A Class-Information-based SNMF Method for Selecting Characteristic Genes

Jin-Xing Liu
School of Information Science and Engineering
Qufu Normal University
Rizhao, China
Email: sdcavell@126.com

Chun-Xia Ma
School of Information Science and Engineering
Qufu Normal University
Rizhao, China
Email: mcxia87@126.com

Ying-Lian Gao
Library of Qufu Normal University
Qufu Normal University
Rizhao, China
Email: yinliangao@126.com

Jian Liu
School of Communication
Qufu Normal University
Rizhao, China
Email:liujiansqjxt@126.com

Chun-Hou Zheng
College of Electrical Engineering and Automation, Anhui University, Hefei, China
E-mail: zhengch99@126.com

***Abstract*—The significant advantage of sparse methods is to reduce the complicacy of genes expression data, which makes them easier to understand and interpret. In this paper, we propose a novel Class-information-based Sparse Non-negative Matrix Factorization (CISNMF) method which introduces the class information by the total scatter matrix. Firstly, the total scatter matrix is obtained via combining the between-class and within-class scatter matrices. Secondly, a new data matrix is constructed via singular values and left singular vectors which can be obtained via decomposing the total scatter matrix. Finally, we decompose the new data matrix by using sparse Non-negative Matrix Factorization and extract characteristic genes. In the end, results on gene expression data sets show that our method can extract more characteristic genes in response to abiotic stresses than conventional gene selection methods.**

Keywords—matrix factorization; scatter matrices; gene expression data; gene selection; abiotic stresses

I. INTRODUCTION

Environmental abiotic stresses have caused many unfavorable effects on plant growth, such as heat, osmotic stress. In order to reduce the negative impact of these environmental conditions, plants have evolved a variety of strategies and they can able to cope with these environmental conditions, including salt, cold, osmotic stress, uv-b light, drought and so on [1]. The fundamental concept is that they have some interacting genes responding to each abiotic stress. Therefore, it is one of the utmost significant topics how to comprehend the abiotic stresses responses in plant science [2].

Many conventional methods, such as RT-PCR [3] or Northern blotting [4, 5] have been used to study the genes responding to abiotic stresses. However, these methods have one defect that only a small part of genes can be studied at the same time. So, the microarray technology has been put forward to overcome this shortcoming. The technology makes it possible to monitor gene expression levels on a genomic scale [6]. With the booming of microarray technology, a large number of mathematical methods have been used to analyze gene expression data [7-16], such as, principal component analysis (PCA), independent component analysis (ICA) and singular value decomposition (SVD). In gene expression data analysis, PCA is an unsupervised method to search the useful eigenassay or eigengene [7]. ICA [10] is a useful extension of PCA. Huang et al. introduced a penalized discriminant method based on ICA for tumor classification [12]. Alter et al. put forward to use SVD for modeling and processing the gene expression data [17]. Although these methods have been widely used in gene expression data, they have some drawbacks. For example, they are not sparse, which makes it hard to interpret the expression data. Therefore, the corresponding sparse algorithms are proposed by researchers to overcome these drawbacks For example, Journée et al. proposed an SPCA method by using generalized power method [8]. Liu et al. proposed the Weighting Principal Components by Singular Values to extract characteristic genes [15]. In [14], Witten et al. proposed a penalized matrix decomposition, which was used to analyze plants gene expression data by Liu et al. [9, 18]. However, they have some common defects: They allow the negative component exists and need to standardize the original data.

To seek a better solution, Lee and Seung firstly introduced Non-negative Matrix Factorization (NMF) method to decompose image matrix in [19]. NMF decomposes a non-negative data matrix into two non-negative factors. Thereinto one matrix is called basis matrix, the other is defined as the coefficient matrix of the corresponding basis matrix. In addition, NMF usually involves some simple operations, so it has a lower computational cost. So far, many algorithms of NMF have been proposed [20, 21]. In addition, the corresponding sparse NMF algorithms have been proposed to give a reasonable sparse representation, such as sparse NMF [22], Fisher NMF [23] and Non-negative Matrix Factorization with Sparse Constraints (SNMF) [24]. NMF has been widely used in gene expression data analysis [25]. The sparse NMF and Fisher NMF have a common side-effect that the sparsity cannot be controlled. However, SNMF, which was first

This work was supported in part by the NSFC under grant Nos. 61370163 and 61373027; the China Postdoctoral Science Foundation funded project, No.2014M560264; the Shandong Provincial Natural Science Foundation, under grant Nos. ZR2013FL016 and ZR2012FM023; Shenzhen Municipal Science and Technology Innovation Council (Nos.JCYJ20130329151843309 and JCYJ20140417172417174).

978-1-4799-7294-4/14/$31.00 ©2014 IEEE

introduced by Patrik O. Hoyer [24], can control the sparsity accurately. SNMF has been applied to images processing [26], genes selection [27, 28] and so on. In order to improve the analytical performance of gene expression data, we put forward a Class-Information-based Sparse Non-negative Matrix Factorization (CISNMF) method, which introduces the class information by using the total scatter matrix. The scheme of CISNMF is given as follows: Firstly, the total scatter matrix is obtained by combining the within-class and between-class scatter matrices; Secondly, we decompose the total scatter matrix by using singular value decomposition (SVD) and construct a new data matrix by singular values and left singular vectors; Thirdly, we decompose the new data matrix by a SNMF and extract the genes according to the sparse loading vectors.

This paper is structured as follows. In section II, we describe the methodology of CISNMF. Section III provides experimental results and discussion. Section IV concludes the paper.

II. METHODS

A. The Mathematical Definition of Scatter Matrices

On the basis of a similarity measure, the pending classification pattern set $\{\vec{x}_1, \vec{x}_2, \cdots, \vec{x}_n\}$ is divided into c categories. The classified pattern denoted as $\left\{\vec{x}_i^{(j)}; j = 1, 2, \cdots, c; i = 1, 2, \cdots, n_j\right\}$. Three matrices are defined as: the within-class scatter matrix $\mathbf{J}_w$, the between-class matrix $\mathbf{J}_b$ and the total scatter matrix $\mathbf{J}_t$. For all the samples of all classes, the three scatter matrices can be written as follows:

1) The first is called within-class scatter matrix $\mathbf{J}_w$ which is described as:

$$\mathbf{J}_w = \sum_{j=1}^{c} \sum_{i=1}^{n_j} (\mathbf{x}_i^j - z_j)(\mathbf{x}_i^j - z_j)^T, \tag{1}$$

where $\mathbf{x}_i^j$ is the i-th sample in class j; n_j is the number of sample in class j, z_j is the mean of class j, c is the number of classes.

2) The next is called between-class scatter matrix $\mathbf{J}_b$ which is given by

$$\mathbf{J}_b = \sum_{j=1}^{c} n_j \left(z_j - z\right)\left(z_j - z\right)^T, \tag{2}$$

where z is the mean of all classes.

3) The goal of clustering makes the within class distance as small as possible, the distance between classes as bigger as possible. So, the total scatter matrix $\mathbf{J}_t$ can be denoted by

$$\mathbf{J}_t = \mathbf{J}_b - \xi \mathbf{J}_w, \tag{3}$$

where $\xi \geq 0$ is an adjustable parameter that stems from a trade-off between $\mathbf{J}_b$ and $\mathbf{J}_w$.

The aim of clustering is: $Tr[\mathbf{J}_b] \Rightarrow \max$ or/and $Tr[\mathbf{J}_w] \Rightarrow \min$ and the traces of scatter matrices could measure the between-class and within-class distances. They are written as follows:

$$\begin{aligned} Tr(\mathbf{J}_w) &= Tr\left[\sum_{j=1}^{c}\sum_{i=1}^{n_j}\left(\mathbf{x}_i^j - z_j\right)\left(\mathbf{x}_i^j - z_j\right)^T\right] \\ &= \lambda_{w1} + \lambda_{w2} + \cdots + \lambda_{wk}. \end{aligned} \tag{4}$$

$$\begin{aligned} Tr(\mathbf{J}_b) &= Tr\left[\sum_{j=1}^{c} n_j \left(z_j - z\right)\left(z_j - z\right)^T\right] \\ &= \lambda_{b1} + \lambda_{b2} + \cdots + \lambda_{bk}. \end{aligned} \tag{5}$$

Here, $Tr(\mathbf{J}_w)$ is used to measure the close degree of the samples within the classes. While $Tr(\mathbf{J}_b)$ is used to measure the degree of separation between the classes. Hence, the adjustable parameter ξ in (3) can be written as follows [29]:

$$\xi = \frac{Tr(\mathbf{J}_b)}{Tr(\mathbf{J}_w)} \tag{6}$$

B. Mathematical Definition of CISNMF

In order to extract the characteristic genes effectively, we introduce a supervised learning method. The new data matrix is obtained by the total scatter matrix $\mathbf{J}_t$. Then the new data matrix is decomposed into two non-negative entries by SNMF.

Firstly, the scatter matrix $\mathbf{J}_t$ is decomposed by using SVD. It can be written as follows:

$$\mathbf{J}_t = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^T, \tag{7}$$

where $\mathbf{\Lambda} = diag(\Lambda_1, \Lambda_2, \cdots, \Lambda_r)$ is a diagonal matrix that consists of singular values and r is the rank of $\mathbf{J}_t$. $\mathbf{U}$ and $\mathbf{V}$ are orthogonal feature vectors.

Secondly, for reducing the computational complexity, we construct a new matrix, and the new data matrix is constructed as follows:

$$\mathbf{Q} = \mathbf{U}\mathbf{\Lambda}^{1/2}. \tag{8}$$

978-1-4799-7294-4/14/$31.00 ©2014 IEEE

Finally, $\mathbf{L}$ is the transpose of $\mathbf{Q}$, and $\mathbf{L}$ is decomposed by using Sparse Non-negative Matrix Factorization (SNMF).

$$\mathbf{L} \sim \mathbf{FP}, \tag{9}$$

where $\mathbf{L}$ is an $m \times n$ non-negative matrix, $\mathbf{F}$ is an $m \times k$ non-negative matrix, $\mathbf{P}$ is a $k \times n$ non-negative matrix and $k < min(m,n)$.

The optimization problem can be described as the following:

$$\underset{\mathbf{F},\mathbf{P}}{minimize} \left\| \mathbf{L} - \mathbf{FP} \right\|^2 = \underset{\mathbf{F},\mathbf{P}}{minimize} \sum_{ij} (\mathbf{L}_{ij} - (\mathbf{FP})_{ij})^2, \forall \mathbf{F}, \mathbf{P} > 0, \tag{10}$$

s. t. optional constraints:

$$sparseness(\mathbf{F}_i) = \varphi, \forall \mathrm{i} \tag{11}$$

$$sparseness(\mathbf{P}_i) = \gamma, \forall \mathrm{i} \tag{12}$$

where $\mathbf{F}_i$ is the i-th column of $\mathbf{F}$, $\mathbf{P}_i$ is the i-th row of $\mathbf{P}$.

C. The Algorithm

The details of the CISNMF algorithm are listed as follows:

1) The total scatter matrix $\mathbf{J}_t$ is obtained via (3).

2) The $\mathbf{U}$, $\mathbf{\Lambda}$ and $\mathbf{V}$ are obtained via decomposing $\mathbf{J}_t$ by SVD in (7).

3) Construct a new data matrix $\mathbf{Q}$ according to (8).

4) Transpose $\mathbf{Q}$ as $\mathbf{L}$, and $\mathbf{L}$ is decomposed into $\mathbf{F}$ and $\mathbf{P}$ by SNMF.

5) Initialize $\mathbf{F}$ and $\mathbf{P}$ to random positive matrices.

6) Iterate until convergence or reach the largest number of iteration.

a) Sparse constraints on $\mathbf{P}$,

$$\mathbf{P} \coloneqq \mathbf{P} - \delta_p \mathbf{F}^T (\mathbf{FP} - \mathbf{L})$$

$$\mathbf{F} \coloneqq \mathbf{F} \otimes \left(\mathbf{LP}^T \right) / \left(\mathbf{FPP}^T \right)$$

b) Sparse constraints on $\mathbf{F}$,

$$\mathbf{F} \coloneqq \mathbf{F} - \delta_f \left(\mathbf{FP} - \mathbf{L} \right) \mathbf{P}^T$$

$$\mathbf{P} \coloneqq \mathbf{P} \otimes \left(\mathbf{F}^T \mathbf{L} \right) / \left(\mathbf{F}^T \mathbf{FP} \right)$$

In this algorithm, $\xi \geq 0$ is an adjustable parameter that stems from a trade-off between $\mathbf{J}_b$ and $\mathbf{J}_w$. The sparseness of $\mathbf{F}$ (φ) and $\mathbf{P}$ (γ) are in the range between (0, 1). δ_f and δ_p are small positive constants (stepsizes) and the two parameters need not be set by the user.

D. Extracting Characteristic Genes by CISNMF

In the research, our goal is to gain the characteristic genes responding to the abiotic stresses. Here, we transpose the gene expression data matrix $\mathbf{Q}$ and defined it as $\mathbf{L}$. Hence, $\mathbf{L} = \mathbf{Q}^T$ and the size of $\mathbf{L}$ is $m \times n$, rows represents the expression level of the n genes in m samples, each column of $\mathbf{L}$ represents the expression level of a gene across all samples. So, the $\mathbf{L}$ can be written as: $\mathbf{L} \sim \mathbf{FP}$, where $\mathbf{F}$ is an $m \times k$ non-negative matrix, $\mathbf{P}$ is a $k \times n$ non-negative matrix and $k < min(m,n)$.

The optimization problem is convex in $\mathbf{F}$ and $\mathbf{P}$ separately and minimize the reconstruction error between $\mathbf{L}$ and $\mathbf{FP}$. Various error functions have been testified in [30], and the squared error (Euclidean distance) function is given as follows:

$$\left\| \mathbf{L} - \mathbf{FP} \right\|^2 = \sum_{ji} (\mathbf{L}_{ji} - (\mathbf{FP})_{ji})^2, \forall \mathbf{F}, \mathbf{P} > 0. \tag{13}$$

The sample expression profile $\mathbf{l}_j$ by (9) can be denoted as:

$$\mathbf{l}_j = \sum_{i=1}^{k} \mathbf{f}_{ji} \mathbf{p}_i, j = 1,2,\cdots,m. \tag{14}$$

Here $\mathbf{l}_j$ is a linear combination of the metasamples $\{\mathbf{p}_i\}$ and $\mathbf{l}_j$ is the row of $\mathbf{L}$, $\mathbf{f}_{ji}$ is the entry of $\mathbf{F}$. Here, we can view the rows of $\mathbf{F}$ as the encoding coefficients and the k rows of $\mathbf{P}$ as basis vectors (metasamples). The data matrix $\mathbf{L}$, coefficients matrix $\mathbf{F}$ and basis matrix $\mathbf{P}$ are shown in Fig. 1.

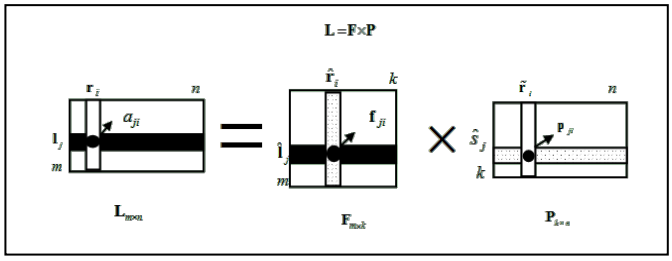

Fig. 1. The graph description of the matrix $\mathbf{L}$ with the factors $\mathbf{F}$ and $\mathbf{P}$.

In Fig. 1, $\mathbf{l}_j$ is the samples characteristics of the matrix $\mathbf{L}$, $\mathbf{r}_i$ represents the feature vectors of $\mathbf{L}$, a_{ji} shows the expression level of the i-th gene in the j-th sample. $\hat{\mathbf{l}}_j$ is the j-th eigensamples of $\mathbf{F}$, $\hat{\mathbf{r}}_i$ is the column vector of $\mathbf{F}$ and

978-1-4799-7294-4/14/$31.00 ©2014 IEEE

indicates the i-th vector in k genes of $\mathbf{F}$. $\tilde{\mathbf{l}}_j$ and $\tilde{\mathbf{r}}_i$ refer to the j-th sample vector and i-th feature of $\mathbf{P}$ which consists of n genes in k samples. In order to reduce the dimension of $\mathbf{L}$, we choose part of the sample characteristics to replace $\mathbf{L}$. Due to the matrix $\mathbf{P}$ contains all genes and it is one subset of metasamples of $\mathbf{L}$, the matrix $\mathbf{P}$ is called the basis vectors. Hence, we can extract characteristic genes from the basis matrix $\mathbf{P}$. So, $\mathbf{l}_j$ can be replaced by $\mathbf{p}_i$. By controlling the parameters of SNMF, the sparse matrix $\mathbf{P}$ can be gained. So, the characteristic genes can be extracted from the non-zero entries in the matrix $\mathbf{P}$.

In the end, we summarize in getting characteristic genes via the CISNMF method as the following:

- Gain the total scatter matrix $\mathbf{J}_t$.
- Decompose the scatter matrix $\mathbf{J}_t$ by SVD.
- Gain a new matrix $\mathbf{Q}$ via executing the SVD, and transpose $\mathbf{Q}$ into $\mathbf{L}$.
- Obtain the matrix $\mathbf{P}$ according to SNMF.
- Extract the characteristic genes via the non-zero entries in $\mathbf{P}$.
- Exploit the GO to check the extracted characteristic genes.

III. Result and Discussion

In this section, we will show the results of exploiting CISNMF method. In this section, the results on gene expression data sets are given. Our method will compared with SNMF [24], SPCA [8] and PMD [9] methods in this section.

A. Data Source

The gene expression data are download from the NASCArrays [http://affy.arabidopsis.info/], reference numbers are: NASCArrays-141, drought stress; NASCArrays-140, salt stress; NASCArrays-144, uv-b light stress; NASCArrays-138, cold stress; NASCArrays-146, heat stress; NASCArrays-139, osmotic stress; NASCArrays-137, control [31]. Here, each sample contains 22810 genes and the sample numbers of each stress are listed in Table I.

TABLE I. The Number of Each Stress Type in the Data Set

stress type	drought	salt	uv-b	cold	heat	osmotic	control
number	7	6	7	6	8	6	8

The background light noise of these data can be adjusted by using the GC-RMA method which was proposed by Wu et al. [32]. The GC-RMA results are collected in a matrix to be further processing. In this paper, the two labels are selected by two data sets (except the control sets) to construct the matrix $\mathbf{J}_t$. For drought set in root, we assign the drought samples to the first class and the other 11 samples as the second class. We use SNMF to process these data, and the extracted genes are verified by GO tools.

B. Selection of the Parameters

In [27, 28] the best results are obtained when the sparseness controlling parameter φ is set to 0.5. So in our experiment, parameter φ is set to 0.5 and the adjusted parameter γ is controlled in range (0-1) [24]. For comparison, 500 genes are selected by CISNMF, SNMF, PMD and SPCA methods.

C. Gene Ontology (GO) Analysis

Terms are the basic concept of Gene Ontology (GO). Each entry in GO has a unique digital label. The Gene Ontology term enrichment tool, including meaningful shared GO terms, can search those genes that may have in common [33]. The analysis of GO Term Finder is modular, which offers valuable information of high-throughput experiments in biological science field. In this research, our method will be evaluated by GO TermFinder, which is freely used at <http://go.princeton.edu/cgi-bin/GOTermFinder> [34]. The threshold parameters are set as listed below: minimum number of gene products is set to 2 and maximum p-value is set to 0.01.

TABLE II. Response to Stimulus (GO: 0050896) in Root Samples

stress	**CISNMF**		**SNMF**		**PMD**		**SPCA**	
type	**PV**	**SF**	**PV**	**SF**	**PV**	**SF**	**PV**	**SF**
drought	1.36E-84	313*, 62.7%	2.39E-88	**318, 63.6%**	3.67E-65	287, 57.4%	1.39E-66	289, 57.8%
salt	1.39E-79	307, 61.4%	5.54E-95	**326, 65.2%**	1.19E-84	313, 62.6%	1.42E-34	237, 47.4%
uv-b	1.59E-65	**288, 57.6%**	3.81E-64	286, 57.2%	8.78E-38	243, 48.6%	9.56E-22	210, 42.0%
cold	9.18E-84	**312, 62.5%**	3.79E-81	309, 61.8%	5.06E-68	291, 58.2%	1.54E-61	281, 56.4%
heat	3.06E-40	**248, 49.6%**	3.71E-21	209, 41.8%	1.18E-19	205, 41.0%	1.23E-17	200, 40.0%
osmotic	4.59E-89	**319, 63.8%**	1.65E-47	260, 52.1%	1.1E-26	221, 44.2%	1.49E-34	237, 47.4%

In this table, the background frequency of response to stimulus (GO: 0050896) in TAIR set is 21.8% (6619/30324). And in the sample frequency, 313* denotes having 313 genes to response to stimulus in the 500 selection genes. PV: p-value, and SF: sample frequency.

In root sample, the response to stimulus (GO: 0050896) is listed in Table II. In TAIR set, the corresponding background frequency is 21.8% (6619/30324). In this experiment, 500 genes are selected by CISNMF, SNMF, PMD and SPCA methods. Table II lists the p-value and sample frequency of various stresses. In our method, 313 genes for drought stress are extracted and the sample frequency is 62.7%, 307 genes for salt stress are extracted (61.4%), 288 genes for uv-b stress are extracted (57.6%), 312 genes for cold stress are extracted (62.5%), 248 genes for heat stress are extracted (49.6%), and 319 genes for osmotic stress are extracted (63.8%). While in SNMF method, 318 genes for drought stress are extracted (63.6%), 326 genes for salt stress are extracted (65.2%), 286 genes for uv-b stress are extracted (57.2%), 309 genes for cold stress are extracted (61.8%), 209 genes for heat stress are extracted (41.8%) and 260 genes for osmotic stress are extracted (52.1%). In Table II, only two of these stresses

978-1-4799-7294-4/14/$31.00 ©2014 IEEE

(drought stress and salt stress) that SNMF method is superior to our method. In other stresses, our method outperforms SNMF. Obviously, the bold fonts in Table II show that our method is far superior to PMD method and SPCA method. From Fig. 2, we can see that our method surpasses other methods.

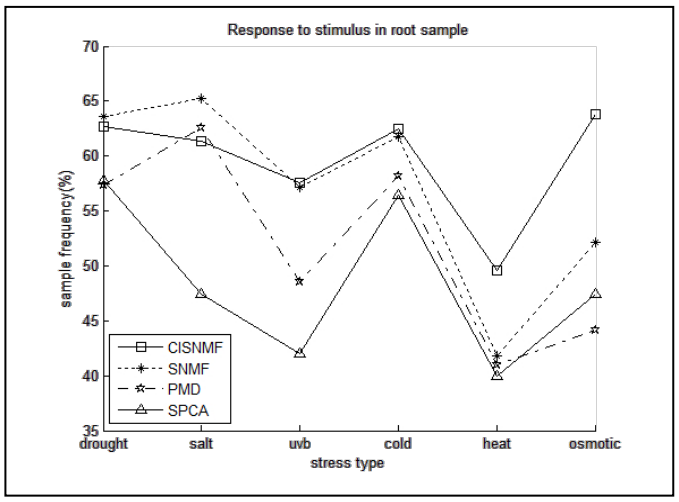

Fig. 2. The response to stimulus (GO: 0050896) in root samples.

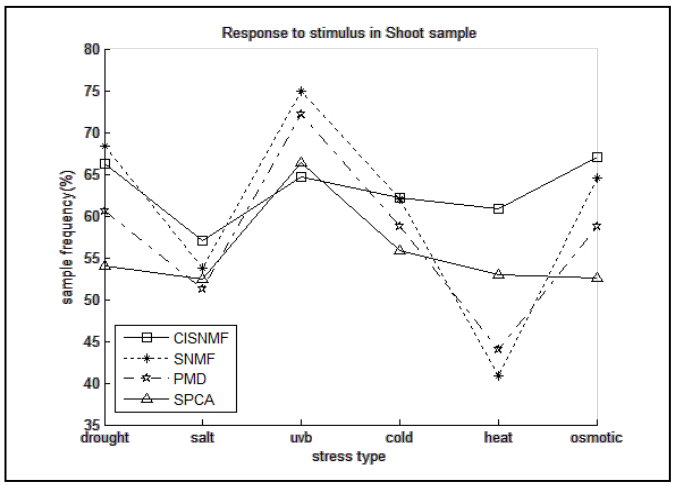

Fig. 3. The response to stimulus (GO: 0050896) in shoot samples.

TABLE III. RESPONSE TO STIMULUS (GO: 0050896) IN SHOOT SAMPLES

stress	CISNMF		SNMF		PMD		SPCA	
type	PV	SF	PV	SF	PV	SF	PV	SF
drought	3.46E-99	331, 66.2%	5.28E-109	**342, 68.4%**	8.09E-77	303, 60.6%	5.45E-54	270, 54.0%
salt	1.81E-63	**285, 57.0%**	2.01E-52	268, 53.7%	1.43E-45	256, 51.2%	9.59E-49	262, 52.4%
uv-b	1.01E-92	323, 64.7%	2.79E-141	**375, 75.0%**	6.4E-128	361, 72.2%	3.12E-101	332, 66.4%
cold	1.6E-82	**311, 62.2%**	1.04E-82	310, 62.0%	1.82E-70	294, 58.8%	3.54E-60	279, 55.8%
heat	1.34E-76	**302, 60.8%**	4.28E-19	204, 40.8%	3.69E-26	220, 44.0%	8.04E-51	265, 53.0%
osmotic	4.72E-103	**335, 67.1%**	1.97E-92	323, 64.6%	3.67E-70	294, 58.8%	2.3E-49	263, 52.6%

In shoot samples, the sample frequencies of response to stimulus (GO: 0050896) are shown in Fig. 3. It can be seen that only in drought stress and uv-b stress data points, SNMF method is superior to our method. In the four remaining data points, our method is better than SNMF method. The specific results are listed in Table III. In addition, in the six data sets, our method has priority over PMD method and SPCA method.

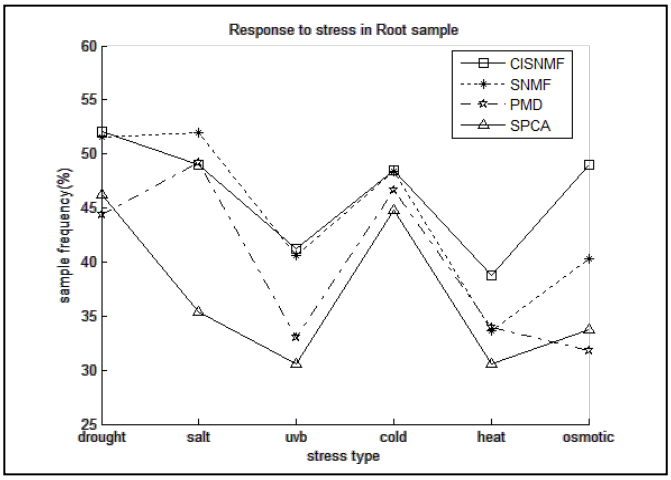

Fig. 4. The response to stress (GO: 0006950) in root samples.

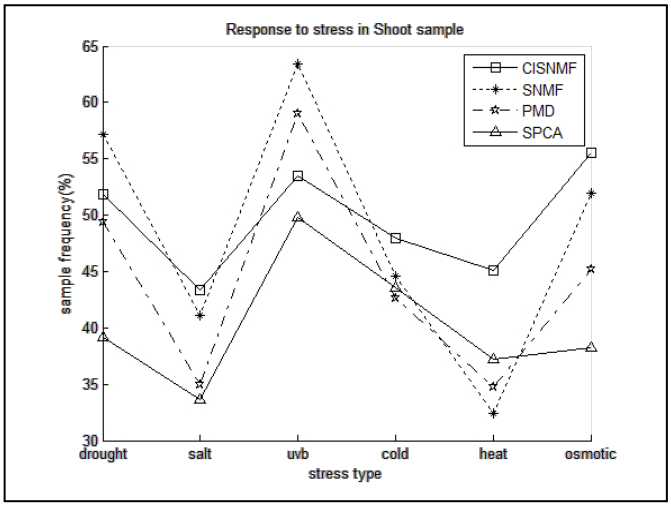

Fig. 5. The response to stress (GO: 0006950) in shoot samples.

Fig. 4 and 5 depict the corresponding sample frequency of response to stress (GO: 0006950) in root and shoot samples, respectively. As shown in Fig. 4, the salt stress point of SNMF method is higher than ours. In other data points, our method outperforms other methods. Fig. 5 shows that only in drought stress and uv-b stress data points, SNMF method gets ahead of our method. In the residuary data points, our method has an advantage over other methods.

The numbers we select and p-value of response to stress (GO: 0006950) in root and shoot samples, respectively, are listed in Table IV and Table V. In TAIR set, the background frequency of response to stress is 13.3% (4028/30324). In Table IV, our method is superior to other four methods except the salt stress data set. SNMF method excels our method in the drought stress and uv-b stress data sets, and the detailed

978-1-4799-7294-4/14/$31.00 ©2014 IEEE

contents are listed in Table V. In other data sets, our method outperforms other methods.

TABLE IV. RESPONSE TO STRESS (GO: 0006950) IN ROOT SAMPLES

stress	CISNMF		SNMF		PMD		SPCA	
type	PV	SF	PV	SF	PV	SF	PV	SF
drought	2.82E-94	**260, 52.1%**	2.5E-92	258, 51.6%	9.89E-64	222, 44.4%	1.49E-70	231, 46.2%
salt	2.57E-81	245, 49.0%	**4.57E-94**	**260, 52.0%**	1.48E-82	246, 49.2%	3.65E-34	177, 35.4%
uv-b	5.27E-52	**206, 41.2%**	5.29E-50	203, 40.6%	1.36E-27	165, 33.0%	1.38E-21	153, 30.6%
cold	4.79E-79	**242, 48.5%**	7E-79	242, 48.4%	4.03E-72	233, 46.6%	7.11E-65	223, 44.8%
heat	4.3E-44	**194, 38.8%**	5.91E-29	168, 33.6%	2.98E-30	170, 34.0%	1.47E-21	153, 30.6%
osmotic	2.99E-81	**245, 49.0%**	7.28E-49	201, 40.3%	1.55E-24	159, 31.8%	1.07E-29	169, 33.8%

TABLE V. RESPONSE TO STRESS (GO: 0006950) IN SHOOT SAMPLES

stress	CISNMF		SNMF		PMD		SPCA	
type	PV	SF	PV	SF	PV	SF	PV	SF
drought	4.03E-93	259, 51.8%	5.03E-118	**286, 57.2%**	2.43E-83	247, 49.4%	7.15E-46	196, 39.2%
salt	9.73E-60	**217, 43.4%**	1.66E-51	205, 41.1%	2.77E-33	175, 35.0%	3.66E-29	168, 33.6%
uv-b	1.84E-100	267, 53.5%	8.29E-150	**317, 63.4%**	9.65E-128	295, 59.0%	1.56E-85	249, 49.8%
cold	7.2E-77	**239, 48.0%**	3.96E-64	223, 44.6%	2.01E-57	213, 42.6%	3.48E-61	218, 43.6%
heat	2.12E-65	**224,45.1%**	8.05E-26	162, 32.4%	1.97E-32	174, 34.8%	1.34E-39	186, 37.2%
osmotic	1.12E-109	**277, 55.5%**	4.94E-94	260, 52.0%	1E-66	226, 45.2%	1.52E-42	191, 38.2%

TABLE VI. RESPONSE TO WATER DEPRIVATION (GO: 0009414) IN ROOT SAMPLES

stress	CISNMF		SNMF		PMD		SPCA	
type	PV	SF	PV	SF	PV	SF	PV	SF
drought	2.6E-39	**64, 12.8%**	4.87E-35	60, 12.0%	5.4E-26	51, 10.2%	1.54E-19	44, 8.8%

To sum up, our method gets ahead of others. In order to further study, the drought stress data set responding to water deprivation (GO: 0009414) in root samples is analyzed in Table VI. The background frequency of response to water deprivation is 1.4%. This table lists the numbers of response to water deprivation and p-value by the four methods, Moreover, the neglected characteristic genes by other methods are listed in Table VII. The literatures of those characteristic genes and the authors of these literatures are noted in the Table. All these genes are relevant to drought stress, and some are relevant to cold or salt and / or osmotic stress. From Table VI, we can see that CISNMF method extracts more characteristic genes than other methods.

TABLE VII. REFERENCES ABOUT CHARACTERISTIC GENES RESPONSE TO WATER DEPRIVATION (GO: 0006950) IN ROOT SAMPLES

Gene name	Response to	References
At2g33380	Drought, cold	Heyndrickx KS, et al. (2012) [35]
At3g45140	Drought	Bell, et al. (1993) [36]
At4g34390	Drought	Heyndrickx KS, et al. (2012) [35]
At2g46680	Drought, cold	Heyndrickx KS, et al. (2012) [35]
At5g27520	Drought	Heyndrickx KS, et al. (2012) [35]
At2g42530	Drought	Heyndrickx KS, et al. (2012) [35]
At5g62470	Drought	Seo, et al. (2009) [37]
At4g26070	Drought	Xing, et al. (2008) [38]
At3g19970	Drought	Heyndrickx KS, et al. (2012) [35]
At5g54490	Drought	Heyndrickx KS, et al. (2012) [35]
At5g27420	Drought	Heyndrickx KS, et al. (2012) [35]
At3g63060	Drought, salt, osmotic	Koops, et al. (2011) [39]
At4g36990	Drought	Heyndrickx KS, et al. (2012) [35]
At4g21440	Drought	Heyndrickx KS, et al. (2012) [35]
At1g73480	Drought, cold	Heyndrickx KS, et al. (2012) [35]
At5g67340	Drought, cold	Heyndrickx KS, et al. (2012) [35]
At4g17500	Drought	Heyndrickx KS, et al. (2012) [35]
At2g41430	Drought	Kiyosue, et al. (1994) [40]
At3g52400	Drought, cold	Heyndrickx KS, et al. (2012) [35]
At4g05100	Drought	Heyndrickx KS, et al. (2012) [35]
At2g46270	Drought, salt	Heyndrickx KS, et al. (2012) [35]
At4g26080	Drought, cold	Heyndrickx KS, et al. (2012) [35]
At5g57050	Drought, cold	Heyndrickx KS, et al. (2012) [35]
At3g57530	Drought, cold	Heyndrickx KS, et al. (2012) [35]
At4g34710	Drought	Heyndrickx KS, et al. (2012) [35]

In a word, these experiments and analyses show our method can extract more genes than other methods. Therefore, our method has more advantages of extracting characteristic genes than other methods.

IV. CONCLUSIONS

A new method (CISNMF) is proposed to extract genes in this paper. CISNMF method introduces the classification information by scatter matrix, so it can get more comprehensible and interpretable results. The extracted characteristic genes are analyzed by GO tools. For gene expression data, CISNMF can extract more characteristic genes than other methods. The experiments demonstrate that our method is effective and suitable for selecting characteristic genes.

In future, we will focus on the biological interpretation of the characteristic genes.

REFERENCES

[1] G. J. Allen, S. P. Chu, K. Schumacher, C. T. Shimazaki, D. Vafeados, A. Kemper, et al., "Alteration of stimulus-specific guard cell calcium oscillations and stomatal closing in Arabidopsis det3 mutant," Science, vol. 289, pp. 2338-2342, 2000.

[2] T. Hirayama and K. Shinozaki, "Research on plant abiotic stress responses in the post‐genome era: past, present and future," The Plant Journal, vol. 61, pp. 1041-1052, 2010.

[3] N. S. Maan, S. Maan, K. Nomikou, D. J. Johnson, M. El Harrak, H. Madani, et al., "RT-PCR assays for seven serotypes of epizootic haemorrhagic disease virus & their use to type strains from the mediterranean region and North America," PLoS One, vol. 5, p. e12782, 2010.

[4] T. Blevins, "Northern blotting techniques for small RNAs," in Plant Epigenetics, ed: Springer, 2010, pp. 87-107.

[5] K. Josefsen and H. Nielsen, "Northern blotting analysis," in RNA, ed: Springer, 2011, pp. 87-105.

[6] M. Seki, M. Narusaka, J. Ishida, T. Nanjo, M. Fujita, Y. Oono, et al., "Monitoring the expression profiles of 7000 Arabidopsis genes under drought, cold and high‐salinity stresses using a full‐length cDNA microarray," The Plant Journal, vol. 31, pp. 279-292, 2002.

[7] O. Alter, P. O. Brown, and D. Botstein, "Singular value decomposition for genome-wide expression data processing and modeling," Proceedings of the National Academy of Sciences, vol. 97, pp. 10101-10106, 2000.

[8] M. Journée, Y. Nesterov, P. Richtárik, and R. Sepulchre, "Generalized power method for sparse principal component analysis," The Journal of Machine Learning Research, vol. 11, pp. 517-553, 2010.

978-1-4799-7294-4/14/$31.00 ©2014 IEEE

[9] J.-X. Liu, C.-H. Zheng, and Y. Xu, "Extracting plants core genes responding to abiotic stresses by penalized matrix decomposition," Computers in Biology and Medicine, vol. 42, pp. 582-589, 2012.
[10] M. S. Bartlett, J. R. Movellan, and T. J. Sejnowski, "Face recognition by independent component analysis," Neural Networks, IEEE Transactions on, vol. 13, pp. 1450-1464, 2002.
[11] J.-X. Liu, J. Liu, Y.-L. Gao, J.-X. Mi, C.-X. Ma, and D. Wang, "A Class-Information-Based Penalized Matrix Decomposition for Identifying Plants Core Genes Responding to Abiotic Stresses," PloS ONE, vol. 9, p. e106097, 2014.
[12] D.-S. Huang and C.-H. Zheng, "Independent component analysis-based penalized discriminant method for tumor classification using gene expression data," Bioinformatics, vol. 22, pp. 1855-1862, 2006.
[13] J. Liu, C. Zheng, and Y. Xu, "Lasso logistic regression based approach for extracting plants coregenes responding to abiotic stresses," in Advanced Computational Intelligence (IWACI), 2011 Fourth International Workshop on, 2011, pp. 461-464.
[14] D. M. Witten, R. Tibshirani, and T. Hastie, "A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis," Biostatistics, vol. 10, pp. 515-534, 2009.
[15] J.-X. Liu, Y. Xu, C.-H. Zheng, Y. Wang, and J.-Y. Yang, "Characteristic Gene Selection via Weighting Principal Components by Singular Values," Plos One, vol. 7, p. e38873, 2012.
[16] J.-X. Liu, Y.-T. Wang, C.-H. Zheng, W. Sha, J.-X. Mi, and Y. Xu, "Robust PCA based method for discovering differentially expressed genes," BMC bioinformatics, vol. 14, pp. 1-10, 2013.
[17] V. M. Aradhya, F. Masulli, and S. Rovetta, "A novel approach for biclustering gene expression data using modular singular value decomposition," in Computational Intelligence Methods for Bioinformatics and Biostatistics, ed: Springer, 2010, pp. 254-265.
[18] J.-X. Liu, Y.-L. Gao, Y. Xu, C.-H. Zheng, and J. You, "Differential Expression Analysis on RNA-Seq Count Data Based on Penalized Matrix Decomposition," IEEE Transactions on NanoBioscience, vol. 13, pp. 12-18, 2014.
[19] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," Nature, vol. 401, pp. 788-791, 1999.
[20] Y. Li and A. Ngom, "The non-negative matrix factorization toolbox for biological data mining," Source code for biology and medicine, vol. 8, pp. 1-15, 2013.
[21] S. Yang and M. Ye, "Global Minima Analysis of Lee and Seung's NMF Algorithms," Neural processing letters, vol. 38, pp. 29-51, 2013.
[22] Y. Gao and G. Church, "Improving molecular cancer class discovery through sparse non-negative matrix factorization," Bioinformatics, vol. 21, pp. 3970-3975, 2005.
[23] Y. Wang and Y. Jia, "Fisher non-negative matrix factorization for learning local features," in In Proc. Asian Conf. on Comp. Vision, 2004.
[24] P. O. Hoyer, "Non-negative matrix factorization with sparseness constraints," The Journal of Machine Learning Research, vol. 5, pp. 1457-1469, 2004.
[25] C.-H. Zheng, T.-Y. Ng, D. Zhang, C.-K. Shiu, and H.-Q. Wang, "Tumor classification based on non-negative matrix factorization using gene expression data," NanoBioscience, IEEE Transactions on, vol. 10, pp. 86-93, 2011.
[26] J.-X. Du, C.-M. Zhai, and Y.-Q. Ye, "Face aging simulation based on NMF algorithm with sparseness constraints," in Advanced Intelligent Computing Theories and Applications. With Aspects of Artificial Intelligence, ed: Springer, 2012, pp. 516-522.
[27] X. Kong, C. Zheng, Y. Wu, and L. Shang, "Molecular cancer class discovery using non-negative matrix factorization with sparseness constraint," in Advanced Intelligent Computing Theories and Applications. With Aspects of Theoretical and Methodological Issues, ed: Springer, 2007, pp. 792-802.
[28] Z. Tang and S. Ding, "Nonnegative dictionary learning by nonnegative matrix factorization with a sparsity constraint," in Advances in Neural Networks–ISNN 2012, ed: Springer, 2012, pp. 92-101.
[29] B. Scholkopft and K.-R. Mullert, "Fisher discriminant analysis with kernels," 1999.
[30] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in Advances in neural information processing systems, 2000, pp. 556-562.
[31] D. J. Craigon, N. James, J. Okyere, J. Higgins, J. Jotham, and S. May, "NASCArrays: a repository for microarray data generated by NASC's transcriptomics service," Nucleic acids research, vol. 32, pp. D575-D577, 2004.
[32] Z. Wu, R. A. Irizarry, R. Gentleman, F. M. Murillo, and F. Spencer, "A model based background adjustment for oligonucleotide expression arrays," 2004.
[33] G. O. Consortium, "The Gene Ontology in 2010: extensions and refinements," Nucleic acids research, vol. 38, pp. D331-D335, 2010.
[34] E. I. Boyle, S. Weng, J. Gollub, H. Jin, D. Botstein, J. M. Cherry, et al., "GO:: TermFinder—open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes," Bioinformatics, vol. 20, pp. 3710-3715, 2004.
[35] K. S. Heyndrickx and K. Vandepoele, "Systematic identification of functional plant modules through the integration of complementary data sources," Plant physiology, vol. 159, pp. 884-901, 2012.
[36] E. Bell and J. E. Mullet, "Characterization of an Arabidopsis lipoxygenase gene responsive to methyl jasmonate and wounding," Plant Physiology, vol. 103, pp. 1133-1137, 1993.
[37] P. J. Seo, F. Xiang, M. Qiao, J.-Y. Park, Y. N. Lee, S.-G. Kim, et al., "The MYB96 transcription factor mediates abscisic acid signaling during drought stress response in Arabidopsis," Plant Physiology, vol. 151, pp. 275-289, 2009.
[38] Y. Xing, W. Jia, and J. Zhang, "AtMKK1 mediates ABA‐induced CAT1 expression and H2O2 production via AtMPK6‐coupled signaling in Arabidopsis," The Plant Journal, vol. 54, pp. 440-451, 2008.
[39] P. Koops, S. Pelser, M. Ignatz, C. Klose, K. Marrocco-Selden, and T. Kretsch, "EDL3 is an F-box protein involved in the regulation of abscisic acid signalling in Arabidopsis thaliana," Journal of experimental botany, vol. 62, pp. 5547-5560, 2011.
[40] T. Kiyosue, K. Yamaguchi-Shinozaki, and K. Shinozaki, "ERD15, a cDNA for a dehydration-induced gene from Arabidopsis thaliana," Plant physiology, vol. 106, p. 1707, 1994.

978-1-4799-7294-4/14/$31.00 ©2014 IEEE