A generalized method for Sparse Partial Least Squares (Dual-SPLS): theory and applications

Louna Alsouki, Francois Wahl, Laurent Duval, Clément Marteau, Rami El-Haddad

A generalized method for Sparse Partial Least Squares (Dual-SPLS): theory and applications

Louna Alsouki[1,3], François Wahl[1,2], Laurent Duval[2], Clément Marteau[1] & Rami El-Haddad[3]

[1] *Université Claude-Bernard Lyon 1, 43 boulevard du 11 Novembre 1918, 69100 Villeurbanne, France,*
[2] *IFP Energies nouvelles, 1-4 avenue de Bois-Préau, 92852 Rueil-Malmaison, France,*
[3] *Université Saint Joseph de Beyrouth, Mar Roukoz – Dekwaneh, B.P. 1514, Liban,*
louna.al-souki@univ-lyon1.fr, francois.wahl@math.univ-lyon1.fr, laurent.duval@ifpen.fr, marteau@univ-lyon1.fr, rami.haddad@usj.edu.lb

Résumé. En analyse des données, la grande dimensionalité est souvent un obstacle délicat à surmonter qui être résolu en représentant les données dans un espace de dimension inférieure en utilisant des méthodes de projection comme la régression des moindres carrés partiels (PLS) [1] ou en ayant recours à des méthodes de sélection de variables comme l'approche lasso [2]. La Sparse Partial Least Squares (SPLS) combine les deux dernières afin de mieux interpréter les résultats grâce à la parcimonie imposée sur les nouvelles directions. Plusieurs implémentations ont été proposées [3, 4, 5]. Cependant, des problèmes de précision de prédictions et de bonne interprétation des coefficients surgissent dans ces travaux. C'est pourquoi nous avons développé la Dual Sparse Partial Least Squares, une méthode flexible qui permet d'obtenir des prédictions plus précises et une meilleure interprétation des coefficients grâce à leur parcimonie. Dans cet article, nous présentons la théorie derrière Dual-SPLS et certains résultats d'applications sur des ensembles de données pétrolières réelles.

Mots-clés. Moindres carrés partiels, parcimonie, régression, norme duale, algorithme lasso.

Abstract. In data analysis, high dimensionality is often a delicate obstacle to overcome which can be solved by representing the data in a lower dimensional space using projection methods like the Partial Least Squares regression (PLS) [1] or by resorting to variable selection methods like the lasso approach [2]. The Sparse Partial Least Squares (SPLS) combines the two latter in order to better interpret the results due to the sparsity imposed on the new directions. Several implementations have been proposed [3, 4, 5]. However, problems of accuracy of predictions and correct interpretation of regression coefficients arise in these approaches. Hence we developed the Dual Sparse Partial Least Squares, a flexible method that results in more accurate predictions and better interpretation of the coefficients due to their sparsity. In this paper we present the theory behind Dual-SPLS and some applicative results on petroleum data sets.

Keywords. Partial Least Squares, sparsity, regression, dual norm, lasso algorithm.

1 Introduction

Regression analysis helps in inferring relationships between data sets, with the additional objective of extracting interpretable information. However, a recurrent problem haunting statistical data analysis is data high dimensionality. One can choose to tackle this issue by using dimension regression methods, like the PLS procedure [1], allowing to represent the data in a lower dimensional space. It reduces the dimensionality by selecting derived components. It is an iterative method that deals with highly correlated data and results in accurate outcomes. Algorithms are generally straightforward and simple to handle without matrix inversion. However, regression coefficients are frequently hard to interpret (see section 3). Another suggestion often considered is variable selection, like in the lasso [2]. It performs regularization in order to enhance the prediction accuracy, while simplifying the interpretation of the regression coefficients due to the sparsity of the representation. Nevertheless, the lasso is very sensitive to the type of data and does not always result in interpretable coefficients: in fact, it selects at most n variables before it saturates [6]. Sparse Partial Least Squares (SPLS) [3, 4, 5] combines both approaches by adding to the PLS framework a selection step inspired by the lasso. It is represented by the following optimization problem:

$$\hat{\mathbf{w}} = \underset{\mathbf{w}\in\mathbb{R}^p}{\operatorname{argmin}}\{-\hat{\mathrm{Cov}}(\mathbf{X}\mathbf{w}, \mathbf{y}) + \lambda_s \|\mathbf{w}\|_1\}, \qquad \text{for } \mathbf{w}^T\mathbf{w} = 1, \tag{1}$$

under the orthogonality constraint of components, with sparsity parameter $\lambda_s > 0$.

Lê Cao *et al.* (2008) [3] and Chun and Keleş (2010) [4] developed SPLS approaches that both give an approximate solution. Thus, Durif *et al.* [5] conceived a similar method in the context of classification that solves exactly Problem (1) in the univariate response case. Nonetheless, it can be applied in the regression. It however appears to be time consuming on high dimensional data.

Inspired by these methodologies, we devised a new strategy called Dual Sparse Partial Least Squares (Dual-SPLS) that provides prediction accuracy equivalent to the PLS method along with easier interpretation of regression coefficients thanks to the sparsity of the results. Moreover, it generalizes the above mentioned approaches on the theoretical point of view.

We first present the main ingredients of the Dual-SPLS. We then show some results of applications on petroleum data sets.

2 Dual Sparse Partial Least Squares

The proposed method originated from noticing the similarity between the variational formulation of the PLS (with the PLS1 methodology) and the expression of the dual norm of a vector.

Let $\Omega(\cdot)$ be a norm on $\mathbb{R}^P$. The associated dual norm, denoted $\Omega^*(\cdot)$, is defined, for any $\mathbf{z} \in \mathbb{R}^P$, as:

$$\Omega^*(\mathbf{z}) = \max_{\mathbf{w}} (\mathbf{z}^T \mathbf{w}) \quad \text{s.t.} \quad \Omega(\mathbf{w}) = 1. \tag{2}$$

Meanwhile, the optimization problem solved by the PLS method for the first component writes:

$$\max_{\mathbf{w}} (\mathbf{y}^T \mathbf{X} \mathbf{w}) \quad \text{s.t.} \quad \|\mathbf{w}\|_2 = 1. \tag{3}$$

Comparing (2) and (3), one notices that optimizing the PLS criterion amounts to finding the vector $\mathbf{w}_1$ that goes with the conjugate of the ℓ_2-norm of $\mathbf{z}$ where $\mathbf{z} = \mathbf{X}^T \mathbf{y}$, which can be exploited in evaluating different norm expressions.
As in the lasso approach, we consider the combination of the ℓ_1 and ℓ_2 norms:

$$\Omega(\mathbf{w}) = \lambda \|\mathbf{w}\|_1 + \|\mathbf{w}\|_2 \,. \tag{4}$$

The closed form solution can be expressed with the soft thresholding operator. Since we are dealing with a vector, we can consider each coordinate $p \in \{1, \ldots, P\}$ of $\mathbf{w}$ and the solution can be written as:

$$\frac{w_p}{\|\mathbf{w}\|_2} = \frac{1}{\mu} \delta_p (|z_p| - \nu)_+ \tag{5}$$

where $\boldsymbol{\delta}$ is the vector of the signs of $\mathbf{z}$, μ guarantees the normality constraint in (2) and $\nu = \lambda \mu$. This is relevant since we can compare ν to z_p, and therefore shrink to zero the coefficients that correspond to the small coordinates of $\mathbf{z}$ (compared to ν), which enforces sparse regression coefficients.

However, the main challenge resides in setting the parameter ν, which affects the amount of shrinkage. We propose to choose it iteratively and adaptively according to the number of variables that we would like to keep in the active set at each iteration. In other words, for each number i of desired components, an optimal ν_i is chosen to impose a given proportion of null coefficients. The Dual-SPLS method is implemented in the form of Algorithm 1.

Algorithm 1 DUAL-SPLS ALGORITHM FOR $\Omega(\mathbf{w}) = \lambda \|\mathbf{w}\|_1 + \|\mathbf{w}\|_2$

Input: $\mathbf{X}_1, \mathbf{y}, I$ (number of components)
for $i = 1, \ldots, I$ **do**
  $\mathbf{z}_i = \mathbf{X}_i^T \mathbf{y}$ (weight vector)
  Find ν in the adaptive way
  $\mathbf{z}_\nu = (\delta_p (|z_p| - \nu)_+)_{p \in}$ (applying the threshold)
  $\mu = \|\mathbf{z}_\nu\|_2 \quad \lambda = \dfrac{\nu}{\mu} \quad \mathbf{w}_i = \dfrac{\mu}{\nu \|\mathbf{z}_\nu\|_1 + \|\mathbf{z}_\nu\|_2^2} \mathbf{z}_\nu$ (loadings)
  $\mathbf{t}_i = \mathbf{X}_i \mathbf{w}_i / \|\mathbf{X}_i \mathbf{w}_i\|$ (scores)
  $\mathbf{X}_{i+1} = \mathbf{X}_i - \mathbf{t}_i^T \mathbf{t}_i \mathbf{X}_i$ (deflation)
end for
Compute $\hat{\boldsymbol{\beta}}$.

3 Results and discussions

3.1 Data sets

The data set is composed of 243 NMR spectra of refined oil samples. Each spectrum is originally represented by more than 65000 variables. However, we have pretreated them by eliminating irrelevant parts, removing repeated observations and normalizing amplitudes between 0 and 1, which leaves us with around 21000 variables and 182 observations. Our aim is to predict the density of these oil samples.

3.2 Benchmark

We assess the efficiency of the Dual-SPLS by computing the root mean square error (RMSE) for prediction performance and then we examine the interpretation of the coefficients by comparing them to the original raw spectra. The evaluation is organized as a benchmark comparing the following methods together: PLS [1], sPLS of Lê Cao *et. al.* (as implemented in mixOmics) [3], SPLS of Keleş *et. al.* (as implemented in spls) [4], SPLS of Durif *et. al.* (as implemented in plsgenomics) [5] and lasso in glmnet package [2]. In Figure 1 the calibration and validation sets are chosen adequately. It is divided into two parts: the left part corresponds to the RMSE values of the validation set according to the number of components and the right part represents the coefficients of each regression, and for PLS related methods we select 6 components. As for Figure 2, the calibration and validation sets are chosen randomly. We applied the regression methods for 100 repetition in order to represent the boxplots and compare the results.

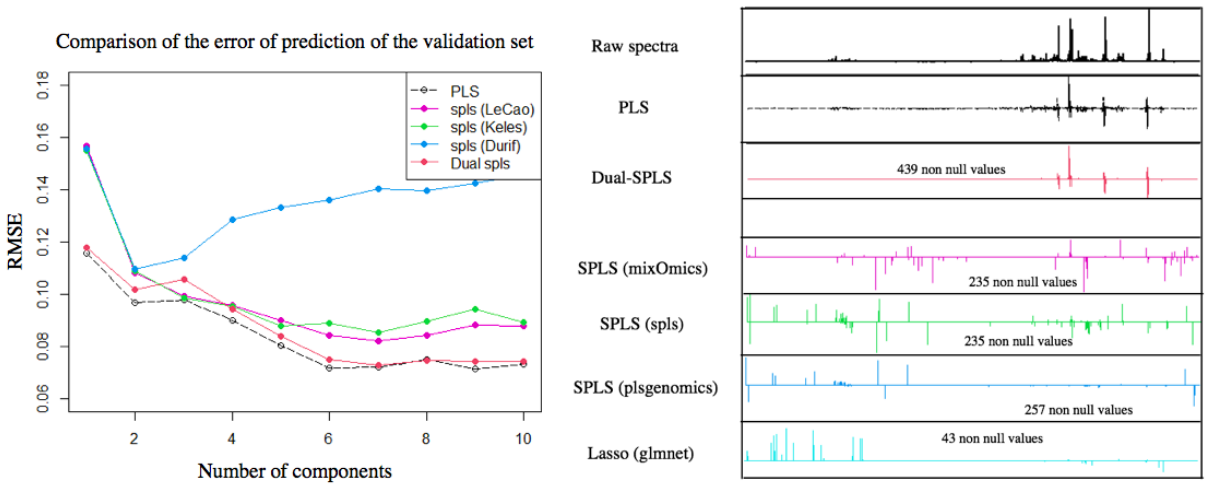

Figure 1: Benchmark of (sparse) PLS methods on the NMR data set: prediction error according to the number of component (left), raw data and coefficients localization (right).

In the Figures 1 and 2 , we require a 99 % proportion of null coefficients while applying the Dual-SPLS and use cross validation to choose the adequate amount of penalization λ_s for each of the other cases. Note that the x-axis is not represented with chemically-sound units due to preprocessing.

From Figure 1 (left), all methods almost match the prediction accuracy of the PLS from two components on, except for spls from the plsgenomics package [5] whose predictions are slightly less accurate. The lasso algorithm provides RMSE values around 0.09 according to the choice of shrinkage parameter. We even notice that the closest results to the PLS are those from the new approach. To compare coefficients localization, we select six components for PLS-related methods as the RMSE curves tend to plateau above this value. On Figure 1 (right), the sparsest results are obtained by the lasso and the proposed Dual-SPLS. However, the lasso does not properly indicate the location of the important variables borne by the highest peaks of the NMR spectra. Their location is better estimated with the Dual-SPLS.

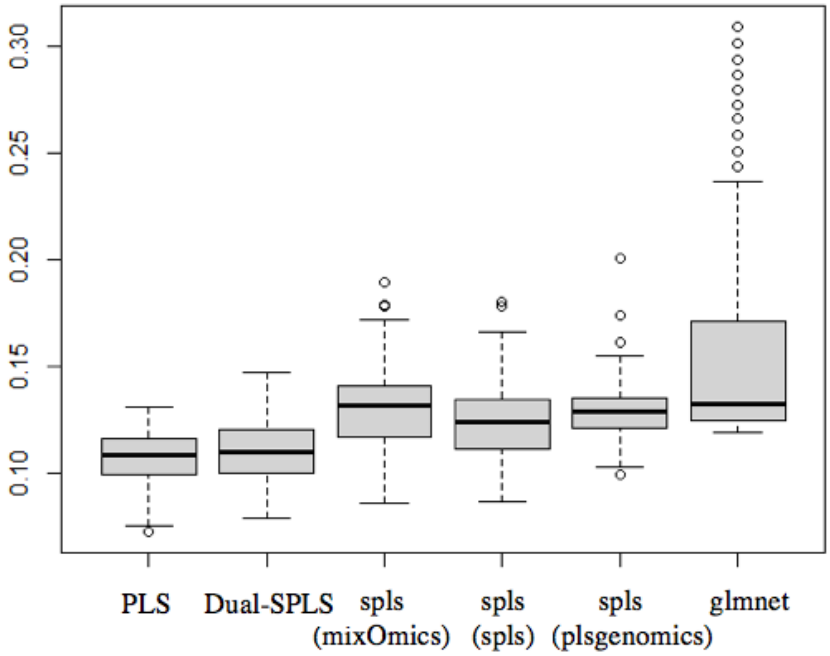

Figure 2: Benchmark of (sparse) PLS methods on the NMR data set: prediction error boxplots using random calibration.

From Figure 2, where we compare the boxplots of the methods applied for six components (for PLS-related methods), we conclude the same: the prediction accuracy of the PLS is the most similar by using the Dual-SPLS.

4 Conclusions

The Dual-SPLS introduces a general framework providing a novel family of regression methods that encompasses the standard PLS method. It offers the possibility to use a quantity of different norm shapes. In the case of a norm inspired by the lasso, it already preserves the prediction accuracy of the PLS and previously proposed sparse PLS methodologies. On NMR data it shows to be even sparser with better localized and more interpretable coefficients. The next steps will consist first in implementing and sharing this method as an R package, and second in evaluating the gain in performance by using other norms mimicking the fused or the group lasso.

References

[1] H. Wold, "Path models with latent variables: The NIPALS approach," in *Quantitative Sociology. International Perspectives on Mathematical and Statistical Modeling*, pp. 307–357. Elsevier, 1975.

[2] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. R. Stat. Soc. Ser. B Stat. Methodol.*, vol. 58, no. 1, pp. 267–288, 1996.

[3] K.-A. Lê Cao, D. Rossouw, C. Robert-Granié, and P. Besse, "A sparse PLS for variable selection when integrating omics data," *Stat. Appl. Genet. Mol. Biol.*, vol. 7, no. 1, pp. 35, 2008.

[4] H. Chun and S. Keleş, "Sparse partial least squares regression for simultaneous dimension reduction and variable selection," *J. R. Stat. Soc. Ser. B Stat. Methodol.*, vol. 72, no. 1, pp. 3–25, 2010.

[5] G. Durif, L. Modolo, J. Michaelsson, J. E. Mold, S. Lambert-Lacroix, and F. Picard, "High dimensional classification with combined adaptive sparse PLS and logistic regression," *Bioinformatics*, vol. 34, no. 3, pp. 485–493, Feb. 2018.

[6] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *J. R. Stat. Soc. Ser. B Stat. Methodol.*, vol. 67, no. 2, pp. 301–320, Apr. 2005.

Acknowledgement

This work was performed within the framework of the LABEX MILYON (ANR-10-LABX-0070) of Université de Lyon, within the program "Investissements d'Avenir" (ANR-11-IDEX-0007) operated by the French National Research Agency (ANR).