Information-Theoretic Linear Feature Extraction based on Kernel Density Estimators: A Review

José M. Leiva-Murillo, *Member, IEEE* and Antonio Artés-Rodríguez, *Senior Member, IEEE*

***Abstract*—In this paper, we provide a unified study of the application of kernel density estimators to supervised linear feature extraction by means of criteria inspired by information and detection theory. We enrich this study by the incorporation of two novel criteria to the study: the mutual information and the likelihood ratio test, and perform both a theoretical and a experimental comparison between the new methods and other ones previously described in the literature. The impact of the bandwidth selection of the density estimator in the classification performance is discussed. Some theoretical results that bound classification performance as a function or mutual information are also compiled. A set of experiments on different real-world datasets allow us to perform an empirical comparison of the methods, in terms of both accuracy and computational complexity. We show the suitability of these methods to determine the dimension of the sub-space that contains the discriminative information.**

***Index Terms*—Machine Learning, Information-Theoretic Learning, Feature Extraction, Kernel Density Estimation**

I. Introduction

Information theory (IT) has become increasingly popular in the machine learning community because it provides a set of tools to measure the redundancy among the variables involved in a problem, as well as their relevance for the prediction of an additional variable. However, working with IT measurements involves two difficulties. First, the entropy and mutual information are defined in terms of the probability distribution of the data. Thus, there exists a need for the estimation of the probability density function (PDF) $p(\mathbf{x})$ if the data are continuous. Because of this, information theoretic learning (ITL) often relies on generative modeling. However, in some cases this estimation may be avoided, such as in the case of the Infomax method for independent component analysis [1] or the maximization of mutual information for feature extraction [2]. The second difficulty arises from the fact that, even when the PDFs involved are accurately estimated, the computation of the IT magnitudes from them may be intractable. This is the case of the entropy: unless $p(\mathbf{x})$ consists of a simple parametric model, the computation of its entropy can be analytically unfeasible [3].

The problem of estimating the PDF of the data is typically addressed in ITL by means of non parametric kernel density estimators (KDE). Although other methods [4] make use of histograms, their applicability is limited by the facts that they are not smooth, and their accuracy decreases soon with the number of variables. The kernel usually considered in KDE is the Gaussian [3], [5], [6].

In this paper we perform a review of some existing ITL criteria and try to examine their similarities and differences. In a machine learning context, the redundancy among variables is defined as the degree of their statistical dependence. It is usually desired to reduce this redundancy to improve the performance of the learning task at hand or the interpretation of the data. On the other hand, this reduction should not remove information of interest contained in the data (v.g., the ability to predict the value of an auxiliary variable).

In supervised learning, the general problem is to estimate the relationship between an input $\mathbf{x}$ and an output y from a dataset $\{\mathbf{x}_i, y_i\}$, $i = 1, \ldots, L$, $\mathbf{x}_i \in \mathbb{R}^N$. In this paper, we consider classification problems, in which y is discrete, i.e. $y \in \{1, 2, \ldots, N_c\}$. In that case, y is referred to as the *class* or the *label*. We are usually interested in reducing the redundancy among the components of $\mathbf{x}$ as well as maximizing their relevance for predicting the value of y by means of a transformation $\mathbf{z} = \mathbf{f}(\mathbf{x})$. There is a number of reasons for performing feature extraction (FE) or dimensionality reduction. Both Kolmogorov's and Cover's theorems [7], [8] suggest that the higher the dimension of the data, the easier the pattern separation; 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 [9]. Also, a projection in a low dimension space helps us to visualize and interpret the underlying structure of data. Moreover, neurophysiological studies on humans and animals reveal the fact that the brain receives a compressed version of the data acquired by the sensory system [10]. This fact suggests that a pattern recognition process can be improved by a proper redundancy elimination via dimension reduction.

The FE is defined by a function $\mathbf{z} = f(\mathbf{x})$, $\mathbf{z} \in \mathbb{R}^N$ that may be linear or non-linear. The choice between one or another is conditioned by the classifier used. Hence, it is a common practice to apply either a linear FE method followed by a non-linear classifier, or a non-linear FE method before a linear classifier. In the first strategy, the responsibility of finding the non-linear separation boundaries relies on the classifier. In the second case, the feature extractor projects the data on a set of variables in which the non linear patterns are *unfolded*, and a linear discrimination function is able to separate the classes [11]. In this paper, we focus on linear FE. As an example of the potential of linear FE, it has been shown that the

The authors are with the Dept. Signal Theory and Communication, Universidad Carlos III de Madrid, Leganés (Madrid), Spain. Email: jose@tsc.uc3m, antonio@tsc.uc3m.es

This work was supported in part by the Spanish Ministry of Science and Innovation under projects CSD2008-00010 and TEC2009-14504-C02-01, and the IST Programme of the European Community, under the PASCAL2 Network of Excellence, IST-2007-216886.

performance of a simple k-nearest-neighbors (KNN) classifier can be remarkably improved by a proper linear transformation on data [12].

The main objectives of this paper are i) to discuss the impact of the bandwidth selection in the performance of methods based on KDEs; ii) to analyze the equivalence between a method previously proposed in the literature -maximum conditional likelihood- and the maximization of the mutual information, iii) to perform a compilation of theoretical results that relate mutual information with classification error; iv) to compare the classification accuracy of different ITL feature extraction methods, and vi) to study their computational complexity.

In the next Section, we describe kernel density estimators and explain why they are appropriate in information-theoretic learning. In Section III, we describe new methods for information-theoretic feature extraction, and compare them to other ones, previously proposed in the literature. In Section IV, a set of experiments are provided on real data to evaluate the classification performance on the variables obtained by the feature extraction methods, as well as an analytical and empirical comparison of their computational complexity. The paper finishes in Section V with some conclusions about the work presented.

II. Kernel Density Estimators

A Kernel Density Estimator (KDE) is a non-parametric PDF model that consists of a linear combination of kernel functions centered on the data (see, for example, [13])

$$\hat{p}_{\boldsymbol{\theta}}(\mathbf{x}) = \frac{1}{N}\sum_{i=1}^{N} k(\mathbf{x}-\mathbf{x}_i|\boldsymbol{\theta}) \tag{1}$$

where $\mathbf{x} \in \mathbb{R}^D$ and $k(\mathbf{x}|\boldsymbol{\theta})$ is the kernel function with a given bandwidth $\boldsymbol{\theta}$. These models are often used in information-theoretic learning because i) we do not need a-priori assumptions on the distribution of the data; ii) the model does not need to be trained, as it only relies on the samples, and iii) it is easy to carry out transformations on the data, such as $\mathbf{z} = \mathbf{f}(\mathbf{x})$, and estimate the PDF in the $\mathbf{z}$-space by a KDE with kernels centered in $\{\mathbf{z}_i = \mathbf{f}(\mathbf{x}_i)\}$.

Although the KDEs are commonly considered as non-parametric models, the kernel function has an adjustable bandwidth defined by $\boldsymbol{\theta}$ that determines the accuracy of the model, so that it can be treated as a parameter to be optimized.

The problem of choosing an appropriate $\boldsymbol{\theta}$ is called *the bandwidth selection problem* and has been intensively studied by the statistics community - see [14] for an exhaustive review of criteria for univariate data. The most extended criteria in bandwidth estimation are the integrated square error (ISE), the mean ISE (MISE), the asymptotic MISE (AMISE), as well as criteria based on L_1-norm [15]. In the one-dimensional case, optimizing these criteria with respect to the bandwidth is not problematic as it involves a global search on one variable which is computationally feasible. This is the main reason why multivariate bandwidth selection has been addressed only in very low dimensional spaces [16].

We are interested in a bandwidth selection that leads to an accurate estimation of $\log \hat{p}(\mathbf{x})$ rather than of $\hat{p}(\mathbf{x})$, because most of ITL methods are based on the computation of log-likelihoods rather than the evaluation of the densities.

For this reason, we make use of the maximum-likelihood leave-one-out (ML-LOO) method for bandwidth selection [17], which maximizes the likelihood, measured in each data point, of a KDE model built with the rest of points

$$\hat{p}_{\boldsymbol{\theta}}(\mathbf{x}_i) = \frac{1}{N-1}\sum_{\substack{j=1 \\ j\neq i}}^{N} G(\mathbf{x}_i-\mathbf{x}_j|\boldsymbol{\theta}). \tag{2}$$

Note that a maximum-likelihood (ML) solution is equivalent to minimum entropy, when the entropy is estimated as $\hat{h}(\mathbf{X}) = -\frac{1}{N}\sum_i \log \hat{p}(\mathbf{x}_i)$. It has been proven in the literature that if $\hat{p}(\mathbf{x})$ is a KDE, the entropy is overestimated [18]; then, a minimum entropy criterion provides the estimation which is closest to the true entropy value among those performed with KDEs. Equivalently, ML provides a bandwidth selection criterion that allows to estimate the log-likehood $\sum_i \log p(\mathbf{x})$ with the highest accuracy achievable with a KDE.

III. Supervised Feature Extraction with KDEs

In this Section, we propose new methods for dimensionality reduction in classification, and study the theoretical connection of the proposed methods with other ones previously presented in the literature. We assume in the following that only spherical kernels are used in the KDE models. There are three main reasons for this. First, the computational burden of both the bandwidth selection and the feature extraction itself is far higher in the full case than in the spherical one. Secondly, the risk of obtaining overfitted models is lower with an spherical KDE, because in this case only one parameter is to be adjusted for each of the classes. Finally, it does not make sense to pay much computational effort to accurately model the density in the $\mathbf{x}$-space if the different criteria are estimated in the $\mathbf{z}$-space.

The criteria described in the following are conditional likelihood, likelihood ratio test, (conventional) mutual information and quadratic mutual information. The conditional likelihood and the quadratic mutual information have been proposed in [6] and [5], respectively. The maximization of the likelihood ratio test and the mutual information constitute original work.

A. Maximum Conditional Likelihood (MCL)

The Informative Discriminant Analysis (IDA) was proposed for linear feature extraction, by using KDEs to model the distribution of the data [6]. Here we rename it as Maximum Conditional Likelihood for ease of interpretation. This method searches for the transformation $\mathbf{z} = \mathbf{W}^T\mathbf{x}$ that maximizes the conditional log-likelihood. Under the i.i.d. assumption, we have

$$\log L(Y|\mathbf{Z}) = \sum_{i=1}^{N} \log \hat{p}(y_i|\mathbf{z}_i). \tag{3}$$

The conditional density is estimated as

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

where each $\hat{p}(\mathbf{z}_i|c_l)$ is a KDE for class c_l. The method for feature extraction consists in finding the transformation matrix $\mathbf{W}$ that maximizes the likelihood in (3). The criterion to be maximized is then

$$\begin{aligned}\hat{\mathbf{W}}_{MCL} &= \arg\max_{\mathbf{W}} \sum_i \log \frac{\hat{p}(\mathbf{z}_i|y_i)P(y_i)}{\sum_l \hat{p}(\mathbf{z}_i|c_l)P(c_l)} \\ &= \arg\max_{\mathbf{W}} \sum_i \log \frac{\sum_{j\in I_{y_i}} G(\mathbf{z}_i - \mathbf{z}_j|\boldsymbol{\theta}_{y_i})}{\sum_{c_l\neq y_i}\sum_{j\in I_l} G(\mathbf{z}_i - \mathbf{z}_j|\boldsymbol{\theta}_l)}.\end{aligned} \tag{4}$$

where I_l is the set of size n_l of samples belonging to class c_l. We have used the empirical estimation $P(c_l) = n_l/N$ for the a-priori probability of class c_l. Since we are using a Gaussian kernel, we know that its parameter set boils down to a covariance matrix $\mathbf{C}_l$.

Now we need to relate the width in the $\mathbf{x}$-space σ_x^2 with the one in the $\mathbf{z}$-space σ_z^2. We assume that the relationship between covariance matrices under a linear transformation, $\boldsymbol{\Sigma}_z = \mathbf{W}^T\boldsymbol{\Sigma}_x\mathbf{W}$, also holds for the kernel bandwidths, $\mathbf{C}_x$ and $\mathbf{C}_z$. Then, we take into account that $\mathbf{W}$ is orthonormal and that we are assuming a spherical bandwidth for $\mathbf{x}$, i.e.,$\mathbf{C}_x = \sigma_x^2\mathbf{I}$. Hence, it follows that $\boldsymbol{\theta}_z = \sigma_z = \sigma_x$.

The proposed maximization can be performed by a gradient ascent, taking into account that the derivatives of the Gaussians are given by

$$\nabla_{\mathbf{W}} G(\mathbf{z}-\mathbf{z}_i|\sigma_z^2) = -\frac{1}{\sigma_z^2}(\mathbf{z}-\mathbf{z}_i)(\mathbf{x}-\mathbf{x}_i)^T G(\mathbf{z}-\mathbf{z}_i|\sigma_z^2). \tag{5}$$

B. Maximum Likelihood Ratio Test (MLRT)

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

$$LT(\mathbf{z}, y) = \frac{p(\mathbf{z}|\mathcal{H}_1)}{p(\mathbf{z}|\mathcal{H}_0)} \underset{\mathcal{H}_1}{\overset{\mathcal{H}_0}{\lessgtr}} \lambda \tag{6}$$

where λ is the threshold established by some criterion as Neymann-Pearson's or Bayes' [19]. 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 value of the test (6) if $\mathcal{H}_1$ is the hypothesis of *correct* classification and $\mathcal{H}_0$ is the *wrong* one. 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. The logarithmic test for the whole set of data can be rewritten as

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

The hypothesis of the sample belonging to a given class is modeled, as in the previous cases, by a KDE built from its samples. The hypothesis that the samples do not belong to the class can be expressed as a linear combination of the rest of classes

$$\hat{p}(\mathbf{z}|\bar{c}_l) = \sum_{k\neq l} \pi_{kl}\hat{p}(\mathbf{z}|c_k)$$

where π_{kl} is a prior that indicates the a-priori probability that a sample belongs to c_k subject to that it does not belong to c_l. In this case, we have: $\pi_{kl} = \frac{n_k}{N-n_l}$, being n_l the number of data points from class c_l. Thus, we can see the similarity between the cost optimized in (4) and the cost for the hypothesis test procedure, since the expression (7) can be rewritten as

$$\hat{\mathbf{W}}_{MLRT} = \arg\max_{\mathbf{W}} \sum_i \Big[\log \hat{p}(\mathbf{z}_i|y_i) - \log \sum_{c_l\neq y_i} \pi_{ly_i}\hat{p}(\mathbf{z}_i|c_l)\Big] \tag{8}$$

$$= \arg\max_{\mathbf{W}} \sum_i \log \frac{\sum_{j\in I_{y_i}} G(\mathbf{z}_i - \mathbf{z}_j|\boldsymbol{\theta}_{y_i})}{\sum_{c_l\neq y_i} \frac{1}{N-n_l}\sum_{j\in I_l} G(\mathbf{z}_i - \mathbf{z}_j|\boldsymbol{\theta}_l)}. \tag{9}$$

C. Maximum Mutual Information (MMI)

Mutual information (MI) is, according to Shannon's Information Theory, a measure of the statistical dependence among several random variables [20]. The MI between a continuous, multidimensional variable $\mathbf{z}$ and a discrete one y may be described in terms of entropy as

$$I(\mathbf{z}, y) = h(\mathbf{z}) - h(\mathbf{z}|y) = h(\mathbf{z}) - \sum_{l=1}^{L} P(c_l)h(\mathbf{z}|c_l) \tag{10}$$

where $h(\mathbf{z}) = -\int p(\mathbf{z})\log p(\mathbf{z})d\mathbf{z}$.

Because the computation of $\hat{h}(\mathbf{z}|c_l)$ is intractable on KDEs, we consider the following sample estimation, which is proven to converge to the entropy as $N \to \infty$ due to the asymptotic equipartition property [20]

$$\hat{h}(\mathbf{z}|c_l) = -\frac{1}{n_l}\sum_{i\in I_l} \log \hat{p}(\mathbf{z}_i|c_l). \tag{11}$$

If $\hat{p}(\mathbf{z})$ is modeled as a linear combination of the $\hat{p}(\mathbf{z}|c_l)$, i.e.

$$\hat{p}(\mathbf{z}) = \sum_l P(c_l)\hat{p}(\mathbf{z}|c_l) = \frac{1}{N}\sum_{i=1}^{N} G(\mathbf{z} - \mathbf{z}_i|\sigma_{y_i}^2) \tag{12}$$

then the projection matrix is given by the maximization problem

$$\begin{aligned}\hat{\mathbf{W}}_{MMI} &= \arg\max_{\mathbf{W}} \hat{I}(\mathbf{z}, y) = \arg\max_{\mathbf{W}} \Big[\hat{h}(\mathbf{z}) - \sum_{l=1}^{L} P(c_l)\hat{h}(\mathbf{z}|c_l)\Big] \\ &= \arg\max_{\mathbf{W}} \frac{1}{N}\sum_{c_l}\sum_{i\in I_l} \log \frac{\frac{1}{n_l}\sum_{j\in I_l} G(\mathbf{z}_i - \mathbf{z}_j|\sigma_l^2)}{\frac{1}{N}\sum_{c_m}\sum_{j\in I_m} G(\mathbf{z}_i - \mathbf{z}_j|\sigma_m^2)}.\end{aligned} \tag{13}$$

By simple manipulation it can be shown that the maximization problem is equivalent to (4), leading to the same solution. For this reason, in the experiments section we will consider the method MCL/MMI referring to both MCL and MMI.

The connection between MI and classification error p_e has been theoretically stated by several authors. A lower bound by Fano [20], and two upper bounds by Feder and Merhav [21] and by Hellman and Raviv [22] respectively have been defined on the error probability, as a function of the MI. In Figure 1, we have added Hellman and Raviv's bound to an example given in [23] for a multiclass classification problem. Note that the inverse proportionality between MI and p_e justifies the maximization of the MI in pattern recognition.

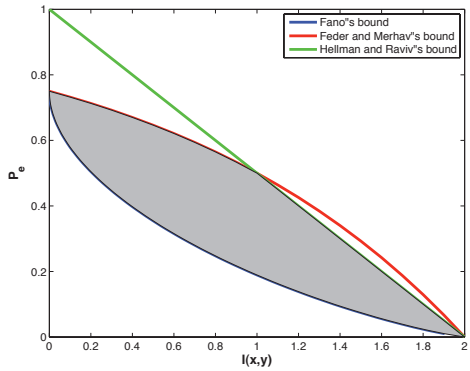

Fig. 1. Upper and lower bounds for the error probability in a 4-class classification problem.

D. Maximum Quadratic Mutual Information

An alternative to the approach introduced above for MI estimation is to avoid the use of Kullback-Leibler divergence, which appears in Shannon's definition of MI. Torkkola's Maximization of Quadratic MI (MQMI) was proposed in [5] with this aim.

MQMI consists in the maximization of the quadratic distance between $p_{ZY}(\mathbf{z},\mathbf{y})$ and $p_Z(\mathbf{z})p_Y(\mathbf{y})$, using the scalar product $\langle p,q\rangle = \int p(\mathbf{x})q(\mathbf{x})d\mathbf{x}$. Then, the quadratic pseudo-mutual information is given by

$$\begin{aligned} I_Q(\mathbf{z},y) &= D_Q\left(\hat{p}(\mathbf{z},y),\hat{p}(\mathbf{z})P(y)\right) \\ &= \sum_{l=1}^{N_c}\int_{\mathbf{z}}\hat{p}^2(\mathbf{z},c_l)d\mathbf{z} + \sum_{l=1}^{N_c}\int_{\mathbf{z}}P^2(c_l)\hat{p}^2(\mathbf{z})d\mathbf{z} \\ &\quad - 2\sum_{l=1}^{N_c}\int_{\mathbf{z}}\hat{p}(\mathbf{z},y)P(c_l)\hat{p}(\mathbf{z})d\mathbf{z}. \end{aligned} \tag{14}$$

where, in absence of additional information, the a-priori probabilities may be set to $P(c_l) = n_l/N$, being n_l the number of samples labeled with c_l and N the size of the whole dataset. The integrals in (14) can be analytically solved by convolving Gaussian kernels, since the PDFs are modeled as KDEs. The result is described in terms of interactions between pairs of data points, and it is referred to as *potential* because of its physical analogy.

IV. Experiments

In this Section, we present some experiments to evaluate the four feature extractors whose performance is evaluated in this section are the ones described in Section III. We make use of the cross-validation maximum-likelihood (ML-LOO) rule for bandwidth selection of the KDEs involved, as described in [17]. However, in order to explore the relevance of this kernel bandwidth choice, we include two versions of the MCL/MMI method, one of them based on the aforementioned method and the other one on Scott rule. For comparison, we also evaluate the performance of two classical statistical methods: Principal Component Analysis (PCA), which is unsupervised, and Linear Discriminant Analysis (LDA), which is supervised.

The maximization of I_Q in Torkkola's method is carried out by means of a stochastic gradient ascent with orthogonality constraints for $\mathbf{W}$. In order to make the optimization procedure as non-parametric as possible, and to plug our optimized bandwidth in the models, we have performed the following modifications on the original Torkkola's scheme and applied them to the experiments with the different techniques:

1) A batch-type gradient ascent is used instead of the stochastic one. This way, we avoid choosing the rate at which the step size is decreased.
2) The kernel width used in the $\mathbf{z}$-space can be different for modeling each of the classes. Thus, each of the models can be more accurately estimated.
3) A simple Gram-Schmidt orthogonalization is performed after each iteration in order to hold the ortho-normallity constraints, instead of the use of Givens rotations as in the original work. We have empirically checked that both methods perform similarly, so that Gram-Schmidt is used because of its lower computational complexity. This procedure has also been applied to the rest of the methods described above.
4) The kernel width is not modified during the optimization. In the original work, the bandwidth is shared by all the classes and it was decreased during the stochastic gradient descend in deterministic annealing fashion. The a-priori choice according to the ML-LOO criterion allows us to assume that the width is adequate at the first stage of the optimization as well as at the end.

First, we show the classification performance of the different FE methods, under different degrees of dimension reduction. Secondly, we analyze the computational complexity of these methods.

A. Classification Performance

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 Parzen models in the $\mathbf{z}$-space, i.e. $\hat{y} = \arg\max_y p(\mathbf{x}|y)$. This criterion has the advantage that it provides us with (estimated) probability values. In the following, we refer to this classification rule as Parzen classification (PC).

Dataset	N_c	D	Train	Test	Ref. Accuracy
Landsat	6	36	4435	2000	90.9
Optdigits	10	64 (40)	3823	1797	98.22
Letter	26	16	16000	4000	97.75
Isolet	26	617 (40)	6238	1559	95.06
Waveform	3	21	300	500	83.40
Segmentation	7	19 (8)	210	2100	92.62

TABLE I
CHARACTERISTICS OF THE PUBLIC DATASETS.

The characteristics of the datasets used are shown in Table I. They have been compiled from the public UCI repository [24]. They show different dimensionality degrees and numbers of classes, in order to evaluate the methods in a variety of pattern recognition scenarios. The datasets have been previously whitened, in order to make the data as spherical as possible before obtaining its spherical bandwidth. The numbers between brackets indicate the dimension after a principal component analysis is applied in order to avoid problems with singular covariance matrices. The reference classification accuracy achievable in each dataset is also provided, computed by a non-linear support vector machine with a radial-basis function as kernel.

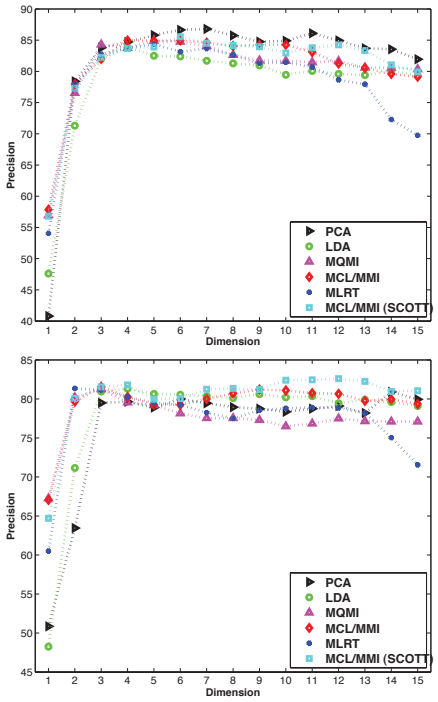

Fig. 2. Classification performance on Landsat dataset. Top: 1NN classification; bottom: Parzen classification

In the Figures 2 to 7, the classification results of the methods proposed are displayed for the different datasets.

The results highlight the superiority of the proposed ITL methods in the wide majority of the datasets and reduction degrees considered, with respect to the classical method LDA. The superiority over PCA is higher, which is expected given that PCA is an unsupervised method. An exception is however found in Landsat data (Fig. 2), due to the fact that, in this case,

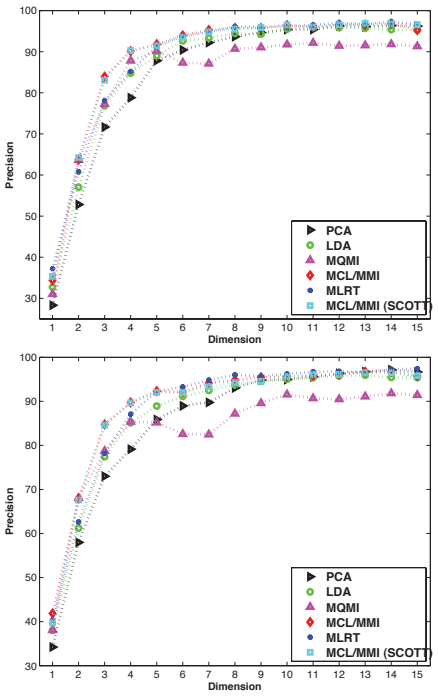

Fig. 3. Classification performance on Optdigits dataset. Top: 1NN classification; bottom: Parzen classification

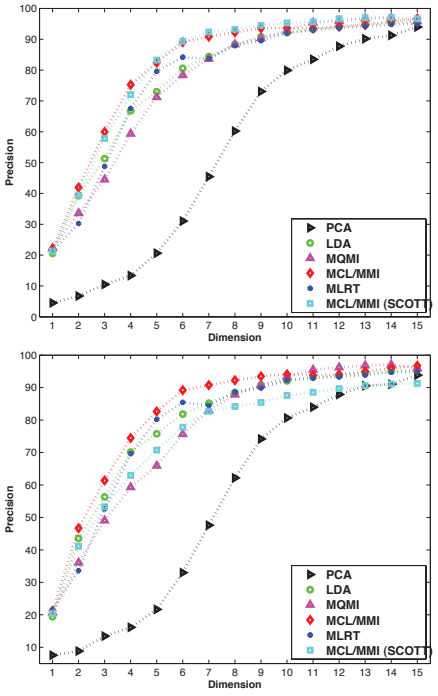

Fig. 4. Classification performance on Letter dataset. Top: 1NN classification; bottom: Parzen classification

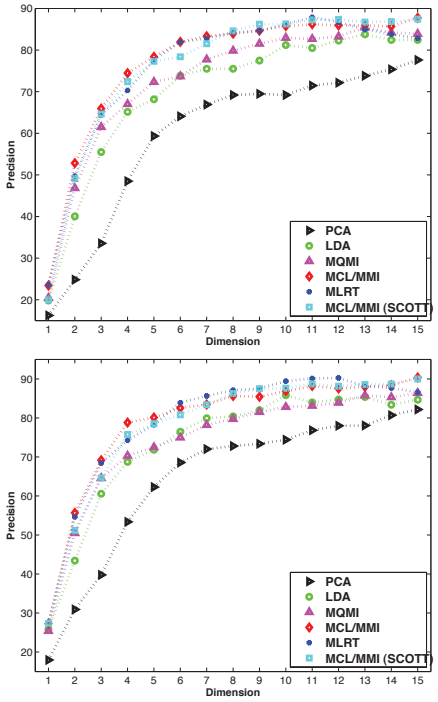

Fig. 5. Classification performance on Isolet dataset. Top: 1NN classification; bottom: Parzen classification

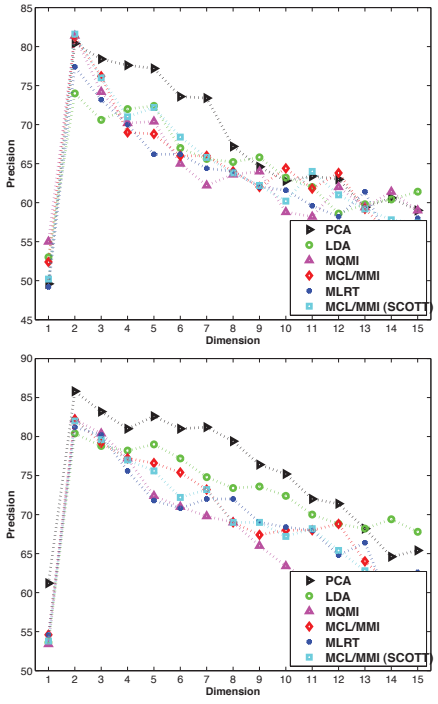

Fig. 6. Classification performance on Waveform dataset. Top: 1NN classification; bottom: Parzen classification

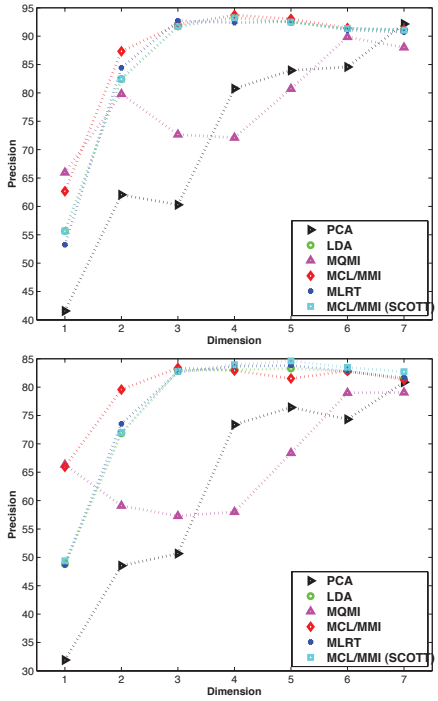

Fig. 7. Classification performance on Segmentation dataset. Top: 1NN classification; bottom: Parzen classification

projections of high energy and projections of high discriminative power are aligned. The superiority of ITL methods suggest a distribution of data far from Gaussian and strongly non-linear discrimination functions. Among the methods proposed, the maximization of the likelihood (or, equivalently the mutual information) MCL/MMI is the one that provides the best results in general.

The curves in the figures allow us to have an idea about the intrinsic dimension of the data, i.e. the dimension of the space that contains all the relevant information needed for discrimination. An extreme case can be seen in Waveform dataset; the curves suggest that this intrinsic dimension is two, since the error probability increases from there on. Because Waveform is a 3-class classification problem, the dimension of the relevant subspace for linear discrimination is 2 (this is given by Vapnik-Chervonenkis dimension of linear classifiers, which is $h = D - 1$). Hence, additional projections add noisy, non-discriminative information to data.

In Segmentation dataset, there are *outliers*, which provoke the degradation in the performance of MQMI. This is due to the maximization of the cost $\|p(\mathbf{z}, y) - P(y)p(\mathbf{z})\|^2$: if some outliers exist, they are pushed away as a way to minimize $p(\mathbf{z})$. MCL/MMI and MLRT are more robust in that sense. No data points can be pushed away because in that case $\log p(\mathbf{z})$ would tend to minus infinite.

Regarding the performance of the two classifiers considered, PC outperforms 1NN at the lowest dimensions (1 or 2 features). When more features are considered, both classifiers provide similar results. Although 1NN is not a state-of-the-art classifier, the fact that PC performs better or similarly in most cases suggests the convenience of its usage, specially in those

cases in which a probability measure or soft output is required. Besides, the computational complexity of both methods 1NN and PC are similar.

When we perform a comparison between ML-LOO and Scott bandwidth selection criteria, we find that in general the ML-LOO criterion performs better, although not in all cases. According to the plots, the performance depends on the FE method more strongly than on the bandwidth selection criterion. This suggests that it can be reasonable to use Scott rule when computational limitations exist -for example, when the number of per-class data points is very high, since the complexity of ML-LOO is with $O(n_l^2)$ for each class.

Finally, in order to visualize how the dimension is reduced while preserving discrimination ability, we show a scatter-plot of the projected data points into a 2-dimensional space by the MCL/MMI method for the Optdigit datasets in Figures 8 and 9. This dataset consists of images of handwritten digits with a 8×8 resolution. The classification accuracy is still far below the optimal, as shown in Fig. 3, but we can notice how the different digits are spatially arranged so that samples from the same class are neighbors after the projection. This locality is also present for the test samples, which makes possible to obtain a decent 60 % of accuracy.

B. Computational Complexity

MCL/MMI, MLRT and MQMI are based on the optimization of a non-convex cost. For this reason, the computational complexity cannot be determined by an entirely analytical study. Hence, we separately describe the complexity of each iteration, which can be determined analytically, and the number of iterations required if a gradient descend method is used together with a backtracking step selection [25].

The computational complexity of each iteration is given by the expressions listed in Table II. These expressions are easy to obtain from Eq. (5), which gives us the complexity of computing the gradient of each kernel computation (roughly $D \cdot d$ operations), multiplied by the number of kernels, as established by Eqs. (4), (8) and (14). For MCL/MMI and MQMI, no inter-class kernels are evaluated, because the KDE of each class is not evaluated on data points from the other classes. In MLRT, such inter-class evaluations actually take place, which is the reason of the higher complexity of this method.

Method	Complexity
MCL/MMI	$O(Dd\sum n_k^2)$
MLRT	$O(DdN^2)$
MQMI	$O(Dd\sum n_k^2)$

TABLE II
COMPTUTATIONAL COMPLEXITY OF AN ITERATION IN EACH ITL METHOD.

Regarding the number of iterations needed to reach the maximum, a gradient descend procedure has been used, with backtracking line search [25]. In Figure 10 we show the number of iterations required to perform each dimensionality reduction, averaged across tasks. The overall complexity, obtained by the combination of the figures in Table II and the number of iterations displayed in Fig. 10 , is shown in Figure 11.

The results stress that both MCL/MMI and MQMI have a computational complexity lower than MLRT's. From Figs. 10 and 11, we note that the higher computational cost of MLRT is mainly due to the complexity of each iteration rather than the number of iterations required.

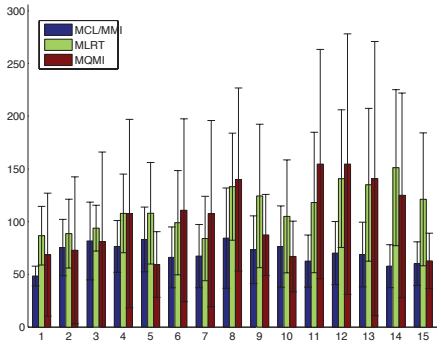

Fig. 10. Number of iterations needed to reach maximum in each FE method.

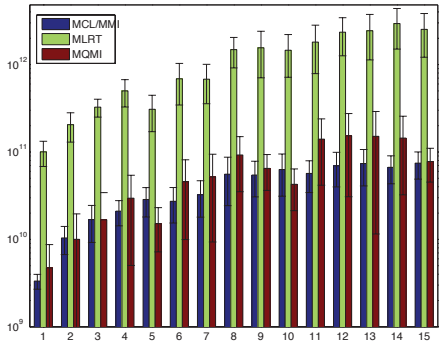

Fig. 11. Overall computational complexity of the methods, and their dependence with the output dimension.

V. CONCLUSIONS

We have provided a survey of methods for supervised feature extraction that make use of kernel density estimators to model the distribution of data. Together with methods existing in the literature, such as MQMI and MCL, we have also proposed two new criteria: the maximization of the mutual information (MMI) and the maximization of the likelihood ratio test (MLRT). An analytical study of MMI has revealed its equivalence to MCL. Unlike the other methods, the theoretical connections between MI and classification error have been found in the literature and reviewed in this paper. The experiments carried out have shown that ITL methods outperform classical methods PCA and LDA. Also, the results have revealed that, although in absence of outliers the methods perform similarly, in the presence of outliers a method based on the use of log-likelihoods - MCL/LOO or MLRT - show more robustness than MQMI.

The evaluation of the different criteria in terms of classification accuracy may allow us to discover the intrinsic dimension

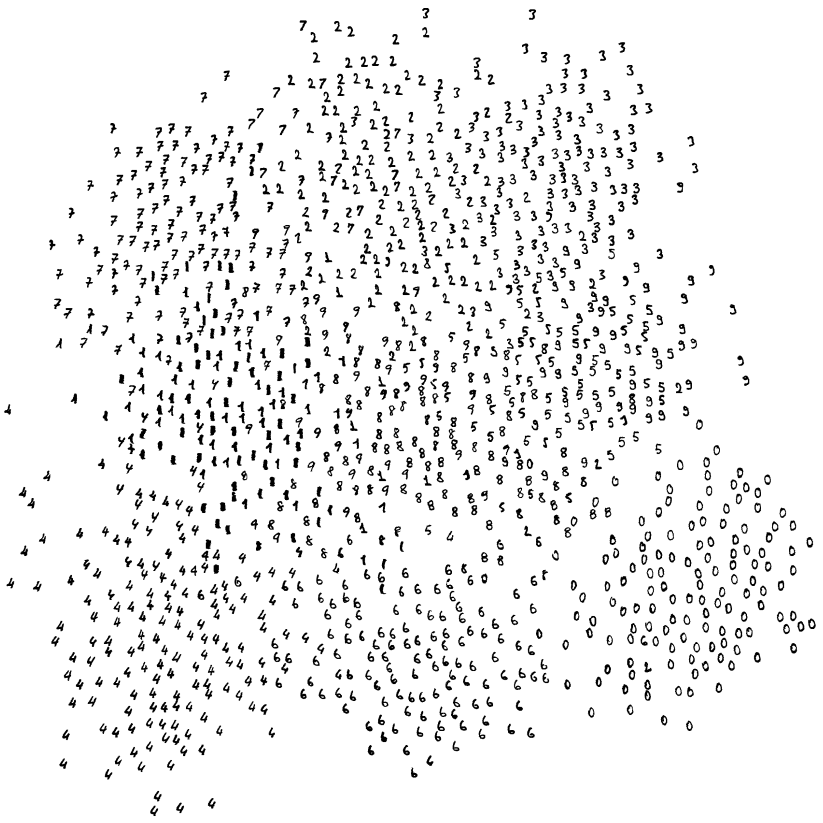

Fig. 8. Optdigits train data mapped to 2 dimensions by the MCL/MMI method.

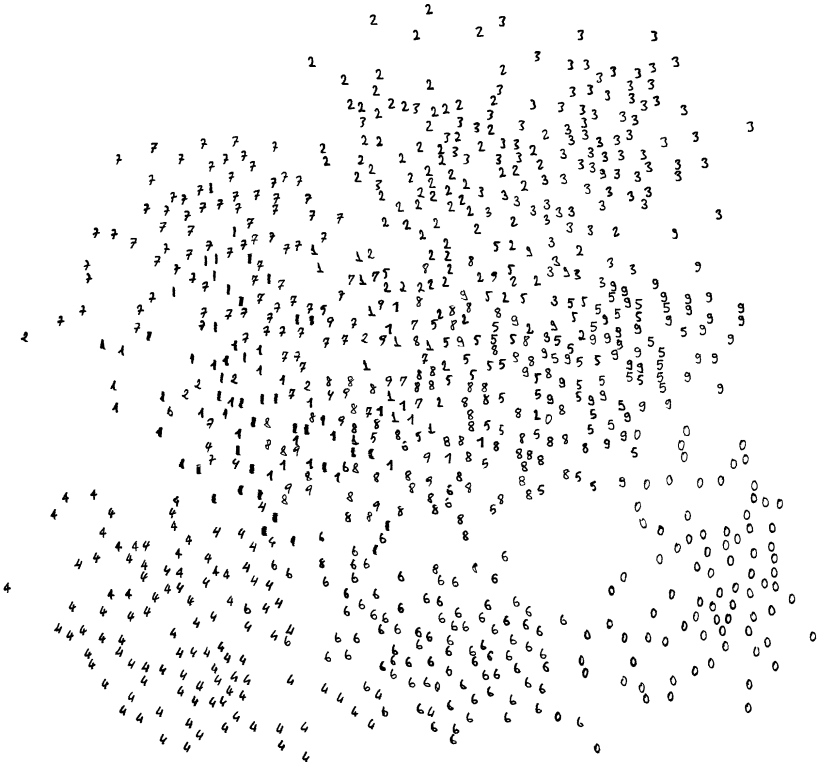

Fig. 9. Optdigits test data mapped to 2 dimensions by the MCL/MMI method.

of data. The empirical comparison suggests a slight superiority of the MCL/MMI method. The analysis of the computational complexity of the different methods show a superiority of MCL/MMI as well.

REFERENCES

[1] A. Bell and T. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *Neural Computation*, vol. 7, no. 6, pp. 1004–1034, 1995.

[2] J.M. Leiva-Murillo and A. Artés-Rodríguez, "Maximization of mutual information for supervised linear feature extraction," *IEEE Transactions on Neural Networks*, vol. 18, no. 5, pp. 1433–1441, 2007.

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

[4] S. Chen; X. Hong and C. Harris, "Probability density estimation with tunable kernels using orthogonal forward regression," *IEEE Trans. on Systems, Man and Cybernetics, Part B: Cybernetics*, vol. 40, no. 4, pp. 1101–1114, 2010.

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

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

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

[8] T. M. Cover, "Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition," *IEEE Trans. on Electronic Computers*, vol. EC-14, pp. 326–334, 1965.

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

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

[11] B. Schölkopf, A. Smola, and K. Müller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural Computation*, vol. 10, no. 15, pp. 1299–1319, 1998.

[12] J. Goldberger, S. Roweis, G. Hinton, and R. Salakhutdinov, "Neighbourhood components analysis," in *Advances in Neural Information Processing Systems, NIPS 17*, Vancouver, Canada, 2005, pp. 513–520.

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

[14] Berwin A. Turlach, "Bandwidth selection in kernel density estimation: A review," in *CORE and Institut de Statistique*, 1993, pp. 1–33.

[15] Luc Devroye and Gábor Lugosi, "A universally acceptable smoothing factor for kernel density estimates," , no. 24, pp. 2499–2512, Dec. 1996.

[16] Tarn Duong and Martin L. Hazelton, "Cross-validation bandwidth matrices for multivariate kernel density estimation," *Scandinavian Journal of Statistics*, vol. 32, pp. 485–506, 2005.

[17] J.M. Leiva-Murillo and A. Artés-Rodríguez, "Algorithms for gaussian bandwidth selection in kernel density estimators," in *Advances in Neural Information Processing System, NIPS, Optimization Workshop*, Wistler, Canada, 2008.

[18] I. Ahmad and P. Lin, "A nonparametric estimation of the entropy for absolutely continuous distributions," *IEEE Transactions on Information Theory*, vol. 22, no. 3, pp. 372–375, 1976.

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

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

[21] M. Feder and N. Merhav, "Relations between entropy and error probability," *IEEETIT: IEEE Transactions on Information Theory*, vol. 40, 1994.

[22] M. Hellman and J. Raviv, "Probability of error, equivocation, and the Chernoff bound," *IEEE Transactions on Information Theory*, vol. 16, no. 4, pp. 368–372, 1970.

[23] M. Grosse-Wentrup and M. Buss, "Multiclass common spatial patterns and information theoretic feature extraction," *Biomedical Engineering, IEEE Transactions on*, vol. 55, no. 8, pp. 1991–2000, 2008.

[24] 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.

[25] Stephen Boyd and Lieven Vandenberghe, *Convex Optimization*, Cambridge University Press, 2004.

LIST OF FIGURES

LIST OF TABLES