Appl. Comput. Harmon. Anal. 45 (2018) 425–435

Contents lists available at ScienceDirect

Applied and Computational Harmonic Analysis

www.elsevier.com/locate/acha

Letter to the Editor

Similarity matrix framework for data from union of subspaces

Akram Aldroubi [a,1], Ali Sekmen [b,*,2], Ahmet Bugra Koku [b,c,2,3], Ahmet Faruk Cakmak [b,2]

[a] Department of Mathematics, Vanderbilt University, Nashville, TN, 37212 USA
[b] Department of Computer Science at Tennessee State University, Nashville, TN 37209, USA
[c] Department of Mechanical Engineering at Middle East Technical University, Ankara, Turkey

ARTICLE INFO

Article history:
Received 29 June 2016
Received in revised form 17 April 2017
Accepted 18 August 2017
Available online 4 September 2017
Communicated by Naoki Saito

Keywords:
Subspace segmentation
Union of subspaces
Data clustering
Similarity matrix
Shape interaction matrix
Skeleton decomposition

ABSTRACT

This paper presents a framework for finding similarity matrices for the segmentation of data $\mathbf{W} = [w_1 \cdots w_N] \subset \mathbb{R}^D$ drawn from a union $\mathcal{U} = \bigcup_{i=1}^M S_i$ of independent subspaces $\{S_i\}_{i=1}^M$ of dimensions $\{d_i\}_{i=1}^M$. It is shown that any factorization of $\mathbf{W} = BP$, where columns of B form a basis for data $\mathbf{W}$ and they also come from $\mathcal{U}$, can be used to produce a similarity matrix $\Xi_{\mathbf{W}}$. In other words, $\Xi_{\mathbf{W}}(i,j) \neq 0$, when the columns w_i and w_j of $\mathbf{W}$ come from the same subspace, and $\Xi_{\mathbf{W}}(i,j) = 0$, when the columns w_i and w_j of $\mathbf{W}$ come from different subspaces. Furthermore, $\Xi_{\mathbf{W}} = Q^{d_{max}}$, where $d_{max} = \max\{d_i\}_{i=1}^M$ and $Q \in \mathbb{R}^{N \times N}$ with $Q(i,j) = |P^T P(i,j)|$. It is shown that a similarity matrix obtained from the reduced row echelon form of $\mathbf{W}$ is a special case of the theory. It is also proven that the Shape Interaction Matrix defined as VV^T, where $\mathbf{W} = U\Sigma V^T$ is the skinny singular value decomposition of $\mathbf{W}$, is not necessarily a similarity matrix. But, taking powers of its absolute value always generates a similarity matrix. An interesting finding of this research is that a similarity matrix can be obtained using a skeleton decomposition of $\mathbf{W}$. First, a square sub-matrix $A \in \mathbb{R}^{r \times r}$ of $\mathbf{W}$ with the same rank r as $\mathbf{W}$ is found. Then, the matrix R corresponding to the rows of $\mathbf{W}$ that contain A is constructed. Finally, a power of the matrix $P^T P$ where $P = A^{-1}R$ provides a similarity matrix $\Xi_{\mathbf{W}}$. Since most of the data matrices are low-rank in many subspace segmentation problems, this is computationally efficient compared to other constructions of similarity matrices.

1. Introduction

In this research, the focus is on the generation of similarity matrices for clustering a set of data points that are drawn from a union of subspaces. Specifically, given a set of data $\mathbf{W} = \{w_1, ..., w_N\} \subset \mathbb{R}^D$ drawn

* Corresponding author.
E-mail addresses: akram.aldroubi@vanderbilt.edu (A. Aldroubi), asekmen@tnstate.edu (A. Sekmen).

[1] The research of A. Aldroubi is supported by NSF Grant NSF/DMS 132099.
[2] The research of A. Sekmen, A. B. Koku, and A. Cakmak is supported by US Army Grant W911NF-15-1-0495.
[3] The research of A. B. Koku is also supported by TUBITAK-2219-1059B191600150.

http://dx.doi.org/10.1016/j.acha.2017.08.006
1063-5203/

from a union of subspaces of the form $\mathcal{U} = \bigcup_{i \in I} S_i$, where $\left\{S_i \subset \mathbb{R}^D\right\}_{i \in I}$ is a set of subspaces, we wish to define a similarity matrix that allows us to

1) determine the number of subspaces $M = |I|$,
2) determine the set of dimension d_i for each subspace S_i,
3) find an orthonormal basis for each subspace S_i,
4) collect the data points belonging to the same subspace into the same cluster.

Union of subspace models have become common in several areas of mathematics and its applications, such as sampling, compressed sensing, and frame theory [1–4]. For example, all images of a given face i with same facial expression, obtained under different illuminations and facial positions, can be modeled as a set of vectors belonging to a low dimensional subspace S_i living in a higher dimensional space $\mathbb{R}^D$ [5,6]. Another example is segmentation of moving rigid objects in videos. Consider a video with F frames of a scene that contains multiple moving rigid objects. Let p be a point on one of these objects and let $x_i(p), y_i(p)$ be the coordinates of p in frame i. Define the *trajectory vector* of p as the vector $w(p) = (x_1(p), y_1(p), x_2(p), y_2(p), \ldots, x_F(p), y_F(p))^T$ in $\mathbb{R}^{2F}$. In this case, the trajectory vectors of multiple independent motions lie in 4-dimensional independent subspaces in $\mathbb{R}^{2F}$ [7,8]. However, it should be noted that independence is a strong assumption for real-world problems (e.g. motion of non-rigid objects or motion of rigid objects on the same planar surface).

In this research, we focus on a way of finding similarity matrices that can be used in clustering algorithms. As such, our aim is to concentrate and understand the first step used in many subspace clustering algorithms. Typically, a method for finding a similarity matrix is used, followed by spectral clustering (a common practice in computer vision and machine learning is to use the absolute value of a similarity matrix as an affinity before spectral clustering). For example, a method related to compressed sensing by Liu et al. [9,10] finds the lowest rank representation of the data matrix. The lowest rank representation is then used to define the similarity of an undirected graph, which is then followed by spectral clustering. It is shown in [11] that the low-rank minimization problem of [9] results in the shape interaction matrix VV^T and the problem is related to factorization rather than sparsity. There are many other methods that produce a similarity matrix as a stage for further processing, such as sparsity methods [12–14,9], algebraic methods [15–17], iterative and statistical methods [18,8,19–22], and spectral clustering methods [13,14,23–29]. Some important methods on subspace clustering are reviewed and their advantages and disadvantages discussed in [28]. Spectral graph partitioning and harmonic analysis on graphs and networks data are explained in [30].

In this research, graph connectivity of data nodes is analyzed to develop theory for a general framework of similarity matrices. Some of the existing techniques generates affinity matrices that has some graph connectivity issues. A discussion on this is given [31]. For example, the well-celebrated Sparse Subspace Clustering (SSC) algorithm produces a sparse affinity matrix whose (i, j)th entry is non-zero only if the data points x_i and x_j are from the same subspaces. However, it is not guaranteed that the data points form the same subspace generates a connected graph. Low-Rank Representation (LRR) method results in an affinity matrix, which is equivalent to Shape Interaction Matrix VV^T [9]. However, in this paper, we show that Shape Interaction Matrix may not always be a similarity matrix and it may lead over-segmentation problem for a set with measure zero.

1.1. Paper contributions

- This paper presents a mathematical framework for finding similarity matrices for segmentating data $\mathbf{W} = [w_1 \cdots w_N] \subset \mathbb{R}^D$ drawn from a union $\mathcal{U} = \bigcup_{i=1}^M S_i$ of independent subspaces $\{S_i\}_{i=1}^M$ of dimensions $\{d_i\}_{i=1}^M$. It is shown that any factorization $\mathbf{W} = BP$, where the columns of B come from $\mathcal{U}$ and form a basis the column space of $\mathbf{W}$, can be used to produce a similarity matrix $\Xi_{\mathbf{W}}$, i.e., if $\Xi_{\mathbf{W}}(i, j) \neq 0$, the columns w_i and w_j of $\mathbf{W}$ come from the same subspace, similarly, and if $\Xi_{\mathbf{W}}(i, j) = 0$, the columns

w_i and w_j of $\mathbf{W}$ come from different subspaces. In particular, $\Xi_{\mathbf{W}} = Q^{d_{max}}$, where $d_{max} = \max\{d_i\}_{i=1}^M$ and $Q \in \mathbb{R}^{N\times N}$ is given by $Q(i,j) = \left|P^T P(i,j)\right|$.

- It is proven that the Shape Interaction Matrix defined as VV^T, where $\mathbf{W} = U\Sigma V^T$ is the skinny singular value decomposition of $\mathbf{W}$, does not always produce a similarity matrix even though VV^T has been widely used as a similarity matrix in subspace segmentation research [32]. However, we show that VV^T can always be used to obtain a similarity matrix as follows: (1) Define $Q \in \mathbb{R}^{N\times N}$ with $Q(i,j) = \left|VV^T(i,j)\right|$, and (2) $\Xi_{\mathbf{W}} = Q^{d_{max}}$, where d_{max} is the dimension of the highest dimensional subspace. For example, $d_{max} = 4$ for segmentation of independent rigid body motion.
- It is shown that a similarity matrix can be obtained from the reduced row echelon form of $\mathbf{W}$ as in [33].
- An interesting finding of this research is that a similarity matrix can be obtained using a skeleton decomposition of $\mathbf{W}$. Let $A \in \mathbb{R}^{r\times r}$ be any square sub-matrix of $\mathbf{W}$ with the same rank r as $\mathbf{W}$, and let R be the rows of $\mathbf{W}$ containing A. Then a similarity matrix can be constructed using the matrix $P^T P$ where $P = A^{-1}R$. Since most of the data matrices are low-rank in many subspace segmentation problems, this is a computationally efficient method compared to other constructions of similarity matrices.

2. Preliminaries

In this paper, we assume that the subspaces are independent and data is generic (as defined below). We have the following definitions and some basic theory on subspaces.

Definition 1 *(Independent subspaces).* Subspaces $\{S_i \subset \mathbb{R}^D\}_{i=1}^M$ are called *independent* if their dimensions satisfy the following relationship:

$$dim(S_1 + \cdots + S_M) = dim(S_1) + \cdots + dim(S_M) \leq D.$$

The definition above is equivalent to the property that any set of non-zero vectors $\{w_1, w_2, \ldots, w_M\}$ such that $w_i \in S_i$, $i = 1, \ldots, M$ are linearly independent.

Definition 2 *(Generic data).* Let S be a linear subspace of $\mathbb{R}^D$ with dimension d. A set of data $\mathbf{W}$ drawn from S is said to be *generic* if (i) $|\mathbf{W}| > d$, and (ii) every d vectors from $\mathbf{W}$ form a basis for S.

Definition 3 *(Similarity matrix).* $\mathbf{W} = [w_1 \cdots w_N] \subset \mathbb{R}^D$ drawn from a union $\mathcal{U} = \bigcup_{i=1}^M S_i$ of independent subspaces $\{S_i\}_{i=1}^M$. We say $\Xi_{\mathbf{W}}$ is a similarity matrix for $\mathbf{W}$ if and only if (i) $\Xi_{\mathbf{W}}$ is symmetric, and (ii) $\Xi_{\mathbf{W}}(i,j) \neq 0$ implies that w_i and w_j come from the same subspace, and $\Xi_{\mathbf{W}}(i,j) = 0$ implies that w_i and w_j come from different subspaces.

Definition 4 *(Pre-similarity matrix).* Let $\mathbf{W} = [w_1 \cdots w_N] \subset \mathbb{R}^D$ drawn from a union $\mathcal{U} = \bigcup_{i=1}^M S_i$ of independent subspaces $\{S_i\}_{i=1}^M$. We say $\Pi_{\mathbf{W}}$ is a pre-similarity matrix for $\mathbf{W}$ if and only if there exists an integer $n \geq 1$ such that $(\Pi)^n$ is a similarity matrix. The order of a pre-similarity matrix is the smallest r such that $(\Pi)^r$ is a similarity matrix.

Note that a pre-similarity matrix is a similarity matrix if and only if its order is 1. Also note that, for a pre-similarity matrix Π of order r, $(\Pi)^n$ is a similarity matrix for $n \geq r$.

Definition 5 *(Absolute value of a matrix).* Let $A \in \mathbb{R}^{m\times n}$. We denote by $abs(A) \in \mathbb{R}^{m\times n}$ the absolute value version of A defined by $abs(A)(i,j) = |A(i,j)|$.

Definition 6 *(Binary version of a matrix).* Let $A \in \mathbb{R}^{m\times n}$. We denote by $bin(A) \in \mathbb{R}^{m\times n}$ the binary version of A defined by $bin(A)(i,j) = 1$ for $A(i,j) \neq 0$ and $bin(A)(i,j) = 0$ otherwise.

We state a well-known definition and theorem from graph theory (please refer to any graph theory textbook such as [34]).

Definition 7 *(Walk on a graph).* Let $G = (V, E)$ be a graph that includes self-loops. A walk on G is an alternating sequence $(v_1, e_{12}, v_2, e_{23}, v_3, ..., v_m)$ of vertices $\{v_1, v_2, v_3, ..., v_m\}$ and edges $\{e_{12}, e_{23}, ...\}$ beginning and ending with vertices, such that each edge's endpoints are the vertices preceding and following it.

Theorem 1 *(nth power of adjacency matrix). Let G be a undirected graph with vertices $\{v_1, ..., v_m\}$ and assume its adjacency matrix is A. Then, $A^n(i, j)$ is the number of walks of length n from v_i to v_j.*

Definition 8 *(Positively weighted adjacency matrix).* Let $G = (V, E)$ be a graph that includes self-loops. A *Positively Weighted Adjacency Matrix* A for G is such that $A(i, j) > 0$ if and only if there is an edge between v_i and v_j.

Definition 9 *(Diameter of a graph).* Let $G = (V, E)$ be a graph with vertices $\{v_1, ..., v_m\}$. The distance between v_i and v_j is the number of edges in the shortest path between v_i and v_j. The diameter d of G is defined as $d(G) = \max_{v_i, v_j \in V} dis(v_i, v_j)$. In other words, the diameter is the largest distance between two vertices in the graph.

Corollary 1 *(dth power of positively weighted adjacency matrix). Let G be an undirected connected graph with vertices $\{v_1, ..., v_m\}$ and assume its positively weighted adjacency matrix is A, i.e., $A(i, j) \geq 0$. Then, $A^d(i, j) > 0$ for all i, j, where d is the diameter of G.*

Proof. We first show by induction that there is a walk from v_i to v_j of length n, if and only if $A^n(i, j) > 0$. By construction, $A(i, j) > 0$ if and only if there is a walk of length one between v_i and v_j. Now assume that for any $v_i, v_j \in V$ there is a walk of length n between v_i and v_j if and only if $A^n(i, j) > 0$. And assume that there is a walk of length $n + 1$ between $v_m, v_l \in V$. Then, there is a k_0 such that there is a walk of length one between v_{k_0} and v_l, and a walk of length n between v_m and v_{k_0}. Thus $A(k_0, l) > 0$ by construction and $A^n(m, k_0) > 0$ by the induction hypothesis. Since $A^{n+1} = A^n A$, we have that $A^{n+1}(m, l) = \sum_k A^n(m, k)A(k, l) \geq A^n(m, k_0)A(k_0, l) > 0$.

For the converse. If $A^{n+1}(m, l) = \sum_k A^n(m, k)A(k, l)$, then there exists a k such that $A^n(m, k) > 0$ and $A(k, l) > 0$. Hence there is a walk of length n between v_m and v_k and a walk of length one between v_k and v_l. Finally, since the diameter of G is d, then there is a walk of length at most d between any two vertices. □

3. Main results

Theorem 2 below provides the relationship between a factorization of $\mathbf{W}$ and an associated similarity matrix $\Xi_{\mathbf{W}}$. We also show that the reduced row echelon of $\mathbf{W}$, singular value decomposition of $\mathbf{W}$, and skeleton decomposition of $\mathbf{W}$ can be used to generate similarity matrices for $\mathbf{W}$.

Assumptions 1 *(Assumptions on data).* In the remaining of this manuscript, we will assume that $\mathcal{U} = \bigcup_{i=1}^{M} S_i$ is a nonlinear set consisting of the union of non-trivial, independent subspaces $\{S_i\}_{i=1}^{M}$ of $\mathbb{R}^D$, with corresponding dimensions $\{d_i\}_{i=1}^{M}$. We will assume that the data matrix $\mathbf{W} = [w_1 \cdots w_N] \in \mathbb{R}^{D \times N}$ has column vectors that are drawn from $\mathcal{U}$, and that the data is drawn from each subspace S_i and that it is generic for it.

Theorem 2. *Let $\mathbf{W} = [w_1 \cdots w_N] \in \mathbb{R}^{D\times N}$ be a matrix whose columns are drawn from $\mathcal{U}$ be factorized as $\mathbf{W} = BP$ where B is a basis for the column space of $\mathbf{W}$ and that the columns of B lie in $\mathcal{U}$. If $Q = abs(P^TP)$ and $d_{max} = \max\{d_i\}_{i=1}^M$, then, $\Xi_{\mathbf{W}} = Q^{d_{max}}$ is a similarity matrix for $\mathbf{W}$.*

The proof of Theorem 2 is based on the following three Lemmas.

Lemma 1. *Assume $\mathbf{W} \in \mathbb{R}^{D\times N}$ is a matrix of generic data drawn from a single subspace S of dimension $r < N$. Consider the factorization $\mathbf{W} = BP$, where the columns of $B \in \mathbb{R}^{D\times r}$ form a basis for the range of $\mathbf{W}$. Let $P = [p_1 \cdots p_N] \in \mathbb{R}^{r\times N}$, where p_i's are columns of P. Then, the columns of P are generic for $\mathbb{R}^r$, i.e., any r number of columns of P is a basis for $\mathbb{R}^r$.*

Proof. Pick any r columns $\{p_{i_1}, \dots, p_{i_r}\}$ of P and assume that $c_1p_{i_1} + c_2p_{i_2} + \dots + c_rp_{i_r} = 0$, for constants $\{c_1.\cdots, c_r\}$. Then, $B(c_1p_{i_1} + c_2p_{i_2} + \dots + c_rp_{i_r}) = 0$. Hence $c_1w_{i_1} + c_2w_{i_2} + \dots + c_rw_{i_r} = 0$. Since $\mathbf{W}$ is generic, we get that $c_i = 0$ for all $i \in \{1, \dots, r\}$. Therefore, $\{p_{i_1}, \dots, p_{i_r}\}$ is linearly independent. □

Lemma 2. *Let $\{p_1, \dots, p_{r+1}\}$ be a set of generic vectors that represent data from a subspace S of dimension $r \geq 1$, and let G be the graph whose nodes are indexed by p_i and whose edges are those p_ip_j such that $Q(i,j) > 0$, where Q is defined as in Theorem 2. Then G is a connected graph.*

Proof of lemma. Let C be the vertices of a non-empty connected component of G. We will show that C^c is empty. Assume that $|C^c| = k > 0$. Then, since C is non-empty, we have $|C^c| \leq r$. Hence C^c is a linearly independent set due to data being generic. Similarly, since $k > 0$, $|C| \leq r$, and therefore C is also linearly independent set. But by construction of G, $\langle p, q\rangle = 0$ for any $p \in C^c$ and $q \in C$. Therefore, $C^c \perp C$. Thus, the set $\{p_1, \dots, p_{r+1}\}$ is linearly independent contradicting the assumption that $\dim(S) = r$. Hence $|C^c| = 0$ and G is connected. □

Lemma 3. *Let $V = \{p_1, \dots, p_N\}$ be a set of generic vectors that represent data from a subspace S of dimension r and $N > r \geq 1$. Let G be the graph whose nodes are indexed by p_i and whose edges are those p_ip_j such that $Q(i,j) > 0$. Then G is connected. Moreover the diameter of G is at most r.*

Proof of lemma. Let $p, q \in V$ and assume that $p \neq q$. Since $N > r$, and V is generic, there exist $r-1$ distinct vectors of $E \subset V$ such that $\{p, q\} \cup E$ form distinct $r+1$ vectors that are generic. Thus, by Lemma 2, their graph is connected. Hence there is a path between p and q of length at most r. Since p, q are arbitrary, the proof is complete. □

Proof of Theorem 2. We simply note that if p and q are representations of data from subspaces S_i and S_j with $i \neq j$, $\langle p, q\rangle = 0$. Hence there is no edge between p and any q. Therefore, by the previous Lemmas, the graph G consists of exactly M disconnected components such that each connected component j corresponds to the vectors that belong the same subspace S_j.

To prove the theorem, we simply note that the diameter of each component is at most d_{max} and $Q^{d_{max}}(i,j) > 0$ by Corollary 1. □

Now, we will discuss a special case in which the binary version of P^TP is a pre-similarity matrix. Corollary 2 has the same conditions as Theorem 2 with the exception that Q is not the absolute value version of P^TP, but it is the binary version of P^TP.

Remark 1. As a direct consequence of Lemma 3, if $V = \{p_1, \dots, p_N\}$ is a set of generic vectors that represent data from S of dimension r and $N > r \geq 1$ and G is the graph whose nodes are indexed by p_i and whose edges are those p_ip_j such that $Q(i,j) = 1$ when $P^TP(i,j) \neq 0$ and $Q(i,j) = 0$ otherwise, then G is connected.

As a consequence of Theorem 2 and Remark 1, we get the following corollary.

Corollary 2. *Assume the same conditions as those of Theorem 2. Let* $Q = bin(P^T P)$ *and* $d_{max} = \max\{d_i\}_{i=1}^M$. *Then,* $\Xi_{\mathbf{W}} = Q^{d_{max}}$ *is a similarity matrix for* $\mathbf{W}$.

3.1. Shape Interaction Matrix – revisited

We begin this section by a definition of the skinny SVD.

Definition 10 *(Skinny SVD).* Let $\mathbf{W}$ be an $D \times N$ matrix with $\mathrm{rank}(\mathbf{W}) = r$, and let the SVD of $\mathbf{W}$ be $U\Sigma V^T$. The skinny SVD of $\mathbf{W}$ is the decomposition $\mathbf{W} = U_r \Sigma_r V_r^T$, where U_r is the first r columns of U, V_r^T is the first r rows of V^T, while Σ_r is the upper left $r \times r$ sub-matrix of Σ.

In [32], it is stated that a similarity matrix (called Shape Interaction Matrix – SIM) can be constructed using the singular value decomposition (SVD). Let the skinny SVD of $\mathbf{W}$ be $U\Sigma V^T$. Then $\mathrm{SIM}(\mathbf{W})$ is defined by

$$\mathrm{SIM}(\mathbf{W}) = VV^T. \tag{3.1}$$

It is also stated that $\mathrm{SIM}(\mathbf{W})$ is block diagonal for sorted $\mathbf{W}$. A proper proof is not provided in [32]. However, we can show, that although $\mathrm{SIM}(\mathbf{W})$ is a similarity matrix most of the time, there are some cases in which it is not. In fact, VV^T is not even a pre-similarity matrix in such cases.

Consider the skinny SVD $\mathbf{W} = U\Sigma V^T$ of a data matrix $\mathbf{W}$. It is not even clear that $abs(VV^t)$ is a pre-similarity matrix, since the basis $U\Sigma$ for $\mathbf{W}$, does not necessarily come from the union of subspaces $\mathcal{U} = \{S_i\}_{i=1}^M$, and Theorem 2 does not apply directly. However, we can still show that $abs(VV^T)$ is in fact a pre-similarity matrix.

Theorem 3. *Let* $\mathbf{W} = [w_1 \cdots w_N] \in \mathbb{R}^{D\times N}$ *be a matrix whose columns are drawn from a union of subspaces* $\mathcal{U}$ *as in Assumptions 1. Let the skinny SVD of* $\mathbf{W}$ *be given by* $\mathbf{W} = U\Sigma V^T$, *and define* $Q = abs(VV^T)$. *Then,* $\Xi_{\mathbf{W}} = Q^{d_{max}}$ *is a similarity matrix for* $\mathbf{W}$, *where* $d_{max} = \max\{d_i\}_{i=1}^M$.

Proof. Let $\mathcal{U} = \{S_i\}_{i=1}^M$, and n_i be the number of data points that come from subspaces S_i with dimensions d_i ($\sum_{i=1}^M n_i = N$). Assume $\mathrm{rank}(W) = r$ (i.e., $\sum_{i=1}^M d_i = r$). Without loss of generality, assume that the data is sorted, that is

$$\mathbf{W} = [\,\mathbf{W}_1 \quad \mathbf{W}_2 \ldots \mathbf{W}_M\,]$$

where $\mathbf{W}_i$'s have columns drawn from S_i's. Consider the following skinny SVDs:

$$\begin{aligned} \mathbf{W} &= U\Sigma V^T \\ \mathbf{W}_i &= U_i \Sigma_i V_i^T \end{aligned} \tag{3.2}$$

where $\mathbf{W} \in \mathbb{R}^{D\times N}$, $U \in \mathbb{R}^{D\times r}$, $\Sigma \in \mathbb{R}^{r\times r}$, $V^T \in \mathbb{R}^{r\times N}$, $\mathbf{W}_i \in \mathbb{R}^{D\times n_i}$, $U_i \in \mathbb{R}^{D\times d_i}$, $\Sigma_i \in \mathbb{R}^{d_i\times d_i}$, $V_i^T \in \mathbb{R}^{d_i\times n_i}$. Then,

$$\mathbf{W} = [\,\mathbf{W}_1 \quad \mathbf{W}_2 \ldots \mathbf{W}_M\,] = \begin{bmatrix} U_1\Sigma_1 V_1^T & U_2\Sigma_2 V_2^T & \ldots & U_M\Sigma_M V_M^T \end{bmatrix}$$

This is equivalent to the following factorization:

$$\mathbf{W} = [U_1\Sigma_1 \quad U_2\Sigma_2 \quad \dots \quad U_M\Sigma_M] \left[\mathrm{diag}(V_1^T, V_2^T, \dots, V_M^T)\right], \tag{3.3}$$

where $\left[\mathrm{diag}(V_1^T, V_2^T, \dots, V_M^T)\right]$ is the block diagonal matrix whose blocks are V_i^T, $i = 1, \dots, M$. Since the columns of $U_i\Sigma_i$'s are bases for S_i's, and since $U_i\Sigma_i$'s are obtained from the skinny SVDs of $\mathbf{W}_i$'s, it follows that the columns of $[U_1\Sigma_1 \quad \dots \quad U_M\Sigma_M]$ live in $\mathcal{U} = \bigcup_{i=1}^M S_i$ and they form a basis for the column space of $\mathbf{W}$ (by the independence assumption of the subspace S_is). Thus, by Theorem 2, we get $B = [U_1\Sigma_1 \quad U_2\Sigma_2 \quad \dots \quad U_M\Sigma_M]$, $P = \left[\mathrm{diag}(V_1^T, V_2^T, \dots, V_M^T)\right]$, and a block diagonal matrix $P^TP = \left[\mathrm{diag}(V_1V_1^T, V_2V_2^T, \dots, V_MV_M^T)\right]$ whose absolute $\Pi_W = abs(P^TP)$ value is a pre-similarity matrix. Although $\Pi_W = abs(P^TP)$ is a pre-similarity matrix, it cannot be computed from $\mathbf{W}$ since we do not know n_i a priori, even if $\mathbf{W} = [\mathbf{W}_1 \quad \mathbf{W}_2 \dots \mathbf{W}_M]$ is sorted. The dimensions of V_i^T is $d_i \times n_i$. Thus, the block diagonal matrix P has dimensions $r \times N$. Moreover, the matrix V in the skinny SVD of $\mathbf{W}$ has dimensions $N \times r$. What is remarkable is that $VV^T = P^TP$ and therefore VV^T is block diagonal and that $abs(VV^T)$ is a pre-similarity. Moreover, VV^T can be computed from $\mathbf{W}$ directly since all we need is to computed the skinny SVD of $\mathbf{W}$ to get V. Thus, to finish the proof of the theorem, we only need to prove the following claim.

Claim 1. *$VV^T = P^TP$.*

Proof of claim. Since $U\Sigma$ is a basis for $\mathbf{W}$ and the columns of $U_i\Sigma_i$ lie in S_i, we can find an $A_i \in \mathbb{R}^{r\times d_i}$ such that $U_i\Sigma_i = U\Sigma A_i$. Therefore, using Equation (3.3), we get

$$\mathbf{W} = U\Sigma [A_1 \quad A_2 \quad \dots \quad A_M] P = U\Sigma AP \tag{3.4}$$

where $P = \left[\mathrm{diag}(V_1^T, V_2^T, \dots, V_M^T)\right]$ and $A = [A_1 \quad A_2 \quad \dots \quad A_M]$. Since $U\Sigma$ is full-rank, using Equation (3.2), we have

$$V^T = AP \quad \text{and} \quad V = P^TA^T. \tag{3.5}$$

Since V is computed from the skinny SVD of $\mathbf{W}$, it follows that $V^TV = I_r$ where I_r is the $r \times r$ identity matrix. Moreover, since the matrices V_is in the blocks of P are computed from the skinny SVD of $\mathbf{W}_i$s it follows that $PP^T = I_r$ where I_r is the $r \times r$ identity matrix. Therefore, $I_r = V^TV = APP^TA^T = AA^T$. This implies that

$$AA^T = I_r$$

and therefore $A^T = A^{-1}$, hence $AA^T = A^TA = I_r$. As a result, we have

$$VV^T = P^TA^TAP = P^TP, \tag{3.6}$$

which finished the proof of the claim. $\square$

Thus the proof of the theorem is complete. $\square$

Although VV^T is often used as a similarity matrix in applications, the following example shows that VV^T does not have to be a similarity or even a pre-similarity matrix, in general.

Example 1. Let the data come from union of two subspaces S_1 (with canonical basis $\{e_1, e_2\}$) and S_2 (with canonical basis $\{e_3\}$). The first 4 columns are from S_1 and the last 2 columns are from S_2.

$$\mathbf{W} = \begin{bmatrix} 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 2 \end{bmatrix}$$

A simple calculation shows that the Shape Interaction Matrix $SIM(\mathbf{W})$ for $\mathbf{W}$ has two blocks and one of the blocks contains zero elements. Since V^TV is the identity matrix, $(VV^T)^n = VV^T$ for any $n > 0$. Hence, $SIM(\mathbf{W})$ is neither a similarity nor a pre similarity matrix. However, $abs(VV^T)$ is a pre-similarity matrix and $(abs(VV^T))^2$ is a similarity matrix in this particular case where $d_{max} = 2$.

However, similarly to the well-known fact that data that is drawn uniformly at random from a subspace is always generic except for a set measure of zero. It can also be shown that if data is randomly drawn from a union of subspaces $\mathbf{W}$, then VV^T will be a similarity matrix except for a set of measure of zero, of which the matrix above is an example.

Lemma 4. *Let S_l be a d_l dimensional subspace of $\mathbb{R}^D$. Then, any $n \leq d_l$ number of data points picked uniformly at random from S_l are linearly independent with probability 1.*

Proof. We will use induction over n. Assume the statement is true for $\{w_1, w_2, \ldots, w_n\}$ and consider $\{w_1, w_2, \ldots, w_n, w_{n+1}\}$. Since $\{w_1, w_2, \ldots, w_n\}$ are linearly independent with probability 1 by assumption, normalized $\{\widetilde{w_1}, \widetilde{w_2}, \ldots, \widetilde{w_n}\}$ lies on a n-dimensional ball in $\mathbb{R}^n$ with probability 1. Now, assume $\{w_1, w_2, \ldots, w_n, w_{n+1}\}$ are not linearly independent. Then, $\{\widetilde{w_1}, \widetilde{w_2}, \ldots, \widetilde{w_n}, \widetilde{w_{n+1}}\}$ still lies on the same n-dimensional ball but this time in $\mathbb{R}^{n+1}$. Such a probability is clearly 0. Finally, the statement of lemma is obviously true for $n = 1$. □

Theorem 4. *Let $\mathbf{W} = [w_1 \cdots w_N] \in \mathbb{R}^{D\times N}$ be a matrix whose columns are drawn from a union of subspaces $\mathcal{U}$ as in Assumptions 1. Assume the normalized data uniformly distributed from each subspace and that it is generic. Let the skinny SVD of $\mathbf{W}$ be given as $\mathbf{W} = U_s\Sigma_sV_s^T$. Then, $Q = V_sV_s^T$ is a similarity matrix for $\mathbf{W}$ except for a set of measure of zero.*

Proof. Let $d = \sum_i d_i$ be the rank of $\mathbf{W}$, and $U_s\Sigma_sV_s^T$ the skinny SVD's of the data matrix $\mathbf{W}$. Then the set $V_sV_s^T$ form an Nd-dimensional manifold. Let $U\Sigma V^T$ and $U_s\Sigma_sV_s^T$ be the regular and the skinny SVD's of the data matrix W, respectively.

$$V = [\, v_1 \quad \ldots \quad v_N \,] \in \mathbb{R}^{N\times N}$$
$$V_s = [\, v_1 \quad \ldots \quad v_d \,] \in \mathbb{R}^{N\times d}$$

Then,

$$VV^T(i,j) = \sum_{k=1}^{D} v_k(i)v_k(j)$$
$$V_sV_s^T(i,j) = \sum_{k=1}^{d} v_k(i)v_k(j)$$

Let each entry $V(i,j)$ be a variable. Then we have a total of N^2 variables. Observe that, VV^T generates $(N^2+N)/2$ equations. Therefore, there are $(N^2-N)/2$ free variables. Since $|V(i,j)| \leq 1$, it means that the solution space is a compact manifold of dimension $(N^2-N)/2$ of $\mathbb{R}^{N^2}$. Let's call this manifold M_V.

We will show that when $VV^T(i,j) = 0$, then $V_sV_s^T(i,j) \neq 0$ with probability 1. Assume $V_sV_s^T(i,j) = 0$, then the number of free variables is decreased by one. The solution to $VV^T = I$ and $V_sV_s^T(i,j) = 0$ lies in a compact manifold in $\mathbb{R}^{N^2}$ with dimension $(N^2-N)/2 - 1$. Let's call this manifold $M_{V'}$. We further know that $M_{V'} \subset M_V$. Since the probability of having such a submanifold $M_{V'}$ in M_V is zero, probability of having $V_sV_s^T(i,j) = 0$ is also 0. Note that the submanifold $M_{V'}$ will be even lower dimensional if more entries of $V_sV_s^T$ is assumed to be 0. Also, since we are dealing with finite number of entries, we can conclude that $V_s V_s^T(i,j) = 0$ only for measure zero. Finally, note that if the columns of W are picked in such a way that the angle between any two columns in uniformly at random, then entries of V will also be uniformly at random. □

Proposition 1. *Let $\mathbf{W} = [w_1 \cdots w_N] \in \mathbb{R}^{D\times N}$ be a matrix whose columns are drawn from a union of subspaces $\mathcal{U}$ as in Assumptions 1. Let the skinny SVD of $\mathbf{W}$ be given by $\mathbf{W} = U\Sigma V^T$. Then, VV^T is a similarity matrix, if and only if it is a pre-similarity matrix.*

Proof. Assume VV^T is a pre-similarity matrix. Without loss of generality, assume that data comes from a single subspace S of dimension d. Then, if VV^T has a 0-entry, then some of the remaining entries must be negative. Otherwise, by Theorem 2, $Q = VV^T$ would be a pre-similarity matrix and $(VV^T)^d = VV^T$ would be a similarity matrix by Corollary 2, and therefore Q has all strictly positive entries, contradicting the fact that VV^T has a 0-entry. Converse is obvious. □

3.2. Similarity matrix via reduced row Echelon form

A subspace segmentation algorithm for independent subspaces based on reduced row echelon form of the data matrix $\mathbf{W}$ was developed in [33]. A performance analysis based on subspace angles and noise was provided. We now use Theorem 2 to provide another view and proof that a similarity matrix $\Xi_{\mathbf{W}}$ can be obtained from reduced row echelon form of $\mathbf{W}$.

The reduced row echelon form of a matrix $\mathbf{W}$ of rank r is obtained by the three elementary row operations on $\mathbf{W}$ to get a matrix of the form rref($\mathbf{W}$)

$$\operatorname{rref}(\mathbf{W}) = \begin{bmatrix} R \\ 0 \end{bmatrix}, \tag{3.7}$$

where the r rows of R are linearly independent. We can obtain a similarity matrix from this decomposition as follows.

Theorem 5. *Let $\mathbf{W} = [w_1 \cdots w_N] \in \mathbb{R}^{D\times N}$ be a matrix whose columns are drawn from a union of subspaces $\mathcal{U}$ as in Assumptions 1. Let R be the matrix in the reduced row echelon form of $\mathbf{W}$ (3.7), and let Q be the binary or absolute value version of $R^T R$, and $d_{max} = \max\{d_i\}_{i=1}^M$. Then, $\Xi_{\mathbf{W}} = Q^{d_{max}}$ is a similarity matrix.*

Proof. In [33], it is shown that $\mathbf{W}$ can be factored as a product of a matrix that is a basis for the column space of $\mathbf{W}$ and another matrix that is concatenated reduced row echelon form of $\mathbf{W}$.

$$\mathbf{W} = BR \tag{3.8}$$

where the columns of B are the columns of $\mathbf{W}$ corresponding to the pivots of rref($\mathbf{W}$) is as in (3.7). Then, the proof follows from Theorem 2. □

Observe that if $\mathbf{W}$ is sorted such that the columns drawn from the same subspace are adjacent, then $\Xi_{\mathbf{W}}$ is block diagonal. This is a consequence of the fact that reordering columns of $\mathbf{W}$ will be reflected in the same order in the reduced row echelon form of $\mathbf{W}$.

3.3. Similarity matrix via skeleton decomposition

Now, we will show that a computationally efficient similarity matrix can be generated using a skeleton decomposition of $\mathbf{W}$. For any rank-r matrix Z, we can find a rank-r square sub-matrix A of Z. Then, Z can be factorized as $Z = CA^{-1}R$, where C and R are the column and row restrictions of Z from A.

Theorem 6. *Let $\mathbf{W} = [w_1 \cdots w_N] \in \mathbb{R}^{D\times N}$ be a matrix whose columns are drawn from a union of subspaces $\mathcal{U}$ as in Assumptions 1. Let $\operatorname{rank}(\mathbf{W}) = r$. Assume A is any rank-r square sub-matrix of $\mathbf{W}$. Further, let $C \in \mathbb{R}^{D\times r}$ and $R \in \mathbb{R}^{r\times D}$ be the column and row restrictions of $\mathbf{W}$ corresponding to A, respectively. Let $P = A^{-1}R$, and Q be the binary or absolute value version of P^TP. Then, $\Xi_{\mathbf{W}} = Q^{d_{max}}$ is a similarity matrix for $\mathbf{W}$, where $d_{max} = \max\{d_i\}_{i=1}^M$.*

Proof. Without loss of generality, assume that $\mathbf{W}$ is sorted as follows:

$$\mathbf{W} = [\,\mathbf{W}_1 \quad \mathbf{W}_2 \ldots \mathbf{W}_M\,]$$

where $\mathbf{W}_i$'s have columns drawn from S_i's. By the independence of subspace and the generic data assumptions, $\operatorname{rank}(\mathbf{W}_i) = d_i$ and $\operatorname{rank}(\mathbf{W}) = r = d_1 + \ldots + d_M$. Let A_i be any sub-matrix of $\mathbf{W}_i$. Then, $\operatorname{rank}(A_i) \le d_i$. Thus, any rank-$r$ sub-matrix A of $\mathbf{W}$ must include exactly d_i columns from each $\mathbf{W}_i$. This implies that the corresponding column restriction of A is a basis for $\mathbf{W}$ and also it comes from $\mathcal{U}$. The proof then directly follows from Theorem 2. □

Algorithm 1 provides a simple algorithm for subspace segmentation. Some of the subspace segmentation problems involve subspaces of low dimensions, e.g., independently moving rigid body motion segmentation problem includes independent subspaces of dimension 4. In such cases, Algorithm 1 is efficient. If the subspaces are only approximations, e.g., face/facial expression recognition problem includes approximated subspaces of dimension 9. In such cases, Algorithm 1 may not be an ideal choice. Note that the dimensions $\{d_1, \ldots, d_M\}$ can be determined after the columns of $\mathbf{W}$ are clustered using $\Xi_{\mathbf{W}}$ (i.e., $\operatorname{rank}(\mathbf{W}_i) = d_i$). For all practical purposes we can set $d_{max} = \operatorname{rank}(\mathbf{W})$ in Algorithm 1 if d_{max} is not known a priori.

Algorithm 1: Construction of similarity matrix using skeleton decomposition.

Data: A data matrix $\mathbf{W} = [w_1 \cdots w_N] \in \mathbb{R}^{D\times N}$ whose columns are drawn from union $\mathcal{U}$ of M independent subspaces with $d_{max} = \max\{d_i\}_{i=1}^M$

Result: Computation of a similarity matrix $\Xi_{\mathbf{W}}$

$r = \operatorname{rank}(\mathbf{W})$
do
  Pick two index vectors v_{row} and v_{col} of size r from $\{1, \ldots, N\}$ (randomly or systematically)
  Construct A such that $A(i,j) = \mathbf{W}(v_{row}(i), v_{col}(j))$, where i and j go from 1 to r
while $rank(A) \neq r$;
Construct a matrix R such that $R(i,j) = \mathbf{W}(v_{row}(i), j)$ with $i \in \{1, \ldots, r\}$ and $j \in \{1, \ldots, N\}$
Construct matrix $P = A^{-1}R$
Construct the binary (or absolute value) version Q of P^TP
Construct matrix $\Xi_{\mathbf{W}} = Q^{d_{max}}$

4. Conclusion

In this research, we developed a framework for finding similarity matrices for data that comes from a union of independent subspaces. We first showed that, the reduced row echelon form of a data matrix can be used to form a similarity matrix. Then, we proved that the shape interaction matrix (widely used in literature) is essentially a special case of this framework. Finally, we proposed a new method to compute a similarity matrix using the skeleton decomposition.

References

[1] J. Cahill, P.G. Casazza, S. Li, Non-orthogonal fusion frames and the sparsity of fusion frame operators, J. Fourier Anal. Appl. 18 (2) (2012) 287–308.
[2] P.G. Casazza, G. Kutyniok, D. Speegle, J.C. Tremain, A decomposition theorem for frames and the Feichtinger conjecture, Proc. Amer. Math. Soc. 136 (6) (2008) 2043–2053.
[3] G. Lerman, T. Zhang, l_p-recovery of the most significant subspace among multiple subspaces with outliers, Constr. Approx. 40 (3) (2014) 329–385.
[4] Y.M. Lu, M.N. Do, A theory for sampling signals from a union of subspaces, IEEE Trans. Signal Process. 56 (6) (2015) 2334–2345.
[5] R. Basri, D.W. Jacobs, Lambertian reflectance and linear subspaces, IEEE Trans. Pattern Anal. Mach. Intell. 25 (2003) 218–233.
[6] J. Ho, M. Yang, J. Lim, D. Kriegman, Clustering appearances of objects under varying illumination conditions, in: IEEE Conference on Computer Vision and Pattern Recognition, 2003, pp. 11–18.
[7] K. Kanatani, C. Matsunaga, Estimating the number of independent motions for multibody motion segmentation, in: 5th Asian Conference on Computer Vision, 2002, pp. 7–9.
[8] A. Aldroubi, K. Zaringhalam, Nonlinear least squares in $\mathbb{R}^n$, Acta Appl. Math. 107 (1–3) (2009) 325–337.
[9] G. Liu, Z. Lin, Y. Yu, Robust subspace segmentation by low-rank representation, in: International Conference on Machine Learning, 2010, pp. 663–670.
[10] G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu, Y. Ma, Robust recovery of subspace structures by low-rank representation, arXiv.org, http://arxiv.org/abs/1010.2955.
[11] S. Wei, Z. Lin, Analysis and improvement of low rank representation for subspace segmentation, arXiv.org, http://arxiv.org/abs/1107.1561.
[12] Y.C. Eldar, M. Mishali, Robust recovery of signals from a structured union of subspaces, IEEE Trans. Inform. Theory 55 (11) (2009) 5302–5316.
[13] E. Elhamifar, R. Vidal, Sparse subspace clustering, in: IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 2790–2797.
[14] E. Elhamifar, R. Vidal, Clustering disjoint subspaces via sparse representation, in: IEEE International Conference on Acoustics, Speech, and Signal Processing, 2010, pp. 1926–1929.
[15] R. Vidal, Y. Ma, S. Sastry, Generalized principal component analysis (GPCA), IEEE Trans. Pattern Anal. Mach. Intell. 27 (12) (2005) 1945–1959.
[16] Y. Ma, A.Y. Yang, H. Derksen, R. Fossum, Estimation of subspace arrangements with applications in modeling and segmenting mixed data, SIAM Rev. 50 (1) (2008) 1–46.
[17] M.C. Tsakiris, R. Vidal, Filtrated algebraic subspace clustering, SIAM J. Imaging Sci. 10 (1) (2017) 372–415.
[18] K. Kanatani, Y. Sugaya, Multi-stage optimization for multi-body motion segmentation, in: IEICE Trans. Inf. and Syst, 2003, pp. 335–349.
[19] A. Aldroubi, C. Cabrelli, U. Molter, Optimal non-linear models for sparsity and sampling, J. Fourier Anal. Appl. 14 (5) (2009) 793–812.
[20] P. Tseng, Nearest q-flat to m points, J. Optim. Theory Appl. 105 (1) (2000) 249–252.
[21] M. Fischler, R. Bolles, Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography, Commun. ACM 24 (6) (1981) 381–395.
[22] N. Silva, J. Costeira, Subspace segmentation with outliers: a grassmannian approach to the maximum consensus subspace, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2008, pp. 1–6.
[23] U. von Luxburg, A tutorial on spectral clustering, Stat. Comput. 17 (4) (2007) 395–416.
[24] G. Chen, G. Lerman, Spectral curvature clustering (SCC), Int. J. Comput. Vis. 81 (2009) 317–330.
[25] F. Lauer, C. Schnorr, Spectral clustering of linear subspaces for motion segmentation, in: IEEE International Conference on Computer Vision, 2009, pp. 678–685.
[26] J. Yan, M. Pollefeys, A general framework for motion segmentation: independent, articulated, rigid, non-rigid, degenerate and nondegenerate, in: 9th European Conference on Computer Vision, 2006, pp. 94–106.
[27] A. Goh, R. Vidal, Segmenting motions of different types by unsupervised manifold clustering, in: IEEE Conference on Computer Vision and Pattern Recognition, 2007, CVPR '07, 2007, pp. 1–6.
[28] R. Vidal, A tutorial on subspace clustering, IEEE Signal Process. Mag. 28 (2010) 52–68.
[29] G. Chen, G. Lerman, Foundations of a multi-way spectral clustering framework for hybrid linear modeling, Found. Comput. Math. 9 (5) (2009) 517–558.
[30] J. Irion, N. Saito, Applied and computational harmonic analysis on graphs and networks, Proc. SPIE 9597, Wavelets and Sparsity XVI.
[31] C. You, D.P. Robinson, R. Vidal, Scalable sparse subspace clustering by orthogonal matching pursuit, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27–30, 2016, 2016, pp. 3918–3927.
[32] J. Costeira, T. Kanade, A multibody factorization method for independently moving objects, Int. J. Comput. Vis. 29 (3) (1998) 159–179.
[33] A. Aldroubi, A. Sekmen, Reduced row echelon form and non-linear approximation for subspace segmentation and high-dimensional data clustering, Appl. Comput. Harmon. Anal. 37 (2) (2014) 271–287.
[34] B. Bollob's, Modern Graph Theory, Graduate Texts in Mathematics, Springer, New York, Berlin, Paris, 1998.