Abstract Algebraic-Geometric Subspace Clustering

Manolis C. Tsakiris René Vidal

***Abstract*— Subspace clustering is the problem of clustering data drawn from a union of linear subspaces. Prior algebraic-geometric approaches to this problem required the subspaces to be of equal dimension, or the number of subspaces to be known. While an algorithm addressing the general case of an unknown number of subspaces of possibly different dimensions had been proposed, a proof for its correctness had not been given. In this paper, we consider an abstract version of the subspace clustering problem, where one is given the algebraic variety of the union of subspaces rather than the data points. Our main contribution is to propose a provably correct algorithm for decomposing the algebraic variety into the constituent subspaces in the general case of an unknown number of subspaces of possibly different dimensions. Our algorithm uses the gradient of a vanishing polynomial at a point in the variety to find a hyperplane containing the subspace passing through that point. By intersecting the variety with this hyperplane and recursively applying the procedure, our algorithm eventually identifies the subspace containing that point. By repeating this procedure for other points, our algorithm eventually identifies all the subspaces and their dimensions.**

I. Introduction

Subspace clustering [15] is an important problem with diverse applications in computer vision [17], systems theory [10] and genomics [8]. Earlier work on subspace clustering featured iterative methods such as *K-subspaces* [14] and *Mixtures of Probabilistic PCA* [13], which required the number of subspaces and their dimensions to be known beforehand and were very sensitive to initialization. Later work explored algebraic-geometric methods such as *Generalized Principal Component Analysis* (GPCA) [19], [18], [20], which required either a known number of subspaces of possibly different dimensions or an unknown number of subspaces of equal dimension. In the case of an unknown number of subspaces of possibly different dimensions, an intuitive recursive version of GPCA (Recursive-GPCA) was proposed in [7]. While this method often performs satisfactorily, no proof of correctness has appeared in the literature and some undesired behavior of "ghost subspaces" has been previously observed. State-of-the-art algorithms such as *Sparse Subspace Clustering* [2], [3], [4] and Low Rank Subspace Clustering [9], [5], [16] rely on notions of sparse and low rank representation theory and spectral clustering. Although these methods are provably correct for low-dimensional subspaces of a high-dimensional ambient space that are sufficiently separated [4], [12], their performance degrades as the dimensions of the subspaces become comparable to the dimension of the ambient space.

This work was partially supported by grants NSF 1447822 and 1218709, and ONR N000141310116.

The authors are with the Center for Imaging Science of The Johns Hopkins University. Emails: {mtsakir1,rvidal}@jhu.edu.

Motivated by the need to develop provably correct subspace clustering algorithms that can handle an unknown number of subspaces of possibly high and different dimensions, in this paper we propose an alternative algorithm to Recursive-GPCA, called *Abstract Algebraic-Geometric Subspace Clustering* (AAGSC). A key difference between our approach and previous work on subspace clustering is that we consider as input to the algorithm the algebraic variety of a union of subspaces, i.e., a subspace arrangement, instead of a finite subset of it. Given a subspace arrangement $\mathcal{A}$, our goal is to decompose it into its irreducible components, which are precisely the subspaces appearing in the union. Our algorithm approaches this problem by selecting a suitable polynomial vanishing on the subspace arrangement. The gradient of this polynomial at a point in $\mathcal{A}$ gives the normal vector to a hyperplane containing the subspace passing through the point. By intersecting the subspace arrangement with the hyperplane, choosing another suitable polynomial vanishing on the intersection, computing the gradient of this new polynomial at the same point, intersecting again with the new hyperplane, and so on, we can eventually find the subspace containing the point and its dimension. By repeating this procedure at another point not in the first subspace, we can identify the second subspace and so on, until all subspaces have been identified. Using results from algebraic geometry, we are able to rigorously prove that this algorithm correctly identifies the number of subspaces, their dimensions and a basis for each subspace.

Notation: For an integer n, we denote the set $\{1,\dots,n\}$ by $[n]$. For any subset $\mathcal{W}$ of $\mathbb{R}^D$, we denote the subspace spanned by all elements of $\mathcal{W}$ by $\langle \mathcal{W} \rangle$.

II. Review of Generalized Principal Component Analysis

A. GPCA Theory and Algorithm

The GPCA algorithm [19], [18], [20], exploits the fact that a union of subspaces $\mathcal{A} = \mathcal{S}_1 \cup \cdots \cup \mathcal{S}_n$ of $\mathbb{R}^D$ is in fact an algebraic set, i.e., it is the zero locus of an ideal $\mathcal{I}_{\mathcal{A}}$[1] of the polynomial ring $\mathbb{R}[x] := \mathbb{R}[x_1,\dots,x_D]$. To see this, note first that a single subspace $\mathcal{S}$ is an algebraic set, since if $\boldsymbol{b}_1,\dots,\boldsymbol{b}_c$ is a basis of the orthogonal complement $\mathcal{S}^\perp$, then S is precisely the zero set of the ideal

$$\mathcal{I}_{\mathcal{S}} = (\boldsymbol{b}_1^\top x,\dots,\boldsymbol{b}_c^\top x). \tag{1}$$

This is a homogeneous ideal, since it is generated by homogeneous polynomials of degree 1, i.e., it admits a

[1]The reader is encouraged to consult the appendix for all relevant algebraic-geometric concepts before proceeding.

decomposition

$$\mathcal{I}_{\mathcal{S}} = \mathcal{I}_{\mathcal{S},1} \oplus \mathcal{I}_{\mathcal{S},2} \oplus \cdots \oplus \mathcal{I}_{\mathcal{S},k} \oplus \cdots \tag{2}$$

where $\mathcal{I}_{\mathcal{S},k}$ is the set of polynomials of degree k that vanish on $\mathcal{S}$. Since a finite union of algebraic sets is algebraic, we see that $\mathcal{A}$ is indeed an algebraic set. The relation of the vanishing polynomials of the subspace arrangement with those of its irreducible components $\mathcal{S}_i$ is

$$\mathcal{I}_{\mathcal{A}} = \mathcal{I}_{\mathcal{S}_1} \cap \cdots \cap \mathcal{I}_{\mathcal{S}_n}, \tag{3}$$

which shows that $\mathcal{I}_{\mathcal{A}}$ is also a homogeneous ideal. We note that each of the ideals $\mathcal{I}_{\mathcal{S}_i}$ is a prime ideal, thus showing that $\mathcal{S}_i$ is an irreducible algebraic variety in the Zariski topology. Hence, as Theorem 5 ensures, the list of subspaces of a subspace arrangement uniquely defines the subspace arrangement. In other words, the problem of retrieving the list of subspaces of a subspace arrangement is well-defined.

As equation (3) suggests, a polynomial vanishing on $\mathcal{A}$ encodes information about the normal vectors to the subspaces $\mathcal{S}_i$. In fact, it has been shown that for a subspace arrangement of n subspaces, the degree n component of $\mathcal{I}_{\mathcal{A}}$ is generated as a vector space over $\mathbb{R}$ by products of linear forms defined by the normals to the subspaces. That is:

Theorem 1 ([11]): Let $\mathcal{A} = \cup_{i=1}^{n} \mathcal{S}_i$ be a transversal[2] subspace arrangement of $\mathbb{R}^D$. Then $\mathcal{I}_{\mathcal{A},n} = \prod_{i=1}^{n} \mathcal{I}_{\mathcal{S}_i,1}$.

Part of the information about the irreducible components encoded by a vanishing polynomial on the subspace arrangement can be retrieved by computing its gradients at points of the subspace arrangement. More specifically:

Lemma 1: Let $\mathcal{A} = \cup_{i=1}^{n} \mathcal{S}_i$ be a subspace arrangement of $\mathbb{R}^D$. For $\mathbf{x} \in \mathbb{R}^D$ let $\nabla \mathcal{I}_{\mathcal{A}}|_{\mathbf{x}} := \{\nabla p|_{\mathbf{x}} : p \in \mathcal{I}_{\mathcal{A}}\} \subseteq \mathbb{R}^D$. Then for $\mathbf{x} \in \mathcal{S}_i$ we have $\langle \nabla \mathcal{I}_{\mathcal{A}}|_{\mathbf{x}} \rangle \subseteq \mathcal{S}_i^{\perp}$.

The previous two results lead to the main theorem justifying the GPCA algorithm [20], [11]:

Theorem 2: Let $\mathcal{A} = \cup_{i=1}^{n} \mathcal{S}_i$ be a transversal subspace arrangement of $\mathbb{R}^D$ and let $\boldsymbol{x} \in \mathcal{S}_i - \cup_{j \neq i} \mathcal{S}_j$. Then $\mathcal{S}_i$ is precisely the orthogonal complement of the subspace spanned by all vectors of the form $\nabla p|_{\boldsymbol{x}}$ where $p \in \mathcal{I}_{\mathcal{A},n}$.

B. Recursive-GPCA Algorithm

As seen from Theorem 2, to apply the GPCA algorithm, we need to know the number of subspaces n. It has also been shown that if n is unknown and the dimensions of the subspaces are equal, we can correctly estimate n [20]. When the number of subspaces is unknown and their dimensions are unknown and possibly different, the problem of identifying the irreducible components becomes more challenging.

To address this general case, an intuitive algorithm has been developed, known as Recursive-GPCA (RGPCA) [7]. RGPCA divides the original set $\mathcal{A}$ into a finite number of subsets as follows. Let k be the smallest degree such that $\mathcal{I}_{\mathcal{A},k} \neq 0$. Let $\boldsymbol{x}_1 \in \mathcal{A}$ be a point not lying in an intersection of irreducible components of $\mathcal{A}$. Then, RGPCA associates a subspace $\mathcal{V}_{\boldsymbol{x}_1,k}$ as the orthogonal complement of the subspace spanned by all vectors of the form $\nabla p|_{\boldsymbol{x}_1}$, $p \in \mathcal{I}_{\mathcal{A},k}$. By Lemma 1, $\mathcal{V}_{\boldsymbol{x}_1,k}$ will contain the irreducible component to which $\boldsymbol{x}_1$ belongs. Then all points belonging to $\mathcal{A} \cap \mathcal{V}_{\boldsymbol{x}_1,k}$ form the first subset $\tilde{A}_1$. Then a different point $\boldsymbol{x}_2 \in \mathcal{A} - \tilde{A}_1$ not lying in an intersection of irreducible components is chosen and a subspace $\mathcal{V}_{\boldsymbol{x}_2,k}$ is associated to it as before. The second subset formed by RGPCA is $\tilde{A}_2 = (\mathcal{A} - \tilde{A}_1) \cap \mathcal{V}_{\boldsymbol{x}_2,k}$. Then a third point $\boldsymbol{x}_3 \in \mathcal{A} - \tilde{A}_1 \cup \tilde{A}_2$ is chosen and the third subset $\tilde{A}_3 = (\mathcal{A} - \tilde{A}_1 \cup \tilde{A}_2) \cap \mathcal{V}_{\boldsymbol{x}_3,k}$ is formed. This process continues until the sets $\mathcal{V}_{\boldsymbol{x}_i,k}$ cover the entire set $\mathcal{A}$. Then the above process is applied to each set $\tilde{A}_i$ until no vanishing polynomials can be found, in which case the irreducible components are identified as the ambient spaces.

Even though no proof of correctness of RGPCA has been given, it has been shown experimentally that RGPCA performs well. However, it exhibits the artifact of *ghost-subspaces*, i.e., subspaces that are not present in the list of irreducible components of the subspace arrangement. In fact, these subspaces arise precisely as intersections of the intermediate subspaces $\mathcal{V}_{\boldsymbol{x}_i,k}$ with irreducible components not associated to point $\boldsymbol{x}_i$. In the case of two lines and a plane in $\mathbb{R}^3$ (see Fig. 2), if we first start with a point $\boldsymbol{x}$ belonging to line $\mathcal{S}_2$, the intermediate space $\mathcal{V}_{\boldsymbol{x},2}$ will be the plane $\mathcal{V}_1$ defined by the two lines $\mathcal{S}_2, \mathcal{S}_3$ and its intersection with $\mathcal{A}$ will yield a ghost-line S_4 (see Fig. 3). When we apply again RGPCA to the set $\mathcal{A} \cap \mathcal{V}_{\boldsymbol{x},2}$, which is a union of three lines, RGPCA will not be able to distinguish the ghost-line from the two lines that are irreducible components of $\mathcal{A}$ and thus will furnish an additional irreducible component.

III. Abstract Algebraic-Geometric Subspace Clustering Algorithm

In this section we present the main contribution of this paper. Given a transversal but otherwise arbitrary arrangement of n subspaces of $\mathbb{R}^D$ of dimensions $d_i = \dim(\mathcal{S}_i)$, $\mathcal{A} = \mathcal{S}_1 \cup \mathcal{S}_2 \cup \cdots \cup \mathcal{S}_n$, we propose a provably correct Abstract Algebraic-Geometric Subspace Clustering (AAGSC) algorithm that identifies the number of subspaces n, their dimensions, and a basis for each subspace.

A. Identifying an Irreducible Component

In this subsection we show how we can isolate a single irreducible component $\mathcal{S}_i$ of $\mathcal{A}$. The key idea is to construct a decreasing chain of subspace sub-arrangements obtained by intersecting $\mathcal{A}$ with a strictly decreasing chain of subspaces $\mathcal{V}_0 \supseteq \mathcal{V}_1 \supseteq \mathcal{V}_2 \supseteq \cdots$ of strictly decreasing dimensions $\dim \mathcal{V}_0 > \dim \mathcal{V}_1 > \dim \mathcal{V}_2 > \cdots$. The decreasing chain of subspaces will essentially be a chain of hyperplanes corresponding to a chain of ambient spaces of decreasing dimension that have the property that contain some fixed irreducible component $\mathcal{S}_i$ for some $i \in [n]$, i.e., for every j we will have $\mathcal{V}_j \supseteq \mathcal{S}_i$. We then show that the decreasing chain of subspaces forces the corresponding decreasing chain of subspace arrangements to stabilize at the irreducible

[2] A subspace arrangement $\mathcal{A} = \cup_{i=1}^{n} \mathcal{S}_i \subseteq \mathbb{R}^D$ is called transversal, if for any subset $\mathfrak{I}$ of $[n]$, the codimension of $\cap_{i \in \mathfrak{I}} \mathcal{S}_i$ is the minimum between D and the sum of the codimensions of all $\mathcal{S}_i$, $i \in \mathfrak{I}$.

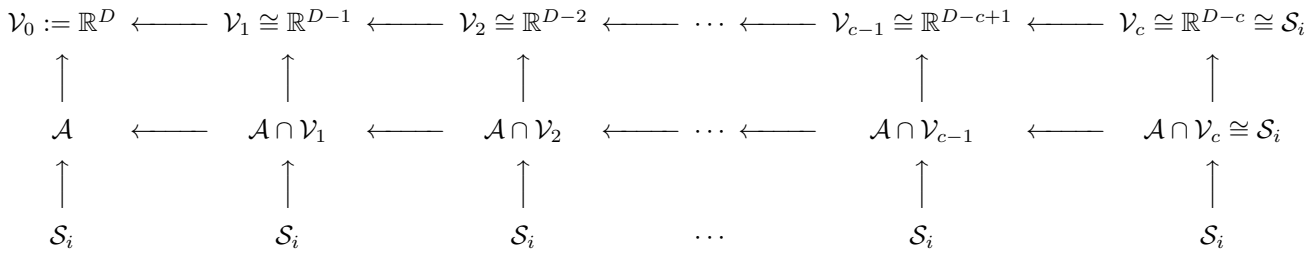

Fig. 1. Commutative diagram of the filtration associated with a reference point $\boldsymbol{x} \in \mathcal{A}$. The arrows denote embeddings. The top row of the diagram shows the strictly decreasing chain of intermediate ambient spaces. The middle row shows the induced decreasing chain of intermediate subspace arrangements. Each such arrangement contains by construction the subspace $\mathcal{S}_i$ associated to the reference point $\boldsymbol{x}$. After a finite number of $c = D - \dim \mathcal{S}_i$ steps all chains stabilize at $\mathcal{S}_i$.

component $\mathcal{S}_i$ of $\mathcal{A}$. The overall process is illustrated by the commutative diagram in Fig. 1.

To construct the decreasing chain of subspaces, we first select $\mathcal{V}_0 := \mathbb{R}^D$. Then, we define a hyperplane $\mathcal{V}_1$ of $\mathbb{R}^D$ as the orthogonal complement of the line spanned by $\nabla p|_{\boldsymbol{x}}$, where p is a non-zero vanishing polynomial of $\mathcal{A}$ of minimal degree and $\boldsymbol{x}$ is a point in $\mathcal{A}$ called the *reference point*. The following lemma ensures that, given p, we can always choose an $\boldsymbol{x}$ such that $\nabla p|_{\boldsymbol{x}} \neq 0$.

Lemma 2: Let $\mathcal{A} = \cup_{i=1}^n \mathcal{S}_i$ be a subspace arrangement of $\mathbb{R}^D$ and let k be the smallest index such that $\mathcal{I}_{\mathcal{A},k} \neq 0$. Then for any non-zero $p \in \mathcal{I}_{\mathcal{A},k}$ there exists $i \in [n]$ and $\boldsymbol{x} \in \mathcal{S}_i - \cup_{j \neq i} \mathcal{S}_j$ such that $\nabla p|_{\boldsymbol{x}} \neq 0$[3].

Note that $\boldsymbol{x} \in \mathcal{S}_i$ for some irreducible component $\mathcal{S}_i$ of $\mathcal{A}$ and we can choose $i = 1$ without loss of generality. By Lemma 1 we will have that $\mathcal{V}_1 \supseteq \mathcal{S}_1$. By intersecting $\mathcal{A}$ with $\mathcal{V}_1$ we obtain a new subspace arrangement $\mathcal{A}_1 = \mathcal{A} \cap \mathcal{V}_1 = \mathcal{S}_1 \cup \left(\cup_{j=2}^n \mathcal{S}_j \cap \mathcal{V}_1\right)$ that lives in the ambient space $\mathcal{V}_1$. Now there are two possibilities: $\mathcal{V}_1 = \mathcal{A}_1$ or $\mathcal{V}_1 \supsetneq \mathcal{A}_1$.

If $\mathcal{V}_1 = \mathcal{A}_1$, then $\mathcal{V}_1 = \mathcal{S}_1$, as shown next.

Proposition 1: If $\mathcal{V}_1 = \mathcal{A}_1$, then $\mathcal{V}_1 = \mathcal{S}_1$.

Proof: If $\mathcal{V}_1 = \mathcal{A}_1$, then the vanishing ideals satisfy

$$\mathcal{I}_{\mathcal{V}_1} = \mathcal{I}_{\mathcal{S}_1} \cap \mathcal{I}_{\mathcal{S}_2 \cap \mathcal{V}_1} \cap \cdots \cap \mathcal{I}_{\mathcal{S}_n \cap \mathcal{V}_1}. \tag{4}$$

Since every ideal appearing in (4) is a vanishing ideal of some subspace and hence prime, by Theorem 3 in the appendix, we have that $\mathcal{I}_{\mathcal{V}_1} \supseteq \mathcal{I}_{\mathcal{S}_j \cap \mathcal{V}_1}$ for some $j \in [n]$. If $j \neq 1$, then $\mathcal{S}_1 \subseteq \mathcal{V}_1 \subseteq \mathcal{V}_1 \cap \mathcal{S}_j \subseteq \mathcal{S}_j$, which is a contradiction on the transversality assumption (a subspace of a transversal subspace arrangement cannot contain another). Hence, we must have that $\mathcal{I}_{\mathcal{V}_1} \supseteq \mathcal{I}_{\mathcal{S}_1}$, which implies (see Proposition 3) that $\mathcal{V}_1 \subseteq \mathcal{S}_1$ and hence $\mathcal{V}_1 = \mathcal{S}_1$. Note also that in this case there are no non-zero polynomials vanishing on $\mathcal{A}_1$ as a subspace arrangement of $\mathcal{V}_1$. ∎

If $\mathcal{V}_1 \supsetneq \mathcal{A}_1$, we can find a hyperplane $\mathcal{V}_2$ of $\mathcal{V}_1$ that contains $\mathcal{S}_1$ by applying the following Proposition to the subspace arrangement $\mathcal{A}_1$ of $\mathcal{V}_1$ and the point $\boldsymbol{x}$.

Proposition 2: Let $\mathcal{A} = \cup_{j=1}^n \mathcal{S}_j$ be a subspace arrangement of $\mathbb{R}^D$ and let $\boldsymbol{x} \in \mathcal{A}$. Then $\boldsymbol{x} \in \mathcal{S}_l - \cup_{j \neq l} \mathcal{S}_j$ for some $l \in [n]$ if and only if there exists $k \leq n$ such that $\langle \nabla \mathcal{I}_{\mathcal{A},k}|_{\boldsymbol{x}} \rangle \neq \{0\}$.

[3]Due to space limitations most proofs are omitted.

Again we obtain a new subspace arrangement of $\mathcal{V}_2$ as $\mathcal{A}_2 = \mathcal{A}_1 \cap \mathcal{V}_2$. As before, if $\mathcal{V}_2 = \mathcal{A}_2$, then $\mathcal{V}_2 = \mathcal{S}_1$, otherwise we can find a hyperplane $\mathcal{V}_3$ of $\mathcal{V}_2$ that contains $\mathcal{S}_1$. By repeating this process, we obtain the chain of decreasing subspaces and sub-arrangements as illustrated in Fig. 1. Since all intermediate spaces $\mathcal{V}_i$ are finite-dimensional, their dimensions are strictly decreasing and they all contain $\mathcal{S}_i$, the chain $\mathcal{V}_0 (= \mathbb{R}^D) \supsetneq \mathcal{V}_1 (\cong \mathbb{R}^{D-1}) \supsetneq \mathcal{V}_2 \cong (\mathbb{R}^{D-2}) \supsetneq \cdots$ will stabilize at $\mathcal{S}_1$ precisely after c steps, i.e., $\mathcal{V}_c = \mathcal{S}_1$, where c is the codimension of $\mathcal{S}_1$. We emphasize that it is crucial to use the same point $\boldsymbol{x}$ for the construction of the above chain (contrast this with RGPCA) and that c is identified as the smallest index m such that there do not exist any non-zero vanishing polynomials of the subspace arrangement $\mathcal{A} \cap \mathcal{V}_m$ with ambient space $\mathcal{V}_m$. Note also that this immediately gives $d_1 = \dim \mathcal{S}_1 = D - c$.

B. Identifying All Irreducible Components

Having identified an irreducible component $\mathcal{S}_1$, we can proceed to construct a new decreasing chain of subspaces and subspace arrangements, this time associated with a different irreducible component of $\mathcal{A}$. Let $p \in \mathcal{I}_{\mathcal{A},k}$ be the vanishing polynomial of minimal degree that we used for the construction of the first step of the chain associated to $\mathcal{S}_1$. Since p vanishes on $\mathcal{A}$, it will also vanish on $\mathcal{S}_2 \cup \cdots \cup S_n \subsetneq \mathcal{A}$, i.e., p is a vanishing polynomial of the subspace arrangement $\cup_{j=2}^n \mathcal{S}_j$. Applying the following Lemma to the subspace arrangement $\cup_{j=2}^n \mathcal{S}_j$ and its vanishing polynomial p, we see that we can find a new reference point $\boldsymbol{x}$ in $\mathcal{A} - \mathcal{S}_1$ and a polynomial q vanishing on $\cup_{j=2}^n \mathcal{S}_j$ such that $\boldsymbol{x}$ does not lie in any intersection of irreducible components and $\nabla q|_{\boldsymbol{x}} \neq 0$.

Lemma 3: Let $\mathcal{A} = \cup_{j=1}^n \mathcal{S}_j$ be a subspace arrangement and let $0 \neq p \in \mathcal{I}_{\mathcal{A},k}$ for some k. Then there exist $i \in [n]$, $\boldsymbol{x} \in \mathcal{S}_i - \cup_{j \neq i} \mathcal{S}_j$, non-negative integer s, distinct indices $i_1, \ldots, i_s$ inside $[D]$, and non-negative integers $\ell_1, \ldots, \ell_s$, such that setting $q := \frac{\partial^\ell p}{\partial x_{i_1}^{\ell_1} \cdots \partial x_{i_s}^{\ell_s}}$, where $\ell = \sum_{j=1}^s \ell_{i_j}$, we have $0 \neq q \in \mathcal{I}_{\mathcal{A},k-\ell}$ and $\nabla q\big|_{\boldsymbol{x}} \neq 0$.

By Lemma 1, the orthogonal complement of the line spanned by $\nabla q|_{\boldsymbol{x}}$ is a hyperplane $\mathcal{V}_1$ of $\mathbb{R}^D$ that contains the irreducible component $\mathcal{S}_2$ to which $\boldsymbol{x}$ belongs. Then $\mathcal{A}_1 = \mathcal{A} \cap \mathcal{V}_1$ is a new subspace arrangement of $\mathcal{V}_1$ and we can apply the procedure of Section III-A to isolate $\mathcal{S}_2$.

In fact, if we have identified t irreducible components $\mathcal{S}_1, \mathcal{S}_2, \cdots, \mathcal{S}_t$ with $t < n$, we can identify $\mathcal{S}_{t+1}$ by applying

Lemma 3 to the subspace arrangement $\cup_{j=t+1}^{n}\mathcal{S}_j$ and its vanishing polynomial p. It is then seen that we can terminate the algorithm precisely when the condition $\mathcal{A} = \cup_{j=1}^{t}\mathcal{S}_j$ is met and identify the number of subspaces n as the corresponding index t.

IV. Illustrative Example

To illustrate the basic ideas behind our algorithm, we consider the example given in Fig. 2, where the unknown subspace arrangement $\mathcal{A}$ consists of a plane $\mathcal{S}_1$ and two lines $\mathcal{S}_2, \mathcal{S}_3$ in general position in $\mathbb{R}^3$.

The first step of the algorithm is to find a homogeneous polynomial p of minimal degree that vanishes on $\mathcal{A} = \mathcal{S}_1 \cup \mathcal{S}_2 \cup \mathcal{S}_3$. In this example, the minimal degree is 2 and the unique vanishing polynomial p with $\deg(p) = 2$ is $p(x) = (\boldsymbol{b}^\top x)(\boldsymbol{f}^\top x)$, where $\boldsymbol{b}$ is the normal to the plane $\mathcal{S}_1$ and $\boldsymbol{f}$ is the normal to the plane spanned by $\mathcal{S}_2$ and $\mathcal{S}_3$ (Fig. 3).

Given p, the algorithm picks a point $\boldsymbol{x} \in \mathcal{A}$, called the *reference point*, such that $\nabla p|_{\boldsymbol{x}} \neq 0$. In this example any point would do, because there are no non-zero intersections between $\mathcal{S}_1, \mathcal{S}_2, \mathcal{S}_3$ and the subspaces are in general position. So suppose for the sake of an argument that the algorithm picks $\boldsymbol{x} \in \mathcal{S}_2$. Then $\nabla p|_{\boldsymbol{x}}$ defines the normal to a plane, that by Lemma 1 contains $\mathcal{S}_2$. A simple computation shows that $\nabla p|_{\boldsymbol{x}}$ is colinear with $\boldsymbol{f}$, and as a result this plane will be precisely $\mathcal{V}_1$, which contains $\mathcal{S}_3$ as well. Notice also that $\mathcal{V}_1$ intersects $\mathcal{S}_1$ in a line $\mathcal{S}_4$ (Fig. 3).

Given $\mathcal{V}_1$, the algorithm intersects the original subspace arrangement $\mathcal{A}$ with the intermediate plane $\mathcal{V}_1$. This yields a new subspace arrangement $\mathcal{A}_1$ that satisfies two important properties: 1) it lives in an ambient space $\mathcal{V}_1$ of dimension one less than the original ambient dimension, and 2) it contains the subspace associated to the reference point $\boldsymbol{x}$, i.e., line $\mathcal{S}_2$. In addition, $\mathcal{A}_1$ contains the projections of the other subspaces onto $\mathcal{V}_1$, which in this case are the line $\mathcal{S}_3$ and the *ghost-line* $\mathcal{S}_4$ (see Fig. 4, *Left*). This concludes the construction of the first step of the filtration.

In the second step, to decide whether the filtration should terminate or one more step should be taken, the algorithm looks at the set of vanishing polynomials of $\mathcal{A}_1$ as a variety of $\mathcal{V}_1$. These can be seen as polynomials in two variables that vanish on $\mathcal{A}_1$ after reducing the coordinate representation of $\mathcal{V}_2$ to 2 coordinates, or equivalently, as polynomials in three variables that vanish on $\mathcal{A}_1$ but do not vanish on $\mathcal{V}_1$. In this discussion we adopt the second interpretation.

Clearly there do exist vanishing polynomials of $\mathcal{A}_1$ that do not vanish on $\mathcal{V}_1$. An example is $q(x) = (\boldsymbol{b}_2^\top x)(\boldsymbol{b}_3^\top x)(\boldsymbol{b}_4^\top x)$, where $\boldsymbol{b}_i$ is the unique normal vector of $\mathcal{V}_1$ that is orthogonal to $\mathcal{S}_i$, for $i = 2, 3, 4$ (see Fig. 4, *Right*). In fact, this is the unique vanishing polynomial of $\mathcal{A}_1$ of minimal degree 3. Because of the general position assumption, none of the lines $\mathcal{S}_2, \mathcal{S}_3, \mathcal{S}_4$ is orthogonal to another. Consequently, $\nabla q|_{\boldsymbol{x}} \neq 0$. According to Lemma 1, $\nabla q|_{\boldsymbol{x}}$ defines the normal to a plane $\mathcal{V}_2$ in $\mathcal{V}_1$ that contains $\mathcal{S}_2$. But a plane in $\mathcal{V}_1$ must be a line and we see that $\nabla q|_{\boldsymbol{x}}$ has to be collinear with $\boldsymbol{b}_2$, which also shows that $\mathcal{V}_2$ must be equal to $\mathcal{S}_2$. This can also be seen by directly computing $\nabla q|_{\boldsymbol{x}} = (\boldsymbol{b}_3^\top \boldsymbol{x})(\boldsymbol{b}_4^\top \boldsymbol{x})\boldsymbol{b}_2$.

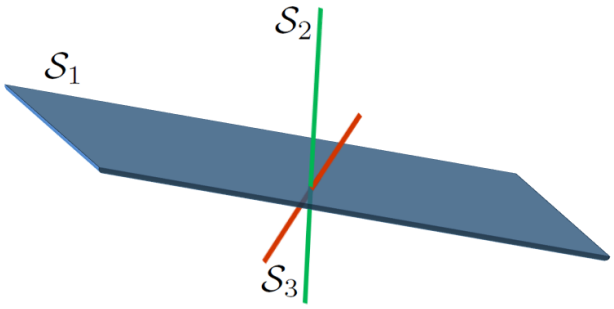

Fig. 2. A union of two lines and one plane in general position in $\mathbb{R}^3$.

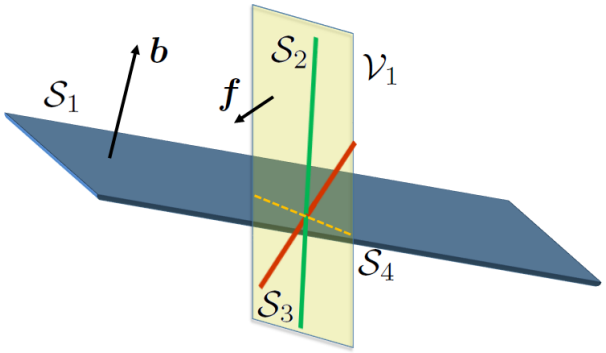

Fig. 3. The unique polynomial of degree 2 that vanishes on $\mathcal{S}_1 \cup \mathcal{S}_2 \cup \mathcal{S}_3$ is $p(x) = (\boldsymbol{b}^\top x)(\boldsymbol{f}^\top x)$, where $\boldsymbol{b}$ is the normal to $\mathcal{S}_1$ and $\boldsymbol{f}$ the normal to the plane spanned by $\mathcal{S}_2$ and $\mathcal{S}_3$.

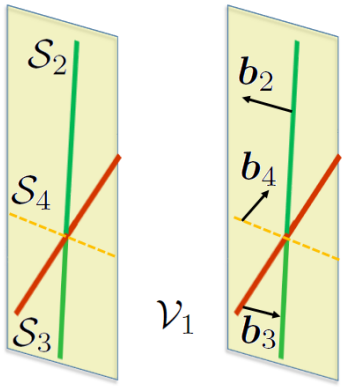

Fig. 4. *Left*: Intersection of the original subspace arrangement with the intermediate ambient space $\mathcal{V}_1$. *Right*: Geometry of the unique degree 3 vanishing polynomial $p(x) = (\boldsymbol{b}_2^\top x)(\boldsymbol{b}_3^\top x)(\boldsymbol{b}_4^\top x)$ of $\mathcal{S}_2 \cup \mathcal{S}_3 \cup \mathcal{S}_4$ in the intermediate ambient space $\mathcal{V}_1$. $\boldsymbol{b}_i \perp \mathcal{S}_i, i = 2, 3, 4$.

Since a second intermediate ambient space $\mathcal{V}_2$ was constructed, the algorithm intersects the subspace arrangement of the previous step $\mathcal{A}_1$ with $\mathcal{V}_2$. This yields a new subspace arrangement $\mathcal{A}_2 = \mathcal{A}_1 \cap \mathcal{V}_1$. Again, the algorithm does not know how $\mathcal{A}_2$ looks like. However by construction, $\mathcal{A}_2$ lives in an ambient space $\mathcal{V}_2$ of dimension one less than the previous ambient space $\mathcal{V}_1$ and it still contains the subspace associated to the reference point $\boldsymbol{x}$, i.e. $\mathcal{A}_2 \supseteq \mathcal{S}_2$.

As before, to decide whether the filtration must be terminated or not, the algorithm looks at the polynomials that vanish on $\mathcal{A}_2$ but not on $\mathcal{V}_2$. However, no such non-

zero vanishing polynomials exist, since there is no non-zero vector of $\mathcal{V}_2$ that is orthogonal to $\mathcal{S}_2$. Hence the algorithm terminates the filtration and returns the ambient space $\mathcal{V}_2 =: \mathcal{V}_2^{(1)}$ as one of the constituent subspaces of the original subspace arrangement $\mathcal{A}$.

Continuing, the algorithm now picks a new reference point $\boldsymbol{x} \in \mathcal{A} - \mathcal{V}_2^{(1)}$, say $\boldsymbol{x} \in \mathcal{S}_1$. A similar process as above will identify $\mathcal{S}_1$ as the intermediate ambient space $\mathcal{V}_1^{(2)}$ after one step of the filtration. Then a third reference point will be chosen as $\boldsymbol{x} \in \mathcal{A} - \mathcal{V}_2^{(1)} \cup \mathcal{V}_1^{(2)}$ and $\mathcal{S}_3$ will be identified as $\mathcal{V}_2^{(3)}$. Since the set $\mathcal{A} - \mathcal{V}_2^{(1)} \cup \mathcal{V}_1^{(2)} \cup \mathcal{V}_2^{(3)}$ is empty, the algorithm will terminate and return $\left\{\mathcal{V}_2^{(1)}, \mathcal{V}_1^{(2)}, \mathcal{V}_2^{(3)}\right\}$, which is up to a permutation a decomposition of the original subspace arrangement into its constituent subspaces.

V. Conclusions

We have presented a new algebraic-geometric algorithm for decomposing the algebraic variety of a union of subspaces into its constituent subspaces. The algorithm can handle varieties consisting of an unknown number of subspaces whose dimensions are arbitrary up to the transversality of their union. Using polynomials that vanish on the union of subspaces and their gradients at suitable points, the algorithm recursively decomposes the variety into its irreducible components, thus identifying the correct number of subspaces, their dimensions and a basis for each subspace. Future research will be concerned with developing algorithmic variants of the presented algorithm, that can operate on a finite subset of a union of subspaces in a robust fashion in the presence of noise, outliers and missing entries, and moreover are scalable for big-data applications.

Appendix

We provide a concise review of basic notions from commutative algebra [1] and algebraic geometry [6] in an effort to make the paper as self-contained as possible.

Definition 1 ((Prime) Ideal): A subset $\mathcal{I}$ of $\mathbb{R}[x] := \mathbb{R}[x_1, \ldots, x_D]$ is called an *ideal* if for every $p, q \in I$ and every $r \in \mathbb{R}[x]$ we have that $p + q \in \mathcal{I}$ and $rp \in \mathcal{I}$. If $p_1, \ldots, p_n$ are elements of $\mathbb{R}[x]$, then the *ideal generated* by these elements is the set of all linear combinations of the p_i with coefficients in $\mathbb{R}[x]$. An ideal $\mathcal{P}$ of $\mathbb{R}[x]$ is called *prime*, if whenever $pq \in \mathcal{P}$, then either $p \in \mathcal{P}$ or $q \in \mathcal{P}$.

Definition 2 (Product of Ideals): Let $\mathcal{I}_1, \mathcal{I}_2$ be ideals of $\mathbb{R}[x]$. The *product* $\mathcal{I}_1\mathcal{I}_2$ is the set of all elements of the form $p_1q_1 + \cdots + p_mq_m$ for any $m \in \mathbb{N}, p_i \in \mathcal{I}_1, q_i \in \mathcal{I}_2$.

Theorem 3: Let $\mathcal{P}, \mathcal{I}_1, \ldots, \mathcal{I}_n$ be ideals of $\mathbb{R}[x]$ with $\mathcal{P}$ prime. If $\mathcal{P} \supseteq \mathcal{I}_1 \cap \cdots \cap \mathcal{I}_n$, then $\mathcal{P} \supseteq \mathcal{I}_i$ for some $i \in [n]$.

Definition 3 (Algebraic Variety): A subset $\mathcal{Y}$ of $\mathbb{R}^D$ is called an *algebraic variety* or *algebraic set* if it is the zero-locus of some ideal $\mathcal{I}$ of $\mathbb{R}[x]$, i.e., $\mathcal{Y} = \left\{\boldsymbol{y} \in \mathbb{R}^D : p(\boldsymbol{y}) = 0, \forall p \in \mathcal{I}\right\}$. A standard notation is to write $\mathcal{Y} = \mathcal{Z}(\mathcal{I})$ where the operator $\mathcal{Z}(\cdot)$ denotes *zero set*.

Definition 4 (Vanishing Ideal): The *vanishing ideal* of an algebraic variety $\mathcal{Y}$ of $\mathbb{R}^D$ is the ideal $\mathcal{I}_\mathcal{Y}$ of all polynomials of $\mathbb{R}[x]$ that vanish on every point of $\mathcal{Y}$.

Definition 5: The *Zariski Topology* on $\mathbb{R}^D$ is the topology in which the closed sets are the algebraic sets.

Definition 6 (Irreducible Variety): A variety $\mathcal{Y}$ is called irreducible if it is not the union of two proper subvarieties.

Theorem 4: An algebraic variety $\mathcal{Y}$ is irreducible if and only if its vanishing ideal $\mathcal{I}_\mathcal{Y}$ is a prime ideal.

Theorem 5: Every algebraic variety $\mathcal{Y}$ of $\mathbb{R}^D$ can be uniquely written as $\mathcal{Y} = \bigcup_{i=1}^n \mathcal{Y}_i$, where $\mathcal{Y}_i$ are irreducible varieties and there are no inclusions $\mathcal{Y}_i \subseteq \mathcal{Y}_j$ for $i \neq j$. The $\mathcal{Y}_i$ are the irreducible components of $\mathcal{Y}$.

Proposition 3: If $\mathcal{Y}_1, \mathcal{Y}_2$ are algebraic varieties with vanishing ideals $\mathcal{I}_{\mathcal{Y}_1}, \mathcal{I}_{\mathcal{Y}_2}$ and $\mathcal{I}_{\mathcal{Y}_1} \subseteq \mathcal{I}_{\mathcal{Y}_2}$, then $\mathcal{Y}_1 \supseteq \mathcal{Y}_2$.

References

[1] M.E. Atiyah and I.G. MacDonald. *Introduction to Commutative Algebra*. Westview Press, 1994.

[2] E. Elhamifar and R. Vidal. Sparse subspace clustering. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[3] E. Elhamifar and R. Vidal. Clustering disjoint subspaces via sparse representation. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2010.

[4] E. Elhamifar and R. Vidal. Sparse subspace clustering: Algorithm, theory, and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(11):2765–2781, 2013.

[5] P. Favaro, R. Vidal, and A. Ravichandran. A closed form solution to robust subspace estimation and clustering. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2011.

[6] R. Hartshorne. *Algebraic Geometry*. Springer, 1977.

[7] K. Huang, Y. Ma, and R. Vidal. Minimum effective dimension for mixtures of subspaces: A robust GPCA algorithm and its applications. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume II, pages 631–638, 2004.

[8] Daxin Jiang, Chun Tang, and Aidong Zhang. Cluster analysis for gene expression data: A survey. *IEEE Trans. on Knowl. and Data Eng.*, 16(11):1370–1386, November 2004.

[9] G. Liu, Z. Lin, and Y. Yu. Robust subspace segmentation by low-rank representation. In *International Conference on Machine Learning*, 2010.

[10] Y. Ma and R. Vidal. Identification of deterministic switched ARX systems via identification of algebraic varieties. In *Hybrid Systems: Computation and Control*, pages 449–465. Springer Verlag, 2005.

[11] Y. Ma, A. Yang, H. Derksen, and R. Fossum. Estimation of subspace arrangements with applications in modeling and segmenting mixed data. *SIAM Review*, 50(3):413–458, 2008.

[12] Mahdi Soltanolkotabi, Ehsan Elhamifar, and Emmanuel J. Candès. Robust subspace clustering. *Annals of Statistics*, 42(2):669–699, 2014.

[13] M. Tipping and C. Bishop. Mixtures of probabilistic principal component analyzers. *Neural Computation*, 11(2):443–482, 1999.

[14] P. Tseng. Nearest q-flat to m points. *Journal of Optimization Theory and Applications*, 105(1):249–252, 2000.

[15] R. Vidal. Subspace clustering. *IEEE Signal Processing Magazine*, 28(3):52–68, March 2011.

[16] R. Vidal and P. Favaro. Low rank subspace clustering (LRSC). *Pattern Recognition Letters*, 43:47–61, 2014.

[17] R. Vidal and R. Hartley. Motion segmentation with missing data by PowerFactorization and Generalized PCA. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume II, pages 310–316, 2004.

[18] R. Vidal, Y. Ma, and J. Piazzi. A new GPCA algorithm for clustering subspaces by fitting, differentiating and dividing polynomials. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume I, pages 510–517, 2004.

[19] R. Vidal, Y. Ma, and S. Sastry. Generalized Principal Component Analysis (GPCA). In *IEEE Conference on Computer Vision and Pattern Recognition*, volume I, pages 621–628, 2003.

[20] R. Vidal, Y. Ma, and S. Sastry. Generalized Principal Component Analysis (GPCA). *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(12):1–15, 2005.