Differentially Expressed Genes Extracted by the Tensor Robust Principal Component Analysis (TRPCA) Method

Yue Hu, Jin-Xing Liu, Ying-Lian Gao, Sheng-Jun Li, and Juan Wang

School of Information Science and Engineering, Qufu Normal University, Rizhao 276826, China
Library of Qufu Normal University, Qufu Normal University, Rizhao 276826, China

Correspondence should be addressed to Jin-Xing Liu; 313339002@qq.com

Received 31 December 2018; Revised 27 April 2019; Accepted 15 May 2019; Published 2 June 2019

Academic Editor: Danilo Comminiello

Copyright © 2019 Yue Hu et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

In the big data era, sequencing technology has produced a large number of biological sequencing data. Different views of the cancer genome data provide sufficient complementary information to explore genetic activity. The identification of differentially expressed genes from multiview cancer gene data is of great importance in cancer diagnosis and treatment. In this paper, we propose a novel method for identifying differentially expressed genes based on tensor robust principal component analysis (TRPCA), which extends the matrix method to the processing of multiway data. To identify differentially expressed genes, the plan is carried out as follows. First, multiview data containing cancer gene expression data from different sources are prepared. Second, the original tensor is decomposed into a sum of a low-rank tensor and a sparse tensor using TRPCA. Third, the differentially expressed genes are considered to be sparse perturbed signals and then identified based on the sparse tensor. Fourth, the differentially expressed genes are evaluated using Gene Ontology and Gene Cards tools. The validity of the TRPCA method was tested using two sets of multiview data. The experimental results showed that our method is superior to the representative methods in efficiency and accuracy aspects.

1. Introduction

In the rapid development of sequencing technology, large amounts of gene expression data have been generated. Cancer (malignant tumor) is the common type of disease in this era and poses a serious threat to human health. Researchers in molecular biology have shown that the human body carries more than 20000 different genes, but few are associated with biological processes. Therefore, the study of gene expression data has become an important trend. The analysis of expression data can help explore the origin of life and understand differences between individuals. Genes are common determinants of in vivo cancer or tumor onset, which are identified as abnormally expressed. Therefore, on the one hand, identifying differentially expressed genes can help people explore the association between different diseases. On the other hand, this information can provide a theoretical basis for medical studies and clinical diagnosis. Techniques to screen differentially expressed genes from gene expression data have gained much attention [1]. These data consist of tens of thousands of genes and hundreds of samples.

It is generally known that analysis of gene expression data is a typical high-dimension-small-size-sample (HD3S) problem. Many researchers have found that only a small portion of genes play key roles in biological processes [2]. Therefore, it is a very great challenge to identify genes related to diseases.

The selection of differentially expressed genes or feature selection is the identification of n from m features. Genomic data are usually contaminated by noise, and thus the identification of differentially expressed genes necessitates the premise of satisfying the system’s optimization criteria [3]. This process requires the identification of disease-related genes and reduces noise, which is an HD3S problem. Moreover, the data are embedded in a high-dimensional space with a low-dimensional flow pattern, so dimension reduction has become an indispensable task [4]. Currently, despite many classic methods that are effectively applied to genomic data, there is still room for improvement. Principal component analysis (PCA) [5] is the most popular method for linear dimension reduction and data analysis. Despite the slight damage from small amounts of noise, the efficiency and effectiveness of PCA data processing are considerable.
An important issue is that PCA is vulnerable to severely damaged data and outliers, especially when the actual data are ubiquitous. In addition, the low-rank representation (LRR) method is also very popular for feature selection. It can decompose the original matrix into the sum of the low-rank matrix and the sparse matrix [6]. Not only in terms of feature selection but also in other directions such as video background separation [7], subspace segmentation [8], image clustering [9], and image denoising [10], the LRR method is also widely used. Although the experimental results of the LRR method are superior, it still has some disadvantages. At present, to solve the above problems, many methods have been proposed to reduce the complexity of the data. The robust principal component analysis (RPCA) [11] method, which was recently proposed to have a strong integrity guarantee, is the first polynomial-time algorithm. Let the size of a given data matrix $A$ be $n_1 \times n_2$, which can be decomposed into the sum of matrices $M_0$ and $P_0$, where $M_0$ is a sparse matrix and $P_0$ is a low-rank matrix. It cannot consider the internal structure of gene expression data, thus overlooking some important information.

The disadvantage of RPCA is that it is a single-view model and can only handle two-order data. In the real world, multidimensional data exist anywhere and are also known as tensors. Like a color image, it is three-dimensional data containing columns, rows, and color models. For another example, the grayscale video contains two spatial vectors and one time vector. A third-order tensor represents the status of a social network. Rows and columns represent different social workers, and the third dimension represents the social modes between them, such as Twitter, Facebook, and the WeChat. To use the RPCA method, preprocessing must be performed to convert multiway data into matrix mode. However, this operation will result in a loss of key information, resulting in poor performance of the experiment. To avoid this problem, many researchers have proposed tensor methods to deal with multiway data. These methods deal with the relationships between the internal structures of tensor data.

To overcome the limitations of the matrix dimension, Lu et al. proposed a tensor robust principal component analysis (TRPCA) method, which extends the known RPCA method to the tensor case [12]. This method has been proven to be effective in many areas, such as image denoising, noise removal, and video separation monitoring [13]. Work [14] proposed a Bayesian robust tensor factorization (BRTF) generation model, which aims to capture global information and local sparse tensor information. The multiview gene expression data are similar to the components in the above fields. Their sparse disturbance signals are similar to the noise in the image.

Benefitting from the development of the big data era, multiple attributes of an object can be easily obtained. For example, an object can contain the color view and the shape view; in camera views of multiple angles of a single object, each camera's characteristics are independent of each other; and the same gene has different levels of gene expression in different cancers. Multiview data contain more information than single-view data for better performance, rather than relying on single-view data [15]. Therefore, the emergence of multiview data has led to the emergence of multiview models. Most available feature selection methods are single-view models, and multiview models are few and far between.

To overcome the above problems, we proposed the TRPCA method to solve multiview data. Although the TRPCA method has been effectively applied for image recovery and removal of random noise from face images, its validity for gene expression data requires confirmation. Gene expression data are close to some low-dimensional subspaces, so it is natural to approximate nondifferentially expressed gene data to a low data rank. Although the human body contains tens of thousands of genes, only a few are in fact related to biological processes. Therefore, the differentially expressed genes are treated as sparsely disturbed signals in the original data.

In this paper, based on TRPCA, a novel approach is proposed for the identification of differentially expressed genes. Unlike the RPCA method, the TRPCA method extends to multiway data. It preserves the intrinsic geometry of the data. Thus, it can select more differentially expressed genes. Nondifferentially expressed genes are considered to be low-rank tensor signals, and differentially expressed genes are treated as sparsely turbulent signals. In the multiview data, tensor $\mathcal{A}$ is decomposed into the sum of the low-rank tensor $\mathcal{M}_0$ and the sparse tensor $\mathcal{P}_0$. Next, differentially expressed genes are identified based on the sparse tensor $\mathcal{P}_0$. Finally, differentially expressed genes are evaluated using the Gene Ontology and the Gene Cards tools.

The main contributions of this paper are as follows.

First, multiview data are innovatively constructed from a variety of cancer gene expression data, attempting to explore the intrinsic geometry structure between coexpressed genes by tensor.

Second, we proposed, for the first time, an approach and idea based on TRPCA, which aims to identify differentially expressed genes in a multiview model. In TRPCA framework, the sparse component contributes to capturing multiple interactions among views, which better preserve the complementary information.

Third, a large number of feature selection experiments are provided to identify differentially expressed genes. The selection of differentially expressed genes can be performed because the sparse tensor can restore common characteristic genes from multiview information. Marking these genes as listed genes will facilitate the diagnosis and treatment of cancers.

The rest of the paper is arranged in the following manner. The second section introduces the tensor-related symbol definitions, as well as detailed description of the TRPCA method. The selection results and analysis of differentially expressed genes are presented in Section 3. Finally, the main points and the future work are summarized.

2. Materials and Methods

2.1. Notations and Preliminaries. In this subsection, some symbols and definitions are given. Throughout this subsection, all symbols are defined according to [12]. We define the tensor symbol in bold Euler script letters, for example, $\mathcal{A}$. 

Complexity
Matrices are represented in bold capital letters, such as \( A \). By analogy, vectors are represented in bold lowercase letters, for instance, \( \mathbf{a} \). Lowercase letters are used to represent scalars such as \( a \). We define the identity matrix as \( I_n \) and the size as \( n \times n \). In this paper, \( \mathbb{R} \) and \( \mathbb{C} \) are used to represent the field of real and complex numbers, respectively. In the third-order tensor, we use \( |i,j,k| \)-elements as \( B_{ijk} \) or \( b_{ijk} \). The MATLAB notations \( \mathbf{B}(i,:,:), \mathbf{B}(;i,:), \) and \( \mathbf{B}(::,i) \) are used to represent horizontal and frontal slices of the \( i \)-th level of the tensor, respectively. Additionally, the tensor front slice \( \mathbf{B}(i,:,:) \) can also be represented by \( \mathbf{B}^{(i)} \). The tensor tube is interpreted as \( \mathbf{B}(i,:,:,:) \).

We specify that the \( L_1 \)-norm is expressed as \( \| \mathbf{B} \|_1 = \sum_{i,j,k} |b_{ijk}| \), the Frobenius norm is defined as \( \| \mathbf{B} \|_F = \sqrt{\sum_{i,j,k} |b_{ijk}|^2} \). The norm of these tensors can be reduced to the norm of the matrices and vectors, when \( \mathbf{B} \) becomes a vector or a matrix. Let \( \mathbf{B} \in \mathbb{R}^{n_1 \times n_2 \times n_3} \), the tensor nuclear norm of \( \mathbf{B} \) denoted by \( \| \mathbf{B} \|_* \). \( \| \mathbf{B} \|_* \) is defined as the average of the sum of the nuclear norms of \( \mathbf{B} \) for each front slice, such as \( \| \mathbf{B} \|_* = (1/n_3) \sum_{i,j} \| \mathbf{B}^{(i)} \|_* \) [12, 16]. The same definition has been theoretically proven in the work [17]. Therefore, it guarantees the theoretical analysis and optimization proof of tensor nuclear norm based on the TRPCA model [18]. We use the fft function in MATLAB to compute the tensor \( \mathbf{B} \in \mathbb{R}^{n_1 \times n_2 \times n_3} \) by \( \mathbf{B} = \text{fft}([\mathbf{B}(:,:,:)]',3) \). The meaning is the result of the Fourier transform of tensor \( \mathbf{B} \) along the third dimension. Similarly, we can calculate \( \mathbf{B} \) to obtain \( \mathbf{B} \) by \( \text{fft}([\mathbf{B}(:,:,:)]',3) \).

Specifically, we define \( \mathbf{B} \) as a block diagonal matrix, where each block diagonally is labeled as \( \mathbf{B}^{(i)} \) of \( \mathbf{B} \), such as

\[
\mathbf{B} = \text{b diag} (\mathbf{B}) = \begin{bmatrix}
\mathbf{B}^{(1)} \\
\mathbf{B}^{(2)} \\
\vdots \\
\mathbf{B}^{(n_3)}
\end{bmatrix}.
\]

An important concept is the block-circulant matrix, which is also known as the new matrix of tensors. The novel tensor-tensor product is defined based on this concept. In concrete terms, the size of the block circulation matrix of a tensor \( \mathbf{B} \) is \( n_1 n_3 \times n_2 n_3 \), as shown below

\[
\text{bcirc} (\mathbf{B}) = \begin{bmatrix}
\mathbf{B}^{(1)} & \mathbf{B}^{(n_3)} & \ldots & \mathbf{B}^{(2)} \\
\mathbf{B}^{(2)} & \mathbf{B}^{(1)} & \ldots & \mathbf{B}^{(3)} \\
\vdots & \vdots & \ddots & \vdots \\
\mathbf{B}^{(n_3)} & \mathbf{B}^{(n_3-1)} & \ldots & \mathbf{B}^{(1)}
\end{bmatrix},
\]

where the tensor \( \mathbf{B} \in \mathbb{R}^{n_1 \times n_2 \times n_3} \).

In addition, we also define the following operations [19]:

\[
\text{unfold} (\mathbf{B}) = \begin{bmatrix}
\mathbf{B}^{(1)} \\
\mathbf{B}^{(2)} \\
\vdots \\
\mathbf{B}^{(n_3)}
\end{bmatrix},
\]

\[
\text{fold} (\text{unfold} (\mathbf{B})) = \mathbf{B}.
\]

More directly, it can be expressed in Figure 1, where \( \mathbf{A} \in \mathbb{R}^{n_1 \times n_2 \times n_3}; A_1 = \mathbf{A}( :, :, i) \).

The tensor-tensor product is an algebraic operation defined between two 3-order tensors, which is defined as \( \mathbf{A} \ast \mathbf{B} = \text{fold} (\text{bcirc} (\mathbf{A}) \ast \text{unfold} (\mathbf{B})) \).

Let \( \mathbf{B} \) be a tensor in the real number range, with a size of \( n_1 \times n_2 \times n_3 \). Then, tensor \( \mathbf{B} \) can be decomposed into

\[
\mathbf{B} = \mathbf{U} \ast \mathbf{S} \ast \mathbf{V}^* ,
\]

where \( \mathbf{U} \) and \( \mathbf{V} \) are orthogonal tensors with sizes of \( n_1 \times n_1 \times n_3 \) and \( n_2 \times n_2 \times n_3 \), respectively. \( \mathbf{S} \) is an F-diagonal tensor with size \( n_1 \times n_2 \times n_3 \) in the real domain.

Figure 2 shows the t-SVD decomposition process for the \( n_1 \times n_2 \times n_3 \) tensor. Thus, t-SVD can be perfectly derived from the matrix SVD in the Fourier domain.

### 2.2. Related Methods and Works

For the processing of high-dimensional and small-sample data, the most commonly used method is RPCA. Let the size of the given data matrix \( \mathbf{A} \) be \( n_1 \times n_2 \), which can be decomposed into the sum of tensors \( \mathbf{M} \) and \( \mathbf{P} \), where \( \mathbf{M} \) is a sparse matrix and \( \mathbf{P} \) is a low-rank matrix. The objective function can be expressed as

\[
\min_{\mathbf{M},\mathbf{P}} \| \mathbf{M} \|_* + \lambda \| \mathbf{P} \|_1,
\]

s.t. \( \mathbf{A} = \mathbf{M} + \mathbf{P} \),

where \( \| \mathbf{M} \|_* \) represents the matrix nuclear norm (the sum of singular values of \( \mathbf{M} \)), \( \| \mathbf{P} \|_1 \) represents the value of the \( L_1 \)-norm (the sum of the absolute values of all entries in \( \mathbf{P} \)), and parameter \( \lambda = 1/\sqrt{\max(n_1, n_2)} \). RPCA and its extensions
have been successfully applied in image segmentation [20], background models [21], and the extraction of characteristic genes from genomic data [9].

Under ideal conditions, we expect to extend the conditions for recovering low-rank matrices to three-dimensional tensors. The tools and methods used to recover matrices can also be extended to the best. However, this achievement is not simple. The numerical algebra of tensor data are filled with hardness results [22]. The definition of a tensor rank is crucial for restoring the tensor effect. The rank of the matrix has always existed. Taking [24] as an example, the CP rank component and hyperspectral images from genomic data [9].

The Tucker rank convex surrogate refers to it.

\[
\min_{\mathcal{A}} \sum_{i=1}^{k} \| \mathbf{A}^{(i)} \|_2^2,
\]

s.t. \( P_{\Omega}(\mathcal{A}) = P_{\Omega}(\mathcal{A}_0) \).

In another work [30] based on the TRPCA model, the SNN algorithm is proposed, and its objective function is

\[
\min_{\mathcal{M}, \mathcal{P}} \sum_{i=1}^{k} \lambda_i \| \mathcal{M}^{(i)} \|_2^2 + \| \mathcal{P} \|_1 + \frac{\tau}{2} \| \mathcal{M} \|_F^2 + \frac{\tau}{2} \| \mathcal{P} \|_F^2,
\]

s.t. \( \mathcal{A} = \mathcal{M} + \mathcal{P}, \mathcal{A} \in R^{n_1 \times n_2 \times n_3} \)

where \( \| \mathcal{P} \|_1 \) is the \( L_1 \)-norm and refers to the sum of the absolute values of all elements in \( \mathcal{P} \). This also ensures that the tensor can be reliably restored to meet certain inconsistencies.

2.3. Objective Function and Solutions Process. The tensor method means that, for a given three-order tensor \( \mathcal{A} \), it can be decomposed into \( \mathcal{A} = \mathcal{M}_0 + \mathcal{P}_0 \), where \( \mathcal{M}_0 \) is the low-rank component and \( \mathcal{P}_0 \) is the sparse component. Under certain suitable assumptions, this problem can be solved by convex optimization problems. The objective function can be expressed as the sum of the weights of the tensor nuclear norm and the \( L_1 \)-norm, i.e.,

\[
\min_{\mathcal{M}, \mathcal{P}} \| \mathcal{M} \|_1 + \lambda \| \mathcal{P} \|_1
\]

s.t. \( \mathcal{A} = \mathcal{M} + \mathcal{P} \),

where \( \| \mathcal{M} \|_1 \) represents the nuclear norm of the tensor \( \mathcal{M} \) and \( \| \mathcal{P} \|_1 \) indicates the \( L_1 \)-norm of the tensor \( \mathcal{P} \). The choice of parameter \( \lambda \) is \( \lambda = 1 / \sqrt{\max(n_1, n_2)} \). It is observed that when \( n_2 \) drops to 1, TRPCA degenerates to RPCA, so TRPCA is also seen as an extension of RPCA.

Due solely to a robust principal component analysis [31], the status of exact recovery cannot be ignored. This situation also applies to TRPCA. For example, define \( \mathcal{A} \) as a tensor that satisfies the following condition \( \mathcal{A} = e_1 \mathcal{M} + \mathcal{P} \), where \( \mathcal{M} \) is the low-rank component and \( \mathcal{P} \) is the sparse component. Therefore, to avoid this thorny problem, we must assume that the low-rank component of \( \mathcal{M}_0 \) is not sparse.

The most common algorithm for solving RPCA-related problems is the Alternating Direction Method of Multipliers (ADMM) algorithm [16]. Therefore, the ADMM algorithm is also used when solving the TRPCA-related convex function problem in this paper. The main content is that the values of \( \mathcal{M}_{k+1} \) and \( \mathcal{P}_{k+1} \) need to be updated simultaneously. It is clear that the cost of the iteration is mainly reflected in the update of \( \mathcal{M}_{k+1} \), because this process requires the solution to FFT and \( n_2 \) SVD of \( n_1 \times n_2 \) matrices.

For (9), the Lagrangian multiplier was introduced to eliminate the equality constraints. According to a previous
work [16], the ADMM algorithm on the Lagrangian function can be expressed as follows:

\[
L(\mathcal{M}, \mathcal{P}, y, \mu) = \| \mathcal{M} \|_* + \lambda \| \mathcal{P} \|_1 \\
+ y^T (\mathcal{M} + \mathcal{P} - \mathcal{A}) \\
+ \frac{\mu}{2} \| \mathcal{M} + \mathcal{P} - \mathcal{A} \|_F^2, \tag{10}
\]

where $\mu$ is a scalar parameter and $\| \cdot \|_F$ is a Frobenius norm.

After several iterations of the TRPCA method, the original tensor was decomposed. After taking the partial derivatives of the ADMM algorithm, we find partial derivatives of $\mathcal{M}$ and $\mathcal{P}$, respectively. Let the partial derivatives be equal to zero and the final iteration formula be

\[
\mathcal{M}_{k+1} = D_{\lambda/\mu} \left( \mathcal{A} - \mathcal{P}_k - \frac{y_k}{\mu_k} \right), \tag{11}
\]

\[
\mathcal{P}_{k+1} = S_{\lambda/\mu} \left( \mathcal{A} - \mathcal{M}_{k+1} - \frac{y_k}{\mu_k} \right), \tag{12}
\]

\[
y_{k+1} = y_k + \mu_k \left( \mathcal{M}_{k+1} + \mathcal{P}_{k+1} - \mathcal{A} \right), \tag{13}
\]

\[
\mu_k = \min \left( \rho \mu_k, \mu_{\text{max}} \right). \tag{14}
\]

The details of the solution algorithm can be found in Algorithm 1.

### 2.4. The TRPCA Model of Gene Expression Data

Considering the gene expression data $\mathcal{A}$ with size $n_1 \times n_2 \times n_3$, each row of the frontal slice in $\mathcal{A}$ represents transcript reactions of one gene in all $n_3$ samples, and each column represents the gene expression level of $n_1$ genes in one sample. Without loss of generality, the matrix size of each front slice should be $n_1 \times n_2$, so this is a classic HDS3 problem.

The purpose of using TRPCA to model multiview data is to discover important genes. As mentioned above, it is reasonable to treat important genes as sparse signals. Thus, the differential expression is regarded as the sparse disturbance signal $\mathcal{P}$, and the nondifferentially expression is regarded as the low-rank tensor $\mathcal{M}$. From this perspective, differentially expressed genes by various cancers can be identified from the sparse disturbance signal $\mathcal{P}$. The multiview model of TRPCA is shown in Figure 3. Three dimensions represent genes, samples, and disease types. Each front slice matrix of the input tensor represents the expression level of all samples of a cancer for all genes, and it is clear that different frontal slices represent different cancer types. The solid color represents the data point equal to zero or close to zero, and a colored noise point denotes a disturbance signal. As shown in Figure 3, the differentially expressed genes in the tensor $\mathcal{P}$ can be recovered from the original tensor gene expression data.

Assume that the tensor decomposition $\mathcal{A} = \mathcal{M} + \mathcal{P}$ has been completed by the TRPCA model. By selecting the appropriate parameter $\lambda$, the sparse disturbance signals can be obtained in the sparse tensor. For example, most of the entries in the sparse tensor is zero or close to zero, and genes that are nonzero can be considered as differentially expressed genes.

### 2.5. Identification of Differentially Expressed Genes

The low-rank tensor $\mathcal{M}$ and sparse tensor $\mathcal{P}$ can be obtained in the experiment. By using the sparse tensor $\mathcal{P}$, differentially expressed genes can be selected. Because we regard the important genes as sparse signals, the differentially expressed genes are treated as sparse perturbation signals. Therefore, differentially expressed genes can be extracted by the sparse
The result of the tensor is to sum each slice and obtain a new vector. Then, the new vectors are arranged in descending order:

\[
\hat{\mathbf{p}} = (\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_m). \tag{16}
\]

Next, we perform the following operations on the descending vectors, filter out the top 500 maximum values, and extract the corresponding genes. Without losing generality, the higher the gene's ranking, the more likely it is to become a differentially expressed gene. Therefore, we selected genes that were only related to the first number in the vector as differentially expressed genes. An important tool for our analysis of genomic data is GO::TermFinder [32]. GO::TermFinder is open source software in which Gene Ontology information and rich Gene Ontology terms can be accessed. When we infuse the gene name into the GO::TermFinder tool, this tool generates a rich vocabulary associated with that gene. The table contains rich biological explanations related to this gene. Performance comparisons of these methods were evaluated using P-values and hit counts. The P-values and the number of input genes were mainly used to measure the superiority of the experiment. The experimental method corresponding to the smaller P-value indicates that the effect of differentially expressed genes is better. The thresholds of its parameters are set in a uniform way: the maximum value of p is set to 0.01.

### 3. Results and Discussion

#### 3.1. The Composition of the Dataset.

The Cancer Genome Atlas (TCGA) maps the genomic variation of cancer using genomics analysis techniques. The TCGA project included the 33 most common cancers and more than 11,000 tumor samples for sequencing. An in-depth study of this information will inspire future clinical trials and treatments. In this paper, we used two multiview datasets to analyze the effectiveness of the proposed method. These multiviews included various cancer types: colon adenocarcinoma (COAD), head and neck squamous cell carcinoma (HNSC), esophageal carcinoma (ESCA), and pancreatic adenocarcinoma (PAAD).

To ensure the versatility of the experiments, the experimental materials were composed of cancer data in the TCGA database (https://cancergenome.nih.gov/).

The information for multiple views in multiview data is rich, and it is significant to study this information in depth. However, one challenging issue is the heterogeneity between different views. In this experiment, we performed the preprocessing for multiview data as follows. First, data from different sources and characteristics were logically synthesized organically. In this process, the common gene parts of the data were extracted, and these common genes showed different expression levels in response to the same type of pathogen. Second, all public genes were aligned in alphabetical order to ensure the validity of the multiview data tags.

Under certain conditions, the tensor can be viewed as an extension of the matrix and vector. When the third dimension of the tensor drops to 1, the tensor degenerates to the matrix. When both the second and third dimensions of the tensor are reduced to 1, the tensor is reduced to a vector. In this paper, tensor refers to the three-order tensor, aiming to study the spatial structure between three-dimensional data and then explore the intrinsic links between various cancer diseases. The frontal section of multiview data is composed of gene expression data from different cancers. Subject to the tensor dimension, the number of samples is determined by the cancer with the fewest number of samples. The details of more multiview datasets are summarized in Table 1. The original tensor was decomposed into a sum of a low-rank tensor and a sparse tensor using TRPCA. The differentially expressed genes were considered to be sparse perturbed signals and then identified based on the sparse tensor.

In many cases, there is a certain connection between substances with similar behavior or changes. There are also many common characteristics between the occurrence of cancer, such as the commonality of disease-causing genes and the upregulation and downregulation of genes. Multiview gene expression data provide support for this study. In this paper, gene expression data in multiview data contain views from multiple cancers. These differences exist in the spatial structure between tensor slices, while the TRPCA method can decompose the original tensor into the low-rank tensor and the sparse tensor without destroying the internal structure of the tensor. The geometry between the internal slices of the tensor obtained from the TRPCA decomposition is not destroyed, retaining valid information. This not only improves the accuracy of feature selection, but also provides a new idea based on the tensor method. Therefore, the tensor-based robust principal component analysis method is able to explore more changes between its data than other methods. Extraction of differentially expressed genes under the influence of multiple views can not only elucidate the

### Table 1: Experimental data.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Number of genes</th>
<th>Number of samples</th>
<th>Dimensions</th>
</tr>
</thead>
<tbody>
<tr>
<td>COAD_HNSC_ESCA_GE</td>
<td>20502</td>
<td>192×3=576</td>
<td>20502×192×3</td>
</tr>
<tr>
<td>COAD_HNSC_PAAD_GE</td>
<td>20502</td>
<td>180×3=540</td>
<td>20502×180×3</td>
</tr>
</tbody>
</table>
commonality of disease-causing genes between cancers but also establish the correlation between cancers. Marking these coexpression genes in the list of genes will provide a targeted orientation for cancer detection.

3.2. Results for the COAD_HNSC_PAAD Data. Gene Ontology is composed of three parts, biological processes, molecular functions, and cellular components. The differentially expressed genes are placed in the GO tool. After the P-values were obtained from the GO tool, they were ranked in ascending order according to the size of the P-values, and the first ten genes were selected for display. Better experimental results are shown in italic type. Table 2 lists the top ten genes generated by the TRPCA method and RPCA, LLRR, PCA, and BRTF methods for the COAD_HNSC_PAAD_GE dataset. The P-value indicates the enrichment degree of the gene. The P-value is the probability or opportunity to observe at least x of the total n genes in a list annotated to a particular GO term, given the proportion of genes annotated to the GO term in the entire genome. The closer the P-value is to zero, the more significant is the specific GO term associated with the genome. The P-values of these 10 genes revealed that the proposed method is superior to other methods. Specifically, the P-value of GO:0006614 was 6.49E-74, which is much smaller than the P-values of other methods. In addition, the largest hit value in genetic terms was also detected in our method. There were 94 genes in the GO:0006614 terminology, and RPCA, LLRR, PCA, and BRTF could detect 51, 51, 60, and 55 genes, respectively. However, 61 genes were identified using the TRPCA method. The corresponding name for GO:0072599 was the establishment of protein localization to the endoplasmic reticulum. It contained TNF, TP53, and other genes that are related to the occurrence of induced tumors. By comparing the P-value and the hit count, we can conclude that our experimental method is superior to other methods.

For the COAD_HNSC_ESCA_GE dataset, 500 genes extracted using TRPCA were compared with those obtained from Gene Cards for the three cancers. Two hundred fifty-five of 500 genes were associated with these three diseases. Many genes, which were previously thought to be unrelated to clinical outcomes, were identified. We list the top 10 differentially expressed genes with higher correlation scores in Table 3. Table 3 lists the gene names, related scores, related GO annotations, and related diseases. In general, among the identified differentially expressed information, the genes closely related to these three diseases were CDH1, MMP9, EPCAM, and MMP2. CCND1 and MMP1 are associated with the occurrence of COAD and HNSC. INS is associated with the development of COAD and PAAD.

Specifically, the official name CDH1 is Cadherin 1, which was the most correlated with this dataset. Recently, works [33, 34] have described mutations in CDH1 in COAD cell lines. Cytoplasmic CDH1 has independent prognostic value in PAAD and provides a new target for prognostic treatment [35]. A meta-analysis of the work [36] indicates that CDH1 promoter methylation is associated with HNSC risk and can be used as a valuable diagnostic biomarker for HNSC. In summary, CDH1 is related to the occurrence of these three cancers. In addition, MMP9 has been identified in a variety of malignancies [37] and as a potential marker for the prognosis of HNSC [38]. Dysregulated MMP9 expression induces invasive growth and metastasis of PAAD [39]. Expression of MMP9 is elevated in a variety of inflammatory and oncological indications and is evident in colitis and colorectal cancer [40]. MMP9 has a huge impact on the occurrence and treatment of three types of cancer, and thus MMP9 is closely related to this dataset. In summary, there was a close correlation between the differentially expressed genes and the cancers contained in the dataset, demonstrating the accuracy of our method.

3.3. Results for the COAD_HNSC_ESCA Data. The results of the experiment using the COAD_HNSC_ESCA dataset are listed in Table 4. TRPCA was compared with the other three methods, and higher expression is indicated in italic. For the gene GO:0005198, the TRPCA results were 1.30E-71, which is significantly less than 8.97E-64, 7.80E-67, 1.77E-70, and 1.02E-66. The name GO:0005198 denotes structural molecule activity. The INS gene has been identified with a multitude of mutant alleles with phenotypic effects. In terms of the number of hits, the TRPCA hit 131 genes, representing a larger number than the 123, 126, 130, and 126 using the other methods, under the premise of a total hit number of 762. The P-value of gene GO:0006413 in TRPCA was 5.51E-64, which is significantly lower than the P-values measured using the three methods RPCA, LLRR, PCA, and BRTF. A global examination of the table revealed only two IDs of the P-values measured by our method that were equal to the PCA method; the rest were better than the three experimental methods. Therefore, in summary, whether we compared the P-values or the number of hits, our experiment performed better than the other three methods.

Table 5 lists GO annotations and related diseases for the top ten differentially expressed genes screened using the TRPCA method for the COAD_HNSC_ESCA data. Of the top ten differentiated genes, most were associated with these three cancers. Overall, 8 of 10 differentially expressed genes were highly correlated with these three cancers. These 8 genes were EGFR, CDH1, ERBB2, CCND1, MMP9, EPCAM, MMP1, and MMP2, respectively. In addition, CTNNB1 was significantly expressed in patients with COAD and ESCA. The last gene, PLAU, could serve as a key biomarker for the accurate diagnosis and prognosis of HNSC, providing a potential target for clinical treatment.

From Table 5 we can see that the official name of “EGFR” is “Epidermal Growth Factor Receptor”, which is related to HNSC [41]. Its correlation score with these three diseases reached 219.78. The higher the relevant score, the greater is the correlation between the genes and the three diseases. Therefore, the correlation between the gene “EGFR” selected by tensing TRPCA and CHOL ESCA HNSC was high. The related disease of EGFR was Inflammatory Skin And Bowel Disease, Neonatal, 2, and Lung Cancer, which is related to ESCA and CHOL [42, 43]. The official name of the gene “EPCAM” is “Epithelial Cell Adhesion Molecule”, which is related to the optimal treatment of HNSC [43] The relevance score with the three diseases was 105.76. Based
Table 2: P-values and hit counts by five methods for the COAD_HNSC_PAAD data.

<table>
<thead>
<tr>
<th>ID</th>
<th>TRPCA P-value</th>
<th>Hit Count</th>
<th>RPCA P-value</th>
<th>Hit Count</th>
<th>LLRR P-value</th>
<th>Hit Count</th>
<th>PCA P-value</th>
<th>Hit Count</th>
<th>BRTF P-value</th>
<th>Hit Count</th>
<th>Count in genome</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO:0006614</td>
<td>6.49E-74</td>
<td>61</td>
<td>5.91E-56</td>
<td>51</td>
<td>5.31E-56</td>
<td>51</td>
<td>4.35E-72</td>
<td>60</td>
<td>5.60E-63</td>
<td>55</td>
<td>94</td>
</tr>
<tr>
<td>GO:0006613</td>
<td>5.10E-71</td>
<td>61</td>
<td>8.65E-54</td>
<td>51</td>
<td>7.76E-54</td>
<td>51</td>
<td>2.84E-69</td>
<td>60</td>
<td>1.53E-60</td>
<td>55</td>
<td>101</td>
</tr>
<tr>
<td>GO:0045047</td>
<td>1.24E-70</td>
<td>61</td>
<td>8.65E-54</td>
<td>51</td>
<td>7.76E-54</td>
<td>51</td>
<td>2.84E-69</td>
<td>60</td>
<td>3.25E-60</td>
<td>55</td>
<td>102</td>
</tr>
<tr>
<td>GO:0072599</td>
<td>3.78E-69</td>
<td>61</td>
<td>2.24E-52</td>
<td>51</td>
<td>2.01E-52</td>
<td>51</td>
<td>1.87E-67</td>
<td>60</td>
<td>5.89E-59</td>
<td>55</td>
<td>106</td>
</tr>
<tr>
<td>GO:007972</td>
<td>4.88E-68</td>
<td>64</td>
<td>1.03E-50</td>
<td>53</td>
<td>9.22E-51</td>
<td>53</td>
<td>4.25E-68</td>
<td>64</td>
<td>6.89E-57</td>
<td>57</td>
<td>125</td>
</tr>
<tr>
<td>GO:0000956</td>
<td>5.56E-57</td>
<td>68</td>
<td>1.27E-41</td>
<td>56</td>
<td>6.89E-43</td>
<td>57</td>
<td>4.81E-57</td>
<td>68</td>
<td>1.19E-46</td>
<td>60</td>
<td>202</td>
</tr>
<tr>
<td>GO:0006612</td>
<td>2.63E-54</td>
<td>65</td>
<td>1.81E-41</td>
<td>55</td>
<td>9.57E-43</td>
<td>56</td>
<td>4.92E-53</td>
<td>64</td>
<td>8.00E-48</td>
<td>60</td>
<td>194</td>
</tr>
<tr>
<td>GO:0006401</td>
<td>2.34E-50</td>
<td>68</td>
<td>1.85E-36</td>
<td>56</td>
<td>1.32E-37</td>
<td>57</td>
<td>2.04E-50</td>
<td>68</td>
<td>5.19E-41</td>
<td>60</td>
<td>247</td>
</tr>
</tbody>
</table>
Table 3: The top ten genes among the COAD_HNSC_PAAD data by TRPCA.

<table>
<thead>
<tr>
<th>Gene name</th>
<th>Relevance score</th>
<th>Related GO annotations</th>
<th>Related diseases</th>
</tr>
</thead>
<tbody>
<tr>
<td>CDH1</td>
<td>183.35</td>
<td>calcium ion binding and protein phosphatase binding</td>
<td>Gastric Cancer, Hereditary Diffuse and Blepharocholeodontic Syndrome 1</td>
</tr>
<tr>
<td>CTNNB1</td>
<td>168.66</td>
<td>DNA binding transcription factor activity and binding</td>
<td>Mental Retardation, Autosomal Dominant 19 and Pilomatrixoma</td>
</tr>
<tr>
<td>CCND1</td>
<td>165.91</td>
<td>protein kinase activity and enzyme binding</td>
<td>Myeloma, Multiple and Von Hippel-Lindau Syndrome</td>
</tr>
<tr>
<td>MMP9</td>
<td>139.39</td>
<td>identical protein binding and metalloendopeptidase activity</td>
<td>Metaphyseal Anadysplasia 2 and Metaphyseal Anadysplasia</td>
</tr>
<tr>
<td>EPCAM</td>
<td>114.16</td>
<td>protein complex binding</td>
<td>Diarrhea 5, with Tufting Enteropathy, Congenital and Colorectal Cancer, Hereditary Nonpolyposis, Type 8</td>
</tr>
<tr>
<td>MMP2</td>
<td>92.35</td>
<td>serine-type endopeptidase activity and metalloendopeptidase activity</td>
<td>Multicentric Osteolysis, Nodulosus, and Arthropathy and Arthropathy</td>
</tr>
<tr>
<td>PLAU</td>
<td>89.58</td>
<td>serine-type endopeptidase activity</td>
<td>Quebec platelet Disorder and Alzheimer Disease</td>
</tr>
<tr>
<td>MMP1</td>
<td>88.33</td>
<td>calcium ion binding and metalloendopeptidase activity</td>
<td>Epidermolysis Bullosa Dystrophica, Autosomal Recessive and Recessive Dystrophic Epidermolysis Bullosa</td>
</tr>
<tr>
<td>IGF2</td>
<td>88.21</td>
<td>growth factor activity and insulin receptor binding</td>
<td>Growth Restriction, Severe, with Distinctive Facies and Silver-Russell Syndrome</td>
</tr>
<tr>
<td>INS</td>
<td>86.99</td>
<td>identical protein binding and protease binding</td>
<td>Hyperproinsulinemia and Diabetes Mellitus, Insulin-dependent, 2</td>
</tr>
</tbody>
</table>

On the table, we can also observe that the GO annotation of EPCAM is protein complex binding. EPCAM, claudin-7, CO-029, and CD44v6 expression were upregulated in COAD and liver metastasis, suggesting that high EPCAM expression is associated with COAD progression [44, 45]. EPCAM expression and release into the circulation can be an effective immunotherapy for ESCA patients [46]. The expression of EPCAM on disseminated tumor cells is significantly associated with the development of lymph node metastasis and significantly reduced overall survival of ESCA patients [47]. Overexpression of EPCAM eventually leads to uncontrolled development of COAD, HNSC, and ESCA. A number of studies have shown that the selected differentially expressed genes are closely related to the disease. Thus, the proposed method is superior for feature selection.

The correlation score refers to the size of the correlation between the selected genes and corresponds to the three diseases. The larger the correlation score, the greater the correlation between the genes and the three diseases. Table 6 lists the experimental results for the COAD_HNSC_ESCA dataset, which contains the number of related genes, the mean of relevant scores and the highest correlation score. Entering the three diseases in the dataset into the Gene Cards (https://www.genecards.org/), we can download a table containing the genes and related scores associated with the diseases. The genes we identified with this table are then compared, and common items are filtered out. The related number refers to the number of hits in the table for the 500 genes identified using the method. The greater the related number, the more relevant is the gene identified by the method. The average of the related scores is the average of all related genes identified. The highest relevant score is the maximum value of the relevant score in all relevant genes.

The number of genes extracted by the TRPCA was 250, and the related numbers of RPCA, LLRR, PCA, and BRTF were 220, 246 215, and 237 respectively. Although the highest correlation score of TRPCA was the same as that of RPCA and LLRR, the mean values of the other three methods were 28.36, 29, 27.29, and 29.41 while TRPCA was 29.59. Therefore, regardless of the number of correlations, the average of the relevant scores or the highest correlation score, our method performed better than the other three methods.

As the result shows, the TRPCA method performs much better than the matrix decomposition method such as RPCA. The validity of the proposed method indicates that our approach is reasonable for processing multiview gene expression data. The reason is that the matrix decomposition method only independently performs matrix recovery on each gene expression data and it cannot use information across views, which ignoring spatial geometric information between the data. The BRTF method decomposes tensor into the low-rank tensor, the sparse tensor, and the noise tensor. Differentially expressed genes are scattered in the sparse tensor and the noise tensor, which in turn affected the feature selection accuracy. The TRPCA approach can take advantage of multidimensional structures to improve the performance, which better preserves the redundant information in multi-view data. This provides a new perspective to study multiview data. Therefore, TRPCA is an effective integration model to consider the intrinsic geometry of multiview data.

4. Conclusions

In this paper, the TRPCA method was applied to identify differentially expressed genes. It combined the TRPCA model
Table 4: P-values and hit counts using the five methods for the COAD_HNSC_ESCA data.

<table>
<thead>
<tr>
<th>ID</th>
<th>TRPCA P-value</th>
<th>Hit Count</th>
<th>TRPCA Hit Count</th>
<th>RPCA P-value</th>
<th>Hit Count</th>
<th>RPCA Hit Count</th>
<th>LLRR P-value</th>
<th>Hit Count</th>
<th>LLRR Hit Count</th>
<th>PCA P-value</th>
<th>Hit Count</th>
<th>PCA Hit Count</th>
<th>BRTF P-value</th>
<th>Hit Count</th>
<th>BRTF Hit Count</th>
<th>Count in genome</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO:0005198</td>
<td>1.30E-71</td>
<td>131</td>
<td>8.97E-64</td>
<td>123</td>
<td>7.80E-67</td>
<td>126</td>
<td>1.77E-70</td>
<td>130</td>
<td>1.02E-66</td>
<td>126</td>
<td>762</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0006664</td>
<td>1.46E-66</td>
<td>57</td>
<td>5.18E-51</td>
<td>48</td>
<td>1.60E-59</td>
<td>53</td>
<td>3.63E-61</td>
<td>54</td>
<td>1.178E-64</td>
<td>56</td>
<td>94</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0006413</td>
<td>5.51E-64</td>
<td>72</td>
<td>1.61E-41</td>
<td>55</td>
<td>3.65E-49</td>
<td>61</td>
<td>1.20E-55</td>
<td>66</td>
<td>6.43E-53</td>
<td>64</td>
<td>194</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0003723</td>
<td>4.51E-63</td>
<td>173</td>
<td>1.44E-29</td>
<td>125</td>
<td>2.10E-40</td>
<td>142</td>
<td>1.56E-64</td>
<td>175</td>
<td>2.73E-40</td>
<td>142</td>
<td>1632</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0000184</td>
<td>4.38E-61</td>
<td>59</td>
<td>3.17E-47</td>
<td>50</td>
<td>7.46E-55</td>
<td>55</td>
<td>2.64E-56</td>
<td>56</td>
<td>2.12E-59</td>
<td>58</td>
<td>121</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0070972</td>
<td>5.13E-60</td>
<td>59</td>
<td>2.30E-46</td>
<td>50</td>
<td>7.19E-54</td>
<td>55</td>
<td>5.81E-60</td>
<td>59</td>
<td>2.447E-58</td>
<td>58</td>
<td>125</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0070973</td>
<td>7.65E-45</td>
<td>60</td>
<td>2.11E-34</td>
<td>51</td>
<td>3.62E-40</td>
<td>56</td>
<td>3.40E-41</td>
<td>57</td>
<td>none</td>
<td>none</td>
<td>116</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 5: The top ten genes in the COAD_HNSC_ESCA dataset by TRPCA.

<table>
<thead>
<tr>
<th>Gene</th>
<th>Relevance score</th>
<th>Related GO Annotations</th>
<th>Related Diseases</th>
</tr>
</thead>
<tbody>
<tr>
<td>EGFR</td>
<td>219.78</td>
<td>identical protein binding and protein kinase activity</td>
<td>Inflammatory Skin And Bowel Disease, Neonatal, 2 and Lung Cancer.</td>
</tr>
<tr>
<td>CDH1</td>
<td>182.64</td>
<td>calcium ion binding and protein phosphatase binding</td>
<td>Gastric Cancer, Hereditary Diffuse and Blepharocheliodontic Syndrome 1</td>
</tr>
<tr>
<td>ERBB2</td>
<td>170.81</td>
<td>identical protein binding and protein kinase activity</td>
<td>Glioma Susceptibility 1 and Lung Cancer</td>
</tr>
<tr>
<td>CTNNB1</td>
<td>165.7</td>
<td>DNA binding transcription factor activity and binding</td>
<td>Mental Retardation, Autosomal Dominant 19 and Pilomatrixoma</td>
</tr>
<tr>
<td>CCND1</td>
<td>165.44</td>
<td>protein kinase activity and enzyme binding</td>
<td>Myeloma, Multiple and Von Hippel-Lindau Syndrome 1</td>
</tr>
<tr>
<td>MMP9</td>
<td>135.93</td>
<td>identical protein binding and metalloendopeptidase activity</td>
<td>Metaphyseal Anadysplasia 2 and Metaphyseal Anadysplasia</td>
</tr>
<tr>
<td>EPCAM</td>
<td>105.76</td>
<td>protein complex binding</td>
<td>Diarrhea 5, With Tufting Enteropathy, Congenital and Colorectal Cancer, Hereditary Nonpolyposis, Type 8</td>
</tr>
<tr>
<td>MMP1</td>
<td>98.28</td>
<td>calcium ion binding and metallopeptidase activity</td>
<td>Epidermolysis Bullosa Dystrophica, Autosomal Recessive and Recessive Dystrophic Epidermolysis Bullosa</td>
</tr>
<tr>
<td>MMP2</td>
<td>89.9</td>
<td>serine-type endopeptidase activity and metallopeptidase activity</td>
<td>Multicentric Osteolysis, Nodulosis, And Arthropathy and Arthropathy</td>
</tr>
<tr>
<td>PLAU</td>
<td>87.24</td>
<td>serine-type endopeptidase activity and metallopeptidase activity</td>
<td>Quebec Platelet Disorder and Alzheimer Disease</td>
</tr>
</tbody>
</table>

Table 6: Relevances scores.

<table>
<thead>
<tr>
<th></th>
<th>TRPCA</th>
<th>RPCA</th>
<th>LLRR</th>
<th>PCA</th>
<th>BRTF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Related number</td>
<td>250</td>
<td>220</td>
<td>246</td>
<td>215</td>
<td>237</td>
</tr>
<tr>
<td>Average of related scores</td>
<td>29.59</td>
<td>28.36</td>
<td>29.00</td>
<td>27.29</td>
<td>29.41</td>
</tr>
<tr>
<td>Highest relevant score</td>
<td>219.78</td>
<td>219.78</td>
<td>219.78</td>
<td>182.64</td>
<td>182.64</td>
</tr>
</tbody>
</table>

References


