Title: Semisupervised Kernel Feature Extraction for Remote Sensing Image Analysis

This paper appears in: IEEE Transactions on Geoscience and Remote Sensing

Date of Publication: 2014

Author(s): Emma Izquierdo-Verdiguier, Luis Gómez-Chova, Lorenzo Bruzzone, and Gustavo Camps-Valls

Volume:-, Issue: -

Page(s): -

DOI: 10.1109/TGRS.2013.2290372

Semi-supervised Kernel Feature Extraction for Remote Sensing Image Analysis

Emma Izquierdo-Verdiguier, *Student Member, IEEE,* Luis Gómez-Chova, *Member, IEEE,* Lorenzo Bruzzone, *Fellow, IEEE* and Gustavo Camps-Valls, *Senior Member, IEEE*

***Abstract*—This paper presents a novel semi-supervised kernel partial least squares (SS-KLPS) algorithm for non-linear feature extraction to tackle both land cover classification and biophysical parameter retrieval problems. The proposed method finds projections of the original input data that align with the target variable (labels) and incorporates the wealth of unlabeled information to deal with low-sized or under-represented datasets. The method relies on combining two kernel functions: the standard radial basis function (RBF) kernel based on labeled information, and a *generative*, i.e. probabilistic, kernel directly learned by clustering the data many times and at different scales across the data manifold. The construction of the kernel is very simple and intuitive: two samples should belong to the same class if they consistently belong to the same clusters at different scales. The effectiveness of the proposed method is successfully illustrated in multi- and hyper-spectral remote sensing image classification and biophysical parameter estimation problems. Accuracy improvements in the range between $+5$ and 15% over standard PCA, $+4$ and 15% over kernel PCA, and $+3$ and 10% over kernel partial least squares (KPLS) are obtained on several images. Average gain in RMSE of $+5\%$ and reductions in bias estimates of $+3\%$ are obtained for biophysical parameter retrieval compared to standard PCA feature extraction.**

***Index Terms*—Classification, biophysical parameter estimation, feature extraction, kernel methods, principal component analysis (PCA), partial least squares (PLS), clustering, semi-supervised learning, generative kernels**

I. Introduction

FEATURE extraction has become an important topic in remote sensing data processing mainly due to the high dimensionality of data, as well as the high redundancy among spectral bands. The problem is ubiquitous and very common in remote sensing image analysis. Moreover, the high-dimensionality of remote sensing data is often increased by stacking spatial, spectral, temporal and multiangular features to the spectral channels for modelling additional information sources. Feature extraction consists of identifying the most discriminative variables for data classification or regression. These variables are often associated with the most relevant directions in the data distribution. For example, feature extraction is typically conducted for reducing the dimensionality of hyperspectral images and infrared sounder imagery before classification and parameter retrieval.

The family of multivariate analysis methods for feature extraction is commonly used to reduce the data dimensionality by projecting examples onto the most relevant directions of the data manifold. Principal component analysis (PCA) [1] and partial least squares (PLS) [2] are two of the most common linear feature extraction methods in remote sensing data analysis. Other methods focus on including information about the noise, such as the minimum noise fraction (MNF) transform [3] or the related noise-adjusted principal components (NAPC) [4].

All previous methods assume that there exists a *linear* relation between the original data. In many situations, this linearity assumption does not hold, and a nonlinear feature extraction is needed to obtain an acceptable performance. Different nonlinear versions of PCA and PLS have been developed, which can address non-linear problems either by local approaches [5], neural networks [6], or kernel-based algorithms [7]. In the last decade, kernel methods have attracted the interest of the remote sensing community because they allow us to develop nonlinear models from linear ones in a very easy and intuitive way, and still require linear algebra [8]. Essentially, kernel methods map the input data into a high dimensional Hilbert space, $\mathcal{H}$, and define a linear method therein. The model results nonlinear with respect to the input space. Interestingly, there is no need to work explicitly with the mapped data, but one computes the nonlinear relations between data via a kernel (similarity) function that reproduces the similarity in $\mathcal{H}$. Kernel methods have in general good performance in the case of high dimensional problems with low number of training examples [7].

The kernel framework has been exploited not only for classification and regression but lately for nonlinear feature extraction. Note that while nonlinear classification or regression methods lead to black-box models, the idea underlying feature extraction is to find an appropriate data representation (typically via projection operators). This different perspective of addressing a problem leads to some interesting properties. The most important is that nonlinear features extracted with kernel methods can be used directly for general tasks including classification, regression, clustering, ranking, compression, or data visualization. Therefore, kernel methods have recently captured the interest of the scientific community for feature

EIV, LGC and GCV are with the Image Processing Laboratory (IPL), University of València, Paterna (València), Spain. E-mails: {emma.izquierdo,luis.gomez-chova,gustavo.camps}@uv.es

LB is with the Dipartimento di Ingegneria e Scienza dell'Informazione, Università degli Studi di Trento, Trento, Italy. E-mail: lorenzo.bruzzone@ing.unitn.it

This work has been partially supported under projects TIN2012-38102-C03-01 (LIFE-VISION) and the Generalitat Valenciana under project GV/2013/079.

extraction. For example, this is the approach used in kernel principal component analysis (KPCA) [7] and kernel partial least squares (KPLS) [9]. The main difference between KPCA and KPLS is that while KPCA finds the projections that maximize the variance of the input data in the feature space, KPLS extracts projections that account for both the projected input and target data (labels). A set of multivariate kernel feature extraction methods, such as kernel PCA (KPCA), kernel PLS, and kernel orthonormalized PLS (KOPLS), were proposed as a preprocessing step for hyperspectral image classification and canopy parameter retrieval [10]. In [11], KPCA was also used for target and anomaly detection, while the kernel nonparametric weighted feature extraction (KNWFE) was introduced for hyperspectral image classification in [12]. Recently, a kernel version of the maximum autocorrelation function (MAF) has been successfully presented for change detection [13], and further extended in [14] to define the signal-to-noise ratio explicitly in the kernel feature space. In [15], the kernel entropy component analysis (KECA) was presented for remote sensing: The method generates nonlinear features that reveal structure related to the Rényi entropy of the input space data set rather than the variance of the projections.

Each of the previous kernel methods focuses on extracting features that optimize a given criterion, e.g. directions that account for the maximum variance (jointly or not with those minimally affected by noise), the maximum entropy, the maximum Fisher's discriminant criterion, etc. In this paper, we focus on the KPLS feature extraction method, which proved to be effective for remote sensing data processing, extracting nonlinear features maximally aligned with the target variables (see e.g. [10], [16]). Data projected onto these features can be directly used in canonical *linear* classification or regression.

Extracting nonlinear features by KPLS is a very complex problem when relatively few labeled data points are available, which is a common situation in remote sensing data analysis problems. Including the information conveyed by unlabeled data via *semi-supervised learning* can potentially improve the feature extraction task. The semi-supervised framework has recently attracted a considerable amount of theoretical [17] as well as remote sensing applied research [8]. In this paper, we present a new semi-supervised KPLS method for nonlinear feature extraction. The features extracted with this method can be used as input for both classification and regression techniques. Our approach considers to modify the kernel similarity function via a kernel defined on the basis of clustering the analyzed image. Specifically, we propose to combine a standard radial basis function (RBF) kernel with a kernel constructed via clustering the data with the Expectation Maximization algorithm assuming a multiscale Gaussian mixture model (GMM). The RBF kernel is a universal kernel that has a stable behaviour and only incorporates one free parameter to be tuned. In this work, we built the RBF kernel only with the labeled samples. The second probabilistic kernel, denoted as *cluster kernel* focuses on the combination of labeled and unlabeled information of the data manifold. It is a parameter-free kernel and captures different (local-to-global) scales of data relations across the manifold.

The novel contributions of this paper consist in: 1) the cluster kernel accounts for the local-to-global structure of the data manifold as it captures similarity between labeled pixels, using unlabeled pixels information at different scales; and 2) as a consequence of the combination between the two kernels considered here, the proposed semi-supervised KPLS (SS-KPLS) method improves results in both very high spatial resolution and hyperspectral image classification scenarios and in biophysical parameter retrieval experiments.

The rest of the paper is outlined as follows. Section 2 reviews the standard formulation of KPLS and highlights the problems encountered when dealing with very few labeled samples. This motivates the introduction of the proposed method in Section 3. Section 4 presents the datasets used and the experimental results in both classification and regression problems. Finally, Section 5 concludes this paper.

II. A REVIEW OF MULTIVARIATE KERNEL METHODS

This section reviews classical multivariate analysis techniques for both linear and nonlinear feature extraction. A family of methods which has been successful in remote sensing data processing, and has recently gathered increasing interest in the remote sensing community, consists of the kernel extensions of multivariate techniques such as PCA and PLS. In this section we pay special attention to kernel partial least squares (KPLS), since it will be the core algorithm of the proposed semi-supervised method.

A. Linear multivariate methods

Notationally, we are given a set of n training data pairs $\{\mathbf{x}_i, y_i\}_{i=1}^n$, with $\mathbf{x}_i \in \mathbb{R}^d$ where d is the number of dimensions. For classification problems, y_i are labels which are converted into vectors via 1-of-M standard encoding, where M is the number of classes, $\mathbf{y}_i \in \mathbb{R}^M$, $\mathbf{y}_{ij} = 1$ if sample $\mathbf{x}_i$ belongs to class j and $\mathbf{y}_{ij} = 0$ otherwise. For regression problems involving a continuous dependent variable, such encoding is not possible and we use directly $y_i \in \mathbb{R}$. By using matrix notation we can write, $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]^\top$ and $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_n]^\top$, where superscript $^\top$ denotes matrix or vector transposition. We denote by $\tilde{\mathbf{X}}$ and $\tilde{\mathbf{Y}}$ the centered versions of $\mathbf{X}$ and $\mathbf{Y}$, respectively. Note that, the operation of centering removes the mean of every variable in the corresponding matrix. Multivariate methods seek for projections of the input data $\mathbf{X}$ optimizing a particular criterion. For instance, PCA looks for projections preserving the maximum variance, while PLS searches for projections of the input data that maximally align with the output data (labels).

1) Principal component analysis (PCA): PCA is a widespread method for dimensionality reduction. It consists in projecting the input data set onto the directions of largest input variance. Thus, PCA only considers the input data and does not take into account any target data set, i.e. it is

an *unsupervised feature extraction* method. The criterion is expressed compactly as:

$$\begin{aligned} \text{PCA:} \qquad & \mathbf{U} = \arg\max_{\mathbf{U}} \ \ \text{Tr}\{\mathbf{U}^\top \mathbf{C}_{xx} \mathbf{U}\} \\ & \text{subject to:} \ \ \mathbf{U}^\top \mathbf{U} = \mathbf{I}, \end{aligned} \tag{1}$$

where $\mathbf{I}$ is the identity matrix of size d_f (number of extracted features), $\mathbf{C}_{xx} = \tilde{\mathbf{X}}^\top \tilde{\mathbf{X}}$ is the sample covariance matrix of input data, and $\mathbf{U}$ is the projection matrix to be estimated.

The main limitation of PCA is that it does not consider the target variables $\mathbf{Y}$ for the input vectors but simply performs a coordinate rotation that aligns the transformed axes with the directions of maximum variance of the original data distribution. Thus, there is no guarantee that the directions of maximum variance will contain good features for discrimination or regression.

2) Partial least squares (PLS): The PLS algorithm, developed by Herman Wold [18], is probably one of the simplest methods for *supervised feature extraction*, since only considers the input data and the target data sets. The central idea of PLS is to find the projection vectors that maximize the covariance between the projected input and output data, whose problem is expressed as:

$$\begin{aligned} \text{PLS:} \qquad & \mathbf{U}, \mathbf{V} = \arg\max_{\mathbf{U},\mathbf{V}} \ \ \text{Tr}\{\mathbf{U}^\top \mathbf{C}_{xy} \mathbf{V}\} \\ & \text{subject to:} \ \ \mathbf{U}^\top \mathbf{U} = \mathbf{V}^\top \mathbf{V} = \mathbf{I}, \end{aligned} \tag{2}$$

where $\mathbf{C}_{xy}$ is the covariance matrix of input and output data, and $\mathbf{V}$ is the projection matrix to be estimated for the output data set. In the literature, there are several variants of the PLS standard formulation (see [2] for a detailed overview). In this work, the singular-value decomposition (SVD) of $\mathbf{C}_{xy}$ has been used in order to solve the problem [19].

B. Kernel multivariate methods

All previous methods assume that there exists a *linear* relation between the original data matrices, $\tilde{\mathbf{X}}$ and $\tilde{\mathbf{Y}}$, and the extracted projections, $\tilde{\mathbf{X}}'$ and $\tilde{\mathbf{Y}}'$, respectively. However, in many situations this linearity assumption does not hold, and nonlinear feature extraction is needed to obtain acceptable performance. In this context, *kernel methods* are promising approaches.

1) Kernel mappings, functions and projections: Notationally, consider we are given a set of pairs $\{\boldsymbol{\phi}(\mathbf{x}_i), \mathbf{y}_i\}_{i=1}^n$, with $\boldsymbol{\phi}(\mathbf{x}) : \mathbb{R}^d \rightarrow \mathcal{H}$ a function that maps the input data into some *feature space* of very large or even infinite dimension. Data matrices for performing the linear feature extraction (e.g. PCA or PLS) in $\mathcal{H}$ are now given by $\boldsymbol{\Phi} = [\boldsymbol{\phi}(\mathbf{x}_1), \ldots, \boldsymbol{\phi}(\mathbf{x}_n)]^\top$ and $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_n]^\top$. As before, the centered versions of these matrices are denoted by $\tilde{\boldsymbol{\Phi}}$ and $\tilde{\mathbf{Y}}$.

Importantly, the projections of the input and output data will be given by $\tilde{\boldsymbol{\Phi}}' = \tilde{\boldsymbol{\Phi}}\mathbf{U}$ and $\tilde{\mathbf{Y}}' = \tilde{\mathbf{Y}}\mathbf{V}$, respectively, where the projection matrix $\mathbf{U}$ is now of size $\dim(\mathcal{H}) \times d_f$. Note, that the input covariance matrix in $\mathcal{H}$, which is usually needed by the different MVA methods, becomes of size $\dim(\mathcal{H}) \times \dim(\mathcal{H})$ and cannot be directly computed. However, making use of the *representer's theorem* [7], we can introduce $\mathbf{U} = \tilde{\boldsymbol{\Phi}}^\top \mathbf{A}$ into the formulation, where $\mathbf{A} = [\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_{d_f}]$ and $\boldsymbol{\alpha}_i$ is an n-length column vector containing the coefficients for the ith projection vector, and the maximization problem can be reformulated in terms of the kernel matrix, which is defined by the dot product of mapped samples $K(\mathbf{x}_i, \mathbf{x}_j) = \langle \boldsymbol{\phi}(\mathbf{x}_i), \boldsymbol{\phi}(\mathbf{x}_j) \rangle$.

Note that in these kernel feature extraction methods, the projection matrix $\mathbf{U}$ in $\mathcal{H}$ might not be explicitly calculated, but the projections of the input data can be obtained implicitly via kernel functions. Therefore, the extracted features for a new input pattern $\mathbf{x}_*$ are given by:

$$\tilde{\boldsymbol{\phi}}'(\mathbf{x}_*) = \tilde{\boldsymbol{\phi}}(\mathbf{x}_*)\mathbf{U} = \tilde{\boldsymbol{\phi}}(\mathbf{x}_*)\tilde{\boldsymbol{\Phi}}^\top \mathbf{A} = \begin{bmatrix} \tilde{\mathbf{K}}(\mathbf{x}_1, \mathbf{x}_*) \\ \vdots \\ \tilde{\mathbf{K}}(\mathbf{x}_n, \mathbf{x}_*) \end{bmatrix} \mathbf{A}, \tag{3}$$

and one can compute the inner products in the feature space $\tilde{K}(\mathbf{x}_i, \mathbf{x}_*) = \tilde{\boldsymbol{\phi}}(\mathbf{x}_i)\tilde{\boldsymbol{\Phi}}^\top$ that contains the inner products between the test point $\mathbf{x}_*$ and all training points $\{\mathbf{x}_i\}_{i=1}^n$ in the feature space, $\tilde{K}(\mathbf{x}_i, \mathbf{x}_*) = \langle \tilde{\boldsymbol{\phi}}(\mathbf{x}_i), \tilde{\boldsymbol{\phi}}(\mathbf{x}_*) \rangle$.

2) Kernel principal component analysis (KPCA): The goal of KPCA is to find the projections that maximize the variance of the input data in the feature space. By simply replacing $\tilde{\mathbf{X}}$ by $\tilde{\boldsymbol{\Phi}}$ in (1), KPCA can be formulated in the following way:

$$\begin{aligned} \text{KPCA:} \qquad & \mathbf{U} = \arg\max_{\mathbf{U}} \ \ \text{Tr}\{\mathbf{U}^\top \tilde{\boldsymbol{\Phi}}^\top \tilde{\boldsymbol{\Phi}} \mathbf{U}\} \\ & \text{subject to:} \ \ \mathbf{U}^\top \mathbf{U} = \mathbf{I}, \end{aligned} \tag{4}$$

where matrix $\tilde{\boldsymbol{\Phi}}^\top$ contains the mapped data centered in the Hilbert space. Making use of the representer's theorem one can introduce $\mathbf{U} = \tilde{\boldsymbol{\Phi}}^\top \mathbf{A}$ into the previous formulation, and the maximization problem can be reformulated as follows:

$$\begin{aligned} \text{KPCA:} \qquad & \mathbf{A} = \arg\max_{\mathbf{A}} \ \ \text{Tr}\{\mathbf{A}^\top \tilde{\mathbf{K}}_x \tilde{\mathbf{K}}_x \mathbf{A}\} \\ & \text{subject to:} \ \ \mathbf{A}^\top \tilde{\mathbf{K}}_x \mathbf{A} = \mathbf{I} \end{aligned} \tag{5}$$

The solution to the above problem can be obtained from the eigendecomposition of $\tilde{\mathbf{K}}_x \tilde{\mathbf{K}}_x$ represented by $\tilde{\mathbf{K}}_x \tilde{\mathbf{K}}_x \boldsymbol{\alpha} = \lambda \tilde{\mathbf{K}}_x \boldsymbol{\alpha}$, which has the same solution as $\tilde{\mathbf{K}}_x \boldsymbol{\alpha} = \lambda \boldsymbol{\alpha}$.

Note that centering in feature space can be done implicitly via the simple kernel matrix operation $\mathbf{K} \leftarrow \mathbf{HKH}$, where $H_{ij} = \delta_{ij} - \frac{1}{n}$, δ represents the Kronecker delta $\delta_{i,j} = 1$ if $i = j$ and zero otherwise. Recently, the centering operation has been questioned in the field of kernel methods and the community has witnessed, for instance, versions of KPCA with uncentered data in kernel feature space. In this paper, however, we stick to the standard kernelization of the methods and do consider centering.

3) Kernel Partial least squares (KPLS): KPLS is the nonlinear kernel-based extension of PLS [10], which is based on maximizing the variance between the projected data into a proper Hilbert space $\mathcal{H}$ and the target data matrix $\tilde{\mathbf{Y}}$ (i.e. the labels):

$$\begin{aligned} \text{KPLS:} \qquad & \mathbf{U}, \mathbf{V} = \arg\max_{\mathbf{U},\mathbf{V}} \ \ \text{Tr}\{(\tilde{\boldsymbol{\Phi}}\mathbf{U})^\top \tilde{\mathbf{Y}} \mathbf{V}\} \\ & \text{subject to:} \ \ \mathbf{U}^\top \mathbf{U} = \mathbf{V}^\top \mathbf{V} = \mathbf{I}, \end{aligned} \tag{6}$$

TABLE I
PROPERTIES OF LINEAR AND NONLINEAR METHODS USED IN THIS PAPER.

	PCA	PLS	KPCA	(SS-)KPLS
Max. Problem	$\mathbf{u}^\top \mathbf{C}_{xx}\mathbf{u}$	$\mathbf{u}^\top \mathbf{C}_{xy}\mathbf{v}$	$\boldsymbol{\alpha}^\top \tilde{\mathbf{K}}_x \tilde{\mathbf{K}}_x \boldsymbol{\alpha}$	$\boldsymbol{\alpha}^\top \tilde{\mathbf{K}}_x \tilde{\mathbf{Y}}\mathbf{v}$
Constraints	$\mathbf{u}^\top \mathbf{u} = 1$	$\mathbf{u}^\top \mathbf{u} = 1$ $\mathbf{v}^\top \mathbf{v} = 1$	$\boldsymbol{\alpha}^\top \tilde{\mathbf{K}}_x \boldsymbol{\alpha} = 1$	$\boldsymbol{\alpha}^\top \tilde{\mathbf{K}}_x \boldsymbol{\alpha} = 1$ $\mathbf{v}^\top \mathbf{v} = 1$
Max. d_f	$rank(\tilde{\mathbf{X}})$	$rank(\mathbf{C}_{xy})$	$rank(\tilde{\mathbf{K}}_x)$	$rank(\tilde{\mathbf{K}}_x \tilde{\mathbf{Y}})$

By using again the representer's theorem, the maximization problem becomes:

$$\text{KPLS (2):} \quad \mathbf{A}, \mathbf{V} = \arg\max_{\mathbf{A},\mathbf{V}} \ \text{Tr}\{\mathbf{A}^\top \tilde{\mathbf{K}}_x \tilde{\mathbf{Y}} \mathbf{V}\} \quad \text{subject to:} \ \mathbf{A}^\top \tilde{\mathbf{K}}_x \mathbf{A} = \mathbf{V}^\top \mathbf{V} = \mathbf{I}, \tag{7}$$

The solution to this problem can be obtained from the singular value decomposition of $\tilde{\mathbf{K}}_x \tilde{\mathbf{Y}}$. Alternatively, the problem can be efficiently solved using the following two-steps iterative procedure (see [7, Sec. 6.7] for more details):

1) Find the largest singular value of $\tilde{\mathbf{K}}_x \tilde{\mathbf{Y}}$, and the associated vector directions: $\{\boldsymbol{\alpha}_i, \mathbf{v}_i\}$.
2) Deflate the kernel matrix and labeled vector using:

$$\tilde{\mathbf{K}}_x \leftarrow \left[\mathbf{I} - \frac{\tilde{\mathbf{K}}_x \boldsymbol{\alpha}_i \boldsymbol{\alpha}_i^\top \tilde{\mathbf{K}}_x}{\boldsymbol{\alpha}_i^\top \tilde{\mathbf{K}}_x \tilde{\mathbf{K}}_x \boldsymbol{\alpha}_i}\right] \tilde{\mathbf{K}}_x \left[\mathbf{I} - \frac{\tilde{\mathbf{K}}_x \boldsymbol{\alpha}_i \boldsymbol{\alpha}_i^\top \tilde{\mathbf{K}}_x}{\boldsymbol{\alpha}_i^\top \tilde{\mathbf{K}}_x \tilde{\mathbf{K}}_x \boldsymbol{\alpha}_i}\right] \tag{8}$$

$$\mathbf{Y} = \mathbf{Y} - \tilde{\mathbf{K}}_x \boldsymbol{\alpha}_i \mathbf{Y} \frac{\tilde{\mathbf{K}}_x \boldsymbol{\alpha}_i}{\|\tilde{\mathbf{K}}_x \boldsymbol{\alpha}\|_2^2} \tag{9}$$

This deflation procedure allows us to extract more features than classes. For a more detailed description as well as implementation details, the reader is referred to [7], [8].

Generally speaking, MVA methods look for projections of the input data that are "maximally aligned" with the targets, and the different methods are characterized by the particular objectives they maximize. Table I compares some of the most important properties of the methods described in this section. An interesting property of linear methods is that they are based on first and second order moments, and that their solutions can be formulated in terms of (generalized) eigenvalue problems. Thus, standard linear algebra methods can be readily applied. This property is shared by kernel methods as well. The table shows the problem to be solved, the constraints involved, and the maximum number of features that each method can extract. Note that the proposed semi-supervised KPLS method (introduced in the next section) has exactly the same characteristics as the standard KPLS.

III. PROPOSED SEMI-SUPERVISED KPLS

This section presents the proposed feature extraction method. The underlying idea of the proposed Semi-supervised KPLS (SS-KPLS) is to construct a kernel function $K(\mathbf{x}_i, \mathbf{x}_j)$ measuring the similarity among labeled samples, taking into account the distribution of all available pixels, i.e. labeled ℓ and unlabeled u. The constructed kernel has two contributions, one using all available $\ell + u$ samples and the other computed with the ℓ labeled samples. The summation of the kernels is a valid kernel, and can be used in any kernel method for classification or regression, such as the standard support vector machine (SVM). Nevertheless, in this paper we plug this kernel into KPLS to extract a desired number of nonlinear features, which are then used for linear classification and regression. The method is simple to apply and relies on our recent developments which are summarized in the next subsections.

A. Bagged Kernel Support Vector Machine

In [20] we exploited the general idea of developing a kernel directly learned from data. The *bagged kernel* [17] was defined by counting the occurrences of two pixels in the same cluster over several runs of an unsupervised algorithm. The algorithm consists of different steps. First compute the standard RBF kernel K_s. Second run t times the k-means algorithm [21] with different initializations but with the same number of clusters k, which results in $p = 1, \ldots, t$ cluster assignments $c_p(\mathbf{x}_i)$ for each sample x_i. Third, we build a bagged kernel K_{bag} based upon the fraction of number of times that $\mathbf{x}_i$ and $\mathbf{x}_j$ are assigned to the same cluster:

$$K_{bag}(\mathbf{x}_i, \mathbf{x}_j) = \frac{1}{t} \sum_{p=1}^{t} [c_p(\mathbf{x}_i) = c_p(\mathbf{x}_j)] \tag{10}$$

where $i, j = 1, \ldots, (\ell + u)$ and operator $[c_p(\mathbf{x}_i) = c_p(\mathbf{x}_j)]$ returns '1' if samples $\mathbf{x}_i$ and $\mathbf{x}_j$ belong to the same cluster according to the pth realization of the clustering, $c_p(\cdot)$, and '0' otherwise. Finally, train a SVM with the sum (or the product) between the standard and the bagged kernels [17]:

$$K(\mathbf{x}_i, \mathbf{x}_j) = \beta K_s(\mathbf{x}_i, \mathbf{x}_j) + (1 - \beta) K_{bag}(\mathbf{x}_i, \mathbf{x}_j), \tag{11}$$

where $i, j = 1, \ldots, \ell$ and the weighting parameter $\beta \in [0, 1]$ provides a trade-off between the supervised and the unsupervised information.

B. Multiscale bagged kernel Support Vector Machine

The previous kernel implements the *cluster assumption* in the sense that samples that repeatedly fall in the same cluster should belong to the same class. However, this quite intuitive idea should hold *independently* of the scale of the relations we look at. Noting that the notion of similarity can be particularly distinctive at different scales, we developed in [22] a *multiscale bagged kernel* for urban very high resolution (VHR) images. The kernel of Eq. (10) was replaced by a kernel using m clusters of t runs of the standard k-means with different values of k (scales). This new averaged kernel accounts for similarities at different scales across the manifold between the pixels. The final kernel is the averaging of the q single-k bagged kernels and encodes multiscale (MS) similarities:

$$K_{bag}^{MS}(\mathbf{x}_i, \mathbf{x}_j) = \frac{1}{q} \sum_{m=1}^{q} K_{bag}^{m}(\mathbf{x}_i, \mathbf{x}_j). \tag{12}$$

This kernel was then linearly combined to the standard supervised kernel K_s [17], as in [20]:

$$K_C^{MS}(\mathbf{x}_i, \mathbf{x}_j) \leftarrow \beta K_s(\mathbf{x}_i, \mathbf{x}_j) + (1-\beta) K_{bag}^{MS}(\mathbf{x}_i, \mathbf{x}_j). \quad (13)$$

C. Proposed Cluster Kernel Feature Extraction

The two previous developments share in common the use of the k-means to decide whether two pixels fall into the same cluster. The k-means clustering algorithm is fast and efficient [21]. Nevertheless, it leads to generating too blocky and sparse kernels as it only gives us hard-decisions. A nice possibility of the framework of bag/cluster kernels is that any kind of clustering algorithm can be used.

In this paper, we propose two modifications of the previous algorithms:

1) The Expectation-Maximization clustering assuming a Gaussian mixture model (GMM) [23] replaces the k-means clustering. The EM-GMM is a probabilistic model to group the data in different subgroups focused on mixture Gaussian densities. Using the general Bayes' rule, it is possible to obtain the posterior probabilities, $\pi_{i,q}$, of the sample $\mathbf{x}_i$ belonging to cluster q as:

$$\pi_{i,q} = \frac{p(\mathbf{x}_i|q)p(q)}{p(\mathbf{x}_i)}, \quad (14)$$

where $p(q)$ is the prior probability and $p(\mathbf{x}_i|q)$ is the conditional probability of sample $\mathbf{x}_i$ given the cluster q. In the case of GMM, $p(\mathbf{x}_i|q)$ is a linear combination of Gaussian probability functions. The mixture parameters are estimated by the classical expectation-maximization method, and the maximum posterior probability is computed. The GMM clustering is almost as fast as k-means, but it also provides posterior membership probabilities. By using these probabilities instead of the hard memberships in k-means, smoother kernels are obtained. Including GMM in the construction of cluster kernels leads to the interesting notion of *probabilistic kernel functions* that account for the local structure of the data manifold.

2) We replace the standard SVM with the KPLS plus linear classification or regression. This has several benefits: i) KPLS allows us to extract nonlinear features maximally aligned with the target variables, ii) KPLS allows us to control the number of features easily, which has a direct effect on the compactness of the solution, and iii) in turn KPLS allows us to describe the data complexity indirectly with the number of needed features to achieve a given level of classification or regression error.

With these two modifications in mind, the proposed cluster kernel will consist of the combination of a kernel on labeled data and a kernel computed from clustering unlabeled data with GMM. The *multiscale probabilistic cluster kernel* is obtained as follows:

1) Compute the kernel function using labeled samples:

$$K_s(\mathbf{x}_i, \mathbf{x}_j) = \langle \boldsymbol{\phi}_s(\mathbf{x}_i), \boldsymbol{\phi}_s(\mathbf{x}_j) \rangle \; i, j = 1, \ldots, \ell \quad (15)$$

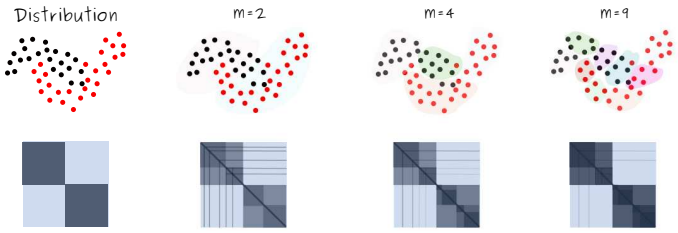

Fig. 1. Illustration of the construction of the probabilistic cluster kernel. The method clusters data with EM-GMM clustering for $m = \{2, 4, 9\}$, the posterior probability vectors are used to compute the dot products leading to the cluster kernel explicitly, and after repeating the process for a number of clusters, it accumulates the similarities in a multi-scale way. Samples with similar probabilities of membership to a grouped should belong to the same class. The multiscale cluster kernel (right kernel) is a better estimation of the optimal ideal kernel $\mathbf{K}_{ideal} = \mathbf{Y}\mathbf{Y}^\top$ (left kernel).

2) Run t times the GMM clustering algorithm with different initializations and with different number of clusters. This results in $q \cdot t$ cluster assignments where each sample $\mathbf{x}_i$ has its corresponding posterior probability vector $\boldsymbol{\pi}_i \in \mathbb{R}^m$ being m the number of clusters.
3) Build a *cluster* kernel K_c based upon the fraction of times that $\mathbf{x}_i$ and $\mathbf{x}_j$ are assigned to the same cluster:

$$K_c(\mathbf{x}_i, \mathbf{x}_j) = \frac{1}{Z} \sum_{p=1}^{t} \sum_{m=2}^{q} (\boldsymbol{\pi}_i^{m\top} \boldsymbol{\pi}_j^m)_p, \quad (16)$$

where $i, j = 1, \ldots, (\ell + u)$ and Z is the maximum value of K_c. An illustrative toy example of the multiscale cluster kernel construction is shown in Fig. 1.
4) Define the final kernel function K as the weighted sum [24] of the standard and the cluster kernels:

$$K(\mathbf{x}_i, \mathbf{x}_j) = \beta K_s(\mathbf{x}_i, \mathbf{x}_j) + (1-\beta) K_c(\mathbf{x}_i, \mathbf{x}_j), \quad (17)$$

where $i, j = 1, \ldots, \ell$ and $\beta \in [0, 1]$ is a scalar parameter.
5) Plug K into the standard KPLS solver (see Section II). KPLS returns the requested number of features d_f, which are used to project data onto them. These (nonlinear) projected data (scores) are then used as inputs to a *linear* classifier or regression method. The application of a linear model to the projected data is not incidental: note that all the features are extracted with a nonlinear method so this is the proper scheme to evaluate the effectiveness of the extracted variables.

The new cluster kernel accounts for *probabilistic* similarities at small and large scales between all labeled samples along the data manifold. Note that finding a proper kernel is equivalent to learn metric relations in the manifold which are defined here through a generative model learned from the data. The proposed kernel generalizes previous approaches based on multiscale cluster kernels. For example, the kernel in eq. (16) reduces to the approach in [22] when only the cluster assignment with maximum posterior probability is considered (hard or crisp clustering). The GMM cluster kernel can be related to the family of Fisher's kernels [25], [26]. Nonetheless, the proposed kernel has the very important advantage that it does

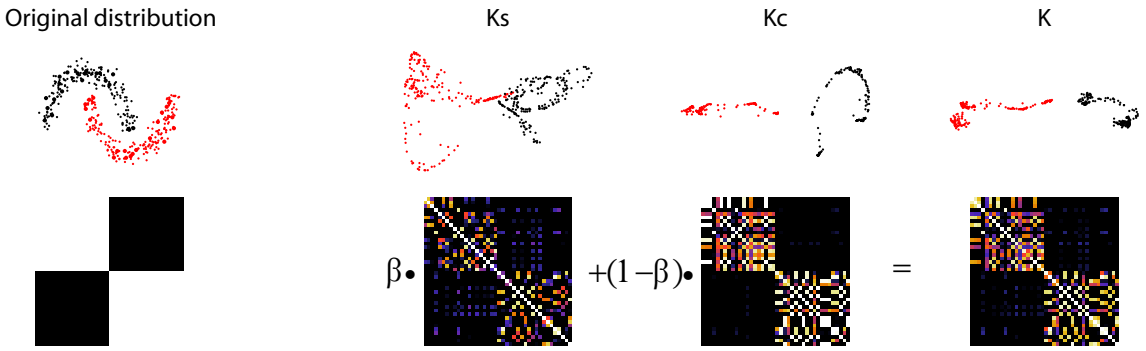

Fig. 2. Illustration of the projections (top) obtained with the KPLS method using different kernels (bottom). The original data and the ideal kernel (left). Ks corresponds to the RBF kernel constructed with labeled samples, Kc is the probabilistic cluster kernel constructed with both labeled and unlabeled samples, and K is the final kernel constructed by a linear combination of the previous kernels (right).

not assume an *ad hoc* parametric form or sophisticated priors and thus is more flexible and general. In addition the method does not require computationally demanding procedures, such as in dynamic programming optimization. Actually, current implementations for GMM scale linearly with the number of examples and data dimensions. All these properties are quite appealing from the practitioner point of view.

Theoretically, it can be shown that the cluster kernel fulfils the existence and uniqueness in a Hilbert space. In addition, it is trivial to show that the cluster kernel performs a linear kernel in a posterior probability space generated by the EM-GMM, it is a positive definite kernel, and the feature map corresponding to the cluster kernel K_c is unique [7].

A toy example of the projections obtained with KPLS method with the three kernels involved is shown in Fig. 2 for a two-dimensional binary classification problem. One could think that the probabilistic cluster kernel alone constitute a good enough metric to find better projections. However, this issue strongly depends on the number of both labeled and unlabeled samples. Figure 3[left] shows the results in this toy example for a fixed number of labeled samples and varying number of unlabeled samples, u: as u is increased the optimal β becomes lower, and hence the cluster kernel becomes relatively more important. Furthermore, we show in Fig. 3[right] the surface of optimal β values for different numbers of labeled and unlabeled samples. It is worth noting that the RBF kernel dominates the linear combination (high β values) when few data (less than 100 labeled and less than 200 unlabeled samples) are available, while for many data available, the probabilistic kernel becomes more important (low β values). This is due to the fact that the cluster kernel is not able to capture good information of the manifold data using low number of samples (labeled and unlabeled) since the clusters obtained by GMM are not representative of the data manifold.

Figure 4 illustrates the features extracted by linear and kernel feature extraction methods in the same nonlinear toy classification problem in a two-dimensional space. Linear methods fail in finding good projections since they cannot cope with the nonlinear nature of the data distribution. Kernel

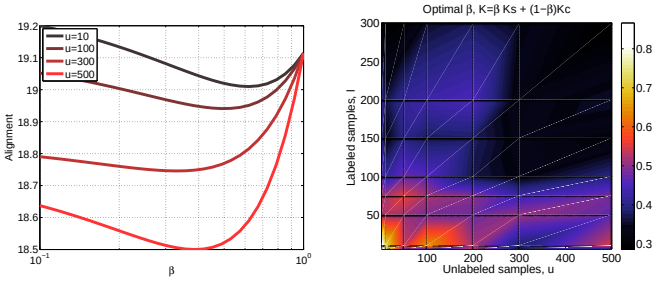

Fig. 3. Left: The alignment obtained with several values of β (weight of linear combination kernels) for a fixed number of labeled samples $\ell = 100$ and different unlabeled samples, $u = \{10, 100, 300, 500\}$. Right: Surface of optimal values of β for different number of labeled and unlabeled samples.

methods find nonlinear projections that better separate the data. The solution of KPCA does not allow to linearly separate the data. This is due to the fact that it becomes very difficult to tune the kernel parameter without labeled data, as previously studied in [27]. Such problem should be alleviated with KPLS but tuning the parameter is hampered by the low number of labeled data. The proposed cluster kernel K_c included in the KPLS method projects the original data such that they become linearly separable. The combination of the supervised and unsupervised kernels in KPLS refines the decision boundaries.

IV. Experimental results

This section presents the results obtained with the proposed SS-KPLS applied to remote sensing image classification and biophysical parameter retrieval problems. For the classification setting, we show results in three multispectral and hyperspectral images acquired by different sensors and involving the identification of different number of land cover classes. For the biophysical parameter retrieval, we consider two particularly relevant problems for land and ocean monitoring: the estimation of oceanic chlorophyll concentration, and of chlorophyll, LAI and fPAR for vegetation monitoring. The method is compared against standard linear and nonlinear feature extraction approaches in terms of accuracy and robustness, and expressive power (compactness of the information). Matlab code and demos are available for the interested reader

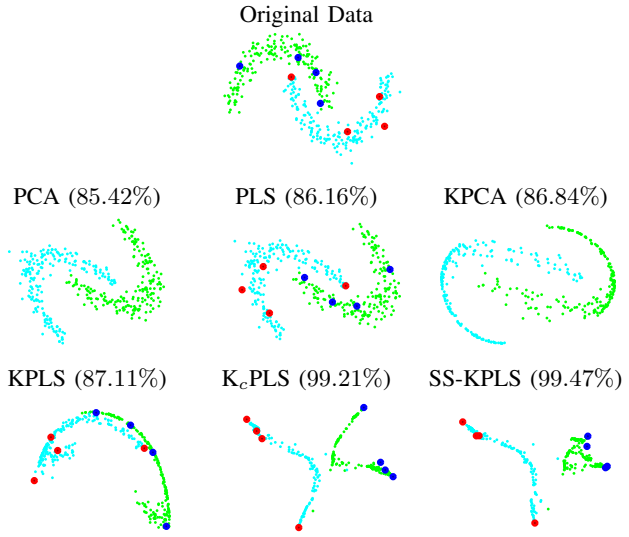

Fig. 4. Projections extracted by different linear and nonlinear feature extraction methods in a binary problem. We indicate the overall accuracy in the test set for comparison. Note that the SS-KPLS method reduces to KPLS for $\beta = 1$ and K$_c$PLS method for $\beta = 0$.

in http://isp.uv.es.

A. Experimental setup

For all experiments, we used ℓ labeled samples and u unlabeled samples in order to define the $(q \cdot t)$ cluster centers and the pixel posterior probabilities for each of the examples $\mathbf{x}_i$, i.e. $\boldsymbol{\pi}_i$. In all cases, we used $t = q = 20$ and the parameters β and σ were optimized by N-fold cross-validation. Given the low number of examples, a common prescription in machine learning is to use a low number of folds; in our case we optimized β and σ with $N = 3$ folds. The parameter β was tuned between $(0, 1)$ in steps of 0.05 and σ was varied between $[0.05, 2] \times s$ (s here represents the mean distance between all labeled data) for each number of extracted features. Once the mixture models are obtained and stored, the posterior probabilities or membership of the samples to each cluster are computed and K_c is constructed following (16). The same assignment is used for predicting the output (class membership for classification or estimated output variable for regression) of an unknown test pixel.

Obtaining projections in feature space for new test data $\mathbf{X}_*$ involves two operations. First, one has to map the data to the feature space, thus yielding $\tilde{\boldsymbol{\Phi}}_*$. Second, one has to project these mapped data onto the projections $\mathbf{U}$, which are expressed as $\mathbf{U} = \tilde{\boldsymbol{\Phi}}^\top \mathbf{A}$. Therefore the projected test data reduces to:

$$\mathcal{P}(\tilde{\boldsymbol{\Phi}}_*) = \tilde{\boldsymbol{\Phi}}_* \mathbf{U} = \tilde{\boldsymbol{\Phi}}_* \tilde{\boldsymbol{\Phi}}^\top \mathbf{A} = \tilde{\mathbf{K}}(\mathbf{X}_*, \mathbf{X})\mathbf{A},$$

where $\mathbf{X}$ is the training data matrix, and $\mathbf{A}$ contains in columns the d_f extracted feature vectors with the particular kernel method. The projected test data $\mathcal{P}(\tilde{\boldsymbol{\Phi}}_*)$ is a finite dimensional matrix of size $n_{test} \times d_f$. We used this projected data (*scores* in the statistics literature) in a simple linear regression model, $\hat{\mathbf{Y}} = \mathcal{P}(\tilde{\boldsymbol{\Phi}}_*)\mathbf{W}$. The weight vectors are obtained through the normal equations, $\mathbf{W} = \mathcal{P}(\tilde{\boldsymbol{\Phi}}_*)^\dagger \tilde{\mathbf{Y}}$, where $\dagger$ is the Moore-Penrose pseudoinverse. This solution is valid for multioutput regression problems. For the particular case of classification, the linear model is followed by a "winner-takes-all" activation function. We used different quality measures to test model's accuracy. In all cases, the quality measures were computed over a total of u unlabeled samples for each number of extracted features. For classification, we used the overall accuracy OA[%] and the estimated Cohen's kappa statistic κ. For regression problems, we evaluated the accuracy of the estimations through the root-means-square error (RMSE) and the Mean Absolute Error (MAE); the bias through the mean error (ME); and the goodness-of-fit through the Pearson's correlation coefficient, R.

B. Semi-supervised Feature Extraction for Classification

This subsection presents the results obtained by applying the proposed SS-KPLS technique to remote sensing multispectral and hyperspectral image classification. The next subsection details the data used in the experiments. Then, we focus our attention on the accuracy and robustness of the proposed algorithm in terms of the number of extracted features. Finally, we analyze the eigenspectrum, structure, and information content of the derived kernels.

1) Data: The first image dataset consists of 4 spectral bands acquired on a residential neighborhood of the city of Zürich by the QuickBird satellite in 2002. The portion of the image analyzed has size (329×347) pixels. The original image has been pansharpened using a Bayesian data fusion method (for more information see [28]) to attain a spatial resolution of 0.6 m. Nine classes of interest have been defined by photointerpretation. According to the good results obtained in previous studies [29], a total of 18 spatial features extracted using morphological opening and closing [30] have been added to the spectral bands, resulting in a final 22-dimensional vector.

The second image was acquired by the DAIS7915 sensor over the city of Pavia (Italy), and constitutes a challenging 9-class urban classification problem dominated by structural features and relatively high spatial resolution (5-meter pixels). Following previous works on classification of this image, we took into account only 40 spectral bands in the range [0.5, 1.76] μm, and thus skipped thermal and middle infrared bands above 1958 nm.

The third image is an AVIRIS hyperspectral image acquired over Salinas valley, an agricultural area of California (USA). A total of 16 crop classes were labeled and 224 spectral bands were used. This is a high-resolution scene with pixels of 3.7 meters. The high number of spectrally similar subclasses makes the classification problem very complex.

2) Results and discussion: For all experiments, ℓ and u are samples *per* class being $\ell = 10$ and $u = 190$ for all images. In order to avoid biased results, a total number of 10 realizations is carried out, and the averaged results are shown. We also provide the classification maps and the accuracies obtained in the whole scenes with the optimal parameters and fixing the number of extracted features.

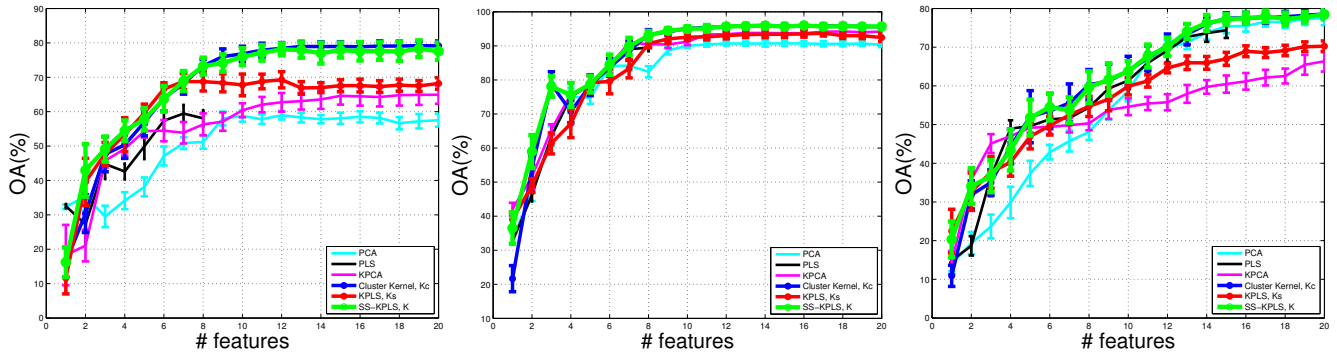

Fig. 5. Comparison between different feature extraction methods (linear and non-linear) using the overall accuracy versus the number of extracted features for the Zürich image (left), Pavia image (center), Salinas image (left).

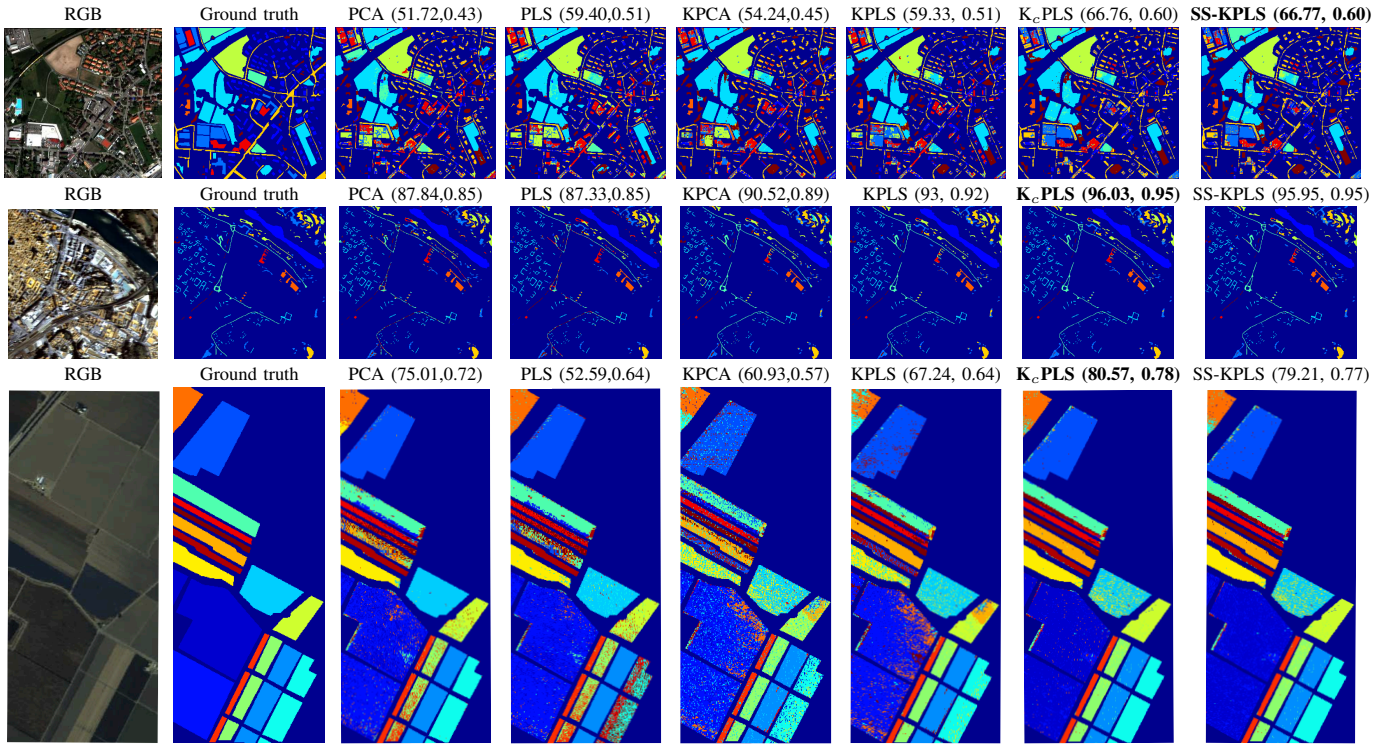

Fig. 6. Left to right: RGB composite, ground truth and three classification maps along with the overall accuracy and kappa for the Zürich image (top) for 11 extracted features, Pavia image (middle) for 16 extracted features and Salinas image (bottom) for 20 extracted features.

We evaluated the accuracy of several methods for a varying number of extracted features: 1) unsupervised linear, PCA, and its nonlinear version, KPCA; 2) supervised feature extraction algorithms (PLS and its nonlinear version KPLS); and 3) the different kernels involved in SS-KPLS. Note that the proposed SS-KPLS generalizes the standard KPLS (when $\beta = 1$).

Mean and standard deviation accuracies are shown in Fig. 5. In general, nonlinear kernel methods (KPCA, KPLS and variants) outperform linear approaches (PCA and PLS). The proposed SS-KPLS improves the results of the standard KPLS and the cluster kernel. The generative cluster kernel proposed here yields higher accuracies than the RBF kernel when increasing the number of features. When a higher number of nonlinear features is extracted, all curves become stable but the proposed SS-KPLS clearly outperforms the standard PCA in a range between $+5$-15%, the more advanced KPCA in a range between $+4$-15% and KPLS in a range between $+3$-10%. The behaviour of PCA and KPCA in the Zürich and Salinas images to be analyzed because higher accuracy is not obtained with higher number of extracted features, revealing a kind of overfitting problem. This effect has been recently reported in the literature [27]. This is not the case of the proposed cluster kernel K_c. These results are confirmed by the visual inspection of the classification maps shown in Fig. 6, which confirm qualitatively the quantitative results in which the SS-KPLS shows a clear and consistent gain over KPLS of about $+7\%$ (Zürich), $+3\%$ (Pavia), $+13\%$ (Salinas).

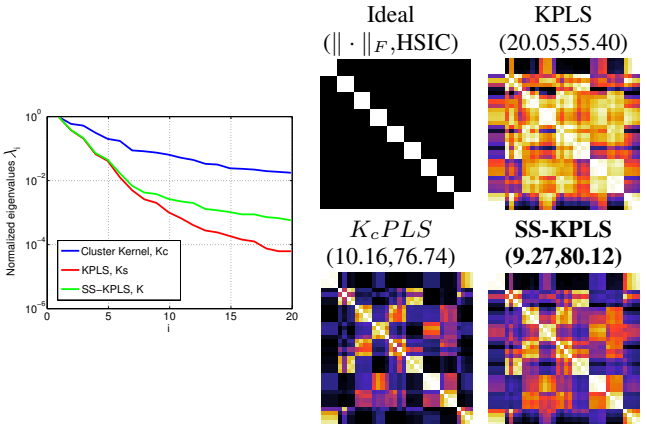

Fig. 7. Left: Normalized eigenvalues for all kernels used in the Pavia dataset. Right: ideal and used kernels, along quantitative measures of error $\|\cdot\|_F$ and dependence (HSIC).

3) Analysis of the kernels: Figure 7 shows the eigenvalues of the best kernels for the Pavia image. The eigendecomposition of the proposed semi-supervised kernel K shows a tradeoff between the RBF and the cluster kernel, as expected. It is worth noting that the eigenvalues of cluster kernel (blue line) show a slower decay because the kernel is indeed quite blocky and sparse. On the other hand, the RBF kernel shows a heavier tail. The introduction of the cluster kernel can be casted as an extra regularization of the RBF kernel. The right plots present the used kernels and their similarity to the ideal one, $\mathbf{K}_{\text{ideal}} = \mathbf{Y}\mathbf{Y}^\top$. Two quantitative measures are given: the Frobenius norm of the difference of these two kernels, $\|\cdot\|_F$, and the Hilbert-Schmidt Independence Criterion (HSIC) between them [31]. The proposed kernel K aligns well with the ideal kernel (lower error, higher dependence), and takes advantage of the sharper structure learned by the Cluster Kernel.

C. Semi-supervised Feature Extraction for Regression

We focus now on two challenging problems of biophysical parameter estimation. In particular, we first tackle the estimation of oceanic chlorophyll concentration from multispectral MERIS measurements, and second the retrieval of land-cover biophysical parameters –leaf chlorophyll content (Chl), leaf area index (LAI), and fractional vegetation cover (fCover)– from CHRIS hyperspectral images. In both cases, satellite-derived data and *in situ* measurements are subjected to high levels of uncertainty, as well as collinearity between the input features (channels) and the output target variables. In these difficult scenarios, a proper (robust) feature extraction is necessary, particularly when their relationship is believed to be non-linear or the target data is scarce thus leading to badly conditioned problems.

1) Oceanic chlorophyll concentration: The first dataset simulates data acquired by the Medium Resolution Imaging Spectrometer (MERIS) on board the Envisat satellite (MERIS dataset), and in particular the spectral behavior of chlorophyll concentration in the subsurface waters. We selected the eight channels in the visible range (412-681 nm) to be used for retrieval. The range of variation of chlorophyll concentration in this dataset is $0.02 - 25mg/m^3$.

TABLE II
ESTIMATED RESULTS FOR THE OCEANIC CHLOROPHYLL CONCENTRATION RETRIEVAL PROBLEM AS A FUNCTION OF THE NUMBER OF EXTRACTED FEATURES.

Model	RMSE	MAE	|ME|	R
PCA ($d_f = 1$)	0.540	0.429	0.173	-0.252
PCA ($d_f = 2$)	0.415	0.329	0.073	0.589
PCA ($d_f = 3$)	0.397	0.318	0.050	0.629
PCA ($d_f = 4$)	0.315	0.232	**0.000**	0.822
PLS	0.540	0.429	0.175	-0.201
KPCA ($d_f = 1$)	0.531	0.425	0.163	0.000
KPCA ($d_f = 2$)	0.523	0.418	0.156	0.109
KPCA ($d_f = 3$)	0.507	0.406	0.171	0.385
KPCA ($d_f = 4$)	0.385	0.310	0.039	0.651
KPLS ($d_f = 1$)	0.451	0.354	0.010	0.462
KPLS ($d_f = 2$)	0.472	0.377	0.033	0.407
KPLS ($d_f = 3$)	0.453	0.355	0.009	0.453
KPLS ($d_f = 4$)	0.436	0.341	0.005	0.513
K$_c$PLS ($d_f = 1$)	0.339	**0.255**	**0.000**	**0.775**
K$_c$PLS ($d_f = 2$)	**0.295**	**0.206**	**0.026**	0.833
K$_c$PLS ($d_f = 3$)	**0.275**	**0.192**	**0.008**	**0.857**
K$_c$PLS ($d_f = 4$)	0.271	0.190	0.006	**0.863**
SS-KPLS ($d_f = 1$)	**0.341**	0.257	**0.000**	0.771
SS-KPLS ($d_f = 2$)	0.298	0.210	0.027	**0.829**
SS-KPLS ($d_f = 3$)	0.280	0.196	0.009	0.853
SS-KPLS ($d_f = 4$)	**0.265**	**0.189**	0.007	**0.864**

In this experiment, we evaluate different quantitative measures of accuracy, bias and goodness-of-fit for a varying number of extracted features. We compare the results obtained by 1) unsupervised linear PCA and its nonlinear kernel version, KPCA; 2) supervised feature extraction algorithms (PLS and its nonlinear version KPLS); and 3) the different kernels involved in SS-KPLS. Table II shows the obtained results with $\ell = 45$ labeled samples and $u = 955$ unlabeled samples to construct the cluster kernel K_c. In general, the nonlinear methods obtain better results than linear approaches. The proposed method reduces the prediction error around 35% with respect to linear PLS and PCA, and KPCA method. In addition, the proposed semi-supervised KPLS reduces the error about 25% for a given number of extracted features. Note that, the good results obtained with semi-supervised KPLS are mainly due to the cluster kernel function (β values are small) which in many cases yields very high accuracies working alone ($\beta = 0$).

This result suggests that K_c could be used as a parameter-free kernel learned from the data, thus constituting an alternative to standard kernel functions.

2) Biophysical parameter retrieval: For the second dataset, we considered data obtained in the SPectra bARrax Campaign (SPARC) in 2003 and 2004 in Barrax, Spain. The test area is an agricultural research facility with an extent of $5 \times 10km$. It is characterized by a flat landscape and large uniform land-use units of irrigated and dry lands. The vegetation biophysical parameters were measured among different crops where a large number of samples on an elementary sampling unit (ESU) were taken and averaged for different parameters,

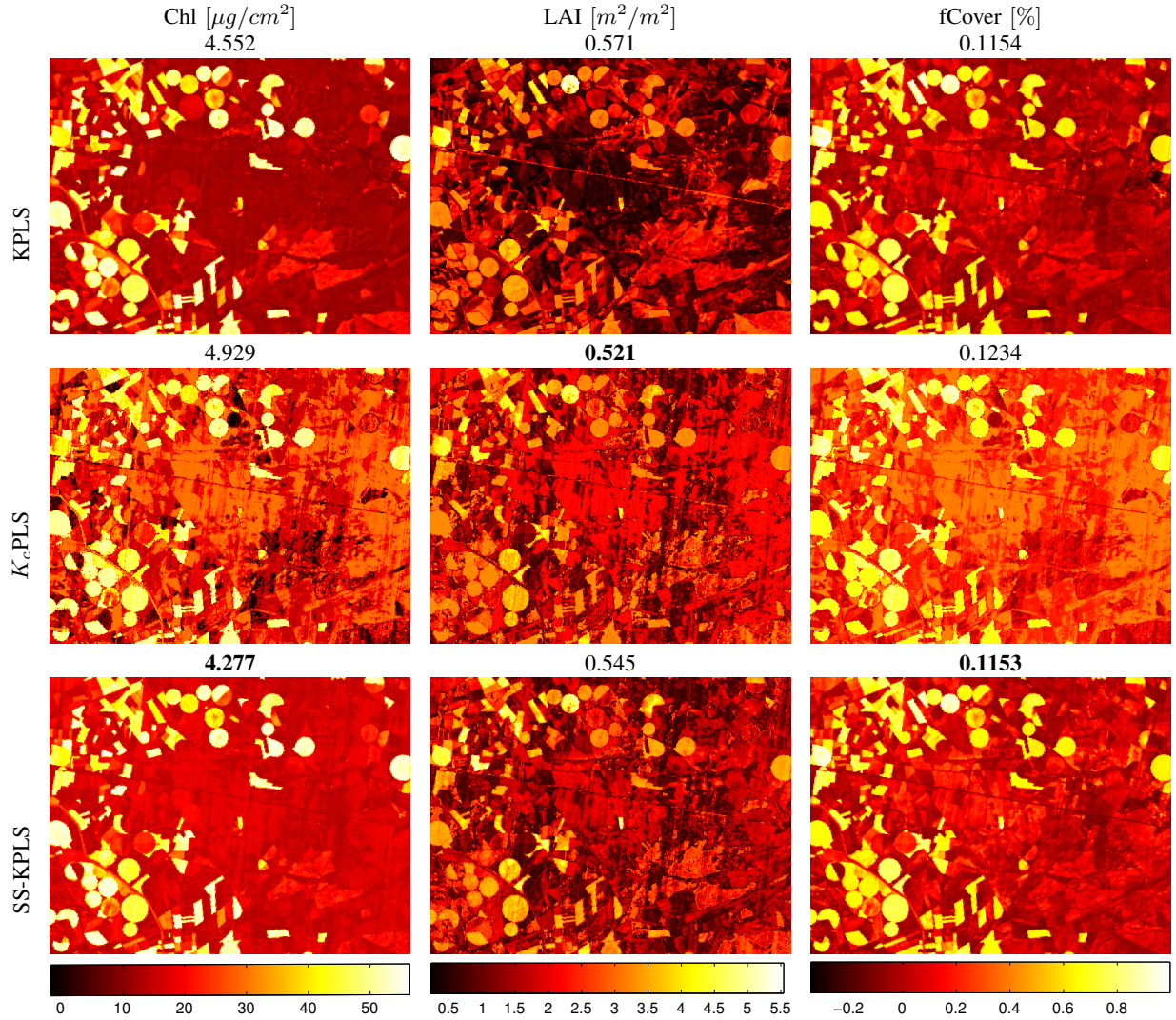

Fig. 8. Estimation maps for Chl, LAI and FCV, for the KPLS, K_cPLS and SS-KPLS feature extractor methods with the RMSE for the small area of CHRIS/PROBA image with 4 features.

obtaining a local characterization. The Chl was measured with a calibrated Minolta CCM-200 from 50 samples per ESU. The LAI was derived from canopy measurements made with a LiCor LAI-2000 at 24 locations per ESU. The fCover was derived from hemispherical photographs taken at the same locations as the LAI measurements. All parameters present standard errors between 3% and 10%. For both years, we have a total of nine crop types (garlic, alfalfa, onion, sunflower, corn, potato, sugar beet, vineyard, and wheat), with field-measured values of LAI that vary between 0.4 and 6.3, Chl between 2 and 55 $\mu g/cm^2$, and fCover between 0 and 1. This makes the dataset representative and well-suited to multioutput regression studies. Simultaneously to the ground sampling, hyperspectral images were collected by the CHRIS/PROBA spaceborne sensor. The data provided have 62 bands in the visible and near-infrared (NIR) region $(400 - 1000$ nm) at a spatial resolution of $34m$. The images selected for this experiment were those acquired from the nadir view sharing similar observation configuration in order to minimize angular and atmospheric effects. The images were geometrically and atmospherically corrected using the official CHRIS/PROBA Toolbox for BEAM [32]. Finally, the database consists of $\ell = 135$ Chl, LAI, and fCover measurements and their associated 62 CHRIS reflectance channels. We used all pixels in the image as unlabeled samples $u = 243648$ to construct the cluster kernel.

The obtained maps and RMSE for the three considered biophysical parameters are shown in Fig. 8. In the three cases, the use of the kernel combination reports slightly better results. Even if the gain is not very big with regard the standard KPLS approach (about $+2\%$), we should note that 1) the built K_c could be used directly for retrieval without the need of tuning kernel parameters; 2) the cluster kernel leads to higher RMSEs than KPLS for Chl and fCover but, since the solutions are complementary, the SS-KPLS benefits from the combination, and 3) the combination makes the final model more robust for LAI as well.

V. Conclusions

This paper proposed a novel nonlinear feature extraction technique for remote sensing image classification and retrieval of biophysical parameters. The method is specifically devised for addressing problems where the number of training samples available is relatively small. Note that these problems are common in operational applications of remote sensing. In such situations, the combination of labeled and unlabeled samples in a semi-supervised framework can significantly improve the representation of the data. The main limitation is that the number of unlabeled samples used to estimate the cluster structure via the EM-GMM algorithm should be high enough, which is usually the case in remote sensing applications.

Good results were obtained on both multispectral and hyperspectral data sets considered in our experiments, where the proposed method performs better than supervised and unsupervised linear and nonlinear approaches, both in classification and regression problems. In this paper we focused on the KPLS method but it is possible to apply and generalize to others supervised kernel feature extraction methods like the Fisher's discriminant family. Future work will consider the direct use of the generative cluster kernel in unsupervised image segmentation as it revealed consistent behaviour in problems with many data available, and the study of the metric space induced by the kernels.

Acknowledgements

The authors would like to acknowledge Prof. Paolo Gamba from the University of Pavia (Italy) for kindly providing both the ROSIS data, and Dr. Devis Tuia from the École Polytechnique Fédérale de Lausanne (EPFL), Switzerland, for kindly providing the Zürich/Brutisellen data.

References

[1] I. T. Jolliffe, *Principal Component Analysis*, Springer, 2nd edition, 2010.

[2] R. Rosipal and N. Krämer, "Overview and recent advances in partial least squares," in *Subspace, Latent Structure and Feature Selection*. 2006, vol. 3940 of *LNCS*, pp. 34–51, Springer.

[3] A.A. Green, M. Berman, P. Switzer, and M.D. Craig, "A transformation for ordering multispectral data in terms of image quality with implications for noise removal," *IEEE Trans. Geosc. Rem. Sens.*, vol. 26, no. 1, pp. 65 –74, jan. 1988.

[4] W.J. Blackwell, "A neural-network technique for the retrieval of atmospheric temperature and moisture profiles from high spectral resolution sounding data," *IEEE Trans. Geosc. Rem. Sens.*, vol. 43, no. 11, 2005.

[5] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, December 2000.

[6] M. A. Kramer, "Nonlinear principal component analysis using autoassociative neural networks," *AIChE Journal*, vol. 37, no. 2, pp. 233–243, 1991.

[7] J. Shawe-Taylor and N. Cristianini, *Kernel Methods for Pattern Analysis*, Cambridge University Press, 2004.

[8] G. Camps-Valls and L. Bruzzone, *Kernel Methods for Remote Sensing Data Analysis*, John Wiley and Sons, 2009.

[9] 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, March 2002.

[10] J. Arenas-García and G. Camps-Valls, "Efficient kernel orthonormalized PLS for remote sensing applications," *IEEE Trans. Geosc. Rem. Sens.*, vol. 46, pp. 2872 –2881, Oct 2008.

[11] Y. Gu, Y. Liu, and Y. Zhang, "A selective KPCA algorithm based on high-order statistics for anomaly detection in hyperspectral imagery," *IEEE Geosc. Rem. Sens. Letters*, vol. 5, no. 1, pp. 43 –47, Jan 2008.

[12] Bor-Chen Kuo, Cheng-Hsuan Li, and Jinn-Min Yang, "Kernel nonparametric weighted feature extraction for hyperspectral image classification," *IEEE Trans. Geosc. Rem. Sens.*, vol. 47, no. 4, pp. 1139 –1155, apr. 2009.

[13] A.A. Nielsen, "Kernel maximum autocorrelation factor and minimum noise fraction transformations," *IEEE Trans. Image Processing*, vol. 20, no. 3, pp. 612–624, Mar 2011.

[14] L. Gómez-Chova, A.A. Nielsen, and G. Camps-Valls, "Explicit signal to noise ratio in reproducing kernel Hilbert spaces," in *IEEE Geosc. Rem. Sens. Symp. (IGARSS)*. Jul 2011, pp. 3570–3570, IEEE.

[15] L. Gómez-Chova, R. Jenssen, and G. Camps-Valls, "Kernel Entropy Component Analysis for Remote Sensing Image Clustering," *IEEE Geosc. Rem. Sens. Letters*, vol. 9, no. 2, pp. 312–316, Mar 2012.

[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," *Signal Processing Magazine, IEEE*, vol. 30, no. 4, pp. 16–29, 2013.

[17] O. Chapelle, B. Schölkopf, and A. Zien, *Semi-Supervised Learning*, MIT Press, Cambridge, 1st edition, 2006.

[18] H. Wold, "Estimation of principal components and related models by iterative least squares.," *Multivariate Analysis*, pp. 391–420, 1966.

[19] P. D. Sampson, A. P. Streissguth, H. M. Barr, and F. L. Bookstein, "Neurobehavioral effects of prenatal alcohol: Partial Least Squares analysis," *Neurotoxicology and teratology*, vol. 11, pp. 477–491, 1989.

[20] D. Tuia and G. Camps-Valls, "Semisupervised remote sensing image classification with cluster kernels," *IEEE Geosc. Rem. Sens. Letters*, vol. 6, no. 2, pp. 224–228, Apr 2009.

[21] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis*, Wiley, New York, USA, 1973.

[22] D. Tuia and G. Camps-Valls, "Urban image classification with semisupervised multiscale cluster kernels," *IEEE JSTARS*, vol. 4, pp. 65–74, Mar 2011.

[23] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*, Springer-Verlag New York, Inc., 2006.

[24] G. Camps-Valls, L. Gómez-Chova, J. Muñoz-Marí, J. Vila-Francés, and J. Calpe-Maravilla, "Composite kernels for hyperspectral image classification," *IEEE Geosc. Rem. Sens. Letters*, vol. 3, no. 1, pp. 93–97, Jan 2006.

[25] T. Jaakkola and D. Haussler, "Exploiting generative models in discriminative classifiers," in *NIPS'11*, 1998, pp. 487–493.

[26] O. Chapelle, J. Weston, and B. Schölkopf, "Cluster Kernels for Semi-Supervised Learning," in *NIPS 2002*, Becker, Ed., Cambridge, MA, USA, 2003, vol. 15, pp. 585–592, MIT Press.

[27] Mikio L. Braun, Joachim Buhmann, and Klaus-Robert Müller, "On relevant dimensions in kernel feature spaces," *Journal of Machine Learning Research*, vol. 9, pp. 1875–1908, Aug 2008.

[28] D. Fasbender, J. Radoux, and P. Bogaert, "Bayesian data fusion for adaptable image pansharpening," *IEEE Trans. Geosc. Rem. Sens.*, vol. 46, no. 6, pp. 1847 –1857, june 2008.

[29] D. Tuia, F. Ratle, A. Pozdnoukhov, and G. Camps-Valls, "Multisource composite kernels for urban-image classification," *IEEE Geosc. Rem. Sens. Letters*, vol. 7, pp. 88–92, 2010.

[30] J. Serra, *Image Analysis and Mathematical Morphology, Volume 2: Theoretical Advances*, Academic press, 1988.

[31] G. Camps-Valls, J. Mooij, and B. Schölkopf, "Remote sensing feature selection by kernel dependence measures," *IEEE Geosc. Rem. Sens. Letters*, vol. 7, no. 3, pp. 587–591, Jul 2010.

[32] L. Alonso, L. Gómez-Chova, J. Moreno, L. Guanter, C. Brockmann, N. Fomferra, R. Quast, and P. Regner, "CHRIS/PROBA toolbox for hyperspectral and multiangular data exploitations," in *IEEE Geosc. Rem. Sens. Symp. (IGARSS)*, Jul 2009, vol. II, pp. 202–205.

Emma Izquierdo-Verdiguier (S'12) received the B.Sc. degree in Physics and the M.Sc. degree in Remote Sensing from the University of Valencia, Valencia, Spain, where she is currently working toward the Ph.D. degree with the Image Processing Laboratory. Her research interests are nonlinear feature extraction based on kernel methods. Previously she worked on automatic identification and classification of multispectral images. In 2012 she ranked second place at student paper competition of the Geoscience and Remote Sensing Symposioum (IGARSS).

Luis Gómez-Chova (S'08-M'09) received the B.Sc. (with first-class honors), M.Sc., and Ph.D. degrees in electronics engineering from the University of Valencia, Valencia, Spain, in 2000, 2002 and 2008, respectively. He was awarded by the Spanish Ministry of Education with the National Award for Electronics Engineering. Since 2000, he has been with the Department of Electronics Engineering, University of Valencia, first enjoying a research scholarship from the Spanish Ministry of Education and currently as an Associate Professor. He is also a researcher at the Image Processing Laboratory (IPL), where his work is mainly related to pattern recognition and machine learning applied to remote sensing multispectral images and cloud screening. He conducts and supervises research on these topics within the frameworks of several national and international projects. He is the author (or coauthor) of 30 international journal papers, more than 90 international conference papers, and several international book chapters. Visit http://www.uv.es/chovago for more information.

Lorenzo Bruzzone (S'95-M'98-SM'03-F'10) received the Laurea (M.S.) degree in electronic engineering (summa cum laude) and the Ph.D. degree in telecommunications from the University of Genoa, Italy, in 1993 and 1998, respectively. He is currently a Full Professor of telecommunications at the University of Trento, Italy, where he teaches remote sensing, radar, pattern recognition, and electrical communications. Dr. Bruzzone is the founder and the director of the Remote Sensing Laboratory in the Department of Information Engineering and Computer Science, University of Trento. His current research interests are in the areas of remote sensing, radar and SAR, signal processing, and pattern recognition. He promotes and supervises research on these topics within the frameworks of many national and international projects. Among the others, he is the Principal Investigator of the *Radar for icy Moon exploration (RIME)* instrument in the framework of the *JUpiter ICy moons Explorer (JUICE)* mission of the European Space Agency. He is the author (or coauthor) of 137 scientific publications in referred international journals (93 in IEEE journals), more than 190 papers in conference proceedings, and 16 book chapters. He is editor/co-editor of 11 books/ conference proceedings and 1 scientific book. His papers are highly cited, as proven form the total number of citations (more than 8500) and the value of the h-index (47) (source: Google Scholar). He was invited as keynote speaker in 24 international conferences and workshops. Since 2009 he is a member of the Administrative Committee of the IEEE Geoscience and Remote Sensing Society. Dr. Bruzzone ranked first place in the Student Prize Paper Competition of the 1998 IEEE International Geoscience and Remote Sensing Symposium (Seattle, July 1998). Since that time he was recipient of many international and national honors and awards. Dr. Bruzzone was a Guest Co-Editor of different Special Issues of international journals. He is the co-founder of the IEEE International Workshop on the Analysis of Multi-Temporal Remote-Sensing Images (MultiTemp) series and is currently a member of the Permanent Steering Committee of this series of workshops. Since 2003 he has been the Chair of the SPIE Conference on Image and Signal Processing for Remote Sensing. Since 2013 he has been the founder Editor-in-Chief of the IEEE Geoscience and Remote Sensing Magazine. Currently he is an Associate Editor for the IEEE Transactions on Geoscience and Remote Sensing and the Canadian Journal of Remote Sensing. Since 2012 he has been appointed *Distinguished Speaker* of the IEEE Geoscience and Remote Sensing Society.

Gustavo Camps-Valls (M'04, SM'07) received a Ph.D. degree in Physics (2002, *summa cum laude*) from the Universitat de València, Spain, where he is currently an Associate Professor in the Electrical Engineering Dep.He teaches time series analysis, image processing, machine learning, and knowledge extraction for remote sensing. His research is conducted as Group Leader of the Image and Signal Processing (ISP) group, http://isp.uv.es, of the same university. He has been Visiting Researcher at the Remote Sensing Laboratory (Univ. Trento, Italy) in 2002, the Max Planck Institute for Biological Cybernetics (Tübingen, Germany) in 2009, and as Invited Professor at the Laboratory of Geographic Information Systems of the École Polytechnique Fédérale de Lausanne (Lausanne, Switzerland) in 2013. His research interests are tied to the development of machine learning algorithms for signal and image processing with special focus on remote sensing data analysis.

He conducts and supervises research within the frameworks of several national and international projects, and he is Evaluator of project proposals and scientific organizations. He is the author (or co-author) of 105 international peer-reviewed journal papers, more than 130 international conference papers, 20 international book chapters, and editor of the books "Kernel methods in bioengineering, signal and image processing" (IGI, 2007), "Kernel methods for remote sensing data analysis" (Wiley & Sons, 2009), and "Remote Sensing Image Processing" (MC, 2011). He's a co-editor of the forthcoming book "Digital Signal Processing with Kernel Methods" (Wiley & sons, 2014). He holds a Hirsch's h index $h = 32$, entered the ISI list of Highly Cited Researchers in 2011, and he is a co-author of the 3 most highly cited papers in relevant remote sensing journals. Thomson Reuters ScienceWatch® identified one of his papers as a Fast Moving Front research.

He is a referee of many international journals and conferences, and currently serves on the Program Committees of International Society for Optical Engineers (SPIE) Europe, International Geoscience and Remote Sensing Symposium (IGARSS),Machine Learning for Signal Processing (MLSP), and International Conference on Image Processing (ICIP) among others. In 2007 he was elevated to IEEE Senior Member, and since 2007 he is member of the Data Fusion technical committee of the IEEE Geoscience and Remote Sensing Society, and since 2009 he is member of the Machine Learning for Signal Processing Technical Committee of the IEEE Signal Processing Society. He is member of the MTG-IRS Science Team (MIST) of the European Organisation for the Exploitation of Meteorological Satellites (EUMETSAT). He is Associate Editor of the "IEEE Transactions on Signal Processing", "IEEE Signal Processing Letters", "IEEE Geoscience and Remote Sensing Letters", and Guest Editor of "IEEE Journal of Selected Topics in Signal Processing" . Visit http://www.uv.es/gcamps for more information.