Interpretable Dual Sparse Partial Least Squares (DS-PLS) regression; Application to NMR/NIR petroleum data sets

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

e- CHIMIOMÉTRIE 2021 Une conférence 100 % en ligne et gratuite 2-3 Février 2021 SFdS GFC

Interpretable Dual Sparse Partial Least Squares (DS-PLS) regression; Application to NMR/NIR petroleum data sets[1]

Louna Alsouki[2] François Wahl[2,3] Laurent Duval[3] Clément Marteau[2] Rami El-Haddad[4]

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

Keywords: Partial Least Squares, sparsity, regression, dual norm.

1 Introduction

Regression analysis helps in inferring relationships between data sets, with the additional objective of extracting interpretable information. Partial Least Squares [1] (PLS) is often used when dealing with NMR or NIR spectra to predict properties of petroleum samples. In spite of its ability to operate with high-dimensional data, and its efficiency in predicting responses, PLS lacks in considering the functional nature of the data and shows weaknesses in result interpretation. In order to improve these two aspects, we developed a new strategy called Dual Sparse Partial Least Squares (DS-PLS) that gives equivalent prediction accuracy along with facilitated interpretation of regression coefficients, due to the sparsity of the representation.

2 Theory

The proposed method was devised from noticing the similarity between finding the PLS components (with the PLS1 methodology) and expressing the dual L_2 norm of a vector.

Let $\Omega(z)$ be a norm. Its dual [2] has the following form:

$$\Omega^*(z) = \max_w (z^T w), \qquad s.t.\ \Omega(w) = 1. \quad \#(1)$$

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

$$\max_w (y^T X w), \qquad s.t.\ \|w\|_2 = 1. \quad \#(2)$$

Comparing (1) and (2), one notices that optimizing the PLS function amounts to finding the vector w_1 that goes with the conjugate of the L_2-norm of z, where $z = X^T y$.
Therefore, we propose to evaluate different norm expressions, notably adding adaptive penalization. An example is the norm $\Omega(w) = \lambda\|w\|_1 + \|w\|_2$. Interestingly, this formulation leads to closed-form expressions as in [3] and requires only slight modifications of the standard PLS1 algorithm: the solution is known as the soft thresholding operator in the lasso [4] literature: $\forall j, w_j = \left(\frac{|z_j|}{\mu} - \lambda\right)_+$ where μ is tuned to guarantee that $\Omega(w) = 1$.

[1]*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).*

Moreover, this framework allows to vary the form of the norm. Another possibility is for example $\Omega(w) = \lambda\|Nw\|_1 + \|w\|_2$ like in fused lasso, where N is a penalty matrix. These constraints would introduce a functional aspect in the treatment.

3 Material and methods

We apply the DS-PLS to 208 samples of NIR spectra represented by 2594 variables and to 243 samples of NMR spectra represented by 20998 variables. After dividing the data sets in two similar sets (calibration and validation) and using the R programming platform, we evaluate the prediction using both Root Mean Squares and Mean Absolute Errors. We also compare the coefficients for each model with the raw data.

4 Results and discussion

Comparing methods, the proposed strategy matches the prediction accuracy of the PLS, and additionally provides a good interpretation of the coefficients (Figure 1) due to the sparsity of its results. In the following figure, we require 99 % of null coefficients while applying DS-PLS and use 6 components for both standard and DS-PLS regressions. Note that x-axis units are not represented due to data preprocessing.

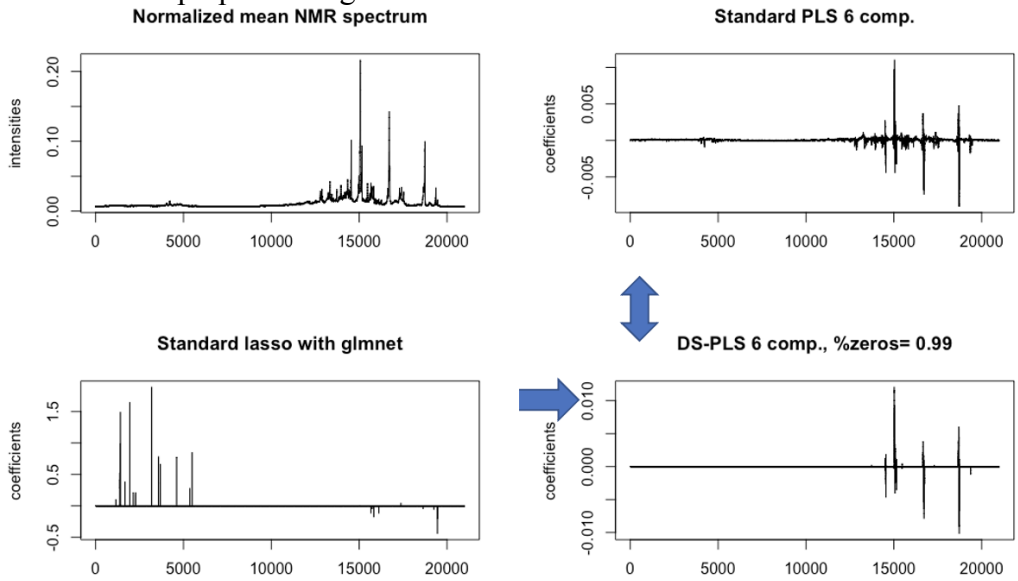

Figure 1 – Comparing coefficients of regression with the normalized mean MNR spectrum.

5 Conclusion

DS-PLS is a novel family of regression methods that provides a general framework: it encompasses the standard PLS method, and gives us the possibility to use other norm shapes. At this point, it preserves the accuracy of prediction of the PLS method and adds on sparsity in the coefficients for interpretation. The next challenge is to evaluate norms like ones for in fused or grouped lasso.

6 References

[1] Tenenhaus M. La régression PLS: théorie et pratique. Paris: Éditions Technip, 1998.

[2] Bach F., Jenatton R., Mairal J., and Obozinski G. Optimization with Sparsity-Inducing Penalties. *Found. Trends Mach. Learn.*, 2012, 4(1), 1–106.

[3] Durif G., Modolo L., Michaelsson J., Mold J., Lambert-Lacroix S., Picard F. High Dimensional Classification with combined Adaptive Sparse PLS and Logistic Regression. *Bioinformatics, Oxford University Press,* 2018, 34 (3), pp.485-493.

[4] Tibshirani R. Regression Shrinkage and Selection via the lasso. *Journal of the Royal Statistical Society. Series B (methodological). Wiley*, 1996, 58 (1): 267–88.