PLOS | ONE

RESEARCH ARTICLE

An NMF-$L_{2,1}$-Norm Constraint Method for Characteristic Gene Selection

Dong Wang[1], Jin-Xing Liu[1,3]*, Ying-Lian Gao[2], Jiguo Yu[1], Chun-Hou Zheng[1], Yong Xu[3]

1 School of Information Science and Engineering, Qufu Normal University, Rizhao, 276826, China, **2** Library of Qufu Normal University, Qufu Normal University, Rizhao, 276826, China, **3** Bio-Computing Research Center, Shenzhen Graduate School, Harbin Institute of Technology, Shenzhen, 518055, China

* sdcavell@126.com

Citation: Wang D, Liu J-X, Gao Y-L, Yu J, Zheng C-H, Xu Y (2016) An NMF-$L_{2,1}$-Norm Constraint Method for Characteristic Gene Selection. PLoS ONE 11(7): e0158494. doi:10.1371/journal.pone.0158494

Editor: Xi Luo, Brown University, UNITED STATES

Received: August 21, 2015

Accepted: June 16, 2016

Published: July 18, 2016

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: This work was supported in part by the NSFC under grant nos. 61572284, 61502272, 61370163 and 61272339; the Shandong Provincial Natural Science Foundation, under grant nos. ZR2013FL016 and BS2014DX004; and Shenzhen Municipal Science and Technology Innovation Council (Nos. JCYJ20140904154645958, JCYJ20140417172417174 and CXZZ20140904154910774). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Abstract

Recent research has demonstrated that characteristic gene selection based on gene expression data remains faced with considerable challenges. This is primarily because gene expression data are typically high dimensional, negative, non-sparse and noisy. However, existing methods for data analysis are able to cope with only some of these challenges. In this paper, we address all of these challenges with a unified method: nonnegative matrix factorization via the $L_{2,1}$-norm (NMF-$L_{2,1}$). While $L_{2,1}$-norm minimization is applied to both the error function and the regularization term, our method is robust to outliers and noise in the data and generates sparse results. The application of our method to plant and tumor gene expression data demonstrates that NMF-$L_{2,1}$ can extract more characteristic genes than other existing state-of-the-art methods.

1 Introduction

The development of microarray technologies makes the study of complex biological gene expression networks possible. Microarray datasets typically contain expression data for the thousands of genes profiled on each chip, and the number of replicates is much smaller than the number of genes, making the selection of genes difficult[1]. In addition, the inclusion of irrelevant or noisy variables may decrease the accuracy of selection[2]. The problem of how to select genes associated with the target terms has become a challenge for scientists[3]. For example, plants are able to cope with environmental challenges such as cold, heat, and salt, which are referred to as abiotic stresses; there must therefore exist specific interacting genes that respond to each abiotic stress. Another typical example is that of cancer, an important cause of human morbidity; the identification of genes that are frequently mutated in cancers and play an essential role in cancer development is critical. Many methods have been proposed for processing gene expression data collected by DNA microarray profiling[4–9]. For example, Liu et al.[10]used a method based on penalized matrix decomposition (PMD) to extract characteristic plant genes, and Zheng et al.[11] applied nonnegative matrix factorization (NMF) to tumor gene selection. Principal component analysis (PCA) and singular value decomposition (SVD) have also been used to analyze gene expression data[12]. Liu et al.[13] proposed a

Competing Interests: The authors have declared that no competing interests exist.

CIPMD algorithm (A Class-Information-Based Penalized Matrix Decomposition) for identifying plants core genes responding to abiotic stresses. This method is PMD method with label information. Liu et al.[14] proposed a PRFE algorithm (A P-Norm Robust Feature Extraction) for identifying differentially expressed genes. Although those methods are all feature selection methods and in widespread use, they present some disadvantages:

1. Although the elements of the initial data matrix are entirely nonnegative, the traditional low-rank algorithm [15]cannot guarantee nonnegative values in the project matrix, thereby complicating their biological interpretation.
2. The high dimensionality of data poses challenges, such as the so-called curse of dimensionality[16,17].
3. Faced with millions of individual data points, it is difficult to interpret gene expression data without sparse constraints.
4. Gene expression data often contains numerous outliers and abundant noise, which traditional methods do not effectively address.

NMF has been widely used in various fields because it can generate low-rank and nonnegative results. The ability to generate a low-rank nonnegative matrix to approximate a given nonnegative data matrix is a significant advantage[18], but the lack of sparsity in data processed via NMF makes this method less than ideal for characteristic gene selection. In high throughput datasets, gene expression data are high dimensional and always contain some redundant information(i.e., not all features are relevant). To address these problems, we sought to incorporate sparsity, or the reduction of certain vector elements to zero. The regular inclusion of sparsity has played a significant role in dimensionality reduction and feature selection[19]. For example, Journée et al.[20] proposed a sparse principal component analysis (SPCA) method using the generalized power method, and Wittenet al.[21]proposed a penalized matrix decomposition (PMD) method, which has been proven useful in microarray analysis by imposing penalization on factor matrices. Nonnegative matrix factorization with sparse constraints (NMFSC), which was first introduced by Patrik O. Hoyer in 2004 [22], accurately controls sparsity. NMFSC has been applied to the problems of imaging and gene selection, among others. However, it does not guarantee that entire rows of a matrix are sparse, which can lead to difficulties during feature selection. To address these issues, the $L_{2,1}$ version of NMF favors the inclusion of a small number of non-zero rows in the factor matrix, which are proposed to generate sparse results for rows.

However, these methods for generating sparsity apply the least square error function, which does not reliably address noise and outliers [23]. When faced with these complications, the error for both features and samples will be squared[24], increasing the effect of large noises or outliers [25]. As a result, the $L_{2,1}$ version of the error function has been proposed to address noisy data [26].

In light of these problems, we propose a novel method called Nonnegative Matrix Factorization with $L_{2,1}$-norm (NMF-$L_{2,1}$), which imposes an $L_{2,1}$-norm constraint on both the error function and a regularization term to solve the aforementioned problems simultaneously. A sparse regularization term avoids the potential problem of over-fitting and selects a sparse subset of features. Rather than use an L_2-norm-based error function, the $L_{2,1}$-norm-based error function diminishes the impact of the outliers and noise in a dataset[25,27].

The main contributions of this paper are the following:

First, $L_{2,1}$-norm is employed for regularization in our method to generate sparse results, making the results easier to interpret.

Second, nonnegative matrix factorization is utilized to generate low-rank results with non-negative values.

Third, the $L_{2,1}$-norm-based error function is used to diminish the outliers and noise inherent in gene expression data.

This paper is organized as follows. The methodology section introduces the NMF-$L_{2,1}$ method and provides an efficient algorithm for estimation. The results and discussion section compares our method with other three methods: PMD, NMFSC and SPCA. Our conclusions are presented in the third section.

2 Methodology

2.1 Mathematical Definition of $L_{2,1}$-norm

This subsection briefly introduces the $L_{2,1}$-norm proposed in [28]. It is defined as

$$\|\mathbf{M}\|_{2,1} = \sum_{i=1}^{n} \sqrt{\sum_{j=1}^{s} \mathbf{m}_{ij}^2} = \sum_{i=1}^{n} \|\mathbf{m}^i\|_2, \tag{1}$$

where $\mathbf{m}^i$ is the i-th row of $\mathbf{M}$, m_{ij} is the (i, j)-th entry in $\mathbf{M}$, $\mathbf{M}$ is an $n \times s$ matrix. An explanation of $L_{2,1}$-norm is as follows. First, we compute the L_2-norm of rows $\mathbf{m}^i$, then compute the L_1-norm of vector $b(\mathbf{M}) = (\|\mathbf{m}^1\|_2, \|\mathbf{m}^2\|_2, \ldots, \|\mathbf{m}^s\|_2)$. The amplitude of the components of vector b($\mathbf{M}$) dictate how important each dimension is $L_{2,1}$-norm favors a small number of non-zero rows in $\mathbf{M}$,ensuring that an appropriate dimensional reduction is achieved[29].

2.2 Extracting Characteristic Genes by NMF-L2,1

In this paper, the matrix $\mathbf{X}$ denotes the initial gene expression dataset, whose size is $n \times c$. Each column of $\mathbf{X}$ represents the transcriptional response of the n genes in one sample. Each row of $\mathbf{X}$ represents the expression level of a gene across all samples. Thus, the $\mathbf{X}$ can be approximated as:

$$\mathbf{X} \approx \mathbf{AY}, \tag{2}$$

where $\mathbf{A}$ is an $n \times d$ matrix, $\mathbf{Y}$ is a $d \times c$ matrix, and $d < \min(n, c)$.

The matrices $\mathbf{Y}$ and $\mathbf{A}$ are called the coefficient and basis matrices, respectively. Given suitable parameters for NMF-$L_{2,1}$, the sparse matrix $\mathbf{A}$ can be obtained. Characteristic genes can then be extracted according to the non-zero entries in $\mathbf{A}$ [30].

2.3 NMF based on L2,1-norm (NMF-L2,1)

Let $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_c) \in R^{n \times c}$, $\mathbf{Y} = (\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_c)^T \in R^{d \times c}$. The error function of standard NMF [31] is

$$\|\mathbf{X} - \mathbf{AY}\|_F^2 = \sum_{i=1}^{c} \|\mathbf{x}_i - \mathbf{A}\mathbf{y}_i\|^2, \quad st.\ \mathbf{Y} > \mathbf{0}, \mathbf{A} > \mathbf{0}. \tag{3}$$

Here, the error for each data point is calculated as a squared residual error in terms of $\|\mathbf{x}_i - \mathbf{A}\mathbf{y}_i\|^2$. As a result, a few outliers with large errors can easily dominate the objection function due to the squared errors. Thus, it is reasonable to propose an NMF-$L_{2,1}$ formulation to reduce the influence of outliers and errors.

The error function of the NMF-$L_{2,1}$ formulation is:

$$\|\mathbf{X}-\mathbf{AY}\|_{2,1}=\sum_{i=1}^{c}\sqrt{\sum_{j=1}^{n}(\mathbf{X}-\mathbf{AY})_{ij}^{2}}=\sum_{i=1}^{c}\|\mathbf{x}_i-\mathbf{A}\mathbf{y}_i\|. \quad (4)$$

In this formulation, the error for each data point is $\|\mathbf{x}_i-\mathbf{A}\mathbf{y}_i\|$, which is not squared; thus, the impact of large errors caused by outliers does not fully dominate the objective function[32].

The NMF-$L_{2,1}$ optimization problem is formulated as

$$\min_{\mathbf{Y}}\|\mathbf{X}-\mathbf{AY}\|_{2,1}+\lambda\|\mathbf{Y}\|_{2,1},\quad st.\ \mathbf{Y}>\mathbf{0},\mathbf{A}>0. \quad (5)$$

The problem in Eq (4) is equivalent to

$$\min_{\mathbf{Y},\mathbf{E}}\|\mathbf{E}\|_{2,1}+\|\mathbf{Y}\|_{2,1},\quad st.\ \mathbf{AY}+\lambda\mathbf{E}=\mathbf{X},\mathbf{Y}>0,\mathbf{A}>0. \quad (6)$$

Thus, the above problem can be rewritten as

$$\min_{\mathbf{Y},\mathbf{E}}\left\|\begin{bmatrix}\mathbf{Y}\\ \mathbf{E}\end{bmatrix}\right\|_{2,1},\qquad st.\ [\mathbf{A}\quad \lambda\mathbf{I}]\begin{bmatrix}\mathbf{Y}\\ \mathbf{E}\end{bmatrix}=\mathbf{X},\mathbf{Y}>\mathbf{0},\mathbf{A}>0, \quad (7)$$

where $\mathbf{I}\in R^{n\times n}$ is an identity matrix, $n=c$. Let $\mathbf{B}=[\mathbf{A}\ \lambda\mathbf{I}]\in R^{n\times b}$ and $\mathbf{U}=[\mathbf{Y}^T\quad \mathbf{E}^T]^T\in R^{b\times c}$, where $b=d+n$.

Then, the problem can be reformulate das

$$\min_{\mathbf{U}>0}\|\mathbf{U}\|_{2,1},\quad st.\ \mathbf{BU}=\mathbf{X}. \quad (8)$$

How can this optimization problem handle high dimensional, nonnegative, noisy and sparse data simultaneously? The reasons are as follows:

1. The $L_{2,1}$-norm error function term is designed to diminish the impact of noise or outliers contained in the original data. As a result, we can expect to obtain cleaner data for subsequent analyses.
2. This clean data may be not sparse–some features may be irrelevant to the learning procedure—so the $L_{2,1}$-norm regularization term[33]can be used to generate a sparse solution.
3. In the next section, we will demonstrate that the above conditions produce more effective models, especially for datasets that are sparse, nonnegative, high dimensional and noisy.

2.4 An Efficient Algorithm for NMF-L2,1

To solve the constrained optimization problem in Eq (7), Nie et al.[34]have provided an efficient algorithm. Here, we briefly introduce the efficient algorithm.

By introducing Lagrangian multiplier $\mathbf{\Lambda}$, we first give the Lagrangian function as follows:

$$L(\mathbf{U})=\sum_{i=1}^{b}\|\mathbf{u}^i\|_2-\mathrm{Tr}(\Lambda^T(\mathbf{BU}-\mathbf{X})), \quad (9)$$

where Tr() is the trace function of a matrix. Here we introduce the augmented cost-function

$$J(\mathbf{U},\mathbf{q})=\mathrm{Tr}(\mathbf{U}^T\mathbf{QU})-\mathrm{Tr}(\Lambda^T(\mathbf{BU}-\mathbf{X})), \quad (10)$$

where $\mathbf{q}\in R^b$ is an auxiliary vector and $\mathbf{Q}=diag(\mathbf{q})$ is a diagonal matrix with the diagonal

element

$$\mathbf{Q} = \mathrm{diag}(\mathbf{q}) = diag\left(\frac{1}{2\|\mathbf{u}^i+\boldsymbol{\varepsilon}\|_2}\right), \tag{11}$$

in which ε is a positive number and infinitely close to, but not equal to, zero.

Taking the derivative of $J(\mathbf{U}, \mathbf{q})$ with respect to $\mathbf{U}$ to zero, we obtain:

$$\frac{\partial J(\mathbf{U},\mathbf{q})}{\partial \mathbf{U}} = 2\mathbf{Q}\mathbf{U} - \mathbf{B}^T\Lambda = 0. \tag{12}$$

By multiplying the two sides of Eq (11) by $\mathbf{B}\mathbf{Q}^{-1}$ and using the constraint $\mathbf{B}\mathbf{U} = \mathbf{X}$, we obtain

$$\begin{aligned} &2\mathbf{B}\mathbf{U} - \mathbf{B}\mathbf{Q}^{-1}\mathbf{B}^T\Lambda = 0 \\ &\rightarrow 2\mathbf{X} - \mathbf{B}\mathbf{Q}^{-1}\mathbf{B}^T\Lambda = 0 \\ &\rightarrow \Lambda = 2(\mathbf{B}\mathbf{Q}^{-1}\mathbf{B}^T)^{-1}\mathbf{X}. \end{aligned} \tag{13}$$

Then, we obtain:

$$\mathbf{U} = \mathbf{Q}^{-1}\mathbf{B}^T(\mathbf{B}\mathbf{Q}^{-1}\mathbf{B}^T)^{-1}\mathbf{X}. \tag{14}$$

More details of the algorithm can be found in[34]. Here, we summarize our method in Box 1. In each iteration, $\mathbf{U}$ is calculated with the current $\mathbf{Q}$. Then, $\mathbf{Q}$ is updated based on the current $\mathbf{U}$. The iteration procedure is repeated until the algorithm converges.

In this paper, the characteristic genes are extracted by the coefficient matrix $\mathbf{A}$. We summarize the NMF-$L_{2,1}$ method to extract core genes as follows:

1. Create the data matrix $\mathbf{X}$ based on gene expression data.
2. Obtain the basis matrix $\mathbf{A}$ by using the NMF-$L_{2,1}$ method.
3. Extract characteristic genes from non-zero entries in matrix $\mathbf{A}$.
4. Exploit the Gene Ontology (GO) tool to investigate the extracted genes.

3 Results and Discussion

In this section, several experiments are carried out. In the first subsection, the NMF-$L_{2,1}$ method is compared with the following methods for a gene expression dataset obtained from plants responding to abiotic stresses: (a) the PMD method (proposed by Witten et al. [15]); (b) the SPCA method (proposed by Journée et al. [35]); and (c) the NMFSC method (proposed by Patrik O. Hoyer[36]);(d) the CIPMD method (proposed by Liu et al.[13]); (e) PRFE method (proposed by Liu et al. [14]). In the second subsection, the six methods are compared for Medulloblastoma and leukemia tumor datasets.

3.1 Results for Plant Gene Expression Data

Plants are continually challenged by environmental parameters such as drought, salt, cold, osmotic pressure, and UV-B light [37]. Among plant genes, there must exist a specific set of interacting genes that respond to each abiotic stress. Thus, it is important but challenging to extract genes responding to each abiotic stress from plant gene expression data.

3.1.1 Data source. The gene expression datasets used in our experiment were downloaded from the *NASC* Arrays [http://affy.arabidopsis.info/link_to_iplant.shtml]. Each sample profiles

Box 1. NMF-$L_{2,1}$ method.

Input: $\mathbf{X} \in R^{n \times c}$ and parameter λ.
Output: $\mathbf{Y} \in R^{d \times c}$, $\mathbf{A} \in R^{n \times d}$.
1: Initialize $\mathbf{Q}_t \in R^{b \times b}$ as an identity matrix and $\mathbf{A} \in R^{n \times d}$ as a nonnegative matrix,
  set t = 0.
2: repeat
  Compute $\mathbf{U}_{t+1} = \mathbf{Q}_t^{-1}\mathbf{B}^T(\mathbf{B}\mathbf{Q}_t^{-1}\mathbf{B}^T)^{-1}\mathbf{X}$.
  Setting $\mathbf{U} > \mathbf{0}$.
  Compute diagonal matrix $\mathbf{Q}_{t+1}$ according to Eq (11).
  t = t+1.
  $\mathbf{A}$ and $\mathbf{Y}$ are obtained from $\mathbf{U}$ according to Eq (7).
Until convergence.

22810 genes. The plant gene expression datasets are shown in supplementary file (S1 Table). Table 1 lists the reference numbers and sample numbers for each stressor.

3.1.2 The selection of parameter λ. In order to obtain the most effective results, we used gene expression data to train the parameter λ. For each sample, the parameter varied from 0–1 with a step of 0.1, and GO Terms were used to select the most appropriate parameter. The results are provided in Table 2.

3.1.3 gene ontology (GO) analysis. In this paper, GO Term is used to evaluate the genes that responded to plant abiotic stressors[38]. GO Term Finder analysis provided information to aid with the biological interpretation of high-throughput experiments. GO Term Finder is available publicly at [http://go.princeton.edu/cgi-bin/GOTermFinderS] [39]; it aims to describe genes in the query/input set and to find the genes that may have something in common.

For the sake of simplicity, 500 genes were selected from the gene expression data by the NMFSC, PMD, SPCA, CIPMD, PRFE and NMF-L_{21} methods. The threshold parameters used were: maximum P-value = 0.01, and minimum number of genes = 2.

3.1.4 Response to stimulus. Table 3 summarizes the results of the response to a stimulus whose background frequency in the TAIR (A.thaliana (common wallcress))set was 6617/30320 (21.8%). The results are presented according to P-value and sample frequency. The P-value was calculated using a hyper-geometric distribution (details can be seen in[40]). The sample frequency denotes the number of the characteristic genes selected. For example, 330/500 denotes that 330 genes corresponding to GO terms out of 500 genes were selected by the method.

Table 1. Reference and Sample Numbers for Stress Types.

Stress type	Drought	Salt	UV-B	Cold	Heat	Osmotic	Control
Reference Number	141	140	144	138	146	139	137
Sample Number	6	7	6	7	8	6	9

doi:10.1371/journal.pone.0158494.t001

Table 2. The Selection of Parameter λ.

Stress type	Drought	Salt	UV-B	Cold	Heat	Osmotic
Shoot	0.3	0.3	0.2	0.5	0.3	0.3
Root	0.3	0.3	0.3	0.3	0.3	0.3

doi:10.1371/journal.pone.0158494.t002

Table 3. Response to Stimulus (GO: 0050896).

Stress type	NMF-L21		NMFSC		PMD		SPCA		CIPMD		PRFE	
	P-value	Sample frequency	P-value	Sample frequency	P-value	Sample frequency	P-value	Sample frequency	P-value	Sample frequency	P-value	Sample frequency
Drought s	**2.77E-122**	**353/500 70.60%**	5.28E-109	342/500 68.40%	1.09E-55	276/500 55.20%	1.19E-55	273/500 54.60%	9.08E-106	338/ 50067.7%	3.39E-93	314/ 50062.8%
Drought r	2.69E-69	293/500 58.60%	2.39E-88	318/500 63.60%	3.67E-65	287/500 57.40%	5.27E-65	289/500 57.70%	**5.54E-101**	**333 \50066.6%**	3.49E-41	240/50048%
Salt s	7.71E-53	271/500 54.20%	2.01E-52	268/500 53.70%	3.78E-48	262/500 52.40%	3.31E-21	262/500 52.40%	**2.00E-81**	**309/ 50061.9%**	1.72E-39	236/ 50047.5%
Salt r	**5.64E-96**	**326/500 65.20%**	3.69E-94	325/500 65.00%	1.25E-80	314/500 62.80%	1.42E-34	237/500 47.40%	1.31E-70	295/50059%	2.95E-55	163/ 50052.6%
UV-B s	**6.26E-152**	**382/500 76.40%**	1.62E-126	360/500 72.00%	4.99E-128	362/500 72.40%	2.59E-33	332/500 66.40%	1.81E-103	335/ 50067.3%	7.72E-89	308/ 50061.8%
UV-B r	5.85E-42	247/500 49.40%	3.81E-64	286/500 57.20%	1.21E-36	242/500 48.20%	9.56E-22	210/500 42.50%	**6.85E-95**	**326/ 50065.2%**	3.16E-34	227/ 50045.5%
Cold s	1.61E-85	312/500 62.40%	3.81E-77	304/500 60.80%	1.58E-66	294/500 58.80%	4.31E-62	283/500 56.60%	**7.52E-98**	**329/ 50065.9%**	1.10E-47	250/ 50050.3%
Cold r	2.56E-81	309/500 61.80%	7.59E-79	306/500 61.20%	8.44E-72	291/500 58.20%	3.08E-61	281/500 56.20%	**6.92E-105**	**337/ 50067.5%**	7.84E-46	248/ 50049.6%
Heat s	3.43E-25	218/500 43.60%	4.28E-19	204/500 40.80%	2.56E-23	220/500 44.00%	2.56E-23	300/500 46.40%	**2.39E-98**	**330/ 50066.0%**	9.16E-31	220/ 50044.2%
Heat r	3.96E-23	214/500 42.80%	3.22E-16	197/500 39.40%	1.18E-19	205/500 41.0%	1.69E-17	200/500 40.00%	**7.27E-105**	**337/ 50067.5%**	2.10E-14	182/ 50036.6%
Osmotic s	5.96E-73	298/500 59.60%	1.02E-82	311/500 62.20%	3.07E-75	294/500 58.80%	4.20E-49	263/500 52.60%	**6.15E-92**	**322/ 50064.5%**	2.17E-51	257/ 50051.4%
Osmotic r	5.84E-55	273/500 54.40%	1.65E-47	260/500 52.10%	2.67E-21	221/500 44.20%	1.12E-34	237/500 47.50%	**2.54E-76**	**302/ 50060.2%**	9.00E-25	208/ 50041.6%

's' denotes shoot samples; 'r' denotes root samples.

doi:10.1371/journal.pone.0158494.t003

As listed in Table 3, the six methods were compared by sample frequency and P-value. NMFSC, NMFL21, PRFE, SPCA and PMD are unsupervised methods, so we first compare the five algorithms. In the 12 terms, the results show that our method could extract more characteristic genes than the other methods for eight of twelve samples. For example, for shoot samples exposed to UV-B stress, the sample frequency was 76.4% by our method, 72% by NMFSC, 72.4% by PMD, 66.4% by SPCA and 61.8% by PRFE. This shows that our method is markedly improved over PRFE, PMD and SPCA. When compared with the supervised method CIPMD, except the salt and UV-B stress, our methods performs worth than CIPMD. Generally speaking, since supervised methods take the class labels into consideration, they usually have better performance than unsupervised methods.

3.1.5 Response to the abiotic stimulus. Table 4 summarizes the results of the six methods for datasets describing the response to abiotic stimulus whose background frequency in the TAIR (A.thaliana (common wallcress)) set is 1539/29556 (5.2%). The numbers of characteristic genes and the P-values of genes responding to an abiotic stimulus (GO:0009628) in root and shoot samples are listed in Table 4.

As described above in the 'Response to stimulus' section, we first compare the five algorithms NMFSC, NMFL21, PRFE, SPCA and PMD. From the results we can see that our method could extract more characteristic genes than the PMD and SPCA methods for all datasets. For the four sample datasets (drought, salt, UV-B and osmotic), our method performed worse than NMFSC, but superior to PMD and SPCA. When compared with the supervised

Table 4. Response to an Abiotic Stimulus (GO:0009628).

Stress type	NMF-L_{21}		NMFSC		PMD		SPCA		CIPMD		PRFE	
	P-value	Sample frequency	P-value	Sample frequency	P-value	Sample frequency	P-value	Sample frequency	P-value	Sample frequency	P-value	Sample frequency
Drought s	6.32E-59	**182/ 50036.4%**	4.89E-36	149/500 29.80%	3.91E-34	107/500 21.40%	7.50E-21	87/500 17.00%	2.28E-50	170/50034%	4.85E-28	136/ 50027.2%
Drought r	1.82E-22	126/500 25.20%	3.65E-23	127/500 25.40%	1.78E-10	68/500 13.60%	4.14E-08	63/500 12.60%	1.21E-55	177/ 50035.5%	8.05E-60	**183/ 50036.6%**
Salt s	2.71E-50	**170/500 34.00%**	9.95E-43	159/500 31.80%	9.93E-39	113/500 22.60%	9.83E-33	105/500 21.00%	5.65E-44	161/ 50032.2%	1.00E-39	114/ 50022.8%
Salt r	7.32E-36	149/500 29.80%	8.58E-47	165/500 33.00%	1.36E-15	78/500 15.60%	6.18E-12	71/500 14.00%	5.52E-57	**295/50059%**	7.90E-39	153/ 50030.8%
UV-B s	9.11E-46	164/500 32.80%	4.95E-49	168/500 33.80%	1.76E-13	74/500 14.80%	7.84E-23	90/500 18.00%	7.04E-55	**176/ 50035.3%**	1.53E-27	135/ 50027.1%
UV-B r	1.17E-14	110/500 22.00%	2.75E-22	100/500 20.00%	5.30E-10	67/500 13.40%	8.00 E-4	52/500 10.00%	5.20E-61	**184/ 50036.9%**	4.30E-51	171/ 50034.3%
Cold s	5.78E-64	**188/ 50037.60%**	3.85E-61	183/500 36.60%	5.82E-35	106/500 21.60%	1.17E-19	85/500 17.00%	8.99E-52	172/ 50034.4%	5.02E-56	178/ 50035.6%
Cold r	1.05E-53	175/ 50035.00%	1.59E-52	173/500 34.60%	2.74E-23	91/500 18.20%	4.10E-19	84/500 16.80%	6.31E-58	180/ 50036.1%	4.24E-61	**184/ 50037.0%**
Heat s	7.16E-22	125/ 50025.00%	3.43E-76	118/500 23.60%	1.44E-24	93/500 18.60%	4.64E-22	89/500 17.80%	1.13E-52	**173/ 50034.6%**	1.37E-35	148/ 50029.8%
Heat r	9.27E-34	145/ 50029.00%	6.04E-32	142/500 28.40%	1.41E-15	78/500 15.60%	1.35E-08	64/500 12.80%	1.64E-52	**173/ 50034.6%**	1.18E-49	169/ 50033.9%
Osmotic s	6.35E-56	178/500 35.60%	8.22E-61	**184/500 36.80%**	6.55E-38	112/500 22.40%	2.02E-18	83/500 16.60%	4.88E-47	165/ 50033.1%	4.76E-28	136/ 50027.2%
Osmotic r	1.32E-49	169/ 50033.80%	2.01E-39	154/500 30.80%	1.40E-14	76/500 15.20%	2.87E-17	81/500 16.20%	1.28E-47	166/ 50033.3%	1.67E-54	**176/ 50035.2%**

's' denotes shoot samples; 'r' denotes root samples.

doi:10.1371/journal.pone.0158494.t004

method CIPMD, except the salt, heat and UV-B stress, CIPMD performs better than other methods.

3.1.6 Characteristic terms. In Table 5, we list the characteristic terms. Our method outperformed SPCA and PMD for all 12 items and outperformed NMFSC for seven items. Only in one item (shoot sample in UV-B) the PRFE outperforms than our method. However, for one of the twelve items (cold in root) our method produced the same result as NMFSC. From these results, it can be concluded that our method is more effective than other unsupervised methods. The response to water deprivation (GO:0009414) for shoot samples is also analyzed in Table 5. The background frequency of the response to water deprivation (GO: 0009414) is 1.4%. It is obvious that NMF-$L_{2,1}$ is able to extract more characteristic genes than the other methods, and the sample frequency in response to water deprivation by our method is 18.1%, while it is 13.2% for NMFSC, 11.9% for PMD, 16.8% for CIPMD, 11.2% for PRFE and 8.2% for SPCA, indicating that our method performs 6.2% better than PMD, 7% better than PRFE and almost 10% better than SPCA.

3.2 Results for tumor datasets. Two tumor datasets were also analyzed to verify the performance of the proposed method. The medulloblastoma dataset contains 34 samples, which can be divided into 25 tumor and 9 normal tissue samples, and assesses the expression of 5893 genes [41]. The leukemia dataset consists of 5000 genes and 38 samples [42]. The samples include 27 tumor and 11 normal tissue samples.

To make a fair comparison, all the six methods extracted 100 genes as characteristic genes from the two tumor datasets. The Gene Ontology (GO) enrichment and functional annotation

Table 5. Characteristic Terms Selected from GO by Algorithms.

Stress type	GO Terms	Background frequency	Sample frequency and ratio					
			NMF-L21	NMFSC	PMD	SPCA	CIPMD	PRFE
Drought s	GO:0009414 response to water deprivation	207/298870.70%	**91/ 50018.20%**	66/500 13.20%	47/500 9.40%	23/500 4.60%	84/ 50016.8%	56/ 50011.2%
Drought r	GO:0009415 response to water deprivation	207/298870.70%	58/500 11.60%	62/500 12.40%	26/500 5.20%	24/500 4.80%	**69/ 50013.8%**	30/5000.6%
Salt s	GO:0009651 response to salt stress	395/29887 1.30%	**80/500 16.00%**	80/500 16.00%	41/500 8.20%	28/500 5.60%	64/ 50012.8%	42/5008.4%
Salt r	GO:0009651 response to salt stress	395/298871.30%	74/500 14.80%	**76/500 15.20%**	33/500 6.60%	22/500 4.40%	57/ 50011.4%	37/5007.4%
UV-B s	GO:0009416 response to light stimulus	557/298871.90%	31/5006.20%	24/ 5004.80%	23/500 4.60%	30/500 6.00%	none	**40/5008%**
UV-B r	GO:0009416 response to light stimulus	557/298871.90%	**34/5006.80%**	24/ 5004.80%	24/500 4.80%	22/ 5004.40%	none	none
Cold s	GO:0009409 response to cold	276/298870.90%	62/500 12.40%	59/500 11.80%	44/500 8.80%	34/500 6.80%	**54/ 50019.8%**	57/ 50011.4%
Cold r	GO:0009409 response to cold	276/29887 0.90%	**67/500 13.40%**	67/500 13.40%	43/500 8.60%	33/500 6.60%	56/ 50011.2%	48/5009.6%
Heat s	GO:0009408 response to heat	140/29887 0.50%	67/500 13.40%	59/ 50011.8%	45/500 9.00%	30/500 6.00%	**97/ 50019.4%**	59/ 50011.8%
Heat r	GO:0009408 response to heat	140/298870.50%	80/500 16.00%	91/500 18.20%	43/500 8.60%	28/500 5.60%	**93/ 50018.6%**	52/ 50010.4%
Osmotic s	GO:0006970 response to osmotic stress	474/29887 1.60%	91/ 50018.20%	94/500 18.80%	55/500 11.00%	29/500 5.80%	**95/50019%**	55/50011%
Osmotic r	GO:0006970 response to osmotic stress	474/298871.60%	**80/500 16.00%**	79/500 15.80%	39/500 7.80%	27/500 5.40%	68/ 50013.6%	47/5009.4%

's' denotes shoot samples; 'r' denotes root samples.

doi:10.1371/journal.pone.0158494.t005

of the extracted genes by all six methods was performed by *ToppFun*, which is publicly available at [http://toppgene.cchmc.org/enrichment.jsp].

Tables 6 and 7 list the top 10 closely related terms with P-value corresponding to different methods for the two tumor datasets. From Table 6, it can be seen that NMF-$L_{2,1}$ outperforms the other three methods for all terms. For example, for the term M11197 in Table 6, the P-value from NMF-$L_{2,1}$ is 7.36E-129 and 70/389 denotes that NMF-$L_{2,1}$ extracts 70 genes corresponding to M11197, whereas NMFSC, PMD,CIPMF, PRFE and SPCA extracted 62, 62,48, 52 and 55 genes corresponding to M11197, respectively, and the total number of genes corresponding to M11197 was 389. In Table 7 we can see that our method outperforms than other method in seven terms, only in other three terms(17092989, 19755675, 11108479) PRFE have a lower P-value than our method. Thus, we can conclude that our method extracts more genes than others.

In summary, we conclude that our method is generally superior to others and is effective for the extraction of genes.

4 Conclusions

In this paper, we proposed an effective method to select characteristic genes with$L_{2,1}$-norm minimization of both the error function and the regularization term. The $L_{2,1}$-norm-based error function is robust to outliers and noise in the data points and is computationally efficient. Furthermore, the $L_{2,1}$-norm-based regularization term is used to generate a sparse solution. We also used the nonnegative factorization method to avoid the problems stemming from the

Table 6. P-value Terms for the Medulloblastoma Dataset.

ID	Name	P-value Sample frequency					
		NMF-L_{21}	NMFSC	PMD	SPCA	CIPMD	PRFE
M11197	Housekeeping genes identified as expressed across 19 normal tissues	**7.63E-129 70/389**	3.28E-121 62/389	5.59E-92 59/389	6.28E-81 55/389	1.227E-6848/389	7.112E-7552/389
12456497	Human Leukemia Durig03 88 genes	**1.35E-11851/81**	1.29E-99 43/81	3.26E-77 38/81	4.90E-79 39/81	1.592E-7738/81	2.331E-5831/81
GO:0022626	Cytosolic ribosome	**9.06E-111 54/96**	8.62E-81 37/96	3.03E-56 32/96	7.30E-68 37/96	1.152E-3825/96	2.955E-4528/96
GO:0006415	Translational termination	**8.43E-108 53/95**	7.19E-78 37/95	6.91E-56 32/95	9.59E-68 37/95	1.388E-3526/95	2.146E-3426/95
GO:0006414	Translational elongation	**1.35E-105 56/130**	1.86E-77 40/130	2.04E-57 35/130	7.25E-66 39/130	1.311E-4935/130	6.78E-3628/130
GO:0006614	SRP-dependent cotranslational protein targeting to membrane	**7.66E-104 53/107**	5.26E-75 37/107	6.38E-54 32/107	2.18E-65 37/107	4.160E-5432/107	1.499E-4026/107
GO:0044391	ribosomal subunit	**1.86E-99 55/148**	1.54E-70 38/148	1.97E-51 33/148	7.75E-62 38/148	1.317E-3023/148	3.727E-3928/148
GO:0003735	Structural constituent of ribosome	**4.11E-95 54/156**	3.35E-67 37/156	2.38E-48 32/156	8.85E-59 37/156	9.511E-4633/156	3.656E-3427/156
GO:0005198	Structural molecule activity	**5.09E-65 58/641**	1.07E-54 44/641	9.98E-37 38/641	2.68E-44 43/641	1.952E-2628/641	1.192E-2935/641
GO:0003723	RNA binding	**9.48E-57 63/1568**	2.16E-39 47/1568	7.72E-26 41/1568	2.37E-30 45/1568	7.033E-2037/1568	1.153E-2037/1568

doi:10.1371/journal.pone.0158494.t006

high-dimensional and nonnegative nature of the data. In summary, our method can cope with high dimensionality, non-negativity, sparseness and noise simultaneously.

Furthermore, the genes selected by our method and others from both plant and tumor datasets were compared using GO enrichment. These results indicate that the proposed NMF-$L_{2,1}$ method is superior to SPCA and PMD for selecting characteristic genes.

Table 7. P-value Terms for the Leukemia Dataset.

ID	Name	P-value sample frequency					
		NMF-L21	NMFSC	PMD	SPCA	CIPMD	PRFE
M11197	Housekeeping genes identified as expressed across 19 normal tissues.	**3.65E-41 40/389**	3.49E-40 33/389	1.57E-30 21/389	5.62E-23 23/389	2.584E-1820/389	1.05E-3230/389
17092989	Human Lymphoma Foge l07 33 genes	1.20E-33 18/33	6.25E-29 14/33	3.49E-32 12/33	2.57E-19 10/33	3.93E-2212/33	**8.43E-5123/33**
19755675	Human Leukemia Li09 419 genes	4.55E-25 25/410	4.27E2220/410	4.48E-21 18/410	none	6.82E-2223/410	**5.88E-2928/410**
19699293	Human Leukemia Bienkowska09 80 genes	**2.17E-22 14/75**	2.88E-15 9/75	8.91E-16 11/75	1.95E-13 9/75	2.58E-1712/75	3.03E-1712/732
15474998	Mouse StemCell Lindmark04 950 genes	**4.71E-21 24/732**	2.45E-20 22/732	4.89E-18 20/732	1.52E-15 20/732	none	none
16872506	Human Leukemia Yukinawa06 2000 genes	**1.54E-20 33/1505**	4.05E-18 28/1505	1.10E-16 23/1505	none	none	none
18689800	Human EmbryonicStemCell Thomas08 1088 genes	**1.97E-20 29/1023**	1.06E-17 26/1023	4.59E-19 21/1023	1.03E-10 20/1023	2.28e-1121/1023	3.62E-1222/1023
11108479	Human Leukemia Ben-Dor00 143 genes	1.22E-18 14/129	1.17E-15 11/129	1.10E-16 12/129	1.61E-15 12/129	1.424E-1714/129	**1.45E-2217/129**
12077300	Human Lymphoma Lossos02 152 genes	**5.43E-17 13/99**	6.91E-14 10/99	8.21E-16 10/99	3.41E-129/99	none	none
M11620	Genes induced in the liver during hepatitis B (HBV) viral clearance in chimpanzees.	**3.20E-15 12/101**	3.27E-12 9/101	3.58E-129/101	4.18E-12 8/101	6.96E-119/101	5.68E-1411/101

doi:10.1371/journal.pone.0158494.t007

Supporting Information

S1 Table. The plant gene expression dataset.
(ZIP)

Author Contributions

Conceived and designed the experiments: JXL. Performed the experiments: DW JXL. Analyzed the data: YLG JGY CHZ. Contributed reagents/materials/analysis tools: DW YX. Wrote the paper: DW JXL YX.

References

1. Zheng CH, Huang DS, Zhang L, Kong XZ (2009) Tumor clustering using nonnegative matrix factorization with gene selection. IEEE Transactions on Information Technology in Biomedicine 13: 599–607. doi: 10.1109/TITB.2009.2018115 PMID: 19369170
2. Hou C, Nie F, Li X, Yi D, Wu Y (2014) Joint embedding learning and sparse regression: A framework for unsupervised feature selection. Cybernetics, IEEE Transactions on 44: 793–804.
3. Nie F, Xiang S, Jia Y, Zhang C, Yan S. Trace Ratio Criterion for Feature Selection; 2008. pp. 671–676.
4. Jauhari S, Rizvi S (2014) Mining gene expression data focusing cancer therapeutics: a digest. IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB) 11: 533–547.
5. Fa R, Nandi AK (2014) Noise resistant generalized parametric validity index of clustering for gene expression data. IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB) 11: 741–752.
6. BALADANDAYUTHAPANI V, Coombes K, Momin A (2014) Latent Feature Decompositions for Integrative Analysis of Diverse High-throughput Genomic Data. IEEE/ACM Transactions on Computational Biology and Bioinformatics: 1.
7. Mazza T, Fusilli C, Saracino C, Mazzoccoli G, Tavano F, Vinciguerra M, et al. (2015) Functional impact of autophagy-related genes on the homeostasis and dynamics of pancreatic cancer cell lines. IEEE/ACM Transactions on Computational Biology and Bioinformatics 1.
8. Fang X, Xu Y, Li X, Fan Z, Liu H, Chen Y (2014) Locality and similarity preserving embedding for feature selection. Neurocomputing 128: 304–315.
9. Nie F, Yuan J, Huang H. Optimal mean robust principal component analysis; 2014. pp. 1062–1070.
10. Liu J-X, Zheng C-H, Xu Y (2012) Extracting plants core genes responding to abiotic stresses by penalized matrix decomposition. Computers in biology and medicine 42: 582–589.
11. Zheng CH, Ng To-Yee V, Zhang L, Shiu CK, Wang HQ (2011) Tumor Classification Based on Non-Negative Matrix Factorization Using Gene Expression Data. IEEE Transactions on NanoBioscience 10: 86–93. doi: 10.1109/TNB.2011.2144998 PMID: 21742573
12. Livak KJ, Schmittgen TD (2001) Analysis of Relative Gene Expression Data Using Real-Time Quantitative PCR and the 2< sup>– ΔΔCT Method. methods 25: 402–408. PMID: 11846609
13. Liu J-X, Liu J, Gao Y-L, Mi J-X, Ma C-X, Wang D (2014) A Class-Information-Based Penalized Matrix Decomposition for Identifying Plants Core Genes Responding to Abiotic Stresses. PloS one 9: e106097. doi: 10.1371/journal.pone.0106097 PMID: 25180509
14. Liu J, Liu J-X, Gao Y-L, Kong X-Z, Wang X-S, Wang D (2015) A P-Norm Robust Feature Extraction Method for Identifying Differentially Expressed Genes. PloS one 10: e0133124. doi: 10.1371/journal.pone.0133124 PMID: 26201006
15. Witten DM, Tibshirani R, Hastie T (2009) A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. Biostatistics: kxp008.
16. Chen D, Cao X, Wen F, Sun J. Blessing of dimensionality: High-dimensional feature and its efficient compression for face verification; 2013. IEEE. pp. 3025–3032 2013.
17. Hall P, Marron J, Neeman A (2005) Geometric representation of high dimension, low sample size data. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67: 427–444.
18. Lee DD, Seung HS. Algorithms for non-negative matrix factorization; 2001. pp. 556–562.
19. Di L, Pagan PE, Packer D, Martin CL, Akther S, Ramrattan G, et al. (2014) BorreliaBase: a phylogeny-centered browser of Borrelia genomes. BMC bioinformatics 15: 233. doi: 10.1186/1471-2105-15-233 PMID: 24994456

20. Journée M, Nesterov Y, Richtarik P, Sepulchre R (2010) Generalized power method for sparse principal component analysis. The Journal of Machine Learning Research 11: 517–553.
21. Yalavarthy PK, Pogue BW, Dehghani H, Paulsen KD (2007) Weight-matrix structured regularization provides optimal generalized least-squares estimate in diffuse optical tomography. Medical physics 34: 2085–2098. PMID: 17654912
22. Hoyer PO (2004) Non-negative matrix factorization with sparseness constraints. The Journal of Machine Learning Research 5: 1457–1469
23. Lin C-f, Wang S-d (2004) Training algorithms for fuzzy support vector machines with noisy data. Pattern recognition letters 25: 1647–1656.
24. Ferson W, Nallareddy S, Xie B (2013) The "out-of-sample" performance of long run risk models. Journal of Financial Economics 107: 537–556.
25. Nikolova M (2004) A variational approach to remove outliers and impulse noise. Journal of Mathematical Imaging and Vision 20: 99–120.
26. Ding H, Wang C, Huang K, Machiraju R (2014) iGPSe: A visual analytic system for integrative genomic based cancer patient stratification. BMC Bioinformatics 15: 203. doi: 10.1186/1471-2105-15-203 PMID: 25000928
27. Utreras F (2013) Optimal smoothing of noisy data using spline functions. SIAM Journal on Scientific and Statistical Computing 2.
28. Kong D, Ding C, Huang H. Robust nonnegative matrix factorization using l21-norm; 2011. ACM. pp. 673–682.
29. Nie F, Huang H, Cai X, Ding C (2010) Efficient and robust feature selection via joint l2, 1-norms minimization. Advances in neural information processing systems 23: 1813–1821.
30. Parra G, Bradnam K, Korf I (2007) CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics 23: 1061–1067. PMID: 17332020
31. Ortega-Martorell S, Lisboa PJ, Vellido A, Julià-Sapé M, Arús C (2012) Non-negative matrix factorisation methods for the spectral decomposition of MRS data from human brain tumours. BMC bioinformatics 13: 38. doi: 10.1186/1471-2105-13-38 PMID: 22401579
32. Liu J, Ji S, Ye J. Multi-task feature learning via efficient l 2,1-norm minimization; 2009. AUAI Press. pp. 339–348.
33. Yang S, Hou C, Zhang C, Wu Y (2013) Robust non-negative matrix factorization via joint sparse and graph regularization for transfer learning. Neural Computing and Applications 23: 541–559.
34. Nie F, Huang H, Cai X, Ding CH. Efficient and robust feature selection via joint ℓ2, 1-norms minimization; 2010. pp. 1813–1821.
35. Nyamundanda G, Gormley IC, Brennan L (2014) A dynamic probabilistic principal components model for the analysis of longitudinal metabolomics data. Journal of the Royal Statistical Society: Series C (Applied Statistics).
36. ZHANG Y, MU Z-c (2006) Ear recognition based on improved NMFSC. Journal of Computer Applications 4: 010.
37. Allen GJ, Chu SP, Schumacher K, Shimazaki CT, Vafeados D, Kemper A, et al. (2000) Alteration of stimulus-specific guard cell calcium oscillations and stomatal closing in Arabidopsis det3 mutant. Science 289: 2338–2342. PMID: 11009417
38. Jenks MA, Hasegawa PM (2008) Plant abiotic stress: John Wiley & Sons.
39. Feigelman J, Theis FJ, Marr C (2014) MCA: Multiresolution Correlation Analysis, a graphical tool for subpopulation identification in single-cell gene expression data. arXiv preprint arXiv:14072112.
40. Dinkla K, El-Kebir M, Bucur C-I, Siderius M, Smit MJ, Westenberg MA, et al. (2014) eXamine: Exploring annotated modules in networks. BMC bioinformatics 15: 201. doi: 10.1186/1471-2105-15-201 PMID: 25002203
41. Pomeroy SL, Tamayo P, Gaasenbeek M, Sturla LM, Angelo M, McLaughlin ME, et al. (2002) Prediction of central nervous system embryonal tumour outcome based on gene expression. Nature 415: 436–442. PMID: 11807556
42. Wu M-Y, Dai D-Q, Zhang X-F, Zhu Y (2013) Cancer Subtype Discovery and Biomarker Identification via a New Robust Network Clustering Algorithm. PloS one 8.