Maximization of Mutual Information for Supervised Linear Feature Extraction

Jose Miguel Leiva-Murillo, *Student Member, IEEE*, and Antonio Artés-Rodríguez, *Senior Member, IEEE*

***Abstract*—In this paper, we present a novel scheme for linear feature extraction in classification. The method is based on the maximization of the mutual information (MI) between the features extracted and the classes. The sum of the MI corresponding to each of the features is taken as an heuristic that approximates the MI of the whole output vector. Then, a component-by-component gradient–ascent method is proposed for the maximization of the MI, similar to the gradient-based entropy optimization used in independent component analysis (ICA). The simulation results show that not only is the method competitive when compared to existing supervised feature extraction methods in all cases studied, but it also remarkably outperform them when the data are characterized by strongly nonlinear boundaries between classes.**

***Index Terms*—Feature extraction, information theory, pattern recognition.**

I. Introduction

THE general problem in supervised learning is to estimate the relationship between an input $\mathbf{x}$ and an output y from a data set $\{\mathbf{x}_i, y_i\}$, $i = 1, \ldots, L$, $\mathbf{x}_i \in \mathbb{R}^N$. In this paper, we focus on *classification* problems, in which y is discrete, i.e., $y \in \{c_1, c_2, \ldots, c_{N_c}\}$. In that case, y is referred to as the *class* or the *label*.

Very often, it is desirable to reduce the dimension of the previous data to the classification. There are several reasons to do it. First, the generalization ability of the resulting machine is improved when the number of variables is low with respect to (w.r.t.) the number of input data samples. Second, a lower dimension leads to faster and computationally cheaper training and testing of the classifier. Finally, finding a relevant subset of variables or a ranking of the most informative ones can be useful for interpretation and explanatory purposes. Thus, the problem of extracting a set of M features from the original N ones $(M < N)$ is referred to as feature extraction (FE). If the FE takes into account the class, the FE is said to be *supervised*.

The FE is defined by a function $\mathbf{z} = f(\mathbf{x})$, $\mathbf{z} \in \mathbb{R}^N$ that may be linear or nonlinear. The use of one or the other is related to the classifier used, so that it is a common practice to apply either a linear FE method followed by a nonlinear classifier, or a nonlinear FE method before a linear classifier. In the first case, the responsibility of finding the nonlinear separation boundaries relies on the classifier. In the second case, the feature extractor projects the data on a set of variables in which the nonlinear patterns are *unfolded*, and a linear discrimination function is able to separate the classes [1]. In this paper, we consider the linear FE. As an example of the potential of linear FE, it has been shown that the performance of a simple k-nearest-neighbors (KNN) classifier can be spectacularly improved by a proper linear transformation on data [2].

Classical supervised FE methods assume that classes are linearly separable, so that the distance between classes is taken as the discrimination criterion. Some of these methods are linear discriminant analysis (LDA) [3], sliced inverse regression (SIR) [4], partial least square regression (PLS) [5], and canonical correlation analysis (CCA) [6].

There is a common limitation of these methods. They rely on first- and second-order statistics, which give an idea about the linear separability between classes (remote means and low variances suggest easy separation). Thus, a method able to see further than linear class separability must take into account additional information, which rely on high-order moments. The FE, in spite of being linear, must preserve the discriminative information even if the classification boundaries are nonlinear. For this purpose, we make use of Shannon's information theory, which provides us with a powerful tool for measuring *information* in statistical terms. In this context, mutual information (MI) measures the degree of statistical dependence among two or more variables. When dealing with FE for classification, the dependence of interest is the one between the features extracted in $\mathbf{z}$ and the classes in y.

The purpose of this paper is to introduce a method that, by making use of MI, is able to extract the most discriminative projections. The validity of our approach is supported by information theoretic results that relate the concepts of MI and error probability. The set of performance experiments show that our method is as good as any of the existing linear methods, and outperforms them when facing data sets characterized by a strongly nonlinear classification function.

The rest of the paper is organized as follows. In Section II, a summary of information theory concepts is provided, together with the current state-of-the-art of information-theoretic feature extraction methods. In Section III, we describe our algorithm for maximizing the MI. In Section IV, a set of experiments are described that show the feasibility of the method for feature extraction. In Section V, we offer some conclusions and suggestions for future work.

Manuscript received December 22, 2005; revised July 19, 2006 and November 6, 2006; accepted December 3, 2006. This work was supported in part by the Ministerio de Educación y Ciencia of Spain under the Projects "DOIRAS" TIC2003-02602 and "MONIN" TEC2006-13514-C02-01, and in part by the Comunidad de Madrid under the Project "PRO-MULTIDIS-CM" S0505/TIC/0223.

The authors are with the Department of Signal Theory and Communications, Universidad Carlos III, Madrid 28911, Spain (e-mail: leiva@ieee.org; antonio@ieee.org).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TNN.2007.891630

II. Background on Information Theoretic Learning

In a classification problem, the MI between the $\mathbf{x}$ and the y is given by

$$I(\mathbf{x}, y) = h(\mathbf{x}) - h(\mathbf{x}|y) \tag{1}$$

where $h(\cdot)$ is Shannon's *differential* entropy, defined as

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

The MI measures the information that one variable contains about another one, i.e., the reduction of uncertainty of a magnitude when another one is known [7]. Intuitively, features containing a high quantity of MI w.r.t. the classes are more suitable for classification than others that contain a lower level. As noted by Torkkola [8], two bounds on Bayes error justify the use of MI between components and classes as a criterion for FE. The first one is Fano's lower bound $p_e \geq (H(y) - I(\mathbf{x}, y) - 1)/\log(N_c)$; the second one is Hellman and Raviv's upper bound $p_e \leq (1/2)(H(y) - I(\mathbf{x}, y))$. Both bounds decrease as the $I(\mathbf{x}, y)$ grows, which makes reasonable to use the MI as a criterion for FE.

On the other hand, the data processing inequality says that, for any deterministic transformation $T(\cdot)$, the following property holds:

$$I(T(\mathbf{x}), y) \leq I(\mathbf{x}, y)$$

where the equality holds only when $T(\cdot)$ is invertible or leads to a sufficient statistic of $\mathbf{x}$ w.r.t. y [7]. Since the MI between the data and the classes cannot be improved, our objective is to obtain the $T(\cdot)$ that preserves the maximum information for a given dimension reduction. In the sequel, we will only consider linear transformations as $\mathbf{z} = T(\mathbf{x}) = \mathbf{W}^T\mathbf{x}$. Formally, the problem can be stated as

$$\mathbf{W}_{\text{opt}} = \arg\max_{\mathbf{W}} I(\mathbf{W}^T\mathbf{x}, y) \tag{3}$$

with $\mathbf{W}$ being a $N \times M$ matrix $(M < N)$. The computation of this quantity is difficult due to two reasons. First, the probability density functions (pdfs) of the data need to be estimated to obtain the entropy. This can be done via a nonparametric estimation, as Parzen window modeling, or by a semiparametric way, such as a Gaussian mixture model (GMM). Both models may suffer from the "curse of dimensionality," which refers to the overfitting of the training data when their dimension is high. Second, the integral in (2) cannot be solved in an analytical way, except for few, analytically known pdfs. Parzen models as well as GMMs describe the pdf as a sum of simple (usually Gaussian) distributions. The sum inside the logarithm makes the problem of integration intractable.[1]

Due to these difficulties, some recent works on information-theoretic learning have proposed the use of alternative measures for MI. The MI can be stated in terms of the Kullback–Leibler's divergence as $I(\mathbf{x}, y) = D_{KL}(p_{\mathbf{x}Y}(\mathbf{x}, y)||p_{\mathbf{x}}(\mathbf{x})P_Y(y))$ [7]. To avoid the difficulty of its computation, alternative divergences based on geometric theorems have been described, as the Cauchy–Schwartz and the Euclidean difference-of-vectors inequalities [10]. The pseudo-MIs obtained have been successfully applied to nonlinear FE [11] as well as linear FE [8] by means of a nonparametric estimation of the pdfs involved. Renyi's MI also avoids some of the computation problems of Shannon's, so that it has been proposed for both unsupervised and supervised learning [11]. Recently, an interesting approach to the information-theoretic FE has been introduced, in which a linear projection on the data is applied that maximizes the likelihood of the conditional (estimated) conditional densities. The criterion is shown to be asymptotically equivalent to the maximization of MI [12].

The estimation of the MI has also been considered from a KNN perspective. Kraskov *et al.* [13] propose a method for estimating the MI among a set of continuous variables based on a KNN evaluation of their interaction. This approach is based on a previous work that estimated the entropy following the same procedure [14]. Unfortunately, a KNN approach cannot been applied to a scheme in which the MI is to be optimized by a structured procedure, as a gradient-based one. A linear transformation on the data can make the nearest neighbor of a sample change, so that there are discontinuities in the derivative; hence, a global search procedure should be used for it optimization.

In spite of FE, the MI as a criterion has been applied to a variety of learning-related problems. Feature selection (the especial case of FE in which the features extracted are a subset of the original ones) has also been analyzed by means of a direct estimation of MI between features and classes [15]. The MI was estimated by building a histogram from each continuous variable and then treating it as a discrete one. The information bottleneck is a clustering method inspired by Shannon's rate distortion theory that maximizes the MI of the selected clusters w.r.t. an auxiliary variable while minimizing the MI w.r.t. the original data [16]. From a neuroscience perspective, neural responses to natural stimuli have been analyzed, in which the information carried by a spike is measured and optimized w.r.t. a projection [17].

The purpose of this paper is to overcome the difficulties of MI estimation and its application to FE, by means of an entropy estimation method that has succeeded in independent component analysis (ICA). Our proposal for MI estimation can be viewed as an heuristic in which the sum of 1-D MIs $\sum_i I(z_i, y)$ is computed instead of the $I(\mathbf{z}, y)$. We obtain the gradient of each variable w.r.t. a projection, so that we can search for the most informative projection from the multidimensional input. Then, we describe the extension of the method to obtain multiple projections.

III. Maximization of Mutual Information for Feature Extraction

In this section, we introduce our maximization of mutual information (MMI) algorithm for FE. We begin explaining the mixture model on which our methodology is based. In Section III-A, we provide a theoretical justification of MI usage when a mixture model with both relevant and noise features is

[1] A method for multidimensional entropy estimation based on GMM modeling has been described [9]. According to this estimation, a maximization of MI is carried out for feature extraction. This method reliably works when very few clusters are enough for the GMMs and the number of dimensions is not very large.

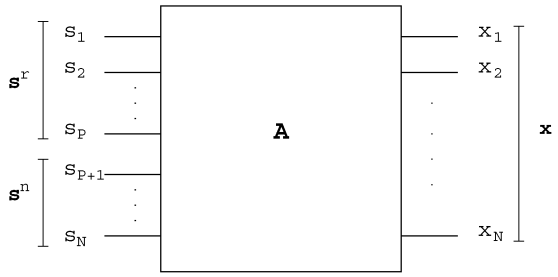


Fig. 1. Model assumed for the mixture of signals. The vector $\mathbf{s}^r$ groups the relevant variables (for classification); $\mathbf{s}^n$ groups the noisy ones. Each x_i is a mixture of signals from both types of variables.

assumed for the data. In Section III-B, we theoretically discuss the feasibility of the alternative criterion given by $\sum_i I(z_i, y)$, instead of $I(\mathbf{z}, y)$. We describe the estimation of both the entropy and the MI in Section III-C, as well as the method for the extraction of each individual feature. After that, we describe in Section III-D how a number of features are obtained.

A. Mixture Model

Each of the features in the observed data can be assumed to be generated by a mixture of unknown signals as shown in Fig. 1. This is similar to the blind source separation (BSS) scheme, in which every observed signal is thought of as a mixture of unknown, statistically independent signals. In our case, the property assumed for the unknown signals is that there are two types of sources s_i: the ones that are *relevant* and contain a certain degree of dependence w.r.t. the class (which are grouped in $\mathbf{s}^r$), and the ones that are independent of the class vector, which can be considered as *noisy* sources (in $\mathbf{s}^n$).

The vector $\mathbf{s}^n$ is characterized by its statistical irrelevancy, i.e.,

$$I(\mathbf{s}^n, y) = 0. \tag{4}$$

In the following, we briefly prove that, in such case:

$$I(\mathbf{s}^r, y) = I(\mathbf{x}, y).$$

To prove it, we start by assuming the matrix $\mathbf{A}$ to be invertible, so that the following holds:

$$I(\mathbf{x},y)=I(\mathbf{s},y)=I\left(\{\mathbf{s}^r,\mathbf{s}^n\},y\right)=\left\langle \log\frac{p\left(\{\mathbf{s}^r,\mathbf{s}^n\},y\right)}{p(\mathbf{s}^r,\mathbf{s}^n)P(y)}\right\rangle_{p(\mathbf{s},y)}.$$

By applying the Bayes rule, we get

$$I(\mathbf{x}, y) = \left\langle \log \frac{p(\mathbf{s}^n|\mathbf{s}^r, y)p(\mathbf{s}^r, y)}{p(\mathbf{s}^r, \mathbf{s}^n)P(y)} \right\rangle_{p(\mathbf{s}^r, \mathbf{s}^n, y)}. \tag{5}$$

According to the theorem of statistical sufficiency, $\mathbf{s}^n$ is said to be (statistically) irrelevant if

$$p(\mathbf{s}^n|\mathbf{s}^r, y) = p(\mathbf{s}^n|\mathbf{s}^r).$$

In that case, (5) can be rewritten as

$$\begin{aligned} I(\mathbf{x}, y) &= \left\langle \log \frac{p(\mathbf{s}^n|\mathbf{s}^r)p(\mathbf{s}^r, y)}{p(\mathbf{s}^r, \mathbf{s}^n)P(y)} \right\rangle_{p(\mathbf{s},y)} \\ &= \left\langle \log \frac{p(\mathbf{s}^r, y)}{p(\mathbf{s}^r)P(y)} \right\rangle_{p(\mathbf{s}^r, \mathbf{s}^n, y)}. \end{aligned}$$

Finally, we conclude the proof by marginalizing w.r.t. $\mathbf{s}^n$

$$I(\mathbf{x}, y) = \left\langle \log \frac{p(\mathbf{s}^r, y)}{p(\mathbf{s}^r)P(y)} \right\rangle_{p(\mathbf{s}^r, y)} = I(\mathbf{s}^r, y).$$

We now demonstrate that an orthonormal matrix is sufficient to recover the separation between relevant and irrelevant features. This is important because it allows us to constrain our search on the *group* of orthonormal matrices, instead of doing it on the *ring* of matrices.

If $\mathbf{A}$ is invertible, $\mathbf{A}^{-1}$ separates $\mathbf{s}^r$ and $\mathbf{s}^n$. Let $\mathbf{QR}$ be the QR decomposition of $\mathbf{A}^{-1}$, so that $\mathbf{Q}$ is orthonormal and $\mathbf{R}$ is lower diagonal. In that case, $\mathbf{s} = \mathbf{A}^{-T}\mathbf{x} = \mathbf{R}^T\mathbf{Q}^T\mathbf{x}$. Let $\mathbf{x}'$ be defined as $\mathbf{x}' = \mathbf{Q}^T\mathbf{x}$, so that $\mathbf{s} = \mathbf{R}^T\mathbf{x}'$. Since $\mathbf{R}^T$ is upper diagonal, $\mathbf{s}$ is separated in relevant and irrelevant features only if $\mathbf{x}'$ has also been separated. Thus, if $\mathbf{x}'$ itself consists of either relevant or irrelevant projections, it means that $\mathbf{Q}$ has the structure $\mathbf{Q} = [\mathbf{W}\ \mathbf{N}]$, being $\mathbf{W}$ the matrix whose vectors project the relevant features and $\mathbf{N}$ the one that projects the noisy ones. Summarizing, there always exists an orthonormal $N \times M$ matrix $\mathbf{W}$ that, provided that $\mathbf{s}$ and $\mathbf{A}$ are as in Fig. 1, fulfills

$$I(\mathbf{x}, y) = I(\mathbf{W}^T\mathbf{x}, y). \tag{6}$$

B. Global MI and Individual MIs

The approximation of $I(\mathbf{z}, y)$ involves the entropy estimation of multidimensional data. From (1), we get

$$I(\mathbf{z}, y) = h(\mathbf{z}) - \sum_k P(c_k)h(\mathbf{z}|c_k)$$

being c_k each of the classes and $P(c_k)$ each of their *a priori* probability.

The estimation of $h(\mathbf{z})$ and each $h(\mathbf{z}|c_k)$ from a set of samples is a problem that has not been successfully solved in an n-dimensional space. However, good estimators exist for 1-D variables. This is why we use an alternative cost function: the sum of MI between individual projections and labels, i.e.,

$$\max_{\{\mathbf{w}_i\}} \sum_i I\left(\mathbf{w}_i^T\mathbf{x}, y\right) = \max_{\{\mathbf{w}_i\}} \sum_i I(z_i, y) \tag{7}$$

where the $\{\mathbf{w}_i\}$s are the columns of the matrix $\mathbf{W}$ as in (6).

In the following, we obtain the relationship between the function to be maximized in (7) and the ideal one, as expressed in (3). The sum of the output MIs is given by

$$\sum_i I(z_i, y) = \sum_i \left[h(z_i) - h(z_i|y)\right]. \tag{8}$$

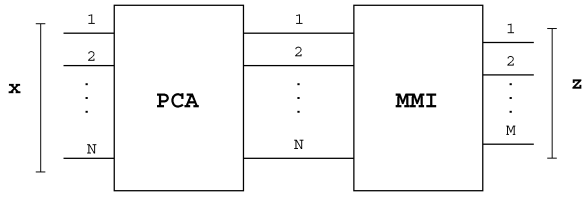


Fig. 2. Proposed scheme for MMI. The previous PCA performs a lossless transformation and whitens the input features. The dimension reduction is actually performed by the MMI block.

On the other hand, the inner MI corresponding to the "interaction" of the features is given by

$$I(\mathbf{z}) = \sum_i h(z_i) - h(\mathbf{z}). \tag{9}$$

Similarly, the conditional MI is given by

$$I(\mathbf{z}|y) = \sum_i h(z_i|y) - h(\mathbf{z}|y). \tag{10}$$

The $I(\mathbf{z}|y)$ denotes the MI between the z_is when conditioned to the classes. By combining the (8)–(10), we obtain

$$\sum_i I(z_i, y) = [I(\mathbf{z}) + h(\mathbf{z})] - [I(\mathbf{z}|y) + h(\mathbf{z}|y)].$$

After regrouping some terms, it leads to

$$\sum_i I(z_i, y) = I(\mathbf{z}, y) + [I(\mathbf{z}) - I(\mathbf{z}|y)]. \tag{11}$$

We have characterized the relationship between our cost function $\sum_i I(z_i, y)$ and the information given by $I(\mathbf{z}, y)$. The offset between both magnitudes is $I(\mathbf{z}) - I(\mathbf{z}|y)$. This offset can be described by infinite correlation coefficients of ascending degree, called *cumulants*. However, in practice, the cumulants until the fourth degree are enough to characterize the MI in an ICA context [18]. The cancellation of these cumulants would lead to the removing of the offset. However, there are two reasons why this is inadvisable. First, its computational burden is high, because working with cumulants of fourth degree involves the use of tensors with four dimensions. Second, the objective of performing ICA while maximizing the FE criterion would excessively constrain the search space.

As an alternative, we propose the application of a previous principal component analysis (PCA) step to reduce the interaction between the output variables, as shown in Fig. 2. This way, the first- and second-order cumulants are canceled at the input, and each x_i becomes normalized and uncorrelated to the rest of variables. If the projections $\mathbf{w}_i$ are forced to be orthogonal, the uncorrelation is preserved at the output, which is the purpose of the whitening. This way, first- and second-order interactions are removed in $\mathbf{z}$. In Section IV-C, a brief experiment is provided that suggests the convenience of this whitening.

C. Algorithm: Estimation of the MI and Its Gradient

We now face the problem of estimating each single $I(z_i, y)$ in (7) by the approximation of the entropies involved in (8). In 1-D, the entropy can be expressed as

$$h(z) = h_G(z) - J(z)$$

where $h_G(z)$ is the entropy with Gaussian assumption, i.e., the entropy that would be obtained if z were Gaussian, with the same mean and variance as in z; $J(z)$ is the so-called *negentropy*, which is the degree of dissimilarity of a variable w.r.t. a Gaussian one. It is a nonnegative magnitude (a Gaussian density is the most entropical among all densities with the same variance). Some approaches to ICA are based on the maximization of the negentropy [19], so that the less similar the pdfs of a source and a Gaussian are, the more likely the source is to be an independent and informative one.

Popular approaches to the problem of estimating the $J(z)$ [and so the $h(z)$] from a finite set of realizations are the Edgeworth and the Gram–Charlier expansions. The Gram–Charlier expansion estimates the pdf of z, so that the entropy is expressed as a linear combination of moments with increasing degree and infinite terms, which must be truncated at a certain degree [20]. In practice, a strong sensitivity to outliers appears due to the terms with the highest degree in the expansion. This leads to an estimation highly determined by the *tails* of the distribution, coming from outstanding, probably erroneous samples.

As an alternative, the use of pairs of nonpolynomial functions has been proposed, so that one of them is even, and the other one is odd [19]

$$J(z) \approx k_1 \left(E\left\{g_1(z)\right\}\right)^2 + k_2 \left(E\left\{g_2(z)\right\} - E\left\{g_2(\nu)\right\}\right)^2$$

where ν is an "equivalent" (equal mean and variance) Gaussian random variable of z. The use of polynomial functions as $g_1(z) = z^3$ and $g_2(z) = z^4$ leads to a Gram–Charlier truncation, which suffers from the lack of robustness mentioned previously. Some attempts have been carried out to develop robust expansions in density estimation [21], but its application to entropy estimation is not straightforward. On the other hand, the choice of nonpolynomial functions as $g_1(z) = z\exp(-z^2/2)$ and $g_2(z) = \exp(-z^2/2)$ has been proven to be robust and accurate. This choice leads to the following estimation of the negentropy:

$$J(z) \approx k_1 \left(E\left\{z\exp\left(\frac{-z^2}{2}\right)\right\}\right)^2 + k_2 \left(E\left\{\exp\left(\frac{-z^2}{2}\right)\right\} - \sqrt{1/2}\right)^2 \tag{12}$$

with the constants $k_1 = 36/(8\sqrt{3} - 9)$ and $k_2 = 24/(16\sqrt{3} - 27)$; this estimation of $J(z)$ has been successfully used in ICA [22]. Besides, it shows a good numerical behavior when being maximized. This is the approximation of the negentropy used in our algorithm.

We now express the MI and its gradient w.r.t. the projections $\mathbf{w}_i$. Each 1-D MI of (8) can be now reexpressed as

$$I(z, y) = h_G(z) - J(z) - \sum_k P(c_k)\left[h_G(z|c_k) - J(z|c_k)\right]$$

where each $z|c_k$ refers to subset of samples belonging to class c_k, and the $J(\cdot)$ is estimated according to (12).

If for each projection $||\mathbf{w}_i|| = 1$ holds, several simplifications can be applied, because the z_i are, in this case, variance-normalized, as described in Section III-B.

By substituting each $h_G(z)$ by its analytical value [7], the MI can be stated as

$$I(z,y) = \log\left((2\pi e)^{1/2}\sigma_z\right) - J(z) - \sum_k P(c_k)\left[\log\left((2\pi e)^{1/2}\sigma_{z|c_k}\right) - J(z|c_k)\right].$$

We can simplify the expression by noting that $\sigma_z = 1$ if data are normalized

$$I(z,y) = \sum_k P(c_k)J(z|c_k) - J(z) - \sum_k P(c_k)\log\sigma_{z|c_k}.$$

The variance of the output signal is given by $\sigma_z^2 = \mathbf{w}^T\mathbf{C}_\mathbf{x}\mathbf{w}$ where $\mathbf{C}_\mathbf{x}$ is the covariance matrix of $\mathbf{x}$. The same relation exists between each $\mathbf{C}_{\mathbf{x}|c_k}$ and each $\sigma^2_{\mathbf{x}|c_k}$. We are now able to set an expression for the gradient

$$\nabla_\mathbf{w} I(z,y) = \sum_k P(c_k)\nabla_\mathbf{w} J(z|c_k) - \nabla_\mathbf{w} J(z) - \sum_k P(c_k)\frac{\mathbf{C}_{\mathbf{x}|c_k}\mathbf{w}}{\mathbf{w}^T\mathbf{C}_{\mathbf{x}|c_k}\mathbf{w}} \quad (13)$$

where

$$\nabla_\mathbf{w} J(z) = 2k_1 E\left\{z\exp\left(\frac{-z^2}{2}\right)\right\} E\left\{\mathbf{x}(1-z^2)\exp\left(\frac{-z^2}{2}\right)\right\} - 2k_2\left(E\left\{\exp\left(\frac{-z^2}{2}\right)\right\} - \sqrt{\frac{1}{2}}\right) E\left\{\mathbf{x}z\exp\frac{-z^2}{2}\right\}. \quad (14)$$

Equivalently, the $\nabla_\mathbf{w} J(z|c_k)$ are obtained from the subset of samples belonging to class c_k. The use of this gradient follows a standard gradient–ascent procedure.

Although it is known that the entropy of a variable is sensitive to its variance (with $\log\sigma$), the estimation proposed in this section scales in a different way, because of our approximation of $J(z)$ [see (12)]. Then, another reason to apply PCA as proposed in Section III-B, Fig. 2 is that it helps us control the variance of the output. Otherwise, the estimation would scale exponentially with the variance, so that the MI estimation would be seriously biased.

D. Extension to Multiple Projections

So far, we have described a method for obtaining a single feature z_i that maximizes the MI between it and the classes. Instead of simultaneously searching for the whole set of relevant features, according to the cost function in (7), we iteratively obtain single projections with a decreasing degree of relevance. This way we avoid some problems with local minima if we do not simultaneously search the whole set of components.

The iterative search of the sequential projections $\mathbf{w}_i$ may take into account the following two constraints.

1) They must be normalized: $\|\mathbf{w}_i\| = 1$, since the scalability of the negentropy $J(z)$ is exponential.
2) Each one must be orthogonal to the projections already obtained, i.e., $\mathbf{w}_i\mathbf{w}_j = 0$ for $1 \le j < i$. There are two reasons for this. First, we avoid to obtain the same projection more than once. Second, if we start from an uncorrelated set of samples such a matrix would preserve this *whitening* as explained in Section III-B.

In summary, the set of projections given by the matrix $\mathbf{W}$ must be orthonormal. In order to reach orthonormalization, we obtain, in each iteration, a projection $\mathbf{w}_i$ that is forced to be orthogonal to the ones previously obtained, as well as normalized. There are different ways of dealing with orthogonality constraints. In our case, a simple Gram–Schmidt orthonormalization has provided good results. Each iteration of the ascent–gradient algorithm for a given projection consists of two steps. We first move in the direction given by the gradient [see (13)]. Second, the resulting vector is normalized and made orthogonal to the ones already obtained, by means of the Gram–Schmidt procedure.

In Sections III-B and III-C, we stressed the reasons why a previous PCA step must be applied: the features must be normalized and the interaction among them is reduced. Thus, a justification for imposing orthonormality constraints is that otherwise the PCA step would not make sense since the uncorrelation would not be preserved at the output.

The procedure already described can be considered as a *top-down* methodology, in which the projections are obtained in a decreasing order of relevance. As an alternative to the scheme described, we might sequentially minimize the MI between a variable and the classes and remove it. Thus, we *push* the information contained in data to the subspace that remains orthogonal to the projections rejected. In this case, the direction of the search is the opposite from the one given in (14). This *bottom-up* methodology has the advantage that the number of features requested can be more flexible: we can reject projections until we notice that no more *uninformative* features can be found. The procedure has also a disadvantage: for databases with a high number of dimensions, the method is computationally intensive if only few features contain most of the information.

IV. Experiments

Here, we perform a set of classification experiments to show the validity of our approximation and the suitableness of our method for FE. First, we provide a brief explanation about other supervised FE methods. Second, we give some details about the data sets involved in the experiments, the preprocessing carried out on them, and the classifiers used. After that, we show the results of our experiments: some considerations about the validity of the MI estimation, a comparison between the top-down and the bottom-up versions, and finally, a set of classification experiments in which MMI is compared to other FE methods.

A. Linear FE Methods Used for Comparison

In the following, we provide a brief explanation of SIR, LDA, CCA, and Torkkola's MRMI. These are the supervised methods that, together with unsupervised PCA, have been compared to MMI.

- **Sliced inverse regression (SIR)**. This method performs a feature extraction by carrying out an inverse regression, in which $\mathbf{x}$ is guessed from y. The domain of y is partitioned in *slices*, which is straightforward in classification: there are as many slices as classes. Then, a PCA-

like procedure is applied to a weighted covariance matrix $\mathbf{M} = \sum_{k=1}^{N_c} P(c_k)\boldsymbol{\mu}_k\boldsymbol{\mu}_k^T$, being each $\boldsymbol{\mu}_k$ the sample mean of class k. The transformation matrix $\mathbf{W}$ is obtained from the eigenvectors of $\mathbf{M}$.

- **Linear discrimination analysis (LDA)**. This method obtains a set of projections according to which the largest distance between classes is achieved, in terms of Fisher discriminant. Let $\mathbf{S}_B$ be the so-called *between-class scatter matrix*, given by $\mathbf{S}_B = \sum_{k=1}^{N_c}(\boldsymbol{\mu}_i - \boldsymbol{\mu})(\boldsymbol{\mu}_i - \boldsymbol{\mu})^T$, where $\boldsymbol{\mu}_\mathbf{i}$ is the inner mean of class k and $\boldsymbol{\mu}_i$ the global mean. Let also $\mathbf{S}_W$ be the *within-class scatter matrix*, given by $\mathbf{S}_W = \sum_{k=1}^{N_c} P(c_k)\mathbf{C}_k$, being $\mathbf{C}_k$ the covariance matrix of samples belonging to class k. The objective is to maximize the cost function $J(\mathbf{w}) = \mathbf{w}^T\mathbf{S}_B\mathbf{w}/\mathbf{w}^T\mathbf{S}_W\mathbf{w}$. The projection that maximizes the criterion is the one given by the eigenvector related to the highest eigenvalue in the singular value decomposition $\mathbf{S}_B\mathbf{w} = \lambda\mathbf{S}_W\mathbf{w}$.
- **Canonical correlation analysis (CCA)**. The aim of CCA is to obtain a pair of linear transformations in such a way that, when applied to the two input signals (in this case, one of them is 1-D), the correlation between the outputs is maximized. CCA searches for a couple of projections $\mathbf{W}_x$ and $\mathbf{W}_y$ such that the correlation between $\mathbf{W}_x^T\mathbf{x}$ and $\mathbf{W}_y^T\mathbf{y}$, given by $\mathbf{W}_x^T\mathbf{C}_{xy}\mathbf{W}_y/(\mathbf{W}_x^T\mathbf{C}_{xx}\mathbf{W}_x)^{1/2}(\mathbf{W}_y^T\mathbf{C}_{yy}\mathbf{W}_y)^{1/2}$, is maximized.
- **Maximization of quadratic information (MRMI)**. This method optimized an alternative measure of information called quadratic information [8], which is maximized in Torkkola's algorithm, as mentioned in Section II. The pdfs involved are estimated by a Parzen model. The maximum is reached by a gradient search on the projection matrix.

Some of these linear FE methods show a limitation that must be stressed: they cannot provide more than $N_c - 1$ relevant projections, being N_c the number of classes. In order to overcome this difficulty, and to perform a fair comparison with our method, we propose a method for obtaining an arbitrary number of projections from these methods for the experiments that need it. This is done by means of a subspace search. The procedure for each z_i is as described in Fig. 3. The first one is obtained by directly projecting the vector $\mathbf{x}$ on $\mathbf{w}_1$. Thus, $\mathbf{w}_1$ becomes the first vector in $\mathbf{W}$. By a Gram–Schmidt orthogonalization, we obtain a matrix $\mathbf{B}$ that transforms $\mathbf{x}$ into an element of the space complementary to $\mathbf{w}$. The search of the second component is carried out in this subspace, and then expanded to the N-dimensional space. Iteratively for the k components, the matrix $\mathbf{B}$ is recomputed and each $\mathbf{w}_i$ obtained in the subspace orthogonal to the vectors $\{\mathbf{w}_j\}_1^{i-1}$.

$\mathbf{B}_0 = \mathbf{I}$

for $k = 1$ to M

$\mathbf{x}_p = \mathbf{B}_k^T\mathbf{x}$

Obtain an optimal projection $\tilde{\mathbf{w}}$ for $\mathbf{x}_p$ (LDA, SIR or CCA)

Express it in the original space: $\mathbf{w} = \mathbf{B}_k\tilde{\mathbf{w}}$

$\mathbf{W}_k = \{\mathbf{W}_{k-1}, \mathbf{w}\}$

Obtain $\mathbf{B}_k$ as the base of the complementary space to $\mathbf{W}_k$

end

Fig. 3. Algorithm for searching on complementary subspaces. Each projection is searched on the nullspace of (i.e., the space complementary to) the features already obtained, so that it is kept orthonormal to them.

B. Data Sets, Preprocessing, and Classifiers

We have used four different data sets to carry out our experiments. The synthetic nonlinear problem is used by Weston *et al.* for testing feature selection techniques [23]. We use it here to evaluate the performance of the FE methods when facing data described by highly nonlinear boundaries. To do so, the synthetic variables have been mixed by means of a random orthonormal matrix. The dimension of the data set is 50. Before the mixing, only two variables are relevant. 1000 samples have been generated for training and 500 for testing. The rest of data sets consist of real, public data from the University of California at Irvine (UCI) repository.[2] The characteristics of the data sets are shown in Table I. The data sets have been chosen to evaluate the method in very different input dimensions as well as numbers of classes.

TABLE I
CHARACTERISTICS OF THE PUBLIC DATA SETS USED: SAMPLE SIZE, DIMENSION, AND NUMBER OF CLASSES

Dataset	Train Samples	Test Sps.	Dim.	Classes
Non Linear	1000	500	50	2
Landsat	4435	2000	36	6
Optdigits	3823	1797	64	10
Letter	16000	4000	16	26

The same set of preprocessing steps has been applied to each data set involved in the following experiments:

- training data are centered, i.e., their mean is removed;
- after centering, training data are whitened, i.e., PCA is applied in order to decorrelate the variables, as illustrated in Fig. 2;
- the same mean and whitening matrix of training data are applied to test data.

The preprocessing with PCA allows a more fair comparison with the other methods. MRMI makes use of a spherical Parzen model for density estimation. A spherical shape of the data makes the performance of this method better than in the case that PCA was not applied. The rest of the methods are not affected by PCA, because they only consider linear discrimination criteria, which are not affected by linear transformations.

The classification performance of the methods considered has been evaluated by means of a support vector machine (SVM) [24] with a Gaussian radial basis function as kernel as well as a KNN classifier [25]. We have chosen the SVM because it is a classifier that is proven to be less sensitive to the "curse of dimensionality" than other methods, so that its performance is highly correlated with the quantity of information that data carry about classes. Thus, the performance of SVM is more fair than other classifiers for comparing FE methods. The hyperparameters of the SVM (the cost C and the width of the kernel σ) have been chosen by a threefold cross-validation procedure on the training data. On the other hand, KNN is a classifier that, in

[2]http://www.ics.uci.edu/~mlearn/MLRepository.html.

TABLE II
ESTIMATION OF THE MI WITH AND WITHOUT PCA AT THE INPUT (USING THE KNN METHOD DESCRIBED IN [13]). PCA IS SHOWN TO REDUCE THE INTERACTION AMONG VARIABLES

Dataset	Without PCA			With PCA		
	$\hat{I}(\mathbf{x})$	$\hat{I}(\mathbf{x}\|y)$	$\hat{I}(\mathbf{x}) - \hat{I}(\mathbf{x}\|y)$	$\hat{I}(\mathbf{x})$	$\hat{I}(\mathbf{x}\|y)$	$\hat{I}(\mathbf{x}) - \hat{I}(\mathbf{x}\|y)$
Landsat	41.3611	20.1816	21.1794	-0.8861	-0.5126	-0.3734
Optdigits	15.3331	8.3862	6.9469	3.7325	3.5366	0.1959
Letter	23.8787	19.6774	4.2013	10.8035	7.318	3.4851

TABLE III
ESTIMATION OF THE OFFSET BETWEEN COST FUNCTIONS OF (3) AND(7) AT THE OUTPUT (USING THE KNN METHOD [13] AND THE PARZEN MODEL-BASED ONE [26]). THE ESTIMATED OFFSET (RIGHTMOST COLUMNS) IS LOW WHEN COMPARED TO THE COST FUNCTION (FIRST COLUMN)

Dataset	$\sum_i I(z_i, y)$	$\hat{I}(\mathbf{z})$		$\hat{I}(\mathbf{z}\|y)$		$\hat{I}(\mathbf{z}) - \hat{I}(\mathbf{z}\|y)$	
	MMI	Parzen	KNN	Parzen	KNN	Parzen	KNN
Landsat	16.1143	2.5736	2.0074	1.6678	0.5661	0.9058	1.4413
Optdigits	28.5263	4.2826	4.6845	1.852	1.2496	2.4306	3.4349
Letter	16.9255	4.7894	8.2885	2.5942	5.9074	2.1953	2.3812

spite of its simplicity, provides good performance in addition to interesting properties concerning the relationship between KNN and Bayes errors. In the experiments, we use KNN with $K = 1$.

C. Accuracy of the Estimated MI

In the first experiment, we evaluate the validity of estimating $\sum_i I(z_i, y)$ instead of $I(\mathbf{z}, y)$, on the components obtained for the public data sets. We measured the offset between both magnitudes [see(11)], in order to justify the use of the former instead of the latter. Also, we experimentally justify the use of PCA according to the reasons exposed in Section III-B. In Table II, we have estimated these magnitudes on the input vectors, by means of the KNN-based method of Kraskov *et al.* [13], mentioned in Section I. The negative numbers in the table are due to the inaccuracy of the KNN MI estimation, but suggest values close to zero. The results show that the use of PCA reduces the offset in the three cases studied, which justify its use as a preprocessing step. In Table III, we have estimated the offset for the output features. In this case, another estimator of the MI has been used, based on a Monte Carlo estimation of the entropy from a Parzen density model of the data sets. The width of the window has been chosen by a cross-validation maximum-likelihood (ML) procedure, as proposed in [26]. Both methods have been used to obtain the offset on the three data sets. The results for the first ten components obtained by top-down MMI (TD-MMI) are shown in Table III. There it can be seen that, according to these estimations, the deviation from the value $I(\mathbf{z}, y)$ can be considered small enough to justify the use of $\sum_i I(z_i, y)$ as a valid cost function. This fact provides us with a criterion for selecting the number of features to be obtained. Since the model presented in Fig. 1 is ideal and does not have to correspond to a real situation, a certain loss of information occurs, independently of the number of features M to be obtained. Let us denote this information loss by δ. A methodology for determining M from δ may consist of first obtaining N projections, and then selecting M from them so that

$$\frac{\sum_{i=1}^{M} I(z_i, y)}{\sum_{i=1}^{N} I(z_i, y)} \geq 1 - \delta.$$

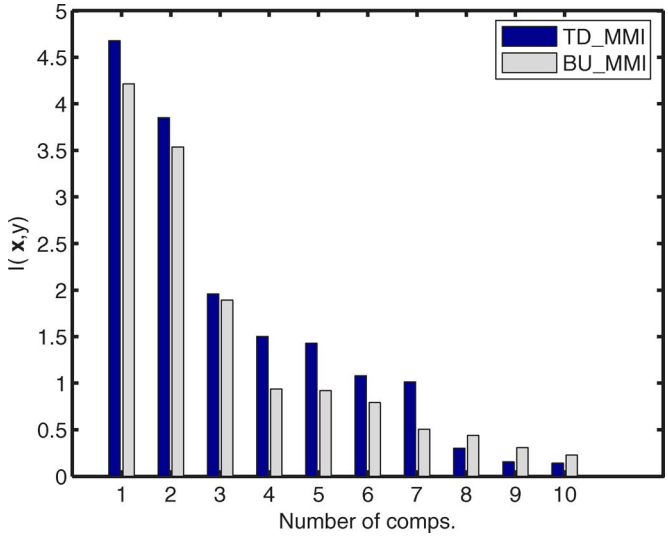


Fig. 4. Comparison between the MI obtained by TD-MMI and BU-MMI for landsat data set.

This would be an alternative to criteria such as Akaike information [27] or minimum description length (MDL) [28], which are difficult to apply in this context.

D. TD-MMI versus BU-MMI

In the second experiment, we compare the top-down version of MMI with the bottom-up one. The landsat data set has been used, because its dimension is not too high, which allows us to use the bottom-up approach with a reasonable computational cost. The value for the MI obtained for each feature is displayed in Fig. 4. The bottom-up MMI (BU-MMI) provides a poor performance, which is not surprising. This version of the algorithm requires a higher number of iterations. The inaccuracy of the MI estimation may be accumulated in each iteration, so that some information can be lost with each projection rejected. As a result, the features preserved have a lower MI than the features obtained by the top-down approach.

TABLE IV
PERCENTILE CLASSIFICATION ACCURACY ON THE NONLINEAR SYNTHETIC DATA SET

N. of Comps:	1	2	3	4	5
PCA/KNN	66.30	66.70	71.30	66.30	65.30
PCA/SVM	77.10	76.90	78.20	76.20	75.80
SIR/KNN	51.90	48.30	49.60	52.50	50.90
SIR/SVM	53.30	52.20	51.90	51.80	51.30
LDA/KNN	49.90	51.40	49.60	53.20	54.30
LDA/SVM	53.00	53.30	53.60	59.60	63.60
CCA/KNN	48.10	50.10	49.70	51.50	54.20
CCA/SVM	52.30	53.00	53.20	53.50	59.20
MRMI/KNN	51.50	71.50	89.00	88.10	85.50
MRMI/SVM	51.80	78.10	91.20	89.70	87.90
MMI/KNN	93.20	95.50	94.50	94.00	91.30
MMI/SVM	**94.90**	**96.70**	**95.70**	**94.90**	**94.10**
Raw Data					50.00

TABLE V
PERCENTILE CLASSIFICATION ACCURACY ON THE LAND SATELLITE DATA SET

N. of Comps:	1	2	3	4	5
PCA/KNN	40.80	78.40	83.80	84.70	85.80
PCA/SVM	51.55	**82.15**	85.75	**88.65**	**89.25**
SIR/KNN	47.15	71.25	82.00	83.95	82.45
SIR/SVM	55.00	79.35	85.65	86.85	87.25
LDA/KNN	47.85	71.35	82.00	83.90	82.50
LDA/SVM	54.90	79.35	85.65	86.85	87.20
CCA/KNN	47.15	71.25	82.00	83.95	82.45
CCA/SVM	55.00	79.35	85.65	86.85	87.25
MRMI/KNN	52.05	69.50	83.65	83.20	84.50
MRMI/SVM	62.55	75.75	85.95	86.40	87.10
MMI/KNN	56.20	74.75	82.80	84.40	84.45
MMI/SVM	**62.20**	81.40	**86.30**	87.55	87.85
Raw Data					86.85

TABLE VI
PERCENTILE CLASSIFICATION ACCURACY ON THE OPTICAL DIGITS DATA SET

N. of Comps:	1	2	3	4	5	6
PCA/KNN	28.32	52.81	71.68	78.80	87.81	90.43
PCA/SVM	35.89	61.83	76.02	82.42	90.87	92.60
SIR/KNN	32.22	59.04	77.30	85.31	89.87	92.60
SIR/SVM	40.57	64.77	**80.80**	**88.04**	91.15	92.77
SIR/KNN	32.22	59.04	77.30	85.31	89.87	92.60
LDA/SVM	40.57	64.77	**80.80**	**88.04**	91.15	92.71
CCA/KNN	32.22	59.04	77.30	85.31	89.87	92.60
CCA/SVM	40.57	64.77	**80.80**	**88.04**	91.15	92.77
MRMI/KNN	35.45	58.82	77.74	23.15	24.04	25.43
MRMI/SVM	41.29	65.16	80.80	33.44	34.84	35.61
MMI/KNN	36.39	60.32	72.12	84.42	89.98	92.65
MMI/SVM	**42.57**	**66.17**	77.57	86.64	**92.15**	**93.43**
Raw Data						97.16

E. Classification Experiments

In the last experiment, the TD-MMI is compared to the methods described in Section IV-A on the four data sets of Table I.

The results, for the classifiers described and several degrees of reduction are displayed in Tables IV–VII. The *raw data* row shows, in each case, the result obtained by an SVM on the original, not reduced data.

TABLE VII
PERCENTILE CLASSIFICATION ACCURACY ON THE LETTER DATA SET

N. of Comps:	1	2	3	4	5	10
PCA/KNN	4.50	6.73	10.50	13.35	20.62	79.93
PCA/SVM	7.72	9.07	13.57	17.12	24.22	85.42
SIR/KNN	15.05	16.70	23.57	37.15	47.95	86.05
SIR/SVM	9.10	13.42	23.35	37.35	52.20	90.17
LDA/KNN	15.05	16.70	23.57	37.15	47.95	86.05
LDA/SVM	9.12	13.42	23.35	37.35	52.20	90.17
CCA/KNN	15.05	16.70	23.57	37.15	47.95	86.05
CCA/SVM	9.12	13.42	23.35	37.35	52.20	90.17
MRMI/KNN	15.15	16.20	24.68	37.35	47.58	89.20
MRMI/SVM	10.17	13.70	24.42	37.17	51.65	**92.27**
MMI/KNN	**15.18**	**16.73**	26.20	35.45	48.20	84.25
MMI/SVM	11.50	14.43	**27.83**	**37.53**	**52.82**	89.47
Raw Data						97.55

The most remarkable result by MMI is the one on the most nonlinear data set, which is the first one. Only MRMI and MMI are able to obtain the discriminative information, and MMI does it better than MRMI. The rest of the methods fail since this data set contains features that are nonlinearly separable.

In the second experiment, MMI and PCA obtain the best results on the landsat data set. In this case, the most discriminative information seem to rely on the most powerful directions on the data, which is why PCA reaches that good result.

In the third case, MMI obtain the best result in most of the cases. However, the difference among the performance of the methods is tighter.

In the last experiment, on the letter data set, MMI reaches again the best results for most of the considered reduction degrees.

These results suggest that MMI is not worse than any of the classical methods in any case. However, in an extreme case of nonlinear class separability, MMI can outperform the other methods. The good behavior of MMI in the nonlinear synthetic data set suggests that the enhancement of MMI w.r.t. traditional methods is higher as 1) data are characterized by higher nonlinear boundaries and 2) the data are more likely to have been generated by a mixing scheme close to the one described in Fig. 1. Although MRMI is also strong at finding the relevant projections, MMI shows a better behavior when facing high-dimensional input data, avoiding the overfitting of MRMI's nonparametric modeling.

V. CONCLUSION

We have presented a novel method for linear FE in classification, based on the maximization of the MI between the features obtained and the classes. The difficulty of the entropy estimation for multidimensional data is overcome by a sequential extraction of the features, so that a 1-D MI estimation can be applied to each single feature. By a set of experiments, the method has been shown to be competitive w.r.t. other existing methods. The best performance of MMI takes place at very high reduction degrees. Besides, the method outperforms classical FE methods in situations in which the boundaries between the classes are strongly nonlinear, without suffering from the overfitting in high-dimensional input spaces as other nonparametric procedures.

REFERENCES

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

[2] J. Goldberger, S. Roweis, G. Hinton, and R. Salakhutdinov, "Neighbourhood components analysis," in *Advances in Neural Information Processing Systems 17*, L. K. Saul, Y. Weiss, and L. Bottou, Eds. Cambridge, MA: MIT Press, 2005, pp. 513–520.

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

[4] K. Li, "Sliced inverse regression for dimension reduction," *J. Amer. Statist. Soc.*, vol. 86, no. 414, pp. 316–327, 1991.

[5] J. Wang, L. Ji, and Z. Kou, "Facial feature point extraction by partial least square regression," *The Pennsylvania State University CiteSeer Archives*, Mar. 19, 2003.

[6] H. Knutsson, M. Borga, and T. Landelius, "Learning canonical correlations," Comp. Vis. Lab., Linköping, Sweden, Rep. LiTH-ISY-R-1761, Jun. 1995.

[7] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.

[8] K. Torkkola, "Feature extraction by non-parametric mutual information maximization," *J. Mach. Learn. Res.*, vol. 3, pp. 1415–1438, 2003.

[9] J. M. Leiva and A. Artés, "A Gaussian mixture based maximization of mutual information for supervised feature extraction," in *Proc. 5th Int. Conf. ICA*, Granada, Spain, 2004, pp. 271–278.

[10] K. Torkkola and W. Campbell, "Mutual information in learning feature transformations," in *Proc. 17th Int. Conf. Mach. Learn.*, 2000, pp. 1015–1022.

[11] J. Principe, D. Xu, and J. W. Fischer, III, *Information-Theoretic Learning*. New York: Wiley, 2000, vol. 1.

[12] J. Peltonen and S. Kaski, "Discriminative components of data," *IEEE Trans. Neural Netw.*, vol. 16, no. 1, pp. 68–83, Jan. 2005.

[13] A. Kraskov, H. Stoegbauer, and P. Grassberger, "Estimating mutual information," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat.*, vol. 69, pp. 066138-1–066138-16, 2004.

[14] L. F. Kozachenko and N. N. Leonenko, "Sample estimate of entropy of a random vector," *Problems Inf. Transmission*, vol. 23, no. 9, pp. 95–101, 1987.

[15] R. Battiti, "Using mutual information for selecting features in supervised neural net learning," *IEEE Trans. Neural Netw.*, vol. 5, no. 4, pp. 537–550, Jul. 1994.

[16] F. C. Pereira, N. Tishby, and W. Bialek, "The information bottleneck method," in *37th Annu. Allerton Conf. Commun., Control, Computing.*, Urbana, IL, 1999, pp. 368–377.

[17] N. Rust, T. Sharpee, and W. Bialek, "Analyzing neural responses to natural signals: Maximally informative dimensions," *Neural Comput.*, vol. 16, pp. 223–250, 2004.

[18] J.-F. Cardoso, "Blind signal separation: Statistical principles," *Proc. IEEE*, vol. 86, no. 10, pp. 2009–2025, Oct. 1998.

[19] J. Karhunen, A. Hyvarinen, and E. Oja, *Independent Component Analysis*. New York: Wiley, 2001.

[20] S. Haykin, *Neural Networks: A comprehensive Foundation*, 2nd ed. Englewood Cliffs, NJ: Prentice-Hall, 1999.

[21] M. Welling, "Robust higher order statistics," in *Proc. Int. Workshop Artif. Intell. Statist. (AISTATS)*, 2005, pp. 405–412.

[22] A. Hyvarinen, "New approximations of differential entropy for independent component analysis," in *Advances in Neural Information Processing System*. Cambridge, MA: MIT Press, 1998, vol. 10, pp. 273–279.

[23] J. Weston, S. Mukherjee, O. Chapelle, M. Pontil, T. Poggio, and V. Vapnik, "Feature selection for SVMs," in *Advances in Neural Information Processing Systems*. Cambridge, MA: MIT Press, 2001.

[24] V. N. Vapnik, *Statistical Learning Theory*. New York: Wiley, 1998.

[25] L. Devroye, L. Györfi, and G. Lugosi, *A Probabilistic Theory of Pattern Recognition*. New York: Springer-Verlag, 1996.

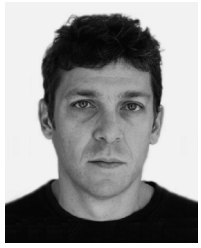
[26] J. M. Leiva and A. Artés, "A fixed-point algorithm for finding the optimal covariance matrix in kernel density modeling," in *Int. Conf. Acoust., Speech, Signal Process.*, 2006, pp. V-705–V-708.

[27] J. L. Marple, *Digital Spectral Analysis With Applications*. Englewood Cliffs, NJ: Prentice-Hall, 1987.

[28] J. Rissanen, "Information theory and neural nets," in *Mathematical Perspectives on Neural Networks*, P. Smolensky, Ed. Upper Saddle River, NJ: Lawrence Erlbaum, 1993.

Jose Miguel Leiva-Murillo (S'02) was born in Almería, Spain, in 1977. He received the M.Sc. degree in telecommunication engineering from the University of Málaga, Málaga, Spain, in 2001.

He is currently an Assistant Professor at the University Carlos III, Madrid, Spain. His research interests include feature extraction, pattern recognition, information theory, and machine learning.



Antonio Artés-Rodríguez (M'89–SM'01) was born in Alhama de Almería, Spain, in 1963. He received the Ingeniero de Telecomunicación and Doctor Ingeniero de Telecomunicación degrees from the Universidad Politécnica de Madrid, Spain, in 1988 and 1992, respectively.

He is a Professor at the Department of Signal Theory and Communications, Universidad Carlos III de Madrid, Spain. Before, he has occupied different teaching positions at Universidad de Vigo, Universidad Politécnica de Madrid, and Universidad de Alcalá, all of them in Spain. He has participated in more than 50 projects and contracts and he has coauthored more that 100 journal and international conference papers. His research interests include signal processing, learning, and information theory methods, and its application to sensor networks, communications, and medical applications.