HAL
open science

Sparse Hilbert Schmidt Independence Criterion and Surrogate-Kernel-Based Feature Selection for Hyperspectral Image Classification

Bharath Bhushan Damodaran, Nicolas Courty, Sébastien Lefèvre

To cite this version:

Bharath Bhushan Damodaran, Nicolas Courty, Sébastien Lefèvre. Sparse Hilbert Schmidt Independence Criterion and Surrogate-Kernel-Based Feature Selection for Hyperspectral Image Classification. IEEE Transactions on Geoscience and Remote Sensing, 2017, 55 (4), pp.2385-2398. 10.1109/TGRS.2016.2642479 . hal-01447452v2

HAL Id: hal-01447452

https://hal.science/hal-01447452v2

Submitted on 27 Mar 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

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

Sparse Hilbert Schmidt Independence Criterion and Surrogate-Kernel-Based Feature Selection for Hyperspectral Image Classification

Bharath Bhushan Damodaran, *Member, IEEE*, Nicolas Courty, *Member, IEEE*, and Sébastien Lefèvre

***Abstract*— Designing an effective criterion to select a subset of features is a challenging problem for hyperspectral image classification. In this paper, we develop a feature selection method to select a subset of class discriminant features for hyperspectral image classification. First, we propose a new class separability measure based on the surrogate kernel and Hilbert Schmidt independence criterion in the reproducing kernel Hilbert space. Second, we employ the proposed class separability measure as an objective function and we model the feature selection problem as a continuous optimization problem using LASSO optimization framework. The combination of the class separability measure and the LASSO model allows selecting the subset of features that increases the class separability information and also avoids a computationally intensive subset search strategy. Experiments conducted with three hyperspectral data sets and different experimental settings show that our proposed method increases the classification accuracy and outperforms the state-of-the-art methods.**

***Index Terms*— Band selection, class separability measure, feature selection, Hilbert Schmidt independence criterion (HSIC), hyperspectral image classification, kernel methods, LASSO, surrogate kernel (SK).**

I. Introduction

HYPERSPECTRAL sensors characterize the information content of the objects under the scene in a large number of spectral bands. For this reason, it has been considered as an important data source for many remote sensing applications (e.g., precision agriculture, land cover/land use mapping, and mineral exploration) [1]–[3]. However, the automatic analysis of hyperspectral data is a complex task mainly due to high dimensionality of the hyperspectral data. This causes the critical issues such as curse of dimensionality, redundant information, and high volume of the data. Addressing the above issues is of prime importance to exploit the potentials of hyperspectral data for real-world applications. Statistical-based models such as the Gaussian mixture model suffer severely due to the curse of dimensionality, leading to an inaccurate estimation of parameters of the model. One effective way to tackle these problems is to reduce the dimensionality of hyperspectral data set without losing the useful information, and this process has the capability to overcome the feature space geometrical and statistical limitations of hyperspectral data [4]. Kernel-based methods such as support vector machine (SVM) have shown to be more effective in handling high dimensionality of hyperspectral data. However, these methods could also benefit from the dimensionality reduction methods [5].

In recent years, several attempts have been made to deal with the dimensionality of the hyperspectral data set using either feature extraction or feature selection methods. Feature extraction projects the high-dimensional image into a lower dimensional subspace [6], [7]. On the other hand, feature selection (or band selection) picks the best subset of bands based on a certain criterion [8]–[10]. The latter is often preferable in remote sensing, since the selected bands are interpretable and usually contain the physical information. Feature selection methods can be categorized into supervised methods [5], [13], [14] and unsupervised [11], [12]. In the unsupervised case, the bands are selected mainly based on feature clustering [10], [11], independence column selection [15], and feature ranking [16]. The performance of these methods might be low compared with that of the supervised case. However, these unsupervised methods are preferred when no ground-truth information is available. On the other hand, in the supervised case, the bands are selected based on wrapper- and filter-based approaches. The former relies on the selection of a subset of bands, which maximizes the classification accuracy in accordance with the underlying classifier [5], [9]. The latter rather selects the subset of bands with respect to the characteristics of the training data. Thus, the selected bands by the filter-based approaches are independent of the classifier, have better generalization power, and are less computationally expensive depending on the underlying filter used.

Designing the criteria for a filter-based supervised feature selection method is a challenging problem. Several criteria have been proposed for feature selection, including distance measures [17], class separability measures [8], [18], information measures, and dependence measures [14], [19]. Among these criteria, the class-separability-based strategies are interesting, since they select the subset of features that maximizes the class separable information. The Jeffries–Matusita distance

Manuscript received October 20, 2016; accepted December 16, 2016. This work was supported in part by the French Agence Nationale de la Recherche under Project ANR-13-JS02-0005-01 (Asterix project) and in part by the People Programme (Marie Curie Actions) of the European Union's Seventh Framework Programme (FP7/2007-2013) under REA Grant PCOFUND-GA-2013-609102, through the PRESTIGE programme coordinated by Campus France.

The authors are with the Université de Bretagne-Sud, UMR 6074, IRISA, 56000 Vannes, France (e-mail: bharath-bhushan.damodaran@irisa.fr; nicolas.courty@irisa.fr; sebastien.lefevre@irisa.fr).

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/TGRS.2016.2642479

0196-2892 © 2017 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

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

measure, Bhattacharya distance, Fisher's ratio measure are the most widely used class separability measures for remote sensing images [20]–[22]. These separability measures compute the class separability information between the land cover classes in the input space. Thus, the above measures might not be so effective in capturing the nonlinear dependence relationship between the land cover classes in remote sensing data while compared with working in the reproducing kernel Hilbert space (RKHS). Moreover, the class separability information is also used to evaluate the training data quality (training sample subset selection) in remote sensing applications [23]. It is known that collecting a sufficient number of training samples to meet the requirements of Hughes Phenomena is very difficult for hyperspectral data and field spectroradiometer data (typically contains more than 1000 spectral channels). In this scenario, the computation of conventional class separability information would not be possible due to inverse of the covariance matrix term involved in the class separability measures. Therefore, it is of prime importance to design a class separability measure in the RKHS to effectively handle the nonlinear dependence of the data and to avoid singularity problems.

In this paper, we propose a new class separability measure in the RKHS as the objective function for feature selection. The proposed measure is defined using the Hilbert Schmidt independence criterion (HSIC) and surrogate kernel (SK) approach. In [24], the similarity measure using SK was derived to evaluate the PerTurbo classifier's performance with the training data characteristics and the similarity measures between the classes and the confusion matrix of the classifier showed some homogeneous structure. We build upon this observation to select the subset of features that increases the diagonal dominance of the similarity matrix, thereby possibly increasing the diagonal dominance of the confusion matrix. The proposed class separability matrix is different from the one in [24], in the sense that we use the HSIC to compare the similarity between the kernel matrices of two classes. The applicability of HSIC for feature selection is shown in [14], [19], and [25]. However, the potential of the HSIC for deriving class similarity measure has not been explored yet.

Beyond the design of the objective function, another key issue in feature selection is the search strategy. Generally, the search-based methods, such as sequential forward and backward search and evolutionary-based search methods, are employed to find the subset of the features. These methods incrementally add or remove bands according to a given criterion and involve a computationally intensive subset search operation. The LASSO or l_1 optimization method has shown to be able to model the discrete band selection problem as a continuous optimization problem [26] and thus avoids a complex subset strategy. Therefore, we adopted in this paper the LASSO model to select the subset of class discriminant features. The proposed feature selection method is developed by coupling the proposed class separability matrix and the LASSO model into a single unified framework called HSIC-SK LASSO. A related method to our HSIC-SK LASSO is given in [26], but the entries in the LASSO model and their length are different. The length of the input and output terms of our proposed method is dependent on the number of classes, i.e., L^2, where L is the number of classes present in the image. Conversely, the length of the entries in [26] is dependent on the number of training samples, i.e., n^2. Let us observe that for hyperspectral data sets, we generally have $n^2 \gg d$, where n is the number of training samples and d is the number of bands of hyperspectral data. As such, the model proposed in [26] might not be computationally feasible for hyperspectral data sets. Indeed, this model was proposed for ultrahigh-dimensional data sets.

Our main contributions are thus: 1) development of a new class separability matrix based on SK and HSIC in the RKHS and 2) effective design of a LASSO-based framework to select class separable features for hyperspectral data classification. The experiments conducted with three airborne hyperspectral data sets with different experimental settings show that our proposed feature selection method outperforms the state-of-the-art methods and it is computationally more efficient than sequential-search-based methods.

The rest of this paper is organized as follows. Section II presents the theoretical background of the proposed feature selection method. Section III proposes the class separability measure and the LASSO feature selection framework. Section IV details the extensive experiments conducted with the proposed method, including the comparison of some methods with the state-of-the-art methods. Finally, the conclusion and some research perspectives are given in Section V.

II. Theoretical Background

This section reviews the necessary theoretical background required for the proposed class separability measure.

Let $\boldsymbol{S} = \{(\mathbf{x_1}, y_1), (\mathbf{x_2}, y_2), \ldots, (\mathbf{x_n}, y_n)\} \in \mathbb{R}^d \times \{\omega_1, \ldots, \omega_L\}$ be the training set with N pairs of training samples $\mathbf{x_i}$ and their corresponding class labels y_i and $\boldsymbol{S_l}$ be the set of all training samples with label ω_l. The geometric structure of the samples belonging to the class ω_l can be characterized using the geometry of the kernel induced feature space. The geometry of the class ω_l can be represented by the Gram matrix in the RKHS as follows:

$$\boldsymbol{K_{ij}}(\boldsymbol{S_l}) = k(\mathbf{x_i}, \mathbf{x_j}) = \phi(\mathbf{x_i})^{\mathbf{T}}.\phi(\mathbf{x_j}) = \exp(-\gamma \|\mathbf{x_i} - \mathbf{x_j}\|^{\mathbf{2}}) \tag{1}$$

where $k(.,.)$ is a Gaussian radial basis kernel function, $\mathbf{x_i}$ and $\mathbf{x_j} \in \boldsymbol{S_l}$, ϕ is the mapping from the original space into RKHS, and γ is the bandwidth parameter of the Gaussian kernel.

A. Surrogate Kernel

Mercer's theorem states that any positive semidefinite kernel can be reconstructed from the kernel's eigenspectra and its continuous eigenfunctions [27]. Thus, Mercer's theorem provides an explicit way to generate the kernel matrix on any arbitrary samples. In general, the kernel matrices cannot be compared directly using standard matrix metrics because they do not share the same size. The SK approach [28] allows us to adapt the Gram matrix based on a new set of support functions, yielding matrices of comparable sizes. The SK [28]

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

can be defined as follows: let $\boldsymbol{X}$ and $\boldsymbol{Z}$ be any given two samples and $k(.)$ be a kernel function, then $\boldsymbol{K_X} \in \mathbb{R}^{|\boldsymbol{X}|\times|\boldsymbol{X}|}$ and $\boldsymbol{K_Z} \in \mathbb{R}^{|\boldsymbol{Z}|\times|\boldsymbol{Z}|}$ are the kernel matrices defined on $\boldsymbol{X}$ and $\boldsymbol{Z}$, respectively, and $\boldsymbol{K_{XZ}} \in \mathbb{R}^{|\boldsymbol{X}|\times|\boldsymbol{Z}|}$ is a kernel matrix defined among $\boldsymbol{X}$ and $\boldsymbol{Z}$. The SK of $\boldsymbol{K_X}$ on the sample $\boldsymbol{Z}$, denoted by $\boldsymbol{K_{Z\leftarrow X}}$, is defined as

$$\boldsymbol{K_{Z\leftarrow X}} = \boldsymbol{K_{ZX}}\boldsymbol{K_X^{-1}}\boldsymbol{K_{XZ}}. \tag{2}$$

The kernel matrix $\boldsymbol{K_X}$[1] and its SK $\boldsymbol{K_{Z\leftarrow X}}$ are built from the same set of set of eigenvalues and eigenfunctions [28], and thus $\boldsymbol{K_{Z\leftarrow X}}$ preserves the key structure of the kernel matrix $\boldsymbol{K_X}$.

B. Hilbert Schmidt Independence Criterion

Another way to compare the geometry of kernel embeddings is found in the use of a statistical independence criterion: the HSIC. It measures the independence between two sets of random variables [29]. Let $\boldsymbol{X}$ and $\boldsymbol{Z}$ be the two random variables, from which the samples $(\mathbf{x}, \mathbf{z})$ can be drawn from the probability density function of $\boldsymbol{X}$ and $\boldsymbol{Z}$. The nonlinear mapping function is defined on each element of $\boldsymbol{X}$, as $\phi(\mathbf{x}) \in F$ from $\mathbf{x} \in \boldsymbol{X}$ to the feature space F, such that the inner product between the features is given by a kernel function $k(\mathbf{x}, \mathbf{x}') = \langle\phi(\mathbf{x}), \phi(\mathbf{x}')\rangle$, and F is the associated RKHS. In a similar manner, let G be the RKHS on $\boldsymbol{Z}$ with kernel $l(.,.)$ and the mapping function $\psi(\mathbf{z})$. Then the cross-covariance operator between these two mapping functions can be defined as a linear operator $C_{xz} : G \rightarrow F$, such that

$$\begin{aligned} C_{\mathbf{xz}} &= E_{\mathbf{xz}}[(\phi(\mathbf{x}) - \mu_{\mathbf{x}}) \otimes (\psi(\mathbf{z}) - \mu_{\mathbf{z}})] \\ &\Rightarrow E_{\mathbf{xz}}[\phi(\mathbf{x}) \otimes \psi(\mathbf{z})] - \mu_{\mathbf{x}} \otimes \mu_{\mathbf{z}} \end{aligned} \tag{3}$$

where $\otimes$ is a tensor product. The HSIC is defined as the squared Hilbert Schmidt norm of (3). It has been shown that the HSIC can be expressed in terms of kernel [29] and the empirical estimate of the HSIC is given as follows:

$$\mathrm{HSIC}(\boldsymbol{Z}, \boldsymbol{F}, \boldsymbol{G}) = (m-1)^{-2} tr(\boldsymbol{KCLC}) \tag{4}$$

where m is the number of observations (samples), $\boldsymbol{C}, \boldsymbol{K}, \boldsymbol{L} \in \mathbb{R}^{m\times m}$, $\boldsymbol{K}_{ij} = k(\mathbf{x_i}, \mathbf{x_j})$, $\boldsymbol{L}_{ij} = l(\mathbf{z_i}, \mathbf{z_j})$, $\boldsymbol{C}_{ij} = \delta_{ij} - m^{-1}$ ($\delta_{ij} = 1$ if $i = j$, zero otherwise) is the centering matrix, and tr is the trace operator. The independence between the two random variables in the Hilbert space can be obtained by (4).

III. Proposed Framework

In this section, we propose a new class separability measure and the LASSO-based feature selection framework.

A. Proposed Similarity Measure in Kernel Hilbert Space

We assume that the samples belonging to different classes follow different probability distributions. Measuring similarity between the classes can then be framed as computing the distances between the probability distributions. Let $\boldsymbol{K}(\boldsymbol{S_l})$ and $\boldsymbol{K}(\boldsymbol{S_{l'}})$ be the kernel matrices of the samples belonging to the classes ω_l and $\omega_{l'}$ in the RKHS as described in (1). Measuring the distribution of the data in the Hilbert space is not straightforward, since the kernel-induced feature maps in the RKHS cannot be expressed explicitly [28], [30]. The kernel matrix has the capability to uniquely describe the feature space geometry of the data samples in the RKHS. Therefore, for two samples (from class ω_l and $\omega_{l'}$) to have similar feature space distributions, we require them to have similar kernel matrices.

Let $\Psi(\boldsymbol{S_l})$ and $\Psi(\boldsymbol{S_{l'}})$ be the empirical kernel maps of the samples belonging to the classes ω_l and $\omega_{l'}$, respectively. If the two kernel matrices are identical, then their corresponding empirical kernel maps will also be identical. This implies that the empirical distributions of the data in the kernel induced feature space will also be equal. In other words

$$\boldsymbol{K}(\boldsymbol{S_l}) = \boldsymbol{K}(\boldsymbol{S_{l'}}) \Rightarrow \Psi(\boldsymbol{S_l}) = \Psi(\boldsymbol{S_{l'}}) \tag{5}$$

$$\Rightarrow p(\Psi(\boldsymbol{S_l})) = p(\Psi(\boldsymbol{S_{l'}})) \tag{6}$$

where $p(\cdot)$ is the probability distribution function in the Hilbert space. Thus, measuring similarity between the data distributions in the Hilbert space (feature space) can be viewed as aligning the two kernel matrices.

In remote sensing images, the spatial distribution of the classes will be different to each other. As a result, the number of samples belonging to each class will also be different. Thus, evaluating the similarity or closeness between the two kernel matrices is difficult, since the kernel matrices of the different classes may have different dimensions. In order to overcome this problem, we generate a new kernel matrix called SK [28]. To compare the kernel matrices of two classes $\boldsymbol{K}(\boldsymbol{S_l})$ and $\boldsymbol{K}(\boldsymbol{S_{l'}})$, we construct the surrogate kernel of $\boldsymbol{S_l}$ based on the sample $\boldsymbol{S_{l'}}$ as follows:

$$\boldsymbol{K}(\boldsymbol{S_{l'}} \leftarrow \boldsymbol{S_l}) = \boldsymbol{K}(\boldsymbol{S_{l'}}, \boldsymbol{S_l})\boldsymbol{K}(\boldsymbol{S_l})^{-1}\boldsymbol{K}(\boldsymbol{S_l}, \boldsymbol{S_{l'}}) \tag{7}$$

where $\boldsymbol{K}(\boldsymbol{S_{l'}} \leftarrow \boldsymbol{S_l})$ is an SK of $\boldsymbol{K}(\boldsymbol{S_l})$. Now the similarity between the class ω_l and $\omega_{l'}$ can be evaluated by measuring the closeness between $\boldsymbol{K}(\boldsymbol{S_{l'}} \leftarrow \boldsymbol{S_l})$ and $\boldsymbol{K}(\boldsymbol{S_{l'}})$. The similarity between $\boldsymbol{K}(\boldsymbol{S_{l'}} \leftarrow \boldsymbol{S_l})$ and $\boldsymbol{K}(\boldsymbol{S_{l'}})$ is measured using the HSIC as

$$\mathrm{HSIC}(\omega_l, \omega_{l'}) = (m_{l'} - 1)^{-2} tr(\boldsymbol{K}(\boldsymbol{S_{l'}} \leftarrow \boldsymbol{S_l})\boldsymbol{C}\boldsymbol{K}(\boldsymbol{S_{l'}})\boldsymbol{C}) \tag{8}$$

where $m_{l'}$ is the number of the samples in the class $\omega_{l'}$ and $\boldsymbol{C}_{ij} = \delta_{ij} - m_{l'}^{-1}$, in which $\delta_{ij} = 1$ if $i = j$ zero otherwise. The pairwise Hilbert Schmidt class similarity matrix between the classes is then defined as

$$\boldsymbol{H} = \begin{bmatrix} \mathrm{HSIC}(\omega_1, \omega_1) & \mathrm{HSIC}(\omega_1, \omega_2) & \cdots & \mathrm{HSIC}(\omega_1, \omega_L) \\ \mathrm{HSIC}(\omega_2, \omega_1) & \mathrm{HSIC}(\omega_2, \omega_2) & \cdots & \mathrm{HSIC}(\omega_2, \omega_L) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{HSIC}(\omega_L, \omega_1) & \mathrm{HSIC}(\omega_L, \omega_2) & \cdots & \mathrm{HSIC}(\omega_L, \omega_L) \end{bmatrix}. \tag{9}$$

The high values of the matrix $\boldsymbol{H}$ indicate that the classes have high similarity and the low values indicate that the classes have less similarity. The pairwise Hilbert Schmidt class similarity

[1] In the implementation, a Tikhonov regularization $\epsilon = 10^{-4}$ is added to $\boldsymbol{K_X}$ before computing inverse to avoid numerical issues.

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

matrix between the classes computed with the kth feature of hyperspectral data is given by

$$\boldsymbol{H}^{\boldsymbol{k}} = \begin{bmatrix} \mathrm{HSIC}^k(\omega_1,\omega_1) & \mathrm{HSIC}^k(\omega_1,\omega_2) & \cdots & \mathrm{HSIC}^k(\omega_1,\omega_L) \\ \mathrm{HSIC}^k(\omega_2,\omega_1) & \mathrm{HSIC}^k(\omega_2,\omega_2) & \cdots & \mathrm{HSIC}^k(\omega_2,\omega_L) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{HSIC}^k(\omega_L,\omega_1) & \mathrm{HSIC}^k(\omega_L,\omega_2) & \cdots & \mathrm{HSIC}^k(\omega_L,\omega_L) \end{bmatrix}. \tag{10}$$

B. Proposed LASSO Feature Selection Framework

If the features have good class discriminative information, then matrix $\boldsymbol{H}$ will be a diagonal dominant matrix. Therefore, the problem is to find the subset of features that maximizes the diagonal dominance of the matrix $\boldsymbol{H}$. The sequential-search-based methods can be employed to find the subset of features; however, they require a computationally intensive subset search strategy. Instead, we formulate in this paper, the discrete band selection problem as a continuous band selection problem using the LASSO optimization method. We denote our proposed LASSO-based band selection framework by HSIC-SK LASSO. It is defined as

$$J(\boldsymbol{\alpha}) = \min_{\boldsymbol{\alpha}} \frac{1}{2} \left\| \hat{\boldsymbol{I}} - \sum_{k=1}^{d} \alpha_k \boldsymbol{H}^k \right\|_F^2 + \lambda \|\boldsymbol{\alpha}\|_1 \tag{11}$$

where $\hat{\boldsymbol{I}} = 1/L(\boldsymbol{I} + \epsilon)$ is a target matrix, $\boldsymbol{I}$ is the identity matrix, $\epsilon = 10^{-4}$ is a constant, $\boldsymbol{H}^k$ is the class separability measure of the kth feature in the hyperspectral data, $\boldsymbol{\alpha}$ is the coefficient vector of length d and nonzero entries of the $\boldsymbol{\alpha}$ correspond to the selected features retained by HSIC-SK LASSO, and λ is the scalar value to control the sparsity of the coefficient vector. The first term of (11) is the squared loss function and the second term is the regularization term. The lengths of the output term ($\hat{\boldsymbol{I}}$) and input term ($\boldsymbol{H}^k$) are dependent on the number of classes present in the image and they are of length L^2. The entries of our proposed method are different from the LASSO model available in [26]. The ideal kernel and featurewise kernel matrix were used as the entries in [26], thus leading to a large number of observations for the LASSO model. The dimensions of output (ideal kernel) and input (featurewise kernel) terms are of length n^2, and thus always $n^2 \gg d$. Due to this, the existing formulation from [26] might not be computationally feasible[2] for feature selection of hyperspectral data. Our proposed model does not depend on the number of training samples but rather on the number of classes, and thus often $L^2 < d$, which is preferable for the LASSO model.

1) Interpretation of HSIC-SK LASSO: The HSIC-SK LASSO formulated in (11) aims to identify highly class separable features that have large inner product between $\hat{\boldsymbol{I}}$ and $\boldsymbol{H}^{\boldsymbol{k}}$. Expanding the first term of (11) and expressing in terms of kernel matrices will provide more insight into our proposed method. For the simplicity, we assume that kernel matrices are centralized, then using (3) the squared loss term in (11) can be expressed as

$$\frac{1}{2} \left\| \hat{\boldsymbol{I}} - \sum_{k=1}^{d} \alpha_k \boldsymbol{H}^k \right\|_F^2 = \Gamma - \underbrace{\sum_{k=1}^{d} \alpha_k \mathbf{A}_k}_{\textbf{class separability measure term}} + \underbrace{\frac{1}{2} \sum_{k,k'=1}^{d} \alpha_k \alpha_{k'} \mathbf{B}_{k,k'}}_{\textbf{correlation measure term}}$$

where

$$\mathbf{A}_k = \sum_{l,l'=1}^{L} \langle K^k(S_l), K^k(S_{l'}\rangle_F \left(\frac{\delta_{l,l'}}{L} + \epsilon \right)$$

and

$$\mathbf{B}_{k,k'} = \sum_{l,l'=1}^{L} \left\langle K^k(S_l) K^{k'}(S_l) \right\rangle_F \left\langle K^k\left(S_l'\right) K^{k'}(S_{l'}) \right\rangle_F. \tag{12}$$

In (12) Γ, $\epsilon = 10^{-4}$ are constant[3] and K^k is the kernel matrix computed with the kth feature of hyperspectral data.[4] The second term in (12) can be interpreted based on class discriminative information of the features; when the kth feature has less within-class variance and high between-class variance, this term takes a large value and correspondingly α_k should also take a large value such that (11) is minimized. On the contrary to that, when the kth feature is not discriminative, the second term takes a smaller value, and thus α_k tends to be eliminated by the regularization term. This implies that important features that have better class separability tend to be selected by our proposed HSIC-SK LASSO. The third term in (12) provides more interesting interpretation as it has the capability to encode the correlation information between the features in a classwise manner. The first two kernel matrices measure correlation between the kth and k'th feature with respect to class ω_l and the remaining two entries correspond to correlation between the kth and k'th features relative to class $\omega_{l'}$. When the k and k'th features are correlated relative to each class, the third term takes a larger value, so either of α_k and $\alpha_{k'}$ tends to be zero. On the other hand, when the features are independent relative to class ω_l and $\omega_{l'}$ or interestingly with either of the classes, then the corresponding features tends to be retained by our method. Thus, (11) eliminates the correlated features in a classwise manner, which is an important property of our proposed method. Furthermore, the second term in (12) can be related to [26], but the third term is different because our method has the advantage of including the feature correlation in a classwise manner.

The proposed formulation given in (11) is solved using the statistical and machine learning toolboxes available in MATLAB.[5]

[2]We have conducted experiments with the model proposed in the work of Yamada *et al.* [26] using the codes provided by them and we found that the model did not converge for hyperspectral image feature selection.

[3]ϵ is added in the target matrix ($\hat{\boldsymbol{I}}$) to include the between-class similarity in the selection of features.

[4]The normalizing terms in (12) are ignored for clarity purpose.

[5]The MATLAB codes of the proposed method will be provided upon request.

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

TABLE I
PAVIA UNIVERSITY TRAINING AND TESTING SAMPLES

No	Class name	Training	Testing
1	Asphalt	548	6641
2	Meadows	540	18649
3	Gravel	392	2099
4	Trees	524	3064
5	Metal Sheets	265	1345
6	Soil	532	5029
7	Bitumen	375	1330
8	Bricks	514	3682
9	Shadows	231	947
	Total	3921	42776

TABLE II
CITY OF PAVIA TRAINING AND TESTING SAMPLES

No	Class name	Training	Testing
1	Water	824	65971
2	Trees	820	7598
3	Gravel	824	3090
4	Trees	808	2685
5	Metal Sheets	820	6584
6	Soil	816	9248
7	Bitumen	808	7287
8	Bricks	1260	42826
9	Shadows	476	2863
	Total	7456	148152

TABLE III
KSC TRAINING AND TESTING SAMPLES

No	Class name	Training	Testing
1	Scrub	114	761
2	Willow	36	243
3	CP Hammock	38	256
4	CP/Oak	37	252
5	Slash Pine	24	161
6	Oak/Broadleaf	34	229
6	Hardwood swamp	15	105
7	Graminoid marsh	64	431
8	Spartina marsh	78	520
9	Cattail Marsh	60	404
10	Salt marsh	62	419
11	Mud flats	75	503
12	Water	139	927
	Total	776	5211

IV. EXPERIMENTAL RESULTS AND DISCUSSION

A. Data Sets

In order to evaluate the proposed HSIC-SK LASSO feature selection method, experiments have been conducted with real-world hyperspectral data sets. A detailed description of these data is provided in the following.

1) Pavia University: The first hyperspectral data considered here were collected over the University of Pavia, Italy, by the ROSIS airborne hyperspectral sensor in the framework of the HySens project managed by DLR (German national aerospace agency). The ROSIS sensor collects images in 115 spectral bands in the spectral range from 0.43 to 0.86 μm with a spatial resolution of 1.3 m/pixel. After the removal of noisy bands, 103 bands were selected for experiments. This data contain 610×340 pixels with nine classes of interest. The training and testing samples are provided along with data and are used to perform quantitative evaluation (see Table I).

2) City of Pavia: The second hyperspectral data were collected over the City of Pavia, Italy, by the ROSIS sensor. The spectral and spatial resolution configurations are similar to the Pavia University data set. After the removal of noisy bands, 102 bands were selected for experiments. This data set contains 1096×715 pixels with nine classes of interest. Similar to Pavia University, the City of Pavia data set comes with some training and testing samples (see Table II).

3) Kennedy Space Center: The third and last hyperspectral data considered here were collected over the Kennedy Space Center (KSC)[6] by the NASA Airborne Visible/Infrared Imaging Spectrometer (AVIRIS). The AVIRIS sensor collects images in 224 bands of 10-nm width with center wavelengths from 400 to 2500 nm. The KSC data, acquired from an altitude of approximately 20 km, have a spatial resolution of 18 m. After removing water absorption and low SNR bands, 176 bands were used for the analysis. This data set consists of 512×614 pixels with 13 classes of interest. Training and testing samples made available with this data set are given in Table III.

[6]Available online: http://www.csr.utexas.edu/hyperspectral/data/KSC/.

B. Experimental Design

In the proposed HSIC-SK LASSO method, we used radial basis function (RBF) kernel with kernel width (σ) set to be the fifth percentile of pairwise distances (PDs) [31], [32]. The λ in (11) are varied in the range $\lambda = 10^R$, where $R = \{-12, -11 \ldots, 1\}$, to select a different number of spectral bands.[7] The high λ value favors to select a more number of spectral bands and low value favors to select a less number of spectral bands.

[7]LASSO may not extract exactly the required number of bands; nevertheless, in most of the cases, it is possible to extract exactly or very close to the required number of spectral bands.

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

The effectiveness of the features (bands) selected by the proposed HSIC-SK LASSO method is assessed by the overall accuracy of supervised classifiers. The classifiers considered in our experiments are PerTurbo [24], SVM [33], extreme learning machine (ELM) [34]–[36], and naive Bayes classifier (NBC) [37]. The choice of these classifiers is motivated by their demonstrated performances in hyperspectral data classification. Besides, our goal is to show the genericity of the selection strategy within a wide range of classifiers. The RBF is used as the kernel function in the SVM, PerTurbo, and ELM classifiers. The SVM parameter cost function $C = 2^{\alpha}$, $\alpha = \{-5, -4, \ldots, 15\}$, and bandwidth $\gamma = 2^{\beta}$, $\beta = \{-15, -13, \ldots, 3\}$, were automatically tuned with the grid search method using fivefold cross validation. Similarly, the parameters of the PerTurbo classifier are automatically tuned with $\tau = \{10^{-6}, 10^{-5}, \ldots, 0.99\}$ and $\gamma = 2^{\beta}$, $\beta = \{-15, -13, \ldots, 3\}$. The performance of the ELM classifier is independent of the number of hidden neurons [38]. However, we have tuned here parameters using fivefold cross validation for optimal performance. The number of hidden neurons varies from 100 to 1000.

The experiments are conducted in three different settings.

1) Feature selection is performed on original spectral bands.
2) Feature selection is performed on the extended multiattribute profile (EMAP) [39] of the hyperspectral data. In order to generate the EMAP features, a principal component analysis is carried out on the hyperspectral data and the first four principal components (99% of total variance) are retained. The EMAPs are then generated from these four PCs using four multiattribute filters, namely, area, diagonal of the region bounding box, moment of inertia, and standard deviation [40], [41]. It leads to a total number of EMAP features equal to 304.
3) Finally, the feature selection experiment is performed with randomization of training samples to assess the stability of selected features.

C. Competitive Methods

We have compared our method with six specific implementations of two state-of-the-art band selection methods, namely, kernel dependence measure (KDM) [14], [19] and constrained band selection (CBS) [16]. The KDM-based feature selection method was considered, since it works similarly to our principle using HSIC. It consists of two different strategies using forward HISC (FOHSIC) and backward HISC (BAHSIC) search. The CBS was considered since it has been widely used to compare the performance of band selection methods. The CBS provides two different approaches named constrained energy minimization (CEM), and linearly constrained minimum variance with four different criteria [band correlation minimization (BCM), band correlation constraint (BCC), band dependence minimization (BDM), and band dependence constraint (BDC)]. It has been shown that BCC and BCM perform identically to BDC and BDM, respectively [16]. Thus, our experimental framework includes six specific implementations to compare the results with our proposed method. Finally, we also report the results obtained using all spectral bands (full-band) as a baseline.

D. Feature Selection on the Original Spectral Bands

We discuss here the experimental results achieved by the feature selection methods when applied on the original spectral bands of the hyperspectral data.

1) Pavia University: For each of the four different classifiers, we report in Fig. 1 the classification accuracy obtained with the different feature selection methods on the Pavia University data set. The classification accuracy using all spectral bands of the hyperspectral data is also included as a baseline (horizontal line) to compare with the feature selection methods. Fig. 1 reveals that when the most informative bands are added, the classification accuracy changes drastically, and when additional (possibly redundant) bands are added, the classification accuracy then changes slowly. The proposed HSIC-SK LASSO feature selection method outperforms its competitors with all the four classifiers. The significance of the proposed method is analyzed in two ways: 1) the number of spectral bands required to approximate full-band classification accuracy (i.e., first peak close to the horizontal line) and 2) the percentage of improvement in classification accuracy compared with the full-band classification accuracy and the number of bands required to achieve the best classification accuracy.

When the number of spectral bands required to approximate full-band accuracy is considered, the proposed method outperforms the existing feature selection methods and is able to approximate the full-band classification accuracy with fewer spectral bands relative to all the four classifiers. Indeed, our method requires only 16% of spectral band for the SVM classifier, 9% for the PerTurbo and ELM classifier, and 3% for NBC to approximate full-band classification accuracy. On the other hand, the existing feature selection methods require a higher number of spectral channels to approximate full-band accuracy. Moreover, their performance is not consistent over the different classifiers. To illustrate, these methods require more than 40% of spectral bands to approximate the full-band accuracy with three classifiers (SVM, PerTurbo, and NBC). Among the competitors, CEM–BCC/BDM offers a better performance with three classifiers compared with other existing methods.

When the improvement in classification accuracy is considered, the proposed method shows about 1–7% increase in classification accuracy in comparison with the full-band classification accuracy. Few of the competitors also improved the classification accuracy. However, they require a much more number of spectral bands to achieve the best performance compared with our proposed method. Among the classifier, the ELM classifier has benefited more from our proposed method, and thus it reaches a similar overall accuracy (80%) with the SVM classifier using all the spectral bands with low computational complexity.[8] The NBC is least benefited from

[8]The computational time of SVM classifier (including cross validation) with all the spectral bands is 716 s, whereas the total computation time of HSIC-SK LASSO and ELM classifiers (including cross validation) with 11 spectral bands is 359 s.

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

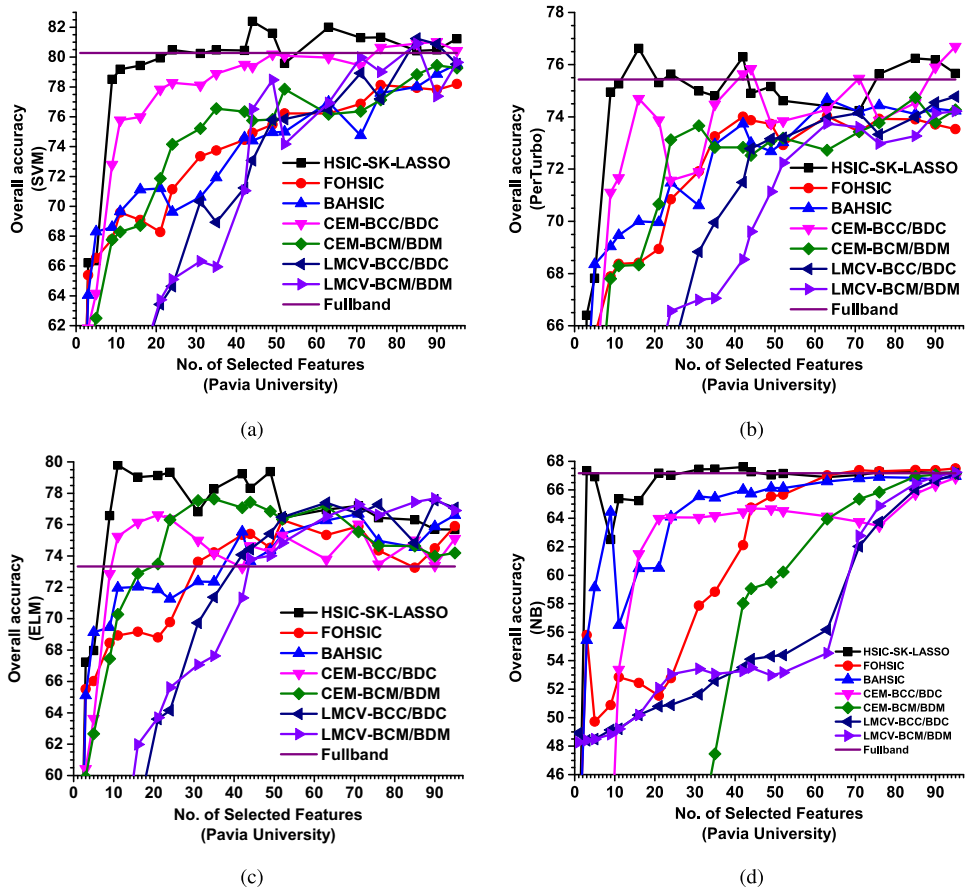

Fig. 1. Classification accuracy of the Pavia University data set with different feature selection methods. (a) SVM classifier. (b) PerTurbo classifier. (c) ELM classifier. (d) NBC.

our proposed method in terms of accuracy improvement, but it is most benefited classifier in achieving full-band accuracy with very few spectral bands. Since it is known that NBC performs better when the feature are independent, we can thus derive that our method provides more informative and less redundant bands.

2) City of Pavia: Fig. 2 reports the classification accuracies for the City of Pavia data set with all the considered feature selection methods and classifiers. We can observe that our proposed method outperforms the existing feature selection methods in two aspects: the minimal number of spectral bands required to approximate the full-band classification accuracy and the improvement in classification accuracy with fewer spectral bands. Furthermore, the proposed method performs better with all the considered state-of-the-art classifiers. The existing methods requires a more number of spectral bands to approximate full-band accuracy and to increase the classification accuracy. Among the existing methods, kernel dependence minimization and CEM–BCM/BDM produce better classification results. The classification results of the NBC are worth to note, since the proposed method is able to approximate the full-band classification accuracy with only three spectral bands.

3) Kennedy Space Center: The classification accuracies for the KSC data set are finally reported in Fig. 3. We can see that our method is able to approximate the full-band classification accuracy with a minimal number of spectral bands. Fig. 3 reveals that our proposed method outperforms the competitors with three classifiers and performs similarly with the PerTurbo classifier. When the improvement in classification accuracy is considered, our method outperforms the competitors with ELM and NBC and performs comparatively equal to the SVM classifier. Moreover, ELM and NBC benefited most from the feature selection methods and improves about 5%–8% accuracy compared with the full-band accuracy. The comparative analysis of feature selection methods highlight that our proposed method behaves well with the minimal number of spectral bands, but it decreases the accuracy when a more number of spectral bands are included. On the other hand, existing methods have converse behavior to our method. That is, the accuracy improves when a more number of spectral bands are included.

E. Interpretation of Selected Features and Computational Efficiency of HSIC-SK LASSO Over Sequential-Search-Based Methods

The selected features by the proposed method and considered existing methods are reported in Fig. 4. For the Pavia University and Pavia Center data sets, 16 selected spectral bands are presented in Fig. 4(a) and (b), and for the KSC data set, 20 selected spectral bands are presented in Fig. 4(c). The longer width of horizontal bar indicates that a more number

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

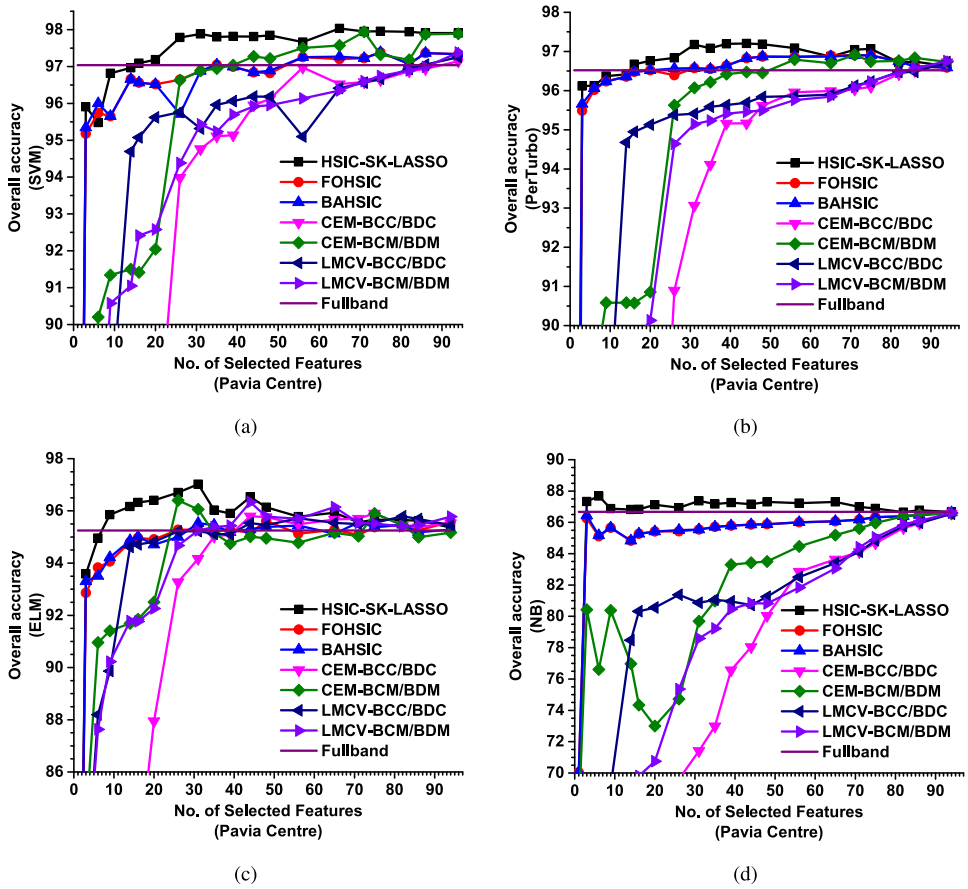

Fig. 2. Classification accuracy of the City of Pavia data set with different feature selection methods. (a) SVM classifier. (b) PerTurbo classifier. (c) ELM classifier. (d) NBC.

of spectral bands are selected in that particular region of the spectrum. The visual inspection of Fig. 4 highlights the potential of our method in selecting the spectral bands from entire region of the spectrum to adapt diverse natural and man-made objects in the Pavia University and Pavia Center data sets, and thus results in nonredundant selection of the features. On the other hand, the selected bands by existing methods are concentrated on the particular region of the spectrum. For the KSC data set, our method selects most of the bands in the lower region of the spectrum with the combination of a few bands in higher region of spectrum. The selected bands in the lower region typically encounter the leaf pigments and chlorophyll absorption (1–30 bands), plant cell structure (NIR region), and higher region for the soil moisture and leaf water content. The HSIC-based methods select most of bands in the NIR region, but it neglects the bands associated with leaf pigments and leaf water content. The LMCV-BCM favors the bands in the mid and higher regions of the spectrum and favors less in the region of chlorophyll absorption and the NIR region. The comparative analysis of Figs. 1–4 demonstrates that the feature selection methods, which uniformly cover the entire region of the spectrum, produce good classification results.

Table IV reports the computational time (in seconds) of our proposed method and HSIC-based sequential search methods.[9] The reported computational time is measured only for the feature selection stage and does not include classification step. Table IV shows that our proposed method is computationally much faster than sequential-search-based methods. Specifically, our method is 8×, 48× times faster than the FOHSIC and BAHSIC[10] methods for the Pavia University data set and this order further increases for the Pavia Center data set. More interestingly, the computational time difference of our method is negligible when a more number of features are selected, whereas the sequential-search-based methods increases the computational time drastically as the number of selected features increases. This computational advantage is due to the potential of our modeling of feature selection problem as a continuous LASSO optimization problem. Our proposed method is computationally very efficient once the design matrix is computed. For instance, the computation of design matrix accounts for 345 s (the value before the plus sign in HSIC-SK LASSO) for the Pavia University data set and the selection of features is performed in negligible time.

F. Feature Selection on the Extended Morphological Attribute Profiles

Having reported the performance of our method on original hyperspectral data, we now discuss its performance when applied on EMAP features. We consider here only the Pavia

[9]The number of selected features mentioned in Table IV may not be exactly the same for the HSIC-SK LASSO method.

[10]The computational complexity of kernel computations involved in our proposed method and other HSIC methods is similar.

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

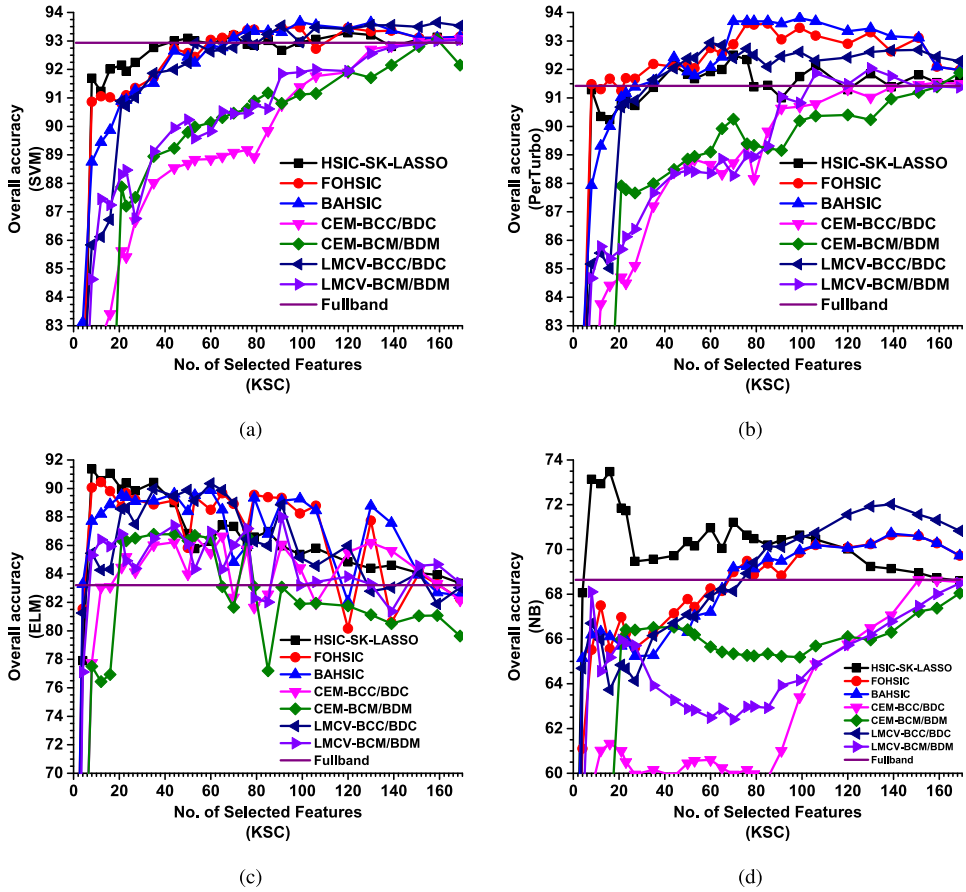

Fig. 3. Classification accuracy of the KSC data set with different feature selection methods. (a) SVM classifier. (b) PerTurbo classifier. (c) ELM classifier. (d) NBC.

TABLE IV

COMPUTATIONAL TIME (IN SECONDS) MEASURED OVER A DIFFERENT NUMBER OF SELECTED FEATURES OF THE PROPOSED HSIC-SK LASSO AND HSIC-BASED SEQUENTIAL SEARCH METHODS. FOR THE HSIC-SK LASSO, THE TIME MENTIONED BEFORE THE PLUS SIGN ACCOUNTS FOR COMPUTING THE DESIGN MATRIX AND AFTER THE PLUS SIGN ACCOUNTS FOR THE SELECTION OF FEATURES

No of Features	Pavia University			Pavia Centre			KSC		
	HSIC-SK LASSO	FOHSIC	BAHSIC	HSIC-SK LASSO	FOHSIC	BAHSIC	HSIC-SK LASSO	FOHSIC	BAHSIC
10	345+06	2934	16808	863+13	16152	91730	20+78	78	723
20	345+08	5642	16293	863+14	30766	88855	20+79	150	716
30	345+08	8230	15473	863+15	43199	84215	20+80	214	704
40	345+09	10023	14343	863+16	54053	77811	20+81	274	688
50	345+12	11769	12887	863+20	63257	69659	20+82	334	667

University and Pavia Center data sets[11] and provide related results in Table V. Classification accuracy is measured for different numbers of EMAP selected by our method, but here we present the results only for a few selected EMAPs. As expected, the classification accuracy significantly increases with the EMAP features compared with the pixel based classification. To illustrate, the PerTurbo classifier achieves 92.8% and 99.15% of classification accuracy with the Pavia University and Pavia Center data sets, respectively. However, we can remark that the classification accuracy significantly decreases with NBC. From Table V, we can clearly see that our method is able to approximate the original EMAP classification accuracy (with 304 EMAP) using only a much smaller number of EMAP features. More precisely, original EMAP classification accuracy is approximated for the Pavia University data set with about 8% of original EMAP features with SVM and PerTurbo, 5% with ELM, and

[11]The KSC data set is not considered here, since the spatial locations of training samples are not provided.

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

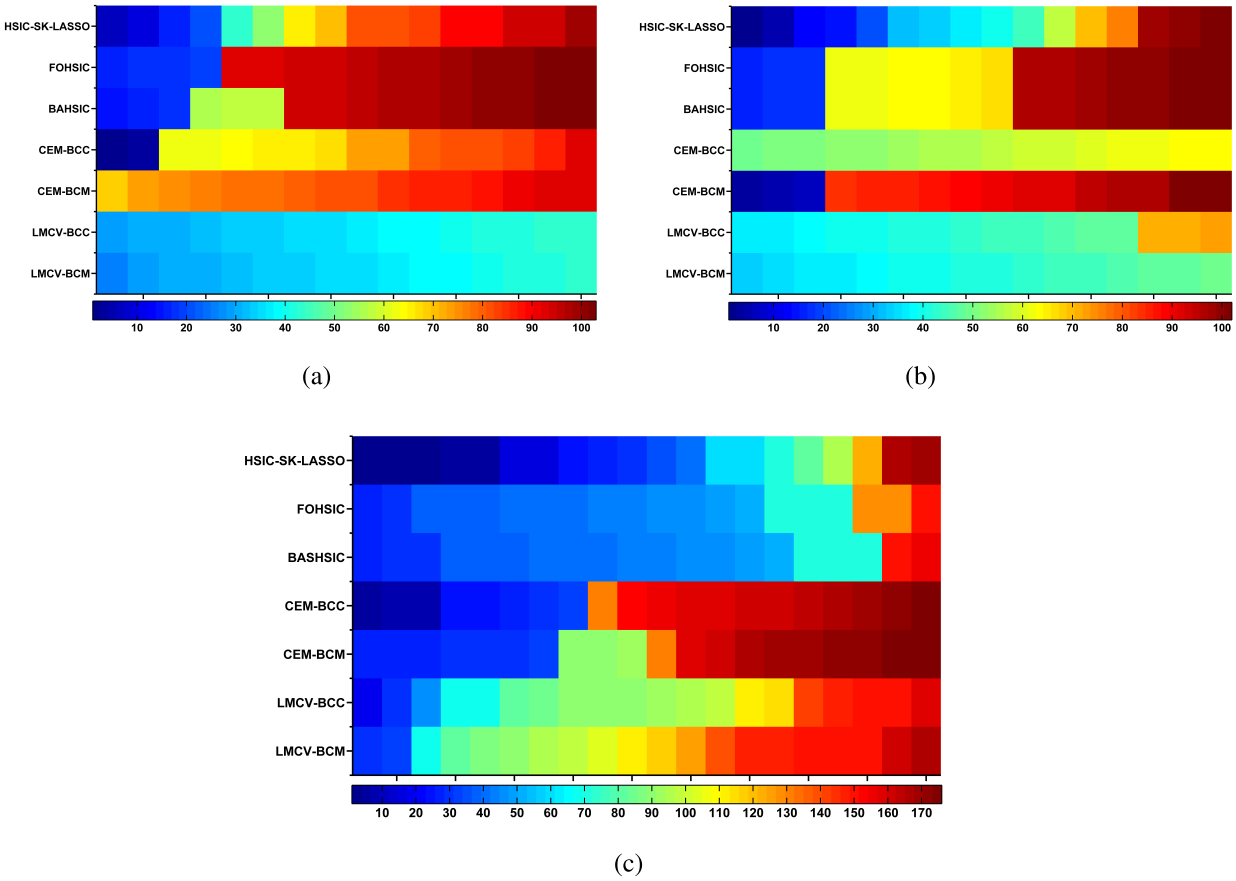

Fig. 4. Selected spectral bands with different feature selection methods. (a) Pavia University (16 selected spectral bands). (b) Pavia Center (16 selected spectral bands). (c) KSC (20 selected spectral bands). The color bar indicates the spectral band indexes, the blue patterns denote the lower band indexes and the red patterns indicate higher spectral band indexes.

TABLE V

CLASSIFICATION PERFORMANCE (OA IN %) OF THE SELECTED FEATURES WITH THE PROPOSED HSIC-SK LASSO METHOD ON THE EMAPS FOR THE PAVIA UNIVERSITY AND PAVIA CENTER DATA SETS

Pavia University					Pavia Centre				
# of features	SVM	PerTurbo	ELM	NB	# of features	SVM	PerTurbo	ELM	NB
3	74.86	78.71	70.81	67.29	2	80.34	75.01	82.94	75.36
6	78.26	80.80	82.80	19.44	8	98.17	99.21	96.72	92.68
10	73.39	75.91	73.40	16.30	10	97.42	99.13	96.60	91.75
15	75.57	81.81	79.61	16.27	15	98.83	99.23	97.69	92.68
20	74.48	80.41	77.02	15.58	20	97.75	98.92	98.53	68.83
26	94.25	93.01	93.97	15.53	26	98.58	99.10	97.93	63.09
31	93.21	95.55	92.52	15.53	31	98.66	99.18	98.58	50.11
36	94.96	96.38	89.88	15.53	36	98.78	99.12	98.21	48.88
304 (all)	94.99	92.80	79.33	15.83	304 (all)	98.88	99.15	97.61	44.53

1% with NBC, respectively. Furthermore, the EMAPs selected by our method are also able to increase the classification accuracy when compared to original EMAP. Among the four classifiers, the improvement in accuracy is higher in magnitude with the ELM and NBCs compared with the SVM and PerTurbo classifiers (e.g., 14% and 51% with Pavia University). The analysis of the selected EMAP features (26 for Pavia University and 15 for Pavia Center) reveals that filters from all the attributes are selected, but it favors only two set of attributes for both the data sets. Among them,

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

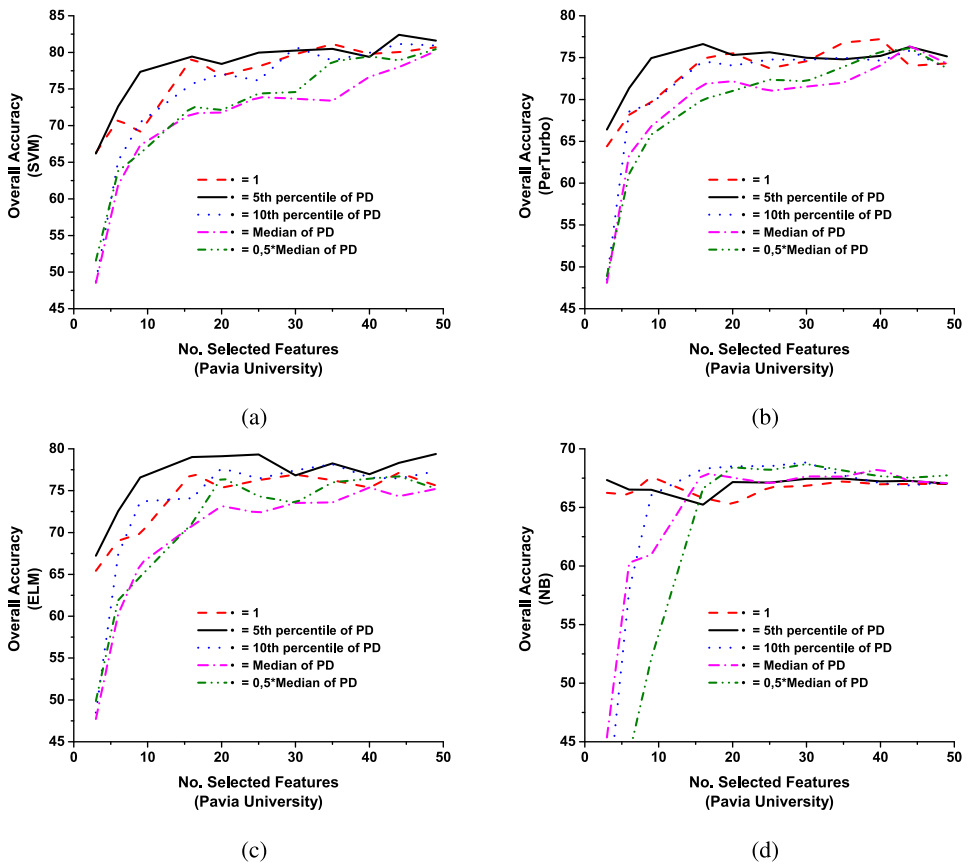

Fig. 5. Classification accuracy of HSIC-SK LASSO method with different heuristic Gaussian bandwidths on the Pavia University data set. (a) SVM classifier. (b) PerTurbo classifier. (c) ELM classifier. (d) NBC.

selection highly favored area and standard deviation attributes for the Pavia University data set, and it highly favored area and moment of inertia attributes for the Pavia Center data set.

G. Influence of the Gaussian Width on the HSIC-SK LASSO Method

In this section, we carried out of set of experiments with two hyperspectral data to investigate the influence of the Gaussian bandwidth parameter on our proposed HSIC-SK LASSO method. For the input $\mathbf{x_i}, \mathbf{x_j} \in \mathbb{R}^n$, the Gaussian RBF kernel is given as $K(\mathbf{x_i}, \mathbf{x_j}) = \exp(-1/2\sigma^2\|\mathbf{x_i} - \mathbf{x_j}\|^2)$, where σ is the Gaussian kernel width. Many studies showed that the heuristic approaches can be used to compute σ of the Gaussian kernel, in this case σ was set to: 1) the dimension of the data (in our case, it is one dimension) [14]; 2) the percentile of PDs [31], [32]; and 3) the median of PDs [19], [42]. Figs. 5 and 6 shows the influence of different kernel width on the proposed method in terms of classification accuracy, again with four classifiers for the Pavia University and KSC data sets. All the heuristic approaches provided a comparatively similar accuracy when a more number of features are selected, but the 5th percentile of PDs approach outperforms other approaches when a less number features are considered for both the images.

H. Impact of Variations in the Training Samples on the Selected Spectral Bands

In section, we analyzed the impact of variations in the training samples on the selected features. In order to introduce variations in the training samples, we randomly choose 100 samples per class from the ground-truth reference (see the right column in Tables I and II) for the Pavia University and Center data sets and 50 samples per class (right column in Table III) for the KSC data set. The feature selection experiments are repeated ten times to produce ten subsets of selected features to analyze the feature stability of feature selection methods. To quantify the feature selection stability, we used two types of measures: 1) feature index measure: Jaccard index [43] and Kuncheva's stability index (KSI) [44] and 2) feature value measure: information stability (IS) [45]. The first one measures the amount of overlap between the feature index values on different subsets, and additionally, KSI corrects overlapping due to the chance, while the latter measures IS over different subsets of features. For additional and implementation details, the reader is referred to [43]–[45].

Table VI reports the feature stability measures for ten selected features computed over ten subsets of selected features. Higher values in Table VI indicate more stability in selected feature subsets. In both feature index and feature value measures, the original HSIC methods provide better stability in selected features over the variations in training samples.

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

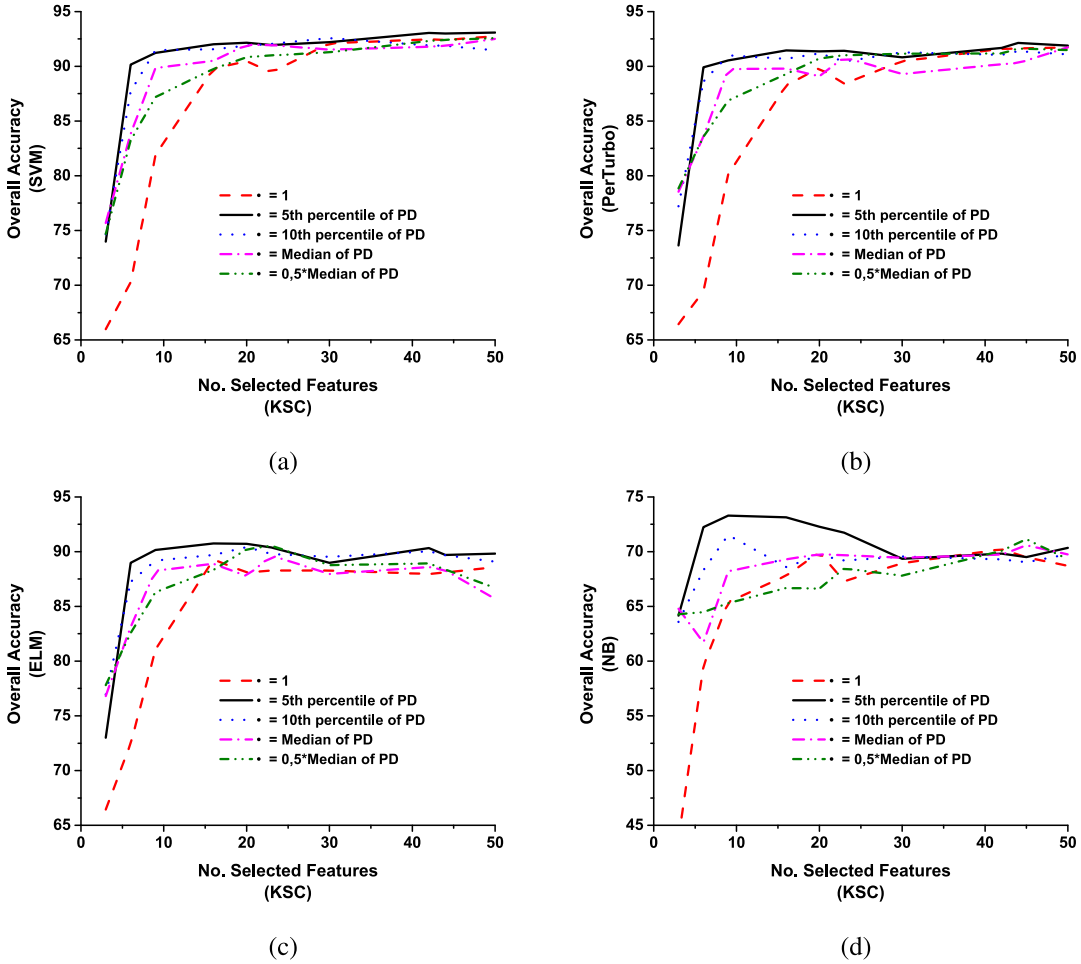

Fig. 6. Classification accuracy of HSIC-SK LASSO method with different heuristic Gaussian bandwidths on the KSC data set. (a) SVM classifier. (b) PerTurbo classifier. (c) ELM classifier. (d) NBC.

TABLE VI

FEATURE STABILITY MEASURES COMPUTED OVER TEN SUBSETS OF SELECTED FEATURES FOR DIFFERENT FEATURE SELECTION METHODS. IN EACH SUBSET, THE TOP TEN SELECTED FEATURES ARE CONSIDERED

Feature selection Methods	Pavia University			Pavia Centre			KSC		
	JI	KSI	IS	JI	KSI	IS	JI	KSI	IS
HSIC-SK LASSO	0.2691	0.3232	0.9228	0.3012	0.397	0.9075	0.2184	0.2813	0.7014
FOHSIC	0.4933	0.6111	0.9551	0.6217	0.7367	0.9534	0.9269	0.9517	0.9875
BAHSIC	0.5733	0.6899	0.9634	0.6711	0.776	0.9598	0.8886	0.9255	0.9831
CEM-BCC	0.0602	-0.0017	0.8826	0.0673	0.018	0.8252	0.1249	0.1017	0.724
CEM-BCM	0.0563	0.0032	0.887	0.0553	-0.0042	0.8241	0.0971	0.0648	0.6949

However, the IS measures reveals that though there are variations in selected features by our method, the information content on subset of features is more or less the same. In other words, instead of selecting the same features, our method selects the nearby features that are valid for hyperspectral data as nearby spectral frequencies have similar spectral behavior. Furthermore, on the KSC data set, the original HSIC method repeats more or less the same set of features over variations in the training set. However, it selected very correlated features (for instance, to quantify the correlation, we measured the average correlation coefficient of selected feature values for the original HSIC method and the measure is 0.49, whereas that for our method is 0.24). The less stability nature of the LASSO model is not an unexpected behavior, as LASSO tends to produce unstable features to the variations in training set. As a future work, we would like to develop upon this limitation on the LASSO model to select more stable features, for instance, by considering elastic netlike strategies.

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

V. Conclusion

This paper developed a new feature selection method for selecting the class separable features for hyperspectral data classification. In the first stage, we proposed a new class separability measure based on SK and HSIC by aligning the empirical kernel maps in the RKHS, so that we can benefit from the nonlinearities present in the data, kernel matrices of different size, and the low number of samples compared with conventional class separability measures. In the second stage, the proposed class separability measure is used as an objective function and the feature selection problem is modeled as the LASSO optimization problem. The nonzero coefficients resulting from our LASSO formulation corresponds to the selected features for hyperspectral image classification. This framework alleviates the computationally intensive subset search strategy involved in selecting the class discriminant features. The extensive experiments conducted with three hyperspectral data sets with different settings shows that our proposed HSIC-SK-LASSO method approximate full-band classification accuracy in fewer spectral channels and increases the classification accuracy over the competitors. In our future work, we would like to overcome the limitation of the proposed method to select more stable features and also to explore the possibility of selecting domain invariant features.

Acknowledgment

The authors would like to thank Prof. P. Gamba for providing the ROSIS hyperspectral images and the anonymous reviewers for the constructive criticism and suggestions.

References

[1] G. Camps-Valls, D. Tuia, L. Bruzzone, and J. A. Benediktsson, "Advances in hyperspectral image classification: Earth monitoring with statistical learning methods," *IEEE Signal Process. Mag.*, vol. 31, no. 1, pp. 45–54, Jan. 2014.

[2] B. B. Damodaran, R. R. Nidamanuri, and Y. Tarabalka, "Dynamic ensemble selection approach for hyperspectral image classification with joint spectral and spatial information," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 8, no. 6, pp. 2405–2417, Jun. 2015.

[3] B. B. Damodaran and R. R. Nidamanuri, "Dynamic linear classifier system for hyperspectral image classification for land cover mapping," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 6, pp. 2080–2093, Jun. 2014.

[4] L. O. Jimenez and D. A. Landgrebe, "Supervised classification in high-dimensional space: Geometrical, statistical, and asymptotical properties of multivariate data," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 28, no. 1, pp. 39–54, Feb. 1998.

[5] M. Pal and G. M. Foody, "Feature selection for classification of hyperspectral data by SVM," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 5, pp. 2297–2307, May 2010.

[6] J. Zabalza *et al.*, "Novel folded-PCA for improved feature extraction and data reduction with hyperspectral imaging and SAR in remote sensing," *ISPRS J. Photogram. Remote Sens.*, vol. 93, pp. 112–122, Jul. 2014.

[7] W. Li, S. Prasad, J. E. Fowler, and L. M. Bruce, "Locality-preserving dimensionality reduction and classification for hyperspectral image analysis," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 4, pp. 1185–1198, Apr. 2012.

[8] L. Bruzzone, F. Roli, and S. B. Serpico, "An extension of the Jeffreys–Matusita distance to multiclass cases for feature selection," *IEEE Trans. Geosci. Remote Sens.*, vol. 33, no. 6, pp. 1318–1321, Nov. 1995.

[9] M. Fauvel, C. Dechesne, A. Zullo, and F. Ferraty, "Fast forward feature selection of hyperspectral images for classification with Gaussian mixture models," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 8, no. 6, pp. 2824–2831, Jun. 2015.

[10] S. Jia, G. Tang, J. Zhu, and Q. Li, "A novel ranking-based clustering approach for hyperspectral band selection," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 1, pp. 1–15, Jan. 2015.

[11] W. Sun, L. Zhang, B. Du, W. Li, and Y. M. Lai, "Band selection using improved sparse subspace clustering for hyperspectral imagery classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 8, no. 6, pp. 2784–2797, Jun. 2015.

[12] M. Gong, M. Zhang, and Y. Yuan, "Unsupervised band selection based on evolutionary multiobjective optimization for hyperspectral images," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 1, pp. 544–557, Jan. 2016.

[13] B. Guo, R. Damper, S. R. Gunn, and J. Nelson, "A fast separability-based feature-selection method for high-dimensional remotely sensed image classification," *Pattern Recognit.*, vol. 41, no. 5, pp. 1653–1662, May 2008.

[14] L. Song, A. Smola, A. Gretton, J. Bedo, and K. Borgwardt, "Feature selection via dependence maximization," *J. Mach. Learn. Res.*, vol. 13, no. 1, pp. 1393–1434, Jan. 2012.

[15] C. Wang, M. Gong, M. Zhang, and Y. Chan, "Unsupervised hyperspectral image band selection via column subset selection," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 7, pp. 1411–1415, Jul. 2015.

[16] C.-I. Chang and S. Wang, "Constrained band selection for hyperspectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 44, no. 6, pp. 1575–1585, Jun. 2006.

[17] N. Keshava, "Distance metrics and band selection in hyperspectral processing with applications to material identification and spectral libraries," *IEEE Trans. Geosci. Remote Sens.*, vol. 42, no. 7, pp. 1552–1565, Jul. 2004.

[18] M. Cui, S. Prasad, M. Mahrooghy, L. M. Bruce, and J. Aanstoos, "Genetic algorithms and linear discriminant analysis based dimensionality reduction for remotely sensed image analysis," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2011, pp. 2373–2376.

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

[20] M. Dabboor, S. Howell, M. Shokr, and J. Yackel, "The Jeffries–Matusita distance for the case of complex Wishart distribution as a separability criterion for fully polarimetric SAR data," *Int. J. Remote Sens.*, vol. 35, no. 19, pp. 6859–6873, Oct. 2014.

[21] A. Ifarraguerri and M. W. Prairie, "Visual method for spectral band selection," *IEEE Geosci. Remote Sens. Lett.*, vol. 1, no. 2, pp. 101–106, Apr. 2004.

[22] T. Kavzoglu and P. M. Mather, "The role of feature selection in artificial neural network applications," *Int. J. Remote Sens.*, vol. 23, no. 15, pp. 2919–2937, Jan. 2002.

[23] J. R. Jensen, *Introductory Digital Image Processing: A Remote Sensing Perspective*. Upper Saddle River, NJ, USA: Prentice-Hall, 1996.

[24] L. Chapel, T. Burger, N. Courty, and S. Lefevre, "PerTurbo manifold learning algorithm for weakly labeled hyperspectral image classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 4, pp. 1070–1078, Apr. 2014.

[25] C. Persello and L. Bruzzone, "Kernel-based domain-invariant feature selection in hyperspectral images for transfer learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 5, pp. 2615–2626, May 2016.

[26] M. Yamada, W. Jitkrittum, L. Sigal, E. P. Xing, and M. Sugiyama, "High-dimensional feature selection by feature-wise kernelized lasso," *Neural Comput.*, vol. 26, no. 1, pp. 185–207, Dec. 2014.

[27] B. Schölkopf and A. J. Smola, *Learning With Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Cambridge, MA, USA: MIT Press, 2001.

[28] Z. Kai, W. Z. Vincent, W. Qiaojun, T. K. James, Y. Qiang, and M. Ivan, "Covariate shift in Hilbert space: A solution via surrogate kernels," in *Proc. 30th Int. Conf. Mach. Learn.*, Atlanta, GA, USA, 2013, pp. 1–8.

[29] A. Gretton, O. Bousquet, A. Smola, and B. Schölkopf, "Measuring statistical dependence with Hilbert–Schmidt norms," in *Algorithmic Learning Theory* (Lecture Notes in Computer Science), S. Jain, H. Simon, and E. Tomita, Eds. Berlin, Germany: Springer, 2005, pp. 63–77.

[30] A. Smola, A. Gretton, L. Song, and B. Schölkopf, "A Hilbert space embedding for distributions," in *Algorithmic Learning Theory* (Lecture Notes in Computer Science), M. Hutter, R. A. Servedio, and E. Takimoto, Eds. Berlin, Germany: Springer, 2007, pp. 13–31.

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

[31] P. K. Mallapragada, R. Jin, and A. Jain, "Non-parametric mixture models for clustering," in *Structural, Syntactic, and Statistical Pattern Recognition* (Lecture Notes in Computer Science), E. R. Hancock, R. C. Wilson, T. Windeatt, I. Ulusoy, and F. Escolano, Eds. Berlin, Germany: Springer, 2010, pp. 334–343.

[32] M. Lin, S. Weng, and C. Zhang, "On the sample complexity of random Fourier features for online learning: How many random fourier features do we need?" *ACM Trans. Knowl. Discovery Data*, vol. 8, no. 3, pp. 13-1–13-19, Jun. 2014.

[33] G. Camps-Valls and L. Bruzzone, "Kernel-based methods for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 6, pp. 1351–1362, Jun. 2005.

[34] G.-B. Huang, H. Zhou, X. Ding, and R. Zhang, "Extreme learning machine for regression and multiclass classification," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 42, no. 2, pp. 513–529, Apr. 2012.

[35] Y. Zhou, J. Peng, and C. L. P. Chen, "Extreme learning machine with composite kernels for hyperspectral image classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 8, no. 6, pp. 2351–2360, Jun. 2015.

[36] A. Samat, P. Du, S. Liu, J. Li, and L. Cheng, "(ELMs)-L-2: Ensemble extreme learning machines for hyperspectral image classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 4, pp. 1060–1069, Apr. 2014.

[37] N. Friedman, D. Geiger, and M. Goldszmidt, "Bayesian network classifiers," *Mach. Learn.*, vol. 29, nos. 2–3, pp. 131–163, 1997.

[38] G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew, "Extreme learning machine: Theory and applications," *Neurocomputing*, vol. 70, nos. 1–3, pp. 489–501, 2006.

[39] M. D. Mura, J. A. Benediktsson, B. Waske, and L. Bruzzone, "Extended profiles with morphological attribute filters for the analysis of hyperspectral data," *Int. J. Remote Sens.*, vol. 31, no. 22, pp. 5975–5991, Dec. 2010.

[40] M. D. Mura, J. A. Benediktsson, B. Waske, and L. Bruzzone, "Morphological attribute profiles for the analysis of very high resolution images," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 10, pp. 3747–3762, Oct. 2010.

[41] M. D. Mura, A. Villa, J. A. Benediktsson, J. Chanussot, and L. Bruzzone, "Classification of hyperspectral images by using extended morphological attribute profiles and independent component analysis," *IEEE Geosci. Remote Sens. Lett.*, vol. 8, no. 3, pp. 542–546, May 2011.

[42] Z. Lu *et al.* (Nov. 14, 2014). "How to scale up kernel methods to be as good as deep neural nets." [Online]. Available: https://arxiv.org/abs/1411.4000

[43] R. Real and J. M. Vargas, "The probabilistic basis of jaccard's index of similarity," *Syst. Biol.*, vol. 45, no. 3, pp. 380–385, 1996.

[44] L. I. Kuncheva, "A stability index for feature selection," in *Artificial Intelligence and Applications*. Anaheim, CA, USA: ACTA Press, 2007, pp. 390–395.

[45] G. Gulgezen, Z. Cataltepe, and L. Yu, "Stable and accurate feature selection," in *Machine Learning and Knowledge Discovery in Databases* (Lecture Notes in Computer Science), W. Buntine, M. Grobelnik, D. Mladenić, and J. Shawe-Taylor, Eds. Berlin, Germany: Springer, 2009, pp. 455–468.

Bharath Bhushan Damodaran (M'16) received the M.Sc. degree in mathematics from Bharathiar University, Coimbatore, India, in 2008, the M.Tech. degree in remote sensing and wireless sensor networks from Amrita Vishwa Vidyapeetham, Coimbatore, in 2010, and the Ph.D. degree in earth and space sciences from the Indian Institute of Space Science and Technology, Trivandrum, India, in 2015.

He is currently a Post-Doctoral Researcher with the OBELIX Team, IRISA, University of Bretagne-Sud, Vannes, France. His research interests include feature selection, large scale kernel learning, multiple classifier system, hyperspectral/multispectral image analysis, machine learning, and image processing.

Dr. Damodaran has been awarded the Prestige and Marie Curie Post-Doctoral Fellowship by the campus France.

Nicolas Courty (M'16) received the habilitation degree from the University of Bretagne-Sud, Vannes, France, in 2013.

He has been an Associate Professor with the University of Bretagne-Sud since 2004. In 2012, he was an Invited Professor with the Senior Chinese Academy of Science Fellowship at the Institute of Automation, Beijing, China. In 2014, he spent two months at LASIG (EPFL), Lausanne, as an Invited Professor from EPFL. His research interests include data analysis/simulation schemes, machine learning, and visualization problems, with applications in computer vision, remote sensing, and computer graphics.

Dr. Courty was one of the recipient of the U.V. Helava ISPRS Award for the 2012–2015 period.

Sébastien Lefèvre received the M.Sc. and Eng. degrees in computer engineering from the University of Technology of Compiegne, Compiegne, France, in 1999, the Ph.D. degree in computer science from the University of Tours, Tours, France, in 2002, and the French HDR (Habilitation to Supervise Doctoral Studies) degree in computer science from the University of Strasbourg, Strasbourg, France, in 2009.

From 2003 to 2010, he was an Associate Professor with the Department of Computer Sciences and the Image Sciences, Computer Sciences and Remote Sensing Laboratory (LSIIT), University of Strasbourg–CNRS. From 2009 to 2010, he was an INRIA invited scientist within the TEXMEX team at the Institute for Research in Computer Science and Random Systems (IRISA)/INRIA Rennes, Rennes, France. In 2010, he joined the University of Bretagne-Sud, Vannes, France, as a Full Professor with the Department of Computer Science, Institute of Technology of Vannes, Vannes, and IRISA. Within IRISA, he is leading the OBELIX team dedicated to image analysis and machine learning for remote sensing and earth observation. He has co-authored more than 100 papers in image analysis and pattern recognition. His research interests include multivariate mathematical morphology, hierarchical models, and machine learning applied to remote sensing of environment.