A P-Norm Singular Value Decomposition Method for Robust Tumor Clustering

Xiang-Zhen Kong, Jin-Xing Liu*, Chun-Hou Zheng*, Mi-Xiao Hou, Yao Lu
School of Information Science and Engineering, Qufu Normal University, Rizhao, 276826
Shandong, China

*Corresponding author
XZX: kongxzheng@163.com, JXG: sdcavell@126.com, CHZ: zhengch999@126.com, MXH: mixiaohou@163.com, YLL: llfzyao136.com

Abstract—Tumor clustering based on biomolecular data plays a very important role for cancer classifications discovery. To further improve the robustness, stability and accuracy of tumor clustering, we develop a novel dimension reduction method named p-norm singular value decomposition (PSVD) to seek a low-rank approximation matrix to the biomolecular data. To enhance the robustness to outliers, the \( L_p \)-norm is taken as the error function and the Schatten \( p \)-norm is used as the regularization function in our optimization model. To evaluate the performance of PSVD, Kmeans clustering method is then employed for tumor clustering based on the low-rank approximation matrix. The extensive experiments are performed on gene expression dataset and cancer genome dataset respectively. All experimental results demonstrate that the PSVD-based method outperforms many existing methods. Especially it is experimentally proved that the proposed method is efficient for processing higher dimensional data with good robustness and superior time performance.

Keywords—dimension reduction; robust tumor clustering; Schatten p-norm; \( L_p \)-norm; singular value decomposition

I. INTRODUCTION

In the past decades, many tumor clustering approaches have been developed and used to perform cancer class discovery from biomolecular data. A reliable and precise identification of the type of tumors is effective for cancer diagnosis and treatment. However, the typical characteristic of the gene expression data or genomic data is high dimension low sample size [1], which makes most of the standard statistical methods lose effective. Firstly, including too many variables may decrease the clustering accuracy of the samples, and make the traditional cluster rules difficult to set. Secondly, the irrelevant or noisy variables in the original data may also degrade the performances of the estimated cluster algorithms.

Despite these difficulties, most of research works for tumor clustering from the areas of statistical machine learning have demonstrated the potential power for tumor-type identification [2] [3] [4] [5] [6] [7]. Most of these research works focus on data reduction and denoising. It means that the clustering works is in two steps. The first step is extracting features from the gene expression data. The second step is using the classic classifier or clustering algorithm to classify the tumor samples based on the extracted features. Zheng et al. extracted features using nonnegative matrix factorization (NMF) [3] [7] and sparse nonnegative matrix factorization (SNMF) [4] to improve the performance of classification [5]. Lee proposed the sparse singular value decomposition (SSVD) for biclustering in gene expression data [8]. More and more dimension reduction methods such as sparse principal component analysis (SPCA) [9] [10] and penalized matrix decomposition (PMD) [11] [12] [13] were used successfully to analyze gene expression data. Most of the methods mentioned above are taking advantage of norm constraint. For example, \( L_0 \)-norm penalty was used to analyze gene expression data by Journee [10]. \( L_p \)-norm was taken as the regularization function in SSVD [8] and PMD [12]. But the non-robust, too longer convergence time, and too many iterations often made them unsatisfactory.

In this paper, a new dimension reduction method base on singular value decomposition (SVD) named \( p \)-norm singular value decomposition (PSVD) is proposed. Compared with SSVD proposed by Lee for biclustering [8], PSVD uses the \( L_p \)-norm in place of the squared Frobenius norm as the error function to improve the robustness to outliers in the data, and the Schatten \( p \)-norm is used as the regularization function in place of \( L_1 \) norm for sparse vectors \( u \) and \( v \). The augmented Lagrangian multiplier (ALM) [14] [15] [16] and alternating direction method (ADM) [17] are employed to deal with the problem nonsmooth and somewhat intractable in the optimization model. The technical details of PSVD will be provided in Section II. To evaluate the validity of PSVD, the traditional and classic unsupervised clustering method—Kmeans is then used for tumor clustering based on the low-rank approximation matrix obtained by PSVD. Four datasets including one gene expression dataset and three genomic datasets from The Cancer Genome Atlas (TCGA) are used in the experiments. Whether compared with other dimension reduction methods such as SSVD, SPCA, SNMF or compared with the ordinary clustering methods such as Kmeans, the experimental results show that our method is efficient and feasible. The clustering accuracy is improved, the convergence time is shorter, and especially it is proved that PSVD is more efficient for processing higher dimensional unbalanced genome data from TCGA.

The rest of the paper is organized as follows. Section II describes the PSVD algorithm to seek low-rank approximation matrix of biomolecular data. The tumor clustering experiments based on the low-rank approximation matrix obtained by PSVD are performed in Section III. Section IV concludes the paper and outlines directions of future research works.
II. METHODOLOGY

A. Definitions of $p$-norm and Schatten $p$-norm

The $L_p$-norm of a matrix $X \in \mathbb{R}^{m \times n}$ is defined as

$$
\|X\|_p = \left(\sum_{i,j} |x_{ij}|^p \right)^{1/p} \quad (0 < p \leq \infty).
$$

Nie et al. used the $L_p$-norm ($0 < p \leq 1$) as the error function to improve the robustness to outliers in given data [18] to solve the robust matrix completion problem.

The extended Schatten $p$-norm ($0 < p < \infty$) of a matrix $X \in \mathbb{R}^{m \times n}$ is defined as

$$
\|X\|_{Sp}^p = \max_{\sigma_i \neq 0} \sigma_i^p \quad (1),
$$

where $\sigma_i$ is the $i$-th singular value of $X$. Equation (2) is the rank of $X$ with the definition of $\sigma_i = 0$. When $p_i \to 0$, the Schatten $p$-norm of $X$ will approximate the rank of $X$ [19].

When $p_i = 1$, the Schatten 1-norm is the nuclear norm, which is usually taken as the form $\|X\|_*$.

B. The Method of PSVD

The PSVD procedure is presented in this section. Let $X$ be an $n \times m$ biomolecular matrix whose every row represents all of genes’ expression level in one sample. The rank-$K$ ($K \leq r$) approximation to $X$ can be written as

$$
X \approx X^{(K)} = \sum_{k=1}^{K} u_k d_k v_k^T.
$$

$X^{(K)}$ gives the rank-$K$ matrix approximation to $X$ in the sense that $X^{(K)}$ minimizes the squared Frobenius norm:

$$
X^{(K)} = \arg \min_{X^{(K)}} \|X - X^{(K)}\|_F,
$$

where $A_k$ is the set of all $n \times m$ matrices of rank $K$ [20].

According to [20], the gene expression data or the genomic data always lies near low dimensional subspace. Therefore, the PSVD is carefully designed to seek low-rank approximation matrix of the original data. That means the rank-$K$ low-rank approximation matrix obtained by PSVD is the summation of the first $K$ layers of SVD in the sense of $p$-norm. Our presentation focuses on how to extract the first PSVD layer, the subsequent layers can be extracted sequentially from the residual matrices by subtracting the preceding layers.

The optimization model of the first layer is as follow,

$$(u_i, d_i, v_i) = \arg \min_{(u_i, d_i, v_i)} \|X - u_i d_i v_i^T\|_F + \lambda_1 \|d_i\|_p + \lambda_2 \|v_i\|_p, \quad (5)$$

where $0 < p \leq 1$ and $0 \leq \lambda_1 \leq 1$ according literature [21], $d$ is a positive scalar, $u$ is a unit $n$-vector, and $v$ is a unit $p$-vector. $\lambda_1$ and $\lambda_2$ are penalty parameters that balance the goodness-of-fit for measure $\|X - u_i d_i v_i^T\|_F$. We note the first PSVD layer $u_i d_i v_i^T$ is the best rank-one matrix approximation of $X$ under the $p$-norm. When $p_i = 2$ and $\lambda = \lambda_2 = 0$, the model of (5) reduces to obtain the plain SVD layers. When $p_i = 2$ and $p_i = 1$, (5) is the SVD model proposed by Lee et al. [8] who obtained the sparse vectors $u$ and $v$ using the lasso regression [22] [23] [24] and Bayesian information criterion (BIC) [25].

To obtain the sparse $u$ and $v$ in the first PSVD layer, (5) is converted to easier solved forms comparatively. Fixing $v$, with respect to $u = du$, minimization of (5) is equivalent to minimizing

$$
\|X - \hat{u} v^T\|_F + \lambda_1 \|\hat{u}\|_p.
$$

Equation (6) can be easily converted to the form:

$$
\|X V - \hat{u} v^T\|_F + \lambda_1 \|\hat{u}\|_p = \|X V - \hat{u}^\prime v^T\| + \lambda_1 \|\hat{u}\|_p.
$$

Similarly, for fixed $u$, minimization of (5) with respect to $\hat{v} = dv$ is equivalent to minimizing

$$
\|X - u d v^T\|_F + \lambda_2 \|v\|_p,
$$

and (8) can be converted to the form:

$$
\|d' X - u' d v^T\|_p + \lambda_2 \|d\|_p = \|d' X - u' \|_p + \lambda_2 \|d\|_p.
$$

Equation (9) is equivalent to minimizing the form:

$$
\|X' u - v\|_F + \lambda_2 \|v\|_p.
$$

Now, to obtain the first PSVD layer, we successfully convert the optimization model (5) to (7) and (10). Next, the main purpose is to solve (7) and (10). These two formulas are essentially of the same form. So we give the solution to (7) in the following presentation, and (10) can be solved in the same manner.
Equation (7) is intractable since both of the two items are nonsmooth. The augmented Lagrangian multiplier (ALM) [14] [15] [16] and alternating direction method (ADM) [17] are recommended to deal with the problem [21]. The ALM method is very attractive because it has been proved that under some rather conditions, the ALM algorithm converges Q-linearly to the optimal solution [16].

To solve (7) using ALM method, we replace $Xv - \tilde{u}$ as $E$, and replace $\tilde{u}$ as $Z$, $\lambda_i$ as $\gamma$. Equation (7) can be equivalently rewritten as:

$$
\min_{E,Z,\gamma} \frac{1}{2} \|E\|_F^2 + \gamma \|Z\|_F^2.
$$

(11)

According to ALM, the following problem is needed to solve:

$$
\min_{E,Z,\gamma} \frac{1}{2} \|E\|_F^2 + \gamma \|Z\|_F^2 + \frac{\mu}{2} \left( \|E - (Xv - \tilde{u})\|_F^2 + \frac{\lambda}{2} \|Z\|_F^2 + \frac{\Omega}{2} \right).
$$

(12)

In (12), it is very difficult and costly with respect to three variables $\tilde{u}$, $E$, $Z$, the ADM can be adopted to deal with this problem suitable. The problem with one variable can be easily optimized when fixing the other two variables. In this way, (12) results in the following three subproblems:

**Problem 1:** When fixing $E$ and $Z$, (12) can be written as:

$$
\min_{\tilde{u}} \frac{\mu}{2} \|\tilde{u}\|_2^2 + \frac{\Omega}{2} \|Z - \tilde{u}\|_2^2 + \frac{\lambda}{2} \|E - \tilde{u}\|_F^2,
$$

(13)

where $\zeta = (Xv - E - \tilde{u})$, and $\tau = (Z - \tilde{u})$. Equation (13) is equivalent to solving a quadratic function, the solution can be easily obtained by

$$
\tilde{u} = \frac{\zeta + \tau}{2}.
$$

(14)

**Problem 2:** When fixing $\tilde{u}$ and $Z$, (12) is simplified as:

$$
\min_{E} \frac{\mu}{2} \|E\|_F^2 + \frac{\Omega}{2} \|E - \tilde{u}\|_F^2,
$$

(15)

where $\eta = (Xv - \tilde{u} - \tilde{u})$.

**Problem 3:** When fixing $\tilde{u}$ and $E$, (12) can be written as:

$$
\min_{Z} \gamma \|Z\|_F^2 + \frac{\lambda}{2} \|Z - \tilde{u}\|_F^2,
$$

(16)

where $\psi = \tilde{u} + \frac{1}{\mu} \Omega$.

The optimal solutions to subproblems in (15) and (16) are described detailedly in literatures [21] [26]. Here we do not repeat it in tautology.

**Algorithm.** Algorithm to extract the first PSVD layer

Input: bimolecular data matrix: $X(X \in \mathbb{R}^{m \times n})$

1. decompose $X(X \in \mathbb{R}^{m \times n})$ using standard SVD, get the first SVD layer, a triple $(d_{old}, u_{old}, v_{old})$
2. $v < v_{old}$, $\tilde{u} \leftarrow d_{old}u_{old}$
3. Set $1 < \rho < 2$. Initialize $\mu > 0$, $\Lambda, \Omega, E, Z$

**while** not convergence do

Update $\tilde{u}$ by (14)

Update $E$ by the optimal solution to (15)

Update $Z$ by the optimal solution to (16)

Update $\Lambda$ by $\Lambda = \Lambda + \mu(\tilde{u} - E)$, update $\Omega = \Omega + \mu(Z - \tilde{u})$

Update $\mu$ by $\mu = \rho \mu$

**end while**

4. $d \leftarrow \tilde{d}_{nov}$, $\tilde{v} \leftarrow d_{nov} \tilde{v}_{nov}$, $\tilde{v}_{nov}$ can be obtained similarly as $\tilde{u}_{nov}$.

5. $\tilde{u} \leftarrow \tilde{u}_{nov} / \|\tilde{d}_{nov}\|$, $\tilde{v} \leftarrow \tilde{v}_{nov} / \|\tilde{v}_{nov}\|$, $\tilde{d} \leftarrow \tilde{d}_{nov}X_{\tilde{v}_{nov}}$, $X_{\tilde{v}_{nov}} \leftarrow \tilde{u}_{nov}d\tilde{v}_{nov}$

After the first $K$ PSVD layers are extracted, the rank-$K$ approximation to the matrix $X$ is computed as $X_{K} = \sum_{i=1}^{K} \tilde{u}_{i}\tilde{d}_{i}\tilde{v}_{i}$.

### III. Experiments and Discussion

Four datasets are used to demonstrate the performance of PSVD including one high-dimensional gene expression dataset and three higher-dimensional genomic datasets from TCGA. For every dataset, we employ the PSVD to extract the first $K$ layers to obtain a rank-$K$ approximation matrix. Based on the low-rank matrix, the classical unsupervised clustering algorithm K-means is used to evaluate the performance of PSVD. As comparisons, experiments are also carried out using the existing methods such as SSVD, SNMF, and SPCA. The cancer type information is only used as a posterior to interpret the analysis results.

**A. Lung Cancer Data**

We especially compare PSVD with SSVD on the same subset of lung cancer gene expression dataset [27] that was
used by Lee to illustrate the SSVD algorithm [8]. This dataset contains 12,625 genes for 56 samples. The samples contain 4 types of lung cancer. There are 20 pulmonary carcinoid samples (Carcinoid), 13 colon metastases (Colon), 17 normal lung samples (Normal) and 6 small cell lung carcinoma samples (SmallCell).

Similar to PSVD, SSVD is also used to seek the low-rank approximation to data matrix through extracting the first $K$ layers. SSVD used Frobenius norm as the error function and used $L_1$ norm as the regularization function for sparse vectors $u$ and $v$. The first three layers are extracted sequentially using PSVD and SSVD respectively. The reason for considering only the first three layers is that the three singular values are much bigger than the rest [8]. For every layer, we compare the convergence time of the two algorithms. Table I lists the results using the Matlab program running on a Windows 7 desktop with Intel® Core™ i5-4590 Duo CPU of a clock speed of 3.30 GHz. It takes 1.98, 2.08, 2.10 seconds using PSVD for the first three layers respectively, however SSVD converges in 401.55, 1127.72, and 1185.09 seconds. Evidently, the time performance of our algorithm is far beyond SSVD. It is very important for dimension reduction of high dimensional data.

To further demonstrate the performance of PSVD, we plot the first three sparse left singular vectors $\hat{u}_k, \hat{v}_k$ using PSVD and SSVD respectively in scatter plots. The subject grouping/clustering can be seen in Fig. 1. Different colors and symbols are used to interpret the cancer types easily. Comparing Fig. 1 (b) with Fig. 1 (a), it is found obviously that the first two vectors in PSVD reveal four sample clusters, however, the first two vectors in SSVD reveal only three with mixing Colon and SmallCell. The next two panels in Fig. 1 (b) and Fig. 1 (a) can also demonstrate the better discrimination of PSVD-vectors compared with SSVD-vectors.

To better analyze the advantage of the proposed method, the Kmeans clustering algorithm is employed to evaluate the performance of PSVD based on the sparse rank-3 approximation matrix $\sum_k\hat{u}_k\hat{v}_k$ of the raw data. The experiment results are also compared with that of the competitive method SSVD. The accuracy rate of clustering of the proposed method PSVD-Kmeans is 96.43% (54/56), only two Carcinoid samples are misclassified to Colon sample. However, the SSVD-Kmeans achieves 83.93% (47/56) accuracy rate. On this dataset, the number of PSVD layers extracted is selected as the same as that recommended in SSVD method [8]. We verified by experiment that the first three layers could get the best clustering results.

| TABLE I. THE RUNNING TIME TO EXTRACT THE FIRST $K$ LAYERS USING SSVD AND PSVD |
|------------------------------------|-------|-------|
| Layers ($K$) | SSVD (Seconds) | PSVD (Seconds) |
| 1 | 401.55 | 1.98 |
| 2 | 1127.72 | 2.08 |
| 3 | 1185.09 | 2.10 |

Fig. 1. (a) Scatter plots of the entries of the first three left sparse singular vectors $\hat{u}_k (k=1,2,3)$ using SSVD. (b) Scatter plots of the entries of the first three left sparse singular vectors $\hat{u}_k (k=1,2,3)$ using PSVD.

From the experiment results on this dataset, also as a dimension reduction method based on extracting sparse vectors, PSVD can make the class structure more evident compared with SSVD. Whether the time performance or the accuracy rate of clustering, PSVD outperforms SSVD.

B. Genome Data

In the experiments on the dataset above, PSVD shows the advantages over other methods. In this section, we employ PSVD on higher dimensional data to confirm the performance for genome data from TCGA. Three datasets of genome are used to evaluate the PSVD-based method comparing with SNMF, SVD and SPCA. SSVD is not taken as a competitor because it is found that the convergence time of SSVD algorithm is too long and sometimes it does not converge with the computer crashed when it is used on higher dimension genomic data. That means SSVD maybe is not suitable to process higher dimensional data. The three datasets are Colorectal Cancer (CRC) dataset, Cholangiocarcinoma (CHOL) dataset and Squamous Cell Carcinoma of Head and Neck (HNSC). All of them include 20, 502 official identified genes, but the number of samples is different. The number of
samples contained respectively in the three datasets is 281 (CRC), 45 (CHOL) and 418 (HNSC). Every dataset is mixed of two types of samples “positive” and “negative”. We simplify “positive” as “TP” which represents diseased sample, “negative” as “NT” represents normal samples. There are 19, 9, and 20 “NT” samples in CRC, CHOL and HNSC datasets, respectively.

Firstly, we study the appropriate number of layers to be extracted by PSVD. The first five layers are extracted sequentially. Based on the obtained rank- $K(K=1,2,3,4,5)$ approximation matrix to the raw data, the PSVD-Kmeans clustering accuracy rates are compared. It is showed in Fig. 2 that for CRC and HNSC datasets, the achieved clustering rate is best when $K = 3$. For CHOL dataset, the results with different $K$ is identical. We surmise that there are distinctly discriminations for the “TP” samples and the “NT” samples in the original dataset.

Next, the PSVD-based method is compared with SNMF, SVD and SPCA. We also adjust the parameters of the competitive algorithm to obtain the best performance. According to the analyses above, the first three layers are extracted using PSVD and SVD to get the best Kmeans clustering performance for these three genomic datasets. As one can see in Table II, PSVD performs better over the competitive methods. For CRC dataset, PSVD performs best with the accuracy of clustering 93.24%. For CHOL dataset, all method can distinguish the “TP” and the “NT” samples correctly. PSVD achieve the best accuracy rate 96.17% over other methods on HNSC dataset.

Finally, we investigate the time performance of PSVD on these datasets. Take HNSC dataset (it contains the largest number of samples in the three datasets) for example, it takes 4.61, 6.33, 6.34 seconds for the first three layers respectively running the Matlab program on a Windows 7 desktop with Intel® Core™ i5-4590 Duo CPU of a clock speed of 3.30 GHz. The running time for the other two datasets to extract the corresponding layer is shorter. It shows that PSVD is very effective for dimension reduction of higher dimensional data.

### IV. CONCLUSIONS AND FUTURE WORKS

In this article, we propose a novel dimension reduction methods based on $L_p$-norm and Schatten $p$-norm named PSVD. The effectiveness of PSVD has been demonstrated through biomolecular data including gene expression data and genome data. Experiments show that our proposed method is especially suitable for processing higher dimensional data with good robustness and excellent time performance. There are a few interesting potential directions for future research works. Firstly, it is interesting to evaluate the usage of PSVD as a dimension reduction tool combining other clustering or classification methods. Secondly, we notice that every extracted PSVD layer has a checkerboard structure. It may be employed also for biclustering as SSVD. We will continue to explore the performance of PSVD in future research.

### ACKNOWLEDGMENT

This work is supported in part by the grants of the National Science Foundation of China, Nos. 61572284, 61502272, 61572283, and 61402258; Foundation of Science and Technology Project of Qufu Normal University, No. xkj201621.

### REFERENCES


<table>
<thead>
<tr>
<th>Genome dataset</th>
<th>Kmeans</th>
<th>SNMF</th>
<th>SVD-Kmeans</th>
<th>SPCA-Kmeans</th>
<th>PSVD-Kmeans</th>
</tr>
</thead>
<tbody>
<tr>
<td>CRC</td>
<td>82.56% (232/281)</td>
<td>82.56% (232/281)</td>
<td>79.00% (222/281)</td>
<td>92.88% (261/281)</td>
<td>93.24% (263/281)</td>
</tr>
<tr>
<td>CHOL</td>
<td>100.00%</td>
<td>100.00%</td>
<td>100.00%</td>
<td>100.00%</td>
<td>100.00%</td>
</tr>
<tr>
<td>HNSC</td>
<td>79.90% (334/418)</td>
<td>95.45% (399/418)</td>
<td>95.45% (399/418)</td>
<td>93.06% (389/418)</td>
<td>96.17% (402/418)</td>
</tr>
</tbody>
</table>

### Fig. 2. Genome data: the accuracy rate of clustering by extracting $K(K=1,2,3,4,5)$ PSVD layers through the three genomic datasets CRC, CHOL and HNSC.


