BMC Bioinformatics

RESEARCH

Open Access

A truncated nuclear norm and graph-Laplacian regularized low-rank representation method for tumor clustering and gene selection

Qi Liu*

From The International Conference on Data Science, Analytics, and Engineering (IDSAE) 2020/2021 Virtual. 24-25 January 2021

*Correspondence:
liuqi_workmail@126.com
College of Computer Science and Engineering, Shandong University of Science and Technology, Qingdao, China

Abstract

Background: Clustering and feature selection act major roles in many communities. As a matrix factorization, Low-Rank Representation (LRR) has attracted lots of attentions in clustering and feature selection, but sometimes its performance is frustrated when the data samples are insufficient or contain a lot of noise.

Results: To address this drawback, a novel LRR model named TGLRR is proposed by integrating the truncated nuclear norm with graph-Laplacian. Different from the nuclear norm minimizing all singular values, the truncated nuclear norm only minimizes some smallest singular values, which can dispel the harm of shrinkage of the leading singular values. Finally, an efficient algorithm based on Linearized Alternating Direction with Adaptive Penalty is applied to resolving the optimization problem.

Conclusions: The results show that the TGLRR method exceeds the existing state-of-the-art methods in aspect of tumor clustering and gene selection on integrated gene expression data.

Keywords: Low-rank representation, Graph-Laplacian, Truncated nuclear norm, Clustering, Gene selection

Background

In most countries, cancer is the first or second cause of death [1]]. Thus, it is a hot topic to prevent and cure cancer effectively in the medical field. Genes can regulate critical movements of organisms, even including the emergence of cancer [2]. As the improvement of gene sequencing technology, plenty of genomic data are available, which is conducive to researching the pathogenesis of cancer [3]. However, the majority of genomic data have the features of high dimension and small sample, which

hinders the advances in medicine studies [4, 5]. Evidently, data dimension reduction acts a momentous role in the process of genomic data analysis.

Data dimension reduction aims to get a significant low-dimensional representation of high-dimensional data, remove redundant features and prevent overfitting. Therefore, it has achieved great successes in many areas, such as characteristic gene selection [6], image analysis [7], and text documents [8]. Principal Component Analysis (PCA) is one of the most classic linear dimension reduction methods [9]. Based on the high efficiency of PCA, it has been widely used on different kinds of data and developed in many fields [5, 10, 11]. To boost the robustness of PCA, Candès et al. developed a new PCA in [10], called Robust PCA (RPCA), which is exploited for background modeling from video and analyzing face images. Moreover, in [5], RPCA is exploited for discovering differentially expressed genes by Liu et al. Though, the above PCA methods all have obtained excellent results, the performance of these methods is corrupted with the noisy observation data.

In [12], Wright et al. introduced a low-rank matrices recovery approach for removing the noise of data. Then, PCA is applied to the low-rank matrix. The robustness of data processing is enhanced significantly through the approach in [12].

The experimental data $\mathbf{X}$ are usually obtained from a union of multiple subspaces $S = \sum S_1, S_2, \ldots, S_k$ rather than a single space, where S_i indicates low-dimensional space hidden in high-dimensional space [13–15]. Since these methods related to PCA prefer to research the data obtained from a single low-dimensional space, Liu et al. proposed a low-rank representation (LRR) model that can excavate the global distribution between data points to study $\mathbf{X}$ [16]. LRR strives to look for the lowest rank matrix representation about original data and has got brilliant results in several applications [16, 17]. However, LRR still exists a few shortages, for instance it cannot reveal the local manifold structures of data obtained from a non-linear low-dimensional manifold. Joyfully, various manifold learning models have been put forward, such as ISOMAP [18], Laplacian Eigenmap (LE) [19], Locally Linear Embedding (LLE) [20], and graph-Laplacian regularization [21].

A graph-Laplacian regularized LRR (LLRR) model [14] was developed, which introduces the graph regularization into LRR. In LLRR model, the useful rules hiding among the data points including the global geometric structure and the internal similarity information are all seized. LLRR only exploits one view of data, *i.e.* data manifold for data analysis. Contrasted with LLRR, Latent LRR (LatLRR) model adds another view, *i.e.* feature manifold to do image processing [22]. For solving these minimization problems of LRR, LLRR and LatLRR, the common point is to use the nuclear norm to approximate the rank operator. Given a data matrix $\mathbf{X}$, the nuclear norm means that the sum of all singular values belonging to $\mathbf{X}$. Since the nuclear norm minimizes the sum of all singular values for accomplishing the minimization problem, all non-zero singular values have different influences for the rank [23]. Thus, the nuclear norm maybe not the best way to approximate the rank of the matrix. To better approximate the rank and handle the non-convex optimizing problems, the truncated nuclear norm (TNN) was proposed in [24] and attracted much attention [13, 23, 25, 26]. The TNN that is the sum of few smallest singular values of a matrix

Table 1 Description about seven integrative gene expression datasets

Datasets	Genes	Samples	Samples classes
PAAD-COAD	20502	176-262	2
HNSC-ESCA	20502	398-183	2
CHOL-HNSC-ESCA	20502	36-398-183	3
COAD-PAAD-ESCA	20502	262-176-183	3
PAAD-ESCA-HNSC	20502	180-192-418	3
HNSC-PAAD-CHOL-ESCA	20502	398-176-36-183	4
ESCA-COAD-CHOL-PAAD	20502	183-262-36-176	4

can dispel the harm of shrinkage of the leading singular values, so it may be a more robust regularization to get the rank of a matrix than the nuclear norm.

To strengthen the efficiency and robustness of the model, in our paper, a novel LRR method is developed, named Truncated nuclear norm and graph-Laplacian regularized Low-Rank Representation model (TGLRR). In the objective function of TGLRR, the nuclear norm is replaced by the TNN for reaching the robust approximation of rank function, a graph-Laplacian regularization is imposed to find the local manifold structure, and the L_1-norm is used for realizing the sparse constraints of outliers. The main contributions of our paper are showed as follows. Firstly, compared with the popular LRR model regularized by the nuclear norm, our TGLRR method can obtain a better performance by the TNN, and solve the non-convex and discontinuous issues. Secondly, the TGLRR method can seize the valuable information lying in data manifold and feature manifold simultaneously. Finally, the TGLRR method can capture the internal similarity information and some underlying affinity among data points by incorporating a graph regularization term and utilizing a linear association of some bases to represent each data point.

The remainder of this article is organized as follows. In the Results section, TGLRR is exploited for clustering and feature selection on integrated gene expression data. In Conclusions section, conclusions and the future work are given. In Methods section, our TGLRR method is put forward and the optimization problem is resolved through an efficient framework based on LADMAP [27].

Results

Integrative gene expression datasets

To validate the performance of TGLRR model, six clustering experiments and a feature selection experiment are conducted. The experimental data are integrative cancer gene expression data instead of single cancer data for avoiding sample imbalance problem. Seven different datasets are produced via integrating five different single gene expression data downloaded from The Cancer Genome Atlas (TCGA, https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga). The pertinent information about the seven integrative datasets is listed in Table 1.

PAAD, ESCA, COAD, CHOL and HNSC are the abbreviations of Pancreatic Ductal Adenocarcinoma, Esophageal Carcinoma, Colorectal Adenocarcinoma, Cholangiocarcinoma, Head and Neck Squamous Cell Carcinoma, respectively. Taking PAAD-COAD

dataset for example, it is only composed of the tumor samples of PAAD and COAD data, in which PAAD data are made up of 176 tumor samples and 4 normal samples, and COAD data consist of 262 tumor samples and 19 normal samples. HNSC-ESCA, CHOL-HNSC-ESCA, COAD-PAAD-ESCA, HNSC-PAAD-CHOL-ESCA and ESCA-COAD-CHOL-PAAD datasets are also made in the production way of PAAD-COAD dataset. But PAAD-ESCA-HNSC dataset is composed of the whole samples of PAAD, ESCA and HNSC.

To eliminate redundant features and avoid over-fitting, the dimension of data matrix $\mathbf{X}$ is reduced before clustering and feature selection experiment, which can also greatly abate the computational cost. PCA is chosen for dimension reduction experiments in our paper. In addition, 2000-dimensional data $\mathbf{X}$ are obtained after dimension reduction.

Parameters selection

There are three important parameters, *i.e.* regularization terms λ, β and r of the TNN in our TGLRR model. The grid search is used to pick up the values of λ, β and r. Figure 1 shows that clustering results are varied with the parameters λ and β on three distinct integrative datasets of tumor gene expression.

X-axis represents the values range of λ, Y-axis represents the values range of β, and Z-axis represents the clustering accuracy in Fig. 1. It can be distinctly found that the effect of clustering accuracy stems from β that is greater than the effect from λ, especially in HNSC-PAAD-CHOL-ESCA dataset. Finally, it can be found that TGLRR performs well when $\lambda = 10$ and $\beta = 10^4$ on PAAD-COAD dataset, $\lambda = 10^{-2}$ and $\beta = 10$ on HNSC-ESCA dataset, $\lambda = 10^{-2}$ and $\beta = 10^2$ on CHOL-HNSC-ESCA dataset, $\lambda = 10^{-1}$ and $\beta = 10^4$ on COAD-PAAD-ESCA dataset, $\lambda = 10^{-2}$ and $\beta = 10^{11}$ on HNSC-PAAD-CHOL-ESCA dataset, and $\lambda = 10^2$ and $\beta = 10^{10}$ on ESCA-COAD-CHOL-PAAD dataset, respectively.

Different from the method in [25] that tries all the possible values to seek the optimal value of r, the method in [28] is used to choose the optimal value of r. A curve graph showing the singular values needs to be drawn. Figure 2 shows the summary curve graph on six datasets applied in clustering experiments.

X-axis indicates the number and Y-axis denotes the singular values in Fig. 2. The value of the first inflection point in each curve is chosen as the value of r corresponding to each dataset. The principle of selecting r is that the singular values before

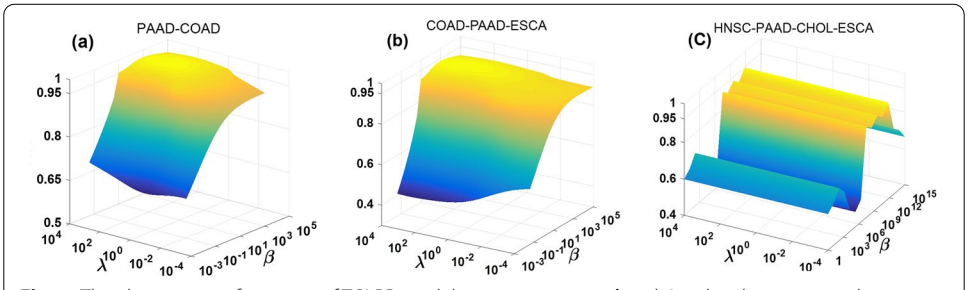

Fig. 1 The clustering performance of TGLRR model versus parameter λ and β. **a** the clustering results on PAAD-COAD dataset, **b** the results on COAD-PAAD-ESCA dataset, **c** the results on HNSC-PAAD-CHOL-ESCA dataset

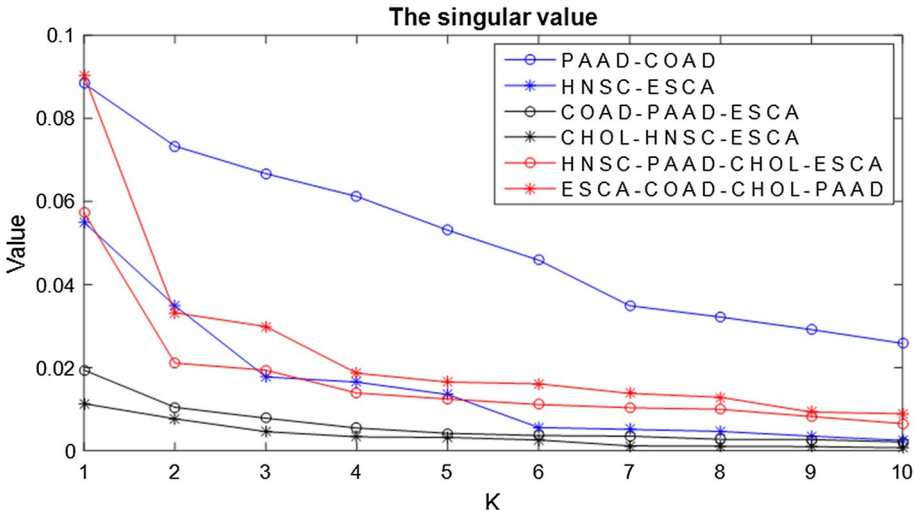

Fig. 2 The singular values on six distinct matrices

inflection point are bigger than the singular values after inflection point. Therefore, the values of r on PAAD-COAD, HNSC-ESCA, CHOL-HNSC-ESCA, COAD-PAAD-ESCA, HNSC-PAAD-CHOL-ESCA and ESCA-COAD-CHOL-PAAD datasets are set as 2, 3, 3, 2, 2 and 3, respectively.

Convergence analysis

Since Algorithm 1 is a practical application based on LADMAP framework whose convergence has been proved in [27], Algorithm 1 should also be convergent. Many approaches are able to demonstrate the convergence property of algorithms [23, 29]. In our paper, an efficient approach in [29] by means of auxiliary function is exploited to validate the convergence property of TGLRR method. The results are exhibited in Fig. 3.

The abscissa in Fig. 3 indicates the iteration number and the ordinate denotes the loss function value. As shown in Fig. 3, our model is convergent. The TGLRR method begins to converge after 30 iterations on two datasets, such as HNSC-ESCA and CHOL-HNSC-ESCA. On other four datasets, the TGLRR method converges in 40 iterations. Here, the HNSC-ESCA dataset may be easily addressed, so our method begins to converge on the two datasets after 30 iterations while it needs 40 iterations on the other four datasets.

Clustering results

In this subsection, the TGLRR is applied for clustering, and compared with K-means, LLRR [14], LRR [30], RPCA [5], DGLRR [31], and LatLRR [22].

In respect of the dictionary matrix $\mathbf{X}$, the optimal solution $\mathbf{Z}^*$ to TGLRR is able to symbolize "the minimum rank representation" of the data matrix $\mathbf{X}$. What's more, the i-th column about $\mathbf{Z}$ could be regarded as a "better" reflection of the i-th column about $\mathbf{X}$ so as to make the subspace structure more easily detectable [31]. Namely, the optimal

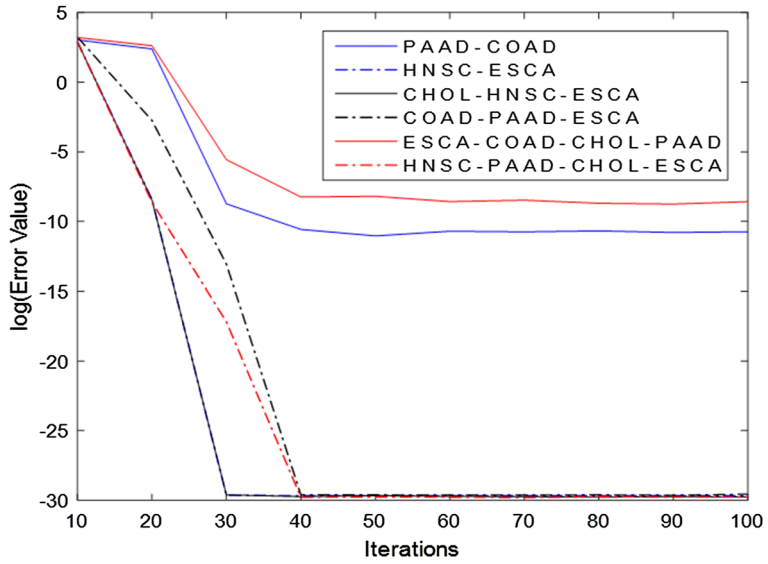

Fig. 3 Convergence curves of TGLRR on gene expression data

solution $\mathbf{Z}^*$ could include almost all the sample information about integrative gene expression data $\mathbf{X}$. Therefore, $\mathbf{Z}^*$ can be used for clustering experiments by K-means.

To measure the performance of our approach, three quantity metrics are adopted in this paper, *i.e.*, accuracy (ACC), normalized mutual information (NMI) and F-measure. As a widely used metric in machine learning field, ACC can be defined as follows:

$$\mathrm{ACC} = \frac{\sum_{i=1}^{n} \delta\left(\hat{\mathrm{L}}(i), \mathrm{Map}(l_i)\right)}{n}, \tag{1}$$

where n indicates the total number of tumor samples in an integrated data, $\delta(x, y)$ is a delta function set to 1 only when $x = y$ and 0 otherwise, $\hat{L}(i)$ denotes true class label of the i-th sample, and l_i represents the cluster label produced by the algorithms. $\mathrm{Map}(l_i)$ is a mapping function permuting every l_i to match real sample label.

The second index of NMI is defined by

$$\mathrm{NMI}\left(\mathrm{T}, \hat{\mathrm{T}}\right) = \mathrm{MI}\left(\mathrm{T}, \hat{\mathrm{T}}\right) \Big/ \max\left(\mathrm{H}(\mathrm{T}), \mathrm{H}\left(\hat{\mathrm{T}}\right)\right), \tag{2}$$

where T and $\hat{T}$ denote two different tumor index sets separately. H(T) and $\mathrm{H}\left(\hat{\mathrm{T}}\right)$ represent the entropy in T and $\hat{T}$, respectively. And $\mathrm{MI}\left(\mathrm{T}, \hat{T}\right) = \sum_{t \in \mathrm{T}} \sum_{\hat{t} \in \hat{T}} \mathrm{P}(t, \hat{t}) \log \mathrm{P}(t, \hat{t}) \big/ \left(\mathrm{P}(t)\mathrm{P}(\hat{t})\right)$ where $\mathrm{P}(t)$ is the marginal probability distribution function, namely, the probabilities that a tumor sample arbitrarily chosen from an integrated dataset belongs to cluster T. In addition, $\mathrm{MI}\left(\mathrm{T}, \hat{\mathrm{T}}\right)$ indicates the joint probabilities that a tumor sample belongs to the two clusters T and $\hat{\mathrm{T}}$ simultaneously.

F-measure is the comprehensive evaluation index considering both precision and recall, and written as:

Table 2 The clustering results on PAAD-COAD and HNSC-ESCA integrative data

	PAAD-COAD			HNSC-ESCA		
	ACC(%)	NMI(%)	F-measure(%)	ACC(%)	NMI(%)	F-measure(%)
K-means	91.57 ± 0.89	68.77 ± 4.24	91.62 ± 1.01	99.36 ± 0.05	98.00 ± 0.50	98.81 ± 0.18
LLRR	93.95 ± 0.29	71.59 ± 1.29	93.83 ± 0.26	99.83 ± 0.00	98.07 ± 0.00	99.80 ± 0.00
LRR	93.63 ± 0.57	70.70 ± 2.52	93.64 ± 0.51	99.83 ± 0.00	98.07 ± 0.00	99.80 ± 0.00
RPCA	93.81 ± 0.46	71.09 ± 2.27	93.81 ± 0.42	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00
DGLRR	94.14 ± 0.40	71.98 ± 1.80	94.13 ± 0.47	99.83 ± 0.00	98.07 ± 0.00	99.80 ± 0.00
LatLRR	93.76 ± 0.33	71.46 ± 1.50	93.77 ± 0.29	99.83 ± 0.00	98.07 ± 0.00	99.80 ± 0.00
TGLRR	95.15 ± 0.00	74.44 ± 0.00	95.10 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00

Table 3 The clustering results on CHOL-HNSC-ESCA and COAD-PAAD-ESCA data

	CHOL-HNSC-ESCA			COAD-PAAD-ESCA		
	ACC (%)	NMI (%)	F-measure (%)	ACC (%)	NMI (%)	F-measure (%)
K-means	83.49 ± 1.77	76.23 ± 3.47	77.05 ± 3.42	83.61 ± 3.25	76.11 ± 2.85	81.95 ± 4.28
LLRR	96.73 ± 0.80	94.80 ± 1.26	94.69 ± 2.09	87.07 ± 2.23	79.50 ± 1.46	86.19 ± 2.94
LRR	97.13 ± 0.42	95.32 ± 0.61	96.16 ± 0.82	88.16 ± 1.98	80.27 ± 1.48	87.47 ± 2.53
RPCA	85.40 ± 2.64	81.43 ± 4.16	81.26 ± 4.59	85.59 ± 2.72	78.85 ± 2.39	83.98 ± 3.49
DGLRR	94.70 ± 1.03	92.33 ± 1.78	91.63 ± 2.62	86.14 ± 2.38	78.67 ± 1.82	84.93 ± 3.11
LatLRR	93.94 ± 1.57	91.37 ± 2.46	91.57 ± 3.32	87.16 ± 2.52	79.33 ± 1.93	86.16 ± 3.30
TGLRR	98.37 ± 0.00	90.58 ± 0.03	96.09 ± 0.01	92.82 ± 0.77	79.51 ± 0.93	92.62 ± 0.91

$$\mathrm{F-measeure} = 2 \cdot (\mathrm{recall} \cdot \mathrm{precision}) \big/ (\mathrm{recall} + \mathrm{precision}), \tag{3}$$

where recall = $\mathrm{TP}/(\mathrm{TP} + \mathrm{FN})$ and precision = $\mathrm{TP}/(\mathrm{TP} + \mathrm{FP})$. TP, FP, TN and FN indicate the true-positive, false-positive, true-negative and false-negative, respectively.

To prove the effectiveness of TGLRR, the detailed clustering results of these methods on integrative tumor gene expression data are listed by three tables. In tables, the values about ACC, NMI and F-measure are the average of 100 clustering results of each approach, and the values on the right of ± are the variance of 100 results.

Table 2 reports the clustering results on PAAD-COAD and HNSC-ESCA datasets. Obviously, our TGLRR method exceeds other six comparison methods on PAAD-COAD dataset. TGLRR is more robust than other six methods on PAAD-COAD dataset from the point of the variance values. HNSC-ESCA data are extraordinary and may be easily addressed, in which the clustering results about all algorithms are good and particularly the evaluation indices of RPCA and TGLRR are 1.

The clustering results on two integrated datasets containing three types of tumors are exhibited in Table 3. It can be seen that the clustering performance of TGLRR model outperforms other models on COAD-PAAD-ESCA dataset. On CHOL-HNSC-ESCA dataset, TGLRR's ACC, NMI and F-measure values are higher than values obtained via other five models except for LRR. Consequently, it still can be said that the TGLRR method outstrips other methods on CHOL-HNSC-ESCA dataset.

From Table 4, our TGLRR method outmatches other six methods on HNSC-PAAD-CHOL-ESCA and ESCA-COAD-CHOL-PAAD datasets.

Table 4 The clustering results on HNSC-PAAD-CHOL-ESCA and ESCA-COAD-CHOL-PAAD data

	HNSC-PAAD-CHOL-ESCA			ESCA-COAD-CHOL-PAAD		
	ACC (%)	NMI (%)	F-measure (%)	ACC (%)	NMI (%)	F-measure (%)
K-means	78.42 ± 0.94	71.34 ± 1.03	72.19 ± 1.92	82.49 ± 2.15	77.01 ± 1.76	75.71 ± 3.30
LLRR	87.66 ± 0.94	75.56 ± 0.40	86.90 ± 2.04	84.41 ± 2.07	80.24 ± 1.22	82.60 ± 2.73
LRR	88.63 ± 0.39	75.89 ± 0.21	89.16 ± 0.81	87.40 ± 2.05	82.52 ± 1.20	87.62 ± 2.17
RPCA	84.85 ± 1.54	80.29 ± 1.50	81.72 ± 2.57	83.39 ± 1.73	79.28 ± 1.31	76.86 ± 3.10
DGLRR	86.68 ± 0.85	75.22 ± 0.41	84.99 ± 1.94	85.99 ± 2.26	81.52 ± 1.39	84.01 ± 3.13
LatLRR	85.14 ± 1.02	73.96 ± 0.43	84.26 ± 2.21	86.04 ± 1.85	81.49 ± 1.14	82.37 ± 1.94
TGLRR	93.46 ± 0.93	82.83 ± 0.75	90.90 ± 1.20	90.62 ± 1.53	79.87 ± 1.49	90.34 ± 1.64

Feature selection

Cancers are commonly relevant to gene mutation or abnormal expression of genes. Thus, in this subsection, the TGLRR method is used to identify co-feature genes of PAAD, ESCA and HNSC from PAAD-ESCA-HNSC dataset.

From the formula (14), a minimum solution $\mathbf{G}^*$ can be got from an integrative gene expression data $\mathbf{X}$ via TGLRR scheme. $\mathbf{G}^*$ can obtain the feature manifold structure lying in data. As a result, it can be applied in feature gene extraction. From the view of cancer, its pathogenesis may be related to gene mutation [32]. It is extremely meaningful to find out the feature genes inducing cancers from gene expression data.

Similar to the subsection of Parameters Selection, 10^{-2}, 10^3 and 4 are assigned to λ, β and r.

Table 5 exhibits the top 10 co-feature genes with the mean of highest relevance score distinguished by the TGLRR method from PAAD-ESCA-HNSC dataset. The related diseases, related pathways and coded proteins about these genes are gotten from GeneCards (https://www.genecards.org/). These genes are most likely to lead to PAAD, ESCA and HNSC simultaneously.

From Table 5, clearly, CDH1 gene with the highest relevance score can result in a host of cancers, which indicates that CDH1 may be a dangerous co-feature gene. What's more, PAAD, ESCA and HNSC are all correlative with CDH1 and RHOA, which can be affirmed from [33–38]. It is a verifiable fact that TGFB1 and RELA all serve as a predictor for PAAD and ESCA via consulting some literatures. Some data show that PTPN11 may induce HNSC and PAAD. From [39, 40], it can be seen that ESCA is relevant to IGF2R and RUNX1. In addition, the related pathways of RUNX1 and EWSR1 include transcriptional misregulation in cancer. So, RUNX1 and EWSR1 may be co-characteristic genes of PAAD, ESCA and HNSC.

All in all, the TGLRR method is successful in identifying co-characteristic genes on the integrative gene expression datasets.

Discussions

The TGLRR method is applied to the tumor clustering and gene selection, and superior to the other methods. Based on above results, it can be affirmed that the TNN could capture more valuable information existed in data than the nuclear norm from data. By comparing the results of DGLRR, a conclusion can be drawn that the graph Laplacian regularization imposed on feature manifold may cause adverse effects for clustering

Table 5 The top 10 genes selected via TGLRR on PAAD-ESCA-HNSC

Gene ED	Relevance score	Related diseases	Coded proteins
CDH1	101.03, 96.95, 124.3, 107.43	Gastric, breast, colorectal, thyroid and ovarian cancer	Cadherin superfamily
TGFB1	73.21, 44.14, 76.66, 64.67	Camurati-Engelmann disease, Encephalopathy, Inflammatory Bowel Disease and Immunodeficiency	Transforming Growth Factor-Beta Superfamily of Proteins
RELA	27.63, 11.33, 41.36, 26.77	Mucocutaneous Ulceration, Chronic and Ependymoma	Transcription Factor
ANXA5	26.80, 10.31, 42.30, 26.47	Pregnancy Loss, Recurrent 3 and Antiphospholipid Syndrome	Calcium-Dependent Phospholipid Binding Proteins
RHOA	27.48, 11.81, 31.46, 23.58	Adenocarcinoma and Peripheral T-Cell Lymphoma	Rho Family of Small GTPases
PTPN11	13.04, 13.56, 43.23, 23.28	Noonan Syndrome 1 and Juvenile Myelomonocytic Leukemia	Protein Tyrosine Phosphatase
CTNNA1	20.94, 19.40, 24.80, 21.71	Macular Dystrophy, Patterned, 2 and Butterfly-Shaped Pigment Dystrophy	Cell Adhesion Process Protein
IGF2R	13.40, 19.07, 25.26, 19.24	Hepatocellular Carcinoma and Inclusion-Cell Disease	Receptor for Both Insulin-Like Growth Factor 2 and Mannose 6-Phosphate
RUNX1	10.85, 12.97, 25.61, 16.48	Platelet Disorder, Familial, with Associated Myeloid Malignancy, leukemia and Isolated Delta-Storage Pool Disease	Transcription Factor
EWSR1	12.55, 9.19, 27.33, 16.36	Ewing Sarcoma and Desmoplastic Small Round Cell Tumor	Multifunctional Protein

Take the contents in the second column of the second row as an example, the first, second and third numeral are the relevance score of CDH1 gene to PAAD, ESCA and HNSC, respectively, and the fourth is the mean

on our integrative datasets. The TGLRR method has some limitations. For example, on HNSC-PAAD-CHOL-ESCA dataset, the variance values of TGLRR is larger than LRR. It may be caused by the integrated datasets and its stability needs to be improved in future.

In a word, these improvements to the prevenient LRR model can help TGLRR catch more useful information concealed in the low-dimensional manifold structure.

Conclusions

The paper proposes a Low-Rank Representation approach called TGLRR. It can capture the global and local geometric structures in data manifold via using the raw data matrix as the dictionary matrix and introducing the graph-Laplacian regularization term. Furthermore, TGLRR can gain a better approximation to the rank operator than the approaches regularized by the nuclear norm. The objective function of TGLRR is perfectly resolved through an iterative algorithm based on LADMAP framework. The efficiency and robustness of our TGLRR method are testified through the encouraging experimental results.

Methods

Related LRR methods

Based on the assumption that the observation data $\mathbf{X}$ are sampled from a union of several low-dimensional subspaces $S = \sum S_1, S_2, \ldots, S_k$ located in a high-dimensional spaces, LRR was raised in [16]. If data are noiseless, the rank minimization problem of LRR is written into

$$\min_{\mathbf{Z}} \operatorname{rank}(\mathbf{Z}),\ s.t.\ \mathbf{X} = \mathbf{DZ}, \tag{4}$$

where $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n] \in \mathbf{R}^{m\times n}$ is the original data matrix and $\mathbf{Z} \in \mathbf{R}^{n\times n}$ is a low-rank matrix recovered from $\mathbf{X}$ via LRR. $\mathbf{D} \in \mathbf{R}^{m\times n}$ is a basis matrix (or named dictionary matrix), which spans the whole data space linearly. The observation data generally exist more or less noise in real life, so the optimization problem (4) may be impracticable. The LRR model with noise is

$$\min_{\mathbf{Z},\mathbf{P}} \operatorname{rank}(\mathbf{Z}) + \lambda\|\mathbf{P}\|_0,\ s.t.\ \mathbf{X} = \mathbf{DZ} + \mathbf{P}, \tag{5}$$

where $\mathbf{P} \in \mathbf{R}^{m\times n}$ is the reconstruction errors matrix (or called noise matrix). λ is a penalty parameter aiming to adjust the sparsity of matrix $\mathbf{P}$ and the reconstruction fidelity of data matrix $\mathbf{X}$ damaged by errors matrix $\mathbf{P}$. $\|\mathbf{P}\|_0$ is the L_0-norm of matrix $\mathbf{P}$, which indicates the number of non-zero elements in matrix $\mathbf{P}$.

Since the rank function is discrete, the problem (5) may have multiple solutions and the L_0-minimization is non-convexity and intractable. Usually, solving the problem (5) is NP-hard [41]. To better solve the above rank minimization problem, the nuclear norm is imposed on the low-rank matrix, and the L_0-norm is replaced with the $L_{2,1}$-norm [17]. The convex optimization problem about LRR model is written as follows:

$$\min_{\mathbf{Z},\mathbf{P}} \|\mathbf{Z}\|_* + \lambda\|\mathbf{P}\|_{2,1},\ s.t.\ \mathbf{X} = \mathbf{DZ} + \mathbf{P}, \tag{6}$$

where $\|\mathbf{Z}\|_* = \sum_i^{\min(m,n)} \sigma_i(\mathbf{Z})$($\delta_i(\mathbf{Z})$ is the i-th largest singular value of $\mathbf{Z}$) denotes the nuclear norm of matrix $\mathbf{Z}$, and $\|\mathbf{P}\|_{2,1} = \sum_{i=1}^{m}\left(\sum_{j=1}^{n} m_{ij}^2\right)^{1/2}$ denotes the $L_{2,1}$-norm of matrix $\mathbf{P}$. To get a self-expression model, the observation data $\mathbf{X}$ are generally installed as the dictionary matrix [13, 14, 22]. The final LRR model becomes

$$\min_{\mathbf{Z},\mathbf{P}} \|\mathbf{Z}\|_* + \lambda\|\mathbf{P}\|_{2,1},\ s.t.\ \mathbf{X} = \mathbf{XZ} + \mathbf{P}. \tag{7}$$

For low-rank matrix $\mathbf{Z} = [\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_n]$, its each element $\mathbf{z}_{ij}$ can reflect the manifold information, *i.e.* the similarity between the data point $\mathbf{x}_i$ and the data point $\mathbf{x}_j$. Therefore, matrix $\mathbf{Z}$ can be seen as an affinity matrix [14]. LRR is devoted to seek the lowest rank representation of the observation data. With the help of an appropriate dictionary matrix, the underlying row space can be recovered via the lowest rank representation such that the true segmentation of data can be correctly revealed. Thus, LRR method can manage the data extracted from a union of multiple subspaces well [17].

Nevertheless, LRR method has to face two issues owing to the raw data $\mathbf{X}$ that are used as the basis. First, LRR method requires that the basis contains adequate data samples from the subspaces so as to possess the capacity of representing the underlying subspaces. Second, LRR method demands that noise of data $\mathbf{X}$ is little, *i.e.* only a part of $\mathbf{X}$ is corrupted. To remedy these two shortcomings of LRR, Liu et al. proposed the following convex optimization LRR problem [22]:

$$\min_{\mathbf{Z},\mathbf{G},\mathbf{P}} \|\mathbf{Z}\|_* + \|\mathbf{G}\|_* + \lambda\|\mathbf{P}\|_1,\ s.t.\ \mathbf{X} = \mathbf{XZ} + \mathbf{GX} + \mathbf{P}, \tag{8}$$

where $\|\mathbf{P}\|_1 = \sum_{i=1}^{n}\sum_{j=1}^{m}\left|\mathrm{p}_{ij}\right|$ is the L_1-norm of matrix $\mathbf{P}$ and $\mathbf{G}$ is the feature matrix separated from the original $\mathbf{X}$. Equation (8) is a state-of-the-art LRR-based subspace learning model, named LatLRR. By means of LatLRR model, the observed sampling can be expressed via many unobserved sampling effectively [42]. In practical application, $\mathbf{Z}$ and $\mathbf{G}$ are applied in cluster analysis and feature selection, respectively.

Truncated nuclear norm (TNN)

The TNN is the summation of a few smaller singular values, *i.e.* the sum of some largest singular values is subtracted from the nuclear norm [24]. As an approximation of a rank operator, the largest r-th singular values could produce minor amount of information, meanwhile, the minimal $(\min(m,n)-r)$-th singular values act a crucial role [23]. Compared to the nuclear norm, the TNN may be a better approximation to the rank operator. Its mathematical formula is

$$\|\mathbf{Z}\|_r = \sum_{i=r+1}^{\min(m,n)} \delta_i(\mathbf{Z}), \tag{9}$$

where $\delta_i(\mathbf{Z})$ denotes the i-th largest singular value belongs to $\mathbf{Z}$ and r is a nonnegative integer and $\mathrm{r} \leq min(m, n)$.

Since the minimization of Eq. (9) is not convex, it cannot be directly resolved through the approaches. For overcoming this issue, Hu et al. come up with a theorem [25]. According to the Theorem, the equivalent transformation of Eq. (9) is achieved.

$$\|\mathbf{Z}\|_r = \sum_{i=r+1}^{\min(m,n)} \delta_i(\mathbf{Z}) = \|\mathbf{Z}\|_* - \max_{\mathbf{A}\mathbf{A}^{\mathrm{T}}=\mathbf{I},\,\mathbf{B}\mathbf{B}^{T}=\mathbf{I}} \mathrm{Tr}\left(\mathbf{A}\mathbf{Z}\mathbf{B}^{T}\right). \tag{10}$$

Graph-Laplacian regularization

Graph-Laplacian regularization is an outstanding manifold learning method, which can uncover the internal geometrical structures among the data points. As a result, naturally, appears a number of LRR models regularized by graph embedding manifold regularization [13, 43].

Given a k-nearest-neighbor graph G, suppose it has n vertices, and each vertex denotes a data point hidden in an underlying sub-manifold M [11]. Then, a symmetric weight matrix $\mathbf{W} \in \mathbf{R}^{n\times n}$ is constructed, where $\mathbf{w}_{ij}$ expresses the i weight of the edge linking vertices i and j. The value of every $\mathbf{w}_{ij}$ can be calculated via

$$\mathbf{w}_{ij}=\begin{cases} 1, & \text{if } \mathbf{y}_i \in \mathrm{N}_k\left(\mathbf{d}_j\right) \text{ or } \mathbf{y}_j \in \mathrm{N}_k(\mathbf{d}_i), \\ 0, & \text{otherwise,} \end{cases} \tag{11}$$

where $\mathrm{N}_k\left(\mathbf{d}_j\right)$ indicates the k-nearest-neighbors of data point $\mathbf{d}_j$.

Next, a diagonal matrix $\mathbf{O}$, termed a degree matrix, need to be established. The value of the i-th member of $\mathbf{O}$ can be calculated by the sum of all the similarities associated with vertex $\mathbf{d}_j$, *i.e.* $\mathbf{o}_{ii} = \sum_j \mathbf{w}_{ij}$. The graph-Laplacian matrix $\mathbf{L}$ can be obtained by

$$\mathbf{L} = \mathbf{O} - \mathbf{W}. \tag{12}$$

Finally, the graph embedding regularization term can be formulated by

$$\mathrm{Tr}\left(\mathbf{ZLZ}^T\right). \tag{13}$$

Truncated nuclear norm and graph-Laplacian regularized low-rank representation method

Motivated by strengthening the robustness of LRR, our method (TGLRR) is put forward. Considering that some data may exist nonlinear geometric structure [14] and the disadvantages about the nuclear norm, the TNN and graph embedding manifold learning are introduced into our rank minimization problem to extract more essential information hidden in data. The objective function of TGLRR is formulated as follows:

$$\begin{aligned} \min \|\mathbf{Z}\|_* - \max_{\mathbf{AA}^{\mathrm{T}}=\mathbf{I},\, \mathbf{BB}^T=\mathbf{I}} \mathrm{Tr}\left(\mathbf{AZB}^T\right) + \|\mathbf{G}\|_* + \frac{\beta}{2}\mathrm{Tr}\left(\mathbf{ZLZ}^T\right) + \lambda\|\mathbf{P}\|_1, \\ s.t.\, \mathbf{X} = \mathbf{XZ} + \mathbf{GX} + \mathbf{P}, \end{aligned} \tag{14}$$

where $\beta \geq 0$ and $\lambda \geq 0$ are the regularization parameters for balancing the contribution of each term.

Essentially, TGLRR can get a more precise approximation to the rank function with the help of the TNN than the nuclear norm. And the underlying low-dimensional structures of data could be captured by the aid of the graph-Laplacian regularization and the basis matrix $\mathbf{X}$.

Optimization solution

To correctly solve the optimization problem of (14), an efficient iterative algorithm based on LADMAP framework is designed. The algorithm (Algorithm 1) is implemented via alternating two iterative procedures till Eq. (14) converges to the minimum.

The first step is to determine matrix $\mathbf{A}$ and $\mathbf{B}$.

Step 1: Given $\mathbf{Z}_k$ (k indicates the k-th updating), the SVD (Singular Value Decomposition) of $\mathbf{Z}_k$ need to be conducted. $[\mathbf{U}_k, \mathbf{\Sigma}_k, \mathbf{V}_k] = \mathrm{SVD}(\mathbf{Z}_k)$, where $\mathbf{U}_k = (\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_m) \in \mathbf{R}^{m\times m}$, $\mathbf{\Sigma}_k \in \mathbf{R}^{m\times n}$ and $\mathbf{V}_k = (\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_m) \in \mathbf{R}^{n\times n}$. $\mathbf{A}_k$ and $\mathbf{B}_k$ are calculated via $\mathbf{A}_k = (\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_r)^T$ and $\mathbf{B}_k = (\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_r)^T$.

The second step is to resolve the following convex optimization problem:

$$\begin{aligned} [\mathbf{Z}, \mathbf{G}, \mathbf{P}] = \arg\min_{\mathbf{Z},\mathbf{G},\mathbf{P}} \|\mathbf{Z}\|_* - \mathrm{Tr}\left(\mathbf{AZB}^T\right) + \|\mathbf{G}\|_* + \frac{\beta}{2}\mathrm{Tr}\left(\mathbf{ZLZ}^T\right) + \lambda\|\mathbf{P}\|_1, \\ s.t.\, \mathbf{X} = \mathbf{XZ} + \mathbf{GX} + \mathbf{P}. \end{aligned} \tag{15}$$

Step 2: To achieve the separation of objective function (15), an auxiliary variable $\mathbf{F}$ is introduced. Equation (15) is rewritten as follows:

$$\begin{aligned} \min \|\mathbf{Z}\|_* - \max_{\mathbf{AA}^{\mathrm{T}}=\mathbf{I},\, \mathbf{BB}^T=\mathbf{I}} \mathrm{Tr}\left(\mathbf{AFB}^T\right) + \|\mathbf{G}\|_* + \frac{\beta}{2}\mathrm{Tr}\left(\mathbf{ZLZ}^T\right) + \lambda\|\mathbf{P}\|_1, \\ s.t.\, \mathbf{X} = \mathbf{XZ} + \mathbf{GX} + \mathbf{P}, \mathbf{Z} = \mathbf{F}. \end{aligned} \tag{16}$$

Equation (16) can be solved through LADMAP method, which introduces two Lagrangian multipliers $\mathbf{Y}^1$ and $\mathbf{Y}^2$. Thus, the augmented Lagrangian function can be defined as

$$
\begin{aligned}
L\left(\mathbf{Z},\,\mathbf{G},\,\mathbf{F},\,\mathbf{P},\,\mu,\,\mathbf{Y}^1,\,\mathbf{Y}^2\right) &= \|\mathbf{Z}\|_* - \mathrm{Tr}\left(\mathbf{AFB}^T\right) + \|\mathbf{G}\|_* \\
&+ \frac{\beta}{2}\mathrm{Tr}\left(\mathbf{ZLZ}^T\right) + \lambda\|\mathbf{P}\|_1 + \left\langle \mathbf{Y}^1,\,\mathbf{X}-\mathbf{XZ}-\mathbf{GX}-\mathbf{P}\right\rangle \\
&+ \left\langle \mathbf{Y}^2,\,\mathbf{Z}-\mathbf{F}\right\rangle + \mu/2\|\mathbf{X}-\mathbf{XZ}-\mathbf{GX}-\mathbf{P}\|_F^2 + \mu/2\|\mathbf{Z}-\mathbf{F}\|_F^2,
\end{aligned} \tag{17}
$$

where μ is the penalty parameter and $\|\cdot\|_F^2$ denotes the Frobenius norm of a matrix that is $\|\mathbf{X}\|_F^2 = \sum_{i=1}^m \sum_{j=1}^n \left|\mathbf{x}_{ij}\right|^2$.

Next, the alternating minimization strategy is adopted to compute **Z**, **F**, **G** and **P**. In the iterative procedure, **Z**, **F**, **G** or **P** is updated when the other three variables are fixed, respectively.

Updating Z

To get the solution of **Z**, the below minimization objective *w.r.t.* **Z** needs to be solved.

$$
\begin{aligned}
\mathbf{Z}_{k+1} &= \arg\min_{\mathbf{Z}} L\left(\mathbf{Z},\,\mathbf{G}_k,\,\mathbf{F}_k,\,\mathbf{P}_k,\,\mu_k,\,\mathbf{Y}_k^1,\,\mathbf{Y}_k^2\right) \\
&= \arg\min_{\mathbf{Z}} \|\mathbf{Z}\|_* + \frac{\beta}{2}\mathrm{Tr}\left(\mathbf{ZLZ}^T\right) + \left\langle \mathbf{Y}_k^1,\,\mathbf{X}-\mathbf{XZ}-\mathbf{G}_k\mathbf{X}-\mathbf{P}_k\right\rangle \\
&\quad + \left\langle \mathbf{Y}_k^2,\,\mathbf{Z}-\mathbf{F}_k\right\rangle + \mu_k/2\|\mathbf{X}-\mathbf{XZ}-\mathbf{G}_k\mathbf{X}-\mathbf{P}_k\|_F^2 + \mu_k/2\|\mathbf{Z}-\mathbf{F}_k\|_F^2.
\end{aligned} \tag{18}
$$

Equation (18) has a closed-form solution:

$$
\mathbf{Z}_{k+1}^* = \Theta_{1/\eta_1\mu_k}\left(\mathbf{Z}_k - \nabla_{zq}(\mathbf{Z}_k)/\eta_1\right), \tag{19}
$$

where $\Theta(\cdot)$ indicates the Singular Value Thresholding operator (SVT), $\nabla_{zq}(\mathbf{Z}_k) = \frac{\beta}{2}\left(\mathbf{ZL}^T + \mathbf{ZL}\right) + \mu_k\left(\mathbf{Z} - \mathbf{F}_k + \mathbf{Y}_k^2/\mu_k\right) + \mu_k\mathbf{X}^T\left(\mathbf{XZ} - \mathbf{X} + \mathbf{G}_k\mathbf{X} + \mathbf{P}_k - \mathbf{Y}_k^1/\mu_k\right)$ and $\eta_1 = \beta\|\mathbf{L}\|_2 + \mu_k\left(1 + \|\mathbf{X}\|_2^2\right)$.

Updating G

Similar to the solution of **Z**, the SVT operator is employed in computing **G**. The optimal solution $\mathbf{G}_{k+1}^*$ is

$$
\mathbf{G}_{k+1}^* = \Theta_{1/\eta_2\mu_k}\left(\mathbf{G}_k - \nabla_{zq}(\mathbf{G}_k)/\eta_2\right), \tag{20}
$$

where $\nabla_{gq}(\mathbf{G}_k) = \mu_k\left(\mathbf{XZ}_{k+1} - \mathbf{X} + \mathbf{G}_k\mathbf{X} + \mathbf{P}_k - \mathbf{Y}_k^2/\mu_k\right)\mathbf{X}^T$ and $\eta_2 = \mu_k\|\mathbf{X}\|_2^2$.

Updating F

The below sub-problem *w.r.t.* **F** is

$$
\begin{aligned}
\mathbf{F}_{k+1} &= \arg\min_{\mathbf{F}} L\left(\mathbf{Z}_{k+1},\,\mathbf{G}_{k+1},\,\mathbf{F},\,\mathbf{P}_k,\,\mu_k,\,\mathbf{Y}_k^1,\,\mathbf{Y}_k^2\right) \\
&= \arg\min_{\mathbf{F}} -\mathrm{Tr}\left(\mathbf{A}_{k+1}\mathbf{FB}_{k+1}\right) + \left\langle \mathbf{Y}_k^2,\,\mathbf{Z}_{k+1}-\mathbf{F}_k\right\rangle + \mu_k/2\left\|\mathbf{Z}_{k+1}-\mathbf{F}_k\right\|_F^2.
\end{aligned} \tag{21}
$$

Equation (21) is the smooth convex planning problem. Different to the solving rules of **Z** and **G**, we can differentiate Eq. (21) and set it to zero to gain the answer of **F**. Its optimal solution is

$$
\mathbf{F}_{k+1} = \mathbf{A}_{k+1}\mathbf{B}_{k+1}^T\Big/\mu_k + \mathbf{Z}_{k+1} + \mathbf{Y}_k^2. \tag{22}
$$

Updating P

Calculating $\mathbf{P}$ has to optimize the following objective:

$$\begin{aligned}\mathbf{P}_{k+1} &= \arg\min_{\mathbf{P}} L\left(\mathbf{Z}_{k+1}, \mathbf{G}_{k+1}, \mathbf{F}_{k+1}, \mathbf{P}, \mu_k, \mathbf{Y}_k^1, \mathbf{Y}_k^2\right) \\ &= \arg\min_{\mathbf{P}} \lambda\|\mathbf{P}\|_1 + \mu_k/2\left\|\mathbf{P} - \left(\mathbf{X} - \mathbf{X}\mathbf{Z}_{k+1} - \mathbf{G}_{k+1}\mathbf{X} + \frac{\mathbf{Y}_k^1}{\mu_k}\right)\right\|_F^2.\end{aligned} \tag{23}$$

The optimal solution to the above sub-problem *w.r.t.* $\mathbf{P}$ can be formulated by

$$\mathbf{P}_{k+1} = S_{\lambda/\mu_k}\left(\mathbf{X} - \mathbf{X}\mathbf{Z}_{k+1} - \mathbf{G}_{k+1}\mathbf{X} + \mathbf{Y}_k^1/\mu_k\right). \tag{24}$$

$S_{\lambda/\mu_k}(\cdot)$ is the shrinkage operator defined as $S_{\lambda/\mu_k}(\cdot) = \mathbf{U}\Sigma_{\lambda/\mu_k}\mathbf{V}^T$, $\mathbf{\Sigma}_{\lambda/\mu_k} = \mathrm{diag}(\max\{\sigma_i - \lambda/\mu_k\}, 0)$.

Updating μ_k, $\mathbf{Y}_k^1$ and $\mathbf{Y}_k^2$

After computing the above variables, two Lagrange multipliers $\mathbf{Y}_k^1$ and $\mathbf{Y}_k^2$ are given by

$$\begin{cases}\mathbf{Y}_{k+1}^1 = \mathbf{Y}_k^1 + \mu_{k+1}(\mathbf{X} - \mathbf{X}\mathbf{Z}_{k+1} - \mathbf{G}_{k+1}\mathbf{X} - \mathbf{P}_{k+1}), \\ \mathbf{Y}_{k+1}^2 = \mathbf{Y}_k^2 + \mu_{k+1}(\mathbf{Z}_{k+1} - \mathbf{F}_{k+1}).\end{cases} \tag{25}$$

The iteration rule about μ_k is

$$\begin{aligned}\mu_{k+1} &= \min(\mu_{\max}, \rho_k\mu_k). \\ \rho_k &= \begin{cases}\rho_0, \text{ if } \mu_k \cdot \max\left\{\begin{array}{c}\eta_1\|\mathbf{Z}_k - \mathbf{Z}_{k+1}\|, \eta_2\|\mathbf{G}_k - \mathbf{G}_{k+1}\| \\ \|\mathbf{F}_k - \mathbf{F}_{k+1}\|, \|\mathbf{P}_k - \mathbf{P}_{k+1}\|\end{array}\right\} \le \varepsilon_2, \\ 1, \qquad \text{otherwise.}\end{cases}\end{aligned} \tag{26}$$

Input: Data matrix $\mathbf{X}$ and parameters $\lambda, \beta, \mu_0, \mu_{max}, \varepsilon_1, \varepsilon_2$.

Output: $\mathbf{Z}, \mathbf{F}, \mathbf{G}$ and $\mathbf{P}$.

While not converged do

Step 1:

Compute $[\mathbf{U}_{k+1}, \mathbf{\Sigma}_{k+1}, \mathbf{V}_{k+1}] = \mathrm{SVD}(\mathbf{Z}_k)$, $\mathbf{A}_{k+1} = (\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_r)^T$ and $\mathbf{B}_{k+1} = (\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_r)^T$.

Step 2:

1) Update $\mathbf{Z}_{k+1}$ by Eq.(19)
2) Update $\mathbf{G}_{k+1}$ by Eq.(20)
3) Update $\mathbf{F}_{k+1}$ by Eq.(21)
4) Update $\mathbf{P}_{k+1}$ by Eq.(24)
5) Update $\mathbf{Y}_{k+1}^1$ and $\mathbf{Y}_{k+1}^2$ by Eq.(25)
6) Update μ_{k+1} by Eq.(26)
7) Verifying convergence, if $\|\mathbf{X} - \mathbf{X}\mathbf{Z}_{k+1} - \mathbf{G}_{k+1}\mathbf{X}\|/\|\mathbf{X}\| < \varepsilon_1$ and $\mu_{k+1} \cdot \max\{\eta_1\|\mathbf{Z}_{k+1} - \mathbf{Z}_k\|, \eta_2\|\mathbf{G}_{k+1} - \mathbf{G}_k\|, \|\mathbf{F}_{k+1} - \mathbf{F}_k\|, \|\mathbf{P}_{k+1} - \mathbf{P}_k\|\} \le \varepsilon_2$

End While

The detailed algorithm about TGLRR model is showed in Algorithm 1.

Time complexity

In this subsection, the time complexity about TGLRR is discussed. Clearly, the main running time of TGLRR is expended on calculating the matrices $\mathbf{Z}$, $\mathbf{F}$, $\mathbf{G}$ and $\mathbf{P}$. For the $m \times n$ input data matrix $\mathbf{X}$, it has m genes and n samples. The time complexity of SVD method with respect to $\mathbf{Z}$ is $\mathrm{O}\left(r_{\mathbf{Z}} n^2\right)$ ($r_{\mathbf{Z}}$ is the lowest rank of $\mathbf{Z}$ decided by algorithm 1). For the same activity, the time complexity of SVD decomposition of $\mathbf{G}$ is $\mathrm{O}\left(r_{\mathbf{G}} m^2\right)$. The optimal solution of $\mathbf{F}$ can be obtained in $\mathrm{O}(rmn)$. In the resolving procedure of $\mathbf{P}$, $\mathbf{Y}^1$ also needs to be updated. The computational cost of $\mathbf{P}$ and $\mathbf{Y}^1$ needs $\mathrm{O}\left(mn^2 + mr_{\mathbf{P}} n\right)$ and $\mathrm{O}\left(nm^2\right)$, respectively. Since, $m >> n$ in our dataset, the total time cost of algorithm 1 is $\mathrm{O}\left(nm^2\right)$.

Abbreviations

LRR: Low-Rank Representation; LADMAP: Linearized Alternating Direction with Adaptive Penalty; PCA: Principal Component Analysis; RPCA: Robust PCA; LE: Laplacian Eigenmap; LLE: Locally Linear Embedding; LLRR: Graph-Laplacian regularized LRR; LatLRR: Latent LRR; TNN: Truncated Nuclear Norm; TGLRR: Truncated nuclear norm and Graph-Laplacian regularized Low-Rank Representation; TCGA: The Cancer Genome Atlas; PAAD: Pancreatic Ductal Adenocarcinoma; ESCA: Esophageal Carcinoma; COAD: Colorectal Adenocarcinoma; CHOL: Cholangiocarcinoma; HNSC: Head and Neck Squamous Cell Carcinoma; ACC: Accuracy; NMI: Normalized Mutual Information; SVT: Singular Value Thresholding; SVD: Singular Value Decomposition.

Acknowledgements

I am grateful to the anonymous reviewers whose suggestions and comments contributed to the significant improvement of this paper. I thank Shandong University of Science and Technology for providing the experimental site.

About this supplement

This article has been published as part of BMC Bioinformatics Volume 22 Supplement 12 2021: Explainable AI methods in biomedical data science. The full contents of the supplement are available at https://bmcbioinformatics.biomedcentral.com/articles/supplements/volume-22-supplement-12.

Authors' contributions

QL contributed to the design of the study, implemented TGLRR, performed the experiments, drafted the manuscript, and approved the final manuscript. All authors read and approved the final manuscript.

Funding

Publication costs are funded by the National Statistical Science Research Project under Grant No. 2019LY49. The funders did not play any role in the design of the study, the collection, analysis, and interpretation of data, or in writing of the manuscript.

Availability of data and materials

The TCGA datasets that support the findings of this study are available in https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga. The results from the free trial version of the GeneCards (https://www.genecards.org/) did only be used for scientific research, but not be used for commercial purposes.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The author declares that I have no competing interests.

Received: 10 August 2021 Accepted: 23 August 2021
Published online: 20 January 2022

References

1. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A: Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians* 2018, 68(6):394–424.
2. Lokody I. Cancer genomics: signature analysis suggests cancer origins. Nat Rev Genet. 2013;14(10):677–677.

3. Liu J-X, Gao Y-L, Zheng C-H, Xu Y, Yu J. Block-constraint robust principal component analysis and its application to integrated analysis of TCGA Data. IEEE Trans Nanobiosci. 2016;15(6):510–6.
4. Yu N, Wu M-J, Liu J-X, Zheng C-H, Xu Y: Correntropy-based hypergraph regularized NMF for clustering and feature selection on multi-cancer integrated data. *IEEE Trans Cybern* 2020.
5. Liu J-X, Wang Y-T, Zheng C-H, Sha W, Mi J-X, Xu Y: Robust PCA based method for discovering differentially expressed genes. *BMC Bioinform.* 2013, 14(S8).
6. Feng C-M, Xu Y, Liu J-X, Gao Y-L, Zheng C-H. Supervised discriminative sparse PCA for com-characteristic gene selection and tumor classification on multiview biological data. IEEE Trans Neural Netw Learn Syst. 2019;30(10):2926–37.
7. Liu W, Yang X, Tao D, Cheng J, Tang Y. Multiview dimension reduction via Hessian multiset canonical correlations. Information Fusion. 2018;41:119–28.
8. Abualigah LM, Khader AT, Al-Betar MA, Alomari OA. Text feature selection with a robust weight scheme and dynamic dimension reduction to text document clustering. Expert Syst Appl. 2017;84:24–36.
9. Abdi H, Williams LJ. Principal component analysis. Wiley Interdisciplinary Rev: Comput Stat. 2010;2(4):433–59.
10. Candes EJ, Li X, Ma Y, Wright J. Robust principal component analysis? J ACM. 2011;58(3):1–37.
11. Feng C-M, Gao Y-L, Liu J-X, Zheng C-H, Yu J. PCA based on graph laplacian regularization and P-norm for gene selection and clustering. IEEE Trans Nanobiosci. 2017;16(7):257–65.
12. Wright J, Ganesh A, Rao S, Ma Y: Robust principal component analysis: Exact recovery of corrupted low-rank matrices via convex optimization. arXiv 2009 Arxiv:0905.0233v1:1-44.
13. Xu X-X, Gao Y-L, Liu J-X, Wang Y-X, Dai L-Y, Kong X-Z, Yuan S-S. A novel low-rank representation method for identifying differentially expressed genes. Int J Data Min Bioinform. 2018;19(3):185–201.
14. Yin M, Gao J, Lin Z. Laplacian regularized low-rank representation and its applications. IEEE Trans Pattern Anal Mach Intell. 2016;38(3):504–17.
15. Jiao C-N, Gao Y-L, Yu N, Liu J-X, Qi L-Y. Hyper-graph regularized constrained NMF for selecting differentially expressed genes and tumor classification. IEEE J Biomed Health Inform. 2020;24(10):3002–11.
16. Liu G, Lin Z, Yu Y. Robust Subspace Segmentation by Low-Rank Representation. In: *International conference on machine learning*: Edited by Fürnkranz J, Joachims T. Omnipress2600 Anderson StMadisonWIUnited States 2010: 663–670.
17. Liu G, Lin Z, Yan S, Sun J, Yu Y, Ma Y. Robust recovery of subspace structures by low-rank representation. IEEE Trans Pattern Anal Mach Intell. 2013;35(1):171–84.
18. Tenenbaum JB, De Silva V, Langford JC. A global geometric framework for nonlinear dimensionality reduction. Science. 2000;290(5500):2319–23.
19. Belkin M, Niyogi P. Laplacian eigenmaps for dimensionality reduction and data representation. Neural Comput. 2003;15(6):1373–96.
20. Donoho DL, Grimes C. Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data. Proc Natl Acad Sci. 2003;100(10):5591–6.
21. Gao S, Tsang IW-H, Chia L-T. Laplacian sparse coding, hypergraph laplacian sparse coding, and applications. *IEEE Trans Pattern Anal Mach Intell.* 2012, 35(1):92–104.
22. Liu G, Yan S: Latent low-rank representation for subspace segmentation and feature extraction. In: *2011 International conference on computer vision*. IEEE 2011: 1615–1622.
23. Cao F, Chen J, Ye H, Zhao J, Zhou Z. Recovering low-rank and sparse matrix based on the truncated nuclear norm. Neural Netw. 2017;85:10–20.
24. Zhang D, Hu Y, Ye J, Li X, He X: Matrix completion by truncated nuclear norm regularization. In: *2012 IEEE conference on computer vision and pattern recognition: 2012*. IEEE 2012: 2192–2199.
25. Yao H, Debing Z, Jieping Y, Xuelong L, Xiaofei H. Fast and accurate matrix completion via truncated nuclear norm regularization. IEEE Trans Pattern Anal Mach Intell. 2013;35(9):2117–30.
26. Liu Q, Lai Z, Zhou Z, Kuang F, Jin Z. A truncated nuclear norm regularization method based on weighted residual error for matrix completion. IEEE Trans Image Process. 2015;25(1):316–30.
27. Lin Z, Liu R, Su Z. Linearized alternating direction method with adaptive penalty for low-rank representation. In: *International conference on neural information processing systems: 2011*. 612–620.
28. Wang Y-X, Gao Y-L, Liu J-X, Kong X-Z, Li H-J. Robust principal component analysis regularized by truncated nuclear norm for identifying differentially expressed genes. IEEE Trans Nanobiosci. 2017;16(6):447–54.
29. Cai D, He X, Han J, Huang TS. Graph regularized nonnegative matrix factorization for data representation. IEEE Trans Pattern Anal Mach Intell. 2011;33(8):1548–60.
30. Cui Y, Zheng C-H, Yang J. Identifying subspace gene clusters from microarray data using low-rank representation. *Plos One* 2013, 8(3):e59377.
31. Yin M, Gao J, Lin Z, Shi Q, Guo Y. Dual graph regularized latent low-rank representation for subspace clustering. IEEE Trans Image Process. 2015;24(12):4918–33.
32. Ponder BA. Cancer genetics. Nature. 2001;411(6835):336–41.
33. Pei Y, Wang P, Liu H, He F, Ming L. FOXQ1 promotes esophageal cancer proliferation and metastasis by negatively modulating CDH1. Biomed Pharmacother. 2015;2015(74):89–94.
34. Schutter HD, Geeraerts H, Verbeken E, Nuyts S. Promoter methylation of TIMP3 and CDH1 predicts better outcome in head and neck squamous cell carcinoma treated by radiotherapy only. Oncol Rep. 2009;21(2):507–13.
35. Zhao L, Wang YX, Xi M, Liu SL, Zhang P, Luo LL, Liu MZ. Association between E-cadherin (CDH1) polymorphisms and pancreatic cancer risk in Han Chinese population. Int J Clin Exp Pathol. 2015;8(5):5753.
36. Xu Z, Zheng X, Yang L, Liu F, Zhang E, Duan W, Bai S, Safdar J, Li Z, Sun C. Chemokine receptor 7 promotes tumor migration and invasiveness via the RhoA/ROCK pathway in metastatic squamous cell carcinoma of the head and neck. Oncol Rep. 2015;33(2):849–55.
37. Faried A, Nakajima M, Sohda M, Miyazaki T, Kato H, Kuwano H. Correlation between RhoA overexpression and tumour progression in esophageal squamous cell carcinoma. Eur J Surg Oncol. 2005;31(4):410–4.

38. Kusama T, Mukai M, Iwasaki T, Tatsuta M, Matsumoto Y, Akedo H, Nakamura H. Inhibition of epidermal growth factor-induced RhoA translocation and invasion of human pancreatic cancer cells by 3-hydroxy-3-methylglutaryl-coenzyme a reductase inhibitors. Can Res. 2001;61(12):4885–91.
39. Wu H, Zheng J, Deng J, Zhang L, Li N, Li W, Li F, Lu J, Zhou Y. LincRNA-uc002yug.2 involves in alternative splicing of RUNX1 and serves as a predictor for esophageal cancer and prognosis. *Oncogene* 2015, 34(36):4723–4734.
40. Cathrine H, Schildkraut JM, Murphy SK, Wong-Ho C, Vaughan TL, Harvey R, Marks JR, Jirtle RL, Brian C, Brian C. IGF2R polymorphisms and risk of esophageal and gastric adenocarcinomas. Int J Cancer. 2009;125(11):2673–8.
41. Natarajan BK. Sparse approximate solutions to linear systems. SIAM J Comput. 1995;24(2):227–34.
42. Fang X, Han N, Wu J, Xu Y, Yang J, Wong WK, Li X. Approximate low-rank projection learning for feature extraction. IEEE Trans Neural Netw Learn Syst. 2018;29(11):5228–41.
43. Wang J, Lu C-H, Liu J-X, Dai L-Y, Kong X-Z. Multi-cancer samples clustering via graph regularized low-rank representation method under sparse and symmetric constraints. BMC Bioinform. 2019;20(1):718.

Publisher's Note