LINEAR DIMENSIONALITY REDUCTION WITH GAUSIAN MIXTURE MODELS

Jose M. Leiva-Murillo and Antonio Artés-Rodríguez

Universidad Carlos III de Madrid
Dept. Signal Theory and Communications
Leganés (Madrid), Spain

ABSTRACT

In this paper, we explore the application of several information-theoretic criteria to the problem of reducing the dimension in pattern recognition. We consider the use of Gaussian mixture models for estimating the distribution of the data. Three algorithms are proposed for linear feature extraction by the maximization of the mutual information, the likelihood or the hypotheses test, respectively. The experiments show that the proposed methods outperform the classical methods based on parametric Gaussian models, and avoid the intense computational complexity of nonparametric kernel density estimators.

***Index Terms*—** Information Theoretic Learning, Feature Extraction, Pattern Recognition.

1. INTRODUCTION

The need of dimensionality reduction in classification problems has been studied from a learning theory point of view. The Kolmogorov's theorem, in a neural network context, suggests that the higher the dimension of the data, the easier the pattern separation [1]. However, the Vapnik's bound from the statistical learning theory establishes that the generalization ability of classifiers gets worse as the rate between the dimension of the data and the number of samples increases [2]. This is why a proper dimension reduction is claimed to be useful for diminishing the error probability of classifiers. There are other reasons to reduce the dimension. First, the computational cost of training the classifiers and using them to classify new samples is reduced. Secondly, a projection in a low dimension space helps us to visualize and interpret the underlying structure of data. Finally, neurophysiological studies on humans and animals reveal the fact that the brain receives a compressed version of the data acquired by the sensory system [3]. This fact suggests that a pattern recognition process can be improved by a proper redundancy elimination via dimension reduction.

The most direct way of reducing the dimension is feature selection: a subset of the original features is selected for classification. On the other hand, feature extraction performs a transformation $\mathbf{z} = \mathbf{f}(\mathbf{x}), \mathbf{f} : \mathcal{R}^D \rightarrow \mathcal{R}^d$, with $d \leq D$. In the case that the transformation is characterized by a matrix, the feature extraction is said to be linear, i.e.: $\mathbf{z} = \mathbf{W}^T\mathbf{x}$. It is easy to note that a feature selection scheme can be described by a linear feature extractor with a given projection matrix.

The choice of using linear or nonlinear feature extraction is determined by the classifier used. If a linear classifier is used, a non linear feature extractor is appropriate to unfold the non linear patterns present in the data. If a nonlinear classifier is applied, a linear feature extractor may be utilized, on condition that the subspace on which the data are projected contains all the discriminative information. As an example of the importance of choosing an appropriate linear transformation, a Mahalanobis distance can be learned so that the performance of a K-nearest-neighbor (KNN) classifier is significantly improved [4]. In this paper, we focus on linear feature extraction.

Linear Discriminant Analysis (LDA) was the first statistical criterion for low rank linear separation, and it is still the most popular supervised linear feature extractor [5]. LDA tries to maximize the dispersion among classes while minimizing the inner dispersion of each class, which is known as Fisher criterion. LDA provides a closed, eigendecomposition-based solution to the maximum likelihood criterion in the homoscedastic case (the same covariance matrix is assumed for each of the classes). Other methods have been proposed that measure the distance among classes, via Chernoff and Bhattacharrya distances [6] [7]. However, these methods assume mono-modality and gaussianity of the data, so that these distances are easy to compute. On the other hand, classification problems determined by nonlinear discrimination boundaries are not successfully solved by these methods, so that other criteria are to be considered instead of linear distances. The need for extending LDA to more complex, multimodal distributions led to a method that makes use of homoscedastic Gaussian mixture models to estimate the distribution of each class [8].

Recently, a number of methods have been proposed that make use of criteria as likelihood or mutual information to learn the features to extract. These methods are commonly claimed to belong to the novel framework of information theo-

This work has been partly supported by Ministerio de Educación y Ciencia of Spain (project 'DOIRAS', id. TIC2003-02602), and the Comunidad de Madrid (project 'PRO-MULTIDIS-CM', id. S0505/TIC/0223).

retic learning (ITL). Although information theory was born in the forties, it has recently become popular in machine learning and neural processing systems. It provides a set of tools for analyzing the statistical dependence among random variables. This is useful in dimensionality reduction because it allows us to measure the relevance of the features extracted with respect to the classes, as well as the redundancy among features themselves. Due to their flexibility, non-parametric kernel density estimators (KDE) are frequently used for modeling the distribution of the data, to estimate the likelihood [9] or an alternative measure of mutual information between features and classes [10] [11], among other proposals.

In this paper we propose the use of semiparametric probability density function (PDF) estimation based on Gaussian mixture models (GMM) for estimating and maximazing several information theoretic criteria. These methods avoid some of the disadvantages of working with KDE, like their computational complexity or the problem of the bandwidth selection for the kernel. In addition to the usage of these models for feature extraction, we also evaluate the performance of a generative classifier based on these models. Thus, a measure of the probability that a sample belongs to each class is provided. In the next Section, Gaussian mixture models are introduced, as well as the model selection criteria for determining the number of mixtures to be used. In Section 2, several criteria are described as well as the procedures for their maximization. Some experiments are carried in Section 4 to compare the performance of the methods proposed. The paper finishes with some conclusions and remarks in Section 5.

2. GAUSSIAN MIXTURE MODELS AND MODEL SELECTION

A Gaussian mixture model is a PDF estimator given by the expression:

$$\hat{p}(\mathbf{x}) = \sum_{k=1}^{K} \alpha_k G(\mathbf{x}|\mathbf{m}_k, \mathbf{C}_k)$$

where K is the number of mixture components and $\mathbf{m}_k$ and $\mathbf{C}_k$ are, respectively, the mean and the covariance matrix of the k-th Gaussian component. Each α_k is a weight parameter, so that $\sum_k \alpha_k = 1$. The Expectation-Maximization algorithm performs a search of the parameters $\{\alpha_k, \mathbf{m}_k, \mathbf{C}_k\}_{k=1}^{K}$ by maximizing the likelihood of the data given the model.

The number of mixtures K must be carefully chosen for each class as well as the whole dataset. The higher the value of K, the higher the likelihood of the GMM obtained. However, a high value of K leads to overfitted models. A widely accepted criterion to apply the Occam's razor and so to penalize the complexity of the model is Akaike information criterion (AIC). Akaike's criterion chooses the model with the highest value of the cost given by:

$$AIC = 2\log\mathcal{L} - 2T \tag{1}$$

where $\mathcal{L}$ is the likelihood and T is the number of free parameters of the model. In the case of GMM, the set of parameters consists of a series of K scalars α_k, K mean vectors $\boldsymbol{\mu}_k$ and K covariance matrices $\mathbf{C}_k$. Because the dimension of $\mathbf{x}$ is D and each covariance matrix has $D(D+1)/2$ unique elements, the total number of parameters in the model is $T = K(D+1)(D+2)/2$. The expression in (1) can be rewritten as:

$$AIC = 2\log\mathcal{L} - K(D+1)(D+2)$$

We use this criterion for establishing the number of mixtures K in each of the models used throughout the paper.

3. CRITERIA FOR SUPERVISED FEATURE EXTRACTION

A pattern recognition problem is defined by a set of samples from a multivariate variable $\mathbf{x}$, each of which comes with a sample of an auxiliary discrete variable $y \in \{c_1, c_2, \ldots, c_L\}$ that indicates the class the sample belongs to. Linear feature extraction consists of finding the $D \times d$ projection matrix $\mathbf{W}$ such that the new variable $\mathbf{z} = \mathbf{W}^T\mathbf{x}$ belongs to a d-dimensional space. Let us denote as x-space the original D-dimensional feature space and z-space the reduced one. In a pattern recognition context, the transformation must provide a classification performance as good as possible when carried out on $\mathbf{z}$.

In this Section, different information-theoretic criteria for feature extraction are described. The methodology for the maximization of these criteria is also provided, once the distribution of the data is modeled by GMMs. These criteria are mutual information, likelihood and hypothesis test.

3.1. Mutual Information

Mutual information (MI) is, according to Shannon's Information Theory, a measure of the statistical dependence among several random variables [12]. The MI between a continuous, multidimensional variable $\mathbf{z}$ and a discrete one y may be described in terms of entropy as:

$$I(\mathbf{z}, y) = h(\mathbf{z}) - h(\mathbf{z}|y) = h(\mathbf{z}) - \sum_{l=1}^{L} P(c_l) h(\mathbf{z}|c_l)$$

where y is the auxiliary variable indicating the class of each sample of data. The entropies involved are given by the expressions:

$$h(\mathbf{z}) = -\int p(\mathbf{z}) \log p(\mathbf{z}) d\mathbf{z} \tag{2}$$

$$h(\mathbf{z}|c_l) = -\int p(\mathbf{z}|c_l) \log p(\mathbf{z}|c_l) d\mathbf{z} \tag{3}$$

Since $p(\mathbf{z})$ and each $p(\mathbf{z}|c_l)$ are unknown we model them by the GMMs $\hat{p}(\mathbf{z})$ and $\hat{p}(\mathbf{z}|c_l)$. The models have been obtained in the *x*-space with parameters $\{\alpha_k, \mathbf{m}_k, \mathbf{C}_k\}$. We need to relate them to the ones in the *z*-space $\{\alpha'_k, \mathbf{m'}_k, \mathbf{C'}_k\}$. The parameters obtained in the *x*-space hold the maximum likelihood property, but there is not a way of relating the likelihood in both the *x* and *z*-spaces. Hence, we take the reasonable transformations of the parameters $\alpha'_k = \alpha_k$, $\mathbf{m}'_k = \mathbf{W}^T\mathbf{m}_k$ and $\mathbf{C'}_k = \mathbf{W}^T\mathbf{C}_k\mathbf{W}$, because these transformations do provide maximum likelihood parameters for individual Gaussians.

Even for a simple distribution model as a GMM, the analytical computation of $h(\mathbf{z})$ is intractable unless the number of mixtures is one. Instead, we propose a sampled estimation. Given the transformed dataset $\mathbf{Z} = \{\mathbf{z}_1, \ldots, \mathbf{z}_N\}$, with $\mathbf{z}_i = \mathbf{W}^T\mathbf{x}_i$, this estimation has the form:

$$\hat{h}(\mathbf{Z}) = -\frac{1}{N}\sum_{i=1}^{N}\log\hat{p}(\mathbf{z}_i)$$

It is easy to show that $E\{\hat{h} - h\} = D_{KL}(p, \hat{p})$, being D_{KL} the Kullback-Leibler divergence. This means that there is a systematic positive bias in the estimation of h. The maximum likelihood procedure followed by the EM algorithm provides us (ignoring local minima problems) with the minimum entropy solution, i.e. the $\hat{p}$ that, among all the GMMs with the same order, minimizes $D_{KL}(p, \hat{p})$. The gradient of this entropy w.r.t. the projection matrix $\mathbf{W}$, in terms of the derivatives of the Gaussians involved, is given by:

$$\nabla_{\mathbf{W}}\hat{h}(\mathbf{z}) = -\frac{1}{N}\sum_i \frac{1}{\hat{p}(\mathbf{z}_i)}\sum_k \alpha_k \nabla_{\mathbf{W}} G(\mathbf{z}_i|\mathbf{m'}_k, \mathbf{C'}_k)$$

where the gradient of each Gaussian is:

$$\begin{aligned}\nabla_{\mathbf{W}}G(\mathbf{z}_i|\mathbf{m'}_k, \mathbf{C'}_k) &= G(\mathbf{z}_i|\mathbf{m'}_k, \mathbf{C'}_k)\times \\ \times\Big[&\big(\mathbf{C'}_k^{-1}(\mathbf{z}_i - \mathbf{m'}_k)(\mathbf{z}_i - \mathbf{m'}_k)^T - \mathbf{I}\big)\mathbf{C'}_k^{-1}\mathbf{W}^T\mathbf{C}_k - \\ &-\mathbf{C'}_k^{-1}(\mathbf{z}_i - \mathbf{m'}_k)(\mathbf{x}_i - \mathbf{m}_k)^T\Big]\end{aligned} \quad (4)$$

The derivative of the estimated MI is given by a linear combination of derivatives of the entropies, so that we can perform a gradient ascent procedure for the optimization of the cost:

$$\begin{aligned}\hat{\mathbf{W}} &= \arg\max_{\mathbf{W}} \hat{I}(\mathbf{Z}, Y) \\ &= \arg\max_{\mathbf{W}}\Big[\hat{h}(\mathbf{Z}) - \sum_{l=1}^{L} P(c_l)\hat{h}(\mathbf{Z}|c_l)\Big]\end{aligned}$$

3.2. Likelihood

The Informative Discriminant Analysis (IDA) has been proposed for linear feature extraction, by using kernel density estimation (KDE) to model the distribution of the data [9]. This method searches for the transformation $\mathbf{z} = \mathbf{W}^T\mathbf{x}$ that maximizes the likelihood:

$$\log L(Y|\mathbf{Z}) = \sum_i \log\hat{p}(y_i|\mathbf{z}_i) \quad (5)$$

The conditional density is estimated as:

$$\hat{p}(y_i|\mathbf{z}_i) = \frac{\hat{p}(\mathbf{z}_i, y_i)}{\sum_l \hat{p}(\mathbf{z}_i, c_l)} = \frac{P(y_i)\hat{p}(\mathbf{z}_i|y_i)}{\sum_l P(c_l)\hat{p}(\mathbf{z}_i, c_l)}$$

where each $\hat{p}(\mathbf{z}_i|c_l)$ is modeled by a GMM, and the $P(c_l)$ is the a-priori probability of class c_l, which can be empirically estimated as $P(c_l) = n_l/N$, i.e. the fraction of training samples belonging to that class. The method for feature extraction consists in finding the transformation matrix $\mathbf{W}$ that maximizes the likelihood in (5), i.e.:

$$\begin{aligned}\hat{\mathbf{W}} &= \arg\max_{\mathbf{W}}\sum_i \log\frac{P(y_i)\hat{p}(\mathbf{z}_i|y_i)}{\sum_l P(c_l)\hat{p}(\mathbf{z}_i, c_l)} \\ &= \arg\max_{\mathbf{W}}\sum_i[\log\hat{p}(\mathbf{z}_i, y_i) - \log\hat{p}(\mathbf{z}_i)]\end{aligned} \quad (6)$$

The proposed optimization of the cost is again given by the derivatives of the Gaussians as explained in (4).

3.3. Likelihood test

A classification problem can be stated in terms of the decision theory. The choice of the class a sample belongs to may be interpreted as the election of one of the hypotheses about the origin of the sample. This methodology has been widely utilized in digital communications, threat detection or clinical diagnosis, among others [13].

In a binary decision problem one must choose between the hypotheses $\mathcal{H}_0$ and $\mathcal{H}_1$. The decision is given by the ratio between the likelihood of the observation given the hypothesis, i.e. by the criterion:

$$\frac{p(\mathbf{z}|\mathcal{H}_1)}{p(\mathbf{z}|\mathcal{H}_0)} \underset{\mathcal{H}_1}{\overset{\mathcal{H}_0}{\lessgtr}} \lambda$$

where λ is the threshold established by some criterion as Neymann-Pearson's or Bayes' [14]. The Neymann-Pearson's lemma states that the test is optimal in the sense that no other criterion can simultaneously reduce both kinds of errors (false alarms and miss detections in the binary case), if both densities $p(\mathbf{z}|\mathcal{H}_0)$ y $p(\mathbf{z}|\mathcal{H}_1)$ are known [12].

Since $\mathbf{z}$ is obtained by the projection $\mathbf{z} = \mathbf{W}^T\mathbf{x}$, a reasonable criterion can be to search for the $\mathbf{W}$ that achieves the maximum of the test between $p(\mathbf{z}|\mathcal{H}_1)$ and $p(\mathbf{z}|\mathcal{H}_0)$ for the training data:

$$LT(\mathbf{z}, y) = \frac{p(\mathbf{z}|\mathcal{H}_1)}{p(\mathbf{z}|\mathcal{H}_0)}$$

where, in the multiclass case, $\mathcal{H}_1$ is the hypothesis that $\mathbf{z}$ belongs to the class given by its label, and $\mathcal{H}_0$ is the hypothesis

that it belongs to any of the other classes. Thus, a one-versus-the-rest learning scheme is applied. The test must be carried out from empirical likelihoods, since the densities $p(\mathbf{z}|\mathcal{H}_0)$ and $p(\mathbf{z}|\mathcal{H}_1)$ must be estimated. Again, GMMs are used to model the distributions. The test for the whole set of data can be rewritten as:

$$\log LT(\mathbf{Z}, Y) = \log \frac{\prod_i \hat{p}(\mathbf{z}_i|y_i)}{\prod_i \hat{p}(\mathbf{z}_i|\bar{y}_i)}$$

Thereby the cost function to maximize is:

$$\hat{\mathbf{W}} = \arg\max_{\mathbf{W}} \sum_i \Big[\log \hat{p}(\mathbf{z}_i|y_i) - \log \hat{p}(\mathbf{z}_i|\bar{y}_i)\Big] \qquad (7)$$

The hypothesis of the sample belonging to a given class is modeled by:

$$\hat{p}(\mathbf{z}|\bar{c}_l) = \sum_{j \neq l} \pi_{j,l} \hat{p}(\mathbf{z}|c_j)$$

where $\pi_{j,l}$ is a prior that indicates the a-priori probability that a sample belongs to c_j subject to that it does not belong to c_l. This prior can be empirically determined from the number of training samples of each class, or by any other a-priori information about the data. In the former case, we would have: $\pi_{j,l} = \frac{N_j}{N - N_l}$. The expression (7) can be rewritten as:

$$\hat{\mathbf{W}} = \arg\max_{\mathbf{W}} \sum_i \Big[\log \hat{p}(\mathbf{z}_i|y_i) - \log \sum_{c_l \neq y_i} \pi_{j,l} \hat{p}(\mathbf{z}_i|c_l)\Big]$$

Again, this cost is derivable as in the previous methods, so that it can be optimized by a gradient ascent procedure. Although a stochastic gradient optimization can be used, in the following experiments a batch-type procedure has been followed.

The complexity of the methods proposed is with $O(N \cdot L \cdot K)$, being K the maximum number of Gaussians used to build each model. This complexity is linear with N, being an advance with respect to methods based on KDE models, which have a complexity $O(N^2)$.

4. EXPERIMENTS

In this Section, we evaluate the performance of the methods proposed and compare them to other feature extraction methods. Also, we evaluate the usage of the GMMs for generative classification

4.1. Datasets and Feature Extraction Methods

The feature extractors whose performance is evaluated in this Section are the ones described in Section 3: the maximization of the mutual information (MMI), the maximization of the likelihood (ML) and the maximization of the likelihood test (LT). For comparison, we also evaluate the following feature extraction methods:

1. Principal Component Analysis (PCA) is an unsupervised method that searches for the most powerful components of the data, and projects them along such directions. It does not take into account the class of the samples.

2. Linear Discriminant Analysis (LDA) looks for the most discriminative projections in terms of first and second order statistics.

3. Maximum Quadratic Mutual Information (MQMI), proposed by K. Torkkola [11], is an ITL method that makes use of KDE to model the densities $p(\mathbf{z}|c_l)$. Shannon's MI is defined in terms of Kullback-Leibler (KL) divergence, which is analytically intractable unless the densities are modeled by simple parametric models. Instead of KL's, the divergence used in MQMI is a quadratic distance that gives place to the pseudo-MI:

$$I_Q(\mathbf{z}, y) = \sum_{c_l} \int [p(\mathbf{z}, c_l) - p(\mathbf{z})P(c_l)]^2 d\mathbf{z}$$

The datasets used in the experiments are from the public UCI repository [15]. They show different dimensionality degrees and numbers of classes, in order to evaluate the methods in a variety of pattern recognition scenarios. Their characteristics are displayed in Table 4.1. Two of the datasets, Optdigits and Isolet, have been previously applied a PCA-based dimension reduction to 40 components, since otherwise singular covariance matrices may appear in the GMMs when applying the EM algorithm.

Data	Train	Test	Dimension	Classes
Landsat	4435	2000	36	6
Optdigits	3823	1797	64 (40)	10
Letter	16000	4000	16	26
Isolet	6238	1559	617 (40)	26

Table 1. Characteristics of the evaluated datasets

4.2. Classifiers

Two classifiers have been used to measure the performance of the methods described. First, a pure discriminative, non parametric K-Nearest-Neighbors (with $K = 1$, 1NN) classifier has been used. Secondly, we propose a generative decision rule given by the GMMs in the z-space. In this case, the classification rule is:

$$\hat{y}_i = \arg\max_l \hat{p}(\mathbf{z}_i|c_l)$$

This criterion has the advantage that it provides us with (estimated) probability values. In the following, we refer to this classification rule as Gaussian mixture classification (GMC).

4.3. Results

In the Tables 2 to 5, the classification results of the methods proposed are displayed for the datasets described in Table 4.1. Several degrees of dimensionality reduction are evaluated, from 1 to 5 projections obtained. The AIC has chosen a number of mixtures for each of the classes between 2 and 3 (Landsat), 2 to 4 (Optdigits), 2 to 5 (Letter) and 1 to 5 (Isolet) . The number of mixtures for the whole set of samples is 11, 26, 12 and 21, respectively for each dataset.

The results highlight the superiority of the GMM-based ITL methods proposed in the majority of the datasets and reduction degrees considered. This superiority is specially remarkable in the datasets Optdigits and Isolet, which suggests that these pattern recognition problems are characterized by strongly non linear classification boundaries. Among the methods proposed, the maximization of the likelihood is the one that provides the best results.

The superiority of the GMC classifier with respect to the 1NN must be stressed. Although 1NN is not a state-of-the-art classifier, the fact that GMC performs better in all the cases suggests the convenience of its usage, specially in those cases in which a probability measure or soft output is required. However, in the classification experiments on the raw data, i.e. without dimension reduction, GMC performs worse than 1NN in some cases, due to the overfitting of GMMs when the dimension is high.

# Comps.	1	2	3	4	5
PCA/1NN	40.80	78.40	83.80	84.70	85.80
PCA/GMC	50.25	81.45	84.15	85.60	87.65
LDA/1NN	47.85	71.35	82.00	83.90	82.50
LDA/GMC	54.85	77.95	85.40	86.10	87.20
MQMI/1NN	52.05	69.50	83.65	83.20	84.50
MQMI/GMC	60.45	75.25	84.15	85.60	87.65
MMI/1NN	65.15	**79.80**	**83.80**	84.70	**85.85**
MMI/GMC	72.50	83.40	**85.95**	**87.25**	**87.95**
ML/1NN	**64.50**	78.45	82.85	**84.95**	84.20
ML/GMC	**72.90**	**83.80**	85.70	87.10	87.30
LT/1NN	55.80	74.85	78.85	81.10	83.00
LT/GMC	63.85	78.80	83.20	84.40	85.40
Raw Data/1NN					89.45
Raw Data/GMC					86.15

Table 2. Landsat Dataset.

In Figure 1 a scatterplot of the features extracted from the Wine dataset is displayed, when two projections are obtained. This dataset contains 178 samples, 13 dimensions and 3 classes. We visualize the results obtained by the three methods proposed and LDA, for comparison. Because of the simplicity of this classification problem, all the methods successfully separate the samples belonging to each class in the z-space.

# Comps.	1	2	3	4	5
PCA/1NN	28.32	52.81	71.68	78.80	87.81
PCA/GMC	35.84	60.16	75.57	81.69	89.15
LDA/1NN	32.22	59.04	77.30	85.31	89.87
LDA/GMC	41.68	64.22	80.97	87.59	90.87
MQMI/1NN	34.06	59.15	76.96	78.80	87.81
MQMI/GMC	44.35	65.66	80.63	81.69	89.15
MMI/1NN	37.28	66.94	82.92	90.09	91.82
MMI/GMC	**47.08**	**73.62**	85.53	90.48	92.32
ML/1NN	**37.90**	**67.84**	**83.58**	**90.32**	**93.27**
ML/GMC	46.52	**73.62**	**86.37**	**91.10**	**94.16**
LT/1NN	33.28	55.26	73.85	86.76	91.26
LT/GMC	42.74	64.05	78.69	87.98	92.10
Raw Data/1NN					95.66
Raw Data/GMC					96.88

Table 3. Optdigits Dataset.

# Comps.	1	2	3	4	5
PCA/1NN	09.05	17.15	35.68	56.15	66.50
PCA/GMC	05.85	18.48	32.70	49.43	59.23
LDA/1NN	17.37	35.90	48.58	63.38	69.73
LDA/GMC	18.25	41.98	52.73	64.00	69.27
MQMI/1NN	17.40	36.30	49.40	56.15	66.50
MQMI/GMC	18.02	42.13	52.73	49.43	59.23
MMI/1NN	18.10	**45.82**	55.75	67.20	74.72
MMI/GMC	21.40	45.45	56.43	66.87	73.55
ML/1NN	16.58	42.08	**58.45**	**71.35**	**80.00**
ML/GMC	20.50	**46.35**	**59.92**	**69.17**	**76.08**
LT/1NN	**18.47**	35.48	48.93	61.22	76.15
LT/GMC	**21.98**	39.05	53.77	62.10	71.92
Raw Data/1NN					95.65
Raw Data/GMC					85.45

Table 4. Letter Dataset.

# Comps.	1	2	3	4	5
PCA/1NN	16.23	24.82	33.55	48.49	59.33
PCA/GMC	20.40	32.20	42.21	58.18	66.97
LDA/1NN	19.82	40.28	55.93	65.62	69.34
LDA/GMC	27.33	49.01	66.07	72.80	76.20
MQMI/1NN	21.55	38.87	58.56	48.49	59.33
MQMI/GMC	25.40	48.30	67.80	58.18	66.97
MMI/1NN	**23.41**	51.38	67.16	77.74	80.76
MMI/GMC	30.60	61.71	75.05	82.75	87.04
ML/1NN	22.58	**52.92**	**69.85**	**77.81**	**83.19**
ML/GMC	31.24	**62.60**	**79.41**	**84.61**	**88.33**
LT/1NN	23.03	41.76	57.67	66.07	74.15
LT/GMC	**31.43**	50.55	65.68	73.70	80.31
Raw Data/1NN					85.31
Raw Data/GMC					90.76

Table 5. Isolet Dataset.

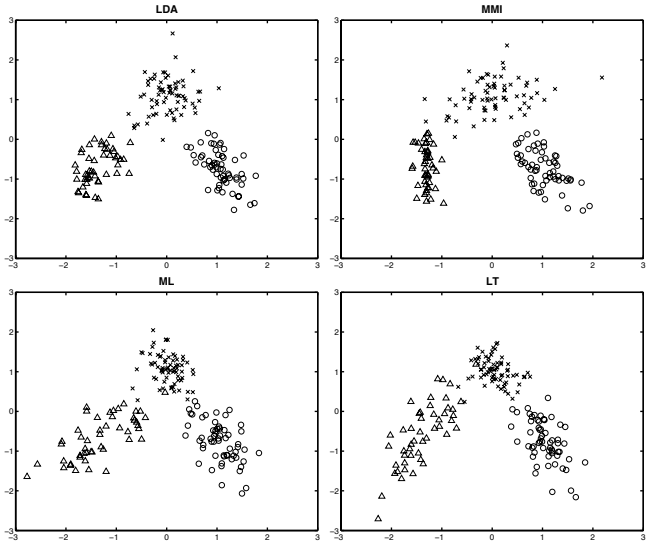

Fig. 1. Scatter plot of the three-class dataset Wine for several feature extraction methods.

5. CONCLUSIONS

We have presented three linear feature extraction methods for pattern recognition that are inspired in ITL criteria, involving concepts from information theory and decision theory. The classification experiments have revealed their better performance with respect to other classical methods as PCA or LDA and even with respect to another ITL method as Torkkola's MQMI.

The methods proposed are based on the previous modeling of the distribution of the data. To do so, Gaussian mixture models have been used. The problem of choosing the complexity of the models (i.e. the number of mixtures) has been solved by the application of Akaike's information criterion. In a set of experiments not included in the paper, another model selection criterion as Bayesian information criterion (BIC) was used, but led to less complex models that provided poor performance results. However, the theorectical study of model-selection criteria for GMM and the proposed ITL methods can be a promising area of research.

The suitability of AIC and the procedure for choosing the parameters in the z-space from the ones in the original x-space is demonstrated by two facts. First, the ability of the proposed feature extraction methods for obtaining the relevant information. Secondly, the good classification performance of the proposed GMM-based generative classifier. The choice of the best dimension reduction, i.e. the dimension of the manifold that contains the discriminative information in each classification problem represents an open problem that could be addressed in a future wok by the interpretation of the values of the likelihood and mutual information.

6. REFERENCES

[1] C. Bishop, *Neural Networks for Pattern Recognition*, Clarendon Press, Oxford, 1995.

[2] V. Vapnik, *Statistical Learning Theory*, John Wiley & Sons, New York, 1998.

[3] J. Atick, "Could information theory provide an ecological theory of sensory processing?," *Network*, vol. 3, pp. 213–251, 1992.

[4] J. Goldberger, S. Roweis, G. Hinton, and R. Salakhutdinov, "Neighbourhood components analysis," in *Advances in Neural Information Processing Systems, NIPS*, Vancouver, Canad, 2005, pp. 513–520.

[5] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, Academic Press, New York, 1990.

[6] M. Loog and R. Duin, "Linear dimensionality reduction via a heteroscedastic extension of LDA: The Chernoff criterion," *IEEE Trans. on PAMI*, vol. 26, no. 6, pp. 732–739, 2004.

[7] P. Hsieh, D. Wang, and C. Hsu, "A linear feature extraction for multiclass classification problems based on class mean and covariance discriminant information," *IEEE Trans. on PAMI*, vol. 28, no. 2, pp. 223–235, 2006.

[8] T. Hastie and R. Tibshirani, "Discriminant analysis by Gaussian mixtures," *Journal of the Royal Statistical Society series B*, vol. 58, pp. 158–176, 1996.

[9] J. Peltonen and S. Kaski, "Discriminative components of data," *IEEE Transactions on Neural Networks*, vol. 16, no. 1, pp. 68–83, 2005.

[10] J. Principe, D. Xu, and J. Fischer, *Information-Theoretic Learning*, vol. 1 of *Unsupervised Adaptive Filtering*, John Wiley & Sons, New York, 2000.

[11] K. Torkkola, "Feature extraction by non-parametric mutual information maximization," *Journal on Machine Learning Research*, vol. 3, pp. 1415–1438, 2003.

[12] T.M. Cover and J.A. Thomas, *Elements of Information Theory, 2nd Edition*, John Wiley & Sons, Hoboken, NJ, 2006.

[13] H. L. Van Trees, *Detection, Estimation, and Modulation Theory, Part I*, John Wiley & Sons, New York, 1968.

[14] S. Kay, *Fundamentals of Statistical Signal Processing. Volumen II, Detection Theory*, Prentice-Hall, New York, 1998.

[15] D.J. Newman, S. Hettich, C.L. Blake, and C.J. Merz, "UCI repository of machine learning databases," Tech. Rep., Univ. of California, Dept. ICS, 1998.