpha \|W\|_F^2 + \beta \|H\|_F^2\}, \; s.t. \; W, H \ge 0, \tag{3}$$

where α and β are regularization parameters. A sparse NMF algorithm using the following least squares,

$$\min_{H} \{\|A - WH\|_F^2 + \beta \|H\|_F^2\}, \tag{4}$$

has appeared in Pauca *et al.* (2004) and Gao and Church (2005). This algorithm sets negative values in H to zero during iterations. However, setting negative values to zero for imposing non-negativity is not recommended, since there is no guarantee that the algorithm converges (Bro and de jong, 1997). The parameter β in Equation (4) has a scaling effect since a large value of β would suppress $\|H\|_F$. As $\|H\|_F$ is suppressed, $\|W\|_F$ may grow relatively large, and therefore, the algorithm needs column normalization of W during iterations. However, the normalization of W changes the objective function, and this makes convergence analysis difficult. It is well known that a quadratic penalty corresponds to Gaussian priors and does not encourage sparsity but rather scales the result giving non-sparse low values. Thus, L_1-norm based formulations would be more appropriate than L_2-norm based formulations so as to control sparsity (Tibshirani, 1996).

2.2 Sparse NMFs based on alternating non-negativity-constrained least squares

In order to enforce sparseness on W or H in the NMF presented in Equation (1), we introduce two formulations and the corresponding algorithms for sparse NMFs, i.e. SNMF/L for sparse W (where 'L' denotes the sparseness imposed on the left factor) and SNMF/R for sparse H (where 'R' denotes the sparseness imposed on the right factor). Our sparse NMF formulations that impose the sparsity on a factor of NMF utilize L_1-norm minimization and the corresponding algorithms are based on alternating non-negativity constrained least squares (ANLS). Each sub-problem is solved by a fast non-negativity constrained least squares (NLS) algorithm (van Benthem and Keenan, 2004) that is improved upon the active set based NLS method. Bro and de Jong (1997) made a substantial speed improvement to Lawson and Hanson's algorithm (Lawson and Hqnson, 1974) for multiple right-hand-side cases. van Benthem and Keenan (2004) devised an algorithm that further improves the performance of NLS.

2.2.1 Formulations for Sparse NMFs **SNMF/R:** To apply sparseness constraints on H, we formulate the following SNMF/R optimization problem:

$$\begin{aligned} \min_{W,H} \quad & \frac{1}{2}\{\|A - WH\|_F^2 + \eta \|W\|_F^2 \\ & + \beta \sum_{j=1}^{n} \|H(:,j)\|_1^2\}, \\ & s.t. W, H \ge 0, \end{aligned} \tag{5}$$

where $H(:, j)$ is the j-th column vector of H, $\eta > 0$ is a parameter to suppress $\|W\|_F^2$, and $\beta > 0$ is a regularization parameter to balance the trade-off between the accuracy of the approximation and the sparseness of H. The SNMF/R algorithm begins with the initialization of W with non-negative values. Then, it iterates the following ANLS until convergence:

$$\min_{H} \left\| \begin{pmatrix} W \\ \sqrt{\beta}\mathbf{e}_{1\times k} \end{pmatrix} H - \begin{pmatrix} A \\ \mathbf{0}_{1\times n} \end{pmatrix} \right\|_F^2, \; s.t. \; H \ge 0, \tag{6}$$

where $\mathbf{e}_{1\times k} \in \mathbb{R}^{1\times k}$ is a row vector with all components equal to one and $\mathbf{0}_{1\times n} \in \mathbb{R}^{1\times n}$ is a zero vector, and

$$\min_{W} \left\| \begin{pmatrix} H^T \\ \sqrt{\eta} I_k \end{pmatrix} W^T - \begin{pmatrix} A^T \\ 0_{k\times m} \end{pmatrix} \right\|_F^2, \; s.t. \; W \ge 0, \tag{7}$$

where I_k is an identity matrix of size $k \times k$ and $0_{k\times m}$ is a zero matrix of size $k \times m$. Equation (6) minimizes L_1-norm of columns of $H \in \mathbb{R}^{k\times n}$ which imposes sparsity on H.

SNMF/L: To impose sparseness constraints on W, we introduce the SNMF/L formulation:

$$\begin{aligned} \min_{W,H} \quad & \frac{1}{2}\{\|A - WH\|_F^2 + \eta \|H\|_F^2 \\ & + \alpha \sum_{i=1}^{m} \|W(i,:)\|_1^2\}, \\ & s.t. \; W, H \ge 0, \end{aligned} \tag{8}$$

where $W(i, :)$ is the i-th row vector of W, $\eta > 0$ is a parameter to suppress $\|H\|_F^2$, and $\alpha > 0$ is a regularization parameter to balance the trade-off between the accuracy of the approximation and the sparseness of W. The algorithm for SNMF/L is also based on ANLS.

2.2.2 Convergence properties of sparse NMF algorithms We show the convergence property of the sparse NMF algorithms. Since the convergence properties of SNMF/L and SNMF/R are essentially the same, we will only discuss the case of SNMF/R in more detail. Under conditions of $\eta > 0$, $\beta > 0$, and $H(i,j) = |H(i,j)|$ due to $H(i,j) \geq 0$, Equation (5) can rewritten as

$$\begin{aligned} \min_{W,H} \quad & \frac{1}{2}\{\|A - WH\|_F^2 + \eta\|W\|_F^2 \\ & + \beta\sum_{j=1}^{n}\left(\sum_{q=1}^{k} H(q,j)\right)^2\}, \\ & s.t.\ W, H \geq 0, \end{aligned} \tag{9}$$

which is differentiable with respect to W or H. SNMF/R contains two sub-problems for two-block minimization scheme. Grippo and Sciandrone (2000) showed that the two-block coordinate descent method does not require each sub-problem to have a unique solution for convergence. The objective function in Equation (9) is coercive on the feasible set; as the feasible set is closed, the intersection of any level set of this function with the feasible set is compact. Therefore, any minimization process that reduces the objective function and preserves feasibility generates points that remain in a compact set. The existence of accumulation points and the differentiability of the objective function in Equation (9) imply that the assumptions of Grippo and Sciandrone's Corollary (Grippo and Sciandrone, 2000) are satisfied, so that we can establish that the two-block minimization process is convergent, in the sense that every accumulation point is a critical point of the problem shown in Equation (9). Similarly, it can be shown that the algorithm SNMF/L converges to a stationary point.

3 RESULTS

3.1 Datasets description

We used the leukemia gene expression dataset (ALLAML) (Golub *et al.*, 1999) and the central nervous system tumors dataset (CNS) (Pomeroy *et al.*, 2002). The ALLAML dataset contains acute lymphoblastic leukemia (ALL) that has B and T cell subtypes, and acute myelogenous leukemia (AML) that occurs more commonly in adults than in children. This gene expression dataset consists of 38 bone marrow samples (19 ALL-B, 8 ALL-T and 11 AML) with 5000 genes. The central nervous system dataset is composed of four categories of CNS tumors with 5597 genes. It consists of 34 samples representing four distinct morphologies: 10 classic medulloblastomas, 10 malignant gliomas, 10 rhabdoids and 4 normal cerebella. All datasets used by us contain only non-negative entries. All algorithms were implemented in Matlab 6.5 (MATLAB, 1992). The Matlab codes for NMF using divergence-based multiplicative update rules were obtained from Brunet *et al.* (2004) and modified to implement *ns*NMF (Puscual-Montano *et al.*, 2006). All our experiments were performed on a P3 600 MHz machine with 512 MB memory.

3.2 Biclustering

We applied the non-negative factorization of Equation (1) to perform clustering analysis of a data matrix. The rows of a microarray data matrix A represent genes and the columns experiments. We can use the basis matrix W to divide the m genes into k gene-clusters and the coefficient matrix H to divide the n samples into k sample-clusters. Typically, gene i is assigned to gene-cluster q if the $W(i,q)$ is the largest element in $W(i, :)$ and sample j is assigned to sample-cluster q if the $H(q,j)$ is the largest element in $H(:, j)$.

3.3 Stopping criteria

NMF using divergence-based multiplicative update rules (i.e. NMF/DUR (Brunet *et al.*, 2004) and *ns*NMF (Pascual-Montano *et al.*, 2006)) in our implementation stops if $\hat{C}$ has not changed for more than 40 convergence tests (each made at 10 iterations), where $\hat{C} = [\hat{c}_{ij}] \in \mathbb{R}^{n\times n}$ is the connectivity matrix of which entry is $\hat{c}_{ij} = 1$ if samples i and j belong to the same sample-cluster, and $\hat{c}_{ij} = 0$ if they belong to different sample-clusters.

For SNMF/L and SNMF/R, we tested convergence at every five iterations by the combined convergence criterion using the Karush-Kuhn-Tucker (KKT) optimality conditions and the convergence of positions of the largest elements in rows of W and columns of H. Our sparse NMF algorithms stop if both the positions of the largest elements in rows of W, i.e. $\tilde{\mathbf{w}} = (\tilde{w}_1, \ldots, \tilde{w}_m)$, and the positions of the largest elements in columns of H, i.e. $\tilde{\mathbf{h}} = (\tilde{h}_1, \ldots, \tilde{h}_n)$, have not changed for more than or equal to 10 convergence tests, where $\tilde{w}_i$ is the position of the largest element in the i-th row of W and $\tilde{h}_j$ is the positions of the largest element in the j-th column of H, and the following KKT conditions are satisfied. The KKT conditions for each objective function $f(W,H)$ with non-negativity constraints $W \geq 0$ and $H \geq 0$ are

$$\begin{aligned} &\text{(C1) } W_{iq} \geq 0, \\ &\text{(C2) } H_{qj} \geq 0, \\ &\text{(C3) } (\partial f(W,H)/\partial W)_{iq} \geq 0, \\ &\text{(C4) } (\partial f(W,H)/\partial H)_{qj} \geq 0, \\ &\text{(C5) } W_{iq} \cdot (\partial f(W,H)/\partial W)_{iq} = 0, \\ &\text{(C6) } H_{qj} \cdot (\partial f(W,H)/\partial H)_{qj} = 0, \forall i,q,j. \end{aligned} \tag{10}$$

These conditions can be rewritten as

$$\begin{aligned} \min(W, \partial f(W,H)/\partial W) = 0, \\ \min(H, \partial f(W,H)/\partial H) = 0, \end{aligned} \tag{11}$$

where the minimum is taken component wise (Gonzales and Zhang, 2005). Let Δ_o be the KKT residual measured by the L_1-vector norm,

$$\begin{aligned} \Delta_o = & \sum_{i=1}^{m}\sum_{q=1}^{k} |\min(W_{iq}, (\partial f(W,H)/\partial W)_{iq})| \\ & + \sum_{q=1}^{k}\sum_{j=1}^{n} |\min(H_{qj}, (\partial f(W,H)/\partial H)_{qj})|. \end{aligned} \tag{12}$$

We count the number of the elements in W that did not converge yet, i.e. $\delta_W = \#(\min(W, \partial f(W,H)/\partial W) \neq 0)$, and the number of the elements in H that did not converge yet,

i.e. $\delta_H = \#(\min(H, \partial f(W,H)/\partial H) \neq 0)$. We define the following normalized KKT residual:

$$\Delta = \frac{\Delta_o}{\delta_W + \delta_H}, \tag{13}$$

which reflects the average of convergence errors for elements in W and H that did not converge. The mathematical convergence criterion is defined as

$$\Delta \leq \epsilon \Delta_1, \tag{14}$$

where Δ_1 is the Δ value in the first iteration and ϵ is a tolerance. We used $\epsilon = 10^{-4}$ for our experiments.

3.4 Clustering performance comparison

To measure the performance of NMFs in clustering, we used purity and entropy. Suppose we are given l categories (true class labels), while NMF generates k clusters. Purity is given by

$$\text{Purity} = \frac{1}{n}\sum_{q=1}^{k} \max_{1 \leq j \leq l}(n_q^j),$$

where n is the total number of samples and n_q^j is the number of samples in the cluster q that belong to original class j ($1 \leq j \leq l$). The larger the value of purity, the better the clustering performance. Entropy is defined as follows:

$$\text{Entropy} = -\frac{1}{n \log_2 l}\sum_{q=1}^{k}\sum_{j=1}^{l} n_q^j \log_2 \frac{n_q^j}{n_q},$$

where l denotes the number of original class labels and n_q is the size of cluster q. The smaller the value of entropy, the better the clustering quality.

The parameter η in Equation (5) is important in keeping $\|W\|_F^2$ small. We set it to be the square of the maximal element in A. For the initialization of W in SNMF/R, the elements in the initial matrix W were randomly chosen and normalized so that the columns of the basis matrix W have unit L_2-norm, i.e. $\|W(:,q)\|_2 = 1$ *for* $1 \leq q \leq k$.

Tables 1 and 2 show the results of SNMF/R with various values of β on the ALLAML dataset with $k = 3$ and on the CNS tumors dataset with $k = 4$, respectively. We compared our proposed SNMF/R with NMF based on divergence-based multiplicative update rules (Brunet *et al.*, 2004; Lee and Seung, 2000). Average sparseness, purity and entropy were computed by running each algorithm five times with different random initializations. By increasing β in SNMF/R, we could obtain a sparser H. SNMF/R algorithm achieved better clustering performance (higher purity, lower entropy) than NMF/DUR within certain range of β with shorter computing time. By increasing α in SNMF/L, we could enhance the sparsity of W (results are not shown). SNMF/L can be applied to obtain parts-based basis vectors.

We compared our methods with other sparse NMF variants on the ALLAML dataset. We tested Hoyer's sparse NMF based on the projected gradient descent method by his Matlab implementation with the sparseness control parameter $s_H = 0.4$ to impose sparsity constraints on only H (see Hoyer (2004) for the details). The average percentages of zero elements in W and H obtained from SNMF/PGD were 0.12 and 21.75%, respectively. SNMF/R showed significantly better clustering performance than SNMF/PGD (average purity = 0.895 and average entropy = 0.280). Moreover, SNMF/R required much shorter computing time and smaller number of iterations than SNMF/PGD (average computing time = 110.1 and average iteration number = 517.8). We tested *ns*NMF with the smoothness control parameter $\theta = 0.5$ suggested in Carmona-Saez *et al.* (2006) for biclustering of gene expression data due to reasonable results from numerous empirical tests. *ns*NMF generated average purity = 0.963 and average entropy = 0.108. The average percentages of elements in the range of $[0, 10^{-8})$ in W and H obtained from *ns*NMF were 8.94 and 25.26%, respectively. It took 79.1 s with 698 iterations on the average. It produced the sparsest W among NMF algorithms we compared. The average percentages of elements in the range of $[0, 10^{-4})$ in W and H obtained from *ns*NMF were 65.94 and 29.30%, respectively, which are much higher than 0.12 and 0.35% obtained from NMF/DUR. *ns*NMF enhanced the sparseness of W as well as H simultaneously. However, the additional sparsity constraints on W is not always helpful. For example, some genes are over-expressed in samples that belong to more than one clusters. NMFs typically generate W whose rows corresponding to these genes have large values

Table 1. Performance dependency of SNMF/R (k=3) on various β values on the leukemia data matrix of size 5000×38. We present the average percentages of zero elements in W and H over five runs with different random initializations. We also present average purity, entropy, computing time (in seconds) and the number of iterations

Leukemia	NMF/DUR	SNMF/R			
β	–	0.001	0.01	0.1	0.5
#(W = 0) (%)	0.10%*	2.43%	2.17%	1.57%	1.09%
#(H = 0) (%)	0.00%*	24.56%	30.70%	44.74%	51.75%
Purity	0.953	0.974	0.974	0.947	0.921
Entropy	0.141	0.095	0.095	0.158	0.210
Number of iterations	502.0	328.0	139.0	77.0	95.0
Computing time	53.6	40.1	17.0	9.4	10.9

Table 2. Performance dependency of SNMF/R (k=4) on various β values on the CNS tumors data matrix of size 5597×34. We present the average percentages of zero elements in W and H over five runs with different random initializations. We also present average purity, entropy, computing time (in seconds) and the number of iterations

CNS tumors	NMF/DUR	SNMF/R			
β	–	0.01	0.1	1.0	2.0
#(W = 0) (%)	1.65%*	8.45%	7.45%	5.06%	4.31%
#(H = 0) (%)	1.47%*	25.74%	28.68%	36.76%	41.91%
Purity	0.941	0.971	0.971	0.971	0.941
Entropy	0.122	0.071	0.071	0.071	0.144
Number of iterations	566.0	319.0	174.0	134.0	103.0
Computing time	63.4	51.6	29.5	20.9	16.0

*For NMF using divergence-based multiplicative update rules (NMF/DUR) (Brunet *et al.*, 2004), the average percentages of the number of very small non-negative elements that are smaller than 10^{-8} in W and H are presented.

in more than one factors. One can obtain a sparser W by imposing strong sparsity constraints on W. However, the sparser W loses the original information and may give us wrong information that these genes are over-expressed in only one factor. Consequently, one may inappropriately select these genes from the sparse W since they are considered as factor-specific genes. We also tested the probabilistic sparse matrix factorization (PSMF) (Dueck *et al.*, 2005). PSMF also produced a sparser W. Moreover, our SNMF/R showed better performance than PSMF in terms of purity, entropy and computing speed.

We used the model selection method proposed by Brunet *et al.* (2004) to determine the number of factors. We ran NMF algorithms 30 times to obtain the average connectivity matrix (i.e. consensus matrix) whose entries reflect the probability that samples i and j belong to the same cluster. To measure the dispersion of a consensus matrix C, we defined the dispersion coefficient (ρ) as

$$\rho = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} 4 * \left(C_{ij} - \frac{1}{2}\right)^2. \quad (15)$$

The value of coefficient is $\rho = 1$ for a perfect consensus matrix (all entries = 0 or 1) and $0 \leq \rho < 1$ for a scattered consensus matrix. After obtaining ρ_k values for various k, we can determine the number of clusters from the maximal ρ_k. Figure 1 illustrates the determination of the number of clusters in the CNS tumors dataset. SNMF/R with $\beta = 0.01$ found perfect consensus matrices for $k = 2, 3, 4$. In other words, SNMF/R with $\beta = 0.01$ generated H matrices that have the same cluster structure in spite of different random initializations of W for $2 \leq k \leq 4$. SNMF/R yielded finer consensus matrices (higher ρ_k) than NMF/DUR for various k values.

3.5 Biological analysis

Figure 2 presents the matrices W and H obtained from SNMF/R with $\beta = 0.01$ on the ALLAML leukemia dataset, which produced the lowest approximation error $\|A - WH\|_F$ after five runs with different random initializations of W. A column vector of the coefficient matrix H has the contributions of k biological processes to the gene expression of a sample. From the matrix H, we can recognize that ALL-B is dominated by the first biological process. ALL-T is almost controlled by the second biological process. The third biological process is the major component for AML cluster. The cluster of each sample was determined by the positions of the largest elements in columns of H. Only one sample (the 29th sample, AML_13) was incorrectly assigned to ALL-B.

A row vector of the basis matrix W indicates the contributions of a gene to the k biological pathways or processes (i.e. k columns of W). Genes can participate in more than one biological process. It is beneficial to investigate genes that have relatively large coefficient in each biological process. We selected factor-specific genes via the non-negative basis matrix $W \in \mathbb{R}^{m \times k}$ obtained from SNMF/R. We define gene_score for the ith gene as

$$\text{Gene\_score}(i) = 1 + \frac{1}{\log_2(k)}\sum_{q=1}^{k} p(i,q)\log_2(p(i,q)), \quad (16)$$

where $p(i, \Omega)$ is a probability that the i-th gene contributes to cluster Ω, i.e. $p(i, \Omega) = W(i, \Omega)/\sum_{q=1}^{k} W(i, q)$. The gene_score is a real value within the range of [0, 1]. The higher the gene_score value, the more factor-specific the corresponding gene. By using the gene_scores obtained from W, we ranked genes and chose genes whose gene_scores were higher than $\hat{\mu} + 3\hat{\sigma}$, where $\hat{\mu}$ and $\hat{\sigma}$ are the median and the median absolute deviation (MAD) of gene_scores respectively, and the maximal values in the

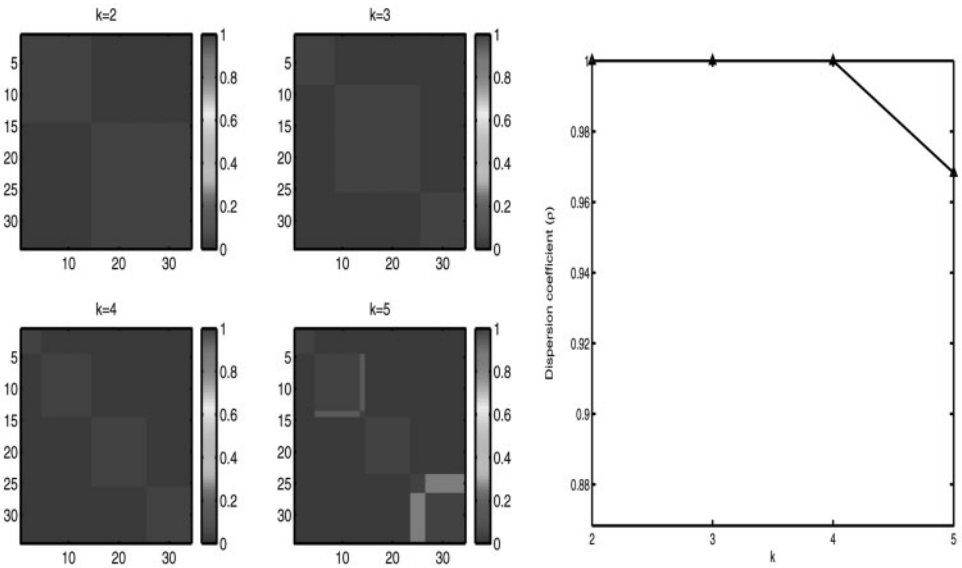

Fig. 1. CNS tumors clustering by SNMF/R with $\beta = 0.01$. (Left) The reordered consensus matrices on the CNS tumors dataset. (Right) The corresponding dispersion coefficients. The dispersion coefficient drops when k increases from 4 to 5, indicating a four-cluster split of the data is more stable than a five-cluster split.

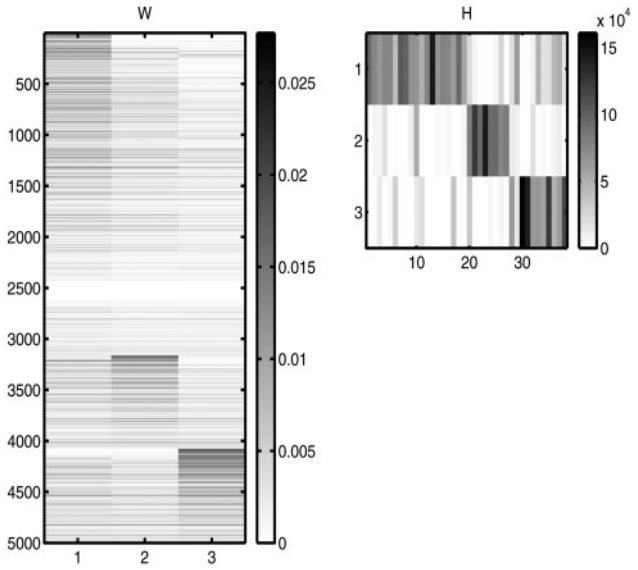

Fig. 2. W (basis matrix) and H (coefficient matrix) obtained from SNMF/R with $\beta = 0.01$ for the ALLAML leukemia dataset (38 samples: 19 ALL-B, 8 ALL-T and 11 AML) with the 5000 most highly varying genes.

corresponding rows of W were larger than the median of all elements in W. Total of 730 genes were selected. The numbers of genes chosen for three factors were 423 genes for ALL-B cluster, 104 genes for ALL-T cluster and 203 genes for AML cluster.

Some chosen genes dominantly contribute to only single biological pathway or process. For instance, MB-1 gene (U05259) is most active in the first process. Transcription factor 7 (T-cell specific, HMG-box) (TCF7, X59871) is active in the second process, which is also known as T cell factor-1 (TCF-1). Some genes play a major role in the third process, for example, Interleukin 8 (IL8, M28130), DF D component of complement (adipsin) (CFD, M84526), Cystatin C (amyloid angiopathy and cerebral hemorrhage (CST3, M27891), Chemokine (C-X-C motif) ligand 2 (CXCL2, M57731), etc. Chemokine is a type of cytokines that bind to a specific cell-surface receptor and is critical to the functioning of both innate and adaptive immune responses. Total of 37 genes including MB-1, IL8, CFD and CST3 were the same genes as those found in Golub *et al.* (1999). Ribosomal protein S3 (RPS3, X57351) simultaneously participates in all three processes. This is reasonable since RPS3 is a housekeeping gene and ribosomal protein genes are usually over-expressed in some cancers. RPS3 encodes a ribosomal protein that is a component of the 40S subunit, where it forms a part of the domain where translation is initiated.

We used the Onto-Express (Draghici *et al.*, 2003; Khatri *et al.*, 2002) to investigate the enrichment of functional annotations of genes selected in each factor. Onto-Express starts by reading the input file that contains a list of GenBank accession numbers, and estimates the statistical significance of the enrichment of Gene-Ontology (GO) terms in the list with respect to a reference list. We used a list of all genes in the dataset as a reference array and hypergeometric distribution.

Table 3. Enrichment of GO categories in genes selected by SNMF/R on the ALLAML leukemia dataset. We present some significant biological processes for each factor, whose P-values are less than 0.01

Factor	Biological process	Number of genes	P-value
Factor 1	Immune response	27	0.0
(423 genes)	Transcription (Tr)	47	6.0×10^{-5}
	Protein biosynthesis	11	3.4×10^{-4}
	B cell activation	2	6.2×10^{-4}
	Regulation of transcription (R-Tr)	53	7.6×10^{-4}
	R-Tr/RNA polymerase II promoter	13	4.5×10^{-3}
	Tr/RNA polymerase II promoter	14	5.6×10^{-3}
Factor 2	T cell activation	2	2.8×10^{-4}
(104 genes)	DNA metabolism	2	4.4×10^{-4}
	DNA replication	4	1.3×10^{-3}
	Cell cycle	7	5.2×10^{-3}
Factor 3	Defense response to bacteria	8	0.0
(203 genes)	Inflammatory response	16	0.0
	Chemotaxis	12	0.0
	Cell-cell signaling	16	0.0
	Response to stimulus	11	0.0
	Anti-apoptosis	8	1.0×10^{-5}
	Cell motility	8	5.1×10^{-4}
	Immune response	12	2.6×10^{-3}
	Apoptosis	8	5.5×10^{-3}
	G-protein coupled receptor pathway	9	7.5×10^{-3}

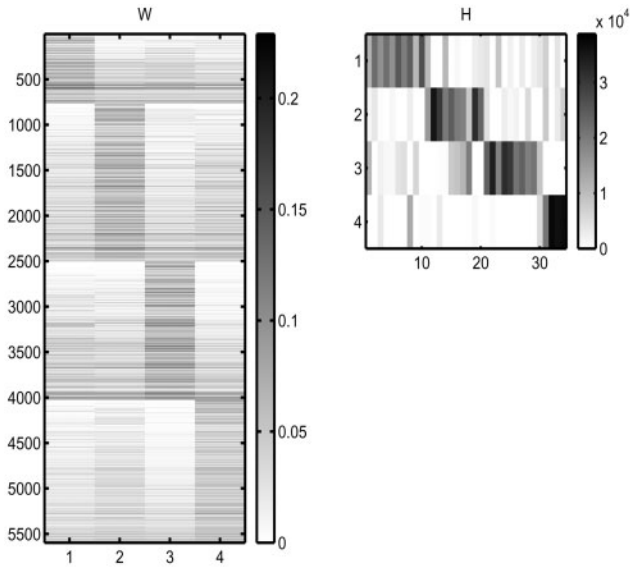

Fig. 3. W (basis matrix) and H (coefficient matrix) obtained from SNMF/R with $\beta = 1.0$ for the CNS tumor dataset (34 samples: 10 classic medulloblastomas, 10 malignant gliomas, 10 rhabdoids, 4 normal cerebella) with the 5597 genes.

Table 3 shows the enrichment of GO terms. We presented some significant biological processes for each factor, whose P-values were less than 0.01.

Figure 3 illustrates the matrices W and H obtained from SNMF/R with $\beta = 1.0$ on the CNS tumors dataset,

Table 4. Enrichment of GO categories in genes selected by SNMF/R on the CNS tumors dataset. We present some significant biological processes for each factor, whose *P*-values are less than 0.01

Factor	Biological process	Number of genes	*P*-value
Factor 1 (42 genes)	Serotonin receptor signaling pathway	2	0.0
	Feeding behavior	2	9.0×10^{-5}
	Central nervous system development	2	2.2×10^{-3}
Factor 2 (93 genes)	Potassium ion transport	4	1.8×10^{-4}
	Synaptic transmission	5	1.8×10^{-3}
	Central nervous system development	3	3.1×10^{-3}
Factor 3 (168 genes)	Cell adhesion	19	10^{-5}
	Cell motility	8	9.4×10^{-5}
	Inflammatory response	8	4.7×10^{-3}
Factor 4 (64 genes)	Potassium ion transport	5	2.0×10^{-5}
	Synaptic transmission	6	1.8×10^{-4}
	Central nervous system development	3	7.6×10^{-4}
	Ion transport	6	1.3×10^{-3}

which produced the lowest approximation error for five runs with different random initializations of W. Only one sample (the 18th sample, Brain_MGlio_8) was incorrectly assigned to the third cluster (rhabdoids). Our gene selection method suggested total of 367 genes (cluster1: 42, cluster2: 93, cluster3: 168, cluster4: 64). To more thoroughly characterize sets of genes dominantly expressed in different factors, we used the Onto-Express. The number of genes corresponding to each GO category was compared with the number of genes expected for the GO category in the Affymetrix HuGeneFL array. Significant differences from the expected were calculated with hypergeometric distribution. Table 4 shows biological processes with a significance of P-value < 0.01. The biological processes showing significant representations in the first factor were serotonin receptor signaling pathway, feeding behavior, and central nervous system development. Serotonin (5-hydroxytryptamine, or 5-HT) is a monoamine neurotransmitter and is known to regulate human mood, emotion, sleep and appetite in the central nervous system. Two GenBank accession numbers (U49516 and M81778) for serotonin receptor signaling pathway were linked to the same gene: 5-hydroxytryptamine (serotonin) receptor 2C (HTR2C). The GO category of feeding behavior seems to be related with childhood brain tumors known as medulloblastomas. Genes involved in the second factor were (4 genes: U52155, M81886, M64752, M81181) for potassium ion transport, (5 genes: X54673, M81886, M64752, M19650, L32961) for synaptic transmission, (3 genes: U62801, Z19002, M93426) for central nervous system development. This second cluster contains malignant glioma that is a tumor arising from glial cells. Genes corresponding to cell adhesion, cell motility and inflammatory response were highly expressed in the third factor. Genes highly expressed in normal cerebella were (5 genes: U79245, U33632, U90065, L36069, D79998) for potassium ion transport, (6 genes: M13577, U92457, L76627, U18244, M58583, U79667) for synaptic transmission, (3 genes: U52969, M13577, U76421) for central nervous system development and (6 genes: S81944, U79245, S95936, U33632, U90065, L36069) for ion transport. Detailed description of the clusters of samples and genes selected for each factor via SNMF/R can be found in supplementary materials. We have shown that SNMF/R can be used for clustering, cancer class discovery, gene selection and biological process analysis.

4 CONCLUSION

We present a novel sparse NMF algorithm via alternating non-negativity-constrained least squares. SNMF/R can be used for cancer class discovery and gene expression data analysis since it shows good biclustering performance and provides us with simple interpretation. This algorithm can be applied to many practical problems in bioinformatics and computational biology such as biomedical text mining and gene/protein microarray data analysis.

ACKNOWLEDGEMENTS

We would like to thank Dr Chris Ding, Dr Jean-Philippe Brunet, Dr Yuan Gao, Prof. Lars Eldén and Prof. Robert J. Plemmons for their valuable comments. In particular, we would also like to thank Prof. Chih-Jen Lin, Prof. Paul Tseng and Prof. Luigi Grippo for discussions on the convergence property. This material is based upon work supported in part by the National Science Foundation Grants ACI-0305543 and CCF-0621889. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Conflict of Interest: none declared.

REFERENCES

Berry,M.W. *et al.* (1999) Matrices, vector spaces, and information retrieval. *SIAM Rev.*, **41**, 335–362.

Berry,M.W. *et al.* (2006) Algorithms and applications for approximate nonnegative matrix factorization. *Comput. Stat. Data Anal.*, (to appear).

Bro,R. and de Jong,S. (1997) A fast non-negativity-constrained least squares algorithm. *J. Chemometrics*, **11**, 393–401.

Brunet,J.P. *et al.* (2004) Metagenes and molecular pattern discovery using matrix factorization. *Proc. Natl Acad. Sci. USA*, **101**, 4164–4169.

Carmona-Saez,P. *et al.* (2006) Biclustering of gene expression data by non-smooth non-negative matrix factorization. *BMC Bioinformatics*, **7**, 78.

Carrasco,D.R. *et al.* (2006) High-resolution genomic profiles define distinct clinico-pathogenetic subgroups of multiple myeloma patients. *Cancer Cell*, **9**, 313–325.

Chagoyen,M. *et al.* (2006) Discovering semantic features in the literature: a foundation for building functional associations. *BMC Bioinformatics*, **7**, 41.

Ding,C. *et al.* (2002) Adaptive dimension reduction for clustering high dimensional data. In *Proceedings of the 2nd IEEE International Conference on Data Mining*. Maebashi, Japan.

Draghici,S. *et al.* (2003) Onto-tools, the toolkit of the modern biologist: Onto-Express, Onto-Compare, Onto-Design and Onto-Translate. *Nucleic Acids Res.*, **31**, 3775–3781.

Dueck,D. *et al.* (2005) Multi-way clustering of microarray data using probabilistic sparse matrix factorization. *Bioinformatics*, **21**(Suppl. 1), i144–i151.

Gao,Y. and Church,G. (2005) Improving molecular cancer class discovery through sparse non-negative matrix factorization. *Bioinformatics*, **21**, 3970–3975.

Golub,T.R. *et al.* (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531–537.

Gonzales,E.F. and Zhang,Y. (2005) Accelerating the Lee-Seung algorithm for non-negative matrix factorization. *Technical report*. Department of Computational and Applied Mathematics, Rice University.

Grippo,L. and Sciandrone,M. (2000) On the convergence of the block nonlinear Gauss-Seidel method under convex constraints. *Operations Res. Lett.*, **26**, 127–136.

Hoyer,P.O. (2004) Non-negative matrix factorization with sparseness constraints. *J. Machine Learning Res.*, **5**, 1457–1469.

Khatri,P. *et al.* (2002) Profiling gene expression using onto-express. *Genomics*, **79**, 266–270.

Kim,H. *et al.* (2005) Dimension reduction in text classification with support vector machines. *J. Machine Learning Res.*, **6**, 37–53.

Kim,P.M. and Tidor,B. (2003) Subsystem identification through dimensionality reduction of large-scale gene expression data. *Genome Res.*, **13**, 1706–1718.

Lawson,C.L. and Hanson,R.J. (1974) *Solving Least Squares Problems*, Prentice-Hall, Englewood Cliffs, NJ.

Lee,D.D. and Seung,H.S. (1999) Learning the parts of objects by non-negative matrix factorization. *Nature*, **401**, 788–791.

Lee,D.D. and Seung,H.S. (2000) Algorithms for non-negative matrix factorization. In *Proceedings of Neural Information Processing Systems*, pp. 556–562.

Li,S.Z. *et al.* (2001) Learning spatially localized parts-based representations. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 207–212.

Maher,E.A. *et al.* (2006) Marked genomic differences characterize primary and secondary glioblastoma subtypes and identify two distinct molecular and clinical secondary glioblastoma entities. *Cancer Res.*, **66**, 11502–11513.

MATLAB (1992) *User's Guide*, The MathWorks, Inc., Natick, MA 01760.

Pascual-Montano,A. *et al.* (2006) Nonsmooth nonnegative matrix factorization (nsNMF). *IEEE, Trans. Pattern Anal. Machine Intell.*, **28**, 403–415.

Pauca,V.P. *et al.* (2004) Text mining using non-negative matrix factorizations. In *Proceedings SIAM International Conference on Data Mining (SDM'04)*.

Pauca,V.P. *et al.* (2006) Nonnegative matrix factorization for spectral data analysis. *Linear Algebra and Applications* (to appear).

Pehkonen,P. *et al.* (2005) Theme discovery from gene lists for identification and viewing of multiple functional groups. *BMC Bioinformatics*, **6**, 162.

Pomeroy,S.L. *et al.* (2002) Prediction of central nervous system embryonal tumour outcome based on gene expression. *Nature*, **415**, 436–442.

Tibshirani,R. (1996) Regression shrinkage and selection via LASSO. *J. Roy. Statist. Soc. B*, **58**, 267–288.

van Benthem,M.H. and Keenan,M.R. (2004) Fast algorithm for the solution of large-scale non-negativity-constrained least squares problems. *J. Chemometrics*, **18**, 441–450.