Dual-sPLS : a Family of Dual Sparse Partial Least Squares Regressions for Feature Selection and Prediction with Tunable Sparsity; Evaluation on Simulated and Near-Infrared (NIR) Data

Louna Alsouki, Laurent Duval, Clément Marteau, Rami El Haddad, François Wahl

Dual-sPLS: a family of Dual Sparse Partial Least Squares regressions for feature selection and prediction with tunable sparsity; evaluation on simulated and near-infrared (NIR) data

Louna Alsouki[a,c], Laurent Duval[b], Clément Marteau[c], Rami El Haddad[a], François Wahl[b,c]

[a] *Laboratoire de Mathématiques et Applications, U.R. Mathématiques et modélisation, Faculté des sciences, Université Saint-Joseph, B.P. 7-5208, Mar Mikhaël Beyrouth, 1104 2020, Liban*

[b] *IFP Energies nouvelles, 1-4 avenue du Bois-Préau, Rueil-Malmaison, 92852, France*

[c] *Institut Camille Jordan, Université Claude Bernard Lyon 1, 43 boulevard du 11 novembre 1918, Villeurbanne, 69100, France*

Abstract

Relating a set of variables $\mathbf{X}$ to a response $\mathbf{y}$ is crucial in chemometrics. A quantitative prediction objective can be enriched by qualitative data interpretation, for instance by locating the most influential features. When high-dimensional problems arise, dimension reduction techniques can be used. Most notable are projections (e.g. Partial Least Squares or PLS) or variable selections (e.g. lasso). Sparse partial least squares combine both strategies, by blending variable selection into PLS. The variant presented in this paper, Dual-sPLS, generalizes the classical PLS1 algorithm. It provides balance between accurate prediction and efficient interpretation. It is based on penalizations inspired by classical regression methods (lasso, group lasso, least squares, ridge) and uses the dual norm notion. The resulting sparsity is enforced by an intuitive shrinking ratio parameter. Dual-sPLS favorably compares to similar regression methods, on simulated and real chemical data.

Keywords: Partial least squares, lasso, ridge, regression, sparsity, dual norm, chemometrics, machine learning

1. Introduction

Two main feats of chemometrics reside in first, providing reliable inference and second, offering interpretability of chemical data sources. On the one hand, one may expect to estimate, within a given precision, responses $\mathbf{Y} \in \mathbb{R}^{N \times Q}$ (e.g. hydrocarbon properties: viscosity, density, cetane number [1]) from spectra or variables represented by quantities $\mathbf{X} \in \mathbb{R}^{N \times P}$ (nuclear magnitude resonance or NMR, chromatography, infrared spectroscopy, etc. [2]). It aims at relating a target $\mathbf{Y}$ to $\mathbf{X}$ through a predictive model: for instance, NMR spectra can be linked to viscosity with predictive purposes. On the other hand, one also wishes to interpret how variables in $\mathbf{X}$ influence quantities $\mathbf{Y}$, i.e. which spectral features are most consistent with response prediction, a question related to wavelength selection. For instance, which spectral bands in NMR, in terms of continuous localization, could be related to the viscosity index estimation (see e.g. [3])? This can be transcribed by a regression model, often considered linear:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}\,, \tag{1}$$

where $\boldsymbol{\epsilon}$ is expected to be independent of $\mathbf{X}$, with zero mean. With the growing size of consolidated analytical chemistry databases, chemometrics still require methodologies to 1) provide accurate predictions 2) extract pertinent knowledge or offer useful insights on measurements 3) combine heterogeneous or high-dimensional data sources. When the number P of variables (samples) is far greater than the number of observations (signals) N ($P \gg N$), naive statistical models risk overfitting. This notably happens in standard least squares optimizations. Dimension reduction techniques are generic approaches to deal with high dimensionality. They include projection methods or variable selection algorithms. Commonly used strategies start with PCA/PCR (principal component analysis/regression), performed only on explanatory variables in $\mathbf{X}$. They however do not incorporate information held by the response $\mathbf{Y}$. Partial least squares (PLS) [4, 5], also called projection onto latent structures, is therefore common in chemometrics, with better prediction-prone latent components. However, PLS sometimes lacks appropriate interpretability.

As for variable selection, one often resorts to the lasso algorithm (least absolute shrinkage and selection operator [6]). Shrinkage induces a form of sparsity, which amounts to selecting important variables. It is however known

to be sensitive to data types. It does not always yield interpretable coefficients. Blends of the two above – dimension reduction and variable selection – have recent avatars called sparse PLS (sPLS). While they enforce lower dimensional decompositions, they do not always provide chemically pertinent feature localization for physico-analytical measurements. Thereby, we propose a dual sparse PLS family dedicated to one dimensional or univariate responses: $\mathbf{y} = \mathbf{Y}$, with $\mathbf{y} \in \mathbb{R}^N$. It generalizes the standard PLS1 algorithm by supplementing it with adequate penalties. This formally provides a unified formulation for regression methods in the spirit of the lasso mentioned above, and also least-squares or ridge, all blended in a PLS formalism. It also allows *variable grouping*: the possibility to gather explanatory variables into more meaningful subsets (contiguous samples around a peak, disjoint spectral bands associated to a compound). This can be used to combine different physico-chemical modalities. Resolution resorts to the dual norm of the chosen Dual-sPLS penalty. This new method has many advantages:

1. predictions match or outperform state-of-the-art or comparable methods,

2. in the different norm options we considered, they additionally yield sparse representations of both simulated and real chemical near infrared data, even singular, a frequent ill-conditioning issue in high dimension,

3. they finally offer a interpretable localization of features from a functional data or statistical point of view.

Those three properties combined offer alternative surrogates to classical approaches (PLS, lasso, least squares, ridge). It permits both accurate inference and pertinent domain-related interpretation.
The paper is structured as follows: setting notations, we briefly revise in Section 2 the background of the PLS, recall classical variable selection methods and evoke their blending in sparse PLS schemes, previously proposed. Then, in Section 3, we explain principles behind the Dual-sPLS family and detail the list of norm penalties and their algorithms in three main instances: the (group) lasso form —being the most important— and least squares and ridge forms. Thereafter Section 4 describes tested data (simulated and real) and the choices of model settings, calibration and validation. Each of the three penalties types are extensively benchmarked in Section 5. We finally draw concluding remarks with perspectives in Section 6 and supplementary material in the appendix.

Notation and definitions

Matrices, vectors and scalars are denoted by boldface uppercase letters, boldface lowercase and light lowercase letters respectively, e.g. $\mathbf{X}$, $\mathbf{y}$ and λ. The transpose of matrix $\mathbf{X}$ is $\mathbf{X}^T$. The identity matrix of size P is represented by I_P. The ℓ_1-norm and the ℓ_2-norm of vector a $\mathbf{w}$ of length P are

$$\|\mathbf{w}\|_1 = \sum_{p=1}^{P} |w_p| \qquad \text{and} \qquad \|\mathbf{w}\|_2 = \sqrt{\sum_{p=1}^{P} |w_p|^2}\,. \tag{2}$$

We denote by $\ell_0(\mathbf{w})$ the sparsity index or count measure [7] of the non-zero coordinates of $\mathbf{w}$ and $\ell_0^c(\mathbf{w})$ its complement i.e. $\ell_0^c(\mathbf{w}) = P - \ell_0(\mathbf{w})$. To choose the number of latent variables we rely on the mean squared error (MSE) expressed as

$$\text{MSE} = \frac{1}{N}\sum_{n=1}^{N} (y_n - \hat{y}_n)^2\,, \tag{3}$$

for a response vector $\mathbf{y}$ of N observations and a given estimate $\hat{\mathbf{y}}$. For performance evaluation, we choose the root mean squares error (RMSE), the mean absolute error (MAE) and the determination coefficient (R^2):

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{n=1}^{N} (y_n - \hat{y}_n)^2} = \frac{1}{\sqrt{N}}\|\mathbf{y} - \hat{\mathbf{y}}\|_2\,, \tag{4}$$

$$\text{MAE} = \frac{1}{N}\sum_{n=1}^{N} |y_n - \hat{y}_n| = \frac{1}{N}\|\mathbf{y} - \hat{\mathbf{y}}\|_1\,, \tag{5}$$

$$\text{R}^2 = \frac{\sum_{n=1}^{N}(y_n - \bar{y})^2}{\sum_{n=1}^{N}(y_n - \hat{y}_n)^2} \qquad \text{where} \qquad \bar{y} = \frac{\sum_{n=1}^{N} y_n}{N}\,. \tag{6}$$

The vector of signs of $\mathbf{w}$ is noted $\text{sign}(\mathbf{w})$, and $(\mathbf{w})_+$ is the vector composed of w_p if $w_p \geq 0$ and 0 if $w_p < 0$[1].

In the following, we assume that the matrix $\mathbf{X} \in \mathbb{R}^{N\times P}$ of independent variables and the response vector $\mathbf{y} \in \mathbb{R}^N$ are mean-centered. We use the convention where columns denote variables and rows observations.

[1]It corresponds to the Rectified Linear Unit (ReLU), a popular activation function for neural networks.

2. Background

2.1. Partial Least Squares (PLS)

PLS originated from econometrics [8]. It was progressively and successfully applied to other fields [9]: social and behavioral sciences, biosciences from bioinformatics [10] to neuroimaging [11], and chemometrics [4, 5]. It denotes a class of methods aimed at explaining the relationship between explanatory data and responses with the help of latent variables. They boast the management of both formative and reflective measurements, require low sample sizes and mild distributional assumptions.

PLS avatars root on projecting response onto a lower M-dimensional space spanned by new orthogonal directions $\{\mathbf{t}_1, \ldots, \mathbf{t}_M\}$ constructed as linear combinations of original variables. Its principle consists in compressing the predictor $\mathbf{X}$ into a smaller score matrix $\mathbf{T}$ of those $M < P$ variables. Thus, PLS computes M weights $\{\mathbf{w}_1, \ldots, \mathbf{w}_M\}$ forming the loading matrix $\mathbf{W}$ such that $\mathbf{T} = \mathbf{XW}$. As a result, loadings form an orthogonal basis. When Principal Component Analysis (PCA) [12] ought to best summarize $\mathbf{X}$ by taking into account only the correlation between the variables in $\mathbf{X}$, the PLS steps up and also consider the covariance between $\mathbf{X}$ and $\mathbf{y}$. Several algorithms have been proposed. NIPALS (nonlinear iterative partial least squares) [13] and SIMPLS [14] are most popular. When applied to a one-dimensional reponse, as in our case, both are shown to be equivalent. They solve the following optimization problem for the first component:

$$\max_{\mathbf{w}} (\mathbf{y}^T \mathbf{X} \mathbf{w}) \quad \text{s.t.} \quad \|\mathbf{w}\|_2 = 1 \,. \tag{7}$$

The convex Problem (7) can be solved with Lagrange multipliers. For $\mu > 0$, it rewrites:

$$\min_{\mathbf{w}} L(\mathbf{w}) \quad \text{where } L(\mathbf{w}) = -\mathbf{z}^T \mathbf{w} + \mu(\|\mathbf{w}\|_2 - 1) \;\text{ and }\; \mathbf{z} = \mathbf{X}^T \mathbf{y} \,. \tag{8}$$

Solving (8) leads to

$$\mathbf{w} = \mathbf{X}^T \mathbf{y} \,. \tag{9}$$

The PLS algorithm uses the weight vector $\mathbf{w}$ to compress regressor $\mathbf{X}$ into score vector $\mathbf{t} = \mathbf{Xw}$. NIPALS iteratively computes weight vectors by deflation while SIMPLS is more direct. Let $\mathcal{P}_{\cdot}$ denotes the orthogonal projection onto the space spanned by components specified in subscript. For instance, scores $\{\mathbf{t}_1, \ldots, \mathbf{t}_{m-1}\}$ span the space corresponding to $\mathcal{P}_{\mathbf{t}_{m-1}}$. The algorithm

considers the part of $\mathbf{X}$ that is orthogonal to $\mathbf{t}_k, k < m$. For the m^{th} component, $\mathbf{X}$ is replaced by $\mathbf{X}_m$ such that:

$$\mathbf{X}_m = \mathbf{X} - \mathcal{P}_{\mathbf{t}_1,\ldots,\mathbf{t}_{m-1}}\mathbf{X} = \mathbf{X}_{m-1} - \mathcal{P}_{\mathbf{t}_{m-1}}\mathbf{X}_{m-1}\,. \tag{10}$$

The NIPALS variant PLS1 for an univariate reponse as given in [15] is described in Algorithm 1.

Algorithm 1 NIPALS PLS1

Input: $\mathbf{X}, \mathbf{y}, M$
$\mathbf{X}_1 = \mathbf{X}$
for $m = 1, \ldots, M$ **do**
  $\mathbf{w}_m = \mathbf{X}_m^T\mathbf{y}$ (weight vector computation)
  $\mathbf{t}_m = \mathbf{X}_m\mathbf{w}_m$ (component construction)
  $\mathbf{X}_{m+1} = \mathbf{X}_m - \mathcal{P}_{\mathbf{t}_m}\mathbf{X}_m$ (deflation)
end for

This algorithm produces a new lower dimensional score matrix $\mathbf{T} \in \mathbb{R}^{N\times M}$. Proposition 1 from [16] explicits the regression coefficients for M components as:

$$\hat{\boldsymbol{\beta}}_M^{PLS} = \mathbf{W}(\mathbf{T}^T\mathbf{X}\mathbf{W})^{-1}\mathbf{T}^T\mathbf{y}\,. \tag{11}$$

The vector of regression fitted values $\hat{\mathbf{y}}$ for M components is the projection of response vector $\mathbf{y}$ onto the space spanned by scores columns of $\mathbf{T}$.

2.2. Least absolute shrinkage and selection operator

By selecting the most important features, variable selection produces a less complicated model. It has the potential advantage of being easier to handle than the complete full set of variables. The optimization problem in standard linear regression is stated as:

$$\arg\min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2\,. \tag{12}$$

Provided $\mathbf{X}$ has full column rank, the ordinary least squares (LS) estimation is $\hat{\mathbf{y}}_{LS} = \mathcal{P}_{[\mathbf{X}]}\mathbf{y}$, where $[\mathbf{X}]$ is the space spanned by the columns of $\mathbf{X}$. In other terms, $\hat{\boldsymbol{\beta}}^{\mathrm{LS}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$ A popular sparsity-based approach is the lasso developed by Tibshirani in 1996 [6]. It is reknown for its ℓ_1 penalty scheme that shrinks less relevant variables to zero. It is obtained by solving:

$$\arg\min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 \quad \text{subject to } \|\boldsymbol{\beta}\|_1 \leq \lambda\,. \tag{13}$$

Threshold parameter $\lambda > 0$ controls the extent of shrinkage applied to the estimate; that is, the number ℓ_0^c of coefficients set to zero. An appropriate λ is important to get interpretable results. If $\hat{\boldsymbol{\beta}}^{\mathrm{LS}}$ exists, as mentioned in [6], then for a $\lambda \geq \|\hat{\boldsymbol{\beta}}^{\mathrm{LS}}\|_1$, the lasso estimate $\hat{\boldsymbol{\beta}}^{\mathrm{l}}$ is equal to the ordinary least square solution. And for $\lambda = \dfrac{\|\hat{\boldsymbol{\beta}}^{\mathrm{LS}}\|_1}{2}$, it selects on average half of the variables. We can reformulate (13) as

$$\arg\min_{\boldsymbol{\beta}} \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + t\|\boldsymbol{\beta}\|_1\,. \tag{14}$$

Note that there is a (non-explicit) correspondence between parameters λ and t. In the orthonormal design case, i.e. $\mathbf{X}^T\mathbf{X} = I_P$, there exists $\hat{\beta}^{\mathrm{l}}$ closed form solution called *soft thresholding* verifying

$$\hat{\beta}_p^{\mathrm{l}} = \mathrm{sign}(\hat{\beta}_p^{\mathrm{LS}})(|\hat{\beta}_p^{\mathrm{LS}}| - \lambda)_+ \quad \forall p \in \{1, \ldots, P\}\,. \tag{15}$$

Coefficients whose magnitude is smaller than λ are set to zero. Amplitudes of the others are shrunk with respect of the threshold. While proved successful for numerous applications, some drawbacks are reported [17, 18]. Some are: 1) non strict convexity of the criterion when the number of predictors exceeds the number of observations ($P > N$) 2) algorithm saturation when N variables have been selected 3) with highly correlated variables, tendency to pick mildly representative ones.
Another shrinking method is ridge regression [19] with optimization problem:

$$\arg\min_{\boldsymbol{\beta}} \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + t\|\boldsymbol{\beta}\|_2\,. \tag{16}$$

Its trick is to add a diagonal matrix to $(\mathbf{X}^T\mathbf{X})$ in order to overcome the singularity problem. Therefore, the solution always exists, expressed as:

$$\hat{\boldsymbol{\beta}}^{\mathrm{r}} = (\mathbf{X}^T\mathbf{X} + tI_P)^{-1}\mathbf{X}^T\mathbf{y}\,. \tag{17}$$

Compared to the lasso, it uses an ℓ_2-norm instead of the ℓ_1 penalization but retains most variables by design.

2.3. Blending methods: sparse Partial Least Squares (sPLS)

Sparse Partial Least Squares (sPLS) denotes a body of works adding a variable selection flavor to the standard PLS framework. We focus here on ones specifically using lasso inspired penalties. An ℓ_1-norm can be incorporated in optimization problem (7). Noting

$$\widehat{\mathrm{Cov}}(\mathbf{Xw}, \mathbf{y}) = \frac{1}{N}\mathbf{w}^T\mathbf{z}, \quad \text{with} \quad \mathbf{z} = \mathbf{X}^T\mathbf{y} = N\widehat{\mathrm{Cov}}(\mathbf{X}, \mathbf{y})\,, \tag{18}$$

adding the coupling parameter $\lambda_s > 0$ and orthogonality constraint on components $\{\mathbf{t}_1, \ldots, \mathbf{t}_M\}$, the sPLS optimization problem is stated as:

$$\hat{\mathbf{w}} = \underset{\mathbf{w}\in\mathbb{R}^p}{\arg\min}\{-\widehat{\mathrm{Cov}}(\mathbf{Xw}, \mathbf{y}) + \lambda_s\|\mathbf{w}\|_1\}, \qquad \text{for } \mathbf{w}^T\mathbf{w} = 1\,. \tag{19}$$

Problem (19) is tackled in 2008 [20] using sparse PCA [21]. Then iterative PLS [22] is combined to singular value decomposition. We denote it as $\mathrm{sPLS}_{\mathrm{LeCao}}$ after the first author. In 2010 [23], Problem (19) is reframed by imposing the ℓ_1 penalty on a surrogate direction close to the original vector $\mathbf{w}$, providing an approximate solution with $\mathrm{sPLS}_{\mathrm{Chun}}$. In 2018, [24] reformulates Problem (19) using recent results from proximal optimization [25] with $\mathrm{sPLS}_{\mathrm{Durif}}$. In this last case, $\mathrm{sPLS}_{\mathrm{Durif}}$ provides an exact and closed-form solution reminiscing the soft threshold operator. Moreover, they suggest an adaptive method for computing the sPLS weight vectors using classical PLS ones.

Along the lines of methods presented above, Dual-sPLS aims at inference and interpretability: accurate predictions combined with sparse localization features for better chemometrics performance. Following [24], we also wish to provide a means to tuning the relative sparsity of the outcome. Finally, as different analytical chemistry modalities provide different insights on chemical mixtures, the Dual-sPLS is designed to naturally allow the combination of heterogeneous datasets as a byproduct of the versatile dual norm approach[2].

[2] Application of this extention is not performed here and is subject to a later work.

3. Dual Sparse Partial Least Squares (Dual-sPLS)

3.1. Motivation and purposes

In statistics and machine learning, it is quite standard to penalize a data fidelity ℓ_2-norm by a penalty involving a specific norm, e.g. ridge (ℓ_2), lasso ℓ_1, etc. (see previous Sections). These penalty options are crucial, they drive the obtention of admissible or reasonable solutions, for instance towards sparsity. The goal of this contribution is to provide a general paradigm for this kind of task. Arbitrary norm choices may not lead to trackable algorithms. However, the concept of dual norm is a means to formulate a unifying optimization framework, for which one can opt for penalties with practical algorithmic properties.

Definition 3.1 *Let $\Omega(\cdot)$ be a norm on $\mathbb{R}^P$. For any $\mathbf{z}\in \mathbb{R}^{\mathrm{P}}$, the associated dual norm, denoted $\Omega^*(\cdot)$, is defined as*

$$\Omega^*(\mathbf{z}) = \max_{\mathbf{w}}(\mathbf{z}^T\mathbf{w}) \quad \text{s.t.} \quad \Omega(\mathbf{w}) = 1\,. \tag{20}$$

Comparing (7) and (20), we find that the optimization of the PLS objective function amounts to finding the vector $\mathbf{w}_1$ that fits the dual norm of the ℓ_2-norm of $\mathbf{z}$, where $\mathbf{z} = \mathbf{X}^T\mathbf{y}$. This gives us the incentive to evaluate different norm expressions that could be used as domain-related penalizations. Thus, for any norm $\Omega(.)$ used, the first component will be:

$$\hat{\mathbf{w}} = \underset{\mathbf{w}\in\mathbb{R}^p}{\arg\min}\{-\mathbf{z}^T\mathbf{w}\}, \qquad \text{s.t. } \Omega(\mathbf{w}) = 1\,. \tag{21}$$

Imposing a form of sparsity on the solution has inspired a quantity of research. The ℓ_1-norm is one of the earliest penalties, that encourages sparsity while remaining convex, and associated with efficient, tractable algorithmic implementations. Our study focuses on norms that 1) have been employed as penalties in previous works and 2) provide explicit, straightforward, and effective algorithms within the PLS framework. Namely, although formulation is generic, we emphasize four types of norms. They make practical sense when dealing with measurements typically available in chemometrics, starting with the lasso analogue, a natural and intuitive approach. We provide the corresponding R [26] package dual.spls [27] with a complete description. It contains the following main functions, each of them being associated to specific penalty:

1. **Dual-sPLS$_{\mathbf{l}}$** *(pseudo-lasso norm, d.spls.lasso()).* Similar to the sPLS Problem (19), an intuitive norm combines ℓ_2 and ℓ_1:

$$\Omega(\mathbf{w}) = \lambda\|\mathbf{w}\|_1 + \|\mathbf{w}\|_2 \,. \tag{22}$$

   Dual-sPLS$_{\mathrm{l}}$ is inspired by ℓ_1 lasso and implemented for situations where we seek selection of features with most impact on the response when dealing with large, highly-correlated data.

2. **Dual-sPLS$_{\mathbf{gl}}$** *(pseudo-group lasso norm, d.spls.GL()).* Inspired by group lasso [28], it combines groups of measurements. It applies pseudo-lasso to each group individually while constraining the total set. For G groups, $\mathbf{w}_g$ represents the variables of the loading vector $\mathbf{w}$ that belongs to group g. The corresponding norm is formulated as:

$$\Omega(\mathbf{w}) = \sum_{g=1}^{G} \alpha_g \|\mathbf{w}_g\|_2 + \lambda_g \|\mathbf{w}_g\|_1 \,, \tag{23}$$

   where $\alpha_g \geq 0, \forall g \in \{1, \ldots, G\}$ and $\sum_{g\in 1,\ldots,G} \alpha_g = 1$. Dual-sPLS$_{\mathrm{gl}}$ is mainly thought for the following not-exclusive cases, akin to multi-block PLS. First, for a single type of measurement $\mathbf{X}$, when G different subsets of scalar variables are expected to contribute jointly to the response, e.g. from wavelength selection or prior analytical chemistry knowledge. Second, when a single response $\mathbf{y}$ can be predited by G distincts sets of measurements $\mathbf{X}_1, \ldots, \mathbf{X}_G$, e.g. different physico-chemical modalities that could be complementary.

3. **Dual-sPLS$_{\mathbf{LS}}$** *(pseudo-least squares norm, d.spls.LS()).* It introduces $\mathbf{N}_1$, a matrix of p columns, and applies when $\mathbf{X}$ is not singular:

$$\Omega(\mathbf{w}) = \lambda\|\mathbf{N_1 w}\|_1 + \|\mathbf{X w}\|_2 \,. \tag{24}$$

   The mild conditions on $\mathbf{N}_1$ are provided in Appendix A.2. Dual-sPLS$_{\mathrm{LS}}$ adds a variable selection flavor to classical least-squares. Therefore, it can be employed when shrinking original LS regression parameters is desired. The classical least squares solution is recovered for $\lambda = 0$.

4. **Dual-sPLS$_{\mathbf{r}}$** *(pseudo-ridge norm, d.spls.ridge()).* It deals with cases where $\mathbf{X}$ is singular and resorts to a ridge-like penalization:

$$\Omega(\mathbf{w}) = \lambda_1\|\mathbf{w}\|_1 + \lambda_2\|\mathbf{X w}\|_2 + \|\mathbf{w}\|_2 \,. \tag{25}$$

The construction of weight vectors $\mathbf{w}_1, \ldots, \mathbf{w}_M$ differs in each of the four cases. It however follows similar steps as for the PLS. Starting with a reformulation of optimization Problem (20) and using Lagrange multipliers, we aim at iteratively minimizing the function $L(\mathbf{w}) = -\mathbf{z}^T\mathbf{w} + \mu(\Omega(\mathbf{w}) - 1)$, for $\mu > 0$. As some norms are not differentiable, we resort to the more generic notion of subgradient $\nabla\Omega(\mathbf{w})$ [25]. It identifies to the classical differential when it is defined. The subgradient of L vanishes for

$$\nabla\Omega(\mathbf{w}) = \frac{\mathbf{z}}{\mu}\,. \tag{26}$$

It is then sufficient to substitute the gradient — when it exists— of the considered norm of $\Omega(\mathbf{w})$ in (26).
We provide in the following a detailed analysis for the pseudo-lasso case of Dual-sPLS (see (22)) and some remarks for the other norms. In all cases we impose that $\mathbf{w}$ and $\mathbf{z}$ lie in the same orthant; it generalizes, in n dimensions, the quadrant in the 2D plane or the octant in the 3D space. In other words, corresponding coordinates of $\mathbf{w}$ and $\mathbf{z}$ have the same sign.

3.2. Norm options (lasso, group lasso, least squares and ridge)

3.2.1. Pseudo-lasso

We rconsider Equation (22). Let $\boldsymbol{\delta}$ be the sign vector of $\mathbf{w}$ and $\mathbf{z}$. By differentiating $\Omega(\mathbf{w})$, we get

$$\nabla\Omega(\mathbf{w}) = \lambda\boldsymbol{\delta} + \frac{w}{\|w\|_2}\,, \tag{27}$$

and by substituting it in (26), we obtain

$$\frac{\mathbf{w}}{\|\mathbf{w}\|_2} = \frac{\mathbf{z}}{\mu} - \lambda\boldsymbol{\delta}\,. \tag{28}$$

The closed-form solution of the Dual-sPLS$_1$ optimization problem consists in zeroing coordinates whose magnitude is lower than the soft threshold λ and in reducing the others toward zero. Thus, for $\nu = \lambda\mu$ and $p \in \{1, \ldots, P\}$, it can be expressed as:

$$\frac{w_p}{\|\mathbf{w}\|_2} = \frac{1}{\mu}\delta_p(|z_p| - \nu)_+\,. \tag{29}$$

A common issue is the choice of the appropriate shrinking parameter. Cross-Validation [29], evoked in Section 4.3, is popularly adopted in sparse regressions. We choose a more intuitive option. We obtain it adaptively, according

to the proportion of variables that we would like to keep in the active set at each iteration. The procedure is illustrated in Figure 1. It represents the empirical cumulative distribution of sorted magnitudes of $|\mathbf{X}^T\mathbf{y}|$ from the real data D_{NIR} described later in Section (4.2). Fixing a shriking ratio ς of expected zero coefficients (e.g. $\varsigma = 80\,\%$), we select the threshold ν at iteration m as depicted. As the cumulative destribution is non-decreasing, we choose the first x-axis value corresponding to ordinate 0.8.

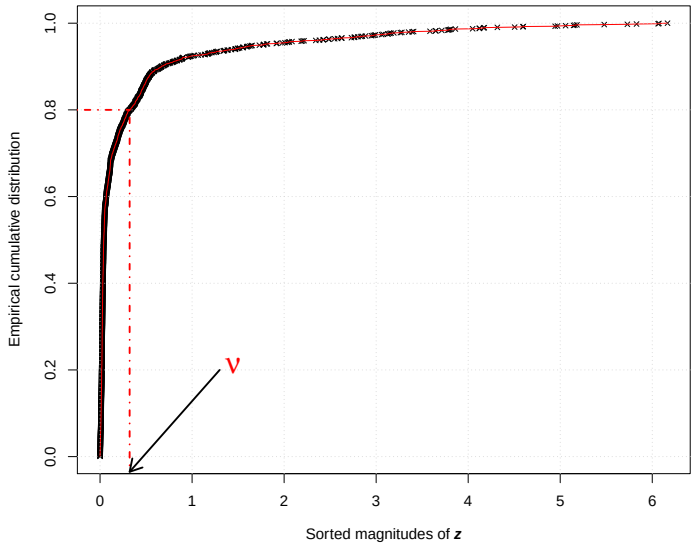

Figure 1: Empirical cumulative distribution of the sorted magnitude of $\mathbf{z} = \mathbf{X}^T\mathbf{y}$ (black crossed connected by solid red line) from real data D_{NIR}. Dotted red line illustrated the selection of appropriate ν for 80 % of sparsity.

To guarantee the unit norm property for $\mathbf{w}$, we set $\mu = \|\mathbf{z}_\nu\|_2$ where $\mathbf{z}_\nu$ is the vector of coordinates $\delta_p(|z_p| - \nu)_+$ for $p \in \{1, \ldots, P\}$. Consequently,

$$\mathbf{w} = \frac{\mu}{\nu\|\mathbf{z}_\nu\|_1 + \|\mathbf{z}_\nu\|_2^2}\mathbf{z}_\nu.$$

The rationale behind constrainting the direction $\mathbf{w}$ instead of the regression coefficients $\hat{\boldsymbol{\beta}}$ is their collinearity. Indeed, the estimator writes $\hat{\boldsymbol{\beta}} = \mathbf{W}(\mathbf{T}^T\mathbf{T})^{-1}\mathbf{T}^T\mathbf{y}$. Being collinear, soft-thresholding $\mathbf{w}$ performs a variable

selection at the same location in $\hat{\boldsymbol{\beta}}$ coordinates. The pseudo-lasso Dual-sPLS is described in Algorithm 2.

Algorithm 2 Dual-sPLS$_\text{l}$

Input: $\mathbf{X}, \mathbf{y}, M$ (number of components desired), ς (shrinking ratio)
$\mathbf{X}_1 = \mathbf{X}$
for $m = 1, \ldots, \text{M}$ **do**
  $\mathbf{z}_m = \mathbf{X}_m^T \mathbf{y}$ (weight vector)
  Find ν adaptatively according to ς
  $\mathbf{z}_\nu = (\delta_p(|z_p| - \nu)_+)_p$ (applying the threshold), $p \in \{1, \ldots, P\}$
  $\mu = \|\mathbf{z}_\nu\|_2$ and $\lambda = \dfrac{\nu}{\mu}$
  $\mathbf{w}_p = \dfrac{\|\mathbf{z}_\nu\|_2}{\nu\|\mathbf{z}_\nu\|_1 + \|\mathbf{z}_\nu\|_2^2}\mathbf{z}_\nu$ (loadings)
  $\mathbf{t}_m = \mathbf{X}_m \mathbf{w}_m$ (component)
  $\mathbf{X}_{m+1} = \mathbf{X}_m - \mathcal{P}_{\mathbf{t}_m}\mathbf{X}_m$ (deflation)
end for
$\hat{\boldsymbol{\beta}} = \mathbf{W}(\mathbf{T}^T\mathbf{T})^{-1}\mathbf{T}^T\mathbf{y}$

Note that as long as $\mathbf{w}$ and $\mathbf{z}_\nu$ are collinear, the sparsity of the results only requires the computation of $\mathbf{w}$, up to a non-zero factor.

3.2.2. Pseudo-group lasso

Response $\mathbf{y}$ may be explained separately by explanatory variables of different nature with prediction models. Combining them appropriately is potentially beneficial both in predictive and interpretative powers. The same reasoning could be used to partition the dataset into groups.
Physico-chemical motivation resides in segmenting a spectrum into homogenous bands or combining complementary modalities (e.g. IR and NMR) to predict the same property (e.g. viscosity, density). We consider G groups, and $\mathbf{z}_g$ sub-vector of $\mathbf{z}$ denotes variables belonging to group g. The group lasso inspired norm is expressed as in Equation (23). The closed-form solution is collinear to the vector $\mathbf{z}_{\nu_g}$. It is given by

$$\mathbf{z}_{\nu_g} = \boldsymbol{\delta}_g(|\mathbf{z}_g| - \nu_g)_+ \quad \text{and } \mathbf{z}_\nu = \left(\mathbf{z}_{\nu_g}\right)_{g\in\{1,\ldots,G\}}, \tag{30}$$

$\boldsymbol{\delta}_g$ being the vector of signs of $\mathbf{w}_g$ and $\nu_g = \lambda_g \mu$ for $g \in \{1, \ldots, G\}$. Each group is driven by its own threshold ν_g. The latter can be obtained similarly

as in Section 3.2.1. Note that this Dual-sPLS version reduces to the pseudo-lasso case when $G = 1$.

3.2.3. Pseudo-least squares and pseudo-ridge

The above can be generalized in many ways, by defining more versatile norm shapes, including notably weighted norms. One such possibility is $\forall \mathbf{w} \in \mathbb{R}^P$

$$\Omega(\mathbf{w}) = \lambda_1 \|\mathbf{N}_1 \mathbf{w}\|_1 + \|\mathbf{N}_2 \mathbf{w}\|_2 + \lambda_2 \|\mathbf{w}\|_2 \,. \tag{31}$$

It is not easily solvable in general. However, an appropriate choice of matrices $\mathbf{N}_1$ and $\mathbf{N}_2$, and factors λ_1 and λ_2 allow us to recover the lasso and group lasso norms, but also several other already known concepts, like fused lasso, least squares or ridge. We focus here on two main situations whose optimization problem resolution can be obtained analytically. An obvious option heavily inspired by least squares regression sets $\mathbf{N}_2 = \mathbf{X}$ and $\lambda_2 = 0$. Its resolution supplements the traditional least squares problem with a more selective shrinkage akin to that of our pseudo-lasso. Namely we first note

$$\Omega(\mathbf{w}) = \lambda \|\mathbf{N_1} \mathbf{w}\|_1 + \|\mathbf{X}\mathbf{w}\|_2 \,. \tag{32}$$

Then for $\nu = \mu\lambda$ and δ the vector of signs of $\mathbf{N_1 w}$ and $\mathbf{N_1 z}$,

$$\frac{\mathbf{w}}{\|\mathbf{X}\mathbf{w}\|_2} = (\mathbf{X}^T\mathbf{X})^{-1}\frac{\mathbf{z}}{\boldsymbol{\mu}} - \lambda(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{N}_1^T\boldsymbol{\delta} \,, \tag{33}$$

where we have implicitly assumed that $\mathbf{X}$ has full rank. Consequently, we penalize $|\hat{\boldsymbol{\beta}}^{\mathrm{LS}}|$ instead of $|\mathbf{z}|$. For equation (33) to take a genuine pseudo-lasso form, it is sufficient that $\mathbf{N}_1$ verifies

$$(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{N}_1^T\boldsymbol{\delta} = \mathrm{sign}\left((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{z}\right) . \tag{34}$$

However, as it does not play a role in loadings' computation, it does not need to be computed explicitly. Thus, the coordinates of the simplified closed-form solution is:

$$\frac{\mathbf{w}_p}{\|\mathbf{X}\mathbf{w}\|_2} = \frac{1}{\mu}\mathrm{sign}(\hat{\beta}_p^{\mathrm{LS}})(|\hat{\beta}_p^{\mathrm{LS}}| - \nu)_+ \,, \tag{35}$$

When $\mathbf{X}$ is singular, the above cannot hold. Meanwhile, this case can be addressed with a regularization inspired by the ridge [19]. By choosing $\mathbf{N}_1 = I_P$, $\mathbf{N}_2 = \lambda_2\mathbf{X}$ and $\lambda_2 = 1$, equation (31) writes

$$\Omega(\mathbf{w}) = \lambda_1 \|\mathbf{w}\|_1 + \lambda_2 \|\mathbf{X}\mathbf{w}\|_2 + \|\mathbf{w}\|_2 \,. \tag{36}$$

It amounts to penalize $|\mathbf{z}_{\nu_2}|$ where $\mathbf{z}_{\nu_2} = \left(\nu_2 \mathbf{X}^T\mathbf{X} + I_P\right)^{-1}$ and $\nu_2 = \lambda_2 \frac{\mathbf{w}}{\|\mathbf{X}\mathbf{w}\|_2}$, instead of $|\mathbf{z}|$ like in the pseudo-lasso. For $\nu_1 = \lambda_1\mu$, the closed-form solution is formulated as:

$$\frac{\mathbf{w}}{\|\mathbf{w}\|_2} = \frac{1}{\mu}\delta(|\mathbf{z}_{\nu_2}| - \nu_1)_+ \,. \tag{37}$$

where $\delta = \text{sign}(\mathbf{z}_{\nu_2}\mathbf{z})$. Adding the diagonal perturbation resolves the non-invertability of $\mathbf{X}^T\mathbf{X}$.

4. Simulated and real data, model settings, evaluation

4.1. Simulated sparse data: Gaussian mixtures D_{SIM} and $\overline{D}_{\text{SIM}}$

For an in-depth analysis of machine learning algorithms, resorting to simulated data allows an unbiased access to ground truth. We thereby propose a parametrized model. It is thought to provide similarities with common analytical chemistry data, with all sparse parameters controlled. We choose a positively weighted mixture of K Gaussians peaks with preset identical scale σ and amplitudes A and locations μ are drawn from uniform distributions. They are summed as follows:

$$\sum_{k=1}^{K} A_k \exp(-\frac{(}{x} - \mu_k)^2 2\sigma^2) \,. \tag{38}$$

and uniformly sampled. The response vector $\mathbf{y}$ is defined by an explicit linear model(affected by a stochastic Gaussian contamination) composed of weighted sums of $\mathbf{X}$ values. Weights can be random or fixed quantities by ranges of indices.

In this work, to evaluate Dual-sPLS in both precision and information location, we devise a sparse additive model with only $S \ll P$ positive weights and $P - S$ null weights. Namely, only S variables are responsible in the construction of response $\mathbf{y}$. This information is especially beneficial to demonstrate the strength of variable selection in sparse methods. Since we deal with high-dimensional situations, we simulated D_{SIM}: 300 mixtures of 30 Gaussians represented by 1000 variables (Figure 2 (left)). Highlighted red areas denote variables involved in the computation of response $\mathbf{y}$. The corresponding matrix of D_{SIM} is singular and used in the evaluation of Dual-sPLS$_{\text{l}}$ and Dual-sPLS$_{\text{r}}$. Since the Dual-sPLS$_{\text{LS}}$ is only operational with invertible matrices, we also simulated non-singular data matrix $\overline{D}_{\text{SIM}}$, 200 mixtures of

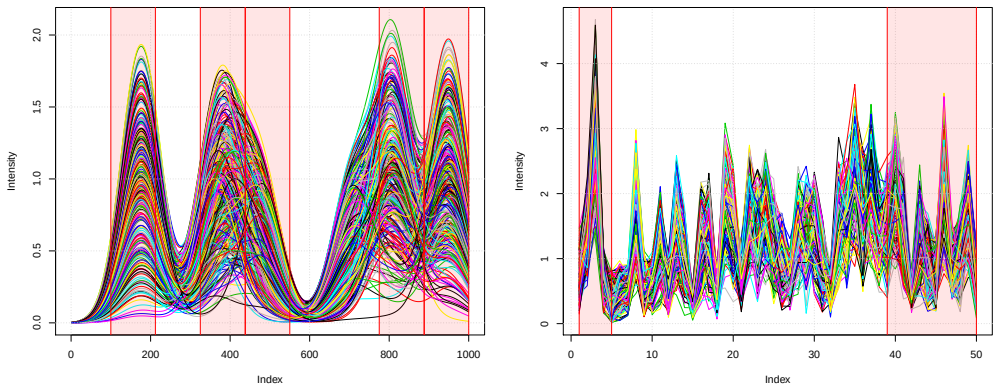

Figure 2: D_{SIM} (left) and $\overline{D}_{\mathrm{SIM}}$ (right) simulated data. Ranges of variables involved in the linear response model $\mathbf{y}$ are highlighted in red.

100 Gaussians represented by 50 variables. The response $\mathbf{y}$ corresponding to $\overline{D}_{\mathrm{SIM}}$ depends only on the first five and last twelve variables as shown in Figure 2 (right).

4.2. Real data: near-infrared (NIR) spectroscopy D_{NIR}

In chemistry, complex mixtures of molecules are analyzed with different physico-chemical methods. Besides, determining macroscopic properties is important to their use.

The evaluation on real data is done using NIR spectra of hydrocarbon samples. NIR is based on the principle of absorption of radiation (infrared) by matter [30]. Infrared radiations correspond to wavenumbers directly lesser than those of the visible light spectrum. The absorption of radiation depends on chemical bonds, therefore a NIR spectrum encodes information about the composition of the sample. We focus on the density property which is obtained by standardized methods. The IFPEN dataset D_{NIR} was partly exposed in [31, 32]. It is available at `http://www.laurent-duval.eu/opus-dual-spls-sparse-pls/` and subject to a forthcoming publication [33]. It is composed of 208 samples with 1557 variables. The corresponding matrix $\mathbf{X}$ is singular. Many chemical data require adequate preprocessing: normalization, baseline removal [34], deconvolution [7]. Here we simply apply a discrete derivative obtained with a Savitzky–Golay smoothing filter [35] of degree 2 and length 15. It serves as both a crude baseline filter and

diversity enhancement operator [36]. The NIR preprocessed dataset D_{NIR} is represented in Figure 3.

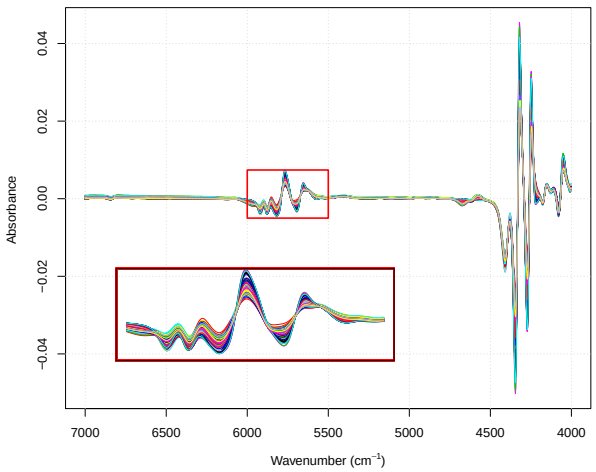

Figure 3: D_{NIR}: first Savitzky-Golay derivatives of the NIR spectra of 208 samples. Bottom subplot: magnification of the red box.

4.3. Model settings: number of latent component selection

Selecting the appropriate number M of latent components is crucial when building a regression model. It balances between model complexity and prediction accuracy (degrees of freedom), preventing the risk of overfitting. This issue is especially important when using PLS and its extensions in chemometrics. In practice, one may use this variant of a proposal in [37], based on cross-validation with multiple random split. First, observations are split randomly several times into calibration and validation sets. Second, candidate models are constructed with different numbers of latent components. Third, each prediction is evaluated on the validation set with MSE. The latter are averaged for each model. Finally, the smallest model with the lowest averaged MSE reveals an adequate number of latent components. With this method, two parameters are necessary: the splitting ratio and the number of times observations are divided.

We do not use this procedure in Section 5. We evaluate Dual-sPLS performance by comparing it to other regression methods, and exploring model

orders. We vary the number of latent components from 1 to 10 and assess each case.

4.4. Calibration and validation

The evaluation of prediction models traditionally divides the dataset into two representative sets called calibration and validation. Three main methods are used. In the first one, observations are randomly selected. The second only considers the distribution of values in the response $\mathbf{y}$ [38, Stratified sampling]. A third class is known as Kennard and Stone (KS) method [39]. It optimizes relative distances between observations according to variables of $\mathbf{X}$. In chemometrics, one may expect the existence of a yet unknown dependence between analytical measurements and properties. Taking both $\mathbf{X}$ and $\mathbf{y}$ values for a proper calibration and validation split would be desirable. The attempts of [40] to consider $\mathbf{X}$ and $\mathbf{y}$ in a single distance with appropriate weights is not straightforward. It is difficult to adequately weight variables that do not belong to the same space. We have recently proposed a CalValXy for that purpose. It consists in dividing the dataset into subgroups according to the repartition of $\mathbf{y}$ and applying the Kennard and Stone to each subgroup. It is summarized in Algorithm 3, and extensively described in [41].

Algorithm 3 Calibration and validation CalValXy

Input: $\mathbf{X}, \mathbf{X}_{type}$(index of which set belongs each observation of $\mathbf{X}$),
Listecal (number of calibration points to pick from each subset)
$G = \text{mean}(\mathbf{X})$ (centroid)
$\text{C}_1 = \max_n \|\mathbf{x}_G - \mathbf{x}_n\|$, $n \in \{1, \ldots, N\}$ (first calibration point)
$s =$ subset where C_1 is located
while Listecal is not empty **do**
  $s \leftarrow s + 1$
  Find the minmax point C in subset s
  Remove C from $\mathbf{X}$ and **Listecal**
  Store C in a vector of calibration index **cal**
end while

5. Comparative evaluation and discussion

We benchmark each proposed Dual-SPLS regression flavor (respectively pseudo-lasso, least squares and ridge) against its classical counterpart, and

comparable sparse SPLSs, when applicable. We follow a common procedure to state the main results. First, we split observations into calibration (80 %) and validation (20 %). We replaced the traditional Kennard and Stone method [39] — using explanatory variables $\mathbf{X}$ only — with CalValXy (cf. Section 4.4 and [41]). The latter incorporates the response variable $\mathbf{y}$ in the splitting and proves slightly better than KS in terms of prediction. Comparative performance is assessed in both accuracy and quality of interpretation. For the first one, common objective metrics are root mean squared error (RMSE), mean absolute error (MAE), or determination coefficient (R^2) (see end of Section 1). As metrics yield similar outcomes, we only compare, in the topmost figures, RMSE values for either calibration (left) or validation (right) CalValXy splits, as we increase the number M of latent components from one to ten. For the second one, we assess both variable selection and localization by vertically stacking regression coefficients for each compared algorithm in the bottom figure. Results are extensively discussed on simulated and real data for Dual-SPLS$_\text{l}$, and in less details for the least squares and ridge flavors. Complementary outcomes are provided in the supplementary materials.

5.1. Dual-sPLS pseudo-lasso evaluation (D_SIM, D_NIR)

Dual-sPLS$_\text{l}$ is compared to standard PLS, three alternative sparse PLS (sPLS$_\text{LeCao}$ [20], sPLS$_\text{Chun}$ [23], sPLS$_\text{Durif}$ [24]) and lasso [6]. Their respective parameters are selected by cross-validation (Section 4.1). Both sPLS$_\text{LeCao}$ and Dual-sPLS$_\text{l}$ explicitely specify a sparsity parameter: the (approximate) proportion of variables ς to be discarded (ℓ_0^c/P). We set it here to 99 %.
We first evaluate Dual-sPLS$_\text{l}$ on simulated data D_SIM (Section 4.1) in Figure 4. Top-left and right plots entail that accuracy (RMSE) globally improves as the number of latent variables M increases for all five PLS-related methods — in both calibration and validation. The lasso performance, independent on the number of components, is represented by the sixth dotted curve. From six to ten latent variables, all curves tend to plateau, with close RMSE values. Dual-SPLS$_\text{l}$, sPLS$_\text{Chun}$ and PLS provide the best results (lowest curves). Thus, adding more components seems uncessary. We choose six latent variables to compare coefficient localization. On Figure 4-bottom, we stack seven panels: original spectra (1) and the coefficients for: PLS (2), Dual-sPLS$_\text{l}$ (3), sPLS$_\text{LeCao}$ (4), sPLS$_\text{Chun}$ (5), sPLS$_\text{Durif}$ (6), lasso (7). PLS coefficients (panel 2) match the shape of the simulated data (panel 1). However, it fails to localize the most important variables, unlike sparse PLS. The ℓ_0 criterion

(Section 1) quantifies the sparsity induced by each method. Dual-sPLS$_\text{l}$, sPLS$_\text{LeCao}$ and lasso perform best, selecting as expected a small number of variables, with an ℓ_0 value around 40 to 60. It however is not sufficient to hint at improvements in interpretability. Looking only at variables affecting the response (shaded red background in panel 1), most compared methods exhibit significant coefficients in many (useless) areas (transparent background). Only Dual-sPLS$_\text{l}$, sPLS$_\text{LeCao}$ present concentrated coefficients that can help chemical interpretation. On this rudimentary yet explainable model, we hint that Dual-sPLS$_\text{l}$ provides a predictive quality comparable to its challengers, and is the best in providing at the same time accurate localization on simulated data, with a verifiable (yet simplified) prediction model. We are now able to evaluate the performance of Dual-sPLS$_\text{l}$ on real near-infrared data D_NIR (4.2) for density prediction. Similarly to D_SIM, RMSE curves in Figure 5 for calibration (top-left) and validation (top-right) globally decrease with an increasing number of components. Errors plateau after six components, indicating that additional latent structure orders might be weakly helpful. The performance gap for sPLS$_\text{Durif}$ could occur as it was mainly designed for classification. Again, we assess model interpretation in Figure 5 (bottom) for six latent vectors. By nature, location of the most influential features of spectra for a specific property is yet to be unveiled. One may expect that most of the meaningful variables are located in the active parts of the signal, e.g. spectral bands with relatively higher intensities, with some others possibly in quieter wavenumber ranges. On the top panel, NIR spectra are mainly active[3] from $4000\,\text{cm}^{-1}$ to $4800\,\text{cm}^{-1}$ and $5500\,\text{cm}^{-1}$ to $6000\,\text{cm}^{-1}$. Meaningful PLS coefficients are visible on a much wider support, provoking ambiguity on the identification of spectral bands related to density. All sPLS actually have smaller support, sPLS$_\text{LeCao}$ and Dual-sPLS$_\text{l}$ being the sparsest with ℓ_0 respectively equal to 88 and 82. The first singularity of Dual-sPLS$_\text{l}$ is the contiguous and smoothness of its coefficients. By contrast, sPLS$_\text{Chun}$ and sPLS$_\text{Durif}$ coefficients location appear to be more scattered across the wavenumber axis, in non-contiguous small chunks and even isolated spikes. The second is the absence of response in the $5500\,\text{cm}^{-1}$ to $6000\,\text{cm}^{-1}$ bands[3] in Dual-sPLS$_\text{l}$. We are not able to chemically explain the discrepancy of absence/presence results in this band. However, Dual-sPLS$_\text{l}$

[3]We do not endeavour a chemical explanation here. It ought to be substanciated in forthcoming paper [33]

does not need it to remain almost as accurate as its competitors.

5.2. Dual-sPLS pseudo-least squares evaluation ($\overline{D}_{\mathrm{SIM}}$)

The Dual-sPLS$_{\mathrm{LS}}$ requires data to be represented by a non-singular matrix $\mathbf{X}$, as explained in Section 3.2.3. Since real data D_{NIR} is singular, we use simulated data $\overline{D}_{\mathrm{SIM}}$ presented in Section 4.1. As the number of variables in $\overline{D}_{\mathrm{SIM}}$ is already small, we only shrink 60 % of its variables to evaluate the Dual-sPLS$_{\mathrm{LS}}$ against classical least squares. The latter is denoted by dashes, as the number of latent components is meaningless in this case.

For calibration (Figure 6 top-left) the RMSE for Dual-sPLS$_{\mathrm{LS}}$ decreases mildly as the number of components increases. It approaches the least squares performance. For validation (Figure 6 top-right) Dual-sPLS$_{\mathrm{LS}}$ performs similarly or better than least squares all over model orders. This contrast in performance might be explained by a tendency to overfit for least squares. A better prediction performance is expected with our model. Similarly to the Dual-sPLS$_{\mathrm{l}}$, we also choose to evaluate it with six components in the bottom of Figure 6. Again, redish regions indicate active variables for the unknown linear model. We observe an overall similarity in the dynamics of both regression coefficients: strong amplitude in the first five and last ten variables corresponding to active regions. The main difference resides in the intermediate part, irrelevant to the response. Least squares as expected shrinks inactive variables towards zero but not as much as Dual-sPLS$_{\mathrm{LS}}$ does. This is exemplified in the zoomed panels, where Dual-sPLS$_{\mathrm{LS}}$ exhibit much less non-zero coefficients.

5.3. Dual-sPLS pseudo-ridge evaluation (D_{SIM}, D_{NIR})

Dual-sPLS$_{\mathrm{r}}$ is compared to classical ridge regression (Section 2.2) either applied to simulated data D_{SIM} or real data D_{NIR}. Ridge hyper parameter t (equation (16)) is fixed using cross-validation. We set λ_2 for Dual-sPLS$_{\mathrm{r}}$ (equation (25)) to $\dfrac{1}{t}$ for easier comparison. All other parameters are kept as for Dual-sPLS$_{\mathrm{l}}$ (Section 5.1). Looking at top-left and -right in Figure 7 Dual-sPLS$_{\mathrm{r}}$ reaches a plateau for D_{SIM} after five latent components. Moreover, its RMSE values are slightly lower than ridge's for both calibration and validation. We can safely select six latent components as before. Reference coefficients for ridge are misleading because the largest ones do not reside in influencing areas. They therefore can not be used for data interpretation.

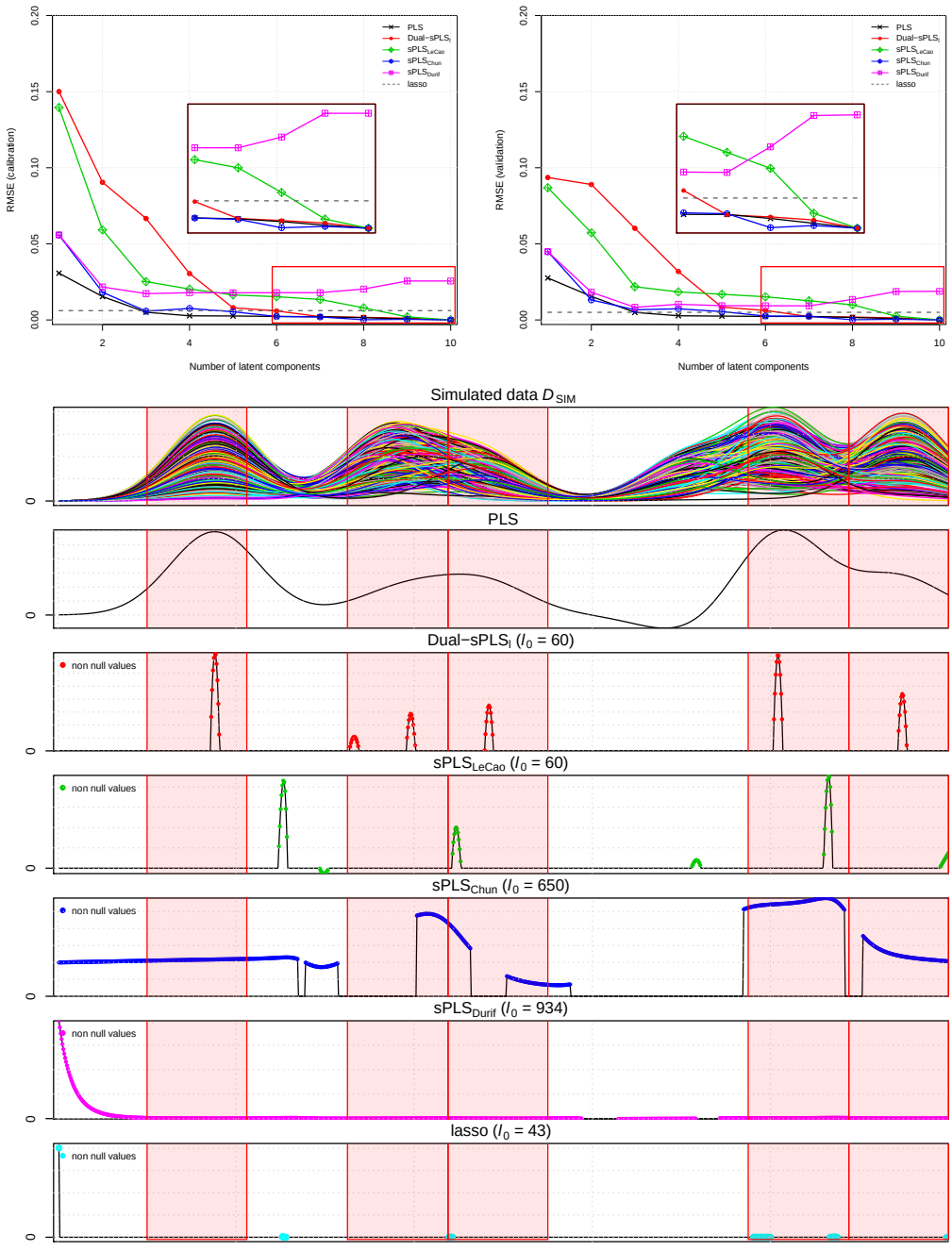

Figure 4: Dual-sPLS$_l$ evaluation on simulated data D_{SIM}. (Top) RMSE values for calibration (left) and validation (right) with respect to the number of latent components. (Bottom) From top to bottom: simulated data D_{SIM}, regression coefficients of PLS, Dual-sPLS$_l$, sPLS$_{\mathrm{LeCao}}$, sPLS$_{\mathrm{Chun}}$, sPLS$_{\mathrm{Durif}}$ for six components, and lasso.

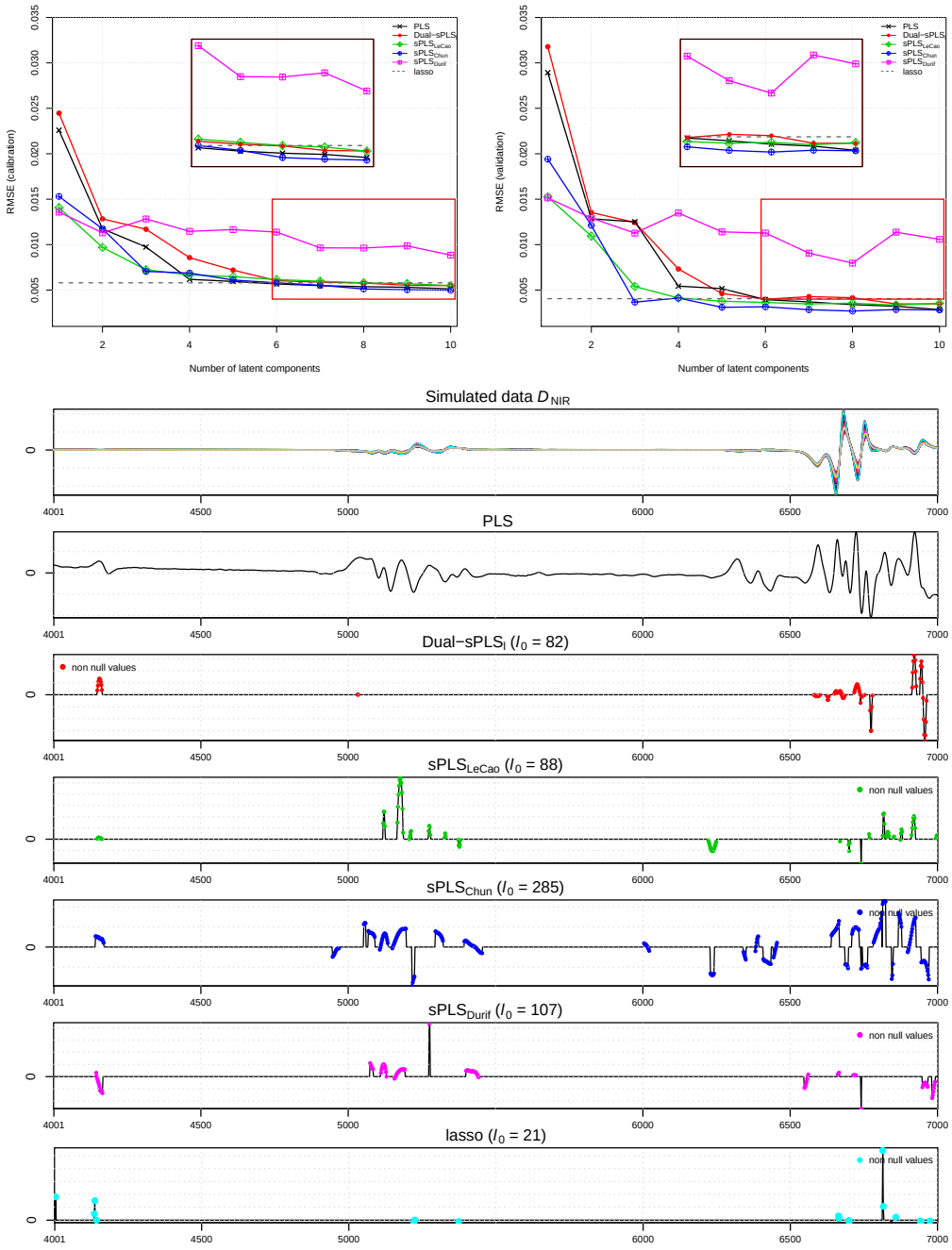

Figure 5: Dual-sPLS$_\text{l}$ evaluation on real data D_NIR. (Top) RMSE values for calibration (left) and validation (right) with respect to the number of latent components. (Bottom) From top to bottom: original data D_NIR, regression coefficients of PLS, Dual-sPLS$_\text{l}$, sPLS$_\text{LeCao}$, sPLS$_\text{Chun}$, sPLS$_\text{Durif}$ for six components, and lasso.

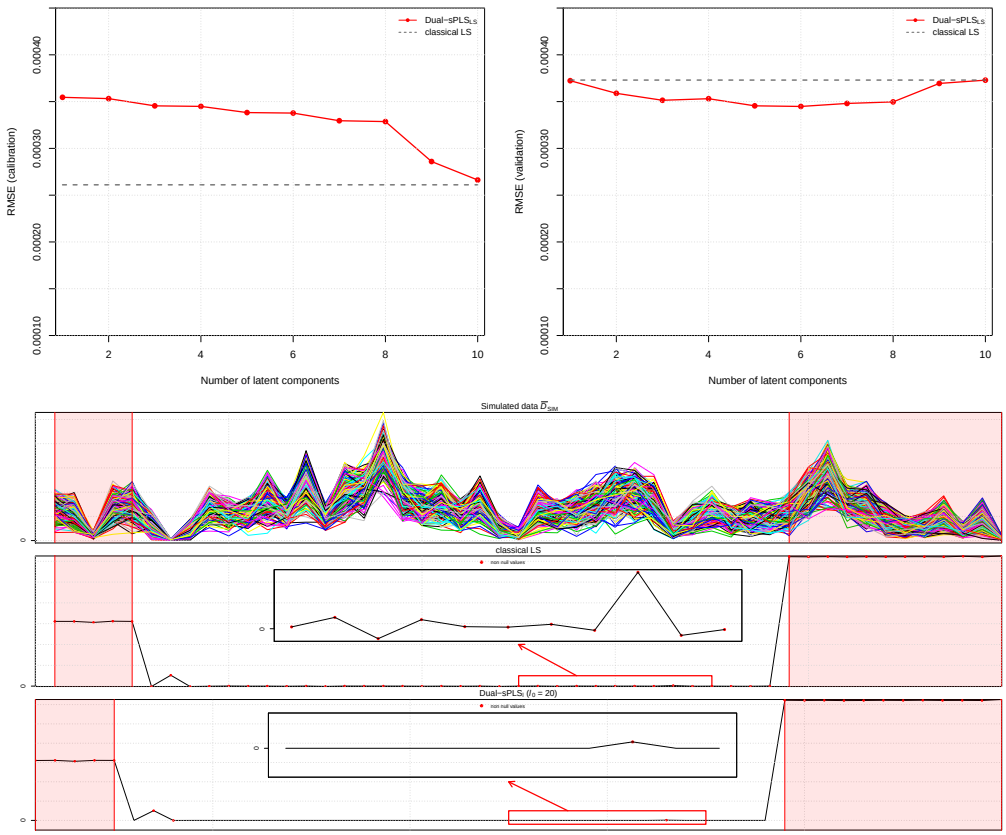

Figure 6: Dual-sPLS$_{\mathrm{LS}}$ evaluation on simulated data $\overline{D}_{\mathrm{SIM}}$. (Top) RMSE values for calibration (left) and validation (right) with respect to the number of latent components. (Bottom) From top to bottom: simulated data $\overline{D}_{\mathrm{SIM}}$, regression coefficients of least squares and Dual-sPLS$_{\mathrm{LS}}$ for five components.

By selecting only fifty variables, located in red regions governing the model, Dual-sPLS$_\mathrm{r}$ better succeeds in both prediction and localization. Similar conclusions can be drawn for real data D_{NIR} on RMSE values. Dual-sPLS$_\mathrm{r}$ even better predicts the response $\mathbf{y}$ with only four components. Regression coefficients (Figure 8 bottom) yield comments akin to above. While ridge apparently emphasizes unimportant features, Dual-sPLS$_\mathrm{r}$ seems more reliable in identifiying of relevant variables to predict density using chemical data.

6. Conclusion and perspectives

We propose a family of dual sparse Partial Least Squares algorithms that broadens the compass of standard PLS. Along with competitive prediction accuracy with respect to PLS as used in chemometrics, we expect additional benefits in dimension reduction or model interpretability. This is achieved by supplementing the traditional optimization problem with well-chosen dual norms.

We chiefly validate this approach by borrowing three classical regression penalties: lasso, least-squares, ridge. Each proposed Dual-sPLS draws close to the reference in calibration/validation performance with a reduced number of latent components. This is assessed in a benchmark on both realistic simulated models and real near infrared spectroscopy data, against a standard baseline and sparse contenders. Coefficients are sieved with a user-defined sparsity target. They are well-located in influential data ranges, suggesting a means for better interpretability of the trained prediction reduced model. Pseudo-lasso and ridge Dual-sPLS avatars exhibit close colocation of selected features in both datasets despite different penalties. This suggests a robust identification of meaningful information in signals.

The Dual-sPLS framework is thus a good candidate for a host of applications. We provide it as an open-source package in R [27]. It can be prolonged to other field-favorite penalties, for instance elastic net. We plan to evaluate the alluded "pseudo-group lasso" option, to refine feature selection on important contiguous areas, or to combine datasets providing complementary information on the predicted response. To improve prediction robustness or reduce the number of necessary latent components (toward three or four instead of six), we explore additional diversity enhancement preprocessing, such as higher-order derivatives and discrete wavelet transforms. Last, as PLS deserves sounder statistical foundations, we endeavor a study of asymptotic convergence bounds.

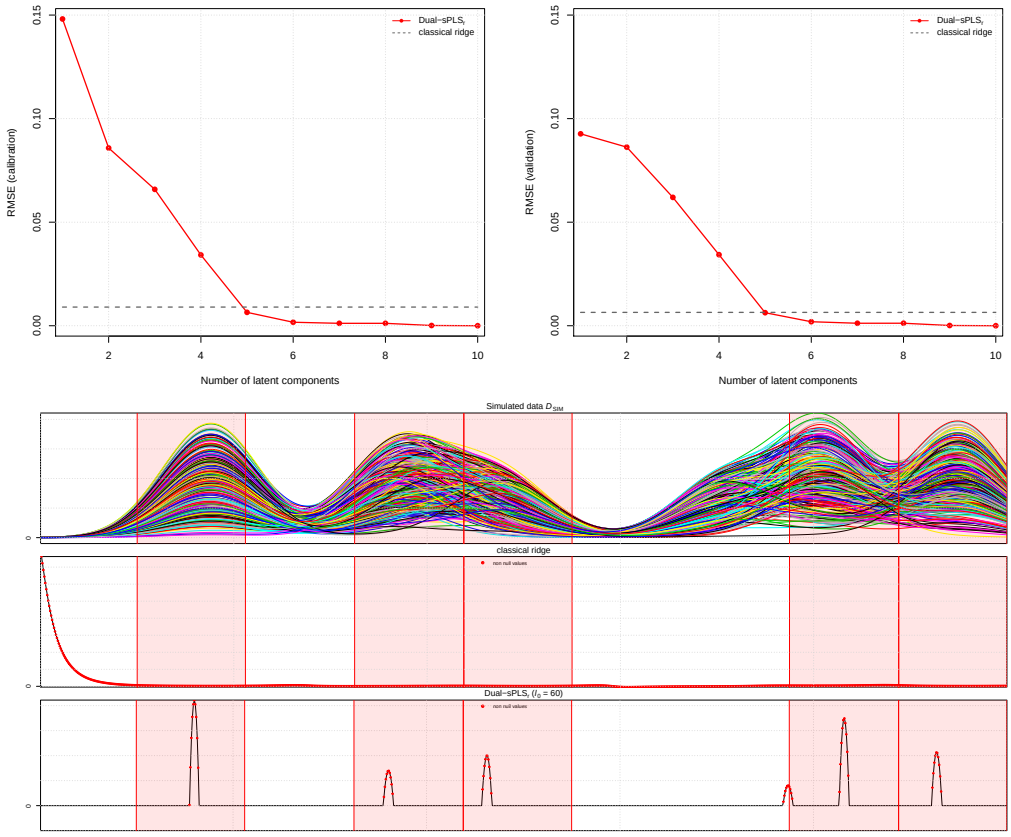

Figure 7: Dual-sPLS$_\mathrm{r}$ evaluation on simulated data D_{SIM}. (Top) RMSE values for calibration (left) and validation (right) with respect to the number of latent components. (Bottom) From top to bottom: original data D_{SIM}, regression coefficients of ridge and Dual-sPLS$_\mathrm{r}$ for five components.

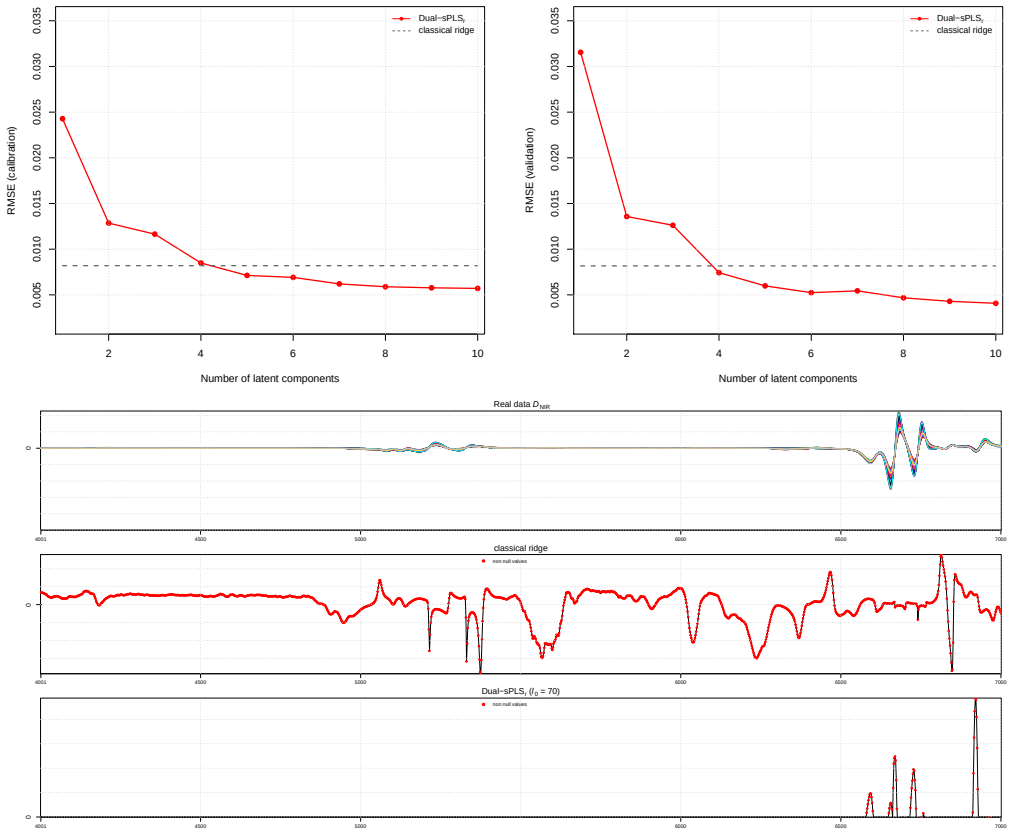

Figure 8: Dual-sPLS$_\mathrm{r}$ evaluation on real data D_{NIR}. (Top) RMSE values for calibration (left) and validation (right) with respect to the number of latent components. (Bottom) From top to bottom: original data D_{NIR}, regression coefficients of ridge and Dual-sPLS$_\mathrm{r}$ for five components.

7. Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

8. Acknowledgements

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). We acknowledge the financial support of the Research Council of the Saint Joseph University of Beirut. Ghislain Durif contributed in the code validation procedure in R. IFPEN provided the real NIR data set used in the applications. We thank Noémie Caillol, Luca Castelli and Irène Gannaz for useful comments. Our mentor, colleague or friend François Wahl passed away unexpectedly during the writing of this paper. He was the driving force of our team.

References

[1] L. F. Ramírez-Verduzco, J. E. Rodríguez-Rodríguez, A. del Rayo Jaramillo-Jacob, Predicting cetane number, kinematic viscosity, density and higher heating value of biodiesel from its fatty acid methyl ester composition, Fuel 91 (1) (2012) 102–111. doi:10.1016/j.fuel.2011.06.070.

[2] H. H. Willard, L. L. Merritt, J. A. Dean, Méthodes physiques de l'analyse chimique, Dunod, 1965.

[3] S. Verdier, J. A. P. Coutinho, A. M. S. Silva, O. F. Alkilde, J. A. Hansen, A critical approach to viscosity index, Fuel 88 (11) (2009) 2199–2206. doi:10.1016/j.fuel.2009.05.016.

[4] S. Wold, Chemometrics; what do we mean with it, and what do we want from it?, Chemometr. Intell. Lab. Syst. 30 (1995) 109–115.

[5] S. Wold, M. Sjöström, L. Eriksson, PLS-regression: a basic tool of chemometrics, Chemometr. Intell. Lab. Syst. 58 (2) (2001) 109–130. doi:10.1016/s0169-7439(01)00155-1.

[6] R. Tibshirani, Regression shrinkage and selection via the lasso, J. R. Stat. Soc. Ser. B Stat. Methodol. 58 (1) (1996) 267–288.
URL http://www.ams.org/mathscinet-getitem?mr=1379242

[7] A. Cherni, E. Chouzenoux, L. Duval, J.-C. Pesquet, SPOQ ℓ_p-over-ℓ_q regularization for sparse signal recovery applied to mass spectrometry, IEEE Trans. Signal Process. 68 (2020) 6070–6084. doi:10.1109/TSP.2020.3025731.

[8] G. Mateos-Aparicio Morales, Partial least squares (PLS) methods: Origins, evolution, and application to social sciences, Commun. Stat. Theory Methods 40 (13) (2011) 2305–2317. doi:10.1080/03610921003778225.

[9] T. Mehmood, B. Ahmed, The diversity in the applications of partial least squares: an overview, J. Chemometrics 30 (1) (2016) 4–17. doi:10.1002/cem.2762.

[10] A.-L. Boulesteix, K. Strimmer, Partial least squares: a versatile tool for the analysis of high-dimensional genomic data, Brief. Bioinform. 8 (1) (2007) 32–44. arXiv:https://academic.oup.com/bib/article-pdf/8/1/32/737013/bbl016.pdf, doi:10.1093/bib/bbl016.
URL https://doi.org/10.1093/bib/bbl016

[11] A. Krishnan, L. J. Williams, A. R. McIntosh, H. Abdi, Partial least squares (PLS) methods for neuroimaging: A tutorial and review, Neuromimage 56 (2) (2011) 455–475. doi:10.1016/j.neuroimage.2010.07.034.

[12] J. Wright, Y. Ma, High-Dimensional Data Analysis with Low-Dimensional Models: Principles, Computation, and Applications, Cambridge University Press, 2022.
URL https://www.ebook.de/de/product/41126998/john_wright_yi_ma_high_dimensional_data_analysis_with_low_dimensional_models_principles_computation_and_applications.html

[13] H. Wold, Path models with latent variables: The NIPALS approach, in: Quantitative Sociology. International Perspectives on Mathematical and Statistical Modeling, Elsevier, 1975, pp. 307–357. doi:10.1016/b978-0-12-103950-9.50017-4.

[14] S. de Jong, SIMPLS: An alternative approach to partial least squares regression, Chemometr. Intell. Lab. Syst. 18 (3) (1993) 251–263. doi:10.1016/0169-7439(93)85002-x.

[15] U. G. Indahl, The geometry of PLS1 explained properly: 10 key notes on mathematical properties of and some alternative algorithmic approaches to PLS1 modelling, J. Chemometrics 28 (3) (2014) 168–180. arXiv:https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/pdf/10.1002/cem.2589, doi:https://doi.org/10.1002/cem.2589.
URL https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/abs/10.1002/cem.2589

[16] N. Krämer, A.-L. Boulesteix, G. Tutz, Penalized partial least squares with applications to B-spline transformations and functional data, Chemometr. Intell. Lab. Syst. 94 (1) (2008) 60–69. doi:https://doi.org/10.1016/j.chemolab.2008.06.009.
URL https://www.sciencedirect.com/science/article/pii/S0169743908001214

[17] H. Zou, T. Hastie, Regularization and variable selection via the elastic net, J. R. Stat. Soc. Ser. B Stat. Methodol. 67 (2) (2005) 301–320. doi:10.1111/j.1467-9868.2005.00503.x.

[18] T. Hastie, R. Tibshirani, M. Wainwright, Statistical Learning with Sparsity: The Lasso and Generalizations, CRC Press, 2015.

[19] A. E. Hoerl, R. W. Kennard, Ridge regression: Applications to nonorthogonal problems, Technometrics 12 (1) (1970) 69–82. doi:10.1080/00401706.1970.10488635.
URL http://dx.doi.org/10.1080/00401706.1970.10488635

[20] K.-A. Lê Cao, D. Rossouw, C. Robert-Granié, P. Besse, A sparse PLS for variable selection when integrating omics data, Stat. Appl. Genet. Mol. Biol. 7 (1) (2008) 35. doi:10.2202/1544-6115.1390.

[21] H. Shen, J. Z. Huang, Sparse principal component analysis via regularized low rank matrix approximation, J. Multivar. Anal. 99 (6) (2008) 1015–1034. doi:10.1016/j.jmva.2007.06.007.

[22] M. Tenenhaus, La régression PLS. Théorie et pratique, Éditions Technip, 1998.

[23] H. Chun, S. Keleş, Sparse partial least squares regression for simultaneous dimension reduction and variable selection, J. R. Stat. Soc. Ser. B Stat. Methodol. 72 (1) (2010) 3–25. doi:10.1111/j.1467-9868.2009.00723.x.
URL http://dx.doi.org/10.1111/j.1467-9868.2009.00723.x

[24] G. Durif, L. Modolo, J. Michaelsson, J. E. Mold, S. Lambert-Lacroix, F. Picard, High dimensional classification with combined adaptive sparse PLS and logistic regression, Bioinformatics 34 (3) (2018) 485–493. doi:10.1093/bioinformatics/btx571.

[25] F. Bach, R. Jenatton, J. Mairal, G. Obozinski, Optimization with sparsity-inducing penalties, Found. Trends Mach. Learn. 4 (1) (2012) 1–106. doi:10.1561/2200000015.

[26] R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria (2021).
URL https://www.R-project.org/

[27] L. Alsouki, F. Wahl, G. Durif, dual.spls: Dual sparse partial least squares regression, CRAN, r package version 0.1.2 (Oct. 2022).
URL https://CRAN.R-project.org/package=dual.spls

[28] N. Simon, J. Friedman, T. Hastie, R. Tibshirani, A sparse-group lasso, J. Comp. Graph. Stat.) 22 (2) (2013) 231–245. doi:10.1080/10618600.2012.681250.

[29] M. Stone, Cross-validatory choice and assessment of statistical predictions, J. R. Stat. Soc. Ser. B Stat. Methodol. 36 (2) (1974) 111–133. doi:10.1111/j.2517-6161.1974.tb00994.x.

[30] J. M. Chalmers, P. R. Griffiths (Eds.), Handbook of Vibrational Spectroscopy, Wiley, 2002.

[31] J. Laxalde, C. Ruckebusch, O. Devos, N. Caillol, F. Wahl, L. Duponchel, Characterisation of heavy oils using near-infrared spectroscopy: Optimisation of pre-processing methods and variable selection, Anal. Chim. Acta 705 (1-2) (2011) 227–234. doi:10.1016/j.aca.2011.05.048.

[32] J. Laxalde, Analyse des produits lourds du pétrole par spectroscopie infrarouge, Ph.D. thesis, Université de Lille 1 (2012).

[33] L. Duval, L. Alsouki, F. Wahl, J. Laxalde, N. Caillol, IFPEN near-infrared spectroscopy dataset for property prediction: 208 NIR hydrocarbon spectra and density response, PREPRINT.

[34] X. Ning, I. W. Selesnick, L. Duval, Chromatogram baseline estimation and denoising using sparsity (BEADS) 139 (2014) 156–167. doi:10.1016/j.chemolab.2014.09.014.

[35] A. Savitzky, M. J. E. Golay, Smoothing and differentiation of data by simplified least squares procedures, Anal. Chem. 36 (8) (1964) 1627—1639.

[36] L. K. DeNoyer, J. G. Dodd, Smoothing and derivatives in spectroscopy. doi:10.1002/0470027320.s4501.

[37] A.-L. Boulesteix, K. Strimmer, Predicting transcription factor activities from combined analysis of microarray and chip data: a partial least squares approach, Theoretical biology and medical modelling 2 (2005) 23. doi:10.1186/1742-4682-2-23.
URL https://europepmc.org/articles/PMC1182396

[38] C.-E. Särndal, B. Swensson, J. Wretman, Model Assisted Survey Sampling, Springer, 2003.

[39] R. W. Kennard, L. A. Stone, Computer aided design of experiments, Technometrics 11 (1) (1969) 137–148. doi:10.1080/00401706.1969.10490666.

[40] H. Tian, L. Zhang, M. Li, Y. Wang, D. Sheng, J. Liu, C. Wang, Weighted SPXY method for calibration set selection for composition analysis based on near-infrared spectroscopy, Infrared Phys. Technol. 95 (2018) 88–92. doi:10.1016/j.infrared.2018.10.030.

[41] L. Alsouki, L. Duval, R. El Haddad, C. Marteau, F. Wahl, CalValXy: well-balanced and stratified calibration/validation splitting using both predictors X and response y, PREPRINT.

Appendix A. Detailed resolution of Dual-sPLSs

Appendix A.1. Dual-sPLS pseudo-group lasso

We recall Equation (23): the Dual-sPLS$_{\text{gl}}$ norm case applied to optimization Problem (20). Note that here

- g represents a group of $P(g)$ index extracted from $\{1, \ldots, P\}$;
- G represents the number of groups;
- $\mathbf{w}_g$ represents the values of index g in the loading vector $\mathbf{w}$.

We denote $\mathbf{z}_g$ the variables of $\mathbf{z}$ belonging to group g. We impose $\mathbf{z}_g$ and $\mathbf{w}_g$ to be in the same orthant. Let $\boldsymbol{\delta}_g$ be their vector of signs. By differentiating equation (23) we obtain

$$\frac{\partial \Omega(\mathbf{w})}{\partial w_g} = \frac{\alpha_g w_g}{\|\mathbf{w}_g\|_2} + \alpha_g \lambda_g \delta_g \,. \tag{A.1}$$

Using Lagrange multipliers as in Section 3.1, we compare (26) to (A.1) and obtain for $g \in \{1, \ldots, G\}$:

$$\frac{\mathbf{w}_g}{\|\mathbf{w}_g\|_2} = \frac{\mathbf{z}_g}{\alpha_g \mu} - \lambda_g \delta_g \,, \tag{A.2}$$

which is simplified by

$$\frac{\mathbf{w}_g}{\|\mathbf{w}_g\|_2} = \frac{1}{\mu \alpha_g} \mathbf{z}_{\nu_g} \,, \tag{A.3}$$

where

$$\mathbf{z}_{\nu_g} = \boldsymbol{\delta}_g (|\mathbf{z}_g| - \nu_g)_+ \quad \text{for } g \in \{1, \ldots, G\} \,. \tag{A.4}$$

Here $\nu_g = \mu \alpha_g \lambda_g$ and controls the amount of variables that we would like to shrink to zero. By applying ℓ_2-norm to (A.3), we conclude that for $g \in \{1, \ldots, G\}$,

$$\mu = \sum_{g=1}^{G} \|\mathbf{z}_{\nu_g}\|_2 \qquad \text{and} \qquad \alpha_g = \frac{\|\mathbf{z}_{\nu_g}\|_2}{\mu} \,. \tag{A.5}$$

The term $\|\mathbf{w}_g\|_2$ is more involved. Thus, we simply use grid search. For each group g, ten possible values are chosen to be tested. The selection is done by detecting the maximum value of $\|\mathbf{w}_g\|_2$ for each group g, denoted

$\|\mathbf{w}_g\|_2^{max}$. The latter is computed by zeroing $\|\mathbf{w}_{g'}\|_2$ for all groups $g' \neq g$ and is expressed as:

$$\|\mathbf{w}_g\|_2^{max} = \frac{\mu}{\Omega_g(\mathbf{z}_{\nu_g})}\,. \tag{A.6}$$

Then, ten values of each group g are selected inside the interval $[0, \|\mathbf{w}_g\|_2^{max}]$. The grid search tests all the possible combinations and retains the one that allows the smallest error. We summarize the methodology with Algorithm 4.

Algorithm 4 Dual-sPLS$_{\text{gl}}$ algorithm

Input: $\mathbf{X}^1, \ldots, \mathbf{X}^G$, $\mathbf{y}$, M (number of components desired), ς (shrinking ratio), $\alpha_1, \ldots, \alpha_g$.
for $m = 1, \ldots, M$ **do**
  $\mathbf{X}_m = (\mathbf{X}^1, \ldots, \mathbf{X}^G)$ (combining data)
  $\mathbf{z}_m = \mathbf{X}_m^T \mathbf{y}$ (weight vector)
  Find ν adaptively according to ς for each group seperatly
  $\mathbf{z}_{\nu_g} = \boldsymbol{\delta}_g(|\mathbf{z}_g| - \nu_g)_+ \quad$ for $g \in \{1, \ldots, G\}$ (applying the threshold)
  $\mu = \sum_{g=1}^{G} \|\mathbf{z}_{\nu_g}\|_2$
  $\alpha_g = \dfrac{\|\mathbf{z}_{\nu_g}\|_2}{\mu}$ and $\lambda_g = \dfrac{\nu_g}{\alpha_g \mu} \quad$ for $g \in \{1, \ldots, G\}$
  $\|\mathbf{w}_g\|_2^{max} = \dfrac{\mu}{\Omega_g(\mathbf{z}_{\nu_g})} \quad$ for $g \in \{1, \ldots, G\}$
  selection of the values of $\|\mathbf{w}_g\|_2$ for each group
  $\mathbf{w}_g = \dfrac{\|\mathbf{w}_g\|_2}{\mu \alpha_g} \mathbf{z}_{\nu_g} \quad$ for $g \in \{1, \ldots, G\}$ (loadings)
  $\mathbf{w}_g = \left(\mathbf{w}_g \right)_{g=1}^{G}$
  $\mathbf{t}_m = \mathbf{X}_m \mathbf{w}_m$ (component)
  $\mathbf{X}_{m+1} = \mathbf{X}_m - \mathcal{P}_{\mathbf{t}_m} \mathbf{X}_m$ (deflation)
end for
Compute $\hat{\boldsymbol{\beta}}$.

Appendix A.2. Dual-sPLS pseudo-least squares

We recall Equation (24): the Dual-sPLS$_{\text{LS}}$ pseudo case applied to optimization Problem (20).
We impose $\mathbf{N}_1\mathbf{z}$ and $\mathbf{N}_1\mathbf{w}$ to be in the same orthant. Let $\boldsymbol{\delta}_2$ be their vector

of signs. By differentiating (24) we obtain

$$\nabla\Omega(\mathbf{w}) = \lambda \mathbf{N}_1^T \boldsymbol{\delta}_2 + \frac{\mathbf{X}^T\mathbf{X}\mathbf{w}}{\|\mathbf{X}\mathbf{w}\|_2}\,. \tag{A.7}$$

Using Lagrange multipliers as in Section 3.1, we compare (26) to (A.7) and obtain

$$\frac{\mathbf{w}}{\|\mathbf{X}\mathbf{w}\|_2} = (\mathbf{X}^T\mathbf{X})^{-1}\frac{\mathbf{z}}{\boldsymbol{\mu}} - \lambda(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{N}_1^T\boldsymbol{\delta}_2\,, \tag{A.8}$$

imposing the invertibility of $\mathbf{X}^T\mathbf{X}$. We choose $\mathbf{N}_1$ such as

$$(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{N}_1^T\boldsymbol{\delta}_2 = \operatorname{sign}\Big((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{z}\Big)\,. \tag{A.9}$$

The resolution steps are be similar to the ones from Dual-sPLS$_\text{l}$ but instead of applying the threshold on $\mathbf{z}$, we apply it on $(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{z}$ which is exactly the classical Least Squares regression coefficients $\hat{\boldsymbol{\beta}}^{LS}$. So, the simplified solution is

$$\frac{\mathbf{w}}{\|\mathbf{X}\mathbf{w}\|_2} = \frac{1}{\mu}\operatorname{sign}(\hat{\boldsymbol{\beta}}_{LS_j})(|\hat{\boldsymbol{\beta}}_{LS_j}| - \nu)_+\,, \tag{A.10}$$

where ν is chosen adaptively.
For a simpler algorithm, $\|\mathbf{X}\mathbf{w}\|_2$ is not computed as it is not mandatory in this case. Additionally, $\mathbf{w}$ only depends on ν and $\hat{\beta}_{LS}$, which means $\mathbf{N}_1$ does not intervene in the computation of the optimal solution. Thus, proving that $\mathbf{N}_1$ exists is enough. (A.9) implies the following

$$\mathbf{N}_1^T\boldsymbol{\delta}_2 = (\mathbf{X}^T\mathbf{X})\operatorname{sign}\Big((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{z}\Big)\,. \tag{A.11}$$

Let $\mathbf{w}$ be an eignvector of $\mathbf{N}_1$, and $\mathbf{N}_1'$ be such as

$$\mathbf{N}_1' = \mathbf{N}_1 - \mathbf{w}^T\mathbf{w} \quad \text{ and } \mathbf{N}_1'\mathbf{w} = 0\,. \tag{A.12}$$

Therefore, using (A.11) we have

$$\mathbf{N}_1'\boldsymbol{\delta}_2 = (\mathbf{X}^T\mathbf{X})\operatorname{sign}\Big((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{z}\Big) - \mathbf{w}\mathbf{w}^T\boldsymbol{\delta}_2 \quad \text{ with } \mathbf{N}_1'\mathbf{w} = 0\,. \tag{A.13}$$

With $\mathbf{N}_1$ a square matrix of P variables, (A.13) is a system of P^2 unknowns, P equations and P contraints. It can be verified by an infinite number of solutions.

The following algorithm reformulates the previous steps:

Algorithm 5 DUAL-sPLS$_{\text{LS}}$ ALGORITHM

Input: $\mathbf{X}, \mathbf{y}, M$ (number of components desired), ς (shrinking ratio)
$\mathbf{X}_1 = \mathbf{X}$
for $m = 1, \ldots, M$ **do**
  $\mathbf{z}_m = \mathbf{X}_m^T \mathbf{y}$ (weight vector)
  $\hat{\beta}_{LS} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{z}$
  Find ν adaptively according to ς and $\hat{\beta}_{LS}$
  $\mathbf{z}_\nu = (\text{sign}(\hat{\boldsymbol{\beta}}_{LS})(|\hat{\boldsymbol{\beta}}_{LS}| - \nu)_+)$ (applying the threshold)
  $\mathbf{w}_m = \dfrac{\mathbf{z}_\nu}{\mu}$ (loadings)
  $\mathbf{w}_m = \dfrac{\mathbf{w}_m}{\|\mathbf{w}\|_2}$ (normalizing loadings)
  $\mathbf{t}_m = \mathbf{X}_m \mathbf{w}_m$ (component)
  $\mathbf{X}_{m+1} = \mathbf{X}_m - \mathcal{P}_{\mathbf{t}_m}\mathbf{X}_m$ (deflation)
end for
Compute $\hat{\boldsymbol{\beta}}$.

Appendix A.3. Dual-sPLS pseudo-ridge

We recall Equation (25): the Dual-sPLS$_{\text{r}}$ pseudo case applied to optimization Problem (20). We impose $\mathbf{z}$ and $\mathbf{w}$ to be in the same orthant. Let $\boldsymbol{\delta}$ be their vector of signs. By differentiating (25), we obtain

$$\nabla\Omega(\mathbf{w}) = \lambda_1\delta + \lambda_2 \frac{\mathbf{X}^T\mathbf{X}\mathbf{w}}{\|\mathbf{X}\mathbf{w}\|_2} + \frac{\mathbf{w}}{\|\mathbf{w}\|_2}\,. \tag{A.14}$$

Using Lagrange multipliers as in Section 3.1, we compare (26) to (A.14) and obtain

$$\frac{\mathbf{w}}{\|\mathbf{w}\|_2} = \left(\nu_2\mathbf{X}^T\mathbf{X} + I_P\right)^{-1}(\mathbf{z} - \nu_1\boldsymbol{\delta})\,, \tag{A.15}$$

where $\nu_1 = \lambda_1\mu$ and $\nu_2 = \lambda_2\dfrac{\|\mathbf{w}\|_2}{\|\mathbf{X}\mathbf{w}\|_2}$.

In line with Dual-sPLS$_{\text{l}}$, we note $\mathbf{z}_{\mathbf{X},\nu_2} = \left(\nu_2\mathbf{X}^T\mathbf{X} + I_P\right)^{-1}\mathbf{z}$ and $\boldsymbol{\delta}_{\mathbf{X}}$ its vector of signs. We exhibit a solution imposing that $\mathbf{w}$ and $\mathbf{z}_{\mathbf{X},\nu_2}$ are in the same orthant, which leads to the following reformulation of (A.15):

$$\frac{\mathbf{w}}{\|\mathbf{w}\|_2} = \frac{1}{\mu}\delta_{\mathbf{X}}(|\mathbf{z}_{\mathbf{X},\nu_2}| - \nu_1)_+\,. \tag{A.16}$$

The threshold ν_1 is chosen with the adaptive procedure described in Section 3.2.1 and Figure 1. However, in this case, we compare ν_1 to $|\mathbf{z}_{\mathbf{X},\nu_2}|$. Since the latter is colinear to $\mathbf{z}$, the shrinkage is adequate. Denoting $\mathbf{z}_\nu = \delta_{\mathbf{X}}(|\mathbf{z}_{\mathbf{X},\nu_2}| - \nu_1)_+$, simple computations lead to

$$\mu = \|\mathbf{z}_\nu\|_2 \,, \tag{A.17}$$

and

$$\mathbf{w} = \frac{\mu}{\nu_1\|\mathbf{z}_\nu\|_1 + \nu_2\|\mathbf{X}\mathbf{z}_\nu\|_2^2 + \mu^2} \,. \tag{A.18}$$

It is summarized in Algorithm 6:

Algorithm 6 DUAL-sPLS$_{\text{R}}$ ALGORITHM

Input: $\mathbf{X}, \mathbf{y}, M$ (number of components desired), ς (shrinking ratio), ν_2
$\mathbf{X}_1 = \mathbf{X}$
for $m = 1, \ldots, M$ **do**
  $\mathbf{z}_m = \mathbf{X}_m^T\mathbf{y}$ (weight vector)
  $\mathbf{z}_{\mathbf{X},\nu_2} = \left(\nu_2\mathbf{X}^T\mathbf{X} + I_P\right)^{-1}\mathbf{z}$
  Find ν adaptively according to ς and $|\mathbf{z}_{\mathbf{X},\nu_2}|$
  $\boldsymbol{\delta}_{\mathbf{X}}$ vector of signs of $\mathbf{z}_{\mathbf{X},\nu_2}$
  $\mathbf{z}_\nu = \delta_{\mathbf{X}}(|\mathbf{z}_{\mathbf{X},\nu_2}| - \nu_1)_+$ (applying the threshold)
  $\mu = \|\mathbf{z}_\nu\|_2$ and $\lambda = \dfrac{\nu}{\mu}$
  $\mathbf{w}_m = \dfrac{\mu}{\nu_1\|\mathbf{z}_\nu\|_1 + \nu_2\|\mathbf{X}\mathbf{z}_\nu\|_2^2 + \mu^2}$ (loadings)
  $\mathbf{t}_m = \mathbf{X}_m\mathbf{w}_m$ (component)
  $\mathbf{X}_{m+1} = \mathbf{X}_m - \mathcal{P}_{\mathbf{t}_m}\mathbf{X}_m$ (deflation)
end for
Compute $\hat{\boldsymbol{\beta}}$.

Appendix B. Complementary plots

As mentioned in Section 5, metrics MAE and R^2 were also computed. They support our findings based on RMSE, as they yield similar results.

Figures B.14 and B.15 represent a clearer perpective on regression coefficients for Dual-sPLS$_{\text{l}}$ applied D_{SIM} and D_{NIR} from Section 5.1.

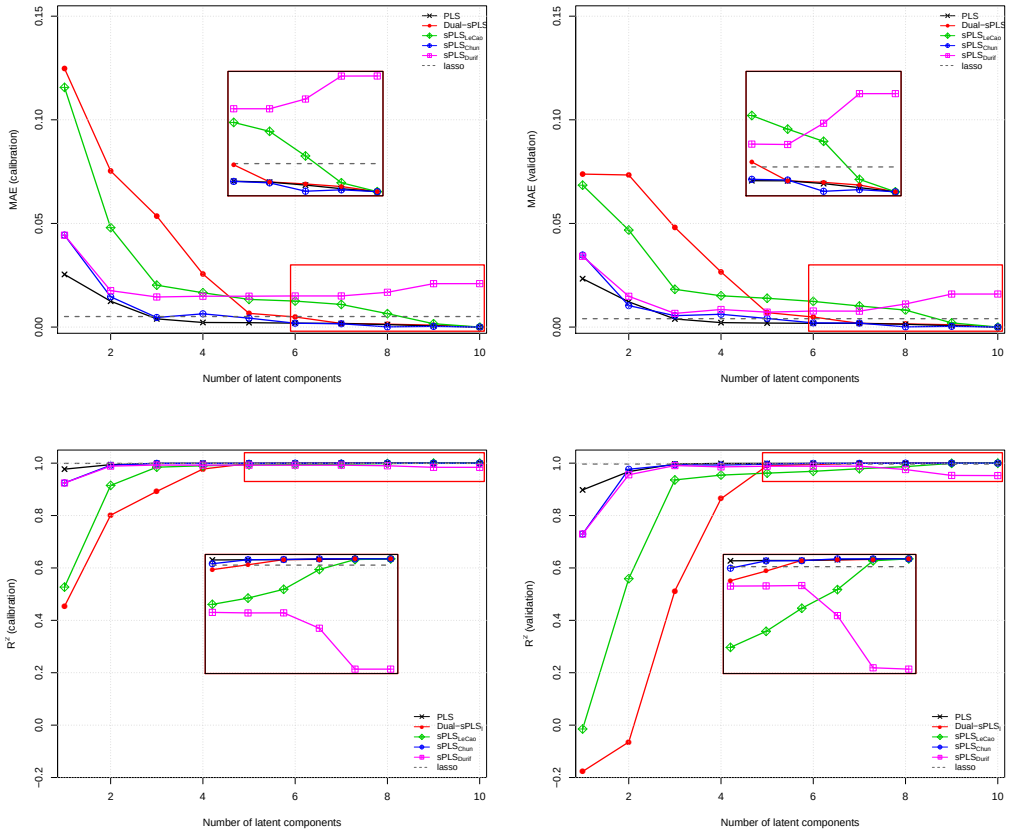

Figure B.9: Dual-sPLS$_l$ evaluation on simulated data D_{SIM}. MAE (top) and R^2 (bottom) values for calibration (left) and validation (right) with respect to the number of latent components derived from PLS, Dual-sPLS$_l$, sPLS$_{\text{LeCao}}$, sPLS$_{\text{Chun}}$, sPLS$_{\text{Durif}}$ and lasso regressions.

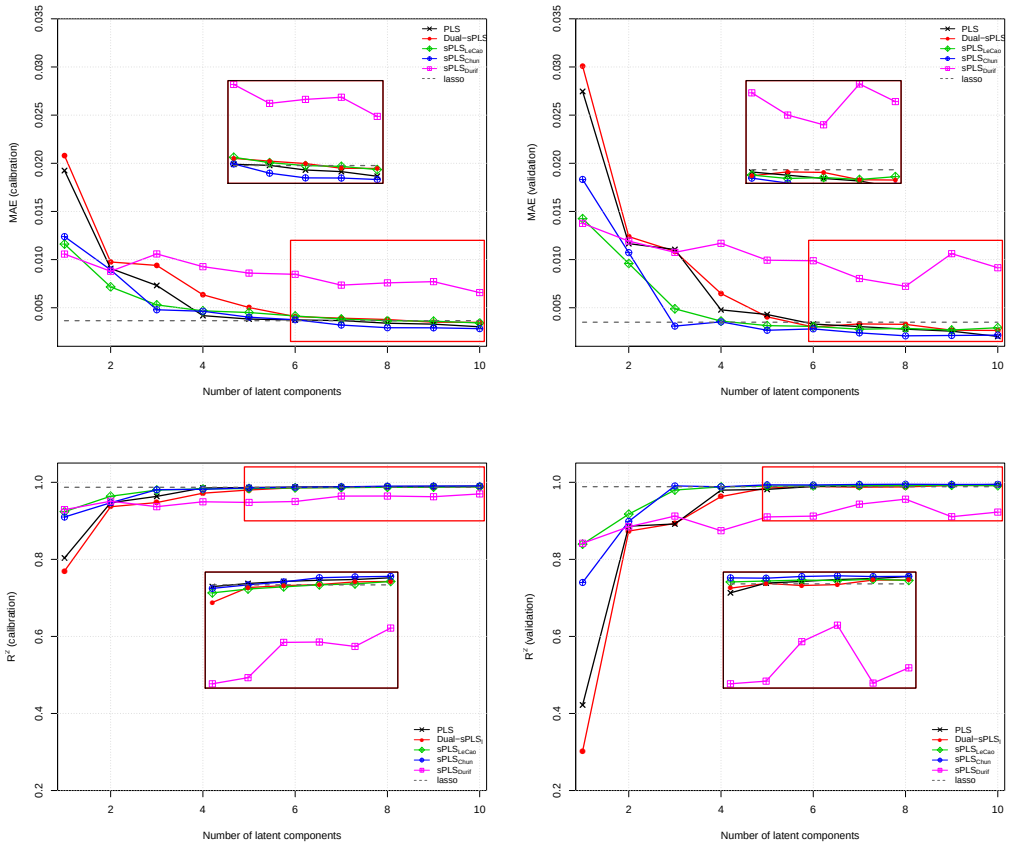

Figure B.10: Dual-sPLS$_l$ evaluation on real data D_{NIR}. MAE (top) and R^2 (bottom) values for calibration (left) and validation (right) with respect to the number of latent components derived from PLS, Dual-sPLS$_l$, sPLS$_{\text{LeCao}}$, sPLS$_{\text{Chun}}$, sPLS$_{\text{Durif}}$ and lasso regressions.

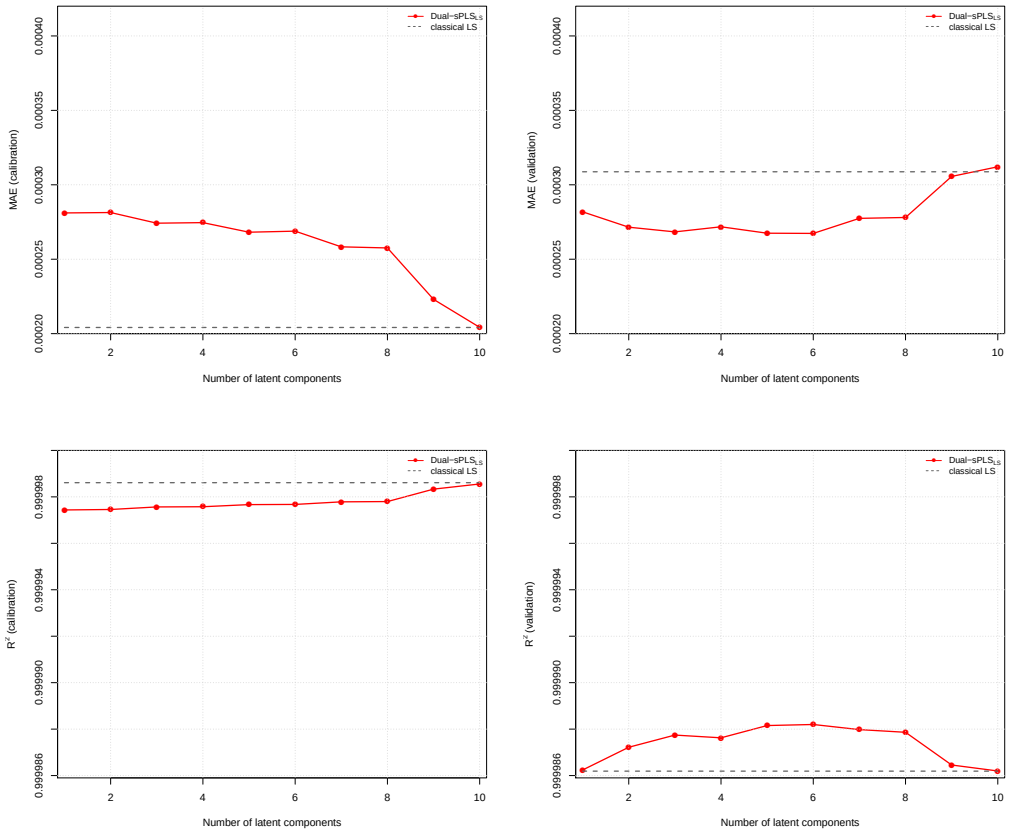

Figure B.11: Dual-sPLS$_{\mathrm{LS}}$ evaluation on simulated data $\overline{D}_{\mathrm{SIM}}$. MAE (top) and R^2 (bottom) values for calibration (left) and validation (right) with respect to the number of latent components derived from Dual-sPLS$_{\mathrm{LS}}$ and least squares regressions.

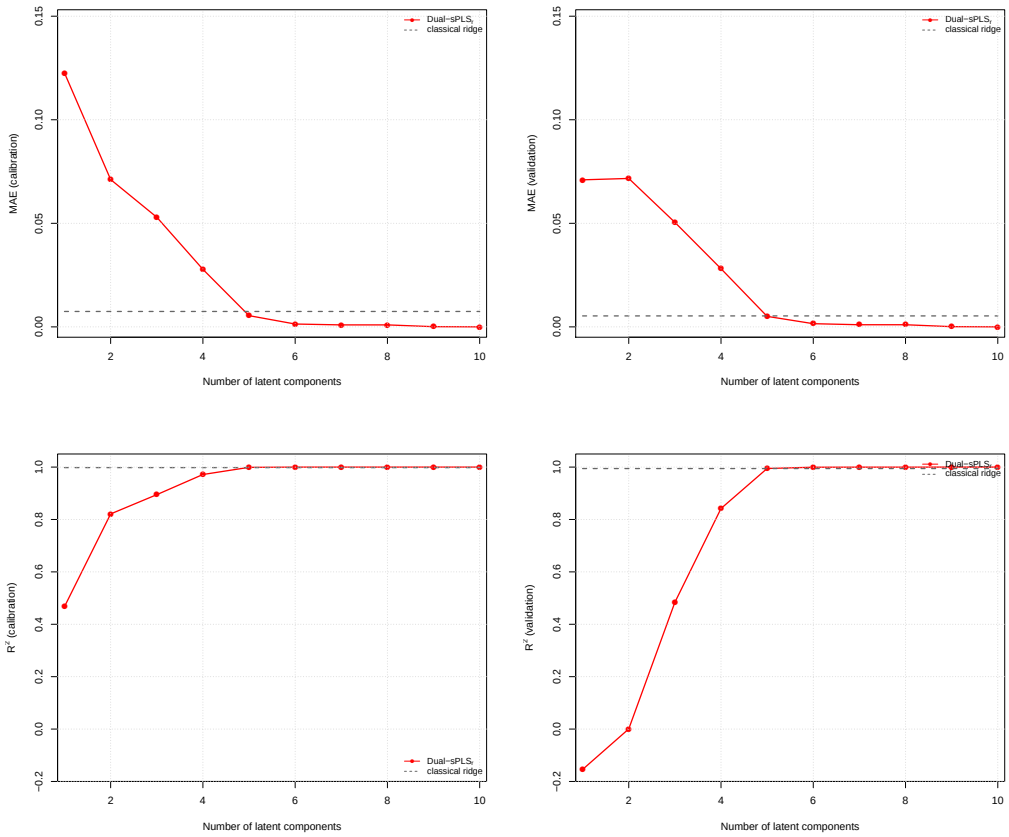

Figure B.12: Dual-sPLS$_\mathrm{r}$ evaluation on simulated data D_{SIM}. MAE (top) and R^2 (bottom) values for calibration (left) and validation (right) with respect to the number of latent components derived from Dual-sPLS$_\mathrm{r}$ and ridge regressions.

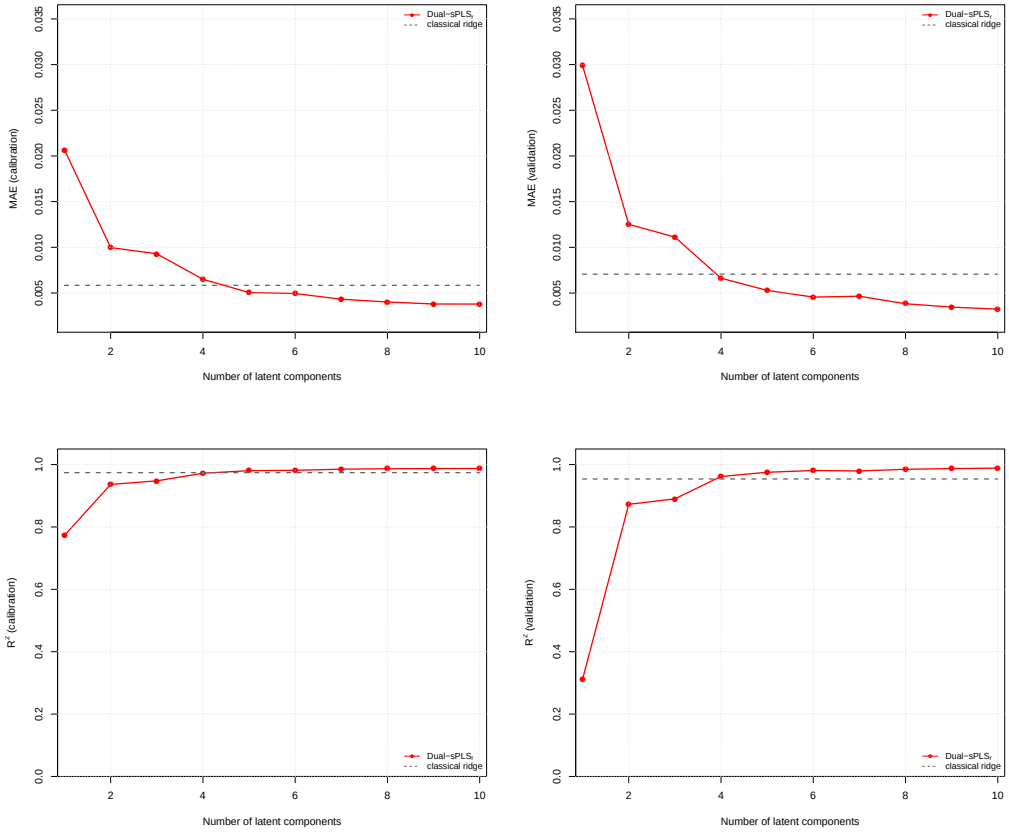

Figure B.13: Dual-sPLS$_\mathrm{r}$ evaluation on real data D_{NIR}. MAE (top) and R^2 (bottom) values for calibration (left) and validation (right) with respect to the number of latent components derived from Dual-sPLS$_\mathrm{r}$ and ridge regressions.

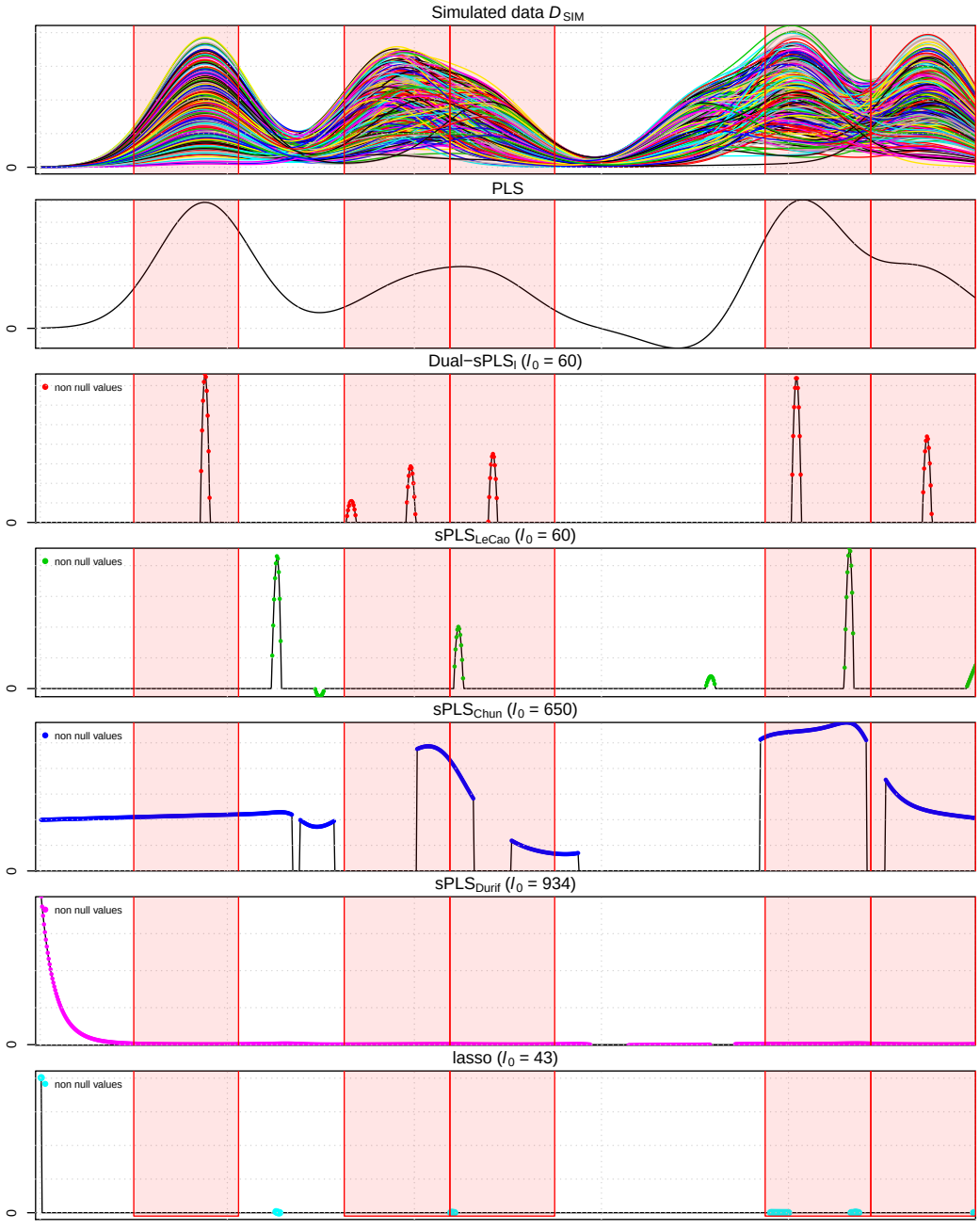

Figure B.14: Dual-sPLS$_l$ evaluation on simulated data D_{SIM}. From top to bottom: simulated data D_{SIM}, regression coefficients of PLS, Dual-sPLS$_l$, sPLS$_{\mathrm{LeCao}}$, sPLS$_{\mathrm{Chun}}$, sPLS$_{\mathrm{Durif}}$ for six components, and lasso.

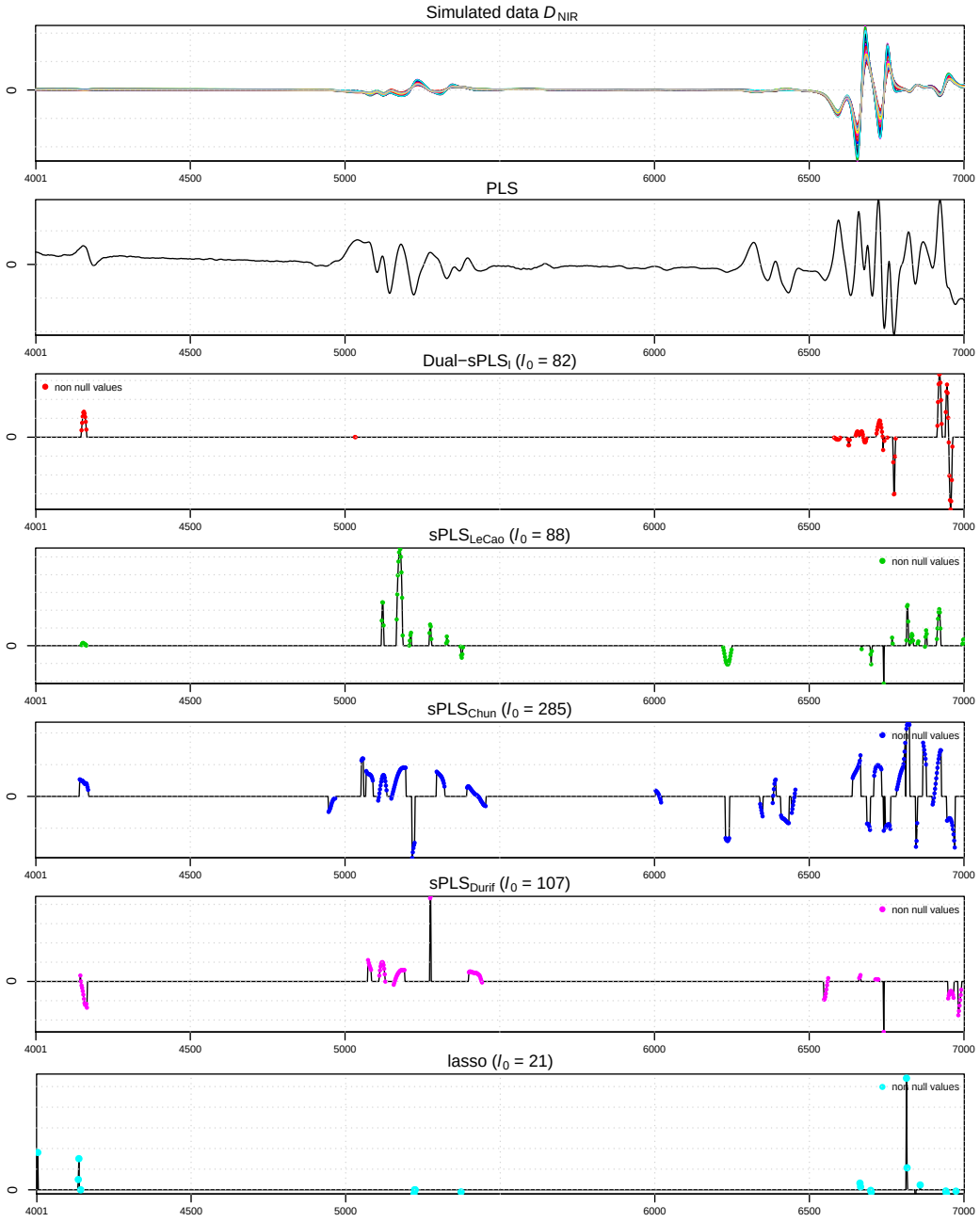

Figure B.15: Dual-sPLS$_l$ evaluation on real data D_{NIR}. From top to bottom: simulated data D_{SIM}, regression coefficients of PLS, Dual-sPLS$_l$, sPLS$_{\mathrm{LeCao}}$, sPLS$_{\mathrm{Chun}}$, sPLS$_{\mathrm{Durif}}$ for six components, and lasso.