Appl. Comput. Harmon. Anal. 37 (2014) 271–287

Contents lists available at ScienceDirect

Applied and Computational Harmonic Analysis

www.elsevier.com/locate/acha

Reduced row echelon form and non-linear approximation for subspace segmentation and high-dimensional data clustering

Akram Aldroubi [a,*], Ali Sekmen [b]

[a] Department of Mathematics, Vanderbilt University, Nashville, TN 37212, United States
[b] Department of Computer Science, Tennessee State University, Nashville, TN 37209, United States

ARTICLE INFO

Article history:
Received 10 May 2012
Received in revised form 9 December 2013
Accepted 15 December 2013
Available online 17 December 2013
Communicated by Jared Tanner

Keywords:
Subspace segmentation
Data clustering
High dimensional data

ABSTRACT

Given a set of data $\mathbf{W} = \{w_1, \ldots, w_N\} \in \mathbb{R}^D$ drawn from a union of subspaces, we focus on determining a nonlinear model of the form $\mathscr{U} = \bigcup_{i\in I} S_i$, where $\{S_i \subset \mathbb{R}^D\}_{i\in I}$ is a set of subspaces, that is nearest to $\mathbf{W}$. The model is then used to classify $\mathbf{W}$ into clusters. Our approach is based on the binary reduced row echelon form of data matrix, combined with an iterative scheme based on a non-linear approximation method. We prove that, in absence of noise, our approach can find the number of subspaces, their dimensions, and an orthonormal basis for each subspace S_i. We provide a comprehensive analysis of our theory and determine its limitations and strengths in presence of outliers and noise.

1. Introduction

In many engineering and mathematics applications, data lives in a union of low dimensional subspaces [1–6]. For instance, the set of all two dimensional images of a given face i, 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 $\mathbb{R}^D$ [4,7,8]. A set of such images from different faces is then a union $\mathscr{U} = \bigcup_{i\in I} S_i$. Similar nonlinear models arise in sampling theory where $\mathbb{R}^D$ is replaced by an infinite dimensional Hilbert space $\mathscr{H}$, e.g., $L^2(\mathbb{R}^D)$ [1,9–12].

The goal of subspace clustering is to identify all of the subspaces that a set of data $\mathbf{W} = \{w_1, \ldots, w_N\} \in \mathbb{R}^D$ is drawn from and assign each data point w_i to the subspace it belongs to. The number of subspaces, their dimensions, and a basis for each subspace are to be determined even in presence of noise, missing data, and outliers. The subspace clustering or segmentation problem can be stated as follows: Let $\mathscr{U} = \bigcup_{i=1}^M S_i$ where $\{S_i \subset \mathscr{B}\}_{i=1}^M$ is a set of subspaces of a Hilbert space or Banach space $\mathscr{B}$. Let $\mathbf{W} = \{w_j \in \mathscr{B}\}_{j=1}^N$ be a set of data points drawn from $\mathscr{U}$. Then,

1. determine the number of subspaces M,
2. determine the set of dimensions $\{d_i\}_{i=1}^M$,

* Corresponding author.

1063-5203/$ – see front matter
http://dx.doi.org/10.1016/j.acha.2013.12.001

3. find an orthonormal basis for each subspace S_i,
4. collect the data points belonging to the same subspace into the same cluster.

Note that often the data may be corrupted by noise, may have outliers or the data may not be complete, e.g., there may be missing data points. In some subspace clustering problems, the number M of subspaces or the dimensions of the subspaces $\{d_i\}_{i=1}^M$ are known. A number of approaches have been devised to solve the problem above or some of its special cases. They are based on sparsity methods [13–18], algebraic methods [19,20], iterative and statistical methods [2,3,10,21–24], and spectral clustering methods [14,15,25–32].

1.1. Paper contributions

- We develop an algebraic method for solving the general subspace segmentation problem for noiseless data. For the case where all the subspaces are four dimensional, Gear observed, without proof, that the reduced echelon form can be used to segment motions in videos [33]. In this paper, we develop this idea and prove that the reduced row echelon form can completely solve the subspace segmentation problem in its most general version. This is the content of Theorem 3.7 in Section 3.1.
- For noisy data, the reduced echelon form method does not work, and a thresholding must be applied. However, the effect of the noise on the reduced echelon form method depends on the noise level and the relative positions of the subspaces. This dependence is analyzed in Section 3.2 and is contained in Theorems 3.9 and 3.11.
- When the dimensions of the subspaces is equal and known, we relate the subspace segmentation problem to the non-linear approximation problem (Problem 1). The existence of a solution as well as an iterative search algorithm for finding the solution is presented in Theorem 2.1. This algorithm works well with noisy data but requires a good initial condition to locate the global minimum.
- The reduced echelon form together with the iterative search algorithm are combined together: A thresholded reduced echelon form algorithm provides the initial condition to the iterative search algorithm. This is done in Section 4.
- In Section 5, the algorithms are tested on synthetic and real data to evaluate the performance and limitations of the methods.

2. Non-linear approximation formulation of subspace segmentation

When M is known, the subspace segmentation problem, for both the finite and infinite dimensional space cases, can be formulated as follows:

Let $\mathscr{B}$ be a Banach space, $\mathbf{W} = \{w_1, \ldots, w_N\}$ a finite set of vectors in $\mathscr{B}$ that correspond to observed data. For $i = 1, \ldots, M$, let $\mathscr{C} = C_1 \times C_2 \times \cdots \times C_M$ be the Cartesian product of M families C_i of closed subspaces of $\mathscr{B}$ each containing the trivial subspace $\{0\}$. Thus, an element $\mathbf{S} \in \mathscr{C}$ is a sequence $\{S_1, \ldots, S_M\}$ of M subspaces of $\mathscr{B}$ with $S_i \in C_i$. For example, when each C_i is the family of all subspaces of dimensions less than or equal to d in the ambient space $\mathscr{B} = \mathbb{R}^D$, then an element $\mathbf{S} \in \mathscr{C}$ is a set of M subspaces $S_i \subset \mathbb{R}^D$, with dimensions $\dim S_i \leqslant d$. Another example is the infinite dimensional case in which $\mathscr{B} = L^2(\mathbb{R})$ and each C_i is a family of closed, shift-invariant subspaces of $L^2(\mathbb{R})$ that are generated by at most $r < \infty$ generators. For example if $r = 1$, $M = 2$, an element $\mathbf{S} \in \mathscr{C}$ may be the subspace S_1 of all bandlimited functions (generated by integer shifts of the generator function $\operatorname{sinc}(x) = \sin(x)/x$), and S_2 the shift invariant space generated by the B-spline functions β^n of degree n. In these cases the subspaces in $S_i \in C_i$ are also infinite dimensional subspaces of L^2.

Problem 1.

1. Given a finite set $\mathbf{W} \subset \mathscr{B}$, a fixed p with $0 < p \leqslant \infty$, and a fixed integer $M \geqslant 1$, find the infimum of the expression

$$e(\mathbf{W},\mathbf{S}) := \sum_{w\in\mathbf{W}} \min_{1\leqslant j\leqslant M} d^p(w,S_j),$$

over $\mathbf{S}=\{S_1,\dots,S_M\}\in\mathscr{C}$, and $d(x,y):=\|x-y\|_{\mathscr{B}}$.

2. Find a sequence of M-subspaces $\mathbf{S}^o=\{S_1^o,\dots,S_M^o\}\in\mathscr{C}$ (if it exists) such that

$$e\big(\mathbf{W},\mathbf{S}^o\big)=\inf\big\{e(\mathbf{W},\mathbf{S})\colon\ \mathbf{S}\in\mathscr{C}\big\}. \tag{1}$$

Given a family $\mathscr{C}$ of closed subspaces of $\mathscr{B}$, a solution to Problem 1 may not exists even in the simple case when $M=1$. For example, assume that $\mathscr{B}=R^2$ and $\mathscr{C}$ is the set of all lines through the origin except the line $x=0$. For this case, a minimizer may exist for certain distribution of data points but not for others. The existence of a solution here means that a minimizer exists for any distribution of any finite number of data points.

In the presence of outliers, it is shown that $p=1$ is a good choice [34] and a good choice for light-tailed noise is $p=2$. The necessary and sufficient conditions for the existence of a solution when $p=2$ and $\mathscr{B}$ is a Hilbert space can be found in [9].

Definition 1. For $0<p\leqslant\infty$, a set of closed subspaces C of a Banach space $\mathscr{B}$ has the Minimum Subspace Approximation Property p-(MSAP) if for every finite subset $\mathbf{W}\subset\mathscr{B}$ there exists an element $S\in C$ that minimizes the expression $e(\mathbf{W},S)=\sum_{w\in\mathbf{W}} d^p(w,S)$ over all $S\in C$.

The MSAP definition was proposed in [9] for finite and infinite dimensional Hilbert spaces (i.e., the 2-MSAP according to Definition 1). It turns out that although MSAP is formulated for approximating data by a single subspace, it is key for the existence of a solution to Problem 1. Necessary and sufficient conditions for the existence of a solution for Problem 1 and its relation to MSAP can be found in [9,35]. Under the assumption that each family of subspaces C_i satisfies p-(MSAP), Problem 1 has a minimizer:

Theorem 2.1. *If for each $i=1,\dots M$, C_i satisfies p-(MSAP), then Problem 1 has a minimizing set of subspaces for all finite sets of data.*

Proof. Let $\mathscr{P}(\mathbf{W})$ be the set of all partitions of $\mathbf{W}$ into M subsets, i.e., $P=\{\mathbf{W}_1,\dots,\mathbf{W}_M\}\in\mathscr{P}(\mathbf{W})$ if $\mathbf{W}=\bigcup_i\mathbf{W}_i$, and $\mathbf{W}_i\cap\mathbf{W}_j=\emptyset$. Let $P=\{\mathbf{W}_1,\dots,\mathbf{W}_M\}$ be a partition in $\mathscr{P}(\mathbf{W})$ (in our definition of partition, we allow one or more of the sets $\mathbf{W}_i$ to be empty). For each subset $\mathbf{W}_i$ in the partition P find the subspace $S_i^o(P)\in C_i$ that minimizes the expression $e(\mathbf{W}_i,S)=\sum_{w\in\mathbf{W}_i} d^p(w,S)$ over all $S\in C_i$. Let $m=\min\{\sum_{i=1}^M e(\mathbf{W}_i,S_i^o(P))\colon\ P\in\mathscr{P}(\mathbf{W})\}$, and denote by $P^o=\{\mathbf{W}_1^o,\dots,\mathbf{W}_M^o\}$ any partition for which $m=\sum_{i=1}^M e(\mathbf{W}_i^o,S_i^o(P^o))$. Then, for any $\mathbf{S}=\{S_1,\dots,S_M\}\in\mathscr{C}$ we have that

$$e(\mathbf{W},\mathbf{S})=\sum_{j=1}^M e(X_j,S_j)\geqslant\sum_{j=1}^M e\big(X_j,S_j^o(P_{\mathbf{S}})\big)\geqslant\sum_{j=1}^M e\big(\mathbf{W}_j^o,S_j^o\big(P^o\big)\big)=e\big(\mathbf{W},\mathbf{S}^o\big)$$

where $P_{\mathbf{S}}=\{X_1,\dots,X_M\}$ is any partition of $\mathbf{W}$ generated using $\mathbf{S}$ by

$$X_j=\big\{w\in\mathbf{W}\colon\ d(w,S_j)\leqslant d(w,S_i),\ i=1,\dots,M\big\}.$$

It follows that $e(\mathbf{W},\mathbf{S}^o)=m=\inf\{e(\mathbf{W},\mathbf{S})\colon\ \mathbf{S}\in\mathscr{C}\}$. □

The proof of Theorem 2.1 suggests a search algorithm (Algorithm 1 below) for the optimal solution $\mathbf{S}^o$. This algorithm is similar to k-subspaces algorithm [8]. Obviously, this solution can be obtained by Algorithm 1. This algorithm will work well if a good initial partition is chosen. Otherwise, the algorithm may terminate in a local minimum instead of the global minimum.

Algorithm 1 Search for optimal solution $\mathbf{S}^o$.

Pick any partition $P \in \mathscr{P}(\mathbf{W})$
For each subset $\mathbf{W}_i$ in the partition P find the subspace $S_i^o(P) \in C_i$ that minimizes the expression $e(\mathbf{W}_i, S) = \sum_{w\in\mathbf{W}_i} d^p(w, S)$
while $\sum_{i=1}^M e(\mathbf{W}_i, S_i^o(P)) > e(\mathbf{W}, \mathbf{S}^o(P))$ **do**
    for all i from 1 to M **do**
        Update $\mathbf{W}_i = \{w \in \mathbf{W}\colon\ d(w, S_i^o(P)) \leqslant d(w, S_k^o(P)),\ k = 1, \dots, M\}$
        Update $S_i^o(P) = \operatorname{argmin}_{S\in C_i} e(\mathbf{W}_i, S)$
    end for
    Update $P = \{\mathbf{W}_1, \dots, \mathbf{W}_M\}$
end while
$\mathbf{S}^o = \{S_1^o(P), \dots, S_M^o(P)\}$

3. Subspace segmentation in finite dimensional space

3.1. Subspace segmentation noiseless case

In this section we consider the problem in which a set of vectors $\mathbf{W} = \{w_1, \dots, w_N\}$ are drawn from a union $\mathscr{U} = \bigcup_{i\in I} S_i$ of M subspaces $S_i \in \mathbb{R}^D$ of dimension d_i. In order to find the M subspaces from the data set $\mathbf{W}$ it is clear that we need enough vectors $\mathbf{W} = \{w_1, \dots, w_N\}$. In particular for the problem of subspace segmentation, it is necessary that the set $\mathbf{W}$ can be partitioned into M sets $\mathbf{W} = \{\mathbf{W}_1, \dots, \mathbf{W}_M\}$ such that $\operatorname{span}\mathbf{W}_i = S_i$, $i = 1, \dots, M$. Thus, we need to assume that we have enough data for solving the problem. Thus, taking $\mathscr{B} = \mathbb{R}^D$, $\mathscr{C} = C_1 \times C_2 \dots C_n$ where each C_i is a family of subspaces of dimensions d, we assume that any $k \leqslant d$ vectors drawn from a subspace $S \in C_i$ of dimension d are linearly independent, and we make the following definition.

Definition 3.1. Let S be a linear subspace of $\mathbb{R}^D$ with dimension d. A set of data $\mathbf{W}$ drawn from $S \subset \mathbb{R}^D$ with dimension d is said to be *generic* if (i) $|\mathbf{W}| > d$, and (ii) every d vectors from $\mathbf{W}$ form a basis for S.

Another assumption that we will make is that the union of subspaces $\mathscr{U} = \bigcup_{i\in I} S_i$ from which the data is drawn consists of independent subspaces:

Definition 3.2 *(Independent subspaces).* Subspaces $\{S_i \subset \mathbb{R}^D\}_{i=1}^n$ are called independent if $\dim(S_1 + \dots + S_n) = \dim(S_1) + \dots + \dim(S_n)$.

In particular, $\{S_i \subset \mathbb{R}^D\}_{i=1}^n$ are independent, then $\sum_{i=1}^n \dim(S_i) \leqslant D$ and $S_i \cap S_j = \{0\}$ for $i \neq j$. Note that if the data $\mathbf{W} = \{w_1, \dots, w_N\}$ is generic and is drawn from a union $\mathscr{U} = \bigcup_{i\in I} S_i$ of M independent subspaces $S_i \in \mathbb{R}^D$ of dimension d_i, then the solution to Problem 1 is precisely the subspaces S_i from which $\mathbf{W}$ is drawn. However, for this case, the solution can be obtained in a more efficient and direct way as will be developed below.

We note that to find the subspaces S_i it would suffice to find the partition $P(\mathbf{W}) = \{\mathbf{W}_1, \dots, \mathbf{W}_M\}$ of the data $\mathbf{W}$. From this partition, the subspaces can be obtained simply by $S_i = \operatorname{span}\mathbf{W}_i$. Conversely, if we knew the subspaces S_i, it would be easy to find the partition $P(\mathbf{W}) = \{\mathbf{W}_1, \dots, \mathbf{W}_M\}$ such that $\mathbf{W}_i \subset S_i$. However, all we are given is the data $\mathbf{W}$, and we do not know the partition $P(\mathbf{W})$ or the subspaces $\mathbf{W}_i$. Our goal for solving Problem 1 from this case is to find the partition $P(\mathbf{W}) = \{\mathbf{W}_1, \dots, \mathbf{W}_M\}$ of $\mathbf{W}$. To do this, we construct a matrix $\mathbf{W} = [w_1, \dots, w_N]$ whose columns are the data vectors $w_i \in \mathbb{R}^D$. The matrix $\mathbf{W}$ is a $D \times N$ matrix, where D may be large, thus our first goal is to replace $\mathbf{W}$ by another matrix $\widetilde{\mathbf{W}}$ while preserving the clusters:

Proposition 3.3. *Let A and B be $m \times n$ and $n \times k$ matrices. Let $C = AB$. Assume $J \subset \{1, 2, \dots, k\}$.*

(i) *If $b_i \in \operatorname{span}\{b_j\colon\ j \in J\}$ then $c_i \in \operatorname{span}\{c_j\colon\ j \in J\}$.*

(ii) *If A is full rank and $m \geqslant n$ then $b_i \in \operatorname{span}\{b_j\colon\ j \in J\} \iff c_i \in \operatorname{span}\{c_j\colon\ j \in J\}$.*

Proof. The relation $b_i = \sum_{j\in J} \alpha_j b_j$ implies that $Ab_i = \sum_{j\in J} \alpha_j Ab_j$, and (i) follows from the fact that the columns c_l of C and b_l of B are related by $c_l = Ab_l$. For (ii), we note that A^tA is invertible and $(A^tA)^{-1}A^tC = B$. We then apply part (i) of the proposition. □

The proposition can be paraphrased by saying that for any matrices A, B, C, a cluster of the columns of B is also a cluster of the columns of $C = AB$. A cluster of C however is not necessarily a cluster B, unless A has full rank. Thus, the proposition above suggest that – for the purpose of column clustering – we can replace a matrix B by matrix C as long as A has the stated properties. Thus by choosing A appropriately the matrix B can be replaced by a more suitable matrix C, e.g. C has fewer rows, is better conditioned or is in a format where columns can be easily clustered. One such useful format is if C is a row echelon form matrix as will be demonstrated below. In fact, the first r rows of the reduced row echelon forms (rref) of $C = AB$ and of B are the same if B has rank r:

Proposition 3.4. *Let A be an $m \times n$ full rank matrix with $m \geqslant n$, and B an $n \times k$ matrix. Then*

$$\operatorname{rref}(AB) = \begin{bmatrix} \operatorname{rref}(B) \\ 0 \end{bmatrix}.$$

In particular, if B has rank r then $\operatorname{rref}(B)$ can be obtained from the first r rows of $\operatorname{rref}(AB)$.

In particular in the absence of noise, a data matrix $\mathbf{W}$ with the SVD $\mathbf{W} = U\Sigma V^t$ has the same reduced row echelon form as that of V^t up to the rank r of $\mathbf{W}$ (see Corollary 3.5 below). This fact together with Proposition 3.3 will help us devise a reduction algorithm for subspace clustering. Before proving Proposition 3.4, recall that there are three elementary row operations that can be used to transform a matrix to its unique reduced row echelon form. The three elementary row operations can be performed by the elementary row operation matrices.

Proof of Proposition 3.4. Since the reduced row echelon form of A can be obtained by product with elementary matrices corresponding to the elementary row operations, we have

$$\operatorname{rref}(A) = E_q \cdots E_1 A = \begin{bmatrix} I_n \\ 0 \end{bmatrix}. \tag{2}$$

Applying the same elementary row operations to AB, we get

$$D := (E_q \cdots E_1)AB = (E_q \cdots E_1 A)B = \begin{bmatrix} I_n \\ 0 \end{bmatrix} B = \begin{bmatrix} B \\ 0 \end{bmatrix}, \tag{3}$$

from which we obtain

$$\operatorname{rref}(D) = \operatorname{rref}(AB) = \operatorname{rref}\left(\begin{bmatrix} B \\ 0 \end{bmatrix}\right) = \begin{bmatrix} \operatorname{rref}(B) \\ 0 \end{bmatrix}. \qquad \square \tag{4}$$

Corollary 3.5 will be utilized in the development of our subspace segmentation algorithm based on the reduced row echelon form.

Corollary 3.5. *Assume that rank$(\mathbf{W}) = r$ and let $U\Sigma V^t$ be the singular value decomposition of $\mathbf{W}$. Then*

$$\operatorname{rref}(\mathbf{W}) = \begin{bmatrix} \operatorname{rref}((V^t)_r) \\ 0 \end{bmatrix},$$

where $(V^t)_r$ *is the first* r *rows of* V^t.

Proof. Using Proposition 3.4 several times, we have that

$$\operatorname{rref}(\mathbf{W}) = \operatorname{rref}\big(U^t\mathbf{W}\big) = \operatorname{rref}\big(\Sigma V^t\big) = \operatorname{rref}\begin{bmatrix} D(V^t)_r \\ 0 \end{bmatrix} = \begin{bmatrix} \operatorname{rref}(D(V^t)_r) \\ 0 \end{bmatrix} = \begin{bmatrix} \operatorname{rref}((V^t)_r) \\ 0 \end{bmatrix}$$

where $D = \operatorname{diag}(\sigma_1, \ldots, \sigma_r)$ is an $r \times r$ diagonal matrix whose diagonal are the r (nonzero) singular values of $\mathbf{W}$. □

Definition 3.6. Matrix R is said to be the *binary reduced row echelon form* of matrix A if all non-pivot column vectors are converted to binary vectors, i.e., non-zero entries are set to one.

Theorem 3.7. *Let* $\{S_i\}_{i=1}^M$ *be a set of non-trivial linearly independent subspaces of* $\mathbb{R}^D$ *with corresponding dimensions* $\{d_i\}_{i=1}^M$. *Let* $\mathbf{W} = [w_1 \cdots w_N] \in \mathbb{R}^{D\times N}$ *be a matrix whose columns are drawn from* $\bigcup_{i=1}^M S_i$. *Assume the data is drawn from each subspace and that it is generic. Let* $\operatorname{Brref}(\mathbf{W})$ *be the binary reduced row echelon form of* $\mathbf{W}$. *Then*

1. *The inner product* $\langle e_i, b_j\rangle$ *of a pivot column* e_i *and a non-pivot column* b_j *in* $\operatorname{Brref}(\mathbf{W})$ *is one, if and only if the corresponding column vectors* $\{w_i, w_j\}$ *in* $\mathbf{W}$ *belong to the same subspace* S_l *for some* $l = 1, \ldots, M$.
2. *Moreover,* $\dim(S_l) = \|b_j\|_1$, *where* $\|b_j\|_1$ *is the* l_1*-norm of* b_j.
3. *Finally,* $w_p \in S_l$ *if and only if* $b_p = b_j$ *or* $\langle b_p, b_j\rangle = 1$.

This theorem suggests a very simple yet effective approach to cluster the data points (Algorithm 2). The data $\mathbf{W}$ can be partitioned into M clusters $\{\mathbf{W}_1, \ldots, \mathbf{W}_M\}$, such that $\operatorname{span}\mathbf{W}_l = S_l$. The clusters can be formed as follows: Pick a non-pivot element b_j in $\operatorname{Brref}(\mathbf{W})$, and group together all columns b_p in $\operatorname{Brref}(\mathbf{W})$ such that $\langle b_j, b_p\rangle > 0$. Repeat the process with a different non-pivot column until all columns are exhausted.

Algorithm 2 Subspace segmentation – row echelon form approach – no noise.

Require: $D \times N$ data matrix $\mathbf{W}$.
Find $\operatorname{rref}(\mathbf{W})$ of $\mathbf{W}$.
Find $\operatorname{Brref}(\mathbf{W})$ of $\mathbf{W}$ by setting all non-zero entries of $\operatorname{rref}(\mathbf{W})$ to 1.
for all j from 1 to N **do**
  Pick the jth column b_j of $\operatorname{Brref}(\mathbf{W})$.
  if b_j is pivot **then**
    continue
  end if
  for all i from 1 to $j-1$ **do**
    if b_i is non-pivot and $\langle b_i, b_j\rangle > 0$ **then**
      Place $\{b_i, b_j\}$ in the same cluster C_i.
      break
    end if
  end for
end for
for all C_i **do**
  Pick any $b \in C_i$.
  Separate b into unit vectors $u_i^1, \ldots, u_i^{d_i}$. {These vectors form a basis for a subspace S_i with dimension d_i.}
  for all k from 1 to N **do**
    if $b_k \in \{u_i^1, \ldots, u_i^{d_i}\}$ **then**
      Place b_k in the same cluster C_i. {This is for handling pivot columns.}
    end if
  end for
  Place the corresponding columns in $\mathbf{W}$ into the same cluster $\mathbf{W}_i$.
end for
Renumber indices i's of S_i starting from 1.

Proof of Theorem 3.7. The reduced row echelon form of $\mathbf{W}$ is of the form

$$\operatorname{rref}(\mathbf{W}) = \begin{bmatrix} R \\ 0 \end{bmatrix}. \tag{5}$$

Let P be an $N \times N$ permutation matrix such that $\mathbf{W}P = [\,U \quad V\,]$, where the columns of U are the columns associated with the pivots $\operatorname{rref}(\mathbf{W})$ and preserving their left to right order. Thus, U forms a basis for $\bigcup_{i=1}^{M} S_i$. This can be done, since the data is drawn from each subspace and it is generic, and that $\{S_i\}_{i=1}^{M}$ are independent. In particular, U includes exactly d_i points from each S_i, and $U \subset \mathbb{R}^{D\times r}$ with rank $r = \sum_{i=1}^{M} d_i$. Moreover, because of the generic assumption of the data, $|V| \geqslant M$. In addition, every column of V is a linear combination of the columns of U, that is, there exists an $r \times (N-r)$ matrix Q with $V = UQ$. Therefore

$$\mathbf{W}P = [\,U \quad V\,] = U\,[\,I_r \quad Q\,], \tag{6}$$

where I_r is $r \times r$ identity matrix. Let $E := E_l \cdots E_1$ be the product of elementary row operation matrices such that $E\mathbf{W}P = \operatorname{rref}(\mathbf{W}P)$. Then,

$$E\mathbf{W}P = EU\,[\,I_r \quad Q\,] = \begin{bmatrix} I_r & X \\ 0 & 0 \end{bmatrix}. \tag{7}$$

Thus $EU = \left[\begin{smallmatrix} I_r \\ 0 \end{smallmatrix}\right]$, and $X = Q$. By the choice of U above, we get that $[\,I_r \quad Q\,] = RP$. It follows that, $\mathbf{W}P = U\,[\,I_r \quad Q\,] = URP$, and since P is invertible, $\mathbf{W} = UR$.

$\langle e_i, b_j\rangle = 1$ if and only if $\langle e_i, r_j\rangle \neq 0$ where r_j is the column in R that corresponds to the column b_j in $\operatorname{Brref}(\mathbf{W})$. Now $r_j = \sum_{i=1}^{r} c_i e_i$. If $\langle e_i, r_j\rangle \neq 0$, then $c_i \neq 0$. Thus $w_j = Ur_j = c_i w_i + \sum_{k\neq i} c_k U e_k$. If $w_i \in S_l$, then $w_i = Ue_i$ is one of the basis vectors of S_l, and since $c_i \neq 0$, independence of the subspaces implies that $w_j \in S_l$. Conversely, if $w_j = Ur_j$ and $w_i = Ue_i$ belong to the same subspace S_l, then $w_j = c_i w_i + \sum_{Ue_k \in S_l\ k\neq i} c_k U e_k$, due to independence of the subspaces. This, together with the assumption that the data is generic implies that $c_i \neq 0$. Hence $r_j = c_i e_i + \sum_k c_k e_k$, and we get $\langle e_i, r_j\rangle = c_i \neq 0$. This proves part (1).

Now let us assume that $w_j \in S_l$. Since the data is generic and subspaces are independent, w_j can be written as a linear combination of exactly d_l columns of U. This means that there are d_l nonzero entries in the corresponding column r_j in R. Since all the nonzero entries are set to 1 for $\operatorname{Brref}(\mathbf{W})$, l_1-norm of the corresponding non-pivot columns must be d_l. This proves part (2).

Finally to prove part (3) if w_p and w_j belong to S_l, then if $w_p = Ue_p$ then part (1) implies $\langle e_p, b_j\rangle = 1$. Otherwise the fact the subspaces are independent and the data generic imply that $b_p = b_j$.

Now let b_p be a column of $\operatorname{Brref}(\mathbf{W})$ with $b_p = b_j$. Let r_p, r_j be the corresponding columns in R. Then, $w_p = Ur_p$ and $w_j = Ur_j$. Since $w_j \in S_l$, and w_p and w_j are in the span of the same column vectors of U corresponding to S_l, it follows, $w_p \in S_l$. Finally if $b_p \neq b_j$ and $\langle b_p, b_j\rangle = 1$, then r_p is a pivot column of R. Part (1) then implies that $\{w_p, w_j\}$ belong to the same subspace S_l. □

3.2. Noisy data case

In practice the data $\mathbf{W}$ is corrupted by noise. In this case, the RREF-based algorithm cannot work, even under the assumption of Theorem 3.7, since the noise will have two effects: 1) The rank of the data corrupted by noise $\mathbf{W}+\eta \subset \mathbb{R}^D$ becomes full; i.e., $\operatorname{rank}(\mathbf{W}+\eta) = D$; and 2) Even under the assumption that $r = D$, none of the entries of the non-pivot columns of $\operatorname{rref}(\mathbf{W}+\eta)$ will be zero. One way of circumventing this problem, is to use the RREF-based algorithm in combination with thresholding to set to zero those entries that are small. The choice of the threshold depends on the noise characteristics and the position of

the subspaces relative to each other. Thus the goal of this section to estimate this error in terms of these factors.

In general, $\dim(\sum_{i=1}^{M} S_i) = \operatorname{rank}(\mathbf{W}) \leqslant D$, where D is the dimension of the ambient space $\mathbb{R}^D$. After projection of $\mathbf{W}$, the new ambient space is isomorphic to $\mathbb{R}^r$, where $r = \operatorname{rank}(\mathbf{W})$, and we may assume that $\operatorname{rank}(\mathbf{W}) = D$. Without loss of generality, let us assume that $\mathbf{W} = [\,A \quad B\,]$ where the columns of A form basis for $\mathbb{R}^D$, i.e., the columns of A consist of d_i linearly independent vectors from each subspace S_i, $i = 1, \ldots, M$. Let $\widetilde{\mathbf{W}} = \mathbf{W} + \mathbf{N}$ be the data with additive noise. Then the reduced echelon form applied to $\widetilde{\mathbf{W}}$ is given by $\operatorname{rref}(\widetilde{\mathbf{W}}) = [\,I \quad \widetilde{A}^{-1}\widetilde{B}\,]$. If b_i and $\tilde{b}_i$ denote the columns of B and $\widetilde{B}$ respectively, $e_i = \widetilde{A}^{-1}\tilde{b}_i - A^{-1}b_i$, $\Delta = \widetilde{A} - A$, and $\nu_i = \tilde{b}_i - b_i$, then we have

$$e_i = \widetilde{A}^{-1}\tilde{b}_i - A^{-1}b_i = \big(I + A^{-1}\Delta\big)^{-1}A^{-1}(b_i + \nu_i) - A^{-1}b_i.$$

Let $\sigma_{\min}$ denote the smallest singular value of A, then if $\|\Delta\| \leqslant \sigma_{\min}(A)$, we get

$$\begin{aligned}
\|e_i\|_2 &= \big\|\big(I - A^{-1}\Delta + \big(A^{-1}\Delta\big)^2 - \big(A^{-1}\Delta\big)^3 + \cdots\big)A^{-1}(b_i + \nu_i) - A^{-1}b_i\big\|_2 \\
&= \big\|A^{-1}\varepsilon + \big(-A^{-1}\Delta A^{-1} + \big(A^{-1}\Delta\big)^2A^{-1} - \big(A^{-1}\Delta\big)^3A^{-1} + \cdots\big)(b_i + \nu_i)\big\|_2 \\
&\leqslant \big\|A^{-1}\big\|\|\nu_i\|_2 + \big(\big\|A^{-1}\big\|^2\|\Delta\| + \big\|A^{-1}\big\|^3\|\Delta\|^2 + \big\|A^{-1}\big\|^4\|\Delta\|^3 + \cdots\big)\big(\|b_i\|_2 + \|\nu_i\|_2\big) \\
&= \frac{\|\nu_i\|_2}{\sigma_{\min}(A)} + \frac{\|\Delta\|}{\sigma_{\min}^2(A)}\left(\frac{1}{1 - \frac{\|\Delta\|}{\sigma_{\min}(A)}}\right)\big(\|b_i\|_2 + \|\nu_i\|_2\big),
\end{aligned} \tag{8}$$

where $\|\cdot\|$ denotes the operator norm $\|\cdot\|_{\ell^2 \to \ell^2}$. Unless specified otherwise, the noise $\mathbf{N}$ will be assumed to consist of entries that are i.i.d. $\mathcal{N}(0, \sigma^2)$ Gaussian noise with zero mean and variance σ^2. For this case, the expected value of $\|\Delta\|$ can be estimated by $\mathbb{E}\|\Delta\| \leqslant C\sqrt{D}\sigma$ [36]. Note that to estimate the error in (8) we still need to estimate $\sigma_{\min}(A)$. This singular value depends on the position of the subspaces $\{S_i\}_{i=1}^M$ relative to each other which can be measured by the principle angles between them. The principle angles between two subspaces $\mathscr{F}, \mathscr{G}$, can be obtained using any pair of orthogonal bases for $\mathscr{F}, \mathscr{G}$ as described in the following lemma [37]:

Lemma 3.8. *Let $\mathscr{F}$ and $\mathscr{G}$ be two subspaces of $\mathbb{R}^D$ with $p = \dim(\mathscr{F}) \leqslant \dim(\mathscr{G}) = q$. Assume that $Q_{\mathscr{F}} \in \mathbb{R}^{D\times p}$ and $Q_{\mathscr{G}} \in \mathbb{R}^{D\times q}$ are matrices whose columns form orthonormal bases for the subspaces $\mathscr{F}$ and $\mathscr{G}$. If $1 \geqslant \sigma_1 \geqslant \sigma_2 \geqslant \cdots \geqslant \sigma_p \geqslant 0$ are the singular values of $Q_{\mathscr{F}}^t Q_{\mathscr{G}}$, then the principle angles are given by*

$$\theta_k = \arccos(\sigma_k) \quad k = 1, \ldots, p. \tag{9}$$

The dependence of the minimum singular value $\sigma_{\min}(A)$ on the principle angles between the subspaces $\{S_i\}_{i=1}^M$ is given in the theorem below, which is one of the two main theorems of this section.

Theorem 3.9. *Assume that $\{S_i\}_{i=1}^M$ are independent subspaces of $\mathbb{R}^D$ with corresponding dimensions $\{d_i\}_{i=1}^M$ such that $\sum_{i=1}^M d_i = D$. Let $\{\theta_j(S_i)\}_{j=1}^{\min(d_i, D-d_i)}$ be the principle angles between S_i and $\sum_{\ell\neq i} S_\ell$. If $A = [\,a_1 \quad \ldots \quad a_D\,]$ is a matrix whose columns $\{a_1, \ldots, a_D\} \subset \bigcup_{i=1}^M S_i$ form a basis for $\mathbb{R}^D$, with $\|a_i\|_2 = 1$, $i = 1, \ldots, D$, then*

$$\sigma_{\min}^2(A) \leqslant \min_i \left(\prod_{j=1}^{\min(d_i, D-d_i)} \big(1 - \cos^2\big(\theta_j(S_i)\big)\big)\right)^{1/D}, \tag{10}$$

where $\sigma_{\min}(A)$ is the smallest singular value of A.

Corollary 3.10. *Under the same conditions of Theorem 3.9, a simpler but possibly larger upper bound is given by*:

$$\sigma^2_{\min}(A) \leqslant \min_i \big(1-\cos\big(\theta_1(S_i)\big)\big)^{1/D} 4^{1/D}, \tag{11}$$

where $\theta_1(S_i)$ is the minimum angle between S_i and $\sum_{\ell\neq i} S_\ell$.

Theorem 3.11. *Assume that $\{S_i\}_{i=1}^M$ are independent subspaces of $\mathbb{R}^D$ with corresponding dimensions $\{d_i\}_{i=1}^M$ such that $\sum_{i=1}^M d_i = D$. Let $\{\theta_j(S_i)\}_{j=1}^{\min(d_i,D-d_i)}$ be the principle angles between S_i and $\sum_{\ell\neq i} S_\ell$. Assume that $\mathbf{W} = [w_1 \quad \cdots \quad w_N] \in \mathbb{R}^{D\times N}$ is a matrix whose columns are drawn from $\bigcup_{i=1}^M S_i$ and the data is generic for each subspace S_i. If P is a permutation matrix such that $\mathbf{W}P = [\,A_P \quad B_P\,]$, and A_P is invertible, then*

$$\sup_P \big\{\sigma^2_{\min}(A_P)\big\} \leqslant \min_i \left(\prod_{j=1}^{\min(d_i,D-d_i)} \big(1-\cos^2\big(\theta_j(S_i)\big)\big) \right)^{1/D}. \tag{12}$$

In particular,

$$\sup_P \big\{\sigma^2_{\min}(A_P)\big\} \leqslant \min_i \big(1-\cos\big(\theta_1(S_i)\big)\big)^{1/D} 4^{1/D}, \tag{13}$$

where $\theta_1(S_i)$ is the minimum angle between S_i and $\sum_{\ell\neq i} S_\ell$.

Remark 3.12. The value $\sigma_{\min}(A_P)$ can be arbitrarily close to zero, thus, one of the goals is to find D columns of $\mathbf{W}$ that form a basis such that $\sigma_{\min}(A_P)$ is as close to the upper bound as possible without an exhaustive search. One possible way to do this is discussed in Section 5.1.

3.3. Proof of Theorem 3.9

The following lemma is essential in the proof of the theorem.

Lemma 3.13. *Assume that S_1 and S_2 are subspaces of $\mathbb{R}^n$ with dimensions d_1 and d_2, respectively, with $d_1 \leqslant d_2$. Let Q_1 and Q_2 be orthonormal bases for S_1 and S_2 and $\lambda_1^2 \geqslant \lambda_2^2, \ldots, \geqslant \lambda_{d_1}^2 \geqslant 0$ be the singular values of $Q_1^t Q_2$. Let $A = [\,Q_1 \quad Q_2\,]$, then,*

1. *If $d_2 > d_1$, then the spectrum $\sigma(A^tA) = \{1\} \cup \{1-\lambda_i^2, 1+\lambda_i^2\}_{i=1}^{d_1}$.*
2. *If $d_2 = d_1$, then the spectrum $\sigma(A^tA) = \{1-\lambda_i^2, 1+\lambda_i^2\}_{i=1}^{d_1}$.*

Proof of Lemma 3.13. A^tA is given by

$$A^tA = \begin{bmatrix} Q_1^t \\ Q_2^t \end{bmatrix} [\,Q_1 \quad Q_2\,] = \begin{bmatrix} Q_1^tQ_1 & Q_1^tQ_2 \\ Q_2^tQ_1 & Q_2^tQ_2 \end{bmatrix} = \begin{bmatrix} I_{d_1} & C \\ C^t & I_{d_2} \end{bmatrix}, \tag{14}$$

where $C := Q_1^tQ_2$, and I_d denotes the $d\times d$ identity matrix. Thus,

$$C^tC = V\Sigma^t\Sigma V^t,$$

where $\Sigma^t\Sigma = \operatorname{diag}\{\lambda_1^4, \lambda_2^4, \ldots, \lambda_{d_1}^4, \underbrace{0, \ldots, 0}_{d_2-d_1}\}$, i.e., the diagonal elements are the eigenvalues of C^tC. Using (14), μ^2 is an eigenvalue of A^tA, if and only if

$$\begin{bmatrix} I_{d_1} & C \\ C^t & I_{d_2} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \mu^2 \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

for some $x = \left[\begin{smallmatrix} x_1 \\ x_2 \end{smallmatrix}\right] \neq 0$, where $x_1 \in \mathbb{R}^{d_1}$ and $x_2 \in \mathbb{R}^{d_2}$. Thus, we have

$$\begin{aligned} Cx_2 &= \big(\mu^2 - 1\big)x_1, \\ C^t x_1 &= \big(\mu^2 - 1\big)x_2, \end{aligned}$$

from which we get, $C^tCx_2 = (\mu^2-1)^2 x_2$. Hence, if $x_2 \neq 0$, then $(\mu^2-1)^2$ belongs to the eigenvalues $\{\lambda_1^4, \lambda_2^4, \ldots, \lambda_{d_1}^4, \underbrace{0, \ldots, 0}_{d_2 - d_1}\}$ of C^tC. If $x_2 = 0$, then $\mu^2 = 1$, and x_1 is an eigenvector for CC^t, corresponding to the eigenvalue $\lambda_{d_1} = 0$.

It follows that if $d_2 > d_1$, then $\sigma(A^tA) \subset \{1\} \cup \{1 - \lambda_i^2, 1 + \lambda_i^2\}_{i=1}^{d_1}$, and if $d_2 = d_1$, then $\sigma(A^tA) \subset \{1 - \lambda_i^2, 1 + \lambda_i^2\}_{i=1}^{d_1}$.

To show the other inclusions, let $\lambda^4 \in \{\lambda_1^4, \lambda_2^4, \ldots, \lambda_{d_1}^4\}$ and let $x_2 \neq 0$ be the corresponding eigenvector. If $\lambda \neq 0$, define $x_1 = \frac{1}{\lambda^2} Cx_2$. Then, using (14) we get

$$\begin{bmatrix} I_{d_1} & C \\ C^t & I_{d_2} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} x_1 + Cx_2 \\ C^t x_1 + x_2 \end{bmatrix}. \tag{15}$$

Since $\lambda^2 x_1 = Cx_2$, we have that $\lambda^2 C^t x_1 = C^tCx_2 = \lambda^4 x_2$ so that we get $C^t x_1 = \lambda^2 x_2$. Thus, for $\lambda \neq 0$ we have

$$\begin{bmatrix} I_{d_1} & C \\ C^t & I_{d_2} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} x_1 + \lambda^2 x_1 \\ \lambda^2 x_2 + x_2 \end{bmatrix} = \big(1 + \lambda^2\big) \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}.$$

In particular $1 + \lambda^2$ is an eigenvalue of A^tA with the eigenvalue $\left[\begin{smallmatrix} x_1 \\ x_2 \end{smallmatrix}\right]$. If $\lambda = 0$, then C^tC is singular. Thus, C is singular as well. Let x_2 be a nonzero vector in the null space of C and define $x_1 = 0$. Then

$$\begin{bmatrix} I_{d_1} & C \\ C^t & I_{d_2} \end{bmatrix} \begin{bmatrix} 0 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0 + Cx_2 \\ 0 + x_2 \end{bmatrix} = \begin{bmatrix} 0 \\ x_2 \end{bmatrix},$$

so that $1 \in \sigma(A^tA)$. Thus in all case $\lambda^4 \in \{\lambda_1^4, \lambda_2^4, \ldots, \lambda_{d_1}^4\}$ implies that $1 + \lambda^2 \in \sigma(A^tA)$. A similar proof yields that $1 - \lambda^2 \in \sigma(A^tA)$.

Finally, if $d_2 > d_1$, C has a nontrivial kernel. Let $x_2 \neq 0$ be such that $Cx_2 = 0$ and $x_1 = 0$. Then, an argument similar to the last one implies that $1 \in \sigma(A^tA)$. Thus we have proved that if $d_2 > d_1$ then $\{1\} \cup \{1 - \lambda_i^2, 1 + \lambda_i^2\}_{i=1}^{d_1} \subset \sigma(A^tA)$, and if $d_2 = d_1$, then $\{1 - \lambda_i^2, 1 + \lambda_i^2\}_{i=1}^{d_1} \subset \sigma(A^tA)$. □

Proof of Theorem 3.9. We first consider two subspaces $\{S_1, S_2\} \subset \mathbb{R}^D$ of dimensions d_1 and d_2 with $d_1 + d_2 = D$. We note that if $A_P = AP$ where P is any permutation matrix, then A_P and A have the same singular values. Thus, without loss of generality, we assume that $A = [\, A_1 \quad A_2 \,]$, where the columns of A_1 and A_2 are unit norm bases of S_1 and S_2, respectively. Using the QR decomposition, we get

$$A = [\, A_1 \quad A_2 \,] = [\, Q_1 R_1 \quad Q_2 R_2 \,] = [\, Q_1 \quad Q_2 \,] \begin{bmatrix} R_1 & 0 \\ 0 & R_2 \end{bmatrix} = QR,$$

where Q_1 and Q_2 are orthonormal and R_1 and R_2 are upper triangular matrices with unit column vectors. It follows that

$$\det\big(A^tA\big) = \det\big(R^tQ^tQR\big) = \det\big(R^tR\big) \det\big(Q^tQ\big) \leqslant \det\big(Q^tQ\big), \tag{16}$$

where for the last inequality we have used the fact that the column vectors of R_1 and R_2 have unit norm. Let $\{\tilde{\mu}_i\}_{i=1}^{D}$ and $\{\mu_i\}_{i=1}^{D}$ be the singular values of A and Q, respectively. Hence, by (16) and Lemma 3.13 we get

$$\prod_{i}^{D}\tilde{\mu}_i^2 \leqslant \prod_{i}^{D}\mu_i^2 = \big(1-\lambda_1^2\big)\big(1-\lambda_2^2\big)\dots\big(1-\lambda_{d_1}^2\big)\big(1+\lambda_{d_1}^2\big)\big(1+\lambda_{d_2-1}^2\big)\dots\big(1+\lambda_1^2\big). \tag{17}$$

Noting that $\tilde{\mu}_1$ is the smallest singular value for A, and using Lemma 3.8 we obtain

$$\sigma_{\min}^{D}(A) = \big(\tilde{\mu}_1^2\big)^D \leqslant \big(1-\lambda_1^4\big)\big(1-\lambda_2^4\big)\big(1-\lambda_3^4\big)\dots\big(1-\lambda_{d_1}^4\big) \tag{18}$$

$$\leqslant \prod_{j=1}^{d_1}\big(1-\cos^2\big(\theta_j(S_1)\big)\big). \tag{19}$$

For the general case of M subspaces, we replace S_1 by S_i, S_2 by $\sum_{\ell\neq i} S_\ell$, d_1 by $\min(d_i, D-d_i)$, and we let i run from 1 to M. □

Proof of Corollary 3.10. As in the previous proof, for two subspaces $\{S_1, S_2\} \subset \mathbb{R}^D$ of dimensions d_1 and d_2 with $d_1 + d_2 = D$, we use (18) to get

$$\begin{aligned}\sigma_{\min}^{D}(A) = \big(\tilde{\mu}_1^2\big)^D &\leqslant \big(1-\lambda_1^2\big)\big(1+\lambda_1^2\big)\big(1-\lambda_2^4\big)\big(1-\lambda_3^4\big)\dots\big(1-\lambda_{d_1}^4\big)\\ &\leqslant \big(1-\lambda_1^2\big)\big(1+\lambda_1^2\big)\big(1-\lambda_{d_1}^4\big)\\ &\leqslant \big(\mu_1^2\big)\big(1-\lambda_{d_1}^2\big)(2)^2.\end{aligned}$$

This implies

$$\sigma_{\min}(A) \leqslant \mu_1^{2/D}4^{1/D} = \big(1-\cos\big(\theta_1(S_1)\big)\big)^{1/D}4^{1/D}.$$

To finish the proof, as before, we replace S_1 by S_i, S_2 by $\sum_{\ell\neq i} S_\ell$, and we let i run from 1 to M. □

4. Subspace segmentation algorithm for noisy data

Algorithm 1 described in Section 1 works perfectly in noiseless data. For noisy data, the success of the algorithm depends on finding a good initial partition. Otherwise, the algorithm may terminate at a local minimum. Algorithm 2 described in Section 3 works perfectly for *noiseless* data (it determines a basis for each subspace and it correctly clusters all of the data points). However, it does not perform very well when sufficiently large noise is present because any threshold value will keep some of the values that need to be zeroed out and will zero out some of the values that need to be kept. However, the thresholded reduced echelon form can be used to determine a set of clusters that can in turn be used to determine a good initial set of subspaces in Algorithm 1.

For example, if the number of subspaces is known and the subspaces have equal and known dimensions (assume that there are M subspaces and each subspace has dimension d), then Algorithm 3 below combines Algorithms 1 and 2 as follows: First, the reduced row echelon form $\operatorname{rref}(\mathbf{W})$ of $\mathbf{W}$ is computed. Since the data is noisy, the non-pivot columns of $\operatorname{rref}(\mathbf{W})$ will most likely have all non-zero entries. The error in those entries will depend on the noise and the positions of the subspaces as in Theorem 3.9. Since each subspace is d-dimensional, the highest d entries of each non-pivot column is set to 1 and all other entries are set to 0. This determines the binary reduced row echelon form $\operatorname{Brref}(\mathbf{W})$ of $\mathbf{W}$ (note that, according to Theorem 3.7, each non-pivot column of $\operatorname{Brref}(\mathbf{W})$ is supposed to have d entries). M groups of the equivalent columns

of Brref($\mathbf{W}$) are determined and used as the initial partition for Algorithm 1. This process is described in Algorithm 3. Note that a dimensionality reduction is also performed (according to Corollary 3.5) to speed up the process.

Algorithm 3 Combined algorithm – optimal solution $\mathbf{S}^o$.

Require: Normalized data matrix $\mathbf{W}$.
1: Set $r = M \times d$.
2: Compute the SVD of $\mathbf{W}$ and find $(V^t)_r$ as in Corollary 3.5.
3: Replace the data matrix $\mathbf{W}$ with $(V^t)_r$.
4: Compute rref($\mathbf{W}$)
5: Compute Brref($\mathbf{W}$) by setting the highest d entries of each non-pivot column to 1 and all the others to 0.
6: Group the non-pivot equivalent columns of Brref($\mathbf{W}$) into M largest clusters $\{\mathbf{W}_1, \ldots, \mathbf{W}_M\}$ and set the initial partition $P = \{\mathbf{W}_1, \ldots, \mathbf{W}_M\}$.
7: For each subset $\mathbf{W}_i$ in the partition P find the subspace $S_i^o(P)$ that minimizes the expression $e(\mathbf{W}_i, S) = \sum_{w \in \mathbf{W}_i} d^p(w, S)$.
8: **while** $\sum_{i=1}^{M} e(\mathbf{W}_i, S_i^o(P)) > e(\mathbf{W}, \mathbf{S}^o(P))$ **do**
9: **for all** i from 1 to M **do**
10: Update $\mathbf{W}_i = \{w \in \mathbf{W}\colon\ d(w, S_i^o(P)) \leqslant d(w, S_k^o(P)),\ k = 1, \ldots, M\}$
11: Update $S_i^o(P) = \operatorname{argmin}_S e(\mathbf{W}_i, S)$
12: **end for**
13: Update $P = \{\mathbf{W}_1, \ldots, \mathbf{W}_M\}$
14: **end while**
15: $\mathbf{S}^o = \{S_1^o(P), \ldots, S_M^o(P)\}$

In Step-7 of Algorithm 3, we find the subspace $S_i^o(P)$ that minimizes the expression $e(\mathbf{W}_i, S) = \sum_{w \in \mathbf{W}_i} d^p(w, S)$ for each subset $\mathbf{W}_i$ in the partition P. For data with light-tailed noise (e.g. Gaussian distributed noise) $p = 2$ is optimal and the minimum in Step-7 can be found using SVD. For heavy-tailed noise (e.g. Laplacian distributed noise), $p = 1$ is the better choice as described in the simulations section.

Remark 4.1. In Step-5 of Algorithm 3, Brref($\mathbf{W}$) is computed by setting the highest d entries of each non-pivot columns to 1 and the others to 0. If we do not know the dimensions of the subspaces, we may need to determine a threshold from the noise characteristics and a priori knowledge of the relative position of subspaces using (8) and (10).

Remark 4.2. In order to reduce the dimensionality of the problem, we compute the SVD of $\mathbf{W} = U\Sigma V^t$, where $U = [\, u_1 \quad u_2 \quad \cdots \quad u_D \,]$ is a $D \times D$ matrix, $V = [\, v_1 \quad v_2 \quad \cdots \quad v_N \,]$ is an $N \times N$ matrix, and Σ is a $D \times N$ diagonal matrix with diagonal entries $\sigma_1, \ldots, \sigma_l$ with $l = \min\{D, N\}$. In Algorithm 3, each subspace is d-dimensional and there are M subspaces. Therefore, it replaces $\mathbf{W}$ by $(V^t)_r$, where $r = M \times d$ is known or estimated rank of $\mathbf{W}$.

5. Simulations and experiments

5.1. Simulations for selection of pivot columns

In order to pick pivots columns from $\mathbf{W}$ to form the basis A in such a way that the value $\sigma_{\min}(A)$ is as close to the upper bound in (10) as possible, at each step of the reduced row echelon form process, we pick a pivot column that has the largest entry. The simulation in this section shows that this is a good technique.

Fig. 1 shows the relationship between the minimum angle and the segmentation rate. For this simulation, 10 data points that come from two 2-dimensional subspaces of $\mathbb{R}^4$ was generated. The angles between the subspaces are computed. Then, some white noise was added to the data in a controlled fashion, i.e., the noise variance was increased from 0.00 to 0.40 with 0.01 increments. The segmentation rate for each step is calculated and then the average segmentation rate is computed. The experiment is repeated 200 times and the scatter plots for three techniques are displayed in Fig. 1. The best-A method refers to the segmentation by using matrix A found after an exhausted search. The modified RREF method refers to the

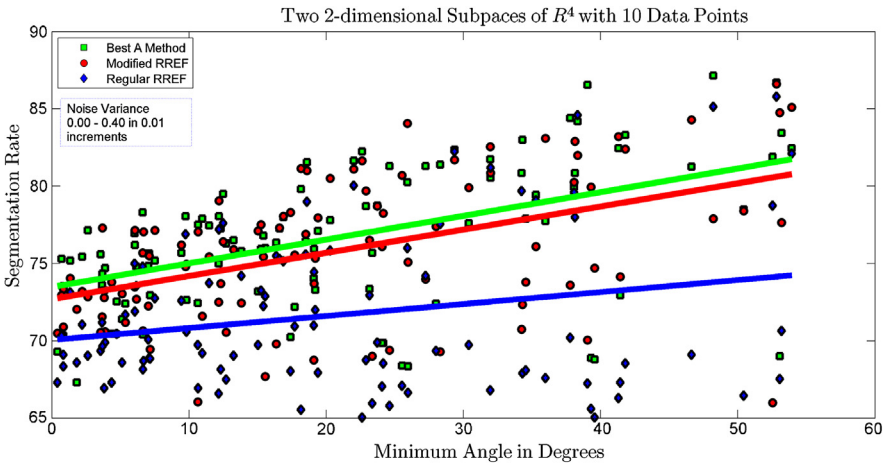

Fig. 1. Comparison of the relationship between minimum angle and segmentation rates in noisy data using three approaches for performing the RREF.

segmentation by giving priority to the highest entries for finding the pivot rows in the reduced row echelon form calculations. The regular RREF method refers to the segmentation using the traditional reduced row echelon form calculation. After computing the reduced row echelon forms using those three techniques, a spectral clustering technique was applied.

5.2. Simulations of subspace segmentation (Algorithm 3)

This section provides various simulations performed on synthetically generated data. The data is first added with Gaussian distributed noise (light-tailed noise). The data is then contaminated with Laplacian distributed noise (heavy-tailed noise). We also evaluated the effect of outliers. In all of the experiments, subspaces with known dimensions are simulated to avoid computing a data driven threshold. Also, the rank of the data matrix is assumed to be known. This is to make sure that simulations evaluate the intended cases properly.

5.2.1. Simulations – light-tailed noise

Let $\mathbf{W}$ be $D \times N$ dimensional matrix of data drawn from a single d dimensional subspace $S \in \mathbb{R}^D$. In order to find S, $\mathbf{W}$ can be factorized as $\mathbf{W} = UV^t$ where the columns of the $D \times d$ matrix U form a basis for S and V^t is a $d \times N$ matrix. However if the data is noisy, we must estimate U.

If the noise is additive and Gaussian, then the maximum likelihood estimation of U (and V) can be obtained as the minimization of the following error [38].

$$E(U,V) = \left\|\mathbf{W} - UV^t\right\|_2^2. \tag{20}$$

It is known that the SVD-based matrix factorization gives the global minimum of (20). Therefore, for light tailed noise, we choose $p = 2$ in Step-7 of Algorithm 3, and apply this approach for each $\mathbf{W}_i$, i.e., we factor $\mathbf{W}_i = U_i\Sigma_i V_i^t$ and assign $S_i^o(P) = \text{span}\{u_{i_1}, \ldots, u_{i_d}\}$ where $\{u_{i_1}, \ldots, u_{i_d}\}$ are the columns of U_i.

Fig. 2 shows a sample result for segmenting data that comes from three 4-dimensional subspaces of $\mathbb{R}^{20}$ with different number of points. Each data point (each column of data matrix) was normalized using l_2-norm. Gaussian distributed noise was added in each step of simulation. Since the data is normalized, noise variance represent approximately the percentage noise added to the data. The algorithm is robust for around 15% noise level, which is a considerably high measurement noise rate.

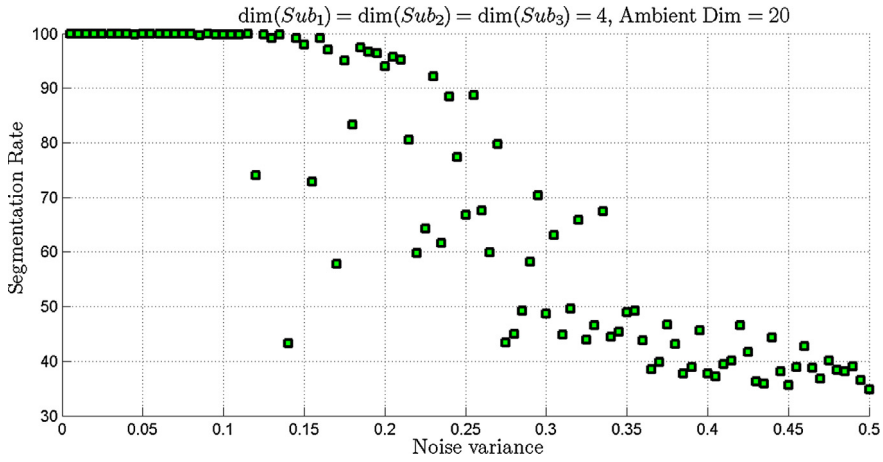

Fig. 2. Segmentation rate for: Three subspaces of $\mathbb{R}^{20}$ with $\dim(S_1) = 4$, $\dim(S_2) = 4$, $\dim(S_3) = 4$, number of data points for $S_1 = 300$, $S_2 = 400$, $S_3 = 500$, and contaminated with Gaussian distributed noise.

5.2.2. Simulations – heavy-tailed noise

In many computer vision applications such as motion segmentation and target tracking, noise is modeled as non-Gaussian heavy-tailed distribution based on empirical studies [39–41]. It is therefore important to analyze this case. Now, assume that noise can be modeled by i.i.d. Laplacian distribution, which is a heavy-tailed distribution [42]. Then, the maximum likelihood estimation of U (and V) can be obtained as the minimization of the following error [38].

$$E(U,V) = \left\|\mathbf{W} - UV^t\right\|_1. \tag{21}$$

This is generally a non-convex optimization problem. However, if U is known, $E(U,V)$ becomes a convex function with respect to V and similarly if V is known, $E(U,V)$ becomes a convex function with respect to U. Therefore, we will need to determine U and V iteratively [38].

Thus, for this case we choose $p = 1$ in Step-7 of Algorithm 3, and we factor $\mathbf{W}_i = U_i V_i^t$ based on ℓ_1-norm approach (of (21)) as described in Algorithm 4 and assign $S_i^o(P) = \text{span}\{u_{i_1}, \ldots, u_{i_d}\}$ where $\{u_{i_1}, \ldots, u_{i_d}\}$ are the columns of U_i. Although, in theory $p = 1$ is a better choice for handling heavy-tailed noise, note that the alternating minimization algorithm used to solve Eq. (21), doesn't guarantee a global optimum, in general. Thus, it is not clear that the alternating minimization algorithm for finding the optimal solution will work better that $p = 2$ in all cases.

Algorithm 4 Iterative solution for (21).

Initialize U by SVD: $\mathbf{W} = U\Sigma V^t$
while not converged **do**
  $V = \text{argmin}_V \|\mathbf{W} - UV^t\|_1$
  for all i from 1 to N **do**
    $v_i = \text{argmin}_v \|W_i - Uv\|_1$. (Note that $\|\mathbf{W} - UV^t\|_1 = \sum_{i=1}^N \|W_i - Uv_i\|_1$ where v_i^t is the ith row of V.)
  end for
  $U = \text{argmin}_U \|\mathbf{W} - UV^t\|_1$
  for all i from 1 to m **do**
    $Q := \mathbf{W}^t$
    $u_i = \text{argmin}_u \|Q_i - Vu\|_1$. (Note that $\|\mathbf{W} - UV^t\|_1 = \|Q - VU^t\|_1 = \sum_{i=1}^d \|Q_i - Vu_i\|_1$ where u_i^t is the ith row of U.)
  end for
end while

Fig. 3 displays a sample result for segmenting data that comes from union of two 4-dimensional subspaces of $\mathbb{R}^{12}$. Each subspace contains 100 data points. We used linear programming software library for implementing Algorithm 4. It is shown that the algorithm is robust for almost 15% noise level.

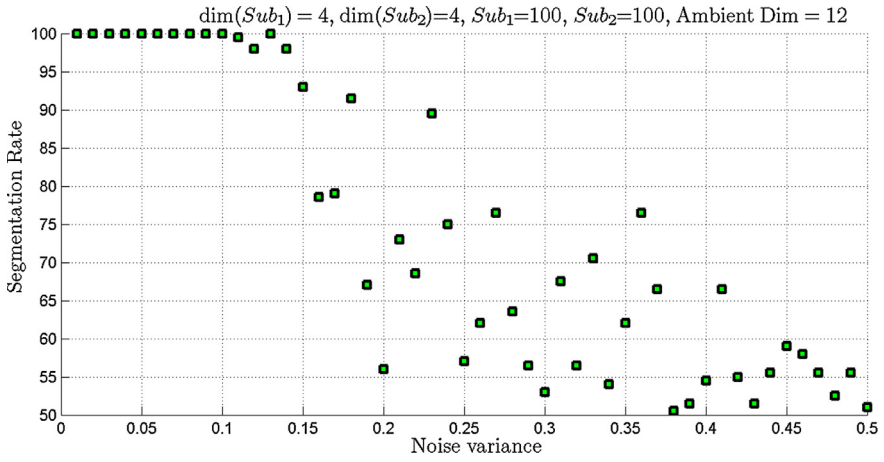

Fig. 3. Segmentation rate for: Two subspaces of $\mathbb{R}^{12}$ with $\dim(S_1) = 4$, $\dim(S_2) = 4$, number of data points for $S_1 = 100$, number of data points for $S_2 = 100$, and contaminated with Laplacian distributed noise.

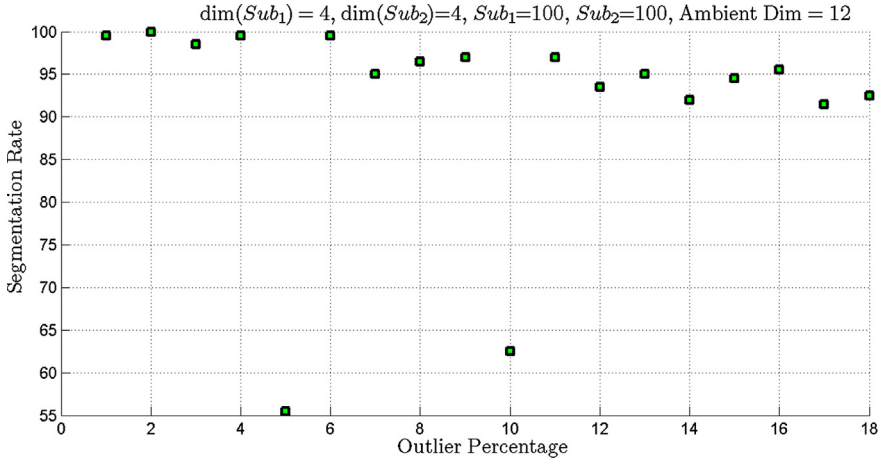

Fig. 4. Segmentation rate for: Two subspaces of $\mathbb{R}^{12}$ with $\dim(S_1) = 4$, $\dim(S_2) = 4$, number of data points for $S_1 = 100$, number of data points for $S_2 = 100$, and contaminated with Laplacian distributed noise.

5.2.3. Simulations – outliers

It is known that SVD-based matrix factorization cannot handle outliers and missing data [43,38,44–46]. ℓ_1-norm factorization approach can handle outliers robustly compared to least square approach (ℓ_2-norm approach). Missing data points can be handled in Algorithm 4 by simply ignoring the missing data points (or steps corresponding to the missing data points).

Fig. 4 shows the segmentation rates for noise-free data with outliers. In order to generate the outliers, certain number of data points from each subspace are randomly picked. Then, those points are randomly corrupted. The data contains only outliers but no noise.

5.3. Experimental results

5.3.1. Motion segmentation problem

Consider a moving affine camera that captures F frames of a scene that contains multiple moving objects. Let p be a point of 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 $\mathbb{R}^{2F}$. It can be shown that the trajectory vectors of all points of an object in a video belong to a vector subspace in

Table 1
% segmentation errors for sequences with two motions.

Checker (78)	RREF-based approach
Average	8.81%
Median	5.44%
Traffic (31)	**RREF-based approach**
Average	16.04%
Median	11.94%
Articulated (11)	**RREF-based approach**
Average	17.25%
Median	12.69%
All (120 seq)	**RREF-based approach**
Average	11.45%
Median	6.78%

$\mathbb{R}^{2F}$ of dimension no larger than 4 [47,48]. Thus, trajectory vectors in videos can be modeled by a union $\mathscr{M} = \bigcup_{i\in I} V_i$ of M subspaces where M is the number of moving objects (background is itself a motion).

The Hopkins 155 Dataset [20] was created as a benchmark database to evaluate motion segmentation algorithms. It contains two and three motion sequences. There are three groups of video sequences in the dataset: (1) 38 sequences of outdoor traffic scenes captured by a moving camera, (2) 104 indoor checker board sequences captured by a handheld camera, and (3) 13 sequences of articulated motions such as head and face motions. Cornerness features that are extracted and tracked across the frames are provided along with the dataset. The ground truth segmentations are also provided for comparison. Table 1 displays the results for the two-motion data from the Hopkins 155 Dataset. The RREF-based algorithm is extremely fast and works well with two-motion video sequences. The average error for all two-motion sequences is 11.45%, while the best results to date is less than 1% [48]. However, the error, as some other methods (e.g. GPCA) is too high for three-motion sequences and it does not work well with such video sequences.

Acknowledgment

The research of Akram Aldroubi is supported in part by NSF Grant DMS-110863. The research of Ali Sekmen is supported in part by NASA Grant NNX12AI14A. We would like to thank Amira Azhari (the mother of Aldroubi) for her culinary support, and Rosy the cat for never leaving our work table while writing this paper.

References

[1] Y.M. Lu, M.N. Do, A theory for sampling signals from a union of subspaces, IEEE Trans. Signal Process. 56 (2008) 2334–2345.
[2] K. Kanatani, Y. Sugaya, Multi-stage optimization for multi-body motion segmentation, in: IEICE Trans. Inf. Syst, 2003, pp. 335–349.
[3] A. Aldroubi, K. Zaringhalam, Nonlinear least squares in $\mathbb{R}^n$, Acta Appl. Math. 107 (2009) 325–337.
[4] R. Vidal, Y. Ma, S. Sastry, Generalized Principal Component Analysis, 2006, unpublished.
[5] G. Chen, A.V. Little, M. Maggioni, L. Rosasco, Some recent advances in multiscale geometric analysis of point clouds, in: J. Cohen, A.I. Zayed (Eds.), Wavelets and Multiscale Analysis, in: Applied and Numerical Harmonic Analysis, Birkhäuser, Boston, 2011, pp. 199–225.
[6] R.G. Baraniuk, V. Cevher, M.F. Duarte, C. Hegde, Model-based compressive sensing, IEEE Trans. Inform. Theory (2010) 1982–2001.
[7] R. Basri, D.W. Jacobs, Lambertian reflectance and linear subspaces, IEEE Trans. Pattern Anal. Mach. Intell. 25 (2003) 218–233.
[8] J. Ho, M. Yang, J. Lim, D. Kriegman, Clustering appearances of objects under varying illumination conditions, in: Computer Vision and Pattern Recognition, 2003, pp. 11–18.
[9] A. Aldroubi, R. Tessera, On the existence of optimal unions of subspaces for data modeling and clustering, Found. Comput. Math. 11 (3) (2011) 363–379, arXiv:1008.4811v1.

[10] A. Aldroubi, C. Cabrelli, U. Molter, Optimal non-linear models for sparsity and sampling, J. Fourier Anal. Appl. 14 (2009) 793–812.
[11] I. Maravic, M. Vetterli, Sampling and reconstruction of signals with finite rate of innovation in the presence of noise, IEEE Trans. Signal Process. 53 (2005) 2788–2805.
[12] T. Blumensath, Sampling and reconstructing signals from a union of linear subspaces, IEEE Trans. Inform. Theory 57 (2011) 4660–4671.
[13] Y.C. Eldar, M. Mishali, Robust recovery of signals from a structured union of subspaces, IEEE Trans. Inform. Theory 55 (2009) 5302–5316.
[14] E. Elhamifar, R. Vidal, Sparse subspace clustering, in: IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 2790–2797.
[15] E. Elhamifar, R. Vidal, Clustering disjoint subspaces via sparse representation, in: IEEE International Conference on Acoustics, Speech, and Signal Processing, 2010.
[16] G. Liu, Z. Lin, Y. Yu, Robust subspace segmentation by low-rank representation, in: International Conference on Machine Learning, 2010, pp. 663–670.
[17] M. Soltanolkotabi, E.J. Candès, A geometric analysis of subspace clustering with outliers, Ann. Statist. 40 (2012) 2195–2238.
[18] M. Soltanolkotabi, E. Elhamifar, E.J. Candès, Robust subspace clustering, arXiv:1301.2603, 2013.
[19] R. Vidal, Y. Ma, S. Sastry, Generalized principal component analysis (GPCA), IEEE Trans. Pattern Anal. Mach. Intell. 27 (2005) 1945–1959.
[20] R. Tron, R. Vidal, A benchmark for the comparison of 3-d motion segmentation algorithms, in: Computer Vision and Pattern Recognition, 2007, pp. 1–8.
[21] P. Tseng, Nearest q-flat to m points, J. Optim. Theory Appl. 105 (2000) 249–252.
[22] M. Fischler, R. Bolles, Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography, Commun. ACM 24 (1981) 381–395.
[23] 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.
[24] Teng Zhang, Arthur Szlam, Yi Wang, Gilad Lerman, Randomized hybrid linear modeling by local best-fit flats, in: IEEE Conference on Computer Vision and Pattern Recognition, 2010, pp. 1927–1934.
[25] U.V. Luxburg, A tutorial on spectral clustering, Statist. Comput. 17 (2007) 395–416.
[26] G. Chen, G. Lerman, Spectral curvature clustering (SCC), Int. J. Comput. Vis. 81 (2009) 317–330.
[27] F. Lauer, C. Schnorr, Spectral clustering of linear subspaces for motion segmentation, in: IEEE International Conference on Computer Vision, 2009.
[28] 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.
[29] A. Goh, R. Vidal, Segmenting motions of different types by unsupervised manifold clustering, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR '07, 2007, pp. 1–6.
[30] R. Vidal, A tutorial on subspace clustering, IEEE Signal Process. Mag. 28 (2011) 52–68.
[31] G. Chen, G. Lerman, Foundations of a multi-way spectral clustering framework for hybrid linear modeling, Found. Comput. Math. 9 (2009) 517–558.
[32] Teng Zhang, Arthur Szlam, Yi Wang, Gilad Lerman, Hybrid linear modeling via local best-fit flats, Int. J. Comput. Vis. 100 (3) (2012) 217–240.
[33] C. Gear, Multibody grouping from motion images, Int. J. Comput. Vis. 29 (1998) 133–150.
[34] G. Lerman, T. Zhang, Robust recovery of multiple subspaces by geometric l_p minimization, Ann. Statist. 39 (2011) 2686–2715.
[35] A. Aldroubi, A review of subspace segmentation: Problem, nonlinear approximations, and applications to motion segmentation, ISRN Signal Process. 2013 (2013) 1–13.
[36] R. Latala, Some estimates of norms of random matrices, Proc. Amer. Math. Soc. 133 (2005) 1273–1282.
[37] G.H. Golub, C.F.V. Loan, Matrix Computations, 3rd edition, Johns Hopkins University Press, 1996.
[38] Q. Ke, T. Kanade, Robust l_1 norm factorization in the presence of outliers and missing data by alternative convex programming, in: IEEE Conference on Computer Vision and Pattern Recognition, 2005, pp. 592–599.
[39] D. Wang, C. Zhang, X. Zhao, Multivariate Laplace filter: A heavy-tailed model for target tracking, in: International Conference on Pattern Recognition, 2008, pp. 1–4.
[40] M.J. Wainwright, E.P. Simoncelli, Scale mixtures of Gaussians and the statistics of natural images, in: Adv. Neural Information Processing Systems, vol. 12, 2005, pp. 855–861.
[41] T. Eltoft, T. Kim, T. Lee, On the multivariate Laplace distribution, IEEE Signal Process. Lett. 13 (2006) 300–303.
[42] R. Marks, G. Wise, D.D. Haldeman, J.L. Whited, Detection in Laplace noise, IEEE Trans. Aerosp. Electron. Syst. AES-14 (1978) 866–872.
[43] F.D.L. Torre, M.J. Black, A framework for robust subspace learning, Int. J. Comput. Vis. 54 (2003) 2003.
[44] A. Baccini, P. Besse, A. de Faguerolles, A l_1-norm PCA and heuristic approach, in: International Conference on Ordinal and Symbolic Data Analysis, 1996, pp. 359–368.
[45] J. Brooks, J. Dula, The l_1-norm best fit hyperplane problem, Appl. Math. Lett. 26 (1) (2013) 51–55.
[46] J. Brooks, J. Dula, E. Boone, A pure l_1-norm principle component analysis, in: Optimization Online, 2010.
[47] K. Kanatani, Motion segmentation by subspace separation and model selection, in: 8th International Conference on Computer Vision, vol. 2, 2001, pp. 301–306.
[48] A. Aldroubi, A. Sekmen, Nearness to local subspace algorithm for subspace and motion segmentation, IEEE Signal Process. Lett. 19 (10) (2012) 704–707.