

Modeling High-Dimensional Probability Distributions via Linear Manifold Clusters

Rave Harpaz^{*}, Robert Haralick

Pattern Recognition Laboratory, CS Dept.

The Graduate Center, City University of New York

365 Fifth Avenue New York, N.Y. 10016

Abstract

One of the ultimate goals of cluster analysis is not only to reveal structure but also to understand it. Most clustering methods focus only on the grouping aspect and do not provide a descriptive model with which the population underlying the data can be described or with which statistical inference such as predictions can be made. Linear manifold clustering seeks to identify groups of points that lie on lower dimensional linear manifolds. In this paper we present a non-parametric density estimation modeling technique by which data that lies in a mixture of linear manifolds can be described and with which statistical inference can be based on. The efficacy of this technique is demonstrated by a target recognition experiment, where image pixels represented by high-dimensional feature vectors are classified with an error rate close to 0.1 using a probabilistic classifier constructed from mixture models of linear manifolds.

Key words: linear manifold, modeling, density estimation.

^{*} Corresponding author. Tel. 1-212-817-8192 Fax 1-212-817-1510

Email addresses: rbharpaz@sci.brooklyn.cuny.edu (Rave Harpaz),

1 Introduction

One of the ultimate goals of clustering is not only to reveal structure but to understand or describe the underlying mechanism or “process” which generated the structure. That is, to model the population of points that may have formed the structure discovered by a clustering algorithm. Describing the population by a model (preferably probabilistic) based on the sample on which the clustering is applied permits us gain insights and to learn the most important aspects of the population. Such a model should also provide us with the capability of making predictions that are consistent with the data. The matter was stated more drastically in [1,2]: a good clustering scheme is one which provides “scientific content”, i.e., a probability model which describes the underlying population, and which statistical inference such as predictions can based on. Lacking scientific content clustering is merely a visualization tool providing different views of the data. Most clustering methods focus only on the grouping aspect of clustering and lack the ability to provide scientific content. Hence, to gain the full potential of cluster analysis, a second step consisting of a *descriptive modeling* process would be beneficial.

Descriptive models can take on many forms. One of the preferred forms of modeling is that of *probability density estimation* with which the density of the population in the vicinity of a single object is estimated. The advantages of describing the data by density estimation models is that it facilitates the ability to make statistical inferences about the population underlying the data, and the ability to devise probabilistic classification schemes for new observations. Moreover, because of the probabilistic setting in which clusters are described,

haralick@netscape.com (Robert Haralick).

statistical methods such as hypothesis testing, permutations tests, maximum likelihood estimation, etc., can be applied systematically to cluster validity problems.

Avoiding unnecessary mathematical abstraction, a *linear manifold* is simply a translated subspace, which can be visualized as a line, plane, hyperplane, etc., depending on its dimensionality. In many problem domains it is assumed that linear models are sufficient enough to describe or capture the data's inherent structure. Typical examples include linear regression, PCA, and subspace clustering, which are all special cases of linear manifold learning. Yet very few remote attempts have been made to devise clustering methods able to identify or learn **mixtures** of linear manifolds. Moreover, very often observed real data is a consequence of a process governed by a small number of factors. In the data space this is manifested by the data points lying or being located close to surfaces such as linear or non-linear manifolds whose intrinsic dimensionality is much smaller than the dimensionality of the data. *Linear manifold clustering* [3] seeks to identify groups of points that fit or are embedded in lower dimensional linear manifolds.

One of the main advantages of the linear manifold clustering paradigm is that it is applicable to a wide range of clustering applications or problem domains. This is because a linear manifold is a generalization of more common and specific cluster models. It can easily be shown that “classical” clusters (hyper-spherical/ellipsoidal in shape) such as the ones sought by the K-means algorithm [4], and subspace clusters such as those discussed in [5–7] are special cases of linear manifolds. In a recent study [8] it was shown that the so called *pattern* clusters or *biclusters* [9–11], which consist of objects with similar *behavior patterns* rather than objects with similar values, are also instances of

linear manifolds. Similarly, it was shown that common to all forms of linear correlation and linear dependencies, is that in the data space they manifest themselves as lines, planes, and generally speaking as linear manifolds [12,13]. Hence, the detection of linear manifolds is a means by which correlations or linear dependencies may also be identified. From a modeling perspective this makes the modeling method presented in this paper appeal to a wide range of clustering schemes and applications.

In the following we present a nonparametric mixture modeling technique that estimates the density of a population of points that lie in lower dimensional linear manifolds. Needless to say the method assumes that the data can indeed be described by linear manifolds. The method also assumes that the partitioning of points into linear manifold clusters is readily available, or can be obtained by any clustering algorithm such as the one proposed in [3] that is able to produce linear manifold clusters. The remainder of the paper is organized as follows: in section 2 we present a formal stochastic linear manifold cluster model, which describes a set of points that fit a linear manifold. Based on this model in section 3 we present a nonparametric linear manifold mixture modeling technique. In section 4 we discuss a Bayesian probabilistic classification scheme that is based on this modeling technique. Finally, in section 5 the efficacy of our modeling technique is demonstrated by a target recognition experiment, where image pixels represented by high-dimensional feature vectors are classified with high accuracy.

2 The Linear Manifold Cluster Model

Fundamental to our method is the model of a linear manifold cluster. Let C be a set of d -dimensional points that fit a k -dimensional linear manifold cluster, where $k < d$. Further, let $\mathbf{b}_1, \dots, \mathbf{b}_d$ be a set of orthonormal vectors that span a d -dimensional space, B be a $d \times k$ matrix whose k columns are a subset of the vectors $\mathbf{b}_1, \dots, \mathbf{b}_d$, and \overline{B} be a $d \times d - k$ matrix whose columns are the remaining vectors.

Definition 1 (Linear Manifold Cluster Model) *Let $\boldsymbol{\mu}$ be some point in \mathbb{R}^d , $\boldsymbol{\lambda}$ be a $k \times 1$ random vector, and $\boldsymbol{\psi}$ be a zero mean $d - k \times 1$ random vector with much smaller variance than that of $\boldsymbol{\lambda}$. Then each point $\mathbf{x} \in C$ is modeled by,*

$$\mathbf{x} = \boldsymbol{\mu} + B\boldsymbol{\lambda} + \overline{B}\boldsymbol{\psi}. \tag{1}$$

Explanation: the idea is that each point in a cluster lies close to a k -dimensional linear manifold of finite extent, defined by the model parameters $\boldsymbol{\mu}$ - a translation vector and the matrix B containing the vectors spanning the manifold. On the manifold the points are assumed to distributed according to some pdf represented by the random vector $\boldsymbol{\lambda}$. Since in reality the points will rarely perfectly fit a linear manifold, the third component of eq. (1) models a small random error associated with each point on the manifold. The idea is that each point may be perturbed in directions that are orthogonal to the manifold defined by the columns of \overline{B} . We model this behavior by requiring that $\boldsymbol{\psi}$ be a random vector distributed according to some pdf with mean zero and small variance relative to the variance of points on the manifold, otherwise the cluster can no longer be characterized as a linear manifold or in signal

processing jargon the “signal” cannot be distinguished from the “noise”.

3 Linear Manifold Mixture Modeling by Density Estimation

Assuming the clustering produced K clusters C_1, C_2, \dots, C_K all following the linear manifold cluster model of eq. (1). Then a mixture density for a point \mathbf{x} is defined as

$$p(\mathbf{x}) = \sum_{i=1}^K p(\mathbf{x}|C_i)P(C_i), \tag{2}$$

where $p(\mathbf{x}|C_i)$ is the density of \mathbf{x} assuming it came from cluster C_i , and $P(C_i)$ is the probability that \mathbf{x} arose from cluster C_i . One way to describe this density model is that each point \mathbf{x} is generated by first randomly picking a linear manifold model or cluster C_i with probability $P(C_i)$ and then drawing the point from the corresponding probability distribution $p(\mathbf{x}|C_i)$.

According to eq. (2) to estimate the density of a point \mathbf{x} we need separate estimates or density models for each of the components of the mixture, i.e., we need to devise density and probability models for each $p(\mathbf{x}|C_i)$ and $P(C_i)$. $P(C_i)$ is estimated by simply computing the fraction of points coming from cluster C_i , that is,

$$P(C_i) = \frac{|C_i|}{\sum_{i=1}^K |C_i|}. \tag{3}$$

According to the linear manifold cluster model an estimate of the probability density function $p(\mathbf{x}|C_i)$ is essentially an estimate of the joint distribution of the random vectors $\boldsymbol{\lambda}$ and $\boldsymbol{\psi}$, representing the distribution of points on the manifold and away from the manifold.

At this point certain assumptions need to be made which create a wide range of modeling possibilities. At one extreme a parametric approach can be taken

where particular functional forms (e.g., Gaussian, Uniform) are assumed in order to model the distributions of points on and off the manifold. Then the problem of modeling the density of the points reduces to the problem of estimating the parameters of the parametric distributions, e.g., estimating a mean and covariance matrix. At the other extreme is a nonparametric approach where the density estimates are data-driven, and where a conditional dependency between the model components is assumed. In this case more elaborate and computationally demanding density estimation methods such as *nearest neighbor estimation* [14], *kernel estimation* [14], *graphical models* [15], or *multi-dimensional histograms* need to be employed. The choice of the method employed depends on a combination of several factors such as the level of model complexity desired, the level of accuracy desired, the computational resources available, and on the final goals of the experiment.

In the following we discuss a nonparametric and computationally efficient approach that is mid-way between the two extremes and that has empirically shown to be effective and accurate enough to model the data. The nonparametric method proposed makes two assumptions that simplify the modeling process and allows the modeling of a joint distribution as a product of marginals. First, it is assumed that the variation of points within and off a manifold are independent of each other. This is a common assumption in many qualitative models, which in signal processing jargon is referred to as the independence of “signal” and “noise”. Second, it is assumed that the distribution of points on each of the vectors spanning the manifold is independent of the other. To support this assumption we note that these vectors are obtained by a “decorrelating” process (PCA, discussed next). And although uncorrelatedness does imply independence from a theoretical standpoint, in practice it can be used

as evidence to support it and to approximate a joint distribution of a set of uncorrelated variables by the product of their marginals. Of course in the case of Gaussian variables uncorrelatedness is in fact equivalent to independence. Alternatively, one can use independent component analysis (instead or as an extra step) to obtain a set of independent (distribution-wise) vectors.

In order to obtain density estimates, we first need to estimate the linear manifold parameters $\boldsymbol{\mu}_i$ and B_i for each linear manifold cluster C_i . $\boldsymbol{\mu}_i$ is estimated by the sample mean of a cluster, i.e.,

$$\boldsymbol{\mu}_i = \frac{1}{|C_i|} \sum_{\mathbf{x} \in C_i} \mathbf{x}. \quad (4)$$

The vectors spanning the manifold (columns vectors of matrix B_i) are obtained through *principle component analysis* (PCA). The aim of PCA can be described in several ways, one of which is to find by an orthonormal transformation a new set of variables (components) in decreasing order of importance that explain the maximum amount of variance in the data. As mentioned, this transformation will result in a set of uncorrelated variables. According to the linear manifold cluster model, the components having the largest amount of variance are the vectors spanning the manifold, i.e., the column vectors of B_i , and the components having the least amount of variance are the vectors spanning the space orthogonal to the manifold, i.e. the column vectors of \bar{B}_i . Hence, estimating the column vectors of B_i is done by finding a set of principle components (vectors) that explain most of the variance in a linear manifold cluster, a problem which can be formulated as an eigenvalue problem. If k_i is the intrinsic dimensionality of linear manifold cluster C_i , then by PCA the k_i vectors spanning the manifold are the k_i eigenvectors $\mathbf{v}_{i1}, \mathbf{v}_{i2}, \dots, \mathbf{v}_{ik_i}$ corresponding to the leading (largest) k_i eigenvalues $\lambda_{i1}, \lambda_{i2}, \dots, \lambda_{ik_i}$ obtained by

an eigen-decomposition of the covariance matrix Σ_i of cluster C_i , where

$$\Sigma_i = \frac{1}{|C_i|} \sum_{\mathbf{x} \in C_i} (\mathbf{x} - \boldsymbol{\mu}_i)(\mathbf{x} - \boldsymbol{\mu}_i)'. \quad (5)$$

However the dimensionality k_i and thus the number of eigenvectors that need to be selected is unknown and must also be estimated. This is typically done by choosing a threshold $\tau \in [0, 1]$ that specifies the total variance in the data we expect the principle components to explain. Since the eigenvalues equal the variances along each principle component, k_i is found by

$$k_i = \min \left\{ r \left| \frac{\sum_{j=1}^r \lambda_{ij}}{\sum_{j=1}^d \lambda_{ij}} \geq \tau, r \in \{1, \dots, d\} \right. \right\}, \quad (6)$$

where typical values for τ are between 0.8 and 0.9.

Having determined the dimensionality k_i and the parameters $\boldsymbol{\mu}_i$ and $B_i = (\mathbf{v}_{i1}, \mathbf{v}_{i2}, \dots, \mathbf{v}_{ik_i})$, the next step is to model the variation of points on and away from each manifold. Because of the independence assumption we chose to make, the joint distribution of points on a manifold (their projection) is modeled by the product of the marginal distributions of the projections of points onto each of the k_i vectors spanning the manifold. The projection of a point \mathbf{x} onto the j -th vector spanning linear manifold cluster C_i is given by

$$\mathbf{v}'_{ij}(\mathbf{x} - \boldsymbol{\mu}_i). \quad (7)$$

To estimate $p_{ij}(\mathbf{v}'_{ij}(\mathbf{x} - \boldsymbol{\mu}_i))$ the true pdf of the projection of points onto the j -th spanning vector of C_i we construct a histogram $h_{ij}(\mathbf{v}'_{ij}(\mathbf{x} - \boldsymbol{\mu}_i))$ of the projections. Hence, the joint distribution of the points (projections) on the manifold is estimated by

$$p_i(B'_i(\mathbf{x} - \boldsymbol{\mu}_i)) = \prod_{j=1}^{k_i} h_{ij}(\mathbf{v}'_{ij}(\mathbf{x} - \boldsymbol{\mu}_i)). \quad (8)$$

The distribution of the points off the manifold are essentially describing the “fuzziness” of the manifolds, which can be modeled in a rather simple way. Instead of modeling this behavior as a multivariate joint distribution, the fuzziness of the manifolds can be modeled by a univariate distribution of the squared distances of points to the manifold. The squared distance of a point \mathbf{x} to linear manifold cluster C_i is given by

$$\|(I - B_i B_i')(\mathbf{x} - \boldsymbol{\mu}_i)\|^2. \quad (9)$$

To estimate the true pdf of the distances of points to linear manifold cluster C_i , we again construct a histogram $h_i(\|(I - B_i B_i')(\mathbf{x} - \boldsymbol{\mu}_i)\|^2)$ of the distances of points in cluster C_i to the manifold in which they are embedded. Because of the assumption that the distribution of points on and off the manifold are independent, the total density estimate for a point \mathbf{x} given that it came from cluster C_i is therefore given by

$$p(\mathbf{x}|C_i) = \left(\prod_{j=1}^{k_i} h_{ij}(\mathbf{v}'_{ij}(\mathbf{x} - \boldsymbol{\mu}_i)) \right) h_i(\|(I - B_i B_i')(\mathbf{x} - \boldsymbol{\mu}_i)\|^2), \quad (10)$$

and therefore the total mixture density estimate of a point \mathbf{x} is given by

$$p(\mathbf{x}) = \sum_{i=1}^K \frac{|C_i|}{\sum_{j=1}^K |C_j|} \left(\prod_{j=1}^{k_i} h_{ij}(\mathbf{v}'_{ij}(\mathbf{x} - \boldsymbol{\mu}_i)) \right) h_i(\|(I - B_i B_i')(\mathbf{x} - \boldsymbol{\mu}_i)\|^2). \quad (11)$$

An illustration of the concept of linear manifold density estimation by histograms is depicted in Fig. 1. The figure shows a one-dimensional linear manifold and the density of points on and away from it. Darker colors indicate regions of higher density. On the manifold the projection of points form an approximate bimodal normal distribution as shown by the corresponding histogram. Off the manifold the distribution of distances follows an approximate chi-squared distribution with a low number of degrees of freedom as illustrated by the smaller histogram.

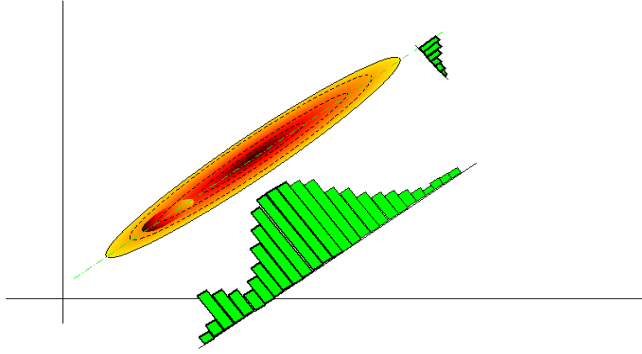


Fig. 1. Nonparametric linear manifold density estimation by histograms. Darker colors indicate regions of higher density.

4 Probabilistic Classification

Having devised density and probability models for $p(\mathbf{x}|C_i)$ and $P(C_i)$, whether by the method proposed in section 3 or any other alternative method, various probabilistic classification schemes or measures of association can be constructed. In the following we choose to use a Bayesian approach. In the Bayesian setting $p(\mathbf{x}|C_i)$ is considered to be the *likelihood* of a point \mathbf{x} given class or cluster C_i , and $P(C_i)$ the *prior* of class or cluster C_i . Then using Bayes' rule the *posterior* probability or *Bayesian measure of association* $P(C_i|\mathbf{x}) \in [0, 1]$ of a point \mathbf{x} to class or cluster C_i is given by

$$P(C_i|\mathbf{x}) = \frac{p(\mathbf{x}|C_i)P(C_i)}{\sum_{j=1}^K p(\mathbf{x}|C_j)P(C_j)}, \quad (12)$$

and the classification rule $\varphi(\mathbf{x}) \in \{1, 2, \dots, K\}$ which assigns a class or cluster label to a point can be formulated as

$$\varphi(\mathbf{x}) = \arg \max_i p(C_i|\mathbf{x}). \quad (13)$$

This rule allows for the classification of unobserved points in a probabilistic setting, providing one of the main motivations for this work. In many applications such as gene expression clustering where some genes may belong to multiple functional categories, a “hard” assignment scheme, which assigns a point to one cluster or class only may not always be adequate. A “soft” assignment allows for objects to belong to multiple clusters or classes, whereas a “fuzzy” assignment associates each object with each cluster or class by its degree of membership typically ranging in $[0, 1]$. One of the advantages of the Bayesian measure of association defined in eq. (12) is that it can easily be extended to devise soft and fuzzy assignment schemes. In fact, the fuzzy assignment measure $\varphi(\mathbf{x}) \in [0, 1]$ is exactly equal to the posterior probability, i.e.,

$$\varphi(\mathbf{x}) = p(C_i|\mathbf{x}). \quad (14)$$

To perform soft assignments a threshold θ must be prespecified, and the soft assignment rule $\varphi(\mathbf{x}) \in \mathcal{P}(\{1, 2, \dots, K\})$ (power set) can be formulated as

$$\varphi(\mathbf{x}) = \{i|p(C_i|\mathbf{x}) \geq \theta\}, \quad (15)$$

i.e., a point is assigned to a cluster or class if its posterior probability is larger than some prespecified threshold.

5 Demonstration of the Method-Target Recognition

The aim of this experiment is twofold. One, is to demonstrate that in certain, not a few cases, high-dimensional data sets can be adequately modeled by mixtures of linear manifolds using the modeling technique presented in this paper. Two, is to give an elaborate real example of how a probabilistic classifier

can be constructed using this modeling technique.

The data sets used in this demonstration consist of E3D (DARPA’s “Exploitation of 3D Data”) images of resolution 75mm and 200mm that were converted by a mathematical morphology technique into 120-dimensional feature vectors, where each feature vector corresponds to a pixel in the image. The goal was to build a classifier that would classify pixels in each of the images as either *target* (military vehicles) or *clutter*. The underlying assumption was that the data in each of the two classes lies on a set of lower dimensional linear manifolds, and would better be described using a mixture of models and not a single model. For each resolution two data sets were produced: target and clutter using the class labels which were known a priori. The target sets contained 136066 75mm feature vectors and 21260 200mm feature vectors. The clutter sets contained 229612 75mm and 76748 200mm feature vectors. The classifier was trained using *2-fold cross-validation*. That is, the data sets were randomly partitioned into two sets of equal size, where one set was used for training the classifier and the other for testing its performance. The training and testing sets were then swapped, and the results were averaged.

The classifier was constructed by first clustering each of the data sets using the linear manifold clustering algorithm presented in ref. [3], and then modeling the data sets using the modeling technique discussed in section 3. The result of this process yielded two mixture density models $p(\mathbf{x}|T)$ and $p(\mathbf{x}|C)$ (corresponding to $p(\mathbf{x})$ of eq. (11), and representing the likelihood that a point (pixel) came from the target and clutter classes respectively. In addition to these two likelihood models the priors $P(T)$ and $P(C)$ were estimated from the size of each of the data sets. Then using Bayes’ rule the posterior probabilities

of the target and clutter classes were computed by

$$P(T|\mathbf{x}) = \frac{p(\mathbf{x}|T) P(T)}{p(\mathbf{x}|T) P(T) + p(\mathbf{x}|C) P(C)} \quad \text{and} \quad P(C|\mathbf{x}) = 1 - P(T|\mathbf{x}). \quad (16)$$

The final classification rule $\varphi(\mathbf{x}) \in \{T, C\}$ was defined by

$$\varphi(\mathbf{x}) = \begin{cases} T & \text{if } \frac{P(T|\mathbf{x})}{P(C|\mathbf{x})} \geq \theta \\ C & \text{otherwise,} \end{cases} \quad (17)$$

where θ is *discrimination threshold* with which a trade off between selectivity and sensitivity (true positives and false positives) is obtained. The value for θ that optimizes a certain criteria such as minimizing the overall classification error rate was obtained through *Receiver Operating Characteristics* (ROC) analysis [14]. ROC analysis examines the performance of a classifier as a function of the *false-alarm* and *mis-detect* rates by varying a discrimination threshold. The results of the analysis are typically illustrated by an *ROC curve*; a plot of the false-alarm rates versus the mis-detect rates for varying values of θ , where a smaller area under the curve generally indicates a better classifier. Having no particular criteria defined in advance the threshold is typically selected as the point that is closest to the origin in the ROC curve, i.e. a threshold that minimizes the false-alarm and mis-detect rates simultaneously without giving preference to any one of them. We constructed the ROC curve as follows: for each feature vector \mathbf{x}_i in the testing set we computed the ratio $\theta_i = \frac{P(T|\mathbf{x}_i)}{P(C|\mathbf{x}_i)}$ with which a set of discrimination thresholds $\Theta = \{\theta_1, \theta_2, \dots, \theta_N\}$ was defined. Then for each $\theta_i \in \Theta$ we applied the classification rule $\varphi(\mathbf{x})$ and computed the following contingency table

		assigned	
		target	clutter
true	target	w	x
	clutter	y	z

were w is the number of target points that were correctly assigned target by $\varphi(\mathbf{x})$ using θ_i , x is the number of target points that were incorrectly assigned clutter, y is the number of clutter points that were incorrectly assigned target, and z is the number of clutter points that were correctly assigned clutter. Using this table we then computed for each $\theta_i \in \Theta$ the probabilities of false-alarm and mis-detect defined as

$$P(\text{false-alarm}) = \frac{y}{y+z} \quad \text{and} \quad P(\text{misdetect}) = \frac{x}{w+x}, \quad (18)$$

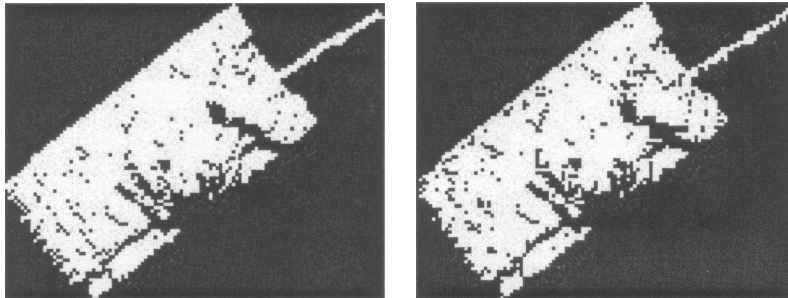
and plotted the corresponding probabilities in the ROC plot.

The final results of this experiment are illustrated by Figs. 2 and 3. Fig. 2 shows a sample target (tank) binary image before (ground-truth of a test sample) and after classification. Fig. 3 shows the ROC curves generated by the classifier. The curves indicate a generally good performance of the classifier, measured by the area under the curve, where at the points closest to the origin the average overall classification error defined as

$$\text{classification error} = \frac{x+y}{w+x+y+z} \quad (19)$$

is very close to 0.1 for 75mm data sets and 0.17 for the 200mm data sets. These results demonstrate not only that the data can be sufficiently modeled by mixtures of linear manifold, but also the efficacy of the modeling tech-

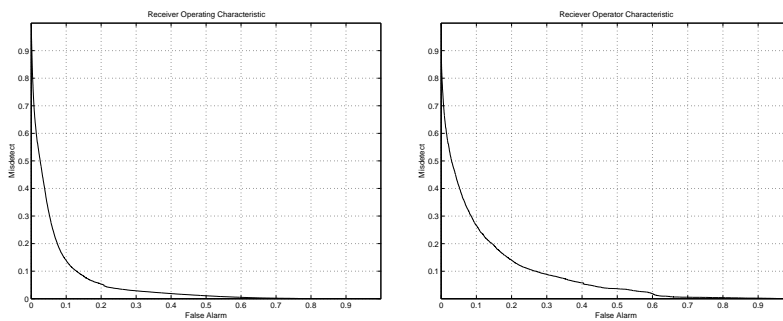
nique presented in this paper. Due to space limitations we refer the reader to ref. [3,8,13] for more examples that demonstrate the applicability of linear manifold mixture models.



(a) ground-truth

(b) after classification

Fig. 2. 75mm binary images of a target tank test sample.



(a) 75mm E3D

(b) 200mm E3D

Fig. 3. ROC curves of the E3D data sets generated by the classifier.

References

- [1] Marcel Brun, Chao Sima, Jianping Hua, James Lowey, Brent Carroll, Edward Suh, and Edward R. Dougherty. Model-based evaluation of clustering validation measures. *Pattern Recognition*, 40(3):807–824, 2007.
- [2] Edward R. Dougherty and Marcel Brun. A probabilistic theory of clustering. *Pattern Recognition*, 37:917–925, 2004.

- [3] Robert Haralick and Rave Harpaz. Linear manifold clustering in high dimensional spaces by stochastic search. *Pattern Recognition*, doi:10.1016/j.patcog.2007.01.020, 2007.
- [4] J. B. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297. Berkeley, University of California Press, 1967.
- [5] Rakesh Agrawal, Johannes Gehrke, Dimitrios Gunopulos, and Prabhakar Raghavan. Automatic subspace clustering of high dimensional data for data mining applications. In *Proceedings of the international conference on Management of data*, pages 94–105, 1998.
- [6] Charu C. Aggarwal, Joel L. Wolf, Philip S. Yu, Cecilia Procopiuc, and Jong Soo Park. Fast algorithms for projected clustering. In *Proceedings of the international conference on Management of data*, pages 61–72, 1999.
- [7] Charu C. Aggarwal and Philip S. Yu. Finding generalized projected clusters in high dimensional spaces. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, pages 70–81, 2000.
- [8] Rave Harpaz and Robert M. Haralick. Exploiting the geometry of gene expression patterns for unsupervised learning. In *Proceedings of the 18th International Conference on Pattern Recognition*, volume 2, pages 670–674, 2006.
- [9] Y. Cheng and G. Church. Biclustering of expression data. In *International Conference on Intelligent Systems for Molecular Biology*, pages 93–103, 2000.
- [10] Jiong Yang, Wei Wang, and Haixun Wang Philip Yu. δ -clusters: Capturing subspace correlation in a large data set. In *Proceedings of the 18th International Conference on Data Engineering*, pages 517–528, 2002.

- [11] Haixun Wang, Wei Wang, Jiong Yang, and Philip S. Yu. Clustering by pattern similarity in large data sets. In *Proceedings of the International Conference on Management of Data*, pages 394–405, 2002.
- [12] Rave Harpaz and Robert M. Haralick. Mining subspace correlations. In *Proceedings of the 1st IEEE Symposium on Computational Intelligence and Data Mining*, 2007. To appear.
- [13] Elke Aichtert, Christian Böhm, Hans-Peter Kriegel, Peer Körger, and Arthur Zimek. Deriving quantitative models for correlation clusters. In *Proceedings of the 12th international conference on Knowledge discovery and data mining*, pages 4–13, 2006.
- [14] Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern Classification, Second Edition*. Wiley, 2000.
- [15] David Edwards. *Introduction to Graphical Modelling*. Springer, 2000.