EVALUATION OF DIFFERENT REGULARIZATION METHODS FOR THE EXTREME LEARNING MACHINE APPLIED TO HYPERSPECTRAL IMAGES

Juan M. Haut, *Student Member, IEEE,* Yi Liu, *Member, IEEE,* Mercedes E. Paoletti, *Student Member, IEEE,* Xiong Xu, *Member, IEEE,* Javier Plaza, *Senior Member, IEEE,* and Antonio Plaza, *Fellow, IEEE*

***Abstract*— During recent years, many regularization techniques have been proposed to deal with ill-posed problems related to hyperspectral image classification, in which the limited number of training samples contrasts with the very high spectral dimensionality. However, the intrinsic structure of a hyperspectral image often depends on the specific scene and spectrometer, although regularizers like Ridge, LASSO, etc, have been widely used in practical applications. Instead of imposing these regularizers to the probabilistic output of a classifier, this work evaluates the use of extreme learning machines (ELM) with output weights of a single-hidden layer feed-forward neural network (SLFN) regularized with Ridge and LASSO priors, respectively. Experimental results with several real hyperspectral images are conducted to compare the performance and adaptation of these two regularizers with the the original ELM in classification scenarios.**

I. INTRODUCTION

The rapidly developing field of hyperspectral remote sensing has reinforced the application of this technology in a wide range of fields [1]. In hyperspectral classification, extreme learning machines [2], as an instance of artificial neural networks, have been one of the most widely used classifiers due to their simplified network structure as well as their high computing performance [3], [4]. One of the prominent advantages of this classifier is its fast estimation of the output weights in the learning process against the singularity of the coefficient matrix.

The issue of singularity, or ill-poseness, of the coefficient matrix has led to great challenges in hyperspectral image classification and regression scenarios, due to the often limited availability of training samples and very high spectral dimensionality of hyperspectral data. In order to deal with this problem, various techniques have been developed in recent years. Intuitively, semi-supervised learning and active learning techniques have been widely explored to automatically search and increase the sample volume for the training set to seek its balance with the high spectral dimensionality [5], [6]. Dimensionality reduction turns out to be another popular method to address the curse of dimensionality via selecting the most relevant bands or features for training a classifier while discarding the rest [7]. On the other hand, spectral partitioning, which pursues dimensionality reduction by reassigning the spectral bands or features into multiple subgroups, where each subgroup consists of much less bands/features, provides an extra mechanism to utilize the original information, especially retaining the physical information to meet specific practical requirements [8]. In addition, spectral-spatial classification techniques [9], [10] has also been proven to be highly effective for improving the classification performance by including the information from the spatial domain of the image.

Besides these techniques, regularized methods are certainly among the most successful ones. These methods impose prior regularizing terms to the learning model, thus being able to address ill-posed problems [11]. Meanwhile, these techniques generally assume certain intrinsic characteristics in the dataset, which often promote the generalization capability of the learning model. Such regularized methods can be straightforwardly used in combination with the rest of techniques.

Among the regularizing priors, Ridge [12] and least absolute shrinkage and selection operator (LASSO) [13] are among the mostly used ones [14], [11]. They address the singularity issues in regression/learning problems and, meanwhile, they render information via Ridge or LASSO regularizing to the coefficient of a learning model. The success of these regularization methods depends, not only on specific datasets, but also on the specific classifiers. This is particularly true for hyperspectral images, whose data volume is usually huge and with complex structure. Taking the example of extreme learning, the output weights are the most relevant variable in the learning process. As an instance of single-hidden layer feed-forward neural network (SLFN), the famous ”black box” effect of the single hidden layer makes the output weight distinct from the coefficients of other classifiers [2]. Motivated by the aforementioned issues, this work developed two new extreme learning machines models based on Ridge and LASSO regularizers in order to explore their capacity to provide regularized estimations of the output weights in hyperspectral classification scenarios. We analyze and compare in detail the performance as well as the adaptation of both, the original ELM and our regularized methods, in the context of hyperspectral image classification

M. E. Paoletti, J. M. Haut, J. Plaza and A. Plaza are with the Hyperspectral Computing Laboratory, Department of Technology of Computers and Communications, Escuela Politécnica, University of Extremadura, PC-10003 Cáceres, Spain.(e-mail: mpaolett@unex.es; juanmariohaut@unex.es; jplaza@unex.es; aplaza@unex.es). Y. Liu is with the Faculty of Information Technology and Electrical Engineering, Norwegian University of Science and Technology, Trondheim 7034, Norway. (e-mail:y.liu@ntnu.no). X. Xu is the College of Surveying and Geo-Informatics, Tongji University, 1239 Siping Road, Shanghai 200092, PR China. (e-mail:xvxiong@tongji.edu.cn).

978-1-5386-7150-4/18/$31.00 ©2018 IEEE

problems.

The remainder of this work is organized as following. Section II introduces the developed regularized methods under the ELM framework. Experimental results are further displayed and discussed in Section III. Finally, Section IV concludes this work and indicates our future research lines.

II. Methodology

A. Extreme Learning Machine (ELM)

Based on artificial neural networks (ANNs), the ELM implements a single-hidden layer feed-forward neural network (SLFN) with topology $d - L - g$, being d the input layer nodes, L the hidden layer nodes and g the output layer nodes (see Fig. 1), whose goal is to find the output weights $\boldsymbol{\beta}$ that best approximate the output of the hidden layer to the desired network output [2]. Given a dataset of m paired data

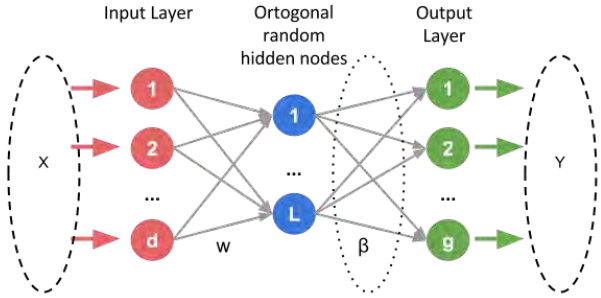


Fig. 1. Extreme learning machine scheme

$\{\mathbf{x}^{(i)}, \mathbf{y}^{(i)}\}_{i=1}^{m}$, where $\mathbf{x}^{(i)} \in \mathbb{R}^d = [x_1^{(i)}, x_2^{(i)}, ..., x_d^{(i)}]$ is a feature vector and $\mathbf{y}^{(i)} \in \mathbb{R}^g = [y_1^{(i)}, y_2^{(i)}, ..., y_g^{(i)}]$ is one of the k possible targets, categories or labels, the desired ELM output can be calculated as $f(\mathbf{x}) = \mathbf{H}(\mathbf{X}, \mathbf{W}, \mathbf{B})\boldsymbol{\beta}$, where $\mathbf{X} = \{\mathbf{x}^{(i)}\}_{i=1}^{m}$ and $f(\mathbf{x}) \simeq \mathbf{Y} = \{\mathbf{y}^{(i)}\}_{i=1}^{m}$ are the input data and the output targets, $\boldsymbol{\beta}$ is the output weights that connect the hidden nodes with the network output nodes, $\mathbf{W}$ and $\mathbf{B}$ are the random-selected weights and biases that connect the input layer nodes with the hidden layer ones and have been generated based on a continuous sampling distribution probability, and $\mathbf{H}$ is the output matrix calculated by the hidden layer as $\mathbf{H} = f(\mathbf{XW} + \mathbf{B})$, where $f(\cdot)$ is an activation function, such as sigmoid, tanh or ReLU, among others. In general, the learning problem of an ELM can be represented by,

$$\mathbf{H}\boldsymbol{\beta} = \mathbf{Y}, \tag{1}$$

whose solution, under the least square error standard, can usually be approximated via

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \|\mathbf{H}\boldsymbol{\beta} - \mathbf{Y}\|_F^2, \tag{2}$$

where $\|\cdot\|_F$ is the Frobenius norm. The imbalance between limited training samples and very high spatial dimensionality of HSI usually lead to the ill-poseness of solving the problem in (1). Therefore, ELM calculates $\boldsymbol{\beta}$ that most minimize the cost while exhibiting the minor norm solving the equation system 3:

$$\widehat{\boldsymbol{\beta}}_{ELM} = \mathbf{H}^{\dagger}\mathbf{Y} = \left(\frac{I}{C} + \mathbf{H}^T\mathbf{H}\right)^{-1}\mathbf{H}^T\mathbf{Y}, \tag{3}$$

where $\mathbf{H}^{\dagger}$ is the Moore-Penrose generalized inverse matrix of $\mathbf{H}$ and C is the corresponding regularization term [3] to increase the robustness and the generalization capability of ELM. Also, equation 3 can be modified in order to add a kernel [4] with the aim of making ELM independent of random weights and bias, such as $\boldsymbol{\beta} = (\frac{I}{C} + \mathbf{K})^{-1}\mathbf{T}$. The target output can be expressed as:

$$f(\mathbf{x}) = \mathbf{H}\boldsymbol{\beta} = \mathbf{h}(\mathbf{X})\left(\frac{I}{C} + \mathbf{H}^T\mathbf{H}\right)^{-1}\mathbf{H}^T\mathbf{Y}, \tag{4}$$

with $\mathbf{h}(\mathbf{X})$ being a feature mapped from the input data. The kernel matrix $\mathbf{K}$ can be expressed as $\mathbf{K}_{ELM} = \mathbf{H}\mathbf{H}^T$: $\mathbf{K}_{i,j} = k(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) = \mathbf{h}(\mathbf{x}^{(i)}) \times \mathbf{h}(\mathbf{x}^{(j)})$, where $k(\mathbf{x}^{(i)}, \mathbf{x}^{(j)})$ is the kernel function, being $\mathbf{x}^{(i)}$ and $\mathbf{x}^{(j)}$ two training samples of dataset $\mathbf{X}$. Leading this equation to the output function expressed in 4 we obtain:

$$f(\mathbf{x^{(i)}}) = \begin{bmatrix} k(\mathbf{x}^{(i)}, \mathbf{x}^{(1)}) \\ k(\mathbf{x}^{(i)}, \mathbf{x}^{(2)}) \\ ... \\ k(\mathbf{x}^{(i)}, \mathbf{x}^{(m)}) \end{bmatrix}^T \left(\frac{I}{C} + \mathbf{K}_{ELM}\right)^{-1}\mathbf{Y}^{(i)} \tag{5}$$

The kernel function $k(\mathbf{x}^{(i)}, \mathbf{x}^{(j)})$ can be implemented as radial basis function (RBF):

$$K(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) = exp(-\frac{\| \mathbf{x}^{(i)}, \mathbf{x}^{(j)} \|^2}{2\sigma^2}) \tag{6}$$

B. Solution via RIDGE regularization

The RIDGE regularization is one of the most commonly used linear regression/classification algorithm, which is utilized here to improve the conditioning of the problem (2) and enforce smoothness if the underlying vector is mostly continuous,

$$\widehat{\boldsymbol{\beta}}_{Ridge} = \arg\min_{\boldsymbol{\beta}} \|\mathbf{H}\boldsymbol{\beta} - \mathbf{Y}\|_F^2 + \lambda_{Ridge}\|\boldsymbol{\beta}\|_F^2, \tag{7}$$

which eases the singularity issue and approximates the solution via Tikhonov regularizing in the following,

$$\widehat{\boldsymbol{\beta}}_{Ridge} = \left(\mathbf{H}^T\mathbf{H} + \lambda_{Ridge}\mathbf{I}\right)^{-1}\mathbf{H}^T\mathbf{Y}, \tag{8}$$

where the Tikhonov regularization matrix is defined as a multiple of the identity matrix that prevents overfitting and underfitting and λ_{Ridge} is a user-defined penalizing term. In our case, Cholesky solver has been further used in order to obtain a closed-form solution [15].

C. Solution via LASSO regularization

The LASSO regularizer is also considered in this work to approximate the solution of the linear regression algorithm of (1), in order to improve the prediction accuracy and interpretability of regression models by altering the model fitting

process to select only a subset of the provided covariates to be used in the final model rather than using all of them [13],

$$\widehat{\boldsymbol{\beta}}_{LASSO} = \arg\min_{\boldsymbol{\beta}} \frac{1}{2}\|\mathbf{H}\boldsymbol{\beta} - \mathbf{Y}\|_F^2 + \lambda_{LASSO}\|\boldsymbol{\beta}\|_1, \quad (9)$$

where $\|\cdot\|_1$ is the 1-norm and λ_{LASSO} is a tuning parameter that controls the amount of regularization. In this work, we solve the object function (9) via the sparse unmixing by variable splitting and augmented Lagrangian (SUNSAL) toolbox without imposing nonnegative and sum-to-one constraints [14].

III. Experimental Results

A. Experimental environment and datasets

Our experiments have been conducted on a hardware environment composed by a 6th Generation Intel® Core™i7-6700K processor with 8M of Cache and up to 4.20GHz (4 cores/8 way multitask processing), 40GB of DDR4 RAM with a serial speed of 2400MHz, a GPU NVIDIA GeForce GTX 1080 with 8GB GDDR5X of video memory and 10Gbps of memory frequency, a Toshiba DT01ACA HDD with 7200RPM and 2TB of capacity, and an ASUS Z170 pro-gaming motherboard. On the other hand, the software environment is composed by Ubuntu 16.04.4 x64 as operating system. Also, the proposal has been implemented in python 2.7, with Numpy as mathematics library.

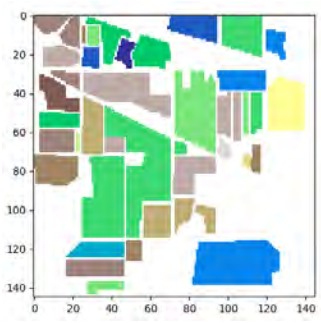
Our experiments have been carried out using two different and well-known hyperspectral datasets, described below:

1) AVIRIS Indian Pines: this scene (see Fig. 2) covers an agricultural site in Northwestern Indiana, and was collected by the Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) sensor [16] in 1992. The data set of is of size $145 \times 145 \times 220$, with spatial resolution of 20 m per pixel and a spectral range from 0.2 to 2.4 microns. Before classification, 20 spectral bands (i.e., 104th-108th, 150th-163rd, and 220th) are discarded due to low SNR. This image contains 16 land-cover classes.
2) AVIRIS Salinas: this scene (see Fig. 3) was also collected by the AVIRIS sensor over the Salinas Valley, California. The data set is of size $512 \times 217 \times 224$, and it has spatial resolution of 3.7 m per pixel with 16 land-cover classes. Before classification, 20 water absorption bands were removed (i.e., 108th-112th, 154th-167th, 224th).

Note that, in this work, we explore the performance and adaptivity of the considered Ridge and LASSO regularizers in the task of learning the output weights of the ELM classifier. Based on the considered hyperspectral images, we seek to analyze whether output weights fit the underlying assumptions of these regularization models.

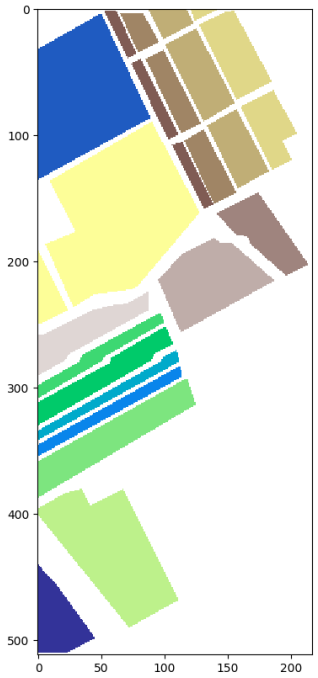
B. Discussion of results

Two experiments have been carried out to test the performance of each method. In the first one, a percentage of samples per class has been randomly selected as training data, using 1%, 2%, 3%, 5%, 10% and 15%. On the other hand in second experiment a fixed number of samples per class has been selected. In this case, we used 1, 3, 5, 10, 20, 50 and 100 samples per class. Also, parameter λ has been selected via grid-search in the range from 10^{-7} to 10^{7}.



Color	Land cover type	Samples
	Background	10659
	Alfalfa	54
	Corn-notill	1434
	Corn-min	834
	Corn	234
	Grass/Pasture	497
	Grass/Trees	747
	Grass/pasture-mowed	26
	Hay-windrowed	489
	Oats	20
	Soybeans-notill	968
	Soybeans-min	2468
	Soybean-clean	614
	Wheat	212
	Woods	1294
	Bldg-Grass-Tree-Drives	380
	Stone-steel towers	95
	Total samples	21025

Fig. 2. Ground-truth of the Indian Pines $145 \times 145 \times 202$ hyperspectral scene.



Color	Land-cover type	Samples
	Background	56975
	Brocoli-green-weeds-1	2009
	Brocoli-green-weeds-2	3726
	Fallow	1976
	Fallow-rough-plow	1394
	Fallow-smooth	2678
	Stubble	3959
	Celery	3579
	Grapes-untrained	11271
	Soil-vinyard-develop	6203
	Corn-senesced-green-weeds	3278
	Lettuce-romaine-4wk	1068
	Lettuce-romaine-5wk	1927
	Lettuce-romaine-6wk	916
	Lettuce-romaine-7wk	1070
	Vinyard-untrained	7268
	Vinyard-vertical-trellis	1807
	Total samples	111104

Fig. 3. Ground-truth of the Salinas $512 \times 217 \times 204$ hyperspectral scene.

Table I shows the obtained overall accuracies (OAs) for each experiment over the Indian Pines hyperspectral dataset, which have been repeated 5 times in order to extract the standard deviations. After imposing the regularizers of Ridge and LASSO, respectively, a decrease of performances in terms of overall accuracies (OAs) can be observed. The original ELM classifier generally acquires the highest OAs, although the regularized methods outperforms the original ELM classifier in some cases. Similar observations can also be obtained when using the AVIRIS Salinas dataset (see TableII). As expected, the classification performances of

different methods show that, although the coefficient smoothness and sparsity of a classifier model have been shown to be helpful to enhance the probabilistic output, a positive enhancement to the output weights β has not been observed in Ridge and LASSO regularizations. This indicates that the output weights do not contain smoothness and sparsity in the learning process of ELM with the considered hyperspectral datasets.

Pixel trainning	Original	Ridge	Lasso
1%	70.06 (0.52)	64.11 (0.90)	66.54 (2.01)
2%	74.61 (1.13)	69.73 (1.08)	70.86 (1.40)
3%	79.59 (0.35)	71.60 (1.27)	73.72 (2.45)
5%	81.64 (1.63)	73.93 (0.83)	77.96 (1.26)
10%	87.00 (0.36)	77.03 (0.73)	81.81 (0.38)
15%	88.67 (0.31)	78.54 (1.15)	83.67 (0.81)
1/class	41.42 (5.18)	45.09 (3.64)	42.71 (4.20)
3/class	54.89 (1.04)	45.03 (3.68)	49.42 (2.40)
5/class	62.91 (0.86)	55.75 (4.98)	52.87 (3.94)
10/class	65.29 (2.22)	65.41 (2.31)	62.93 (1.24)
20/class	74.18 (1.17)	70.49 (2.06)	69.17 (1.24)
50/class	80.74 (1.04)	74.77 (0.92)	75.14 (1.28)
100/class	84.74 (0.41)	76.68 (0.57)	78.51 (0.65)

TABLE I

OVERALL ACCURACIES (AND STANDARD DEVIATION) FOR K-ELM USING DIFFERENT REGULARIZATION FOR INDIAN PINES SCENE.

On the other hand, Table II presents the OAs reached by each method over the Salinas valley hyperspectral dataset.

Pixel trainning	Original	Ridge	Lasso
1%	90.80 (0.40)	86.41 (0.35)	88.83 (0.23)
2%	91.65 (0.12)	87.85 (0.34)	90.36 (0.14)
3%	92.19 (0.06)	88.39 (0.34)	90.61 (0.19)
5%	93.14 (0.04)	88.80 (0.31)	91.34 (0.15)
1/class	78.00 (1.81)	79.57 (0.46)	63.78 (3.92)
3/class	80.22 (2.60)	78.22 (2.49)	78.86 (3.02)
5/class	87.01 (0.35)	82.69 (0.98)	79.90 (1.84)
10/class	87.06 (0.33)	83.80 (0.65)	82.81 (1.74)
20/class	89.75 (0.36)	83.54 (2.14)	84.72 (0.72)
50/class	90.54 (0.19)	87.19 (0.55)	87.90 (0.52)
100/class	91.05 (0.30)	87.68 (0.40)	87.38 (0.27)

TABLE II

OVERALL ACCURACIES (AND STANDARD DEVIATION) FOR K-ELM USING DIFFERENT REGULARIZATION FOR SALINAS SCENE.

IV. CONCLUSIONS

In this work, we developed two ELM instances with the Ridge and LASSO regularizers in order to explore the characteristics of the output weights β of the single hidden layer of the neural network in hyperspectral classification scenarios. With the two considered hyperspectral images, smoothness and sparsity have not been significantly observed in the output weights as other coefficients of the learning process. In future work, our research line will focus on exploring in more details the potential characteristics of the output weights of the single hidden layer of ELMs, as well as to develop more thorough comparisons with different types of classifiers when using remotely sensed hyperspectral imagery.

V. ACKNOWLEDGEMENT

This work has been supported by Ministerio de Educación (Resolución de 26 de diciembre de 2014 y de 19 de noviembre de 2015, de la Secretaría de Estado de Educación, Formación Profesional y Universidades, por la que se convocan ayudas para la formación de profesorado universitario, de los subprogramas de Formación y de Movilidad incluidos en el Programa Estatal de Promoción del Talento y su Empleabilidad, en el marco del Plan Estatal de Investigación Científica y Técnica y de Innovación 2013-2016. This work has also been supported by Junta de Extremadura (decreto 297/2014, ayudas para la realización de actividades de investigación y desarrollo tecnológico, de divulgación y de transferencia de conocimiento por los Grupos de Investigación de Extremadura, Ref. GR15005).

REFERENCES

[1] A. Plaza et al., "Recent advances in techniques for hyperspectral image processing," *Remote Sens. Environ.*, vol. 113, pp. S110–S122, 2009.

[2] G-B. Huang et al., "Extreme Learning Machine: A New Learning Scheme of Feedforward Neural Networks," in *2004 IEEE International Joint Conference on Neural Networks (IEEE Cat. No.04CH37541)*, 2004, vol. 2, pp. 985–990.

[3] G-B. Huang, "Extreme Learning Machine for Regression and Multiclass Classification," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 42, no. 2, pp. 513–529, 2012.

[4] G-B. Huang, "An Insight into Extreme Learning Machines: Random Neurons, Random Features and Kernels," *Cognitive Computation*, vol. 6, no. 3, pp. 376–390, 2014.

[5] I. Dópido and et. al, "Semisupervised self-learning for hyperspectral image classification," *IEEE transactions on geoscience and remote sensing*, vol. 51, no. 7, pp. 4032–4044, 2013.

[6] D. Tuia et al., "Active learning methods for remote sensing image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 47, no. 7, pp. 2218–2232, 2009.

[7] Q. Du, J. Fowler, and B. Ma, "Random-projection-based dimensionality reduction and decision fusion for hyperspectral target detection," in *Geoscience and Remote Sensing Symposium (IGARSS), 2011 IEEE International*. IEEE, 2011, pp. 1790–1793.

[8] Y. Liu and et. al, "Class-oriented spectral partitioning for remotely sensed hyperspectral image classification," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 10, no. 2, pp. 691–711, 2017.

[9] G. Zhang, X. Jia, and J. Hu, "Superpixel-based graphical model for remote sensing image mapping," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 11, pp. 5861–5871, 2015.

[10] M. Dalla, A. Villa, J. Benediktsson, et al., "Classification of hyperspectral images by using extended morphological attribute profiles and independent component analysis," *IEEE GRSL*, vol. 8, no. 3, pp. 542–546, 2011.

[11] D. Tuia, R. Flamary, and M. Barlaud, "Nonconvex regularization in remote sensing," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 11, pp. 6470–6480, 2016.

[12] A. Tarantola, *Inverse problem theory and methods for model parameter estimation*, SIAM, 2005.

[13] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.

[14] J. Bioucas-Dias and M.Figueiredo, "Alternating direction algorithms for constrained sparse regression: Application to hyperspectral unmixing," in *2010 2nd Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing*, June 2010, pp. 1–4.

[15] F. Sun et al., *Extreme Learning Machines 2013: Algorithms and Applications*, vol. 16, Springer, 2014.

[16] R. Green et al., "Imaging spectroscopy and the Airborne Visible/Infrared Imaging Spectrometer (AVIRIS)," *Remote Sens. Environ.*, vol. 65, no. 3, pp. 227–248, Sep. 1998.