ITERATIVE ORTHONORMALIZED PARTIAL LEAST SQUARES WITH SPARSITY CONSTRAINTS

Sergio Muñoz-Romero, Jerónimo Arenas-García, and Vanessa Gómez-Verdejo

Univ. Carlos III de Madrid, Spain,
{smunoz,vanessa,jarenas}@tsc.uc3m.es

ABSTRACT

Orthonormalized partial least squares (OPLS) is a popular multivariate analysis method to perform supervised feature extraction. In this paper, we propose a novel scheme to solve Orthonormalized Partial Least Squares (OPLS) that can be easily modified to include additional constraints over the input data projection vectors. This scheme is used to implement an OPLS method with sparsity constraints (SOPLS), which allows to obtain more interpretable projection vectors that depend only on a few of the original input variables. The discriminative power of the sparse features extracted by SOPLS is analyzed on a benchmark of classification problems, where the method shows very competitive performance in terms of classification error.

***Index Terms*—** Partial least squares (PLS), orthonormalized PLS, sparse solutions, *lasso* regularization, feature extraction

1. INTRODUCTION

Multivariate analysis (MVA) methods [1] are extensively used in many different areas such as machine learning [2], biomedical engineering [3, 4], remote sensing [5, 6], or chemometrics [7], among others. The need for performing feature extraction and dimensionality reduction is especially important when dealing with high-dimensional data and/or collinearity among variables. Furthermore, it could be useful to remove irrelevant or noisy features, providing more interpretable solutions. One way to do so is by enforcing sparse solutions, which justifies the large number of research papers on that topic during the last years [8, 9].

Here, we focus on an MVA method known as Orthonormalized Partial Least Squares (OPLS) which is known to be optimum in the mean square error sense for performing multilinear regression [10, 11]. This method and its kernel counterpart have shown to be very competitive also as a pre-processing step in classification problems [11, 12]. Several recent works have also tried to establish the connections between OPLS and other MVA and discriminative methods (see, e.g., [13]). In an attempt to improve the interpretability of the solution, a sparse OPLS method was proposed in [3]. Unfortunately, this method does not guarantee orthogonality of the projected input data, and thus it does not converge to the true OPLS solution, even when the sparsity constraints are removed.

In this paper, we propose a novel formulation for OPLS which allows an iterative implementation involving two steps: the solution to an eigenvalue decomposition and a standard least squares problem. The latter step can easily be modified to include additional constraints. To illustrate the flexibility of the method, we use it to implement a novel OPLS algorithm with sparsity constraints on the projection vectors (SOPLS). We present both a block method where all projection vectors are extracted at once, and a sequential procedure that extracts projection vectors one at a time.

The discriminative power of the sparse features extracted by SOPLS is analyzed on a benchmark classification problems, showing very competitive performance in terms of classification error. The degree of sparsity of the achieved solutions is also illustrated.

This work was partly supported by MICINN projects TEC2011-22480 and PRI-PIBIN-2011-1266.

2. OPLS FORMULATION

This section reviews the most commonly used formulation of OPLS and, as an alternative to it, presents a novel method to obtain the OPLS solution whose main advantage is that it can be easily modified to include additional constraints. Before that, we briefly review the notation that will be used in the description of the methods.

Let us assume a supervised learning scenario, where the goal is to learn relevant features from input data using a set of N training data $\{\boldsymbol{x}_i, \boldsymbol{y}_i\}$, for $i = 1, \ldots, N$, where $\boldsymbol{x}_i \in \Re^n$ and $\boldsymbol{y}_i \in \Re^m$ are considered as the input and output vectors, respectively. Therefore, n and m denote the dimension of the input and output spaces. In classification problems, y_i will be used to denote the class membership of the ith pattern, e.g., using 1-of-C encoding [14].

For notational convenience, we define the input and output data matrices: $\mathbf{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N]$ and $\mathbf{Y} = [\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N]$. It will be assumed throughout the paper that these matrices are centered to remove any correlation between variables produced by a shift of their centers of mass [2]. Sample estimation of the input and output data covariance matrices, as well as 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 a scaling factor N, and superscript $\top$ denotes vector or matrix transposition.

Input data features are calculated as $\mathbf{X}' = \mathbf{U}^\top\mathbf{X}$, where $\mathbf{U} = [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_{n_p}]$ is a projection matrix where projection vectors are arranged columnwise, and $n_p < n$ is the number of projections. The goal of OPLS is to find the projection vectors so that the projected data best approximate the output data in a mean square error sense (MSE); i.e., OPLS minimizes the following loss function [10],

$$\mathcal{L}(\mathbf{W}, \mathbf{U}) = \|\mathbf{Y} - \mathbf{W}\mathbf{U}^\top\mathbf{X}\|_F^2, \tag{1}$$

where $\mathbf{W}$ is an $m \times n_p$ matrix of regression coefficients that can alternatively be seen as the projection matrix of the output data, $\|\mathbf{A}\|_F = \mathrm{Tr}\{\mathbf{A}\mathbf{A}^T\}$ denotes the Frobenius norm of $\mathbf{A}$, and $\mathrm{Tr}\{\cdot\}$ is the trace operator. Note that the above problem is different from standard least squares regression since matrix $\mathbf{U}$ imposes a representation bottleneck [10, 13]. Note also that the solution to (1) is not unique since, e.g., $\mathbf{W}$ can compensate any scaling of matrix $\mathbf{U}$. In the next two subsections we will pay attention to two different constraints that can be used to make OPLS solution unique.

2.1. OPLS as a generalized eigenvalue decomposition problem

Rewriting (1) as

$$\mathcal{L} = \mathrm{Tr}\{\mathbf{C_{YY}}\} - 2\,\mathrm{Tr}\{\mathbf{W}^\top \mathbf{C}_{\mathbf{XY}}^\top \mathbf{U}\} + \mathrm{Tr}\{\mathbf{U}^\top \mathbf{C_{XX}} \mathbf{U} \mathbf{W}^\top \mathbf{W}\} \tag{2}$$

suggests that $\mathbf{U}^T \mathbf{C_{XX}} \mathbf{U} = \mathbf{I}$ can be used to simplify the problem and to make the minimum of $\mathcal{L}$ unique. In fact, these uncorrelation and unit variance conditions on the projected data have been the most frequently used assumptions in the literature to solve OPLS [10, 13]. If we further note that the optimal $\mathbf{W}$ is uniquely determined for fixed $\mathbf{U}$ as the solution to the LS problem stated in (1)

$$\mathbf{W} = \mathbf{C}_{\mathbf{XY}}^\top \mathbf{U} (\mathbf{U}^\top \mathbf{C_{XX}} \mathbf{U})^{-1} = \mathbf{C}_{\mathbf{XY}}^\top \mathbf{U},$$

it is possible to formulate the OPLS problem as a constrained minimization problem over $\mathbf{U}$ only:

$$\begin{aligned} \mathbf{U}_{\text{GEV}} = \quad & \arg\min_{\mathbf{U}} \mathcal{L}(\mathbf{W}, \mathbf{U}) \Big|_{\mathbf{W} = \mathbf{C}_{\mathbf{XY}}^\top \mathbf{U}} \\ \text{s.t.:} \quad & \mathbf{U}^\top \mathbf{C_{XX}} \mathbf{U} = \mathbf{I} \end{aligned} \tag{3}$$

Once the optimum projection matrix has been calculated, the corresponding regression matrix will be given by $\mathbf{W}_{\text{GEV}} = \mathbf{C}_{\mathbf{XY}}^\top \mathbf{U}_{\text{GEV}}$. Here, subscript 'GEV' has been used since, as we show below, the solution of $\mathbf{U}$ can be given in terms of a generalized eigenvalue problem.

Inserting the constraints indicated in (3) into (2), we see that OPLS can be written down as the solution to the following maximization problem:

$$\begin{aligned} \mathbf{U}_{\text{GEV}} = \quad & \arg\max_{\mathbf{U}} \mathrm{Tr}\{\mathbf{U}^\top \mathbf{C_{XY}} \mathbf{C}_{\mathbf{XY}}^\top \mathbf{U}\} \\ \text{s.t.:} \quad & \mathbf{U}^\top \mathbf{C_{XX}} \mathbf{U} = \mathbf{I} \end{aligned} \tag{4}$$

which is given in terms of the following generalized eigenvalue decomposition problem:

$$\mathbf{C_{XY}} \mathbf{C}_{\mathbf{XY}}^\top \mathbf{U}_{\text{GEV}} = \mathbf{C_{XX}} \mathbf{U}_{\text{GEV}} \mathbf{\Lambda}_{\text{GEV}} \tag{5}$$

A further interesting property of the above solution to OPLS can be obtained if we premultiply both terms of (5) by $\mathbf{U}_{\text{GEV}}^\top$. Since $\mathbf{W}_{\text{GEV}} = \mathbf{C}_{\mathbf{XY}}^\top \mathbf{U}_{\text{GEV}}$ and $\mathbf{U}_{\text{GEV}}^\top \mathbf{C_{XX}} \mathbf{U}_{\text{GEV}} = \mathbf{I}$, it is straightforward to conclude that $\mathbf{W}_{\text{GEV}}^\top \mathbf{W}_{\text{GEV}} = \mathbf{\Lambda}_{\text{GEV}}$, i.e., the columns of $\mathbf{W}_{\text{GEV}}$ are orthogonal.

2.2. OPLS as an eigenvalue decomposition problem

In this subsection, we propose a novel solution to the OPLS problem that opens the door to more efficient implementations, and to modified versions that can easily incorporate constraints into the projection vectors $\boldsymbol{u}_i$. The above orthogonality condition over the regression coefficient vectors and a detailed examination of (2) suggest that assuming $\mathbf{W}^\top \mathbf{W} = \mathbf{I}$ can also be convenient to make the solution of OPLS unique. When $\mathbf{W}$ is given, it can be seen that the projection matrix that minimizes (1) is

$$\mathbf{U} = \mathbf{C}_{\mathbf{XX}}^{-1} \mathbf{C_{XY}} \mathbf{W}. \tag{6}$$

We can now formulate OPLS as a constrained minimization problem over $\mathbf{W}$ only:

$$\begin{aligned} \mathbf{W}_{\text{EVD}} = \quad & \arg\min_{\mathbf{W}} \mathcal{L}(\mathbf{W}, \mathbf{U}) \Big|_{\mathbf{U} = \mathbf{C}_{\mathbf{XX}}^{-1} \mathbf{C_{XY}} \mathbf{W}} \\ \text{s.t.:} \quad & \mathbf{W}^\top \mathbf{W} = \mathbf{I} \end{aligned} \tag{7}$$

where the subscript 'EVD' has been used to denote this solution since, as we show next, it can be obtained from the eigenvalue decomposition of a certain matrix.

Inserting the orthonormality constraint on $\mathbf{W}$ and the expression for $\mathbf{U}$ as a function of $\mathbf{W}$ into (2), we can rewrite (7) as

$$\begin{aligned} \mathbf{W}_{\text{EVD}} = \quad & \arg\max_{\mathbf{W}} \mathrm{Tr}\{\mathbf{W}^\top \mathbf{C}_{\mathbf{XY}}^\top \mathbf{C}_{\mathbf{XX}}^{-1} \mathbf{C_{XY}} \mathbf{W}\} \\ \text{s.t.:} \quad & \mathbf{W}^\top \mathbf{W} = \mathbf{I} \end{aligned} \tag{8}$$

Therefore, the columns of $\mathbf{W}_{\text{EVD}}$ can be obtained as the n_p leading eigenvectors of the following problem:

$$\mathbf{C}_{\mathbf{XY}}^\top \mathbf{C}_{\mathbf{XX}}^{-1} \mathbf{C_{XY}} \mathbf{W}_{\text{EVD}} = \mathbf{W}_{\text{EVD}} \mathbf{\Lambda}_{\text{EVD}}. \tag{9}$$

If we compare (9) and (5), we can observe a first advantage of the proposed OPLS solution: the dimension of the eigenvalue problems are m and n, respectively, meaning that the new solution is computationally more efficient for the common case $m < n$.

An important property for the input projected data can be derived by first premultiplying both terms of (9) by $\mathbf{W}_{\text{EVD}}^\top$, which leads to

$$\mathbf{W}_{\text{EVD}}^\top \mathbf{C}_{\mathbf{XY}}^\top \mathbf{U}_{\text{EVD}} = \mathbf{\Lambda}_{\text{EVD}}$$

where in simplifying we have used the fact that the columns of $\mathbf{W}_{\text{EVD}}$ are orthonormal ($\mathbf{W}_{\text{EVD}}^\top \mathbf{W}_{\text{EVD}} = \mathbf{I}$). If we further note that according to (6) $\mathbf{C_{XY}} \mathbf{W}_{\text{EVD}} = \mathbf{C_{XX}} \mathbf{U}_{\text{EVD}}$, we arrive at

$$\mathbf{U}_{\text{EVD}}^\top \mathbf{C_{XX}} \mathbf{U}_{\text{EVD}} = \mathbf{\Lambda}_{\text{EVD}}.$$

In other words, the OPLS solution guarantees orthogonality of the projected input data.

Finally, it is easy to see that the solutions to (7) and (3) should both provide the same value of the cost function $\mathcal{L}(\mathbf{W}, \mathbf{U})$. This means that the columns of $\mathbf{U}_{\text{GEV}}$ and $\mathbf{U}_{\text{EVD}}$ should span the same subspace. In fact, it can be shown that the ith columns of $\mathbf{U}_{\text{GEV}}$ and $\mathbf{U}_{\text{EVD}}$ have the same direction, and differ only in a scaling factor. More explicitly, it can be proved with some simple algebraic manipulations that the following relationship exists between the two OPLS solutions derived in this section: $\mathbf{\Lambda}_{\text{EVD}} = \mathbf{\Lambda}_{\text{GEV}} = \mathbf{\Lambda}$, $\mathbf{U}_{\text{EVD}} = \mathbf{U}_{\text{GEV}} \mathbf{\Lambda}^{1/2}$, and $\mathbf{W}_{\text{EVD}} = \mathbf{W}_{\text{GEV}} \mathbf{\Lambda}^{-1/2}$.

3. ITERATIVE OPLS SOLUTION

In this section, we propose a novel iterative scheme for OPLS which is based on two coupled steps in which $\mathbf{W}$ and $\mathbf{U}$ are updated by means of an eigenvalue decomposition and a least squares (LS) problem.

To start with, let us introduce (6) into the left-hand-side term of (9) and multiply the resulting expression by its transpose,

$$\mathbf{C}_{\mathbf{XY}}^\top \mathbf{U}_{\text{EVD}} \mathbf{U}_{\text{EVD}}^\top \mathbf{C_{XY}} = \mathbf{W}_{\text{EVD}} \mathbf{\Lambda}^2 \mathbf{W}_{\text{EVD}}^\top$$

If we further postmultiply both terms by $\mathbf{W}_{\text{EVD}}$, we get an alternative eigenvalue decomposition problem that has to be satisfied by $\mathbf{W}_{\text{EVD}}$:

$$\mathbf{C}_{\mathbf{XY}}^\top \mathbf{U}_{\text{EVD}} \mathbf{U}_{\text{EVD}}^\top \mathbf{C_{XY}} \mathbf{W}_{\text{EVD}} = \mathbf{W}_{\text{EVD}} \mathbf{\Lambda}' \tag{10}$$

where $\mathbf{\Lambda}' = \mathbf{\Lambda}^2$. Furthermore, (6) provides another expression relating the optimum OPLS solutions for the projection and regression matrices.

The proposed iterative solution consists in initializing $\mathbf{U}$ and $\mathbf{W}$, and then repeatedly apply (10) and (6) until some convergence criterion is met. Initialization of the algorithm is not critical, and we

1. $\mathbf{C_{XX}} = \mathbf{X}\mathbf{X}^\top$, $\mathbf{C_{XY}} = \mathbf{X}\mathbf{Y}^\top$. Compute $\mathbf{C}_{\mathbf{XX}}^{-1}$.
2. Initialize $\mathbf{U}(0) = \mathbf{I} \in \Re^{n \times n_p}$. $k = 0$.
3. $k = k + 1$.
4. Obtain $\mathbf{W}(k)$ as the solution to the EVD problem stated in (10).
5. Update $\mathbf{U}$ using (6): $\mathbf{U}(k) = \mathbf{C}_{\mathbf{XX}}^{-1}\mathbf{C_{XY}}\mathbf{W}(k)$.
6. Repeat 3–5 until convergence criterion is met.

Table 1. Pseudocode for iterative method for solving OPLS.

will simply initialize $\mathbf{U}$ as the identity matrix of size $n \times n_p$. Table 1 summarizes the main steps of the proposed algorithm.

The main advantage of this iterative procedure with respect to the original method in Subsection 2.2 is that the second step [i.e., the update of $\mathbf{U}$ using (6)] can be reinterpreted as the solution to the following LS problem

$$\mathbf{U}_{\text{EVD}} = \arg\min_{\mathbf{U}} \|\mathbf{W}_{\text{EVD}}^\top \mathbf{Y} - \mathbf{U}^\top \mathbf{X}\|_F^2. \tag{11}$$

In this way, we could easily modify this step incorporating additional constraints to the LS problem (11). In the next section we will add sparsity constraints, thus allowing us to obtain sparse projection vectors $\boldsymbol{u}_i$.

A somewhat similar scheme can be found in [3] implementing orthogonal Procrustes solution. Our scheme and the one in [3] differ in the $\mathbf{W}$-update step. However, it can be shown that the Procrustes solution does not provide orthogonality in the projected input data, preventing it from convergence to the true OPLS solution.

3.1. Sequential OPLS with deflation

In many cases, it is preferred to obtain the projection vectors one by one. When doing so, the data matrices need to be deflated after each new projection vector has been obtained (see, e.g. [2]). In our case, it suffices to deflate the cross-covariance matrix according to

$$\mathbf{C_{XY}} \leftarrow \mathbf{C_{XY}} - \mathbf{C_{XX}}\boldsymbol{u}_i\boldsymbol{w}_i^\top, \tag{12}$$

where $\boldsymbol{u}_i$ and $\boldsymbol{w}_i$ are, respectively, the ith projection and regression vectors provided by the iterative OPLS scheme.

This deflation scheme can be better understood if we notice that it is equivalent to the following deflation of the output data matrix:

$$\mathbf{Y} \leftarrow \mathbf{Y} - \boldsymbol{w}_i\boldsymbol{u}_i^\top\mathbf{X}.$$

In other words, we remove from $\mathbf{Y}$ the best approximation that can be achieved with the ith projection of the input data, $\boldsymbol{u}_i^\top\mathbf{X}$.

When we compute one projection vector at a time, the two steps of the iterative optimization procedure are significantly simplified:

- The eigenvalue decomposition problem (10) becomes

  $$\mathbf{C}_{\mathbf{XY}}^\top\boldsymbol{u}_i\boldsymbol{u}_i^\top\mathbf{C_{XY}}\boldsymbol{w}_i = \boldsymbol{w}_i\boldsymbol{\lambda}_i',$$

  and it is straightforward to check that the solution is given by

  $$\boldsymbol{w}_i = \frac{\mathbf{C}_{\mathbf{XY}}^\top\boldsymbol{u}_i}{\|\mathbf{C}_{\mathbf{XY}}^\top\boldsymbol{u}_i\|_2}, \tag{13}$$

  where $\|\cdot\|_2$ represents the Euclidean norm of a vector.
- The LS problem (11) becomes

  $$\boldsymbol{u}_i = \arg\min_{\boldsymbol{u}} \|\boldsymbol{w}_i^\top\mathbf{Y} - \boldsymbol{u}^\top\mathbf{X}\|_2^2$$

  and the solution is given by

  $$\boldsymbol{u}_i = \mathbf{C}_{\mathbf{XX}}^{-1}\mathbf{C_{XY}}\boldsymbol{w}_i. \tag{14}$$

	$N_{\text{train}}/N_{\text{test}}$	n	m
letter	10000 / 10000	16	26
optdigits	3823 / 1797	64	10
pendigits	7494 / 3498	16	10
satellite	4435 / 2000	36	6
segment	1310 / 1000	18	7
vehicle	500 / 346	18	4

Table 2. Main properties of the selected benchmark problems.

Therefore, when we wish to obtain the projection vectors in a sequential manner (i.e., one at a time), we will simply iterate (13) and (14) until some convergence criterion is satisfied. The next projection vector can then be obtained in a similar manner, once the cross-covariance matrix $\mathbf{C_{XY}}$ has been deflated according to (12).

4. SPARSE OPLS

In this section we explain how the iterative methods for OPLS we have just described can be modified to incorporate additional constraints over the projection vectors; in particular, we will consider the introduction of sparsity constraints, both for the block case and for the sequential scheme using deflation.

It is well-known that adding L_1 regularization favors sparse solutions, making coefficients associated with irrelevant variables zero. When L_1 constraints are considered for the projection vectors, we just need to modify the corresponding cost functions which are minimized in the corresponding step of the block or sequential schemes:

$$\text{Block: } \mathbf{U}_{L_1} = \arg\min_{\mathbf{U}} \|\mathbf{W}^\top\mathbf{Y} - \mathbf{U}^\top\mathbf{X}\|_F^2 + \lambda_1|\mathbf{U}| \tag{15}$$

$$\text{Seq.: } \boldsymbol{u}_{i,L_1} = \arg\min_{\boldsymbol{u}} \|\boldsymbol{w}^\top\mathbf{Y} - \boldsymbol{u}^\top\mathbf{X}\|_2^2 + \lambda_1\|\boldsymbol{u}\|_1 \tag{16}$$

where $\|\boldsymbol{u}\|_1$ and $|\mathbf{U}|$ denote the sum of the absolute values of the components of vector $\boldsymbol{u}$ and matrix $\mathbf{U}$, respectively, and λ_1 is the regularization parameter. We refer the reader to [15, 16] for good reviews of optimization methods to solve the above problems.

5. EXPERIMENTS

This section analyzes the capability of the proposed SOPLS approach to extract a subset of informative and discriminative features by means of a sparse combination of the original ones. For this analysis, we have selected six multi-class classification problems from [17]. Table 2 summarizes their main characteristics.

We will study the performance of the proposed SOPLS method against the EVD-OPLS (hereafter referred as OPLS) and the sparse OPLS algorithm proposed in [3]. The latter uses the Procrustes problem solution to compute $\mathbf{W}$ in the iterative SOPLS algorithm, for this reason, we will denote is as P-SOPLS (Procrustes Sparse OPLS).

Regarding implementation details, OPLS method follows the steps described in Eq. (6-9) and P-SOPLS follows the procedure described in [3]. The proposed SOPLS approach uses its sequential formulation (Eq. (13-16)), initializing matrix $\mathbf{W}$ as the identity matrix and stopping its iterative process when either the cosine distance,

$$d(\boldsymbol{u}_i(k), \boldsymbol{u}_i(k-1)) = \frac{\boldsymbol{u}_i^\top(k)\boldsymbol{u}_i(k-1)}{\|\boldsymbol{u}_i(k)\|\|\boldsymbol{u}_i(k-1)\|},$$

	OPLS	P-SOPLS		SOPLS	
	OA(%)	OA(%)	SR(%)	OA(%)	SR(%)
letter	84,89	84,85	11,33	**85,05**	10,94
optdigits	94,21	94,27	42,47	**95,05**	29,93
pendigits	92,08	91,68	39,58	**92,22**	43,06
satellite	85,7	85,90	17,22	**86,10**	27,22
segment	92,8	**95,60**	90,74	94,90	93,52
vehicle	78,32	77,17	25,93	**78,03**	1,85

Table 3. Overall accuracy (OA) achieved by OPLS, P-SOPLS and SOPLS algorithms. Sparsity rates (SR) of P-SOPLS and SOPLS also are included.

achieves a tolerance level of $\delta = 10^{-12}$ or 500 iterations have been completed. The regularization parameter, λ_1, used by SOPLS and P-SOPLS approaches has been adjusted by a 10-fold Cross Validation (CV) process selecting its value from the set $\{0, 10^{-4}, 10^{-3.9}, \ldots, 10^{-1.1}, 10^{-1}\}$.

To test the discrimination capability of the set of features provided for each feature extraction approach, a linear support vector machine (SVM) has been trained using as inputs these new features and selecting parameter C among the set of values $\{1, 10, 100, 1000\}$ with a 10-fold CV. In this paper, we use the LIBSVM implementation [18].

Table 3 shows the overall accuracy (OA) provided by these three feature selection techniques when the maximum number of projections ($r = \text{rank}\{\mathbf{C_{YX}}\}$) is used to train the SVM. In P-SOPLS and SOPLS methods, the sparsity rate (SR) of the projections vectors, defined as the ratio between the number of zero elements in $\mathbf{U}$ and the total number of entries, is also included.

It is important to note that problem *segment* is ill-conditioned ($\text{rank}\{\mathbf{C_{XX}}\} < n$) preventing the application of OPLS; for this reason, PCA has been applied as preprocessing step to reduce the input data dimension to $\text{rank}\{\mathbf{C_{XX}}\}$, after which OPLS algorithm can be applied. This was not necessary for the sparse approaches (P-SOPLS and SOPLS) since the included L_1 regularizer makes possible to solve ill-conditioned problems without any preprocessing step.

Table 3 shows the advantages, in terms of accuracy, of the proposed SOPLS method against OPLS and P-SOPLS. When SOPLS features are used to train the SVM, OPLS is outperformed in all the datasets, whereas it improves the P-SOPLS method in terms of OA for five out of the six problems.

Apart from its increased discrimination capability, the main advantage of the proposed SOPLS method relies on its sparse formulation that makes it easier to analyze which features do not contribute to the new projected ones. To carry out this analysis, Figure 1 depicts the projection matrices, $\mathbf{U}$, obtained by OPLS, SOPLS, and P-SOPLS solutions in three representative problems. Looking at these figures, one can see that in problems presenting a high SR, such as *segment*, the feature extraction becomes close to feature selection, since most features are associated with just one of the original ones. In *satellite*, features 8, 31, 32 and 36 are removed from the first projection vectors (the most important ones) of the SOPLS algorithm. An additional advantage of the proposed method, in comparison to P-SOPLS, is that the solution provided by SOPLS tends to be more similar to that of OPLS, as it can be seen in *letter* and *satellite*.

This last advantage can be analyzed in detail in Fig. 2 where we display the OA against the number of used projections ($1 \leq k \leq r$) in four problems. The proposed SOPLS method generally outperforms P-SOPLS when less than r features are extracted. This increased performance is due to the orthogonality imposed by the proposed SOPLS formulation, which is not enforced by the P-SOPLS solution.

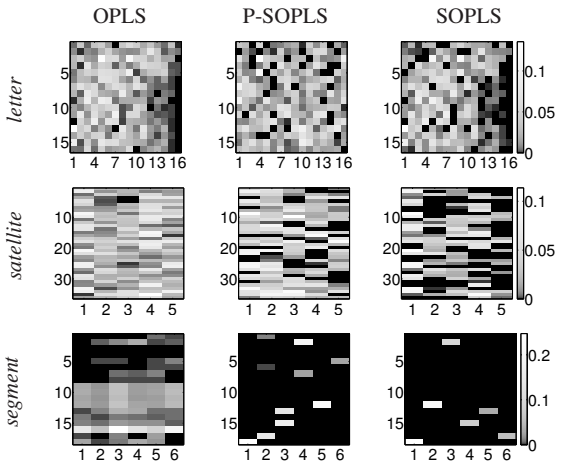

Fig. 1. Representation of the projection matrix $\mathbf{U}$ ($n \times n_p$) in OPLS, P-SOPLS, and SOPLS for three representative problems.

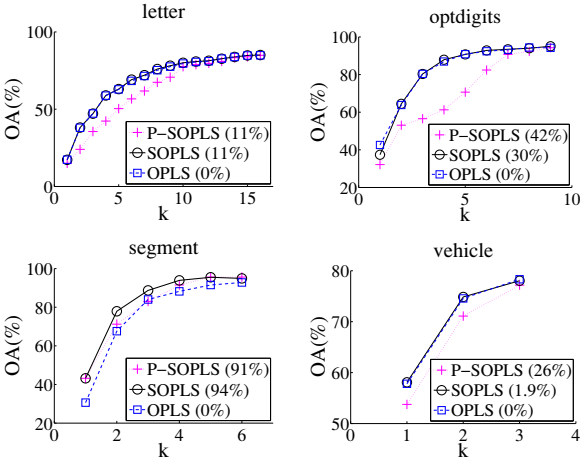

Fig. 2. Overall Accuracy (OA) (%) provided by OPLS, SOPLS, and P-SOPLS algorithms for different number of features k. SR achieved when all projections ($k = r$) are used is shown in the legend.

6. CONCLUSIONS AND FUTURE WORK

This paper proposes a novel OPLS formulation which allows an iterative implementation combining two steps: a standard LS problem to obtain the projection vectors, and an eigenvalue decomposition to compute the regression coefficients. The main advantage of this formulation is its flexibility to include additional constraints. In particular, we analyze its extension to a sparse OPLS formulation (SOPLS). Experimental results show the discriminative power of the SOPLS features in comparison to those provided by OPLS and a previous sparse OPLS approach.

Future work deals with non-linear formulations, as well as an extension with group-lasso constraints to combine the feature extraction process with a selection one.

7. REFERENCES

[1] H. Wold, *Multivariate analysis*, Academic Press, 1966.

[2] J. Shawe-Taylor and N. Cristianini, *Kernel methods for pattern analysis*, Cambridge University Press, 2004.

[3] 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. 26017–26027, 2012.

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

[5] 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.

[6] 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. 2009, Wiley.

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

[8] Z. J. Xiang and P. J. Ramadge, "Fast lasso screening tests based on correlations," in *IEEE Intl. Conf. on Acoustics, Speech and Signal Process. (ICASSP)*. IEEE, 2012, pp. 2137–2140.

[9] J. Salamon, B. Rocha, and E. Gómez, "Musical genre classification using melody features extracted from polyphonic music signals," in *IEEE Intl. Conf. on Acoustics, Speech and Signal Process. (ICASSP)*. IEEE, 2012, pp. 81–84.

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

[11] J. Arenas-García, K. B. Petersen, and L. K. Hansen, "Sparse kernel orthonormalized PLS for feature extraction in large data sets," *Advances in Neural Information Processing Systems*, vol. 19, pp. 33–40, 2007.

[12] C. Dhanjal, S. R. Gunn, and J. Shawe-Taylor, "Efficient sparse kernel feature extraction based on partial least squares," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 8, pp. 1347–1361, 2009.

[13] L. Sun, S. Ji, S. Yu, and J. Ye, "On the equivalence between canonical correlation analysis and orthonormalized partial least squares," in *Proc. of the 21st International Joint Conference on Artificial Intelligence*, 2009, pp. 1230–1235.

[14] C.M. Bishop, *Neural Networks for Pattern Recognition*, Oxford University Press, New York (NY), 1995.

[15] G. X. Yuan, K. W. Chang, C. J. Hsieh, and C. J. Lin, "A comparison of optimization methods and software for large-scale L1-regularized linear classification," *Journal of Machine Learning Research*, vol. 11, pp. 3183–3234, 2010.

[16] F. Bach, R. Jenatton, J. Mairal, and G. Obozinski, "Convex optimization with sparsity-inducing norms," in *Optimization for Machine Learning*, S. Sra, S. Nowozin, and S. J. Wright, Eds. 2011, pp. 19–53, MIT Press.

[17] A. Frank and A. Asuncion, "UCI machine learning repository," [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science, 2010.

[18] Chih-Chung Chang and Chih-Jen Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011, Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.