# Minimum Effective Dimension for Mixtures of Subspaces: A Robust GPCA Algorithm and its Applications

Kun Huang and Yi Ma
Dept. of Electrical & Computer Engineering
University of Illinois at Urbana-Champaign
Urbana, IL 61801
{kunhuang,yima}@uiuc.edu

René Vidal
Dept. of Biomedical Engineering
Johns Hopkins University
Baltimore, MD 21218
rvidal@cis.jhu.edu

## Abstract

*In this paper, we propose a robust model selection criterion for mixtures of subspaces called minimum effective dimension (MED). Previous information-theoretic model selection criteria typically assume that data can be modelled with a parametric model of certain (possibly differing) dimension and a known error distribution. However, for mixtures of subspaces with different dimensions, a generalized notion of dimensionality is needed and hence introduced in this paper. The proposed MED criterion minimizes this geometric dimension subject to a given error tolerance (regardless of the noise distribution). Furthermore, combined with a purely algebraic approach to clustering mixtures of subspaces, namely the Generalized PCA (GPCA), the MED is designed to also respect the global algebraic and geometric structure of the data. The result is a non-iterative algorithm called robust GPCA that estimates from noisy data an unknown number of subspaces with unknown and possibly different dimensions subject to a maximum error bound. We test the algorithm on synthetic noisy data and in applications such as motion/image/video segmentation.*

## 1. Introduction

Many segmentation and compression problems in computer vision (e.g., image or motion segmentation, video segmentation, feature clustering) require to group data into multiple subsets so that each subset can be fit with a single parametric model (which is typically a cluster, a probability distribution, or a manifold). Without knowing either the segmentation of the data or the corresponding model parameters, this problem is traditionally approached within the framework of maximum likelihood estimation by assuming a known error distribution. An optimal solution is typically sought in an iterative manner via the expectation-maximization (EM) algorithm.

However, it has been recently shown that if the underlying model is a mixture of linear subspaces,[1] then the segmentation problem can be solved *non-iteratively* via algebraic geometric means using a generalization of principal component analysis (PCA) to mixtures of subspaces called generalized PCA (GPCA) [14]. Although this algorithm clearly reveals the algebraic nature of the problem and provides an elegant solution when the number of models is *known* and the data presents a moderate level of noise, in practical applications one must also deal with the following issues:

- **Selectivity.** For a given set of data, if the number of subspaces and their dimension are not given a priori, then there might be more than one model that fits the data. For instance, (see Fig. 1) sample points drawn from two intersecting lines in $\mathbb{R}^3$ can also be fit by the plane spanned by the two lines. In such cases, the purely algebraic method does not provide a guiding criterion for making a choice among the possible models.
- **Robustness.** Although the purely algebraic algorithm can tolerate a moderate amount of noise in the data [14], the fact that large amounts of noise and outliers are commonplace for many segmentation problems in vision calls for improving the robustness of the GPCA algorithm.

For the selection of an "optimal" model among a class of models, many useful criteria have been developed in the algorithmic complexity and statistics communities, such as minimum message length (MML) [16], minimum description length (MDL) [8, 6], Bayesian information criterion (BIC), Akaike information criterion (AIC) [1] and Geometric AIC (G-AIC) [7]. All these criteria are very similar in nature: they all try to strike a balance between the model complexity (measured say by the dimension of the parameter space) and the data fidelity to the chosen model (typically measured as the sum of squares of the residuals). However, there are several technical difficulties that prevent us from directly applying these model selection criteria to a mixtures of subspaces:

1. These criteria are mostly based on maximum likelihood (ML) or maximum a posteriori (MAP) estimation,[2] hence they typically assume a known

[1] For instance, perspective images of different textures, images of multiple faces under varying illumination, affine trajectories of points undergoing multiple rigid-body motions, etc., all span different subspaces.

[2] Although MML is measured as the algorithmic complexity of the

probability distribution for the data and the models. However, there are many GPCA problems for which we *do not know a priori the statistical nature of the models and the data,* e.g., in the estimation of multiple epipoles from epipolar lines [13]. Furthermore, we do not even know the number and dimensions of the subspaces.

2. All these criteria reduce to minimizing the "average" of the error residuals (and the model length) [6]. But as for the *robustness of the solution*, we often need to impose a hard bound on the maximum error residual even for the "worst" fit data.
3. These criteria are designed to find the most compressed model for the data in the information-theoretic sense, which however does not necessarily *respect the global algebraic and geometric structure* of the data.

**Contributions of this paper.** In this paper, we show how to solve these difficulties by introducing a model selection criterion that is specifically designed for GPCA-type models. In particular, we address the first difficulty by minimizing the "effective dimension" of the data, which only depends on the geometric configuration of the data and its model and does not assume any particular probability distributions for either the data or the models. As for the second difficulty, we propose to minimize the effective dimension with respect to a maximum error tolerance so as to improve the robustness of the resulting algorithm. The last difficulty is resolved by minimizing the effective dimension over the subset of models that can be derived from the algebraic GPCA method hence restricting the possible solutions to those which are geometrically and algebraically correct. The combination of MED with maximum error tolerance and algebraic GPCA also leads to a non-iterative algorithm, while other robust techniques such as [11] typically require iterative optimization/minimization when searching for the best model.

The final result is a *non-iterative and robust* algorithm for estimating a unknown number of subspaces of varying dimensions from given data. We demonstrate the performance (especially the robustness) of the algorithm via a variety of simulations on synthetic data and a wide spectrum of applications in motion, image, and video segmentation.

**Relations to prior work.** Various methods have been developed in the literature in an effort to extract multiple-subspace type models from data. When the subspaces are linearly independent, geometric methods have been proposed [7, 2, 4]. When the subspaces either independent or partially dependent, [15, 14] showed that one can resort to an algebraic embedding of the data into a high-dimensional space and obtain the segmentation from the factors of the polynomials that fit the data (GPCA). When the data in each one of the subspaces are assumed to have a Gaussian distribution, [10] developed an EM-like algorithm called probabilistic PCA (PPCA) to simultaneously perform data segmentation and model estimation. This is generalized to arbitrary distributions in the exponential family in [3]. If the underlying model is a manifold, [9] shows how one can still apply PCA after embedding the data into a high-dimensional space via a nonlinear kernel function (KPCA). In [7], general information theoretic criteria such as MDL [8] and AIC [1] have been applied to select a mixture of geometric models from data. [11] has proposed to improve the robustness of these criteria via the use of Huber function (robust AIC). Our paper shows how to modify and improve these model selection criteria and robust techniques for GPCA-type model selection. The result is a non-iterative algorithm that not only selects an optimal model for data with a guaranteed error bound, but also respects the global algebraic structure encoded in the data.

## 2 Minimum effective dimension for GPCA-type model selection

**Definition 1 (Effective dimension)** *Given $n$ subspaces*[3] *$S = \cup_{i=1}^n S_i$ in $\mathbb{R}^K$ of dimension $k_i < K$, and $N_i$ sample points $\boldsymbol{X}_i = \{X_{ij}\}_{j=1}^{N_i}$ drawn from each subspace $S_i$, the* effective dimension *of the entire set of $N = \sum_{i=1}^n N_i$ sample points, $\boldsymbol{X} = \cup_{i=1}^n \boldsymbol{X}_i$, is defined to be:*

$$\mathrm{ED}(\boldsymbol{X}, S) \doteq \frac{1}{N}\sum_{i=1}^{n} k_i(K-k_i) + \frac{1}{N}\sum_{i=1}^{n} N_i k_i. \quad (1)$$

We contend that $\mathrm{ED}(\boldsymbol{X}, S)$ is the "average" number of (unquantized) real numbers that one needs to assign to $\boldsymbol{X}$ per sample point in order to specify the configurations of the $n$ subspaces and the relative locations of the sample points in the subspaces.[4] In the first term of equation (1), $k_i(K-k_i)$ is the total number of real numbers needed to specify a $k_i$-dimensional subspace $S_i$ in $\mathbb{R}^K$;[5] in the second term of (1), $N_i k_i$ is the total number of real numbers needed to specify the $k_i$ coordinates of the $N_i$ sample points in the subspace $S_i$. In general $\mathrm{ED}(\boldsymbol{X}, S)$ can be a rational number, instead of an integer for a conventional "dimension."

data, it can be easily associated with a probabilistic (MAP) interpretation via Shannon's optimal coding theory.

[3] For affine subspaces (which do not necessarily pass the origin), we first make them subspaces using the homogeneous coordinates.

[4] We here choose real numbers as the basic "units" for measuring complexity in a similar fashion to binary numbers, "bits," traditionally used in algorithmic complexity or coding theory.

[5] $k_i(K-k_i)$ is the dimension of the Grassmannian manifold of $k_i$-dimensional subspaces in $\mathbb{R}^K$. To specify a subspace, one can use the so-called Grassmannian coordinates which need exactly $k_i(K-k_i)$ entries: starting with a $K \times k_i$ matrix whose columns form a basis for the subspace, perform column-reduction so that the first $k_i \times k_i$ block is the identity matrix. Then, one only needs to give the rest $(K-k_i) \times k_i$ entries to specify the subspace.

Notice that in the above definition, the effective dimension of $\boldsymbol{X}$ depends on the subspaces $S$. This is because in general, there could be many subspace structures that can fit $\boldsymbol{X}$. For example, we could interpret the whole data set as lying in $n = 1$ $K$-dimensional subspace and we would obtain an effective dimension $K$. On the other hand, we could interpret every point in $\boldsymbol{X}$ as lying in a one-dimensional subspace spanned by itself. Then there will be $N$ such one-dimensional subspaces in total and the effective dimension, according to the above formula, will also be $K$. In general, such interpretations are obviously over-fitting. Therefore, we define the *effective dimension* of a given sample set $\boldsymbol{X}$ to be the minimum one among all possible subspace-structures that can fit the data set:[6]

$$\mathrm{MED}(\boldsymbol{X}) \doteq \min_{S:\boldsymbol{X}\subset S} \mathrm{ED}(\boldsymbol{X}, S). \tag{2}$$

**Example 1 (One plane and two lines)** *Fig. 1 shows data points drawn from one plane and two lines in $\mathbb{R}^3$. Obviously, the points in the two lines can also be viewed as lying in the plane that is spanned by the two lines. However, that interpretation would result in an increase of the effective dimension since one would need two coordinates to specify a point in a plane, as opposed to one in a line. For instance, suppose there are fifteen points in each line; and thirty points in the plane. When we use two planes to represent the data, the effective dimension is: $\frac{1}{60}(2\times 2\times 3 - 2\times 2^2 + 60\times 2) = 2.07$; when we use one plane and two lines, the effective dimension is reduced to: $\frac{1}{60}(2\times2\times3-2^2-2\times1+30\times1+30\times2) = 1.6$. In general, if the number of points $N$ is arbitrarily large (say approaching to infinity), depending on the distributions of points on the lines or the plane, the effective dimension may approach arbitrarily close to either 1 or 2, which reveals the true dimensions of the subspaces.*

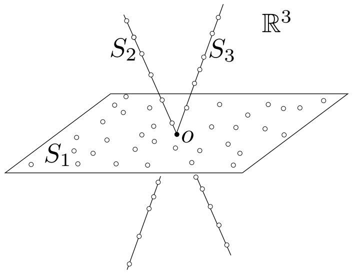

Figure 1: Data points drawn from a mixture of one plane and two lines (through the origin $o$) in $\mathbb{R}^3$.

As suggested by this intuitive example, the subspace-structure that leads to the minimum effective dimension normally corresponds to an "efficient" and hence "natural" interpretation of the data in the sense that it achieves the best compression (or dimension reduction) among all permissible subspace-structures.

# 3. Minimum effective dimension with error tolerance

In practice, real data are corrupted with noise, hence one cannot perfectly fit any (multiple-)subspace model except for the extreme cases – all points are viewed as lying on one $K$-dimensional space, or every point is viewed as lying on a one-dimensional subspace. The conventional wisdom is to strike a good balance between the complexity of the chosen model and the data fidelity (to the model). This is the same rationale that has been adopted in the classic principal component analysis (PCA) – finding a *single* lower-dimensional subspace to approximate the data. If a set of sample points $\boldsymbol{X} = [X_1, X_2, \ldots, X_N] \in \mathbb{R}^{K\times N}$ indeed fall close to a $k$-dimensional subspace $S$ in a $K$-dimensional ambient space, the ordered singular values of the data matrix $\boldsymbol{X}$ versus the dimension $k$ of the subspace will resemble the plot shown in Fig. 2. Notice that there is a significant drop in the singular value right after the "correct" dimension $k$, which is called the "knee point" of the plot. The remaining singular values indicate the residual sum of squares error of the data

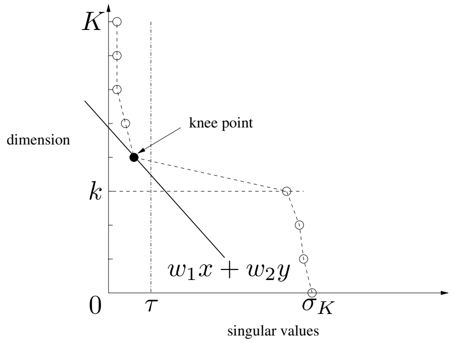

Figure 2: Dimension of the subspace versus singular values.

$$\|\boldsymbol{X}-\hat{\boldsymbol{X}}\|^2 \doteq \sum_{i=1}^{N} \|X_i-\hat{X}_i\|^2, \; \hat{X}_i \doteq \arg\min_{X\in S} \|X-X_i\|^2,$$

after it has been approximated by the $k$-dimensional subspace $S$. If we view the dimension of the subspace as the model complexity and $\|\boldsymbol{X} - \hat{\boldsymbol{X}}\|^2$ as the fidelity of the data, then the knee point of the plot has a special property: for $w_1, w_2 > 0$ in some proper range of values, it minimizes a class of objective functions of the following form:

$$J_{PCA}(S) \doteq w_1 \cdot \|\boldsymbol{X} - \hat{\boldsymbol{X}}\|^2 + w_2 \cdot \dim(S). \tag{3}$$

## 3.1 Information-theoretic model selection criteria

It turns out that the above objective function (3) is not an isolated incidence – almost all model selection criteria result in a similar form only with probably different balancing weights $w_1, w_2$ between the data fidelity and the model complexity (often $\dim(S)$ is replaced with

[6] All such subspace-structures topologically form a compact and closed set, hence the minimum effective dimension is always achievable and hence well-defined.

a more principled complexity measure of the model). For example, using the Grassmannian coordinates, the dimension of the parameter space for a $k$-dimensional subspace in $\mathbb{R}^K$ should be $Kk - k^2$ (see footnote 6). Therefore, for the class of parametric models of dimension $k$ with a Gaussian noise of variance $\sigma^2$, the MDL criterion (equivalent to BIC in this case) [8, 6] minimizes

$$\text{MDL} = \text{BIC} \doteq \frac{1}{2\sigma^2}\|\boldsymbol{X} - \hat{\boldsymbol{X}}\|^2 + \frac{(Kk - k^2)}{2}\log N.$$

More recently, Kanatani proposed the geometric AIC [7] which minimizes

$$\text{G-AIC} \doteq \frac{1}{2\sigma^2}\|\boldsymbol{X} - \hat{\boldsymbol{X}}\|^2 + (Kk - k^2 + Nk),$$

where the extra term $Nk$ accounts for the complexity in representing the data with respect to the chosen model.

We will refer to the above criteria loosely as *information-theoretic* model selection criteria, in the sense that most of these objectives can be interpreted as variations to minimizing the optimal code length for both the model and the data given a distribution and a coding scheme [6].[7]

## 3.2 Robust model selection and minimum effective dimension

Since the models that we consider in this paper for the data can be *mixtures of an unknown number of subspaces with unknown and possibly different dimensions*, the above criteria need to be modified accordingly. For instance, given a data set $\boldsymbol{X}$ and a multiple-subspace model $S$, the geometric Akaike information criterion can be generalized to

$$\text{G-AIC}(\boldsymbol{X}, S) \doteq \frac{1}{2\sigma^2}\|\boldsymbol{X} - \hat{\boldsymbol{X}}\|^2 + \sum_i^n (k_i K - k_i^2 + N_i k_i).$$

Notice that the second term is the effective dimension $\text{ED}(\boldsymbol{X}, S)$ (multiplied by $N$) defined in the previous section. By minimizing the G-AIC criterion among all possible $S$, in principle we can find the "optimal" multiple-subspace model $S^*$ for the data $\boldsymbol{X}$.

However, there are several difficulties that prevent us from directly adopting such generalized information-theoretic criteria to the GPCA problem:

- *Unknown model distribution:* Although we restrict our models to the class of multiple-subspace models, we typically do not know the number or the dimensions of the subspaces a priori. Furthermore, it is extremely difficult to introduce any a priori distribution to a class of models of this kind.[8]
- *Unknown data distribution:* In the context of GPCA, we are mostly interested in extracting the subspace structure from the data. Hence we do not necessarily want to impose a specific probability distribution of the sample points inside the subspaces.[9]
- *Lack of robustness:* Even if some statistical nature of the models and the data, say the noise variance $\sigma^2$, is somehow known, all the information-theoretic criteria minimize the "average" residue of the data while the maximum error $\max_{X \in \boldsymbol{X}} \|X - \hat{X}\|$ might be very large, especially when there are outliers in the data.

How should we modify the information-theoretical criteria, say the G-AIC, so as to improve the robustness of the selected model? The key is to observe that, when the data is noisy, the effective dimension should in fact be a notion that depends the maximum allowable error residue, which we denote as error tolerance $\tau$. Even for the one-subspace PCA model, if we lower the error tolerance to the left of the singular value at the knee point, we will be forced to increase the dimension of the subspace in order to reduce the error residue to meet the tolerance.

Therefore for noisy data, the resulting "effective dimension" of the optimal model in general depends on the given error tolerance. In the extreme, if the error tolerance is arbitrarily large, the "optimal" subspace-model for any data set can simply be the zero-dimensional origin; if the error tolerance is zero instead, for data with random noise, most sample points need to be treated either as one-dimensional subspaces in $\mathbb{R}^K$ or as points in the ambient space $\mathbb{R}^K$ directly. Both ways bring the effective dimension up close to $K$. Therefore, we need to modify the definition (2) of minimum effective dimension to allow the chosen model to tolerance certain error $\tau$. We therefore define the minimum effective dimension with error tolerance as:

$$\text{MED}(\boldsymbol{X}, \tau) \doteq \min_{S:\, \|\boldsymbol{X} - \hat{\boldsymbol{X}}\|_\infty \le \tau} \text{ED}(\hat{\boldsymbol{X}}, S), \qquad (4)$$

where $\hat{\boldsymbol{X}}$ is the projection of $\boldsymbol{X}$ onto the subspaces $S$,[10] and the error norm $\|\cdot\|_\infty$ indicates the maximum norm: $\|\boldsymbol{X} - \hat{\boldsymbol{X}}\|_\infty = \max_{X \in \boldsymbol{X}} \|X - \hat{X}\|$. The aim of robust GPCA model selection is then to find a multiple-subspace model which leads to the lowest effective dimension for a given error tolerance, hence the *minimum effective dimension (MED) criterion.*

**Comment 1 (MED and Robust AIC)** *In the work of [11], the AIC criterion is modified in order to improve its robustness via the robust AIC criterion:*[11]

[7] Even if one chooses to compare models by their algorithmic complexity, such as the minimum message length (MML) criterion [16] (an extension of the Kolmogrov complexity to model selection), a strong connection with the above information-theoretic criteria, such as MDL, can be readily established via Shannon's optimal coding theory.

[8] This class of models, although seemingly simple, is not one of those regular families that can be easily dealt with in the MDL framework [8]. For instance, the space of parameters that describe such models is not even of the same dimension.

[9] Although, in principle, knowing this distribution can further reduce the code length according to Shannon's optimal coding theory, in practice, e.g., in MDL and MML, such knowledge is typically ignored in the two-part coding for the model and the data.

[10] That is, each data point $X$ in $\boldsymbol{X}$ is projected to the closest point $\hat{X}$ in one of the subspaces of $S$.

[11] A similar version can be defined for the G-AIC criterion [11].

$$\mathrm{AICR} \doteq \frac{1}{2\sigma^2}\|\boldsymbol{X} - \hat{\boldsymbol{X}}\|_\rho + (Kk - k^2),$$

*where the norm $\|\cdot\|_\rho$ is defined by the so-called Huber function $\rho(\cdot)$ which behaves like the 2-norm , i.e. $\rho(X - \hat{X}) = \|X - \hat{X}\|^2$, for data $X$ with low residuals and a constant penalty, typically $\rho(X - \hat{X}) = \lambda(K - k)$ (a scaled version of the codimension of the model), for data $X$ with high residuals (outliers), see [11]. However, such a penalty for the outliers is difficult to generalize to or justify for the case of mixtures of multiple subspaces: there is no longer a well-defined notion of "codimension" because the union of all the subspaces often span the entire ambient space. Hence we are left with the option to either fit the outliers by additional subspaces with the same error bound $\tau$ or to completely reject the outliers from the rest of the model as long as the fitting error is large enough. As one will soon see, both options will be exploited when we show how to implement the MED criterion with the algebraic GPCA algorithm in the next section.*

*Another reason why we choose the maximum norm $\|\cdot\|_\infty$ instead of $\|\cdot\|_\rho$ is because we can use it to reject or accept a model directly without comparing to all the other models. As we will soon see, this allows us to easily combine the MED criterion with the* non-iterative *algebraic GPCA scheme to select the optimal GPCA model, whereas AICR and all the other information-theoretic criteria typically require iterative optimization.*

Notice that, given an error tolerance $\tau$, in PCA-type model selection, the dimension of the resulting single-subspace model is always discrete; but in GPCA-type model selection, the effective dimension of the resulting multiple-subspace model can be any rational number since $S$ is in general a mixture of subspaces of different dimensions. The plot of MED versus error tolerance will then be a continuous curve across all values between 0 and $K$ when the error tolerance ranges from the diameter of the data set $\tau_{max}$ to zero, as shown in Fig. 3. Then, as in the case of PCA, a "good" GPCA model takes place at the "knee point" of the plot, $(\tau^*, \mathrm{MED}^*)$, right before which we see a sharp drop of the MED, and after which the MED stabilizes when further increasing the error tolerance. Similarly to the PCA case, this knee point has a special property: for $w_1, w_2 > 0$ in some proper range of values, it minimizes a class of objective functions of the following form:

$$J_{GPCA}(S) \doteq w_1 \cdot \|\boldsymbol{X} - \hat{\boldsymbol{X}}\|_\infty + w_2 \cdot \mathrm{MED}(\boldsymbol{X}, S), \tag{5}$$

which is obviously a natural generalization of the objective $J_{PCA}$ given in (3).

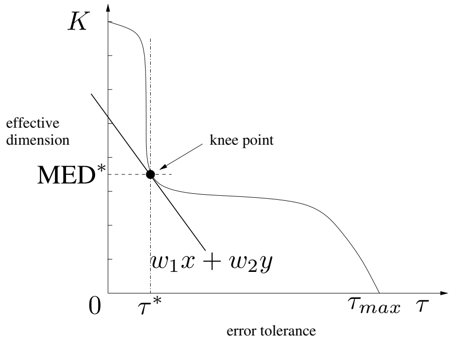

Figure 3: Minimum effective dimension versus error tolerance.

# 4 A Robust recursive GPCA algorithm

Notice that MED, as any other model selection technique, is designed only to best "compress" the representation of the data among the class of models considered. But in doing so, it will not favor any particular model which may capture better the *global* algebraic or geometric structure in the data.[12] Therefore, in order to design an effective algorithm for fitting a mixture of subspaces to data, we also need to consider the algebraic and geometric properties of the mixture model.

In this section, we show how to combine the MED criterion with the algebraic GPCA scheme to develop a robust algorithm that automatically finds a mixture of subspaces for given data within a maximum error bound. We begin by summarizing the main results about the algebraic GPCA algorithm proposed in [14], which provides a closed-form solution to the problem of finding a basis for each one of the subspaces when the number of subspaces is known and the data is perfect.

**Theorem 1 (Algebraic GPCA)** *A collection of $n$ subspaces can be described as the set of points satisfying a set of homogeneous polynomials of degree $n$*

$$f(X) = \prod_{j=1}^{n}(b_j^T X) = \boldsymbol{b}^T \nu_n(X) = 0, \ \forall X \in \boldsymbol{X}, \tag{6}$$

*where $b_j \in \mathbb{R}^K$ is a normal vector to the $j$th subspace $S_j$ and $\nu_n(X) \in \mathbb{R}^{M_n}$, where $M_n = \begin{pmatrix} n+K-1 \\ n \end{pmatrix}$ and $\nu_n(X)$ is the vector of all monomials of degree $n$ in the entries of $X$, also known as the Veronese embedding of degree $n$. When $n$ is known, one can estimate the coefficients $\boldsymbol{b}$ of all such polynomials from the null-space of the embedded data matrix $L_n = [\nu_n(X_1) \ldots \nu_n(X_N)]^T$, and the normal vectors $b_j$ to the $jth$ subspace from the derivative of the polynomials $Df(X)$ at a point $X = Y_j$ in the $jth$ subspace. A basis for the complement of $S_j$ can be obtained from $span_f Df(Y_j)$, hence the dimension of $S_j$ is given by $k_j = K - rank(span_f Df(Y_j))$.*

Let us now consider the case in which $n$ is unknown. In this case, the main difficulty of directly applying Theorem 1 (maybe for multiple values of $n$) is that when the subspaces have different dimensions, there are polynomials of degree $d < n$ that also fit the

[12] For example, if $N$ samples are scattered more or less uniformly within an area of unit length and width in the plane, one can show that MED will always partition the plane into $\frac{1}{2\tau}$ (parallel) lines given that the tolerance $\tau \geq \frac{3}{2N}$. This is a phenomenon that we often observe in any algorithm that tries to aggressively reduce the effective dimension, and we call it the "stripping effect."

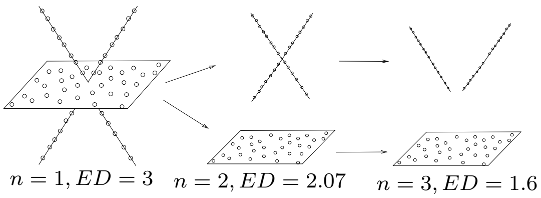

Figure 4: Recursive segmentation of the data points on one plane and two lines.

data. In Example 1, for instance, the interpretation of two planes instead of a plane and two lines leads to a polynomial of degree $2 < 3$. Therefore, we need to test if the data points on the two planes can be further segmented into lines. In other words, we can use a *recursive* scheme to search over all possible collections of subspaces that both respect the algebraic properties of the subspace structure, as stated by Theorem 1, and minimize the effective dimension. Fig. 4 demonstrates the process for the data set in Example 1.

Specifically, in the absence of noise, the recursive searching scheme is to start with $n = 1$ and increase $n$ until there is at least one polynomial of degree $n$ fitting all the data, i.e. until the matrix $L_n$ drops rank ($M_n \leq N$ imposes a constraint on the maximum possible number of groups $n_{max}$). For such an $n$, we can use Theorem 1 to separate the data into $n$ subspaces. Then we can further separate each one of these $n$ groups of points using the same procedure. Hence the process will reduce the effective dimension until there are no lower-dimensional subspaces in each group (or the total number of groups $n$ has reached the maximum $n_{max}$). One can rigorously show that if sufficient sample points are drawn from each subspace, the proposed scheme guarantees to find the correct mixture of subspaces [12].

However, the above scheme is purely algebraic and only works when the samples are noise free. When the samples are noisy, as we have contended in the previous sections, we need to adopt a *robust* approach in our algorithm, i.e. we fit the data to $n$ (subject to search) subspaces within a given error tolerance $\tau$. For a fixed $n$, we identify the $n$ subspaces one at a time. We first identify one subspace and then assign points to with an error less than $\tau$ to that subspace. We repeat this process by fitting the remaining subspaces to remaining data points. To identify a subspace, as in [14], we first find a point on that subspace from the data points that have not been assigned to any previously identified subspaces and then determine the orthogonal complement of the subspace based on this point and the null space of the matrix $L_n$. To determine the null space of $L_n$ with noisy data, we essentially need to perform PCA on the Veronese embedded data. As we have discussed in Section 3, the choice of $\tau$ determines whether the rank of the corresponding matrix $L_n$ can be correctly identified. We need to consider the following two scenarios:

1. If the rank of $L_n$ is under-estimated, then the set of normal vectors for each subspace will be over-estimated and the resulting subspaces will not be able to cover all the samples, within the given error tolerance. We call this situation *over-estimating* (the null space of $L_n$).
2. If the rank of $L_n$ is over-estimated, then the set of normal vectors for each subspace will be under-estimated and the dimension of each subspace will become too large. As a result, we may have already assigned all the data points to some subspaces before we can identify all the $n$ subspaces. We call this situation *under-estimating* (the null space of $L_n$).

Therefore, by determining if the null space of $L_n$ is over- or under-estimated within a range of possible rank values (e.g., between $r_{min}$ and $r_{max}$), we should modify our estimate of the rank. If none of the choice leads to a segmentation of the data into $n$ subspaces while satisfying the given error tolerance $\tau$, the number of subspaces $n$ should be increased. The search for the correct rank and the number of subspaces will certainly increase the amount of computation, in exchange for improved robustness of the resulting solution. We hence obtain the Robust-GPCA function below.

```
function **Robust-GPCA($\boldsymbol{X}, \tau$)**
$n = 1$; success = false;
**repeat**
    set $L_n(\boldsymbol{X}) \doteq [\nu_n(X_1), \ldots, \nu_n(X_N)]^T \in \mathbb{R}^{M_n \times N}$;
    $r_{max} = M_n - 1, r_{min} = \arg\min_i \left\{ \frac{\sigma_i(L_n)}{\sum_{j=1}^{i-1} \sigma_j(L_n)} \leq 0.02 \right\}$;
    **while** $(r_{min} \leq r_{max})$ AND (NOT success) **do**
        over-estimate = false; under-estimate = false;
        $r = \lfloor (r_{min} + r_{max})/2 \rfloor$;
        compute the last $M_n - r$ eigenvectors $\{\boldsymbol{b}_i\}$ of $L_n$;
        obtain polynomials $\{f_i(X) \doteq \boldsymbol{b}_i^T \nu_n(X)\}$;
        find $n$ points $X_j$ at which the subspaces spanned by
        the derivatives $\{Df_i(X_j)\}$ have subspace angle larger
        than $2\tau$, and if fail, under-estimate = true;
        **if** under-estimate **then**
            $r_{max} \leftarrow r - 1$;
        **else**
            assign each point in $\boldsymbol{X}$ to its closest subspace
            within the error tolerance $\tau$ and obtain the $n$ groups
            $\boldsymbol{X}_j$, and if fail, over-estimate = true;
            **if** over-estimate **then**
                $r_{min} \leftarrow r + 1$;
            **else**
                success = true;
            **end if**
        **end if**
    **end while**
    **if** success **then**
        **for** $j = 1 : n$ **do**
            **Robust-GPCA**($\boldsymbol{X}_j, \tau$);
        **end for**
    **else**
        $n \leftarrow n + 1$;
    **end if**
**until** (success) OR ($n \geq n_{max}$).
```

**Comment 2 (Outliers)** *To further improve the robustness of the algorithm, we can also assign a permissible percentage of outliers that may not be fit by the subspace-models with the given error tolerance $\tau$. Setting aside the outliers will allow the GPCA search process to identify the dominant subspace-structure that the majority of the sample points admit, without being "side-tracked" by a few very bad data points. This is important especially when the permissible error tolerance has to be small in some problems. We typically allow about $10\%$ of outliers in our experiments.*

## 5. Experiments and applications

We now demonstrate the performance of the proposed robust GPCA algorithm via a wide spectrum of numerical simulations and applications to feature segmentation, image segmentation, motion segmentation, and video segmentation. However, it is *not* our intention to convince the reader that the proposed GPCA algorithm offers an optimal solution to each of these problems.[13] We merely wish to point out that many data sets dealt with in these classic problems indeed exhibit multiple-subspace structures; if so, the proposed GPCA algorithm is an effective tool that can automatically detect such structures in a non-iterative fashion.

**Simulation results on clustering points.** Fig. 5 demonstrates the robust GPCA algorithm on segmenting a set of synthetic data drawn from two lines and one plane in $\mathbb{R}^3$ corrupted with 5% uniform noise (Fig. 5 top-left). The algorithm stops after two levels of recursion (Fig. 5 top-right). Note that the pink line or the group 4 (Fig. 5 bottom-left) is a "ghost" line at the virtual intersection of the original plane and the plane spanned by the two lines.[14] Fig. 5 bottom-right is the plot of MED versus different error tolerance for the same data set, as anticipated from Fig. 3.

**Motion segmentation via feature point clustering.** [13] has shown that all types of motion segmentation problems can be solved by GPCA. Fig. 6 demonstrates applying the robust GPCA algorithm to segment feature points on two independently moving objects from two images – the board has only a translation relative to the camera; the cube has both rotation and translation. Each pair of corresponding features $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathbb{R}^3$ satisfy the epipolar constraint $\boldsymbol{x}_2^T F \boldsymbol{x}_1 = 0$ where $F$ is the fundamental matrix corresponds to the motion for that point. $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ can be embedded in $\mathbb{R}^9$ as a single point via the Kronecker product $\boldsymbol{x}_1 \otimes \boldsymbol{x}_2$, and it satisfies $(\boldsymbol{x}_1 \otimes \boldsymbol{x}_2)^T F^s = 0$, where $F^s$ is the vector obtained by stacking the columns of $F$. Therefore, the so-obtained nine-dimensional points reside on different subspaces for different motions. At the first level of recursion, the GPCA algorithm segments the features into two groups (blue and red in Fig. 6 middle) corresponding to a 3-D motion and a planar motion (homography); and at the second level, it further segments the features on the two faces of the cube (blue and green in Fig. 6 right) since they correspond to different planar motions (homographies). Only one feature point is miss-grouped in this way.

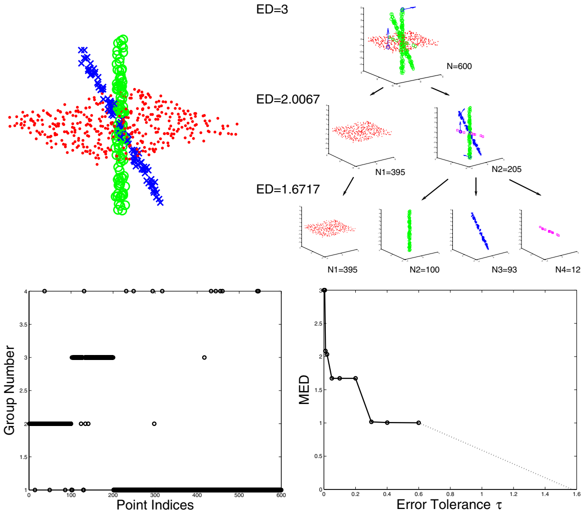

Figure 5: Simulation results. Top-left: sample points drawn from two lines and a plane in $\mathbb{R}^3$ with $5\%$ uniform noise; Top-right: the process of recursive segmentation by the GPCA algorithm at the error tolerance $\tau = 0.05$; Bottom-left: group assignment for the points; Bottom-right: plot of MED versus error tolerance.

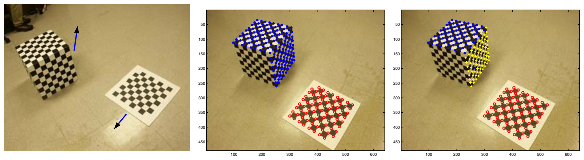

Figure 6: Feature points clustering by different 3-D motions. Left: An image showing two objects with different motions. Middle: Results for the first level of recursion for the GPCA algorithm. Feature points corresponding to the same motion are marked as the same group. Right: Results for the second level of recursion.

**Image segmentation via pixel clustering.** Fig. 7 shows the results of applying the GPCA algorithm to the segmentation of some images from the Berkeley image database. Classic image representation techniques such as Karhunen-Leove transformation usually model the imagery signals (e.g. blocks of pixels) using a single linear model [5]. However, for images with different color and texture components, multiple linear models can be more realistic and GPCA algorithm is designed for such a task. By associating a $16 \times 16$ window to every pixel, we can use GPCA to segment the pixels based on the local color and texture information around them. In the experiments, PCA is first applied to all the windows (which are $16 \times 16 \times 3 = 768$ dimensional data points), and the first twelve eigenvectors are

[13] In fact, one can easily obtain better segmentation results by using algorithms/systems specially designed for these tasks.

[14] This is exactly what we would have expected since the GPCA first segments the data into two planes. The points on the ghost line can be merged with the plane by some simple post-processing.

chosen as the coordinates for the set of new data points that are segmented by the GPCA algorithm. The pixels are then grouped according to the segmentation of the above twelve-dimensional data sets. In these experiments, the GPCA algorithm is set to run only for one level of recursion. Different choices in the error tolerance, window size, and color space (HSV or RGB) may affect the segmentation results. We adopt that RGB color space for these images.

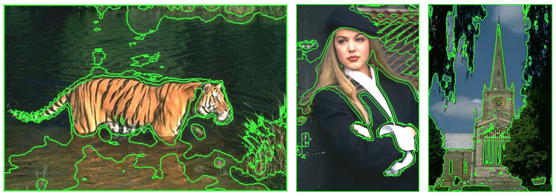

Figure 7: Image segmentation results obtained from the robust GPCA algorithm.

**Video segmentation via frame clustering.** Fig. 9 shows the results of applying our GPCA algorithm on a traffic sequence (Fig. 8) of 200 frames sampled at 3Hz. The images are first down-sampled to 64 × 48 and converted to grayscale. Each image is then treated as a point in the $64 \times 48 = 3072$ dimensional space. We first apply PCA to these points in $\mathbb{R}^{3072}$ and only the first few components are used for GPCA. For this video sequence, experiments show that four components are sufficient. The 3-D projection of these components shows that the data display a nice piece-wise smooth structure (Fig. 9 left). The GPCA algorithm segments the sequence at frames: 3, 17, 24, 83, 87, 113, 121, 124, 157 (Fig. 9 right). Interestingly, all these frames are related to car coming in or out of the scene, as shown in Fig. 8. In fact, all but one of the incoming/outgoing-car events in this video sequence are correctly detected.

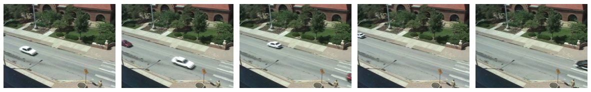

Figure 8: Key frames 3, 24, 83, 113, 157 detected in a video sequence by the GPCA algorithm, which correspond to appearing or disappearing of cars in the scene.

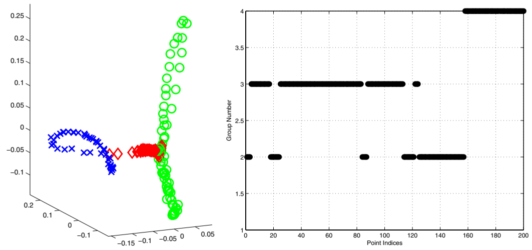

Figure 9: Segmentation of a video sequence of 200 frames. Left: The projection of the image frames to 3-D space. Each color is for a different segment as shown in Right. Right: The segmentation of the sequence of frames into three groups over time (the $x$-axis).

# 6. Conclusions

We have presented the notion of minimum effective dimension (MED) as a new model selection criteria for mixtures of subspaces. MED is a robust measure of the complexity of the mixture model that does not depend on the probabilistic model generating the data. We combined MED with Generalized PCA to propose a robust GPCA algorithm that fits an *unknown* number of subspaces of *unknown* dimensions to sample data. We presented various applications of the algorithm in motion/image segmentation, and video segmentation.

# References

[1] H. Akaike. A new look at the statistical model selection. *IEEE Transactions on Automatic Control*, 16(6):716–723, 1977.

[2] T.E. Boult and L.G. Brown. Factorization-based segmentation of motions. In *Proc. of the IEEE Workshop on Motion Understanding*, pages 179–186, 1991.

[3] M. Collins, S. Dasgupta, and R. Schapire. A generalization of principal component analysis to the exponential family. In *Neural Information Processing Systems*, volume 14, 2001.

[4] J. Costeira and T. Kanade. A multibody factorization method for independently moving objects. *International Journal of Computer Vision*, 29(3):159–179, 1998.

[5] M. Effros and P.A. Chou. Weighted universal transform coding: Universal image compression with the Karhunen-Loeve transform. In *ICIP*, volume II, pages 61–64, 1995.

[6] M.H. Hansen and B. Yu. Model selection and the principle of minimum description length. *Journal of American Statistical Association*, 96:746–774, 2001.

[7] K. Kanatani. Motion segmentation by subspace separation and model selection. In *ICCV*, volume 2, pages 586–91, 2001.

[8] J.J. Rissanen. Modeling by shortest data description. *Automatica*, 14:465–471, 1978.

[9] B. Scholkopf, A. Smola, and K.-R. Muller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10:1299–1319, 1998.

[10] M. Tipping and C. Bishop. Mixtures of probabilistic principal component analyzers. *Neural Computation*, 11(2), 1999.

[11] P. H. S. Torr. Geometric motion segmentation and model selection. *Phil. Trans. Royal Society of London*, 356(1740):1321–1340, 1998.

[12] R. Vidal. Generalized Principal Component Analysis (GPCA): an Algebraic Geometric Approach to Subspace Clustering and Motion Segmentation. Ph.D. Dissertation, UC Berkeley, 2003.

[13] R. Vidal and Y. Ma. A unified algebraic approach to 2-D and 3-D motion segmentation. In *ECCV*, 2004.

[14] R. Vidal, Y. Ma, and J. Piazzi. A new GPCA algorithm for clustering subspaces by fitting, differentiating and dividing polynomials. In *CVPR*, 2004.

[15] R. Vidal, Y. Ma, and S. Sastry. Generalized Principal Component Analysis (GPCA). In *CVPR*, 2003.

[16] C.S. Wallace and D.M. Boulton. An information measure for classification. *The Computer Journal*, 11:185–194, 1968.