This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

Optimized Kernel Entropy Components

Emma Izquierdo-Verdiguier, Valero Laparra, Robert Jenssen, *Senior Member, IEEE*, Luis Gómez-Chova, *Senior Member, IEEE*, and Gustau Camps-Valls, *Senior Member, IEEE*

***Abstract*— This brief addresses two main issues of the standard kernel entropy component analysis (KECA) algorithm: the optimization of the kernel decomposition and the optimization of the Gaussian kernel parameter. KECA roughly reduces to a sorting of the importance of kernel eigenvectors by entropy instead of variance, as in the kernel principal components analysis. In this brief, we propose an extension of the KECA method, named optimized KECA (OKECA), that directly extracts the optimal features retaining most of the data entropy by means of compacting the information in very few features (often in just one or two). The proposed method produces features which have higher expressive power. In particular, it is based on the independent component analysis framework, and introduces an extra rotation to the eigen decomposition, which is optimized via gradient-ascent search. This maximum entropy preservation suggests that OKECA features are more efficient than KECA features for density estimation. In addition, a critical issue in both the methods is the selection of the kernel parameter, since it critically affects the resulting performance. Here, we analyze the most common kernel length-scale selection criteria. The results of both the methods are illustrated in different synthetic and real problems. Results show that OKECA returns projections with more expressive power than KECA, the most successful rule for estimating the kernel parameter is based on maximum likelihood, and OKECA is more robust to the selection of the length-scale parameter in kernel density estimation.**

***Index Terms*— Density estimation, entropy component analysis, feature extraction, kernel methods.**

I. Introduction

The kernel entropy component analysis (KECA) [1], [2] was recently proposed as a general information-theoretic method for feature extraction and dimensionality reduction in pattern analysis and machine intelligence. It has proven useful in different applications, e.g., remote sensing data analysis [3]–[5], face recognition [6], chemical processes modeling [7], high-dimensional celestial spectra reduction [8], and audio processing [9]. Several extensions have been proposed for feature selection [10], class-dependent feature extraction [11], and semisupervised learning as well [12]. The KECA algorithm is fundamentally different from, but still intimately related to, the vastly successful spectral kernel multivariate signal processing methods, such as kernel principal components analysis (KPCA) [13], kernel canonical correlation analysis (KCCA) [14], and kernel partial least squares (KPLS) [15], just to name a few [16].

Manuscript received May 18, 2015; revised February 2, 2016; accepted February 4, 2016. This work was supported in part by the Fonds Européen de Développement Économique et Régional, in part by the Spanish Ministry of Economy and Competitiveness through the LIVE-VISON Project under Grant TIN2012-38102-C03-01, in part by Generalitat Valenciana under Project GV/2013/079, in part by the European Research Council through the SEDAL Project under the ERC consolidator grant SEDAL-647423, and in part by the Research Council of Norway through the FRIPRO/IKTPLUSS Project under Grant 234498.

E. Izquierdo-Verdiguier, V. Laparra, L. Gómez-Chova, and G. Camps-Valls are with the Image Processing Laboratory, Universitat de València, Valencia 46980, Spain (e-mail: emma.izquierdo@uv.es; valero.laparra@uv.es; luis.gomez-chova@uv.es; gustau.camps@uv.es).

R. Jenssen is with the University of Tromsø, Tromsø 9037, Norway (e-mail: robert.jenssen@uit.no).

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/TNNLS.2016.2530403

One distinguishing feature of KECA is that the method originates from kernel density estimation (KDE) [17]–[19], as do, e.g., principal curves estimation [20] and the family of information-theoretic learning methods [21]. In the KDE, the key is the kernel function, locally approximating the underlying probability density function (pdf). This in turn enables the estimation of entropy, and a quantity that describes the shape of the pdf [22]. The KDE kernel must be a nonnegative function that integrates to one (i.e., a density) but need not be positive semidefinite (PSD). The KDE kernel is versatile in that sense. However, many KDE kernels are PSD, and well-known examples include the Gaussian kernel, the Student kernel, and the Laplacian kernel [23] functions.

If the KDE kernel used in the KECA is PSD, then there are close relations to the aforementioned kernel signal processing methods, in the sense that the kernel computes an inner product in a reproducing kernel Hilbert space (RKHS). In this situation, KECA, KPCA, KCCA, and KPLS are based on RKHS learning algorithms to maximize, e.g., the feature space variance, correlation or alignment with the output variables. PSD-KECA hence bridges KDE, information-theoretic learning, and RKHS learning.

Although both the KDE and RKHS kernel methods have experienced great success, all kernel-based methods, including the one in this brief, are sensitive to the kernel function used. For instance, many kernel methods depend heavily on a bandwidth, or length-scale, parameter. In addition, all the aforementioned spectral methods may need a considerable number of components (eigenvalues and eigenvectors) in order to properly describe the data. This may be undesirable, e.g., in compression and data visualization contexts.

In this brief, we take advantage of the KDE foundation of KECA (see also [18] for further details), and introduce an optimization procedure aiming at compressing the entropy information into optimal directions in feature space. The proposed approach is the first kernel-based unsupervised feature extraction method in an information-theoretical sense. The approach introduces a rotation procedure that resembles the one in fast independent component analysis (ICA) [24]. Extracted optimized KECA (OKECA) components present two major advantages.

1) OKECA shows great robustness to the kernel bandwidth parameter. This is important, as there is no universally accepted kernel size selection procedure for unsupervised KDE-based kernel methods.
2) We use OKECA in order to improve the KDE. This is achieved with far fewer components compared with the KECA.

The rest of this brief is organized as follows. Section II presents the OKECA formulation and proposes a density estimation that exploits kernel feature characteristics. Section III is devoted to the analysis of the results. We use OKECA as a feature extraction method, analyze the retained entropy, show the estimated pdf, and perform data classification. Finally, the conclusion is drawn in Section IV.

II. Optimized Kernel Entropy Components

The KECA relies on the eigen decomposition of the uncentered kernel matrix (shown in the following) and sorts the eigenvectors according to the so-called entropy values of the projections. This is tightly related to information-theoretic concepts and KDE. The entropy-relevant dimensionality reduction transforms the data set in a way that reveals cluster structure and hence information about the underlying classes in the data [2], [25].

A. Kernel Entropy Components

To be more precise, the measure of information used in [1] is Renyi's second-order entropy, given by

$$H = -\log \int p^2(\mathbf{x})d\mathbf{x} \tag{1}$$

where $p(\mathbf{x})$ is the pdf generating the data. Given a data set $\mathcal{D} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ of dimensionality d, the entropy may be estimated through the KDE [17] (see Section II-D) as $-\log \nu$, where ν is the so-called information potential [21]

$$\nu = \frac{1}{n^2}\mathbf{1}_n^\top \mathbf{K}\mathbf{1}_n \tag{2}$$

where $\mathbf{K}_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$ is any valid KDE kernel comprising the $(n \times n)$ kernel matrix, and $\mathbf{1}_n$ is an n-dimensional vector of ones. Using the kernel decomposition introduced in [1]

$$\mathbf{K} = \mathbf{A}\mathbf{A}^\top = (\mathbf{E}\mathbf{D}^{\frac{1}{2}})(\mathbf{D}^{\frac{1}{2}}\mathbf{E}^\top) \tag{3}$$

we may write

$$\nu(N_c) = \sum_{j=1}^{N_c}\left(\sum_{i=1}^{n}\mathbf{A}_{ij}\right)^2 = \sum_{j=1}^{N_c}\left(\lambda_j^{\frac{1}{2}}\mathbf{1}_n^\top \mathbf{e}_j\right)^2. \tag{4}$$

In this expression, $\mathbf{E}$ contains the eigenvectors in columns, $\mathbf{E} = [\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n]$, and $\mathbf{D}$ is a diagonal matrix containing the eigenvalues of $\mathbf{K}$, i.e., $\mathbf{D}_{ii} = \lambda_i$, and $N_c \leq n$ is the number of retained components. The terms $(\lambda_j^{(1/2)}\mathbf{1}_n^\top \mathbf{e}_j)^2$ are denoted entropy values.

As mentioned earlier, if the KDE kernel is PSD, then there is a close connection between the KECA and the uncentered KPCA, since the kernel function in that case reproduces the dot product between two samples mapped to an RKHS $\mathcal{H}$ via $\boldsymbol{\phi}(\cdot)$, i.e., $\mathbf{K}_{ij} = k(\mathbf{x}_i, \mathbf{x}_j) = \boldsymbol{\phi}(\mathbf{x}_i)^\top \boldsymbol{\phi}(\mathbf{x}_j)$. Note that centering of the kernel matrix $\mathbf{K}$ makes no sense in the KDE and entropy context of (2), as this would correspond to $\nu = 0$, i.e., infinite entropy. Hence, $\mathbf{D}^{(1/2)}\mathbf{E}^\top$ is the uncentered projection of the feature space data $\mathcal{D}_\mathcal{H} = \{\boldsymbol{\phi}(\mathbf{x}_1), \ldots, \boldsymbol{\phi}(\mathbf{x}_n)\}$ onto all the principal axes in the feature space as [1] and [2]. These projections may be sorted according to their contribution to the input space entropy as measured by the information potential [the entropy values in (4)], constituting the KECA procedure.

However, the projections and their entropy content are fully dependent on the quality of the KDE performed via the kernel function. Moreover, using the eigen-decomposition procedure may not be optimal to find the best projections from an entropy perspective.

B. Proposed Optimized KECA

The novel approach proposed in this brief searches for a basis that maximizes the information potential as few components as possible. Toward that end, we propose a new optimal (in information-theoretic sense) feature extraction and unsupervised dimensionality reduction method. The procedure corresponds to optimally capturing in these components the high information potential part of the data (low entropy), which typically corresponds to the structure of the data in terms of class or cluster information.

Unlike the KECA method which only applies a different sorting of KPCA features, the proposed method is motivated by the classical ICA formulation [26] in which, after the whitening step (applying $\mathbf{D}^{(1/2)}\mathbf{E}^\top$), there is an extra rotation (applying $\mathbf{W}^\top$) that maximizes the independence between the components. Note that $\mathbf{W}$ is an orthonormal linear transformation, i.e., $\mathbf{W}\mathbf{W}^\top = \mathbf{I}$. Similar ideas have been applied in the kernel-based components analysis [27]. Following the ICA rationale, we now aim at a new kernel matrix decomposition:

$$\mathbf{K} = \mathbf{B}\mathbf{B}^\top = (\mathbf{E}\mathbf{D}^{\frac{1}{2}}\mathbf{W})(\mathbf{W}^\top \mathbf{D}^{\frac{1}{2}}\mathbf{E}^\top). \tag{5}$$

Note that the kernel matrix does not change, but the modification allows us to directly find the basis that maximize the information potential with respect to the number of retained components. Therefore, for each column vector $\mathbf{w}_k$ in $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_n]$, we maximize

$$\mathcal{L} = \left(\mathbf{1}_n^\top \mathbf{E}\mathbf{D}^{\frac{1}{2}}\mathbf{w}_k\right)^2 \tag{6}$$

where each $\mathbf{w}_k$ is restricted to be normal $\|\mathbf{w}_k\|_2 = 1$ and to be orthonormal to the previous $\mathbf{w}_l$, $\forall l < k$. This deflationary procedure ensures that the obtained solution retains more (or equal) information potential than the one obtained by the standard KECA in fewer components.

In order to solve the OKECA optimization problem in (6), a gradient-ascent approach can be followed:

$$\mathbf{w}_k(t+1) = \mathbf{w}_k(t) + \tau \frac{\partial \mathcal{L}}{\partial \mathbf{w}_k(t)} \tag{7}$$

where τ is the step size and the gradient is

$$\frac{\partial \mathcal{L}}{\partial \mathbf{w}_k} = 2\left(\mathbf{1}_n^\top \mathbf{E}\mathbf{D}^{\frac{1}{2}}\mathbf{w}_k\right)\left(\mathbf{1}_n^\top \mathbf{E}\mathbf{D}^{\frac{1}{2}}\right)^\top. \tag{8}$$

In this brief, we adopted a simple scheme for early stopping that ensures that the gradient achieves a region, where additional iterations did not modify the cost function significantly. A pseudocode summary of the OKECA feature extraction procedure is given in Algorithm 1. A MATLAB implementation of the algorithm is available at http://isp.uv.es/code/okeca.htm for the interested reader. While other more sophisticated optimization algorithms could be deployed here, in our experiments, we observed that this simple gradient-ascent strategy consistently performed even in the presence of noise.

Algorithm 1 OKECA Feature Extraction

input $\mathbf{K}$
output $\mathbf{B}$, $\mathbf{W}$
$[\mathbf{E}, \mathbf{D}] \leftarrow$ eig of $\mathbf{K}$
Initialize τ, $\mathbf{W}$
for t iterations **do**
  $\mathbf{dJ} \leftarrow 2(\mathbf{1}_n^\top \mathbf{E}\mathbf{D}^{\frac{1}{2}}\mathbf{w}_k)(\mathbf{1}_n^\top \mathbf{E}\mathbf{D}^{\frac{1}{2}})^\top$
  $\mathbf{w}_k(t+1) \leftarrow \mathbf{w}_k(t) + \tau\ \mathbf{dJ}$
  $\mathbf{E} \leftarrow (\mathbf{E}\mathbf{D}^{\frac{1}{2}})\ \mathbf{w}_k(t+1)$
end for
$\mathbf{B} \leftarrow \mathbf{E}\mathbf{D}^{\frac{1}{2}}\mathbf{W}$
$\nu \leftarrow \sum_j (\lambda_j^{\frac{1}{2}}\mathbf{1}^\top \mathbf{e}_j)^2$
$\mathbf{B} \leftarrow \mathrm{sort}_\nu\ \mathbf{B}$, $\mathbf{W} \leftarrow \mathrm{sort}_\nu\ \mathbf{W}$

C. Computational Cost

The proposed method is particularly promising for dimensionality reduction, since it compacts the information in very few features with higher expressive power. When using this method as the first step in a data processing pipeline, its properties will help to reduce the burden and time computing in the next processing steps. The application of the OKECA method only requires the kernel generation and a matrix multiplication, as in the KPCA and KECA. That means that once it is trained its application to new samples is not demanding. However, the training step is more computationally demanding than KECA. While KECA method is based on an eigen decomposition of a kernel matrix, OKECA additionally requires the gradient-ascent procedure in order to refine the obtained features. In particular, the KECA method requires $O(n^3)$, while OKECA spends $O(n^3 + 4tn^2)$, where n is the number of samples and t is the number of iterations of the gradient-ascent process. Note that the number of iterations to converge t depends on the particular problem at hand, but importantly, this just has to be computed once. After the kernel is optimized, entropic decomposition is learned, and the application is straightforward and as demanding as the KPCA and KECA counterparts.

D. Kernel Decomposition in Density Estimation

This section illustrates the benefits of using the proposed decomposition for KDE [28]. KDE is a classical method for estimating a pdf in a nonparametric way. In particular, KDE defines the pdf as a sum of kernel functions, $k(\cdot, \mathbf{x}_i)$, defined over the training samples $\mathbf{x}_i$ of the data set $\mathcal{D}$ as follows:

$$\hat{p}(\mathbf{x}_*) = \frac{1}{n}\sum_{i=1}^{n} k(\mathbf{x}_*, \mathbf{x}_i). \quad (9)$$

As mentioned earlier, KDE kernel functions need not in general be PSD but have to be nonnegative and integrate to one to ensure that $\hat{p}$ is a valid pdf. A classical example of such a kernel function is the Gaussian distribution, $k(\mathbf{x}_*, \mathbf{x}_i) = (2\pi\sigma^d)^{-1/2} \exp(-\|\mathbf{x}_* - \mathbf{x}_i\|^2/(2\sigma^2))$, but as mentioned, other choices exist. Then, the corresponding kernel matrix can be used for KDE

$$\hat{p}(\mathbf{x}_*) = \frac{1}{n}\sum_{i=1}^{n} k(\mathbf{x}_*, \mathbf{x}_i) = \frac{1}{n}\mathbf{1}_n^\top \mathbf{k}_* \quad (10)$$

where $\mathbf{k}_*$ is the vector of kernel evaluations between the point of interest $\mathbf{x}_*$ and all samples in the training data set $\mathcal{D}$. As explained in [18], if the decomposition of the uncentered kernel matrix follows the form $\mathbf{K} = \mathbf{E}\mathbf{D}\mathbf{E}^\top$, where $\mathbf{E}$ is orthonormal and $\mathbf{D}$ is a diagonal matrix, then the KDE may be expressed as:

$$\hat{p}(\mathbf{x}_*) = \mathbf{1}_n^\top \mathbf{E}_r \mathbf{E}_r^\top \mathbf{k}_* \quad (11)$$

where $\mathbf{E}_r$ is the reduced version of $\mathbf{E}$ by keeping columns for $r < n$. Note that when using $\mathbf{E}$ instead of $\mathbf{E}_r$, (11) reduces to (10). This shows that retained KECA components may be used for KDE [18], by selecting the dimensions that maximize the information potential in (4).

A novel aspect in this brief is to use the OKECA components for KDE in a similar manner. Note that the KECA decomposition in (3) is not exactly the same as the proposed OKECA in (5). Nevertheless, it is easy to find a basis that fulfills the same decomposition form, i.e., $\widetilde{\mathbf{E}}\widetilde{\mathbf{D}} = \mathbf{B}$, where $\widetilde{\mathbf{E}}$ is the $\mathbf{B}$ matrix with normalized column vectors and $\widetilde{\mathbf{D}}$ is diagonal matrix containing the norms of each column in $\mathbf{B}$. Therefore, (11) is $\hat{p}(\mathbf{x}_*) = \mathbf{1}_n^\top \widetilde{\mathbf{E}}_r \widetilde{\mathbf{E}}_r^\top \mathbf{k}_*$. In Section III-C, we will empirically show that, even though the OKECA basis is not orthonormal in principle, it is possible to obtain an accurate pdf estimation.

E. Model Estimation

In this brief, we consider the Gaussian RBF kernel, since it is the most common in both the RKHS kernel methods and the KDE [28]. This kernel induces a probabilistic Gaussian mixture model, and it only introduces one scalar free parameter, σ. Note that more complicated models could be considered in both the frameworks. However, a recurrent and unsolved problem in both the approaches is the estimation of the length-scale parameter σ.

A plethora of heuristics and rules for estimating the length-scale have been proposed in the machine learning and statistics literatures. Roughly speaking, one finds two main approximations. The first approach considers maximizing a particular objective function through a cross-validation procedure. The objective function may be optimized using unsupervised (e.g., maximum likelihood [19], denoted by σ_{ML} in the experiments) or supervised (e.g., a classification accuracy score, denoted by σ_{class} in the experiments) approaches. The second approach resorts to empirical rules of performance or theoretical bounds. Good examples of this second approach, which are considered in this brief, are Silverman's rule [17], which is the classical rule of thumb in the KDE, σ_{Silv}, in the experiments, the mean distance between training points, which is a common approach in kernel methods for classification, σ_{d1}, in the experiments, and the 15% of the median distance between points, which is the classical employed in KECA, σ_{d2}, in the experiments.

III. EXPERIMENTS

We compare the performance of the standard KECA and the proposed OKECA for both density estimation and data classification. We analyze the methods in terms of the retained information potential as a function of the extracted features, the impact of the model selection criteria, and the classification accuracies in synthetic and real data sets.

A. OKECA for Optimally Entropic Representations

The first experiment considers three well-known 2-D toy examples for analyzing the methods: a ring-shaped distribution consisting of one class only, and the binary two moons and pinwheel data sets. In this section, we illustrate the ability of the proposed method to obtain projections that maximize the information potential, hence minimizing the squared Rényi entropy. In the results, we used 80, 20, and 45 training samples for each problem, respectively.

Fig. 1 shows the original data distributions and the estimated cumulative information potential [ν and $\mathcal{L}$ defined in (4) and (6), respectively] attained by the KECA and OKECA as a function of the ten top components and all the considered kernel length-scale selection criteria. For all data sets and for all σ values, OKECA reaches almost the maximum entropy value with just one feature, whereas KECA cumulative entropy values need five or more components to saturate. This behavior is almost independent of the chosen criterion to set the σ parameter. The higher information content may translate into more informative features potentially useful for density estimation and classification, as we illustrate in Sections III-B and III-C

B. OKECA for pdf Estimation

Fig. 2 shows the ability of KECA and OKECA for density estimation in the ring data set (see Fig. 1 to analyze the ring data set distribution). We merely applied (11) for different numbers of components r in $\mathbf{A}_r$. Note that, for the proposed OKECA, the first projection concentrates most of the entropy information. This agrees with the fact that just one dimension is needed to obtain a good pdf estimation. On the contrary, KECA cannot correctly estimate the pdf using only the first component and actually needs at least five components. This issue is even more dramatic when using σ_{ML} (see Fig. 2). It is worth noting that σ_{ML} and σ_{d2} give rise to the best pdf estimates.

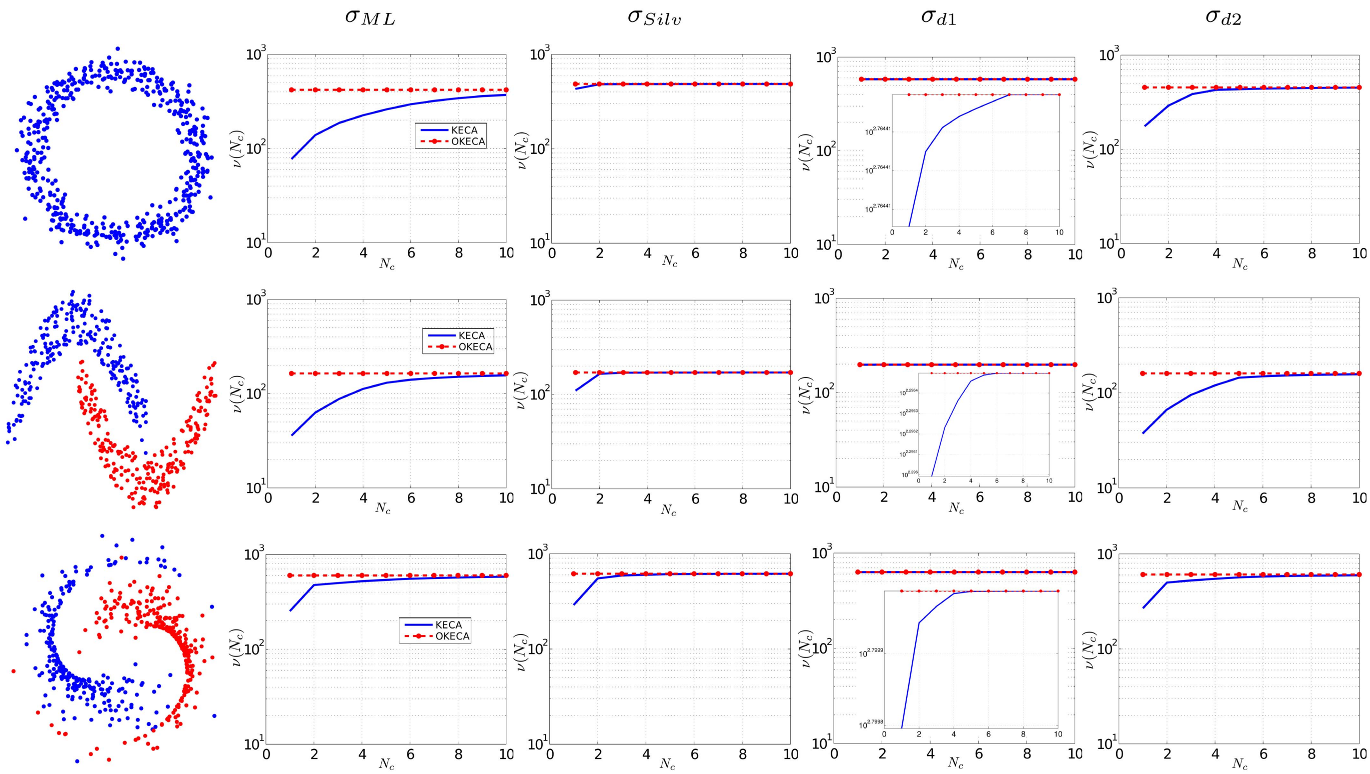

Fig. 1. Cumulative estimated information potential [$\nu(N_c)$] versus the number of dimension components for different toy data sets, using KECA and OKECA and different σ estimation approaches.

C. OKECA Components for Data Classification

We here illustrate the capabilities of OKECA for data classification. A great many feature extraction methods exist, both linear, such as PCA [29] or ICA-based methods [24], [30], and nonlinear, such as the family of kernel multivariate analysis [16]. In this brief, however, for the sake of a fare comparison, we restrict to compare our OKECA proposal to the original KECA counterpart, which are the only existing unsupervised feature extraction kernel methods based on the same principle of entropy maximization. The experiments are conducted on a wide range of synthetic and real problems: 1) the two moons and the pinwheel data sets considered previously in Section III-A; 2) six real data sets from the University California Irvine (UCI) Machine Learning Repository[1]; and 3) a real satellite multispectral image classification problem. In order to evaluate the data classification, we have used the overall classification accuracy (OA) which is obtained as the average of samples correctly predicted in percentage terms. While one could classify on top of the extracted features, we here rely intentionally on the class-dependent estimated densities and perform maximum *a posteriori* (MAP) classification.

1) Synthetic Data Sets: Fig. 3 shows the test OA obtained with different σ values and different numbers of retained dimensions with the KECA and the proposed OKECA on the two moons and pinwheel data sets. The five bars for every number of retained features are from left to right: σ_{d1}, σ_{d2}, σ_{Silv}, σ_{ML}, and σ_{class}. The value of σ_{class} has been optimized for classification using all features in a fivefold cross-validation scheme. We used 20 samples and 45 samples per class for training two moons and pinwheel, respectively, and 500 per class for testing the models and computing the test OA in both the data sets. Note that the OKECA method achieves better classification results than KECA for all σ values, confirming that to seek for optimally entropic data descriptors may benefit classification. Smaller differences between the methods are observed as the number of components increases. When all n features are used, OKECA and KECA are trivially equivalent.

In the following, we discuss the capabilities of OKECA in the presence of distorted distributions. The question raised is how sensitive is the optimization algorithm to the presence of noise. To this end, a toy example of the KECA and OKECA projections in the presence of noise is analyzed. We used 50 samples of two moons data set for training and 500 samples to test the classifier. Gaussian noise was added to the original data distributions by varying the dimensionwise standard deviation of the Gaussian noise σ_n from 0.001 to 0.091. Numerical results are shown in Fig. 4 (left). Note that one main aim of feature extraction and dimensional reduction methods is achieved using the proposed method: KECA needs at least 12 components to obtain similar results to the ones obtained by the OKECA with just one component, even in high-noise regime.

2) UCI Benchmark Data Sets: We used six data sets from the UCI machine learning repository of different sizes and dimensionality: the `Ionosphere` data set is a binary classification problem of the quality of the radar signal returned from the ionosphere; the goal for the `Letter` data set is to detect each of a large number of black-and-white rectangular pixel displays as one of the 26 capital letters in the English alphabet; the `Pendigits` problem deals with the recognition of pen-based handwritten digits; the `Pima-Indians` data set constitutes a classical problem of diabetes diagnosis in patients from clinical variables; the `Vowel` data set deals with the vowels detection problem in japanese and contains data from a large number of time series of cepstrum coefficients taken from speakers; and finally `wdbc` is another clinical problem for the diagnosis of breast cancer in malignant/benign classes. The data sets were

[1]http://archive.ics.uci.edu/ml/datasets.html

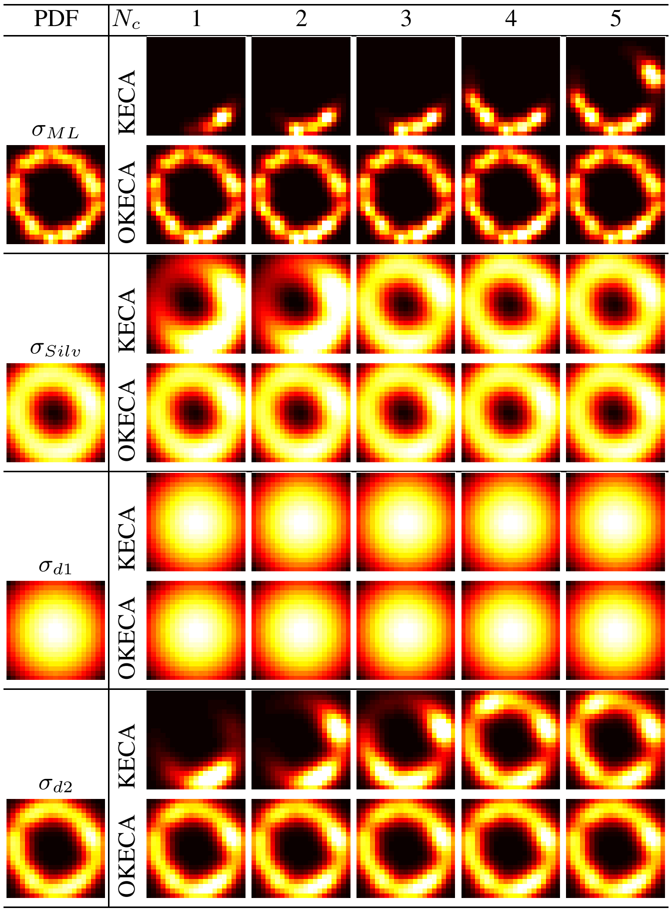

Fig. 2. Density estimation for the ring data set by the KECA and the OKECA using different numbers of extracted features N_c and approaches to estimate the kernel length-scale parameter σ. Black color: low pdf values. Yellow color: high pdf values.

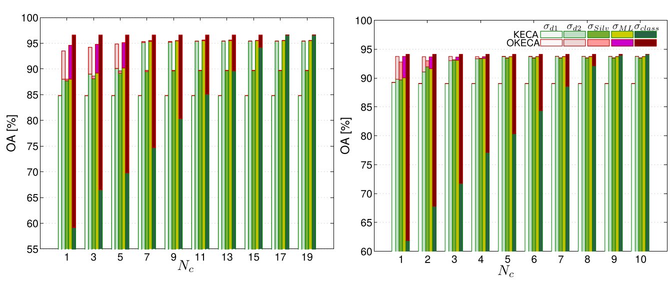

Fig. 3. Overall accuracy obtained with two moons (left) and pinwheel (right) data sets. The five bars for every number of retained features are from left to right: σ_{d1}, σ_{d2}, σ_{Silv}, σ_{ML}, and σ_{class}.

intentionally selected either because of the observed high collinearity between input features or because of the diversity in the number of classes. Table I gives details on the dimensionality, number of classes, and training and test samples used in the experiments that follow.

We run KECA and OKECA for all the data sets for different numbers of extracted components. The average of the OA for the ten first dimensions is shown in Fig. 4 (right). In this case, we restrict ourselves to σ_{ML} because of the good performance in the previous experiments and for the sake of simplicity. In general, the OKECA method outperforms the KECA method and, as observed before, OKECA saturates its performance with just the first extracted dimension.

TABLE I

UCI Database Description (d: Number of Dimensions, n_c: Number of Class, N_{train}: Number of Training Samples, and N_{test}: Number of Test Samples)

Database	m	d	n_c	N_{train}	N_{test}
Ionosphere	351	33	2	60	172
Letter	20000	16	26	780	3874
Pendigits	10992	16	9	450	3498
Pima-Indians	768	8	2	180	330
Vowel	990	12	10	100	330
wdbc	569	30	2	60	344

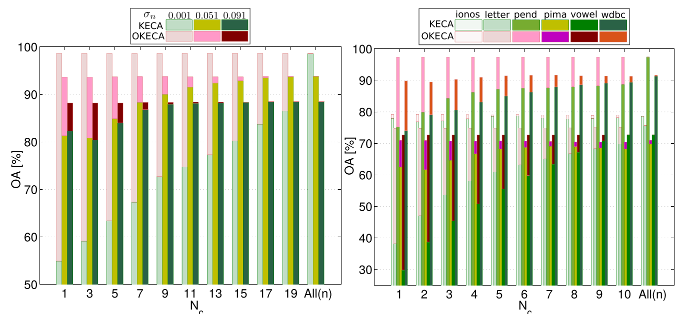

Fig. 4. Overall accuracy obtained by the KECA and OKECA methods using different values of noise (left) and UCI database (right) with different numbers of dimensions. The bars for every number of retained features are different noise values (left) and different databases (right), respectively.

3) Satellite Image Classification: In this experiment, we apply KECA and OKECA to the segmentation of remotely sensed multispectral images. Nowadays, sensors mounted on satellite or airborne platforms may acquire the reflected energy by the earth with high spatial detail and in several wavelengths or spectral channels [31]. This allows an improved detection and classification of the pixels in the scene. We consider a real multispectral image acquired over a residential neighborhood of the city of Zürich by the QuickBird satellite in 2002. The analyzed image has 329 pixel $\times$ 347 pixel. Additional spatial information was added by means of morphological operators, so the data set has 22 input features. The images contain nine classes of interest: water, meadows, trees, asphalt, brick roofs, bitumen, parking lots, bare soil, and shadows. The classes of training samples have been labeled by photointerpretation. The considered data not only are high dimensional but also show high collinearity, since spectral and spatial features are stacked together at a pixel level. The problem may be quite challenging for classification and feature extraction.

The KECA and OKECA cumulative information potential values follow similar trends to the toy examples (see Fig. 5). OKECA reaches the maximum with just one feature, while KECA needs much more components to achieve similar informative content, especially noticeable for σ_{ML} and σ_{d2} criteria. Such dependence with the criterion is not shared by OKECA. These results suggest that the sharpness in the component selection made by OKECA is relevant in the cases of high feature redundancy as well.

Fig. 6 shows the classification results obtained using different σ values and different numbers of retained dimensions. In this case, we use 22 and 200 samples per class for training and testing the models, respectively. Both the methods achieve the best results using σ_{ML} and σ_{class} criteria. Finally, note that σ_{d1}, which is a common choice in unsupervised kernel methods, provides very poor results for both the methods. Fig. 7 shows the classification MAPs obtained using three retained features and σ_{class} for both the methods. Note how OKECA outperforms KECA in general for all the classes.

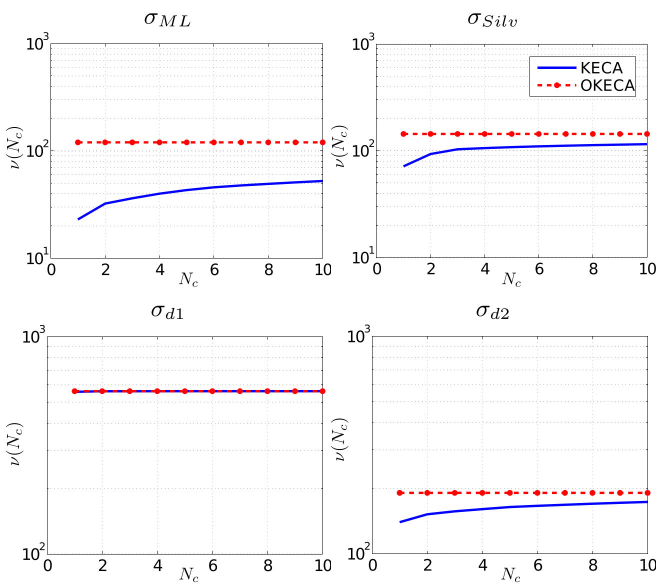

Fig. 5. Cumulative information potential for the multispectral image data set using the KECA and OKECA and different σ estimation approaches. Results for KECA and OKECA in σ_{d1} are equal (appear overlapped).

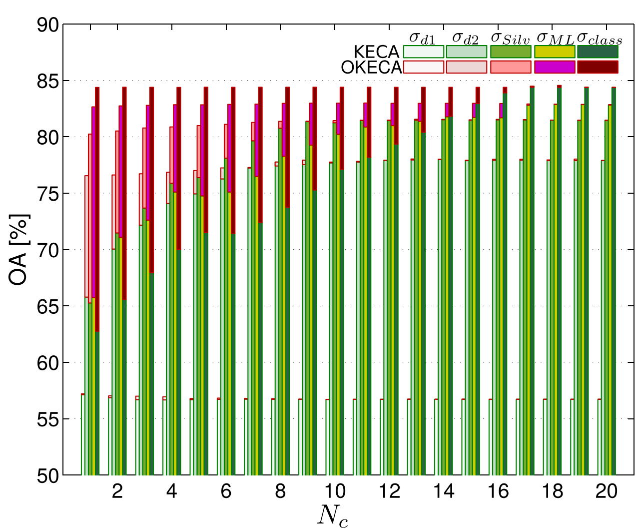

Fig. 6. Classification results for the Zürich QuickBird satellite image for different σ values and numbers of retained dimensions by KECA and OKECA. The five bars for every number of retained features are from left to right: σ_{d1}, σ_{d2}, σ_{Silv}, σ_{ML}, and σ_{class}.

D. Discussion

The OKECA potential has been shown in the different experiments. In all of them, the proposed method presents an extraordinary advantage: the information is compacted in very few features (often in just one or two) with higher expressive power optimizing the information potential. That is demonstrated from experimental viewpoint, not only in pdf estimation but also in classification tasks. As we have shown, the proposed method reduces the number of features clearly required for improving the results. OKECA correctly estimates the pdf using one dimension concentrating most of the entropy information better than the KECA method, and it is more robust to the selection of the kernel parameter. Furthermore, the experiments show an improvement of the classification results even in the presence of noise. In some situations, using just one OKECA feature is enough to achieve the best overall performance, extracting more components does not add new complementary information and classification results do not change significantly.

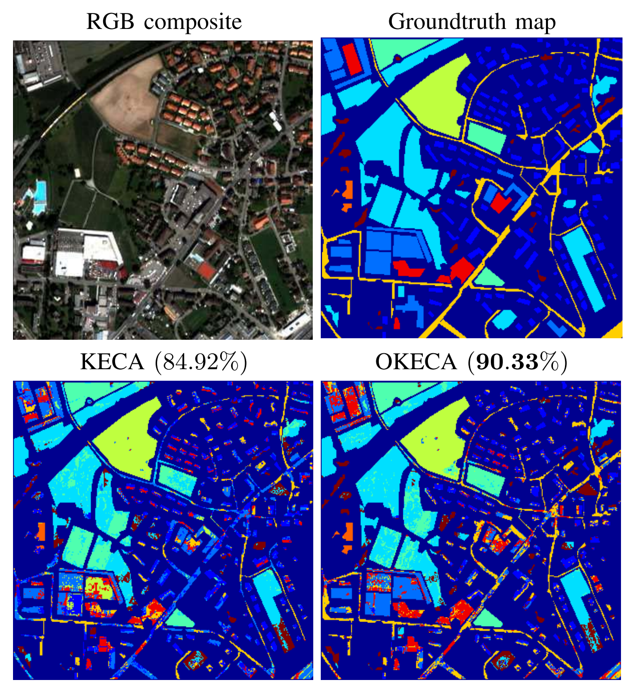

Fig. 7. Classification MAPs for the Zürich QuickBird satellite image using three features and σ_{class} in the KECA and OKECA. Top-left: RBG version of the original image. Top-right: ground truth classification MAP (each color: different land-cover classes). Bottom-left: classification MAP obtained with KECA. Bottom-right: classification MAP obtained with OKECA.

IV. Conclusion

We proposed a highly efficient modification of the KECA algorithm for the optimal extraction of entropic kernel components. While KECA reduces to sort the kernel eigenvectors by entropy, OKECA explicitly searches for the features that retain most informative content. We have illustrated the ability of OKECA to retain more information in pdf estimation and classification on both synthetic and real examples. Results consistently showed that the OKECA outperforms KECA in terms of information content and robustness. In fact, in many experiments, just one or two OKECA components retain almost all the relevant information for data description. Accounting for optimal entropic features allows us to improve the description of the density shape, which is in turn the core for pdf estimation and pdf-based classification. Furthermore, we have analyzed the effect of using different unsupervised rules to fit the RBF kernel length-scale parameter on KECA performance and OKECA performance. In general, the maximum likelihood approach showed the best performance.

References

[1] R. Jenssen, "Kernel entropy component analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 5, pp. 847–860, May 2010.

[2] R. Jenssen, "Entropy-relevant dimensions in the kernel feature space: Cluster-capturing dimensionality reduction," *IEEE Signal Process. Mag.*, vol. 30, no. 4, pp. 30–39, Jul. 2013.

[3] L. Gómez-Chova, R. Jenssen, and G. Camps-Valls, "Kernel entropy component analysis for remote sensing image clustering," *IEEE Geosci. Remote Sens. Lett.*, vol. 9, no. 2, pp. 312–316, Mar. 2012.

[4] X. Luo and X. Wu, "Fusing remote sensing images using a statistical model," *Appl. Mech. Mater.*, vols. 263–266, pp. 416–420, Dec. 2012.

[5] X. Luo, X. Wu, and Z. Zhang, "Regional and entropy component analysis based remote sensing images fusion," *J. Intell. Fuzzy Syst.*, vol. 26, no. 3, pp. 1279–1287, 2013.

[6] B. H. Shekar, M. S. Kumari, L. M. Mestetskiy, and N. F. Dyshkant, "Face recognition using kernel entropy component analysis," *Neurocomputing*, vol. 74, no. 6, pp. 1053–1057, 2011.

[7] Q. Jiang, X. Yan, Z. Lv, and M. Guo, "Fault detection in nonlinear chemical processes based on kernel entropy component analysis and angular structure," *Korean J. Chem. Eng.*, vol. 30, no. 6, pp. 1181–1186, 2013.

[8] Y. D. Hu, J. C. Pan, and X. Tan, "High-dimensional data dimension reduction based on KECA," *Appl. Mech. Mater.*, vols. 303–306, pp. 1101–1104, Feb. 2013.

[9] Z. Xie and L. Guan, "Multimodal information fusion of audio emotion recognition based on kernel entropy component analysis," in *Proc. IEEE Int. Symp. Multimedia*, Dec. 2012, pp. 1–8.

[10] Z. Zhang and E. R. Hancock, "Kernel entropy-based unsupervised spectral feature selection," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 26, no. 5, p. 1260002, 2012.

[11] M. Cheng, C.-M. Pun, and Y. Y. Tang, "Nonnegative class-specific entropy component analysis with adaptive step search criterion," *Pattern Anal. Appl.*, vol. 17, no. 1, pp. 113–127, 2011.

[12] J. N. Myhre and R. Jenssen, "Mixture weight influence on kernel entropy component analysis and semi-supervised learning using the Lasso," in *Proc. IEEE Int. Workshop Mach. Learn. Signal Process.*, Santander, Spain, Sep. 2012, pp. 1–6.

[13] B. Schölkopf, A. Smola, and K. Müller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural Comput.*, vol. 10, no. 5, pp. 1299–1319, 1998.

[14] P. L. Lai and C. Fyfe, "Kernel and nonlinear canonical correlation analysis," *Int. J. Neural Syst.*, vol. 10, pp. 365–377, Oct. 2000.

[15] R. Rosipal and L. J. Trejo, "Kernel partial least squares regression in reproducing kernel Hilbert space," *J. Mach. Learn. Res.*, vol. 2, pp. 97–123, Mar. 2001.

[16] J. Arenas-García, K. Petersen, G. Camps-Valls, and L. K. Hansen, "Kernel multivariate analysis framework for supervised subspace learning: A tutorial on linear and kernel multivariate methods," *IEEE Signal Process. Mag.*, vol. 30, no. 4, pp. 16–29, Jul. 2013.

[17] B. W. Silverman, *Density Estimation for Statistics and Data Analysis*. London, U.K.: Chapman & Hall, 1986.

[18] M. Girolami, "Orthogonal series density estimation and the kernel eigenvalue problem," *Neural Comput.*, vol. 14, no. 3, pp. 669–688, Mar. 2002.

[19] R. P. W. Duin, "On the choice of smoothing parameters for Parzen estimators of probability density functions," *IEEE Trans. Comput.*, vol. C-25, no. 11, pp. 1175–1179, Nov. 1976.

[20] U. Ozertem and D. Erdogmus, "Locally defined principal curves and surfaces," *J. Mach. Learn. Res.*, vol. 12, pp. 1249–1286, Feb. 2011. [Online]. Available: http://dl.acm.org/citation.cfm?id=1953048.2021041.

[21] J. C. Principe, *Information Theoretic Learning: Renyi's Entropy and Kernel Perspectives*. New York, NY, USA: Springer-Verlag, 2010.

[22] T. M. Cover and J. A. Thomas, "Entropy, relative entropy, and mutual information," in *Elements of Information Theory*. New York, NY, USA: Wiley, 2005.

[23] J. Kim and C. D. Scott, "Robust kernel density estimation," *J. Mach. Learn. Res.*, vol. 13, no. 1, pp. 2529–2565, Jan. 2012. [Online]. Available: http://dl.acm.org/citation.cfm?id=2503308.2503323.

[24] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*. New York, NY, USA: Wiley, 2001.

[25] R. Jenssen, "Information theoretic learning and kernel methods," in *Information Theory and Statistical Learning*, F. Emmert-Streib and M. Dehmer, Eds. New York, NY, USA: Springer, 2009, pp. 209–230. [Online]. Available: http://dx.doi.org/10.1007/978-0-387-84816-7_9.

[26] A. Hyvärinen and E. Oja, "Independent component analysis: A tutorial," *Neural Netw.*, vol. 13, nos. 4–5, pp. 411–430, 2000.

[27] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang, "Domain adaptation via transfer component analysis," *IEEE Trans. Neural Netw.*, vol. 22, no. 2, pp. 199–210, Feb. 2011.

[28] E. Parzen, "On estimation of a probability density function and mode," *Ann. Math. Statist.*, vol. 33, no. 3, pp. 1065–1076, Sep. 1962.

[29] I. T. Jolliffe, *Principal Component Analysis*. New York, NY, USA: Springer-Verlag, 1986.

[30] N. Mammone, F. La Foresta, and F. C. Morabito, "Automatic artifact rejection from multichannel scalp EEG by wavelet ICA," *IEEE Sensors J.*, vol. 12, no. 3, pp. 533–542, Mar. 2012.

[31] G. Camps-Valls, D. Tuia, L. Gómez-Chova, S. Jiménez, and J. Malo, *Remote Sensing Image Processing* (Synthesis Lectures on Image, Video, and Multimedia Processing). San Rafael, CA, USA: Morgan & Claypool, 2011.