

**Algebraic-Geometric and Probabilistic Approaches for
Clustering and Dimension Reduction of Mixtures of
Principle Component Subspaces**

**ECE842
Course Project Report**

Changfang Zhu

Dec. 14, 2004

Algebraic-Geometric and Probabilistic Approach for Clustering and Dimension Reduction of Mixtures of Principle Component Subspaces

**ECE842
Course Project Report**

Changfang Zhu

Abstract

Generalized Principal Component Analysis (GPCA) and Probabilistic Principal Component Analysis (PPCA) are two extensions of PCA approaches to the mixtures of principal subspaces. GPCA is an algebraic geometric framework in which the collection of linear subspaces is represented by a set of homogeneous polynomials whose degree corresponds to the number of subspaces and whose factors (roots) encode the subspace parameter. PPCA is a probabilistic approach where the principal component analysis is viewed as a maximum-likelihood procedure based on a probability density model of the observed data. Both techniques are capable of estimating a mixture of subspaces from sample data points, thus useful for data clustering and dimension reduction problems in multivariate data mining. The primary goal of this project is to carry out a conceptual study, to explore the principles and features of the algebraic-geometrical and probabilistic approaches to mixtures of principal component subspaces, and learn from hand-on experience through computational implementation of these techniques. A polynomial factorization algorithm (PFA) for GPCA and an expectation-maximization (EM) for PPCA were implemented using MATLAB codes. The implemented algorithms have been tested on synthetic data sets. It was shown that the PFA algorithm for GPCA can successfully identify the number of subspaces in the mixture, and estimate the normal vectors of the subspaces, if successful, with a relative high correlation. However, the implemented algorithm is not robust as it is data dependent. The potential problems of this implementation were discussed in the report. The implemented EM algorithm for PPCA showed that a probabilistic mixture model can identify the clusters, and assign the cluster association of each data point correctly. Both techniques estimated the component subspaces of lower dimensionality, thus data dimensions can be reduced and underlying clusters can be recovered. In this project, the implemented algorithms were only tested on synthetic 3-dimensional data and not yet tested on higher dimensional data or real data, and the algorithms are far from comprehensive for practical use. However, the computational implementation helped a lot in the understanding of the two approaches for mixtures of principal component subspaces.

1. Introduction

In the analysis of multivariate (multi-dimensional) data sets, group segmentation and cluster information often reveals insight that is useful in knowledge discovery from the complex data set, which are often high in dimension, multi-modal, and lack of prior knowledge. Clustering decomposition may enable the use of relatively simple models for each of the local clustering structures, offering great ease of interpretation as well as the benefits of analytical and computational simplification [1]. On the other hand, although it is now possible to analyze large amounts of high-dimensional data through the use of high-performance computers, in general, however, several problems occur when the number of dimensions becomes high. These problems include the explosion of execution time, difficulty in the selection of explanatory variables [2]. Therefore, data clustering and dimension reduction are important problems in multivariate data mining.

Data clustering and dimensional reduction are correlated to each other. Usually not all the data are useful for producing a desired clustering, i.e. some features may be redundant, and some may be irrelevant. Many clustering algorithms fail when dealing with high dimensional data. In this case, identifying and retaining only those features that are most relevant to the desired clustering would facilitate the multi-dimensional data analysis. If the data clusters can be visualized in a lower dimensional subspace, it will allow better interpretation and less computational command.

Principle Component Analysis (PCA) [2][3] is a very popular method used in dimension reduction, data visualization and exploratory data analysis. The idea is that a d -dimensional data set can be reduced into a set of q -dimensional data using q linear combination of bases of d -dimension. The linear combination is considered as linear projection or linear transformation. The original d -dimensional feature space is transformed to a new q -dimensional ($q < d$) feature subspace. The new feature spaces are called principal component subspace. The advantage of PCA is twofold: 1) the original data is represented by fewer variables with minimal mean square error, which reduces the dimensionality of the data set; 2) the transformation maximized the separation of data clusters. However, one of the limitations of PCA is that it only defines a single global projection of the data. For more complex data, different clusters may require different projection directions. The other limitation of PCA is that the original data should have a linear or near-linear structure, to ensure the singularity of the data matrix. If the data have a non-linear structure, the linear PCA may not be adequate in exploring the data.

Many extensions of PCA have been developed to determine the principal subspace. In this project, we studied the two extensions of PCA to the mixtures of subspaces: Generalized Principal Component Analysis (GPCA) [4][5] and Probabilistic Principal Analysis (PPCA) [6][7]. Generalized principal component analysis is an algebraic-geometric approach, which has been proposed in the computer vision community, primarily in the context of 3-D motion segmentation. Extensive work on GPCA has been carried out by Vidal, et al. [5] and two algorithms, the polynomial factorization algorithm (PFA) and the polynomial differentiation algorithm (PDA) have been proposed. Probabilistic principal component analysis is understood in a probabilistic formulation of

PCA from a Gaussian latent variable model, which is closely related to statistical factor analysis [6]. The primary goal of this project is to explore the principles and features of the algebraic-geometrical and probabilistic approaches for clustering and dimension reduction of mixtures of principal component subspaces, and learn from hand-on experience through computational implementation of these techniques. The polynomial factorization algorithm (PFA) for GPCA and an expectation-maximization (EM) algorithm for PPCA were implemented in MATLAB code.

2. Geometric approach to mixtures of principal component subspaces: GPCA

2.1. Principles of GPCA

In the generalized principal component analysis, the sample data points $\{\mathbf{x}^j \text{ in } \mathbf{R}^K\}$, $j = 1, 2, \dots, N$, are drawn from n k -dimensional linear subspace of \mathbf{R}^K , $\{S_i\}$, $i = 1, \dots, n$. The problem is to identify each subspace without knowing which sample points belong to which subspace. The union of these n linear subspaces of \mathbf{R}^K can be viewed as corresponding to the projective algebraic set defined by one or more homogeneous polynomials of degree n in K variables. Hence, estimating a collection of subspace is equivalent to estimating the algebraic variety defined by such a set of polynomials.

In the case when the subspace has dimensionality of $k = K - 1$. Vidal, et al. [4][5] has shown that the union of n such subspace is defined by a unique homogeneous polynomial $p_n(\mathbf{x})$. The degree of $p_n(\mathbf{x})$ is then the number of hyperplanes n and each one of the n factors of $p_n(\mathbf{x})$ corresponds to each one of the n hyperplanes. Therefore the problem of identifying a collection of hyperplanes is reduced to estimating and factoring $p_n(\mathbf{x})$. Since every sample point \mathbf{x} in \mathbf{R}^K must lie on one of the subspaces, S_i , every \mathbf{x} must also satisfy $p_n(\mathbf{x}) = 0$. Then one can retrieve $p_n(\mathbf{x})$ directly from the given data samples without knowing the segmentation of the data point. Vidal [5] also showed that in fact the number n of subspaces is exactly the lowest degree of $p_n(\mathbf{x})$ such that $p_n(\mathbf{x}) = 0$ for all sample points. This leads to a simple matrix rank condition which determines the number of hyperplanes n . Given n , the polynomial is determined from the solution of a set of linear equations. Given $p_n(\mathbf{x})$, the estimation of the hyperplanes is essentially equivalent to factoring $p_n(\mathbf{x})$ into a product of n linear factors.

2.2. Representing mixtures of subspace as algebraic sets and varieties

One of the important concept underlying GPCA problem is representing the mixture of subspace as algebraic sets and varieties. Noticed that every $(K-1)$ -dimensional space S_i in \mathbf{R}^K can be represented by a nonzero normal vector \mathbf{b}_i in \mathbf{R}^K as: $S_i = \{\mathbf{x} \text{ in } \mathbf{R}^K: \mathbf{b}_i^T \mathbf{x} = 0\}$. Since the subspaces S_i are all distinct from each other, the normal vectors $\{\mathbf{b}_i\}$, $i = 1 \dots n$, are pairwise linearly independent. Given that every sample point \mathbf{x} in \mathbf{R}^K lying on one of the subspaces S_i , such a point satisfies the formula:

$$(\mathbf{b}_1^T \mathbf{x} = 0) \cup (\mathbf{b}_2^T \mathbf{x} = 0) \cup (\mathbf{b}_3^T \mathbf{x} = 0) \dots \cup (\mathbf{b}_n^T \mathbf{x} = 0),$$

which is equivalent to the following homogeneous polynomial of degree n in \mathbf{x} with real coefficients:

$$p_n(\mathbf{x}) = \prod_{i=1}^n (\mathbf{b}_i^T \mathbf{x}) = 0$$

This nonlinear equation is the multiplication of n linear equations in x_i (or 1 order multivariate polynomial), and can be expressed in a linear formula as:

$$p_n(x) = v_n(x)^T \mathbf{c} = \sum c_{n_1, n_2, \dots, n_K} x_1^{n_1} x_2^{n_2} \dots x_K^{n_K} = 0,$$

where $v_n : [x_1, \dots, x_K]^T \mapsto [\dots, x_1^{n_1} x_2^{n_2} \dots x_K^{n_K}, \dots]^T$ is called Veronese map and the item of $x_1^{n_1} x_2^{n_2} \dots x_K^{n_K}$ is a monomial with n_1, n_2, \dots, n_K chosen from the degree-lexicographic order. The coefficients of the form c_{n_1, n_2, \dots, n_K} are functions of the entries of $\{\mathbf{b}_i\}$, $i = 1 \dots n$. The problem of GPCA is then to recover $\{\mathbf{b}_i\}$, given the coefficients of \mathbf{c} of the polynomial $p_n(\mathbf{x})$.

The nonlinear Veronese map maps the original data $\{\mathbf{x}^j\}$ $j = 1, 2, \dots, N$ with dimension of K into an embedded data space with higher dimension of M_n ($M_n = \binom{n+K-1}{K-1} = \binom{n+K-1}{n}$), which is very similar to the commonly used kernel approach. But the merit is that it transforms the nonlinear equation of $p_n(\mathbf{x})$ into a linear equation on the vectors of coefficients \mathbf{c} . When the number of subspace is unknown, it can be determined from the rank of the Veronese map matrix \mathbf{L}_n of the form $[v_n(x^1)^T, v_n(x^2)^T, \dots, v_n(x^N)^T]^T$. The monomials $x_1^{n_1} x_2^{n_2} \dots x_K^{n_K}$ can be calculated from the given data samples, then solving for \mathbf{c} is actually a problem of solving a set of N linear equations, where N is the total number of sample points. The remaining problems is to factorize the polynomial $p_n(\mathbf{x})$ with coefficients of \mathbf{c} to find the entries of $\{\mathbf{b}_i\}$, $i = 1, \dots, n$. Each factor will give an estimation of a subspace (hyperplane).

2.3. Polynomial factorization algorithm (PFA) for GPCA

Vidal, et al. described the polynomial factorization algorithm for GPCA in detail [4]. In this project, the algorithm for the case in the absence of noise and each subspace has dimension of $k = K - 1$ has been implemented. The algorithm implemented in this project is summarized as following:

Given sample points $\{\mathbf{x}^j\}$ $j = 1, 2, \dots, N$ lying on a collection of hyperplanes $\{S_i$ in $\mathbf{R}^K\}$, $i = 1, \dots, n$, find the number of hyperplanes n and the normal vector to each hyperplane $\{\mathbf{b}_i$ in $\mathbf{R}^K\}$, $i = 1, \dots, n$ as follows:

- 1) Apply the Veronese map of order i , for $i = 1, 2, \dots$, to the vectors $\{\mathbf{x}^j\}$ $j = 1, 2, \dots, N$ and form the matrix \mathbf{L}_i . Calculate the rank of each obtained \mathbf{L}_i . When $\text{rank}(\mathbf{L}_i) = M_i - 1$, stop the Veronese mapping, and the number of hyperplanes n is set to be current i . Then solve for \mathbf{c}_n from $\mathbf{L}_n \mathbf{c}_n = 0$ and normalize so that $\|\mathbf{c}_n\| = 1$.
- 2) Get the coefficients of the univariate polynomial $q_n(t)$ from the last $n + 1$ entries of \mathbf{c}_n .

- 3) If the first l ($0 \leq l \leq n$) coefficients of $q_n(t)$ are equal to zero, set $(b_{iK-1}, b_{iK}) = (0, 1)$ for $i = 1, \dots, l$. Then solve an n order polynomial equation $q_n(t) = 0$, and set $(b_{iK-1}, b_{iK}) = (1, -t_j)$ for $j = l+1, \dots, n$ from the $n - l$ roots of $q_n(t)$.
- 4) If all the coefficients of $q_n(t)$ are zero, just set $(b_{iK-1}, b_{iK}) = (0, 0)$ for $i = 1, \dots, n$.
- 5) After obtaining (b_{iK-1}, b_{iK}) for $i = 1, \dots, n$, solve for $\{b_{ij}\}$ $i = 1, \dots, n$ for $J = K - 2, \dots, 1$, by solving a linear system.

A practical PFA algorithm will have to consider the cases such as (1) the dimension of subspace k is smaller than $K - 1$ ($k < K - 1$); (2) degenerate cases in which vectors $(b_{rJ+1}, b_{rJ+2}, \dots, b_{rK})$ are not pairwise linearly independent; and (3) presence of noise. However, these were not explored in this course project.

3. Probabilistic approach to mixtures of principal component subspaces: PPCA

3.1. Principles of PPCA

Conventional PCA seeks a q -dimensional ($q < d$) linear projection that best represents the data in a least-square sense. For a given data set D of observed d -dimensional vector $D = \{t_n\}$, $n = 1, \dots, N$, the sample covariance matrix \mathbf{S} is first calculated, which is used for Singular Value Decomposition (SVD) or eigen-analysis to find a set of eigenvalues and corresponding eigenvectors. Then, the q dominant eigenvectors \mathbf{u}_j can be used to faithfully represent the original data with minimal loss of information, and provide the q principal projection axes. The projected data \mathbf{x}_n is given by $\mathbf{x}_n = \mathbf{U}_q^T (t_n - \mu)$, where $\mathbf{U}_q = [\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_q]$. This is a linear projection and it maximizes the variance in the projected space.

Probabilistic PCA defines a probability model [6][7], where the observations \mathbf{t} is defined as a linear transformation of a latent variable \mathbf{x} with probability distribution of $p(\mathbf{x})$, with additional noise e : $\mathbf{t} = \mathbf{W}\mathbf{x} + \mu + e$. \mathbf{W} is a $d \times q$ linear transformation matrix, μ is a d -dimensional vector that allows \mathbf{t} to have a non-zero mean. In most studies, \mathbf{x} and e are assumed to have Gaussian distribution $p(\mathbf{x}) \sim \mathcal{N}(0, \mathbf{I}_q)$ and $p(e) \sim \mathcal{N}(0, s^2 \mathbf{I}_d)$. Then the distribution of \mathbf{t} is also Gaussian of the distribution $p(\mathbf{t}) \sim \mathcal{N}(\mu, \mathbf{W}\mathbf{W}^T + s^2 \mathbf{I}_d)$.

Given the above probabilistic model of the data, one can always compute the maximum-likelihood estimator for the parameters μ , s^2 and \mathbf{W} from the data samples D , and the maximum-likelihood estimates of these parameters are:

$$\mathbf{m}_{ML} = \frac{1}{N} \sum_{n=1}^N t_n$$

$$s_{ML}^2 = \frac{1}{d - q} \sum_{i=q+1}^d \lambda_i$$

$$\mathbf{W}_{ML} = \mathbf{U}_q (\lambda_{q+1} - s_{ML}^2)^{1/2} \mathbf{R}$$

where $\lambda_{q+1}, \dots, \lambda_d$ are the smallest eigenvalues of the sample covariance matrix \mathbf{S} , the q columns in the $d \times q$ orthogonal matrix \mathbf{U}_q are the q dominant eigenvectors of \mathbf{S} , diagonal

matrix \mathbf{Q} contains the corresponding q largest eigenvalues, and \mathbf{R} is an arbitrary $q \times q$ orthogonal matrix. To simplify the problem, \mathbf{R} can be chosen as identity matrix \mathbf{I} .

3.2. Mixture of PPCA

Usually data can be generated from a mixture of components of different probability density. In the clustering using finite mixture models, each component density function $p(\mathbf{t}|i)$ represents a cluster. With the probabilistic model defined in PPCA, one can model each mixture component as a single PPCA. The observed data then has a probabilistic distribution, and the probability density of the observed data is modeled as the weighted sum of a number of Gaussian distributions and expressed as:

$$p(\mathbf{t}) = \sum_{i=1}^{k_0} p_i p(\mathbf{t} | \boldsymbol{\mu}_i, s_i^2, \mathbf{W}_i),$$

where $p(\mathbf{t} | \boldsymbol{\mu}_i, s_i^2, \mathbf{W}_i)$ denotes a PPCA density function for component i , k_0 is the total number of components, and p_i is the mixing proportion (weight) of the mixture components i (subject to the constraints: $p_i \geq 0$ and $\sum(p_i, i = 1, \dots, k_0) = 1$). Therefore, the maximum-likelihood estimation of the model parameters should maximize the log-likelihood of the observed data, which is given by:

$$L = \sum_{n=1}^N \log(p(\mathbf{t}_n)) = \sum_{n=1}^N \log\left\{ \sum_{i=1}^{k_0} p_i p(\mathbf{t}_n | \boldsymbol{\mu}_i, s_i^2, \mathbf{W}_i) \right\}.$$

Using a Expectation-Maximization (EM) algorithm [7][8], we can compute the maximum-likelihood estimation for parameters p_i , $\boldsymbol{\mu}_i$, s_i^2 and \mathbf{W}_i , recursively. This then gives the mixture components and the mixing weight of each component in the mixture.

Once the model parameters are determined, the linear relation between observation and model components given by $\mathbf{t} = \mathbf{W}\mathbf{x} + \boldsymbol{\mu} + e$ is completely defined. Then the observed data \mathbf{t} can be projected into \mathbf{x} space, as $\mathbf{x}_{ni} = z_{ni} \mathbf{W}_i^T (\mathbf{t}_n - \boldsymbol{\mu}_i)$, which is a q -dimensional reduced representation of i^{th} -cluster focused vector \mathbf{t}_n . Plot the vector \mathbf{x}_{ni} will create a i^{th} -cluster focused projection in i -subspace, and z_{ni} gives the proportion of contribution the point \mathbf{t}_n has to the i -subspace.

3.3. EM algorithm for mixture of PPCA

Expectation-Maximization (EM) refers to an iterative optimization method to estimate some unknown parameters \mathbf{T} , given measurement data \mathbf{U} [8]. In the mixture of PPCA problem, we want to estimate the set of $\{p_i, \boldsymbol{\mu}_i, s_i^2, \mathbf{W}_i\}$, $i = 1, \dots, k_0$, using the observed data D . So EM would be an idea method to solve the problem.

The schematic summary of the algorithm is as follows:

- 1) Initialization: In this step, the initial estimate of the parameters $\{p_{i0}, \boldsymbol{\mu}_{i0}, s_{i0}^2, \mathbf{W}_{i0}\}$ are randomly selected.
- 2) Using EM to compute the estimation of parameters that maximizes the log-likelihood of the observed data D .
- 3) For $k = 1, 2, \dots$

E-step: Using the current estimation of parameters, calculate the posterior probability (R_{ni}) of data \mathbf{t}_n belonging to the i^{th} component given by:

$$R_{n,i,k} = \frac{\mathbf{p}_i p(t_n | \mathbf{m}_{i,k}, \mathbf{s}_{i,k}^2, \mathbf{W}_{i,k})}{p(t_n)}, i = 1, \dots, k_0, n = 1, \dots, N$$

M-step: Using the posterior probability obtained from E-step, calculate the new estimation of parameters as following:

$$\mathbf{p}_{i,k+1} = \frac{1}{N} \sum_{n=1}^N R_{n,i,k}$$

$$\mathbf{m}_{i,k+1} = \frac{\sum_{n=1}^N R_{n,i,k} t_n}{\sum_{n=1}^N R_{n,i,k}}$$

Then using the new estimation of $\mu_{i,k+1}$, $i = 1, \dots, k_0$, compute the weighted sample covariance matrices as:

$$\mathbf{S}_i = \frac{\sum R_{n,i,k} (t_n - \mathbf{m}_{i,k+1})(t_n - \mathbf{m}_{i,k+1})^T}{\sum_{n=1}^N R_{n,i,k}}$$

then compute the eigenvalues and eigenvectors of \mathbf{S}_i , and update the estimate of s_i^2 and \mathbf{W}_i as:

$$\mathbf{s}_{i,k+1}^2 = \frac{1}{d_i - q_i} \sum_{j=q_i+1}^{d_i} \mathbf{I}_j$$

$$\mathbf{W}_{i,k+1} = \mathbf{U}_{q_i} (\mathbf{I}_{q_i} - s_{i,k+1}^2 \mathbf{I}_{q_i})^{1/2}$$

- 4) When iteration completes, calculate i^{th} -cluster focused projection in i -subspace, \mathbf{x}_{ni} , of each sample \mathbf{t}_n : $\mathbf{x}_{ni} = R_{ni} \mathbf{W}_i^T (\mathbf{t}_n - \mu_i)$

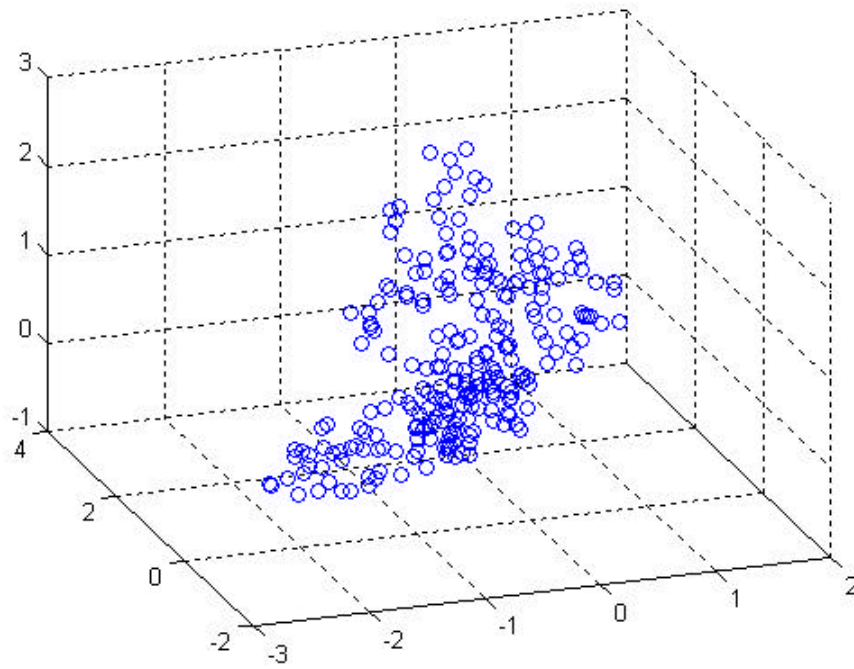
4. Computational experiments

In the computational experiments, we 1) implemented the polynomial factorization algorithm (PFA) for GPCA and an expectation-maximization (EM) algorithm for PPCA in MATLAB codes; and 2) validated the capability of these methods in discovering the clusters in the subspaces.

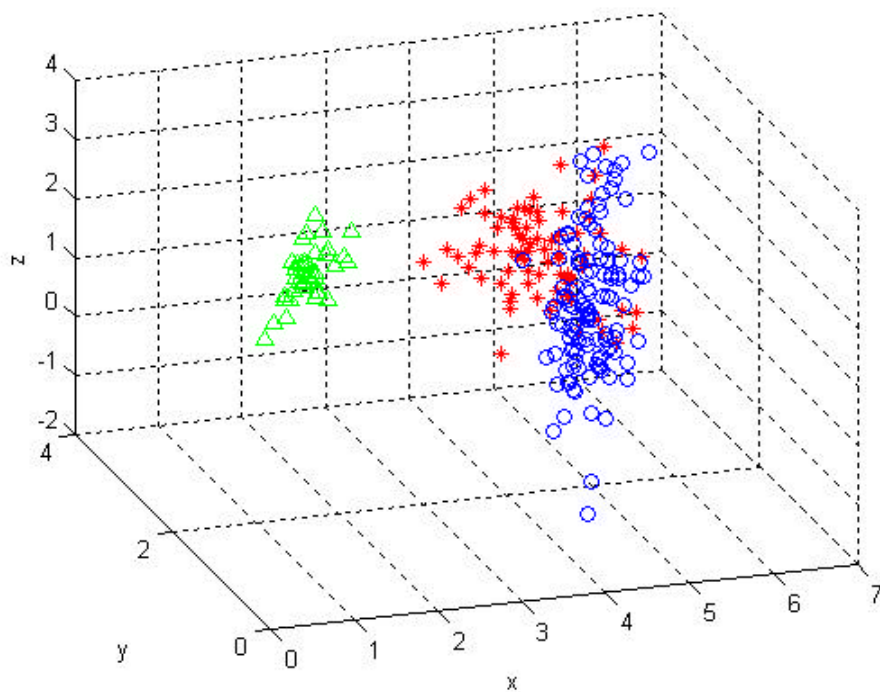
4.1. Synthetic data sets

The implemented algorithms were tested on a simple synthetic data set. Figure 1(a) shows the data set consisting of 240 3-dimensional data points generated for the GPCA test (referred to as Set 1). The data were generated from a linear combination of 3 2-dimensional linear subspaces. Each subspace is represented by a randomly selected normal vector. In order to test whether the algorithm can identify the number of subspaces correctly, data were generated from linear combination of randomly selected $n = 2, 3, 4, 6$ subspaces. In all the cases tested in this study, no noise is added to the generated data. Figure 1(b) displays the data set generated for PPCA test (referred to as Set 2). The data set consists of 240 data points generated from a mixture of three

Gaussians in 3-dimensional space. Two of the clusters are closely spaced and the third is well separated from the first two.



(a)



(b)

Figure 1 (a) Synthetic data set for GPCA test. Data were generated from a combination of 4 linear subspaces; (b) Synthetic data set for PPCA test. Data were generated from a mixture of 3 Gaussians.

4.2. Applying GPCA to data Set 1

The implemented PFA algorithm for GPCA was applied to the synthetic data set 1. It showed that for all the cases with $n = 2,3,4,6$, the algorithm can find the number of subspaces correctly. However, finding the normal vector of each subspace is not a trivial work. The difficulties may come from two facts: 1) the algorithm involves solving for the roots of polynomial equations. It is likely that for some cases, complex roots are obtained; and 2) the algorithm involves solving multivariate linear systems, i.e. solving for \mathbf{x} in $\mathbf{Ax} = \mathbf{b}$, the successful estimation of the normal vector components then depends on the condition number of matrix \mathbf{A} . When the matrix is ill-conditioned, we could obtain an incorrect solution to \mathbf{x} . If the randomly generated data does not impose these ill-conditioned problems to the GPCA procedure, the normal vector of each subspace can be estimated. As an example, the 4 randomly selected normal vectors $\{\mathbf{b}_i\}$ $i = 1,2,3,4$, of the subspaces from which data Set 1 generates are:

$$\begin{array}{cccc} 0 & 0 & 1 & 0 \\ 1 & -1 & -1 & 1 \\ 0 & 1 & 1 & 1 \end{array}$$

and the estimated normal vectors $\{\hat{\mathbf{b}}_i\}$ are:

$$\begin{array}{cccc} -0.0000 & -0.2041 & -1.0000 & 0.0000 \\ 1.0000 & 1.0000 & 1.0000 & 1.0000 \\ 1.0000 & -1.0000 & -1.0000 & -0.0000 \end{array}$$

Note the estimated normal vectors are not in the same order of the actual normal vectors, and the normal vectors can be different with a factor of (-1).

Table 1 listed the correlations (corr) between the actual normal vector $\{\mathbf{b}_i\}$ and the estimated normal vector $\{\hat{\mathbf{b}}_i\}$ in 5 successful estimations of subspaces, for the four cases with the total number of subspaces $n = 2,3,4,6$, respectively. The average and standard deviations of the absolute values of correlations were also listed in the table. The correlation between the actual normal vector $\{\mathbf{b}_i\}$ and the estimated normal vector $\{\hat{\mathbf{b}}_i\}$ is calculated as

$$\text{corr} = \frac{1}{n} \sum_{i=1}^n \mathbf{b}_i^T \hat{\mathbf{b}}_i$$

A minus sign indicates the estimated normal vector is in the opposite direction (or symmetric about the origin) relative to the actual normal vector.

Table 1 Correlation (corr) between the actual normal vector $\{\mathbf{b}_i\}$ and the estimated normal vector $\{\hat{\mathbf{b}}_i\}$ in 5 successful estimations of subspaces, for the four cases with the total number of subspaces $n = 2, 3, 4, 6$, respectively.

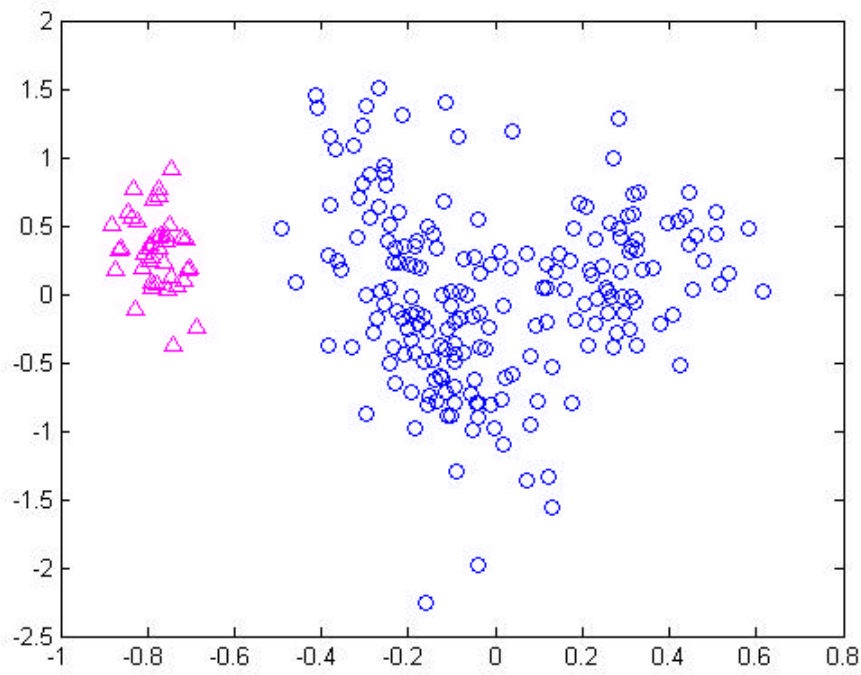
	$n = 2$	$n = 3$	$n = 4$	$n = 6$
1	0.809	0.6762	-0.9837	0.7027
2	-0.5	-0.7721	0.5625	-0.6408
3	0.5693	-0.7294	0.4949	0.4839
4	-0.75	0.7603	-0.9599	0.6872
5	1	-0.9801	0.501	0.8333
AVG (corr)	0.72566	0.78362	0.7004	0.66958
STD (corr)	0.198854	0.115931	0.249302	0.126013

Experiment on the synthetic data showed that the algorithm implemented here can successfully identify the number of subspaces in the mixture, and also estimate the normal vectors of the subspaces, if successful, with a relative high correlation (~ 0.7 in this study). Once the normal vectors of the subspaces are determined, the original data can be represented in the lower dimensional subspaces, and further analysis can be carried out on each subspace separately. However, the implementation is yet not robust, since it does depend on the randomly generated data.

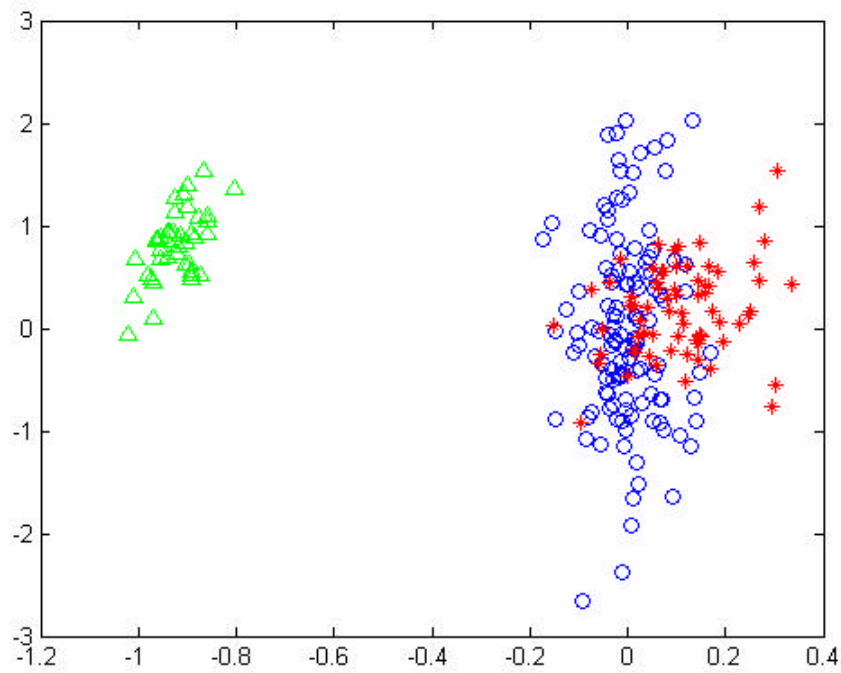
4.3. Applying PPCA to Set 2

The generated data Set 2 is a mixture of three Gaussians, with two clusters closely placed and one cluster placed separately. We applied the EM algorithm to this data set, first assuming there are only two clusters (subspaces), and then assuming there are three clusters (subspaces). Figure 2 shows the projected data in \mathbf{x} -space for the cases (a) assuming 2 clusters; and (b) assuming 3 clusters. Different colors and markers are used to indicate the group association of each data point to the subspaces. It is shown that the probabilistic mixture model can find out the clusters, and assign the cluster association of each data point correctly. Also the original data \mathbf{t} can be reduced to a 2-dimensional data set \mathbf{x} .

In this computational experiment, we have assumed the number of subspaces. However, this information usually is unknown and cannot be assumed arbitrarily. In a practical unsupervised cluster decomposition, it would be desirable to select the structural parameter k_0 of the model automatically and correctly. Wang, et al [1] proposed using two information theoretic criteria, i.e. the Akaike information criterion (AIC) and minimum description length criterion (MDL), to guide the model selection. This allows an optimal model to be selected from several competing model candidates such that the selected model best fits the observed data D . This technique is not implemented in this project.



(a)



(b)

Figure 2 Projected data in \mathbf{x} -space of the observations \mathbf{t} for the cases (a) assuming 2 clusters; and (b) assuming 3 clusters.

5. Discussions

In this project, we explored the principles and features of the algebraic-geometrical (GPCA) and probabilistic approaches (PPCA) for clustering and dimension reduction of mixtures of principal component subspaces, and implemented these two techniques in MATLAB codes for hand-on experience.

In the absence of noise, the GPCA can be casted in an algebraic geometric framework in which the collection of subspaces is represented by a set of homogeneous polynomials whose degree n corresponds to the number of subspaces and whose factors (roots) encode the subspace parameter [5]. The number of subspaces can be determined from the rank condition of the Veronese map matrix of the original data, and the estimation of the hyperplanes is equivalent to factoring the polynomial of degree n into a product of n linear factors. The polynomial factorization algorithm (PFA) proposed by Vidal et al. [4][5] is implemented in the project. There is another algorithm also proposed by Vidal [5], which is called polynomial differentiation algorithm (PDA). The PDA algorithm is designed for subspaces of arbitrary dimensions and obtains a basis for each subspace by evaluating the derivative of the set of polynomials representing the subspaces at a collection of n points in each one of the subspaces. Vidal et al. have shown that PDA algorithm gives about half of the error of the PFA algorithm, and also improves the performance of iterative techniques, such as K-subspace and EM, by about 50% with respect to random initialization. However, this algorithm was not implemented in this study.

The experiment on the synthetic data shows that the PFA algorithm implemented in this study can successfully identify the number of subspaces in the mixture, and estimate the normal vectors of the subspaces, if successful, with a relative high correlation (~ 0.7 in this study). Once the normal vectors of the subspaces are determined, the original data can be represented in the lower dimensional subspaces, and further analysis can be carried out on each subspace separately. However, the implementation is yet not robust since it is data dependent. This may be due to two facts: 1) the algorithm involves solving for the roots of polynomial equations. It is likely that for some cases, complex roots are obtained; and 2) the algorithm involves solving multivariate linear systems, i.e. solving for \mathbf{x} in $\mathbf{Ax} = \mathbf{b}$, the successful estimation of the normal vector components then depends on the condition number of matrix \mathbf{A} . When the matrix is ill-conditioned, we could obtain an incorrect solution to \mathbf{x} .

In PPCA, the principal component analysis is viewed as a maximum-likelihood procedure based on a probability density model of the observed data. The probability model is Gaussian, and determination of model parameters only requires the computing of the eigenvectors and eigenvalues of the sample covariance matrix. A mixture model of PPCA is considered when multiple clusters (subspaces) present. In this case, an EM algorithm is used to find the principal subspaces by iteratively maximizing the likelihood function.

The EM algorithm is implemented and tested on synthetic data set in the study. It is shown that the probabilistic mixture model can find out the clusters, and assign the cluster association of each data point correctly. The PPCA approach however, has some disadvantages [5]: 1) It is hard to analyze the existence and uniqueness of a solution to the problem; 2) The approach is restricted to certain classes of distributions or independence assumptions; and 3) The convergence of EM is in general very sensitive to initialization, thus there is no guarantee that it will converge to the optimal solution.

As a conceptual study, the PPCA decomposition implemented in this study is only completed on a single level. Many groups [1][7] have extended the mixture of PPCA models into a hierarchical mixture model. In their method, each PPCA component i in the lower level can be extended to a group $g_{ij} j = 1, \dots, J$ of PPCA components in the next higher level. The EM algorithm can be applied again to the decomposition in the higher level. In this way, the multiple clustered can be separated recursively to generate a hierarchy of mixtures of PPCA with a number of levels. This hierarchical model will allow the clusters to be visualized in different perceptual level, thus is very useful in multi-dimensional data visualization.

The GPCA and PPCA are two different views of the mixtures of principal components. It is not easy to compare these two methods directly, but both techniques have the capability of identifying clusters and subspaces, so that the original data can be represented in the subspaces with lower dimensionality. They can be applied in a variety of estimation problems, such as 3-D motion segmentation in computer vision, and dimension reduction problems such as data compression and feature extraction. In this project, the implemented algorithms were only tested on synthetic 3-dimensional data and not yet tested on higher dimensional data or real data, and the algorithms are far from comprehensive for practical use. However, the computational implementation helped a lot in the understanding of the two approaches for mixtures of principal component subspaces.

6. References

[1] Y. Wang, L. Luo, M.T. Freedman and S.Y. Kung, Probabilistic Principle Component Subspaces: a Hierarchical Finite Mixture Model for Data Visualization, IEEE Transactions on Neural networks, pp. 625-636, Vol.11, No.3, May 2000

[2] M. Mizuta, Dimension Reduction Methods, <http://www.quantlet.com/mdstat/scripts/csa/html/node156.html>

[3] R.A. Johnson and D.W. Wichern, Applied Multivariate Statistical Analysis, pp.xiii, 594, Prentice-Hall, Englewood Cliffs, N.J. (1982)

[4] R. Vidal, Y. Ma and S. Sastry, Generalized Principle Component Analysis (GPCA), 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '03), pp. 621-628, vol.1, June 18-20, 2003, Madison WI

[5] R. Vidal, Generalized Principal Component Analysis (GPCA): an Algebraic Geometric Approach to Subspace Clustering and Motion Segmentation, PhD thesis, University of California at Berkeley, 2003

[6] M.E. Tipping and C.M. Bishop, Probabilistic Principal Component Analysis, Technical Report NCRG/97/010, Neural Computing Research Group, Aston University, September 1997.

[7] T. Su and J. Dy, Automated Hierarchical Mixtures of Probabilistic Principal Component Analyzers, Proceedings of the 21st International Conference on Machine Learning, Article No.98, Banff, Canada, July 04-08, 2004

[8] S. Roweis, EM algorithms for PCA and SPCA, Proceedings of the 1997 Conference on Advances in Neural Information Processing System, pp. 626-632, Denver, Colorado, 1998