Trends in Computational and Applied Mathematics

Trends in Computational and Applied Mathematics, **23**, N. 3 (2022), 595-605
Sociedade Brasileira de Matemática Aplicada e Computacional
Online version ISSN 2676-0029
www.scielo.br/tcam
doi: 10.5540/tcam.2022.023.03.00595

Sparse Estimation of the Precision Matrix and Plug-In Principle in Linear Discriminant Analysis for Hyperspectral Image Classification

M. L. PICCO[1*] and M. S. RUIZ[2]

Received on September 30, 2021 / Accepted on March 24, 2022

ABSTRACT. In this paper, a new method for supervised classification of hyperspectral images is proposed for the case in which the size of the training sample is small. It consists of replacing in the Mahalanobis distance the maximum likelihood estimator of the precision matrix by a sparse estimator. The method is compared with two other existing versions of *LDA* sparse, both in real and simulated images.

Keywords: hight dimensionality, linear discriminant analysis, hyperspectral image.

1 INTRODUCTION

Hyperspectral sensors collect information from the earth's surface in hundreds of bands of the electromagnetic spectrum whit relatively narrow bandwidths. Each pixel of a hyperspectral image can be regarded as a high-dimensional vector, $\boldsymbol{x} = (x_1, \ldots, x_p)'$, called feature vector, whose entries x_j, $1 \leq j \leq p$ corresponds to the spectral reflectance collected in the band j of the electromagnetic spectrum.

Image classification is an essential step wich consists in classification rules that assign every data point $\boldsymbol{x}$ into one of K classes based on a features vector. In supervised image classification, a classifier is trained on a training data set of size n. Since the collection of such training data is either expensive or time-consuming, there is usually only a small amount of labeled data available.

The huge number of features (p) of hyperspectral data and the limited availability of the training sample data (n) greatly worsen the traditional classification methods performance such as linear discriminant analysis (LDA) as [21] emphasize. This problem is typical in high-dimensional data analysis.

For example, if $p > n$ the sample covariance matrix S is singular and LDA cannot be applied. If n is not much greater than p, S can be calculated but due to the large number of estimated

*Corresponding author: Mery Lucia Picco – E-mail: mpicco@exa.unrc.edu.ar

[1] University of Rio Cuarto, Ruta Nac. 36, Km. 601, Río Cuarto, Córdoba, Argentina – E-mail: mpicco@exa.unrc.edu.ar https://orcid.org/0000-0002-6374-5910

[2] University of Rio Cuarto, Ruta Nac. 36, Km. 601, Río Cuarto, Córdoba, Argentina – E-mail: mruiz@exa.unrc.edu.ar https://orcid.org/0000-0002-2336-9316

parameters $(p(p+1)/2)$, it will have a significant amount of sample error, and its inverse will be a highly biased estimator of Σ^{-1} ([1]). On the other hand, because the bandwidth decreases as p grows, many bands are highly correlated and hyperspectral data can be contain a significant amount of redundant spectral information. Recently, two algorithms have been proposed called sparse discriminant analysis (SDA) ([5]) and penalized LDA ([20]) which simultaneously solve the problem of singularity of the covariance matrix (when $p > n$) and allow variable selection by imposing a penalty of type ℓ_1 to the optimal scoring approach from which one can derive the LDA rule and the discriminant vectors in Fisher's discriminant problem respectively. In the context of high dimensionality, the estimation of the covariance matrix and its inverse, the precision matrix, is already a research area ofimportant results. Most of the proposals rely on sparsity assumptions and requires some kind of regularization strategy. In this framework, it is assumed that many of the matrix entries to be estimate are zeros. ([10]; [8]; [15]; [9]). Using estimation of sparse precision matrices and plug-in principle in the Mahalanobis distance allows to extend LDA to high dimensional problems [7]. This strategy, will be called *plug-in LDA*.

On the other hand, we are interested in non-paametric tree-based classification methods introduced by Breiman, Friedman, Olshen and Stone in mid 1980's. Therefore, we can expect this approach outperforms *LDA* when the decision boundary is highly non-linear. However these methods can lead to a phenomenon known as overfitting the data, which essentially means that on the training data set they follow the errors too closely having a very poor perfomance on new observations. For this reason trees can be very sensitive to small changes in the training sample dramatically affecting the final estimated tree [11].

The aim of this paper is to compare the performance of *SDA* and *penalized- LDA* versus *plug-in-LDA* using Kashlak proposal based on precision matrix estimation [13]. Since *random forest* (*RF*) is a very popular tree-based classification methods it will be also included in this paper for comparative purposes [2]; [6]; [14].

This paper is organized as follows. We review *LDA*, *SDA* and *penalized LDA* in Section 2 and in Section 3 we present our proposal, the *plug-in-LDA*. A simulation study and applications to real hyperspectral images are presented in Section 5. Section 6 contains a discussion.

2 REVIEW OF LDA AND PENALIZED LDA

There are three distinct arguments that result in the LDA classifier: the multivariate Gaussian model, Fisher's discriminant problem, and the optimal scoring problem.

Consider K populations (classes) where a p-dimensional random vector $\boldsymbol{x}$ is defined and assume that $\boldsymbol{x} \sim N(\boldsymbol{\mu}_k, \Sigma_w)$ where $\boldsymbol{\mu}_k$ is the mean vector for class k and Σ_w is a pooled within-class covariance matrix common to all K classes. The Bayesian approach classifies a new observation $\boldsymbol{x}^*$ to the class k that maximizes Mahalanobis's distance

$$\delta_k(\boldsymbol{x}^*) = (\boldsymbol{x}^* - \widehat{\boldsymbol{\mu}}_k)' \ \widehat{\Sigma}_w^{-1} \ (\boldsymbol{x}^* - \widehat{\boldsymbol{\mu}}_k) - 2log\pi_k \tag{2.1}$$

where $\widehat{\boldsymbol{\mu}}_k$ and $\widehat{\Sigma}_w$ denote maximum likelihood estimates for $\boldsymbol{\mu}_k$ and Σ_w, obtained from a training sample $(\boldsymbol{x}'_1,\dots\boldsymbol{x}'_n)$

LDA, considered as arising from Fisher's discriminant problem, consists in finding K vectors, called discriminant vectors, $\widehat{\boldsymbol{\beta}}_1,\dots\widehat{\boldsymbol{\beta}}_K$ such that sequentially solve

$$\begin{aligned}&\text{maximize}_{\boldsymbol{\beta}_k\in\mathbb{R}^p}\left\{\boldsymbol{\beta}'_k\widehat{\Sigma}_b\boldsymbol{\beta}_k\right\}\\ &\text{subject to}\\ &\boldsymbol{\beta}'_k\widehat{\Sigma}_w\boldsymbol{\beta}_k=1 \text{ and } \boldsymbol{\beta}'_j\widehat{\Sigma}_w\boldsymbol{\beta}_k=0,\forall j<k,\end{aligned}\tag{2.2}$$

where $\widehat{\Sigma}_b$ is the standard estimate for the *between-class covariance matrix*. Since $\widehat{\Sigma}_b$ has rank at most $K-1$ then there exits at most $K-1$ non trivial solutions to (2.2), $\widehat{\boldsymbol{\beta}}_1,\dots,\widehat{\boldsymbol{\beta}}_{K-1}$ called discriminant vectors. the classification rule is reduced to projecting a new observation on the space defined by the discriminant vectors and assigning it to the class whose projected mean is closest.

A third formulation that yields the *LDA* classification rule is *optimal scoring*, introduced by [5], that transform the classification problem into a regression problem by solving

$$\begin{aligned}&\text{minimize}_{\boldsymbol{\theta}_k,\boldsymbol{\beta}_k}\{\|Y\boldsymbol{\theta}_k-\mathbf{X}\boldsymbol{\beta}_k\|^2\}\\ &\text{subject to } \boldsymbol{\theta}'_kY'Y\boldsymbol{\theta}_k=1.\end{aligned}\tag{2.3}$$

where $Y=(y_{ik})$ denotes a $n\times K$ matrix with $y_{ik}=1$ if the i-th observation belongs to the $k-$th class and $y_{ik}=0$ otherwise. Moreover, $\boldsymbol{\theta}_K$ is a scores vector of dimension K and $Y\boldsymbol{\theta}$ represents the scores vector of the training data. The solution $\hat{\boldsymbol{\beta}}_k$ to (2.3) are proportional to solution to (2.2) (see [19]).

Among the proposals to extend LDA to the problems of high dimensionality, those that produce more interpretable models are of particular interest. We need that these extensions allow variable selection and classification simultaneously. In linear regression model, lasso is a regularization method that, imposing a penality at the size of the estimated parameters vector, shrunks some coefficients estimates exactly to zero, producing variable selection. Using this approach [5] propose SDA by imposing to (2.3) an additional penalization:

$$\begin{aligned}&\text{minimize}_{\theta_k,\beta_k}\{\|Y\boldsymbol{\theta}_k-\mathbf{X}\boldsymbol{\beta}_k\|^2+\gamma\boldsymbol{\beta}'_k\Omega\boldsymbol{\beta}_k+\lambda\,\|\boldsymbol{\beta}_k\|_1\}\\ &\text{subject to } \boldsymbol{\theta}'_kY'Y\boldsymbol{\theta}_k=1,\boldsymbol{\theta}'_kY'Y\boldsymbol{\theta}_i=0,\forall i<k\end{aligned}\tag{2.4}$$

where Ω is a positive definite matrix and γ and λ are nonnegative tuning parameters.*SDA* produce sparse discriminant vectors since if λ is large, some entries of β_k will be zero.

Penalized Fisher's linear discriminant, proposed by [20], is based on imposing an ℓ_1 penalty on equation (2.2):

$$\begin{aligned}&\text{maximize}_{\boldsymbol{\beta}_k}\{\boldsymbol{\beta}'_k\widehat{\Sigma}_b\boldsymbol{\beta}_k-\lambda\,\|\boldsymbol{\beta}_k\|_1\}\\ &\text{subject to}\\ &\boldsymbol{\beta}'_k\widehat{\Sigma}_w\boldsymbol{\beta}_k=1 \text{ and } \boldsymbol{\beta}'_j\widehat{\Sigma}_w\boldsymbol{\beta}_k=0,\forall j<k,\end{aligned}\tag{2.5}$$

where $\tilde{\Sigma}_w$ is some full rank estimate for Σ_w, such as $\tilde{\Sigma}_w = \text{diag}(\widehat{\sigma_1}^2, \cdots, \widehat{\sigma_p}^2)$, where $\widehat{\sigma_j}^2$ is the j-th diagonal element of $\widehat{\Sigma}_w$. This method also produce sparse discriminant vectors. Although there is a correspondence between the critical points of problems (2.4) and (2.5), the solutions are not necessarily the same [20].

3 PLUG-IN LDA

In high dimensionality, to estimate the precision matrix associated to a multivariate Gaussian distribution it usual to assume that it is sparse. The most commonly estimation methods employ an ℓ_1-penalized maximum likelihood ([10]; [8]; [4]). As an alternative, [13] propose a novel distribution free estimator that controls the false positive rate in the selection of the non zero entries of the estimated precision matrix. This proposal begins with an initial estimator $\widehat{\Omega}_0$- such as debiased graphical lasso ([12]) or debiased ridge estimator ([3]) and iteratively using a binary search procedure locates the densest estimator near I_p or alternatively the sparsest estimator close to $\widehat{\Omega}_0$. Considering these sparse estimates and plug-in principle in the Mahalanobis distance we propose a new version of LDA sparse which we will call *plug-in LDA*.

4 ACCURACY ASSESSMENT MEASURES

In selecting the best classification algorithm a very important task is the choice of an appropriate performance evaluation measures. Accuracy assessment is traditionally conducted using the sample confusion matrix, in which classification results of the validation dataset are compared to “ground truth.” Based on the confusion matrix, a variety of measures can be calculated, including Cohen’s Kappa coefficient [8] and the $F_1 score$ [18]

Cohen kappa statistic is a chance-corrected method for assessing agreement. The values of range lie in $[-1;1]$ with 1 presenting complete agreement and 0 meaning no agreement or independence. A negative statistic implies that the agreement is worse than random.

Table 1: The confusion matrix for a multi-class classification involving k classes.

True class Predicted class	1	2	$\cdots$	k	Total
1	n_{11}	n_{12}	$\cdots$	n_{1k}	$n_{1.}$
2	n_{21}	n_{22}	$\cdots$	n_{2k}	$n_{2.}$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
k	n_{k1}	n_{k2}	$\cdots$	n_{kk}	$n_{k.}$
Total	$n_{.1}$	$n_{.2}$	$\cdots$	$n_{.k}$	N

Using notation similar to Cohen (1960), the kappa coefficient of agreement,(κ), is estimated from confusion matrix:

$$\hat{\kappa} = \frac{p_0 - p_e}{1 - p_e} \tag{4.1}$$

where p_0 is the proportion of cases correctly classified (i.e. overall accuracy) and p_e is the expected proportion of cases correctly classified by chance (under independency hipotesis):

$$p_0 = \frac{\sum_{i=1}^{k} n_{ii}}{N} \tag{4.2}$$

$$p_e = \frac{\sum_{i=1}^{k} n_{i.}n_{.i}}{N^2} \tag{4.3}$$

Precision and recall (sensitivity) are two important model evaluation measures. For a multiclass problem we can compute the *per-class* precisión and recall as follows:

$$recall_i = \frac{n_{ii}}{n_{.i}}; \qquad 1 \leq i \leq k \tag{4.4}$$

$$precision_i = \frac{n_{ii}}{n_{i.}}; \qquad 1 \leq i \leq k \tag{4.5}$$

The F_1 value is used to combine the precision and recall measurements into a single value. This is practical because it makes it easier to compare the combined performance of precision and completeness between various algorithms. $F1$ score is a harmonic mean of precision and recall:

$$(F_1)_i = \frac{2}{\frac{1}{recall_i} + \frac{1}{precision_i}}; \qquad 1 \leq i \leq k \tag{4.6}$$

The weighted F_1 is defined as

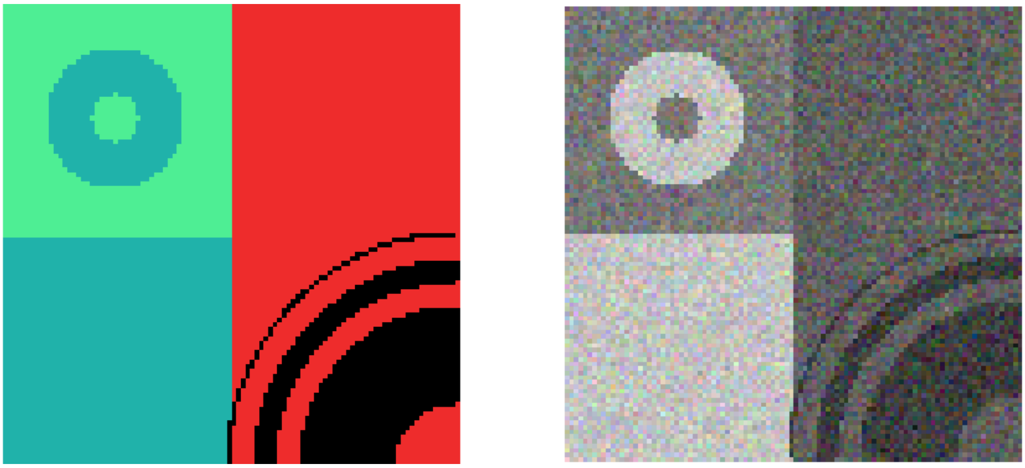
$$wF_1 = \frac{\sum_{i=1}^{k} n_{.i}F1_i}{N} \tag{4.7}$$

To get a more accurate estimate of the measurements, the training data set was randomly divided into three groups, or folds, of equal size ($n = p$). The first fold is treated as a training set, and remaining folds as validation set. This process results in three estimates of each measure, which are averaged in order to obtain a unique value [11].

5 APPLICATIONS

Experiments were conducted to test the performance of the *SDA*; *penalized- LDA*; *plug-in-LDA* and *RF* algorithms with one simulated image and one real image.

For SDA, Penalized-LDA and RF we will use the R-packages `sda`, `sparseMatEst` and `randomForest`, respectively.



(a)

(b)

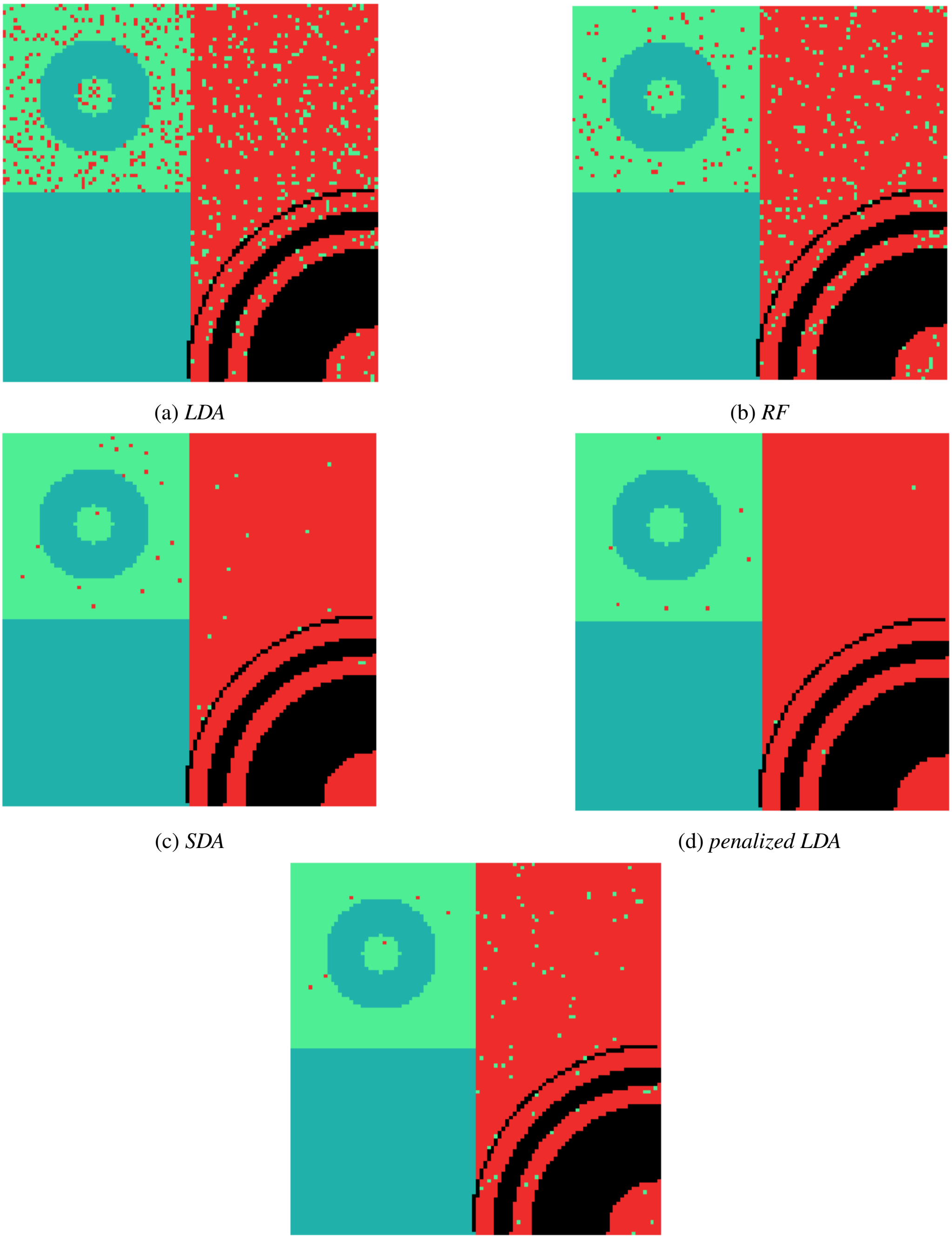
Figure 1: (a) Ideal class image with four regions. (b) RGB-composition of simulated random image assuming a normal multivariate distribution.

5.1 Simulated Data

The first experimental dataset (simulated image) consist of 100×100 pixels, 200 bands and four classes. Each pixel $\boldsymbol{x}_{ij}$; $1 \leq i, j \leq 100$ is a p-dimensional random vector, where p represent the number of bands. We assume that $\boldsymbol{x}_{ij} \sim N(\boldsymbol{\mu}_k; \Sigma)$ if pixel (i, j) belongs to class k, $1 \leq k \leq 4$. Figure 1a and 1b represent an ideal class image with four regions, acting as the reference image, and the RGB-composition of simulated random image assuming a normal distribution. The size of the training set is 160 (40 pixels of each class). Table 2 shows the Frobenius norm and kappa estimated coefficient for the five classifiers. The validation is made with the entire image. The Figure 2 and the Table 2 show that, *penalized-LDA* has the best performance followed by *SDA* and *plug-in-LDA*, while *RF* has the worst performance.

Table 2: Frobenius norm. (kappa estimated coefficient).

training sample	RF	SDA	Pen-LDA	Plug-in LDA
	36.3 (0.953)	11.3 (0.995)	6.3 (0.998)	17.4 (0.989)



(a) *LDA*

(b) *RF*

(c) *SDA*

(d) *penalized LDA*

(e) *LDA plug-in*

Figure 2: Classification results simulated image.

5.2 Real Data

Figure 3 shows an 209 x 167 extract of an EO-1 Hyperion image obtained in http://eo1.usgs.gov. The study area is $EO1H2290822012007110P1_1T$. 44 non-informative bands were removed, as well as water absorption bands 120–132, 165–182, 185–187, 221–224. So, 160 bands were established to be useful for further analysis. By prior knowledge of the area, it was decided to classify the image into four classes: water, urban area, high vegetation index area, and low vegetation index area (blue, white, green and brown, respectively). The size of the training sample is 160.

Table 3: Kappa estimated coeficient and weighted F_1 score averaged over $k = 3$ folds.

	LDA	RF	SDA	Pen-LDA	Plug-in LDA
wF_1 score	0.43	0.90	0.94	0.94	0.92
$\hat{\kappa}$	0.34	0.87	0.92	0.92	0.90

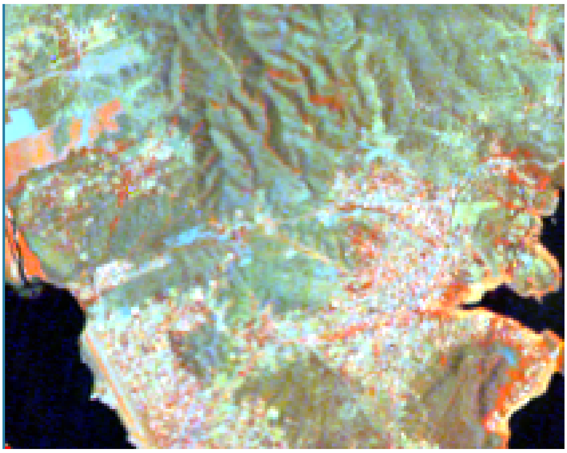
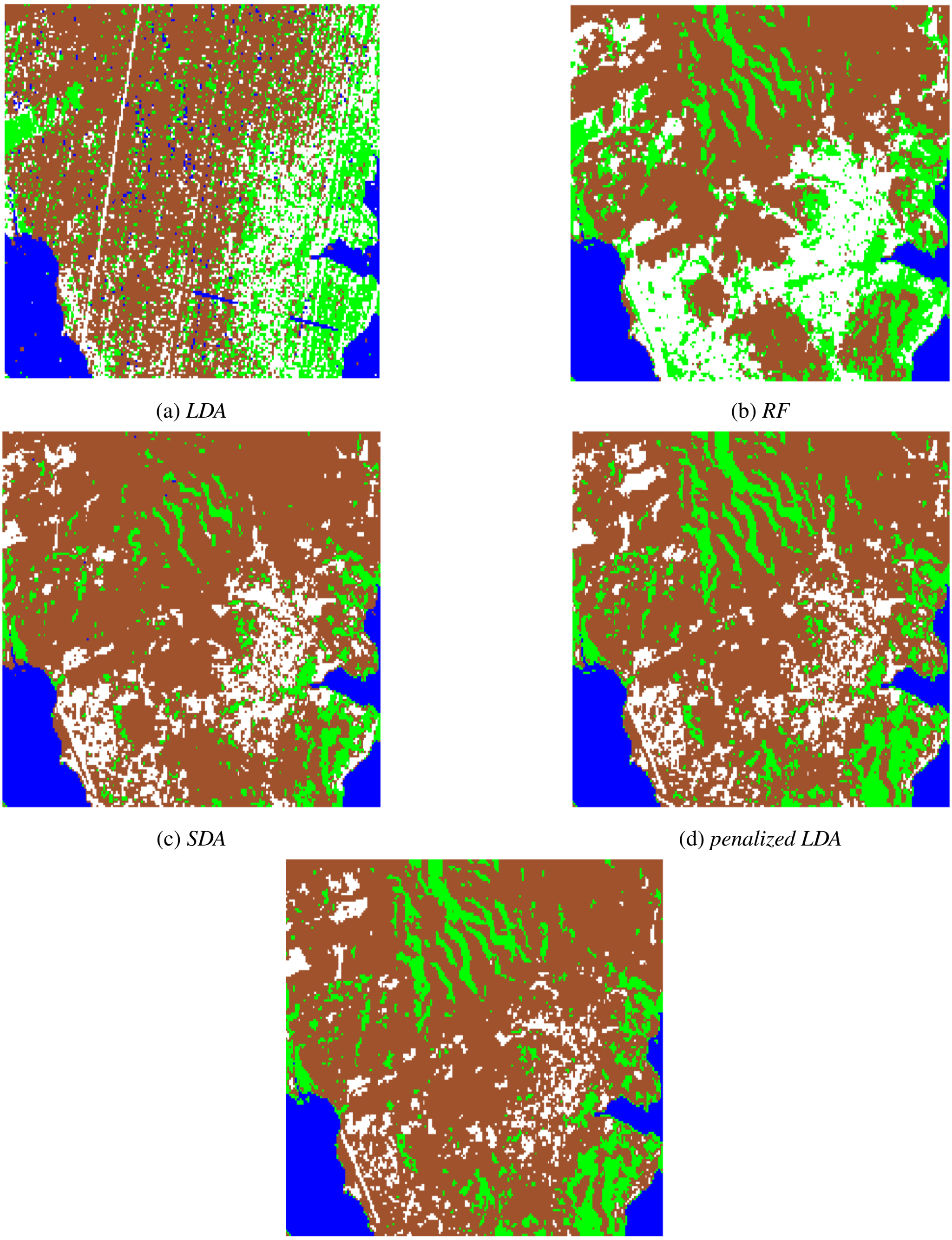


Figure 3: Real image: false color composite image (R-G-B = band 120-60-2).

(a)

(b)

Figure 4: Google maps extract (a) city area (b) mountain area.



(a) *LDA*

(b) *RF*

(c) *SDA*

(d) *penalized LDA*

(e) *plug-in-LDA*

Figure 5: Classification results real image.

Figure 5 show the thematic maps produced by the five classifiers. When executing the LDA function of the R-package MASS, the following error message appears: ”variables are collinear”. This problem, caused by the insufficient size of the training sample to estimate the covariance matrix, produces an unacceptable thematic map (see Figure 5a).

If the algorithms are ranking based on adequacy measures presented in the table 3, the first one are *SDA* and *penalized-LDA*, followed by *plug-in-LDA* and finally *RF*. A visual comparison of the thematic maps (see Figure 5c, 5d, and 5e) with the truth of the terrain (Figure 4a and 4b), seems to show that *SDA* had a better performance, being very similar to *penalized-LDA* and *plug-in-LDA*, while *RF* had the worst performance.

6 CONCLUSION

This paper proposes a new version of sparse *LDA*, called *plug-in LDA*, for the supervised classification of hyperspectral images. It consists of replacing the maximum likelihood estimator of the precision matrix used in Mahalanobis distance with a sparse estimate recently proposed. The results obtained show that *plug-in LDA* outperforms to *RF* and has a similar behavior to *SDA* and *Penalized LDA*, both in the real image and in the simulated one.

In the context of hiperspectral image classification, extracting totally pure training samples is difficult, since there are often incorrectly labeled pixels, wich can have a strong negative impact on the classification result [16]; [17]. The main advantage of the *plug-in* strategy is the possibility of replacing in the Mahalanobis’s distance the estimators of both the position and the covariance matrix with robust alternatives which could turn *plug-in* into a robust classification algorithm.

REFERENCES

[1] J. Bai & S. Shi. Estimating High Dimensional Covariance Matrices and Its Applications. *Annals of Economics and finance*, **12**(2) (2011), 199–215.

[2] L. Breiman. Random forests. *Machine Learning*, **1** (2001), 5–32.

[3] P. Buhlmann. Statistical significance in high-dimensional linear models. *Bernoulli*, **19** (2015), 1212–1242.

[4] T. Cai & W. Liu. Adaptive thresholding for sparse covariance matrix estimation. *Journal of the American Statistical Association*, **106** (2011), 672–684.

[5] L. Clemmensen, T. Hastie, D. Witten & B. Ersbøll. Sparse Discriminant Analysis. *Technometrics*, **53**(4) (2011), 406–413. doi:10.1198/TECH.2011.08118.

[6] T.N. Do, P. Lenca, S. Lallich & N.K. Pham. Classifying Very-High-Dimensional Data with Random Forests of Oblique Decision Trees. *Advances in knowledge discovery and management*, **1** (2009).

[7] J. Fan, Y. Feng & Y. Wu. Network exploration via the adaptive LASSO and SCAD penalties. *The Annals of Applied Statistics*, **3**(2) (2009), 521–541. doi:10.1214/08-AOAS215.

[8] O.B.L. Ghaoui & A. d'Aspremont. Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary Data. *Journal of Machine Learning Research*, **9** (2008), 485–516.

[9] Y.L. J. Fan & H. Liu. An overview on the estimation of large covariance and precision matrices. *The Econometrics Journal*, **19**(1) (2016), 1–34. doi:https://doi.org/10.1111/ectj.12061.

[10] T.H. J. Friedman & R.Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, **9**(3) (2007), 432–441.

[11] G. James, D. Witten, T. Hastie & R. Tibshirani. "An Introduction to Statistical Learning with Applications in R". Springer, 2 ed. (2017).

[12] J. Jankova & S. Van De Geer. Confidence intervals for high-dimensional inverse covariance estimation. *Electronic Journal of Statistics*, **9** (2015), 1205–1229.

[13] A. Kashlak. Non-asymptotic error controlled sparse high dimensional precision matrix estimation. *arXiv:1903.10988 [stat.ME]*, (2019).

[14] C. Li. The Application of high-dimensional Data Classification by Random Forest based on Hadoop Cloud Computing Platform. *Chemical Engineering Transactions*, **51** (2016).

[15] W. Liu & X. Luo. Fast and adaptive sparse precision matrix estimation in high dimensions. *Journal of Multivariate Analysis*, **135** (2015), 153–162. doi:https://doi.org/10.1016/j.jmva.2014.11.005. URL `https://www.sciencedirect.com/science/article/pii/S0047259X14002607`.

[16] M. Popov, S. Alpert, V. Podorvan, M. Topolnytskyi & S. Mieshkov. Method of Hyperspectral Satellite Image Classification under Contaminated Training Samples Based on Dempster-Shafer's Paradigm. *Central European Researchers Journal*, **1** (2015), 82–93.

[17] L. Qingming, Y. Haiou & J. Junjun. Label Noise Cleansing with Sparse Graph for Hyperspectral Image Classification. *Remote Sensing*, **11** (2019), 1116. doi:10.3390/rs11091116.

[18] Y. Sasaki. The truth of the F-measure. *Teach Tutor Mater*, (2007), 1–5.

[19] A.B. T. Hastie & R. Tibshirani. Penalized Discriminant Analysis. *The Annals of Statistics*, **23**(1) (1995), 73–102.

[20] D. Witten & R. Tibshirani. Penalized classification using Fisher's linear discriminant. *J R Stat Soc Series B Stat Methodol*, **73**(54) (2011), 753–772. doi:0.1111/j.1467-9868.2011.00783.x.

[21] H. Zou. Classification with high dimensional features. *WIREs computational Statistics*, **11** (2019). doi:https://doi.org/10.1002/wics.1453.