Munoz-Romero, S., Gomez-Verdejo, V. & Arenas-Garcia, J. (2016). Regularized Multivariate Analysis Framework for Interpretable High-Dimensional Variable Selection. *IEEE Computational Intelligence Magazine, 11*(4), 24–35.

DOI: 10.1109/mci.2016.2601701

Regularized Multivariate Analysis Framework for Interpretable High-Dimensional Variable Selection

Sergio Muñoz-Romero, Department of Signal Processing and Communications, Universidad Rey Juan Carlos, Madrid, SPAIN
Vanessa Gómez-Verdejo, and Jerónimo Arenas-García, Department of Signal Processing and Communications, Universidad Carlos III de Madrid, Madrid, SPAIN

***Abstract*—Multivariate Analysis (MVA) comprises a family of well-known methods for feature extraction which exploit correlations among input variables representing the data. One important property that is enjoyed by most such methods is uncorrelation among the extracted features. Recently, regularized versions of MVA methods have appeared in the literature, mainly with the goal to gain interpretability of the solution. In these cases, the solutions can no longer be obtained in a closed manner, and more complex optimization methods that rely on the iteration of two steps are frequently used. This paper recurs to an alternative approach to solve efficiently this iterative problem. The main novelty of this approach lies in preserving several properties of the original methods, most notably the uncorrelation of the extracted features. Under this framework, we propose a novel method that takes advantage of the $\ell_{2,1}$ norm to perform variable selection during the feature extraction process. Experimental results over different problems corroborate the advantages of the proposed formulation in comparison to state of the art formulations.**

I. Introduction

Multivariate Analysis (MVA) comprises a collection of tools that play a fundamental role in statistical data analysis. These techniques have become increasingly popular since the proposal of Principal Component Analysis (PCA) in 1901 [1]. PCA was proposed as a simple and efficient way to reduce data dimension by projecting the data over the largest variance directions. As illustrated in Fig. 1, PCA learns from a given dataset a set of projection vectors, so that data can be represented in a low-dimensional space that preserves the directions of the input space where the data shows the largest variance. A typical example to illustrate PCA is face recognition, where the projection vectors are known as eigenfaces [2]. Nevertheless, PCA has been used in many other applications, and can indeed be considered as one of the most widely-used tools for feature extraction.

Other MVA algorithms have emerged that are especially suited to supervised learning tasks (e.g., in regression and classification). In these problems, the goal is not just to represent the input data as efficiently as possible, but it actually becomes of major importance to keep the directions of the input space that are more highly correlated with the label information. This is the case of algorithms such as Canonical Correlation Analysis (CCA) [3], Partial Least Squares (PLS) approaches [4], [5], and Orthonormalized PLS (OPLS) [6]. Consider for instance a toy classification problem in Fig. 2. In this problem, the direction of the maximum variance extracted by PCA (left subplot) results in overlapping distributions of the two classes along this direction, while a supervised method like OPLS (right subplot) successfully identifies the most discriminative information. Although this toy example is based on a classification task, the same advantages of supervised MVA over standard PCA are encountered in regression tasks—see [7] for a detailed theoretical and experimental review of these methods.

The simplicity of these methods, as well as the availability of highly-optimized libraries for solving the linear algebra problems they involve, justifies the extensive use of MVA in many application fields, such as biomedical engineering [8], [9], remote sensing [10], [11], or chemometrics [12], among many others (see also [7] for a more detailed review of application-oriented research in the field).

An important property of PCA, OPLS, and CCA is that they lead to uncorrelated variables, so that the feature extraction process provides additional advantages:

- The relevance of each extracted feature is directly given by the magnitude of its associated eigenvalue, which simplifies the selection of a reduced subset of features, if necessary.
- Subsequent learning tasks are simplified, more notably, when the covariance matrix inversion is required. This is the case of least-square based problems, such as Ridge Regression or lasso (least absolute shrinkage and selection operator) [13].

Standard versions of MVA methods implement just a feature extraction process, in the sense that all original variables are used to build the new features. However, over the last few years there have been many significant contributions to this field that have focused on gaining interpretability of the extracted features by incorporating sparsity-inducing norms, such as the ℓ_1 and $\ell_{2,1}$ norms [14], as a penalty term in the minimization problem. When these regularization terms are included, the projection vectors are favored to include zeros in some of their components, making it easier to understand the process to build the new features and thus gaining in interpretability. In fact, the $\ell_{2,1}$ rewards solutions that perform a real variable selection process, in the sense that some of the original variables are excluded from all projection vectors at once. In other words, only a subset of the original variables

Corresponding Author: Sergio Muñoz-Romero (e-mail: sergio.munoz@urjc.es)

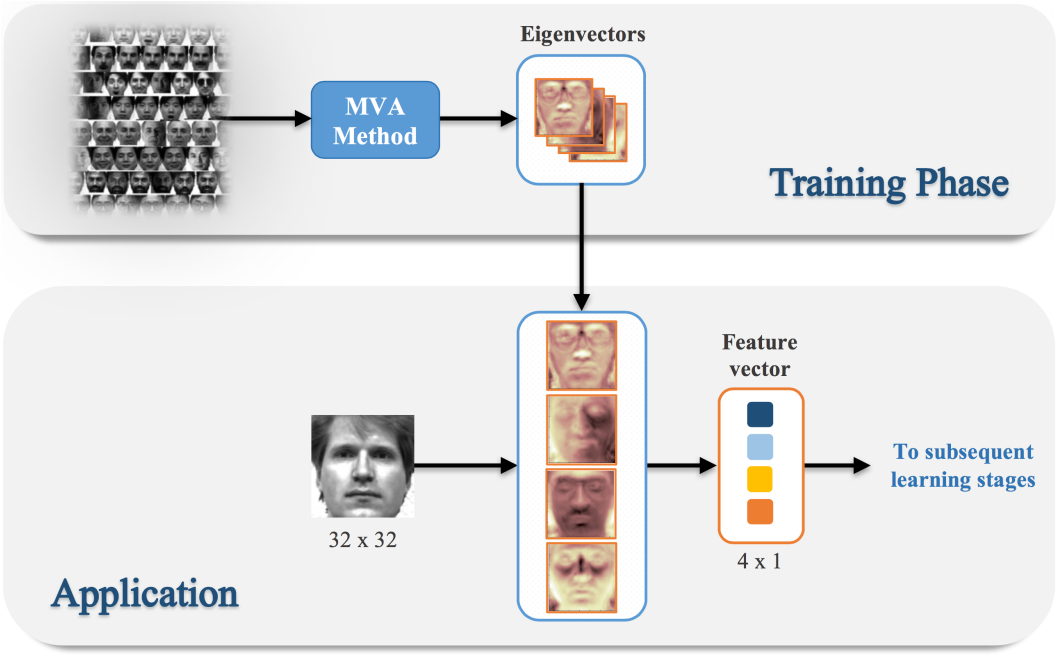
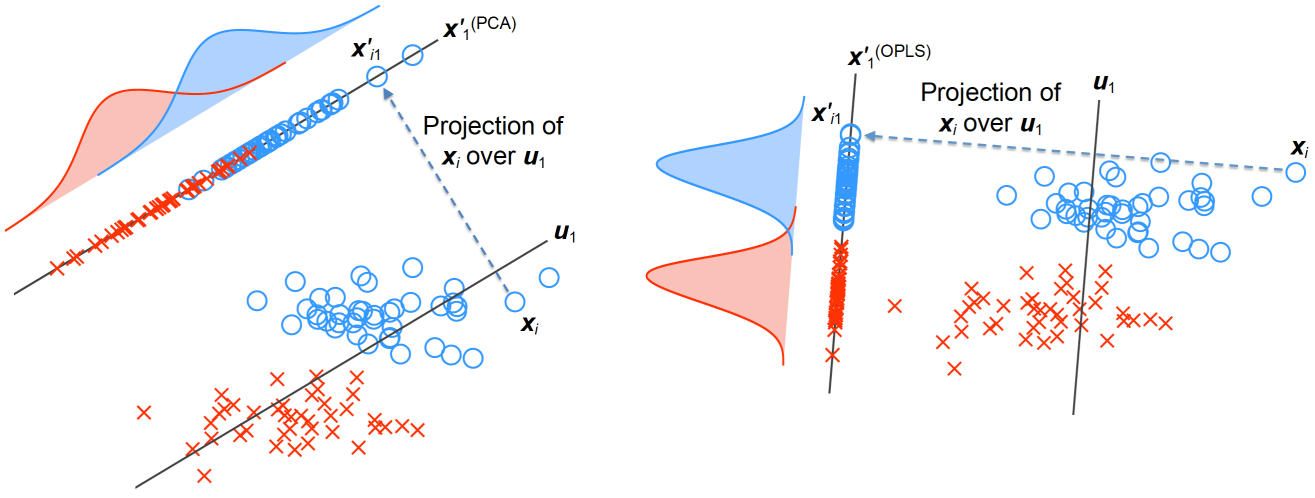


Fig. 1: Feature extraction with Multivariate Analysis. During the training phase, MVA methods learn the most relevant directions for a particular dataset (in face recognition these vectors are known as *eigenfaces* when the PCA method is applied). Feature extraction is then carried out multiplying any vector in the input representation space with these eigenvectors. Subsequent learning tools can then be applied on this new subspace of reduced dimensionality.



(a) PCA (b) OPLS

Fig. 2: Projected data over the first eigenvector of PCA and OPLS in a binary classification task.

are used to build the new features.

Some of the most significant contributions in this direction are sparse PCA [15], sparse OPLS [8], group-lasso penalized OPLS (also known as Sparse Reduced Rank Regression, SRRR) [16], and $\ell_{2,1}$-regularized CCA (or L21SDA) [17]. All these approaches are based on an iterative process which combines the optimization of two coupled least-squares problems, one of them subject to a minimization constraint. Since the inspiring work [15], this constrained least-squares minimization has been typically treated as an orthogonal Procrustes problem [18], an approach that can still be considered mainstream (see, e.g., the very recent works [19], [20]).

A first objective of this paper is to highlight and make the computational intelligence community aware of some limitations derived from the use of orthogonal Procrustes in the context of regularized MVA methods. As explained in [21], these methods

1) do not converge to their associated non-regularized MVA solutions when the penalty term is removed,
2) are highly dependent on initialization, and may even fail to progress towards a solution,
3) do not in general obtain uncorrelated features.

As solution to these problems, [21] proposes an alternative optimization procedure avoiding the use of the Procrustes solution. In this paper, we will briefly review the framework presented in [21] to derive regularized MVA versions, and illustrate the approach and its associated advantages by introducing a novel MVA method using the $\ell_{2,1}$ norm [14] as regularization term. Apart from the advantages we have already discussed implying variable selection over the original variables, this norm holds the property of rotational invariance, a fact that we will exploit to significantly reduce the computational cost of the training phase. Although some authors have already adapted the robust variable selection method [14] to the MVA scenario (see, e.g., the group-lasso penalized OPLS method [16] or the $\ell_{2,1}$-regularized CCA [17]), these adaptations are based on orthogonal Procrustes and the rotational invariance property of the $\ell_{2,1}$ norm is not exploited, taking unnecessary extra computational burden.

In short, the main contributions of this paper can be summarized as:

- Review a framework for regularized MVA methods, and explain an alternative to the most commonly used Procrustes solution to overcome the limitations of this approach.
- Obtain novel MVA algorithms based on $\ell_{2,1}$ regularization.
- Illustrate the effectiveness of these algorithms to carry out feature extraction and, at the same time, obtain some understanding of the original input variables.

The rest of the paper is organized as follows. Section II reviews the common framework for regularized multivariate analysis, and explains an advantageous alternative to the use of orthogonal Procrustes in this context. Then, Section III particularizes the MVA framework by including an $\ell_{2,1}$ norm penalty, explaining in detail how to derive a computationally efficient solution and pointing the differences between our proposal and other existing solutions. Section IV is devoted to experiments and Section V draws the main conclusions of our work.

II. Regularized MVA framework: enforcing feature uncorrelation

Let us assume a supervised learning scenario, where the goal is to carry out feature extraction in the input space, learning the projection vectors from a training dataset of N input-output pairs $\{\boldsymbol{x}_i, \boldsymbol{y}_i\}_{i=1}^N$, where $\boldsymbol{x}_i \in \Re^n$ and $\boldsymbol{y}_i \in \Re^m$ are the input and output vectors, respectively. Therefore, n and m denote the dimensions of the input and output spaces. For notational convenience, we define the input and output data matrices: $\mathbf{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N]$ and $\mathbf{Y} = [\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N]$, with columnwise arranged patterns. It will be assumed throughout the paper that these matrices are centered [22], so that sample estimations of the input and output data covariance matrices, as well as of their cross-covariance matrix, can be calculated as $\mathbf{C_{XX}} = \mathbf{X}\mathbf{X}^\top$, $\mathbf{C_{YY}} = \mathbf{Y}\mathbf{Y}^\top$ and $\mathbf{C_{XY}} = \mathbf{X}\mathbf{Y}^\top$, where we have neglected the scaling factor $\frac{1}{N}$, and superscript $^\top$ denotes vector or matrix transposition. The goal of linear MVA methods is to find n_f relevant features by combining the original variables, i.e., $\mathbf{X}' = \mathbf{U}^\top\mathbf{X}$, where the kth column of $\mathbf{U} = [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_{n_f}]$ is a vector containing the coefficients associated to the kth extracted feature. Note that we are referring to the components of $\mathbf{x}$ as **variables**, whereas the components of $\mathbf{x}' = \mathbf{u}^\top\mathbf{x}$ are being referred to as **features**. Consequently, feature extraction implies obtaining $\mathbf{x}'$ from $\mathbf{x}$, whereas variable selection is the process of selecting a subset of the original variables in $\mathbf{x}$. Besides, the feature extraction process can also imply variable selection when the projection matrix $\mathbf{U}$ has some of their rows equal to zero.

In this paper, we deal with MVA methods which force the extracted features to be uncorrelated; this applies, at least, to PCA, CCA, and OPLS. MVA methods that do not enforce feature uncorrelation, more notably PLS, are therefore left outside the scope of this paper. A common framework for these regularized MVA methods can be set including an uncorrelation constraint, $\mathbf{U}^\top\mathbf{C_{XX}}\mathbf{U} = \mathbf{I}$, over the formulation of [23]:

$$\begin{aligned} \text{minimize}_{\,\mathbf{W},\mathbf{U}} \quad & \|\boldsymbol{\Omega}^{\frac{1}{2}}\left(\mathbf{Y} - \mathbf{W}\mathbf{U}^\top\mathbf{X}\right)\|_F^2 + \gamma R(\mathbf{U}) \\ \text{subject to} \quad & \mathbf{U}^\top\mathbf{C_{XX}}\mathbf{U} = \mathbf{I} \end{aligned} \tag{1}$$

where $\mathbf{W}$ is an $m \times n_f$ matrix of regression coefficients, parameter γ trades off the importance of the regularization term $R(\mathbf{U})$, $\|\mathbf{A}\|_F = \mathrm{Tr}\{\mathbf{A}\mathbf{A}^\top\}$ denotes the Frobenius norm of matrix $\mathbf{A}$, and $\mathrm{Tr}\{\cdot\}$ is the trace operator. Finally, different selections of matrix $\boldsymbol{\Omega}$ give rise to the considered MVA methods, in particular $\boldsymbol{\Omega} = \mathbf{C}_{\mathbf{YY}}^{-1}$ for CCA, $\boldsymbol{\Omega} = \mathbf{I}$ for OPLS, and $\boldsymbol{\Omega} = \mathbf{I}$ with $\mathbf{Y} = \mathbf{X}$ for PCA [23], [24].

The objective function in (1) is composed of two terms. The first term tries to minimize the reconstruction error when matrix $\mathbf{Y}$ is estimated from the projected data as $\mathbf{W}\mathbf{X}' = \mathbf{W}\mathbf{U}^\top\mathbf{X}$. Note that this is different from standard least-squares since the introduction of matrix $\mathbf{U}$ imposes a representation bottleneck [25], i.e., matrix $\mathbf{Y}$ needs to be approximated from a matrix $\mathbf{X}'$ with less features than the original matrix $\mathbf{X}$. The regularization term $R(\mathbf{U})$ is usually a particular matrix norm that gives a desired property to the solution. Three common regularization terms are:

- $R(\mathbf{U}) = \|\mathbf{U}\|_F^2 = \sum_{ij} U_{ij}^2$, where U_{ij} is the element in the ith-row and jth-column of $\mathbf{U}$. This term is known as Tikhonov, ridge, or ℓ_2 regularization, and it is used to improve the conditioning of the solution.
- $R(\mathbf{U}) = \|\mathbf{U}\|_1 = \max_j \sum_{i=1}^N |U_{ij}|$, where $|U_{ij}|$ is the absolute value of U_{ij}. This term is known as lasso regularization [13] and it is frequently used to induce sparsity on the solution matrix (i.e., to nullify some elements of $\mathbf{U}$).
- $R(\mathbf{U}) = \|\mathbf{U}\|_{2,1} = \sum_{i=1}^n \|\boldsymbol{u}^i\|_2$, being $\boldsymbol{u}^i$ the ith row of $\mathbf{U}$. This is known as $\ell_{2,1}$ regularization and penalizes all n_f coefficients corresponding to a single variable as a whole, making them drop to zero jointly, thus favoring variable selection.

Fig. 3 depicts the solution matrix $\mathbf{U}$ for the above regularization terms over a toy problem. As it can be seen, ℓ_1 and $\ell_{2,1}$ penalties result in many elements of $\mathbf{U}$ dropping to zero. Furthermore, $\ell_{2,1}$ norm provides the sparsity in a structured way, i.e., the coefficients of $\mathbf{U}$ are annulled by rows. Since

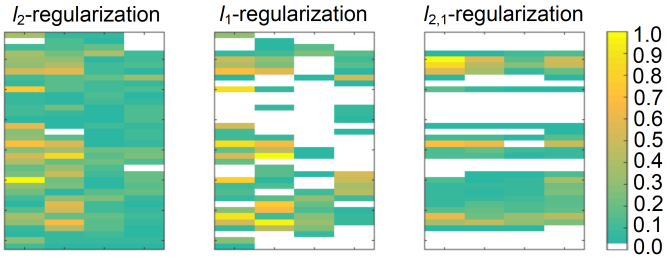


Fig. 3: Examples of the provided projection matrices $\mathbf{U}$ for a case with $n = 40$ and $n_f = 4$ and considering the regularizations ℓ_2, ℓ_1, and $\ell_{2,1}$. Coefficients that take a zero value have been identified and represented in white.

each row of $\mathbf{U}$ is associated to a different input variable, this structured sparsity implies that many input variables are completely ignored during the feature extraction, so that variable selection is also achieved in this case.

When using non-derivable penalties, such as ℓ_1 or $\ell_{2,1}$ norms, the solution of the minimization problem in (1) cannot be obtained in closed-form. However, as shown in [24] an equivalent formulation can be obtained by replacing the uncorrelation constraint in (1) by $\mathbf{W}^\top \mathbf{\Omega} \mathbf{W} = \mathbf{I}$, leading to

$$\begin{aligned} \text{minimize}_{\mathbf{W},\mathbf{U}} \quad & \|\mathbf{\Omega}^{\frac{1}{2}} \left(\mathbf{Y} - \mathbf{W}\mathbf{U}^\top \mathbf{X}\right)\|_F^2 + \gamma R(\mathbf{U}) \\ \text{subject to} \quad & \mathbf{W}^\top \mathbf{\Omega} \mathbf{W} = \mathbf{I}. \end{aligned} \tag{2}$$

As demonstrated in [24], (1) and (2) provide the same solution, but using (2) is normally more efficient for the common case $m \ll n$. Furthermore, placing the constraint on matrix $\mathbf{W}$ allows to solve the problem with the two-step iterative method that we describe in Algorithm 1 (please, refer to [24] for further details). For simplicity, the solution of the $\mathbf{W}$-step has been written in terms of a new matrix $\mathbf{V} = \mathbf{\Omega}^{\frac{1}{2}}\mathbf{W}$. As it can be seen, the minimization problems involved by the $\mathbf{U}$- and $\mathbf{W}$-steps are coupled, so it becomes necessary to iterate both steps until some convergence criterion is met.

The $\mathbf{U}$-step is a regularized least-squares problem that can take advantage of a great variety of existing efficient solvers [14], [26], [27]. With respect to the $\mathbf{W}$-step, it is important to point out that uncorrelation has to be enforced. When this is not the case, the above steps can provide infinite coupled pairs of solutions which are rotated versions of the desired ones, and the uncorrelation property of the extracted features is lost.

In fact, in the literature $\mathbf{W}$-step is typically solved by using the orthogonal Procrustes approach. As it has been proved in [21], this solution neglects the uncorrelation among the extracted features. In spite of this limitation, since its initial proposal in [15] for the sparse PCA algorithm, it has been later extended to supervised approaches such as sparse OPLS [8], group-lasso penalized OPLS [16], and $\ell_{2,1}$-regularized CCA (or L21SDA) [17]. In this respect, this paper aims (1) to highlight the limitations of Procrustes when used as part of the above iterative method, and (2) to encourage the adoption of an alternate method for the $\mathbf{W}$-step that pursues feature uncorrelation.

A. Solving the $\mathbf{W}$-step with Orthogonal Procrustes

The minimization problem in the $\mathbf{W}$-step is known as orthogonal Procrustes, and its optimal solution is given by $\mathbf{V}_\text{P} = \mathbf{Q}\mathbf{P}^\top$, where $\mathbf{Q}$ and $\mathbf{P}$ are the matrices of the singular value decomposition $\mathbf{C}_{\bar{\mathbf{Y}}\mathbf{X}'} = \mathbf{Q}\mathbf{\Sigma}\mathbf{P}^\top$ [18], with $\mathbf{C}_{\bar{\mathbf{Y}}\mathbf{X}'} = \bar{\mathbf{Y}}\mathbf{X}'^\top$, and where $\bar{\mathbf{Y}}$ was defined in Algorithm 1.

However, when the Procrustes solution is adopted, the uncorrelation of the extracted features is not explicitly imposed during this step. In the simplest case when the regularization is not used ($\gamma = 0$), this causes that the extracted features differ from those of the corresponding standard MVA formulation (see [21] for further details and experimental results). For the general case in which $\gamma > 0$, the Procrustes solution results in higher correlation among the features than when using the alternative method described in the next subsection, as will also be illustrated in Section IV.

Apart from the correlation among the extracted features, the use of Procrustes also makes the algorithm highly dependent on initialization. For instance, it can be shown that when the regularization is removed and $\mathbf{V}$ is initialized as an orthogonal matrix the algorithm fails to progress at all.

B. Solving the $\mathbf{W}$-step as an Eigenvalue Problem

In [21], it is proven that the uncorrelation of the extracted features can be obtained if the $\mathbf{W}$-step is solved by means of the following eigenvalue problem:

$$\mathbf{\Omega}^{\frac{1}{2}}\mathbf{C}_{\mathbf{XY}}^\top \mathbf{U}\mathbf{U}^\top \mathbf{C}_{\mathbf{XY}}\mathbf{\Omega}^{\frac{1}{2}}\mathbf{V} = \mathbf{V}\mathbf{\Lambda}, \tag{3}$$

being $\mathbf{\Lambda}$ the diagonal matrix containing the n_f largest eigenvalues arranged in decreasing order.

The desired uncorrelation is obtained due to this eigenvalue problem forces the diagonalization of the matrix $\mathbf{U}^\top \mathbf{C}_{\mathbf{XY}}\mathbf{\Omega}^{\frac{1}{2}}\mathbf{V}$, which is a necessary condition to meet the uncorrelated extracted features.

Table I includes a summary of the $\mathbf{U}$- and $\mathbf{W}$-steps for the particular cases of regularized CCA, OPLS and PCA, when formulating the $\mathbf{W}$-step as an eigenvalue problem. Remember that $\mathbf{W}$ can be straightforwardly computed from $\mathbf{V}$ using the relation $\mathbf{W} = \mathbf{\Omega}^{-\frac{1}{2}}\mathbf{V}$.

III. MVA methods with $\ell_{2,1}$ penalty

In this section, we particularize the presented MVA framework for the $\ell_{2,1}$ regularization norm. In this way, we can take advantage of the variable selection property enjoyed by this norm and obtain an algorithm that can simultaneously perform dimensionality reduction and variable selection.

For this purpose, let us replace $R(\mathbf{U}) = \|\mathbf{U}\|_{2,1}$ in (2). Rewriting also the minimization problem in terms of $\mathbf{V} = \mathbf{\Omega}^{\frac{1}{2}}\mathbf{W}$, we arrive at

$$\begin{aligned} \text{minimize}_{\mathbf{V},\mathbf{U}} \quad & \|\mathbf{Y}' - \mathbf{V}\mathbf{U}^\top \mathbf{X}\|_F^2 + \gamma \|\mathbf{U}\|_{2,1}, \\ \text{subject to} \quad & \mathbf{V}^\top \mathbf{V} = \mathbf{I}, \end{aligned} \tag{4}$$

where $\mathbf{Y}' = \mathbf{\Omega}^{\frac{1}{2}}\mathbf{Y}$ is the new output matrix.

Considering the iterative solution detailed in Subsection II-B, the solution of (4) can be obtained by an iterative procedure consisting of two coupled steps:

1.- Input: $\mathbf{X}$, $\mathbf{Y}$, $\mathbf{\Omega}$.
2.- Optimization algorithm:
  2.1.- Initialize $\mathbf{W}^{(0)} = \mathbf{I}$.
  2.2.- For $k = 1, 2, \ldots$
    2.2.1.- ($\mathbf{U}$-Step) For fixed $\mathbf{W}$ (satisfying $\mathbf{W}^\top \mathbf{\Omega} \mathbf{W} = \mathbf{I}$):
      $\mathbf{U}^* = \arg\min_{\mathbf{U}} \|\mathbf{Y}' - \mathbf{U}^\top \mathbf{X}\|_F^2 + \gamma R(\mathbf{U})$
      with $\mathbf{Y}' = \mathbf{W}^\top \mathbf{\Omega} \mathbf{Y}$.
    2.2.2.- ($\mathbf{W}$-Step) For fixed $\mathbf{U}$:
      $\mathbf{V}^* = \arg\min_{\mathbf{V}} \|\bar{\mathbf{Y}} - \mathbf{V}\mathbf{X}'\|_F^2$
      subject to $\mathbf{V}^\top \mathbf{V} = \mathbf{I}$
      with $\mathbf{V} = \mathbf{\Omega}^{\frac{1}{2}} \mathbf{W}$, $\bar{\mathbf{Y}} = \mathbf{\Omega}^{-\frac{1}{2}} \mathbf{Y}$, and $\mathbf{X}' = \mathbf{U}^\top \mathbf{X}$.
    2.2.3.- Back to 2.2.1. until convergence criterion is met.
3.- Output: $\mathbf{U}$, $\mathbf{V}$.

Algorithm 1: Summary of the two steps involved in the minimization problem (2).

TABLE I: Proposed solution for the two coupled steps of most popular regularized MVA methods.

	$\mathbf{U}$-step (reg. LS)	$\mathbf{W}$-step (eigenvalue problem)
reg. CCA	$\arg\min_{\mathbf{U}} \|\mathbf{Y}' - \mathbf{U}^\top \mathbf{X}\|_F^2 + \gamma R(\mathbf{U})$	$\mathbf{C}_{\mathbf{YY}}^{-\frac{1}{2}} \mathbf{C}_{\mathbf{XY}}^\top \mathbf{U}\mathbf{U}^\top \mathbf{C}_{\mathbf{XY}} \mathbf{C}_{\mathbf{YY}}^{-\frac{1}{2}} \mathbf{V} = \mathbf{V}\mathbf{\Lambda}$
reg. OPLS	$\arg\min_{\mathbf{U}} \|\mathbf{Y}' - \mathbf{U}^\top \mathbf{X}\|_F^2 + \gamma R(\mathbf{U})$	$\mathbf{C}_{\mathbf{XY}}^\top \mathbf{U}\mathbf{U}^\top \mathbf{C}_{\mathbf{XY}} \mathbf{V} = \mathbf{V}\mathbf{\Lambda}$
reg. PCA	$\arg\min_{\mathbf{U}} \|\mathbf{X}' - \mathbf{U}^\top \mathbf{X}\|_F^2 + \gamma R(\mathbf{U})$	$\mathbf{C}_{\mathbf{XX}}^\top \mathbf{U}\mathbf{U}^\top \mathbf{C}_{\mathbf{XX}} \mathbf{V} = \mathbf{V}\mathbf{\Lambda}$

1) $\mathbf{U}$-step. For fixed $\mathbf{V}$, find the matrix $\mathbf{U}$ that minimizes the following regularized least-squares problem,

$$\|\mathbf{Y}' - \mathbf{V}\mathbf{U}^\top \mathbf{X}\|_F^2 + \gamma \|\mathbf{U}\|_{2,1}. \tag{5}$$

   In the next subsection, we will further analyze this minimization problem to obtain an efficient solution that exploits the properties of $\ell_{2,1}$ regularization.
2) $\mathbf{W}$-step. For fixed $\mathbf{U}$, matrix $\mathbf{V}$ is obtained solving the eigenvalue problem (3). As already discussed, existing algorithms solve this step by using orthogonal Procrustes with the undesired consequences described in previous sections.

A. An Efficient Implementation for the $\mathbf{U}$-step

To solve the $\mathbf{U}$-step, we start from the iterative solution proposed in [14], where $\mathbf{U}$ is redefined as $\mathbf{U} = \mathbf{U}'\mathbf{V}$ and $\mathbf{U}'$ is obtained as:

$$\mathbf{U}' = \begin{cases} (\mathbf{C}_{\mathbf{XX}} + \gamma \mathbf{G})^{-1} \mathbf{C}_{\mathbf{XY}'} & \text{if } n < N \\ \mathbf{G}^{-1}\mathbf{X}(\mathbf{X}^\top \mathbf{G}^{-1} \mathbf{X} + \gamma \mathbf{I})^{-1} \mathbf{Y}'^\top & \text{if } n > N, \end{cases} \tag{6}$$

being $\mathbf{G}$ a diagonal matrix, where its ith diagonal element is $G_{ii} = \frac{1}{2\|\boldsymbol{u}^i\|_2}$, $\boldsymbol{u}^i$ is the ith row of $\mathbf{U}$, n is the number of input variables (i.e., the number of rows of $\mathbf{U}$), and N is the number of training data. The straightforward application of this solution would result in a $\mathbf{U}$-step involving two coupled iterative processes: one between $\mathbf{U}$ and $\mathbf{G}$, and other between $\mathbf{U}$ and $\mathbf{V}$ (note that they are coupled by means of matrix $\mathbf{U}$).

However, these processes can be decoupled by taking advantage of the fact that $\mathbf{V}$ is the solution of an eigenvalue problem (i.e., $\mathbf{V}\mathbf{V}^\top = \mathbf{I}$) and rewriting each diagonal term of $\mathbf{G}$ as a function of $\mathbf{U}'$:

$$G_{ii} = \frac{1}{2\|\boldsymbol{u}^i\|_2} = \frac{1}{2\|\boldsymbol{u}'^i \mathbf{V}\|_2} = \frac{1}{2\sqrt{\boldsymbol{u}'\mathbf{V}\mathbf{V}^\top \boldsymbol{u}'^\top}} = \frac{1}{2\|\boldsymbol{u}'^i\|_2}. \tag{7}$$

In this way, the solution of the $\mathbf{U}$-step is independent of matrix $\mathbf{V}$. This result, known in the literature as the rotational invariance property for rows of the $\ell_{2,1}$ norm [14], allows us to follow this simplified procedure:

- Find the optimum $\mathbf{U}'$ by iterating expressions (7) (for $i = 1, \ldots, n_f$) and (6) until a stopping criterion is met.[1]
- Compute $\mathbf{V}$ in a single step by solving:

$$\mathbf{C}_{\mathbf{XY}'}^\top \mathbf{U}' \mathbf{U}'^\top \mathbf{C}_{\mathbf{XY}'} \mathbf{V} = \mathbf{V}\mathbf{\Lambda}^2,$$

  which results from (2) considering that $\mathbf{U} = \mathbf{U}'\mathbf{V}$ and $\mathbf{V}^\top \mathbf{V} = \mathbf{I}$.

In this way, we can obtain important computational savings (as we will analyze in the experimental section). Algorithm 2 summarizes this algorithm. This approach let us formulate $\ell_{2,1}$ based methods such as $\ell_{2,1}$-OPLS and $\ell_{2,1}$-PCA, where $\mathbf{\Omega} = \mathbf{I}$ and the new output matrix is $\mathbf{Y}' = \mathbf{Y}$ ($\ell_{2,1}$-OPLS) or $\mathbf{Y}' = \mathbf{X}$ ($\ell_{2,1}$-PCA); or $\ell_{2,1}$-CCA for $\mathbf{\Omega} = \mathbf{C}_{\mathbf{YY}}^{-1}$ and $\mathbf{Y}' = \mathbf{C}_{\mathbf{YY}}^{-\frac{1}{2}} \mathbf{Y}$.

B. Differences with State of the Art Approaches

In principle, previously proposed L21SDA [17] and SRRR [16] algorithms attempt to solve the same problems as the algorithms $\ell_{2,1}$-CCA and $\ell_{2,1}$-OPLS presented in this paper.

[1] Although other criteria could be considered, we stop the method when $\mathrm{Tr}\{\mathbf{G}^{(k)} - \mathbf{G}^{(k-1)}\} \leq \delta$, where the superscripts denote the iteration index and δ is a small constant, or when a maximum number of iterations have been completed.

1.- Input: $\mathbf{X}$, $\mathbf{Y}$, $\mathbf{\Omega}$, γ.
2.- Optimization algorithm:
2.1.- Initialize $\mathbf{G}^{(0)} = \mathbf{I}$ and $\mathbf{Y}' = \mathbf{\Omega}^{\frac{1}{2}}\mathbf{Y}$.
2.2.- For $k = 1, 2, \ldots$
2.2.1.- $\mathbf{U}'^{(k)} = \begin{cases} (\mathbf{C_{XX}} + \gamma\mathbf{G}^{(k-1)})^{-1}\mathbf{C_{XY'}} & \text{if } n < N \\ \mathbf{G}^{(k-1)^{-1}}\mathbf{X}(\mathbf{X}^\top\mathbf{G}^{(k-1)^{-1}}\mathbf{X} + \gamma\mathbf{I})^{-1}\mathbf{Y}'^\top & \text{if } n > N. \end{cases}$
2.2.2.- $G_{ii}^{(k)} = \dfrac{1}{2\|\boldsymbol{u}'^{i(k)}\|_2}$, for $i = 1, \ldots, n$.
2.2.3.- If the convergence criterion is met, go to 2.3.
2.3.- Solve eigenvalue problem $\mathbf{C}_{\mathbf{XY'}}^\top\mathbf{U}'\mathbf{U}'^\top\mathbf{C_{XY'}}\mathbf{V} = \mathbf{V}\mathbf{\Lambda}^2$.
2.4.- $\mathbf{U} = \mathbf{U}'\mathbf{\Omega}^{\frac{1}{2}}\mathbf{V}$.
3.- Output: $\mathbf{U}$, $\mathbf{V}$.

Algorithm 2: Pseudocode of MVA methods with $\ell_{2,1}$ penalty.

However, due to the procedure followed by their resolution, these state of the art algorithms suffer from the following important inconveniences:

- First, they present the aforementioned drawbacks of all Procrustes based MVA methods.
- Second, they do not exploit the rotational invariance property resulting in considerably larger computational burden in comparison with our proposal. Whereas our proposed solution completely decouples both iterative procedures and gets to reduce them to just one iterative process, where $\mathbf{V}$ can be computed at the end, L21SDA method has to obtain the value of $\mathbf{V}$ inside the iterative procedure. The case of SRRR algorithm is even worse, since it does not merge the two iterative processes, causing a much more expensive solution. The following section will analyze these issues over some real problems.

IV. Experiments

This section analyzes the advantages of the proposed $\ell_{2,1}$-MVA framework from different points of view. To this purpose, we have split it into three subsections so that each one focuses on a different advantage of our proposal. The first subsection shows the advantages of including $\ell_{2,1}$ regularization to provide MVA methods with variable selection capabilities. Then, we will analyze the ability of the MVA approaches and, in particular, the $\ell_{2,1}$-MVA methods for dealing with data that have high multicollinearity among the input variables. This is a difficult situation for MVA methods, since linear dependency among the input variables can cause large fluctuations in the solution. Finally, we will show the importance of avoiding the Procrustes solution by comparing our proposals with state of the art L21SDA and SRRR.

A. Variable Selection by means of $\ell_{2,1}$ Regularization

In this section we are going to deal with a hyperspectral image segmentation and classification problem. This data set comes from a set of hyperspectral sensors mounted on satellite or airborne platform which acquire the reflected energy by the Earth with high spatial detail and in several wavelengths. In particular, we have selected the standard Airborne Visible/Infrared Imaging Spectrometer image taken over Northwest Indianas Indian Pine in June 1992 [28]. This dataset consists of 220 spectral bands, with 20 noisy bands covering the region of water absorption. Discriminating among the major crop classes in the area can be very difficult (in particular, given the moderate spatial resolution of 20 m), making the scene a challenging benchmark to validate classification accuracy of hyperspectral imaging algorithms. Besides, the large number of narrow spectral bands induce a high collinearity among variables, making MVA approaches a powerful tool for this application.

The selected hyperspectral image has 145×145 pixels and contains 17 quite unbalanced classes (ranging from 20 to 10,776 pixels). Among the available 21,025 labeled pixels, 70% were used for training the feature extractors and classifiers, and the remaining 30% were taken apart for testing purposes. The discriminative power of all extracted features was tested using a linear SVM classifier.

To analyze the advantages of the $\ell_{2,1}$ penalty as variable selection tool, we are going to analyze the performance of our proposed regularized MVA framework when different regularizations are used: ridge norm or ℓ_2 penalty, lasso regularization or ℓ_1 norm, and the $\ell_{2,1}$ penalty. For this first study, only OPLS methods have been considered. Fig. 4 shows the reconstructed image or the classification map for these three regularized versions of OPLS, including its overall accuracy (OA) over the test data and the percentage of selected bands ($\%$ band). In this case, regularization parameters, as well as the number of selected variables of the $\ell_{2,1}$-OPLS method, were adjusted using five-fold cross-validation in the training set. In particular, we have explored a rectangular grid taking values from the sets $\{10^{-6}, 5 \cdot 10^{-6}, 10^{-5}, 5 \cdot 10^{-5}, \ldots, 50, 100, 500, 1000\}$ and $\{1, 10, 100, 1000\}$ for γ and the SVM regularization parameter, respectively. We have checked that these intervals are sufficiently large to ensure that the limits were not selected as a result of the CV. After this validation process, all methods show similar accuracy, with ℓ_2 and ℓ_1 OPLS versions achieving an accuracy of 73% using almost all bands (their percentage of selected bands is around 99%). The performance of the $\ell_{2,1}$-OPLS method, is only slightly better with an accuracy of 73.5%, although it uses only 80% of the spectral

bands.

In order to go deeper into the structured sparsity obtained by $\ell_{2,1}$-OPLS, and its variable selection capabilities, Fig. 5 depicts, for the three regularized OPLS versions, the importance of each variable (calculated as $\|\boldsymbol{u}^i\|^2$) over their first three eigenvectors ($k = 1, 2, 3$). The total importance of each variable, given as its averaged importance over all the eigenvectors, is included at the last row of the plot for each algorithm and denoted as "TOT". In this case, to better analyze the sparsity properties of the different regularization terms, the regularization parameter (γ) has been fixed in such a way the both ℓ_1 and $\ell_{2,1}$ norms provide similar number of zeros over all the projection vectors. Furthermore, for comparison purposes, all components with a relevance value lower than 10^{-4} have been drop to zero. As expected, ℓ_1-OPLS is able to nullify some of the eigenvector components and, in a few cases, this provides a variable selection (a 6% of the bands have zero in all their associated components); however, $\ell_{2,1}$-OPLS presents this sparsity in a structured way (all columns are zero), causing 35% of the input bands are not used by their associated eigenvectors. Regarding ℓ_2-OPLS, it is important to remark that it also seems to provide sparsity over its solutions, but this is mainly due to the 10^{-4} threshold applied over the relevant values; even so, it only removes 1% of the input bands. Furthermore, ℓ_2 and ℓ_1 OPLS versions are selecting some positions of the noisy water bands (marked with three shadow rectangles) in the figure, whereas $\ell_{2,1}$-OPLS is able to remove all of them.

Finally, to analyze the influence of the regularization term over the number of selected bands, Fig. 6 displays the total importance of each variable (averaged over all eigenvectors) for $\ell_{2,1}$-OPLS when the parameter γ is varied from 0 to 0.5. When $\gamma = 0$ (lack of regularization), the method is recovering standard OPLS solution, which is quite similar to ℓ_2 versions in terms of sparsity. However, as larger gamma values are considered less bands are used. In particular, values of γ close to 0.25 are enough to remove all water absorption bands.

B. Dealing with Multicollinearity of the Input Variables

The aim of this subsection is to illustrate the advantages of combining the variable selection and feature extraction processes when there exist multicollinearity among the input variables. This is a difficult situation for MVA methods since highly correlated variables can cause large fluctuations in the solution in response to small changes in the model or data.

For this purpose, we compare the proposed methods against the state-of-the-art Robust Feature Selection (RFS) algorithm [14], which is an efficient and an outlier-robust implementation of the least squares problem with an $\ell_{2,1}$ penalization term. We start this study with a toy regression problem and, next, we carry out a similar evaluation over two classification problems with high dimensionality and multicollinearity among their variables. The main characteristics of these problems are summarized in Table II.

TABLE II: Main properties of the datasets: number of training (N_{train}) and test (N_{test}) samples, number of input (n) and output (m) variables, and number of training images per person (p).

	N_{train}/N_{test}	n	m
Carcinomas	139 / 35	9182	11
Yale ($p = 8$)	120 / 45	1024	15

Regression Toy Problem with High Multicollinearity

This toy problem consists of a simple artificial regression problem with three types of input variables: relevant, redundant and noisy. In particular, this problem considers $n = 4000$ random variables where $n_{relev} = 500$ are the relevant ones, which are generated following a Gaussian distribution with zero mean and variance randomly selected from 0 to 4; $n_{redund} = \frac{n}{2} = 2000$ variables can be considered redundant, since they are obtained as a linear combination of the relevant ones; the model also includes $n_{noisy} = 1500$ noisy variables, generated as independent Gaussian variables with zero mean and unit variance. Therefore, defining the observation $\boldsymbol{x} = \left(\boldsymbol{x}_{relev}^\top, \boldsymbol{x}_{redund}^\top, \boldsymbol{x}_{noisy}^\top\right)$ and the output vector $\boldsymbol{y} \in \mathbb{R}^{m\times 1}$, $m = 10$ being the number of output variables, the regression model is given by:

$$\boldsymbol{y} = \begin{pmatrix} \mathbf{W}_{relev} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \boldsymbol{x} + \epsilon,$$

where ϵ is a vector of Gaussian noise with mean 0 and variance 10^{-6}, $\mathbf{W}_{relev} \in \mathbb{R}^{m\times n_{relev}}$ is a fixed matrix with random elements selected from an uniform distribution between -1 and $+1$, and $\mathbf{0}$ is a zero-matrix with the appropriate size. Thus, the regression coefficient matrix is built such that $\boldsymbol{y}$ depends only on the relevant input variables.

Following the above model, we build a set of $N = 500$ training samples and apply a 70/30 (%) partitioning to obtain the training and test sets, respectively. Then, we normalize both sets to zero mean and unitary standard deviation. This process is repeated over 10 random executions, obtaining independent datasets, to average the final results over these runs.

Variable selection is carried out taking the best $n_s < n$ variables after sorting them by relevance according to the corresponding values of $\|\boldsymbol{u}'_i\|$ or $\|\boldsymbol{u}_i\|$ (with $i = 1, \ldots, n$) for RFS and the $\ell_{2,1}$-MVA methods, respectively. Once the n_s variables have been obtained, an optimal Least Squares (LS) regression model is adjusted by using as inputs either the n_s selected original variables (in the case of RFS) or the n_f features extracted from the n_s variables selected by $\ell_{2,1}$-MVA algorithms. The iterative process of $\ell_{2,1}$-MVA methods is stopped when a maximum of 50 iterations are reached or when the Frobenius norm of the difference between the solutions obtained in two consecutive iterations is less than a tolerance value $\delta = 10^{-6}$.

In Fig. 7, MSE obtained by the proposed $\ell_{2,1}$-CCA and $\ell_{2,1}$-OPLS algorithms using all extracted features and the reference algorithm RFS are shown according to the number of selected variables for two values of the penalty parameter: (a) $\gamma = 0.5$ selected by cross-validation, and (b) $\gamma = 100$ selected to illustrate the robustness of $\ell_{2,1}$-OPLS with respect to the selection of this parameter. As can be seen, multicollinearity

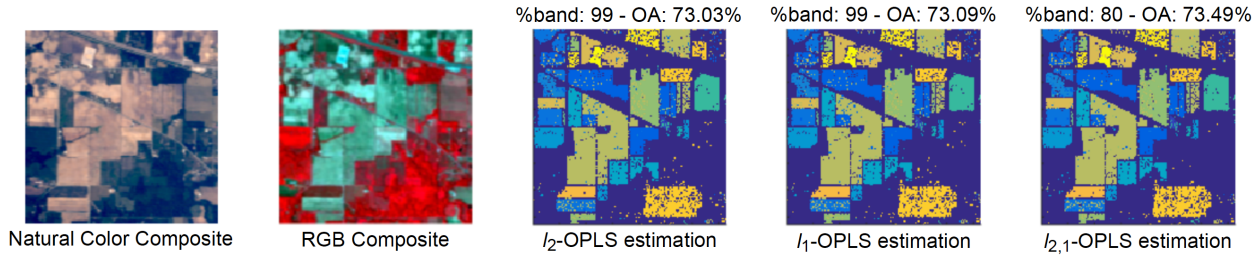


Fig. 4: Natural and RGB composite (by using 50, 27 and 17 channels) hyperspectral images and its reconstructed images using ℓ_2-OPLS, ℓ_1-OPLS and $\ell_{2,1}$-OPLS algorithms.

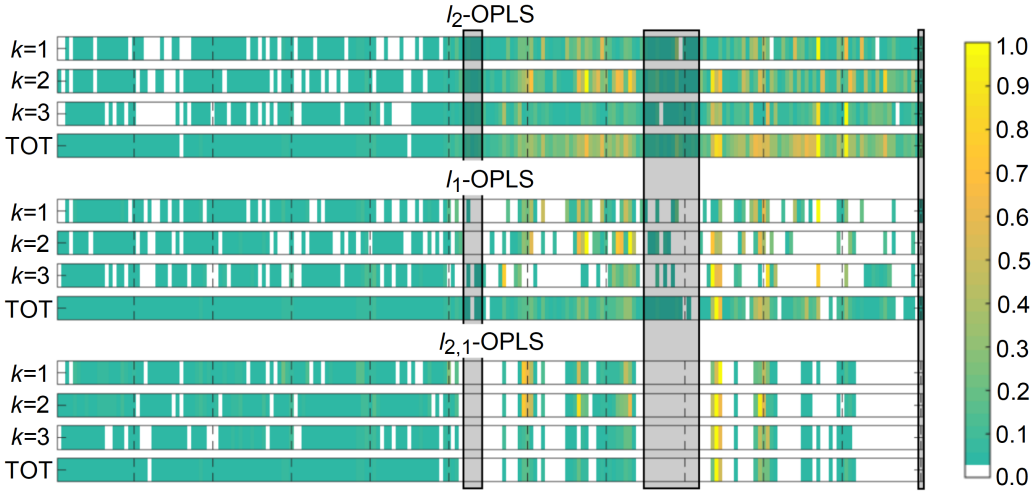


Fig. 5: Analysis of the band selection capabilities for ℓ_2-OPLS, ℓ_1-OPLS and $\ell_{2,1}$-OPLS algorithms. For each algorithm, each row plots the relevance of the components of the first three eigenvectors and the last row includes the average relevance over all eigenvectors (TOT). Water absorption bands, that should be discarded for the classification tasks, have been highlighted in grey color.

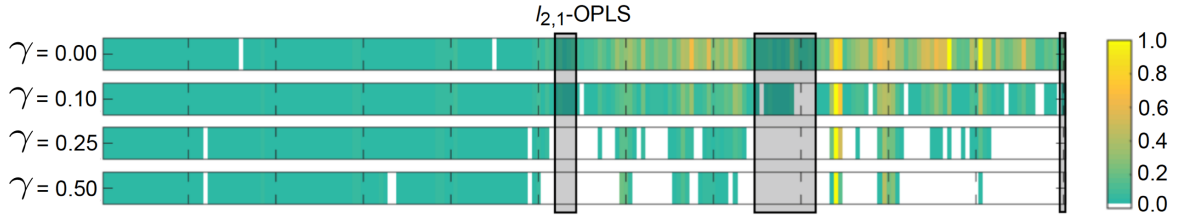


Fig. 6: Analysis of the band selection capabilities for the $\ell_{2,1}$-OPLS method according to penalization term value.

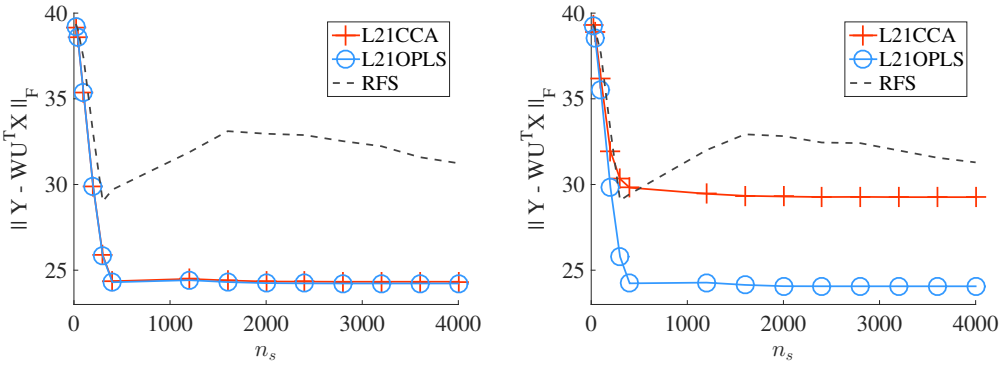
can cause serious problems of overfitting, in fact, this is the case of the RFS method that, although it is a robust method in the presence of outliers, suffers from serious overfitting caused by redundant variables of the problem. On the contrary, MVA methods can successfully deal with such problems. We can see that they improve on RFS performance for $n_s < 500$, and remain stable after that point with no significant degradation. It can be shown that for $n_s = 500$ our methods successfully identify the relevant variables in all cases. In addition to this, $\ell_{2,1}$-MVA extracted features remain mostly orthogonal, as a consequence of the proposed optimization method.

Real World Classification Problems

This subsection analyzes the advantages of the proposed $\ell_{2,1}$-MVA methods, in comparison to RFS method, over two real classification problems with high dimensionality and multicollinearities among their variables: *Carcinomas* and *Yale*.

To make a fair comparison between the methods under study, the free parameter of these models (γ) is selected through a 10 fold cross-validation. For this study, $\ell_{2,1}$-MVA methods use all the extracted features with the corresponding selected variables. The extracted features (both from $\ell_{2,1}$-MVA methods and RFS) are then fed to a linear SVM whose accuracy is used to evaluate the performance of each method. We explored the same ranges of values for γ and SVM regularization parameter C as we did in IV-A.

Fig. 8 displays the overall accuracy (OA) as a function of the number of selected variables (n_s) for $\ell_{2,1}$-OPLS, $\ell_{2,1}$-CCA and RFS. These results corroborate the conclusions derived from the toy problem, that is, multicollinearity among



(a) $\gamma = 0.5$ (b) $\gamma = 100$

Fig. 7: Comparative curves in terms of MSE according to the number of selected variables (n_s) for (a) $\gamma = 0.5$ and (b) $\gamma = 100$.

variables causes RFS to suffer from overfitting problems, which is more evident in the *Carcinomas* dataset; on the contrary, $\ell_{2,1}$-MVA approaches overcome this drawback. It is also interesting to see that $\ell_{2,1}$-OPLS clearly outperforms the other methods. According to the $\ell_{2,1}$-OPLS and $\ell_{2,1}$-CCA curves, one might conclude that all relevant information of the *Carcinomas* dataset lies within 2% of the variables, which is where these algorithms reach their maximum performance.

C. Comparison with L21SDA and SRRR

Whereas the previous subsection compared our approaches with a pure variable selection method, in this subsection, we carry out a comparison against the state of the art approaches based on the Procrustes solution. Remember that SRRR [16] can be seen as an OPLS with $\ell_{2,1}$ penalty, whereas L21SDA [17] is a CCA version including the same penalty term. To the best of our knowledge, no PCA method with $\ell_{2,1}$ penalty has appeared in the literature to date, but its derivation following a Procrustes formulation is straightforward, and is considered here for the sake of completeness.

Here, the experimental procedure is the same as in the previous subsection, but the curves shown below are made based on the number of features extracted instead of the number of selected variables. Fig. 9 shows the OA obtained according to the number of extracted features. When a low number of features is extracted ($n'_f < n_f$), $\ell_{2,1}$-CCA and $\ell_{2,1}$-OPLS clearly outperform L21SDA and SRRR methods. This advantage is due to the ability of the proposed framework to extract a set of mostly uncorrelated features, making easier the training of the subsequent classifier and straightforward the selection of an optimum reduced subset of features.

To analyze in detail this issue, Figs. 10 and 11 show the correlation matrices of he projected data and the discrimination capabilities of these methods for Carcinomas dataset. As expected, the proposed $\ell_{2,1}$-MVA approaches, unlike Procrustes-based schemes, are able to obtain an almost complete uncorrelation among the projected data (note that non-diagonal terms are almost null in our proposed methods). This fact directly provides an improvement of the discrimination capability of new projected features, as Fig. 11 reveals. In this plot, we can check that $\ell_{2,1}$-CCA algorithm is able to project the data into a two dimensional space without overlapping among classes, making easier the subsequent classification task (in this case, the OA is close to 60%); whereas, L21SDA projects most of the classes over the same region and, therefore, the classifier accuracy is reduced by half (OA is around 30%).

Note that, when all the extracted features are used, the results are the same, since the final classifier (SVM) uses all the projected information, which is just a reconstruction from the original space.

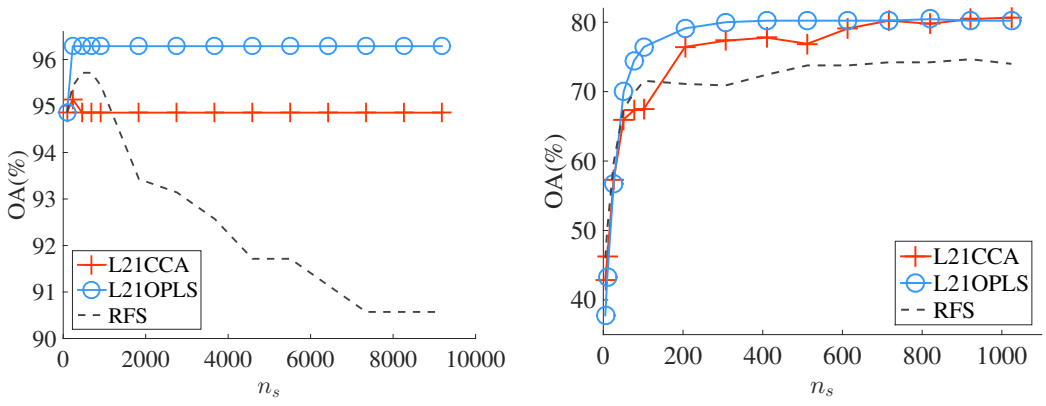
Finally, a comparative study of the computational burden is also shown in Fig. 12. As expected, proposed methods are computationally more efficient, as a direct consequence of exploiting the rotational invariance of the $\ell_{2,1}$ norm, as explained in Subsection III-A.

V. Conclusions

Solutions of regularized MVA approaches are based on an iterative approach consisting of two coupled steps. Whereas the first step eases the inclusion of regularization terms, the second results in a constrained minimization problem which has been typically solved as an orthogonal Procrustes problem. Despite the extended use of this scheme, it fails in obtaining a new subspace of uncorrelated features, this being a desired property of MVA solutions. In this paper we have analyzed the drawbacks of these schemes, recurring to an alternative to the Procrustes solution for the second step, that forces uncorrelation among the extracted features, and thus overcome the drawbacks of previous schemes.

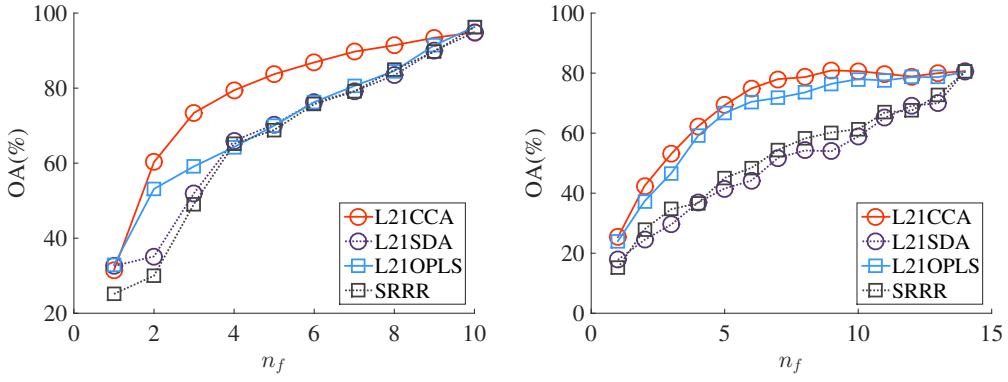
In order to show the practical advantages of our regularized MVA solution, this paper particularizes the proposed method to derive MVA methods implementing $\ell_{2,1}$ regularization. These proposed $\ell_{2,1}$-MVA methods provide an efficient selection of the relevant variables of the problem exploiting the rotational invariance of the $\ell_{2,1}$ norm. At the same time, they can deal with the multicollinearity problems using feature extraction and providing mostly uncorrelated features.

Finally, experimental results over high dimensional problems show that the methods included in this MVA framework



(a) *Carcinomas* (b) *Yale*

Fig. 8: Comparative curves in terms of overall accuracy as a function of the number of selected variables (n_s).



(a) *Carcinomas* (b) *Yale*

Fig. 9: Comparative curves in terms of OA according to the number of extracted features (n_f) for $\ell_{2,1}$-CCA, $\ell_{2,1}$-OPLS and reference methods L21SDA and SRRR.

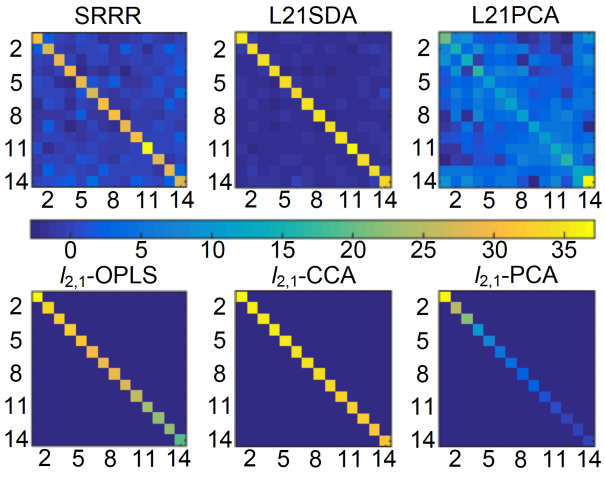


Fig. 10: Comparison among the feature correlation matrices of our proposed methods and those proposed in the literature, which use Procrustes approach for the *Carcinomas* dataset.

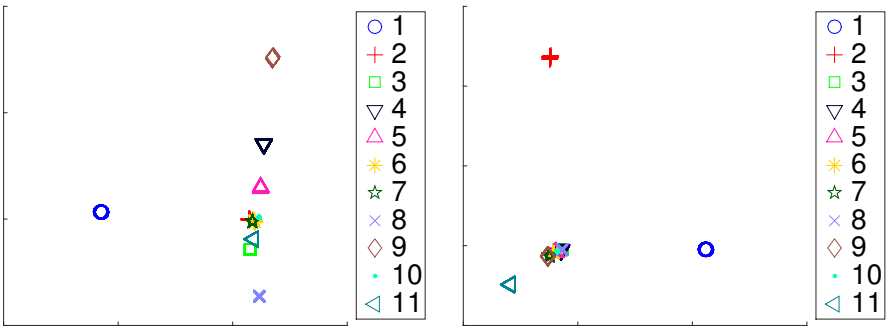
are not only computationally more efficient than previous state of the art solutions, but also can improve their performance.

ACKNOWLEDGMENT

This work has been partly supported by MINECO projects TEC2013-48439-C4-1-R, TEC2014-52289-R and TEC2016-75161-C2-2-R, and Comunidad de Madrid projects PRICAM P2013/ICE-2933 and S2013/ICE-2933.

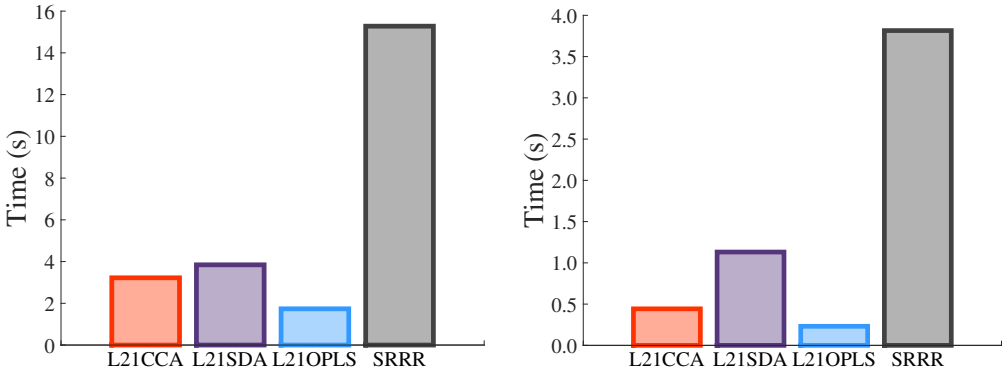
REFERENCES

[1] K. Pearson, "On lines and planes of closest fit to systems of points in space," *Philosophical Magazine*, vol. 2, no. 6, pp. 559–572, 1901.
[2] M. Turk and A. Pentland, "Eigenfaces for recognition," *Journal of cognitive neuroscience*, vol. 3, no. 1, pp. 71–86, 1991.
[3] H. Hotelling, "Relations between two sets of variates," *Biometrika*, vol. 28, pp. 321–377, 1936.
[4] H. Wold, "Estimation of principal components and related models by iterative least squares," in *Multivariate Analysis*. Academic Press, 1966, pp. 391–420.
[5] ——, "Non-linear estimation by iterative least squares procedures," in *Research Papers in Statistics*. Wiley, 1966, pp. 411–444.
[6] K. Worsley, J. Poline, K. Friston, and A. Evans., "Characterizing the response of PET and fMRI data using multivariate linear models (MLM)," *Neuroimage*, vol. 6, pp. 305–319, 1998.
[7] J. Arenas-García, K. B. Petersen, G. Camps-Valls, and L. K. Hansen, "Kernel multivariate analysis framework for supervised subspace learning: A tutorial on linear and kernel multivariate methods," *IEEE Signal Processing Magazine*, vol. 30, no. 4, pp. 16–29, July 2013.
[8] M. A. J. van Gerven, Z. C. Chao, and T. Heskes, "On the decoding of intracranial data using sparse orthonormalized partial least squares," *Journal of Neural Engineering*, vol. 9, no. 2, pp. 26 017–26 027, 2012.



(a) L21CCA (b) L21SDA

Fig. 11: Discriminative power comparison between $\ell_{2,1}$-CCA and L21SDA when only using the two first extracted features in *Carcinomas* dataset.



(a) *Carcinomas* (b) *Yale*

Fig. 12: Training time (in seconds) required by $\ell_{2,1}$-CCA, $\ell_{2,1}$-OPLS, L21SDA and SRRR algorithms.

[9] L. K. Hansen, "Multivariate strategies in functional magnetic resonance imaging," *Brain and Language*, vol. 102, no. 2, pp. 186–191, 2007.

[10] J. Arenas-García and G. Camps-Valls, "Efficient kernel orthonormalized PLS for remote sensing applications," *IEEE Trans. Geosci. Remote Sens.*, vol. 44, pp. 2872–2881, 2008.

[11] J. Arenas-García and K. B. Petersen, "Kernel multivariate analysis in remote sensing feature extraction," in *Kernel Methods for Remote Sensing Data Analysis*, G. Camps-Valls and L. Bruzzone, Eds. Wiley, 2009.

[12] M. Barker and W. Rayens, "Partial least squares for discrimination," *Journal of Chemometrics*, vol. 17, no. 3, pp. 166–173, 2003.

[13] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society, Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[14] F. Nie, H. Huang, X. Cai, and C. Ding, "Efficient and robust feature selection via joint $\ell_{2,1}$-norms minimization," in *Advances in Neural Information Processing Systems 23*. The MIT Press, 2010, pp. 1813–1821.

[15] H. Zou, T. Hastie, and R. Tibshirani, "Sparse principal component analysis," *Journal of Computational and Graphical Statistics*, vol. 15, no. 2, pp. 265–286, 2006.

[16] L. Chen and J. Z. Huang, "Sparse reduced-rank regression for simultaneous dimension reduction and variable selection," *Journal of the American Statistical Association*, vol. 107, no. 500, pp. 1533–1545, 2012.

[17] X. Shi, Y. Yang, Z. Guo, and Z. Lai, "Face recognition by sparse discriminant analysis via joint $L_{2,1}$-norm minimization," *Pattern Recognition*, vol. 47, no. 7, pp. 2447–2453, 2014.

[18] P. H. Schönemann, "A generalized solution of the orthogonal procrustes problem," *Psychometrika*, vol. 31, no. 1, pp. 1–10, 1966.

[19] Z. Lai, W. K. Wong, Y. Xu, J. Yang, and D. Zhang, "Approximate orthogonal sparse embedding for dimensionality reduction," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 27, no. 4, pp. 723–735, April 2016.

[20] Z. Hu, G. Pan, Y. Wang, and Z. Wu, "Sparse principal component analysis via rotation and truncation," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 27, no. 4, pp. 875–890, April 2016.

[21] S. Muñoz-Romero, V. Gómez-Verdejo, and J. Arenas-García, "Why (and how) avoid orthogonal procrustes in regularized multivariate analysis," *arXiv preprint, arXiv:submit/1555588*, 2016.

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

[23] G. C. Reinsel and R. P. Velu, *Multivariate Reduced-Rank Regression: Theory and Applications*. Springer New York, 1998.

[24] S. Muñoz-Romero, J. Arenas-García, and V. Gómez-Verdejo, "Sparse and kernel OPLS feature extraction based on eigenvalue problem solving," *Pattern Recognition*, vol. 48, no. 5, pp. 1797 – 1811, 2015.

[25] S. Roweis and C. Brody, "Linear heteroencoders," Gatsby Computational Neuroscience Unit, Tech. Rep. 1999-002, 1999.

[26] M. Grant and S. Boyd, "CVX: Matlab software for disciplined convex programming, version 2.1," http://cvxr.com/cvx, Mar. 2014.

[27] J. Kim and H. Park, "Toward faster nonnegative matrix factorization: A new algorithm and comparisons," in *Proc. 8th IEEE Intl. Conf. on Data Mining (ICDM'08)*. Pisa, Italy: IEEE, December 2008, pp. 353–362.

[28] (2013) Purdue Univ. Web site. College of Engineering. [Online]. Available: http://dynamo.ecn.purdue.edu/~biehl/MultiSpec