Simultaneous Variable Selection for the Classification of Near Infrared Spectra

Leila Belmerhnia[b,a], El-Hadi Djermoune[a,*], Cédric Carteret[c], David Brie[a]

[a] *Université de Lorraine, CNRS, CRAN, F-54000 Nancy, France*
[b] *TVT Innovation, Maison du numérique et de l'innovation, Place Georges Pompidou, F-83000, Toulon, France*
[c] *Université de Lorraine, CNRS, LCPME, F-54000 Nancy, France*

Abstract

We propose a simultaneous variable selection method for material sorting based on near infrared spectroscopic data. Our objective is to perform fast classification in industrial wood recycling processes based only on a few spectral bands. The spectra are first jointly modeled as linear combinations of explanatory variables drawn from a collection of Gaussian-shaped functions. The aim is to select a common subset of wavebands shared by several spectra. The variable selection is then formulated as an unconstrained simultaneous sparse approximation problem in which the coefficients related to different spectra are encouraged to be piecewise constant, *i.e.* the coefficients associated to successive spectra should have comparable magnitudes. We also investigate the case where the coefficients are constrained to be nonnegative. These problems are solved using the fast iterative shrinkage-thresholding algorithm. The proposed approaches are illustrated on a dataset of 290 spectra of wood wastes; each spectrum is composed of 1647 wavelengths. We show that the selected variables lead to better classification performances as compared to standard approaches.

Keywords: Variable selection; Simultaneous sparse approximation; NIR spectroscopy; Classification

1. Introduction

Near infrared (NIR) spectroscopy is a vibrational spectroscopy which provides information about the molecular composition and interactions within the studied material sample [1, 2]. As the sample spectrum is a kind of signature characterizing the material, NIR spectroscopy is used in a wide range of applications, including material identification, characterization and non destructive evaluation [3, 4]. In this work the targeted application is material (wood wastes) sorting, which is envisaged as a supervised classification problem. To face the curse of dimensionality and avoid overfitting problems, feature selection has been recognized as a key step, especially when the variables are highly correlated. The problem is to find a small subset of variables describing the main characteristics of the different classes. Popular features selection approaches include best subset selection [5, 6], forward and backward stepwise regression [7, 8], forward stagewise regression and sparse linear regression known as the Lasso (*Least absolute shrinkage and selection operator*) [9].

Sparse representations have been widely studied over the last decade, and applied to different problems such as data compression [10, 11], pattern recognition [12], classification and clustering [13, 14], and hyperspectral image unmixing and classification [15, 16, 17]. It is based on the assumption that the essential characteristics of the data can be approximated by a linear combination of a few atoms (also named explanatory variables) drawn from an overcomplete dictionary. Feature selection may also be viewed as dimensionality reduction problem that can be tackled using a sparse approximation. The idea is simply to select the set of atoms corresponding to nonzero coefficients resulting from the approximation. More recently, group sparsity was introduced to enforce certain structural constraints. Restricting our attention to simultaneous sparsity, a particular instance of group sparsity, we seek at finding a set of coefficients explaining jointly the observed variables. As it involves an ℓ_0 "norm", solving the exact simultaneous sparse approximation problem yields to an NP-hard problem for which the greedy methods provide a good compromise between reconstruction accuracy and computational cost [18, 19, 20]. Convex relaxations of the simultaneous sparse approximation was also proposed in [21].

In this paper, we propose a simultaneous variable selection strategy for NIR spectra based on sparse decomposition. Given a set of training spectra, the core idea consists in finding a small subset of wavebands/variables that captures the main spectral components shared by several measurements. The wavebands are picked from a dictionary containing Gaussian features of various centers and widths. The regression coefficients associated to the selected variables may then be used to perform classification of candidate spectra. Some similar approaches have been already proposed in the literature. For example, Turlach *et al.* [22] presented a simultaneous variable selection algorithm and applied it to NIR spectra. Unlike [22], the sparse decomposition problem considered in this work incorporates a regularization term enforcing the rows of the coefficient matrix to be

☆This work is supported by the French FUI AAP15 Trispirabois project funded by BPI France and Région Lorraine.

*Corresponding author

Email addresses: belmerhnia@tvt.fr (Leila Belmerhnia), el-hadi.djermoune@univ-lorraine.fr (El-Hadi Djermoune), cedric.carteret@univ-lorraine.fr (Cédric Carteret), david.brie@univ-lorraine.fr (David Brie)

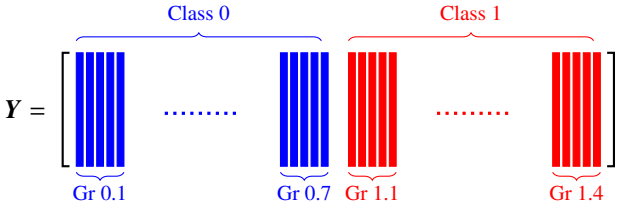


Figure 1: Illustration of spectra ordering in data matrix $\boldsymbol{Y}$. The spectra are ordered according to their groups and class labels.

piecewise constant. The intuition behind the proposed approach is quite simple. Consider the situation where the samples can be divided into groups and that a class includes different groups. Rather than randomly gathering the samples into the data matrix, we propose to order them according of their group labels. Figure 1 illustrates this ordering for a two class problem with 11 groups. Consecutive samples belong to the same group and are expected to share common features. This will be captured by enforcing piecewise constant coefficients and group sparsity. Malli and Natschläger [23] also proposed a waveband selection algorithm for spectroscopy based on fused Lasso [24]. The fused penalty encourages the selection of connected wavelengths resulting in the so-called "wavebands". On the contrary, the method presented here consists in modeling the spectra with Gaussian-shaped functions. By doing so, not only the algorithm is structurally able to select wavebands rather than individual wavelengths but it also allows to reduce the number of spectral features (variables). These properties are particularly suitable for high speed industrial classification because the computational cost of the regression coefficients, associated to a small number of variables, is pretty low.

The paper is organized as follows. In Section 2, we present the regularized simultaneous sparse approximation problem involving an ℓ_0/ℓ_2 mixed pseudo-norm. The details of the convex relaxation approaches are described in Section 3. Specifically, we first propose the ℓ_1/ℓ_1 relaxation and then the ℓ_1/ℓ_2 surrogate of the ℓ_0/ℓ_2 norm. Both relaxations lead to algorithms that enjoy a decomposition property allowing one to compute an efficient solution even for large scale problems. We also propose a nonnegative version of these algorithms. An application to wood wastes sorting based on NIR measurements is provided in Section 4. Finally, conclusions and drawn in Section 5.

Notation. Scalars are denoted by regular letters (N, s, λ), column vectors by lower-case bold-face letters ($\boldsymbol{x}, \boldsymbol{\phi}$), and matrices as bold-face capitals ($\boldsymbol{X}, \boldsymbol{\Phi}$). $\boldsymbol{x}_i$ is i-th column of $\boldsymbol{X}$ and $\boldsymbol{x}^i$ denotes the transpose of the i-th row. Notation $(\cdot)^\top$ stands for matrix or vector transposition. $\|\boldsymbol{A}\|_{p,q}$ is the mixed ℓ_p/ℓ_q-norm and $\|\boldsymbol{A}\|_F$ is the Frobenius norm of matrix $\boldsymbol{A}$. The symbols "$\otimes$", "$*$", and "$\circ$" denote the Kronecker product, the Hadamard (entrywise) product, and the composition operator, respectively.

2. Problem formulation

Suppose that K response variables (spectra) are collected and stacked in the columns of a data matrix $\boldsymbol{Y} \in \mathbb{R}^{M\times K}$ where M is the number of observations in each spectrum. The matrix $\boldsymbol{Y}$ is assumed to be group ordered as illustrated in fig. 1. We seek to decompose the matrix $\boldsymbol{Y}$ such that:

$$\boldsymbol{Y} \approx \boldsymbol{\Phi}\boldsymbol{X}, \tag{1}$$

where $\boldsymbol{X} \in \mathbb{R}^{N\times K}$ is a *sparse* coefficient matrix meaning that only a small subset of its rows is nonzero. The columns of the redundant dictionary $\boldsymbol{\Phi} = [\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_N] \in \mathbb{R}^{M\times N}$ represents the explanatory variables (also called *atoms*). This dictionary is designed to concentrate the energy of the signals in $\boldsymbol{Y}$ over a set of a few atoms. Its choice depends essentially on the application at hand. As in NIR spectra, the observed peaks are typically very broad, we assume in the present work that the $\boldsymbol{\phi}_n$'s are Gaussiand-shaped functions whose locations (central wavelengths) and widths cover all the NIR range. In other words, each atom is used as a model representing the most significant spectral bands in the available data.

The simultaneous sparse approximation [25, 26] consists in finding a solution $\boldsymbol{X}$ having a limited number of active rows. The problem can be formulated as

$$\underset{\boldsymbol{X}}{\text{minimize}} \quad \tfrac{1}{2}\|\boldsymbol{Y} - \boldsymbol{\Phi}\boldsymbol{X}\|_F^2, \tag{2a}$$

$$\text{subject to} \quad \|\boldsymbol{X}\|_{0,2} \leq s, \tag{2b}$$

where $\|\boldsymbol{X}\|_{0,2}$ is the mixed ℓ_0/ℓ_2 pseudo-norm of $\boldsymbol{X}$ (*i.e.* the number of rows with nonzero ℓ_2-norm) and $s \ll N$ is the sparsity parameter which is related to the support of $\boldsymbol{X}$: $\text{supp}(\boldsymbol{X}) = \{1 \leq n \leq N \,|\, \boldsymbol{x}^n \neq \boldsymbol{0}\}$. The rationale behind simultaneous reconstruction for variable selection is to find predictors for all input signals in $\boldsymbol{Y}$ from a *common subset* of active variables [22] which are indexed by the support of the solution $\boldsymbol{X}$.

The regularized simultaneous sparse approximation aims at reconstructing piecewise constant rows. In that respect, as in [23], we propose to include a regularization term leading to the following problem

$$\underset{\boldsymbol{X}}{\text{minimize}} \quad \tfrac{1}{2}\|\boldsymbol{Y} - \boldsymbol{\Phi}\boldsymbol{X}\|_F^2 + \lambda_2\|\boldsymbol{D}\boldsymbol{X}^\top\|_{1,1}, \tag{3a}$$

$$\text{subject to} \quad \|\boldsymbol{X}\|_{0,2} \leq s, \tag{3b}$$

where $\boldsymbol{D} \in \mathbb{R}^{(K-1)\times K}$ is a matrix of finite differences of order 1:

$$\boldsymbol{D} = \begin{bmatrix} -1 & 1 & & \boldsymbol{0} \\ & \ddots & \ddots & \\ \boldsymbol{0} & & -1 & 1 \end{bmatrix}. \tag{4}$$

Criterion (3) includes an additional total variation-like penalty enforcing sparsity in the differences between successive columns of $\boldsymbol{X}$: $\|\boldsymbol{D}\boldsymbol{X}^\top\|_{1,1} = \sum_{i=1}^{K-1}\|\boldsymbol{x}_{i+1} - \boldsymbol{x}_i\|_1$. This, in fact, encourages the reconstruction of piecewise constant rows. It is important to note that this penalty makes sense only if the signals in $\boldsymbol{Y}$ have a meaningful ordering. This is for example the case when the signals are ordered according to their class or group labels in the training phase. In hyperspectral image classification, the signals are naturally ordered according to their spatial position. Due to the ℓ_1 penalty, consecutive columns of $\boldsymbol{X}$ tend to be similar which helps to decrease intragroup variance. Following the terminology of the fused Lasso, this term

will be referred to as the fusion penalty. The formulation in (3) involving the mixed ℓ_0/ℓ_2 norm will lead to an NP-hard problem, thus making the resolution not easy. A suboptimal but simple approach to solve (3) was proposed in [27]. In the next section, we propose a convex relaxation of the ℓ_0/ℓ_2 norm and the resulting problem is solved using fast and effective algorithms.

3. Convex relaxations

3.1. Fused Sparse Lasso

The first relaxation of the constrained problem (3) consists in minimizing the following penalized criterion:

$$J_{FSL}(\boldsymbol{X}) = \tfrac{1}{2}\|\boldsymbol{Y} - \boldsymbol{\Phi}\boldsymbol{X}\|_F^2 + \lambda_1\|\boldsymbol{X}\|_{1,1} + \lambda_2\|\boldsymbol{D}\boldsymbol{X}^\top\|_{1,1} \quad (5)$$

where $\|\boldsymbol{X}\|_{1,1} = \sum_{i=1}^{K}\|\boldsymbol{x}_i\|_1 = \sum_{n=1}^{N}\|\boldsymbol{x}^n\|_1$. The parameters $\lambda_1, \lambda_2 \geq 0$ are controlling the tradeoff between data fitting, the sparsity term $\|\boldsymbol{X}\|_{1,1}$, and the fusion penalty $\|\boldsymbol{D}\boldsymbol{X}^\top\|_{1,1}$. This criterion is in fact an extension to the multiple measurement vector setting of the sparse fused Lasso which was studied in [24] and solved using a two-phase active set algorithm [28] designed for quadratic programming problems with linear sparsity constraints. In [29], the problem is extended to general graphs where the fusion term is promoting constant coefficients over neighboring variables. This approach does not include the additional sparsity term. It is referred to as the generalized fused Lasso (GFL). In [30] it is proposed to solve the generalized sparse fused Lasso problem (including both sparsity and fusion terms) in the special case where the dictionary $\boldsymbol{\Phi}$ is an identity matrix. This work was then extended in [31] to general dictionary. Here, we propose to solve this problem in the special case where the fusion term only acts on the rows of $\boldsymbol{X}$. According to this specific structure it is possible to obtain a computationally efficient implementation of the minimization problem using the proximal gradient method FISTA (Fast Iterative Shrinkage-Thresholding Algorithm) [32]. However, before going further, let us give a few comments on criterion (5). In fact, the sparsity term $\|\boldsymbol{X}\|_{1,1}$ does not correspond to a proper convex relaxation of $\|\boldsymbol{X}\|_{0,2}$. As will be explained in section 3.2, the mixed norm $\|\boldsymbol{X}\|_{1,2}$ is more appropriate. But combining $\|\boldsymbol{X}\|_{1,1}$ to $\|\boldsymbol{D}\boldsymbol{X}^\top\|_{1,1}$ yields to a kind of simultaneous sparse approximation: the simultaneity is actually enforced by the row regularization term $\|\boldsymbol{D}\boldsymbol{X}^\top\|_{1,1}$, but there is no direct control on the number of active rows.

Let vec$(\cdot)$ be the vectorization operator that converts a matrix into a vector by stacking the columns of the matrix on top of one another. We set $\boldsymbol{x} = \text{vec}(\boldsymbol{X}^\top)$ and $\boldsymbol{y} = \text{vec}(\boldsymbol{Y}^\top)$. Then, criterion (5) can be rewritten as:

$$J_{FSL}(\boldsymbol{x}) = \tfrac{1}{2}\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2^2 + \lambda_1\|\boldsymbol{x}\|_1 + \lambda_2\|\boldsymbol{F}\boldsymbol{x}\|_1 \quad (6)$$

with $\boldsymbol{A} = \boldsymbol{\Phi} \otimes \boldsymbol{I}_K \in \mathbb{R}^{NK\times MK}$, $\boldsymbol{F} = \boldsymbol{I}_N \otimes \boldsymbol{D} \in \mathbb{R}^{N(K-1)\times NK}$, and $\boldsymbol{I}_N$ denotes the $N \times N$ identity matrix. Note that (6) is also a generalization to the multiple measurement vector setting of the fused Lasso criterion already proposed for variable selection in spectroscopy by Malli and Natschläger [23]. To minimize (6), we use FISTA which is an extension of the Nesterov's gradient-based method ISTA used to solve convex optimization problems including both smooth and non-smooth terms. Let

$$f(x) = \tfrac{1}{2}\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2^2, \quad (7)$$

$$g(\boldsymbol{x}) = \lambda_1\|\boldsymbol{x}\|_1 + \lambda_2\|\boldsymbol{F}\boldsymbol{x}\|_1. \quad (8)$$

Then, following [31], the update of vector $\boldsymbol{x}$ at iteration $k+1$ is:

$$\boldsymbol{x}_{(k+1)} = \arg\min_{\boldsymbol{x}\in\mathbb{R}^{NK}} \left(g(\boldsymbol{x}) + \tfrac{L}{2}\|\boldsymbol{x} - \boldsymbol{v}_{(k)}\|_2^2\right) \quad (9)$$

where

$$\boldsymbol{v}_{(k)} = \boldsymbol{x}_{(k)} - \tfrac{1}{L}\nabla f(\boldsymbol{x}_{(k)}) \quad (10)$$

and $\nabla f(\boldsymbol{x}) = \boldsymbol{A}^\top(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y})$ is the gradient of $f(x)$. L is the Lipschitz constant of $\nabla f(\boldsymbol{x})$. Note that the update of $\boldsymbol{v}_{(k)}$ according to (10) involves the calculation and storage of $\boldsymbol{A}^\top\boldsymbol{A}$ and $\boldsymbol{A}^\top\boldsymbol{y}$. Instead, to save on computational costs, we can update the matrix $\boldsymbol{V}_{(k)}$ according to:

$$\boldsymbol{V}_{(k)} = \boldsymbol{X}_{(k)} - \tfrac{1}{L}\boldsymbol{\Phi}^\top(\boldsymbol{\Phi}\boldsymbol{X}_{(k)} - \boldsymbol{Y}), \quad (11)$$

where $\boldsymbol{V}_{(k)}$ is the matrix satisfying $\boldsymbol{v}_{(k)} = \text{vec}(\boldsymbol{V}_{(k)}^\top)$. As a consequence, only lower dimension matrices, $\boldsymbol{\Phi}^\top\boldsymbol{\Phi}$ and $\boldsymbol{\Phi}^\top\boldsymbol{Y}$, need to be computed and stored. Solving (9) is similar to the 1D fused Lasso signal approximator (FLSA) [33]. Moreover, due to the block diagonal structure of $\boldsymbol{F}$, it is obvious that $\|\boldsymbol{F}\boldsymbol{x}\|_1 = \sum_{n=1}^{N}\|\boldsymbol{D}\boldsymbol{x}^n\|_1$. Therefore, problem (9) can be solved separately for each row $\boldsymbol{x}^n$ of $\boldsymbol{X}$:

$$\boldsymbol{x}^n_{(k+1)} = \arg\min_{\boldsymbol{x}\in\mathbb{R}^K} \tfrac{1}{2}\|\boldsymbol{x} - \boldsymbol{v}^n_{(k)}\|_2^2 + \tfrac{\lambda_1}{L}\|\boldsymbol{x}\|_1 + \tfrac{\lambda_2}{L}\|\boldsymbol{D}\boldsymbol{x}\|_1. \quad (12)$$

The solution to (12) is obtained by using subgradient technique. Indeed, any solution corresponding to (λ_1, λ_2) is obtained by a soft thresholding of the solution obtained for $(\lambda_1 = 0, \lambda_2)$. This is stated by the following theorem.

Theorem 1 (Friedman *et al.* [30], Liu *et al.* [34]). *Let*

$$\boldsymbol{x}(\lambda_1, \lambda_2) = \arg\min_{\boldsymbol{x}} \tfrac{1}{2}\|\boldsymbol{x} - \boldsymbol{v}\|_2^2 + \lambda_1\|\boldsymbol{x}\|_1 + \lambda_2\|\boldsymbol{D}\boldsymbol{x}\|_1. \quad (13)$$

For all λ_1, $\lambda_2 \geq 0$, we have:

$$\boldsymbol{x}(\lambda_1, \lambda_2) = \text{sign}(\boldsymbol{x}(0, \lambda_2)) * \max(|\boldsymbol{x}(0, \lambda_2)| - \lambda_1, 0). \quad (14)$$

where $*$ denotes the element-wise product operator. □

This observation is used in [30, 33] to propose algorithms for solving the FLSA problem over a path of λ_2 values, keeping λ_1 fixed (typically $\lambda_1 = 0$). In our case, each problem in (12) is solved using the FLSA routine implemented in SLEP package[1]. Finally, the main steps of the Fused Sparse Lasso (FSL) algorithm are presented in Algorithm 1, where $\boldsymbol{Z}$ is a linear combination of two consecutive estimates of $\boldsymbol{X}$; it is updated at each FISTA iteration.

[1] `http://yelab.net/software/SLEP/`

Algorithm 1: Fused Sparse Lasso (FSL)

Input : $\boldsymbol{Y} \in \mathbb{C}^{M\times K}$, $\boldsymbol{\Phi} \in \mathbb{C}^{M\times N}$, λ_1, λ_2, *maxiter*

Initialization: $\boldsymbol{X}_{(0)} = \boldsymbol{0}$, $\boldsymbol{Z}_{(1)} = \boldsymbol{0}$, $t_{(1)} = 1$;
for $k \leftarrow 1$ **to** *maxiter* **do**
  $\boldsymbol{V}_{(k)} \leftarrow \boldsymbol{Z}_{(k)} - \frac{1}{L}\boldsymbol{\Phi}^\top(\boldsymbol{\Phi}\boldsymbol{Z}_{(k)} - \boldsymbol{Y})$;
  for $n \leftarrow 1$ **to** N **do**
    $\mathbf{x}^n_{(k)} \leftarrow \arg\min_{\boldsymbol{x}} \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{v}^i_{(k)}\|_2^2 + \frac{\lambda_1}{L}\|\boldsymbol{x}\|_1 + \frac{\lambda_2}{L}\|\boldsymbol{D}\boldsymbol{x}\|_1$;
  end
  $t_{(k+1)} \leftarrow \frac{1+\sqrt{1+4t^2_{(k)}}}{2}$;
  $\boldsymbol{Z}_{(k+1)} \leftarrow \boldsymbol{X}_{(k)} + \frac{t_{(k)}-1}{t_{(k+1)}}(\boldsymbol{X}_{(k)} - \boldsymbol{X}_{(k-1)})$;
end

Output: $\boldsymbol{X} \in \mathbb{R}^{N\times K}$

3.2. Fused Sparse Group Lasso

As mentioned before, the fused sparse Lasso is not a proper relaxation of the problem in (3). Indeed, the term $\|\boldsymbol{X}\|_{1,1}$ does not allow to control the number of active rows. Here, we propose to relax the ℓ_0/ℓ_2 pseudo-norm into the ℓ_1/ℓ_2 mixed norm defined by: $\|\boldsymbol{X}\|_{1,2} = \sum_{n=1}^N \|\boldsymbol{x}^n\|_2$, which is a particular instance of the group Lasso penalty. So we propose the following criterion:

$$J_{FSGL}(\boldsymbol{x}) = \tfrac{1}{2}\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2^2 + \lambda_1\|\boldsymbol{x}\|_1 + \lambda_2\|\boldsymbol{F}\boldsymbol{x}\|_1 + \lambda_3\sum_{n=1}^N \|\boldsymbol{x}^n\|_2, \tag{15}$$

as a convex relaxation of problem (3). Note that the $\|\boldsymbol{x}\|_1$ penalty is maintained to eventually control the global sparsity of the solution. The proximal operator associated with the composite of non-smooth penalties in the fused sparse group Lasso (FSGL) is defined as:

$$\begin{aligned}\text{prox}_{FSGL}(\boldsymbol{v}) = \arg\min_{\boldsymbol{x}} &\tfrac{1}{2}\|\boldsymbol{x} - \boldsymbol{v}\|_2^2 \\ &+ \tfrac{\lambda_1}{L}\|\boldsymbol{x}\|_1 + \tfrac{\lambda_2}{L}\|\boldsymbol{F}\boldsymbol{x}\|_1 + \tfrac{\lambda_3}{L}\sum_{n=1}^N \|\boldsymbol{x}^n\|_2.\end{aligned} \tag{16}$$

Here again, it is clear that each row of $\boldsymbol{X}$ is decoupled in (16). So we only need to solve the following optimization problem for each row $n = 1, \ldots, N$:

$$\begin{aligned}\text{prox}_{FSGL}(\boldsymbol{v}^n) = \arg\min_{\boldsymbol{x}^n} &\tfrac{1}{2}\|\boldsymbol{x}^n - \boldsymbol{v}^n\|_2^2 \\ &+ \tfrac{\lambda_1}{L}\|\boldsymbol{x}^n\|_1 + \tfrac{\lambda_2}{L}\|\boldsymbol{D}\boldsymbol{x}^n\|_1 + \tfrac{\lambda_3}{L}\|\boldsymbol{x}^n\|_2.\end{aligned} \tag{17}$$

Now, with the three non-smooth terms in the objective function, the proximal operator may be computed as suggested in [35]. In fact, the proximal operator in (17) has a decomposition property that allows to compute it in two steps based on the following theorem.

Theorem 2 (Zhou *et al.* [35]). *Define*

$$\text{prox}_{FSL}(\boldsymbol{v}) = \arg\min_{\boldsymbol{x}} \tfrac{1}{2}\|\boldsymbol{x} - \boldsymbol{v}\|_2^2 + \lambda_1\|\boldsymbol{x}\|_1 + \lambda_2\|\boldsymbol{D}\boldsymbol{x}\|_1 \tag{18}$$

$$\text{prox}_{GL}(\boldsymbol{v}) = \arg\min_{\boldsymbol{x}} \tfrac{1}{2}\|\boldsymbol{x} - \boldsymbol{v}\|_2^2 + \lambda_3\|\boldsymbol{x}^n\|_2. \tag{19}$$

Algorithm 2: Fused Sparse Group Lasso (FSGL)

Input : $\boldsymbol{Y} \in \mathbb{C}^{M\times K}$, $\boldsymbol{\Phi} \in \mathbb{C}^{M\times N}$, λ_1, λ_2, λ_3, *maxiter*

Initialization: $\boldsymbol{X}_{(0)} = \boldsymbol{0}$, $\boldsymbol{Z}_{(1)} = \boldsymbol{0}$, $t_{(1)} = 1$;
for $k \leftarrow 1$ **to** *maxiter* **do**
  $\boldsymbol{V}_{(k)} \leftarrow \boldsymbol{Z}_{(k)} - \frac{1}{L}\boldsymbol{\Phi}^\top(\boldsymbol{\Phi}\boldsymbol{Z}_{(k)} - \boldsymbol{Y})$;
  for $n \leftarrow 1$ **to** N **do**
    $\mathbf{w}^n_{(k)} \leftarrow \arg\min_{\boldsymbol{x}} \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{v}^n_{(k)}\|_2^2 + \frac{\lambda_1}{L}\|\boldsymbol{x}\|_1 + \frac{\lambda_2}{L}\|\boldsymbol{D}\boldsymbol{x}\|_1$;
    $\boldsymbol{x}^n_{(k)} \leftarrow \arg\min_{\boldsymbol{x}} \frac{1}{2}\|\boldsymbol{x} - \mathbf{w}^n_{(k)}\|_2^2 + \frac{\lambda_3}{L}\|\boldsymbol{x}\|_2$;
  end
  $t_{(k+1)} \leftarrow \frac{1+\sqrt{1+4t^2_{(k)}}}{2}$;
  $\boldsymbol{Z}_{(k+1)} \leftarrow \boldsymbol{X}_{(k)} + \frac{t_{(k)}-1}{t_{(k+1)}}(\boldsymbol{X}_{(k)} - \boldsymbol{X}_{(k-1)})$;
end

Output: $\boldsymbol{X} \in \mathbb{R}^{N\times K}$

Then, the following holds for all $\lambda_1, \lambda_2, \lambda_3 \geq 0$:

$$\text{prox}_{FSGL}(\boldsymbol{v}) = (prox_{GL} \circ \text{prox}_{FSL})(\boldsymbol{v}). \tag{20}$$

where $\circ$ is the composition operator. □

This result implies that we can first compute the proximal operator associated to the fused sparse Lasso as in the previous section. The solution is then plugged in the proximal operator associated to the group Lasso. The latter is finally computed using the ALTRA routine also available in the SLEP package. The resulting algorithm (FSGL) is summarized in Algorithm 2.

3.3. Nonnegative Fused Sparse Group Lasso

As we deal with positive data, it is suitable to impose a nonnegativity constraint on the solution. Indeed, the solution proposed above may induce artifacts due to bad conditioning of matrices, causing the appearance of negative values. From a physical point of view, such a solution is unacceptable and a rigorous recovery process must take into account this additional constraint. So, we propose here to minimize the nonnegative version of the fused sparse group Lasso algorithm. The constrained problem expresses as:

$$\underset{\boldsymbol{x}}{\text{minimize}} \quad J_{FSGL}(\boldsymbol{x}), \tag{21a}$$

$$\text{subject to} \quad \boldsymbol{x} \geq 0. \tag{21b}$$

First, we include a slack variable $\boldsymbol{u} \in \mathbb{R}^{NK}$ to the objective function which leads to:

$$\underset{\boldsymbol{x}}{\text{minimize}} \quad J_{FSGL}(\boldsymbol{x}), \tag{22a}$$

$$\text{subject to} \quad \boldsymbol{x} - \boldsymbol{u} = 0, \boldsymbol{u} \geq 0. \tag{22b}$$

The equality constraint in (22b) can be handled by using the quadratic penalty method [36]. The new objective in then:

$$\begin{aligned}J_{NN-FSGL}(\boldsymbol{x}, \boldsymbol{u}) = &\tfrac{1}{2}\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2^2 + \tfrac{\xi}{2}\|\boldsymbol{x} - \boldsymbol{u}\|_2^2 \\ &+ \lambda_1\|\boldsymbol{x}\|_1 + \lambda_2\|\boldsymbol{F}\boldsymbol{x}\|_1 + \lambda_3\sum_{n=1}^N \|\boldsymbol{x}^n\|_2, \quad \boldsymbol{u} \geq 0\end{aligned} \tag{23}$$

where ξ is the parameter that penalizes the constraint violations in the sense that, when $\xi \to \infty$, the entries of the vector $\boldsymbol{x}$ tend toward those of the vector $\boldsymbol{u}$ making the inequality constraint $\boldsymbol{x} \geq 0$ satisfied asymptotically. The surrogate problem (23) is unconstrained with respect to $\boldsymbol{x}$. Hence, by stacking the two quadratic terms of the objective $J_3(\cdot)$ we obtain:

$$J_{NN-FSGL}(\boldsymbol{x}, \boldsymbol{u}) = \tfrac{1}{2}\|\boldsymbol{y}'(\boldsymbol{u}) - \boldsymbol{B}\boldsymbol{x}\|_2^2 + g'(\boldsymbol{x}), \quad \boldsymbol{u} \geq 0 \tag{24}$$

where $\boldsymbol{B} = [\boldsymbol{A}^\top, \sqrt{\xi}\boldsymbol{I}^\top]^\top$, $\boldsymbol{y}'(\boldsymbol{u}) = [\boldsymbol{y}^\top, \sqrt{\xi}\boldsymbol{u}^\top]^\top$, $\boldsymbol{I}$ is an identity matrix of the same size as $\boldsymbol{A}$, and $g'(\boldsymbol{x})$ is defined by:

$$g'(\boldsymbol{x}) = \lambda_1\|\boldsymbol{x}\|_1 + \lambda_2\|\boldsymbol{F}\boldsymbol{x}\|_1 + \lambda_3 \sum_{i=1}^{N} \|\boldsymbol{x}^i\|_2. \tag{25}$$

We should now minimize the cost function $J_{NN-FSGL}(\boldsymbol{x}, \boldsymbol{u})$ with respect to $\boldsymbol{x}$ (without constraint) and $\boldsymbol{u}$ (with the nonnegativity constraint). The minimization with respect to $\boldsymbol{x}$ leads to an iteration similar to that of FSGL:

$$\begin{cases} \boldsymbol{v}_{(k)} = \boldsymbol{x}_{(k)} - \frac{1}{L'}\nabla f'(\boldsymbol{x}_{(k)}) \\ \boldsymbol{x}_{(k+1)} = \arg\min_{\boldsymbol{x}} \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{v}_{(k)}\|_2^2 + \frac{1}{L'}g'(\boldsymbol{x}), \end{cases} \tag{26}$$

where $f'(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{y}'(\boldsymbol{u}) - \boldsymbol{B}\boldsymbol{x}\|_2^2$, $\nabla f'(\boldsymbol{x}) = \boldsymbol{B}^\top(\boldsymbol{B}\boldsymbol{x} - \boldsymbol{y}'(\boldsymbol{u}))$ and $L' = L + \xi$ is the Lipschitz constant of $\nabla f'(\boldsymbol{x})$. Once again, the optimization is separable for each row $\boldsymbol{x}^n$. Define matrix $\boldsymbol{U}$ such that $\boldsymbol{u} = \mathrm{vec}(\boldsymbol{U}^\top)$. Replacing $\boldsymbol{B}$ and $\boldsymbol{y}'(\boldsymbol{u})$ by their expressions yields:

$$\begin{cases} \boldsymbol{V}_{(k)} & = \boldsymbol{X}_{(k)} - \frac{1}{L'}\boldsymbol{\Phi}^\top(\boldsymbol{\Phi}\boldsymbol{X}_{(k)} - \boldsymbol{Y}) - \frac{\xi}{L'}(\boldsymbol{X}_{(k)} - \mathbf{U}_{(\ell)}), \\ \boldsymbol{x}^n_{(k+1)} & = \arg\min_{\boldsymbol{x}\in\mathbb{R}^K} \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{v}^n_{(k)}\|_2^2 + \frac{\lambda_1}{L'}\|\boldsymbol{x}\|_1 + \frac{\lambda_2}{L'}\|\boldsymbol{D}\boldsymbol{x}\|_1 + \frac{\lambda_3}{L'}\|\boldsymbol{x}\|_2, \\ & \text{for } n = 1, \ldots, N. \end{cases} \tag{27}$$

Hence, an external loop (ℓ) is added to update the variable $\boldsymbol{u}$. The minimization of $J_{NN-FSGL}(\boldsymbol{x}, \boldsymbol{u})$ with respect to the slack variable $\boldsymbol{u}$ is simply a hard thresholding operation:

$$\boldsymbol{u}_{(\ell+1)} = \max(0, \boldsymbol{x}^*), \tag{28}$$

where $\boldsymbol{x}^*$ is the value of $\boldsymbol{x}_{(k)}$ when the final iteration on k is completed. The tuning parameter ξ is updated in the loop with the classical linear rule: $\xi_{(\ell+1)} = \beta\xi_{(\ell)}$, with $\beta > 1$ and $\xi_1 = 1$. The complete NN-FSGL algorithm is summarized in Algorithm 3.

Algorithm 3: Nonnegative Fused Sparse Group Lasso (NN-FSGL)

Input : $\boldsymbol{Y} \in \mathbb{C}^{M\times K}$, $\boldsymbol{\Phi} \in \mathbb{C}^{M\times N}$, λ_1, λ_2, λ_3, β, *maxiter*, *nniter*

Initialization: $\boldsymbol{X}_{(0)} = \mathbf{0}$, $\boldsymbol{Z}_{(1)} = \mathbf{0}$, $t_{(1)} = 1$, $\boldsymbol{u} = 0$, $\xi_{(1)} = 1$;
for $\ell \leftarrow 1$ **to** *nniter* **do**
  $L' \leftarrow L + \xi_{(\ell)}$;
  for $k \leftarrow 1$ **to** *maxiter* **do**
    $\boldsymbol{V}_{(k)} \leftarrow \boldsymbol{Z}_{(k)} - \frac{1}{L'}\boldsymbol{\Phi}^\top(\boldsymbol{\Phi}\boldsymbol{Z}_{(k)} - \boldsymbol{Y}) - \frac{\xi_{(\ell)}}{L'}(\boldsymbol{Z} - \mathbf{U}_{(\ell)})$;
    for $n \leftarrow 1$ **to** N **do**
      $\mathbf{w}^n_{(k)} \leftarrow \arg\min_{\boldsymbol{x}} \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{v}^n_{(k)}\|_2^2 + \frac{\lambda_1}{L'}\|\boldsymbol{x}\|_1 + \frac{\lambda_2}{L'}\|\boldsymbol{D}\boldsymbol{x}\|_1$;
      $\boldsymbol{x}^n_{(k)} \leftarrow \arg\min_{\boldsymbol{x}} \frac{1}{2}\|\boldsymbol{x} - \mathbf{w}^n_{(k)}\|_2^2 + \frac{\lambda_3}{L'}\|\boldsymbol{x}\|_2$;
    end
    $t_{(k+1)} \leftarrow \frac{1+\sqrt{1+4t^2_{(k)}}}{2}$;
    $\boldsymbol{Z}_{(k+1)} \leftarrow \boldsymbol{X}_{(k)} + \frac{t_{(k)}-1}{t_{(k+1)}}(\boldsymbol{X}_{(k)} - \boldsymbol{X}_{(k-1)})$;
  end
  $\boldsymbol{u}_{(\ell+1)} \leftarrow \max(0, \boldsymbol{x}_{(maxiter)})$;
  $\xi_{(\ell+1)} \leftarrow \beta\xi_{(\ell)}$;
end

Output: $\boldsymbol{X} \in \mathbb{R}^{N\times K}$

3.4. Sofware

An open source Matlab implementation of FSL, FSGL and NN-FSGL can be downloaded from `http://w3.cran.univ-lorraine.fr/el-hadi.djermoune/?q=content/publications`. The software also contains a test program and the experimental NIR data used in the next section.

4. Wood wastes sorting

4.1. Motivations

One of the most promising application of spectroscopy and hyperspectral imaging in industry is material sorting [37, 38, 39] and quality control [40, 41]. In this work, we are interested in sorting wood wastes which have to be separated into two broad classes: recyclable and non recyclable. Each class includes a number of wood wastes types called "groups" as given in Table 1. The wood wastes sorting is addressed as a binary classification of NIR spectra. A single spectrum is acquired for each wood sample and the classifier has to decide whether it is recyclable or not recyclable.

The goal of this section is to show the effectiveness of the algorithms presented before in variable selection and classification. These algorithms are primarily intended at selecting the explanatory variables used in classifiers. Here we restrict our attention to the kernel SVM classifier which proved to be among the most effective for the considered problem, and the question at hand is: is it possible to improve the classification rates and decrease the computational burden by performing a proper variable selection?

4.2. Data acquisition and pre-processing

We collected several hundred samples of wood in a waste park amongst which 290 were gathered by experts into 11 labeled groups as shown in Table 1. The data acquisition was carried in reflectance mode on a Nicolet 8700 FTIR spectrometer equipped with a MCT detector and a CaF2 beam splitter. Near infrared reflectance spectra cover the spectral range $[3562, 10000]$ cm^{-1} (corresponding to $[1000, 2800]$ nm). The spectral resolution is 16 cm^{-1}. The spectral sampling step is 4 cm^{-1} yielding 1647 spectral bands. Each spectrum is obtained

Table 1: Composition of the two classes of wood wastes

(a) Class 0: recyclable

Group	Type	Samples
0.1	raw wood	32
0.2	painted solid wood	36
0.3	varnished solid wood	35
0.4	raw plywood	18
0.5	varnished plywood	18
0.6	raw particle board	28
0.7	painted particle board	6

(b) Class 1: non-recyclable

Group	Type	Samples
1.1	raw wood metallic salts	35
1.2	MDF-HDF	28
1.3	painted MDF-HDF	50
1.4	raw fiber board	8

by averaging 100 scans. The data pre-processing includes baseline removal using the method proposed in [42], offset correction ensuring zero lower bound, and unit energy normalization. Some spectra from these different groups are shown in Figure 2. It appears that the discriminant features cannot be determined by a simple visual examination.

The data are then gathered in matrix $\boldsymbol{Y} \in \mathbb{R}^{1647\times 290}$. Note that the spectra are ordered according to the group they belong to. The spectra from class 0 are put in the first columns of $\boldsymbol{Y}$ starting from group 0.1 through group 0.7. In the same manner, the spectra from class 1 are put in the last columns. This is a very important point since it is this ordering which enables the rows of the coefficient matrix $\boldsymbol{X}$ to be piecewise constant when $\lambda_2 > 0$.

4.3. Dictionary

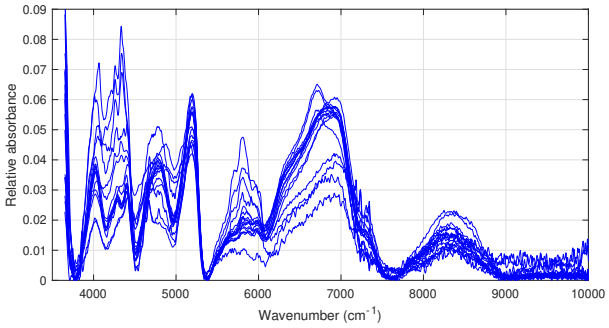
The dictionary $\boldsymbol{\Phi}$ is composed of normalized Gaussian-shaped functions whose means $m_i \in [3660, 10000]$ cm^{-1} and widths $\sigma_j \in [30, 600]$ cm^{-1} are covering uniformly their respective intervals. The discretization leads to 20 different values for σ_j. For each σ_j, the interval $[3660, 10000]$ cm^{-1} is discretized such that two adjacent m_i's are separated by σ_j. As a consequence, the dictionary is composed of 773 atoms with mutual coherence 0.9995.

4.4. Variable selection

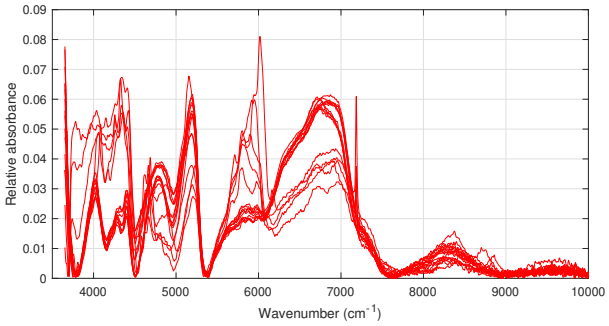
Here we compare FSL and FSGL to the simultaneous variable selection algorithm (SVS) proposed by Turlach *et al.* [22]). This algorithm is an extension of Lasso strategy and corresponds to the following optimization problem:

$$\min_{\boldsymbol{X}} \tfrac{1}{2}\|\boldsymbol{Y} - \boldsymbol{\Phi}\boldsymbol{X}\|_F^2 \quad \text{s. t.} \quad \sum_{n=1}^{N} \|\boldsymbol{x}^n\|_\infty \leq t, \tag{29}$$

where t is a user parameter controlling the sparsity of the solution. The problem is solved using an interior point method and



(a) Class 0: recyclable



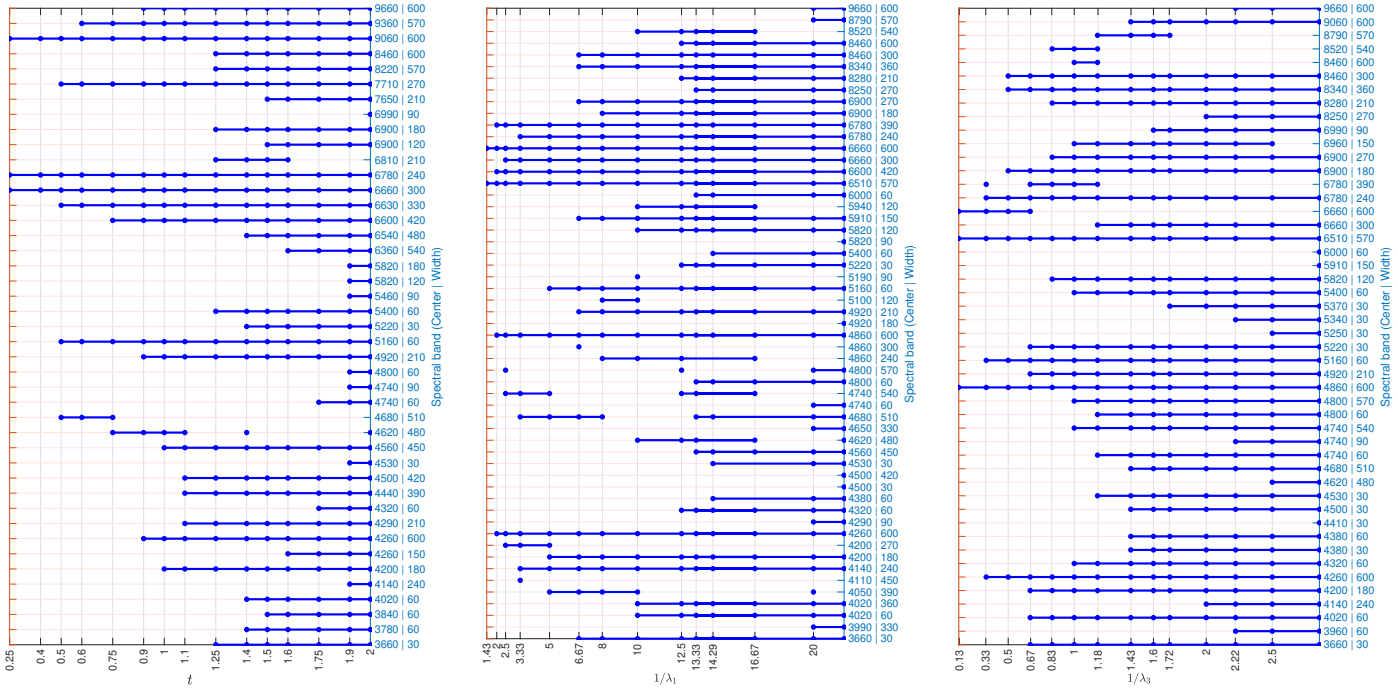
(b) Class 1: non-recyclable

Figure 2: Some (pre-processed) NIR spectra from the two classes of wood wastes.

the C implementation is kindly provided by the author (Berwin A. Turlach).

Figure 3 displays the selected variables obtained by SVS, FSL and FSGL for some values of t, $1/\lambda_1$ and $1/\lambda_3$, respectively. Note that the number of active variables returned by SVS increases when t increases whereas, for FSL and FSGL, it decreases when λ_1 or λ_3 increases. The horizontal lines connect two adjacent values of the parameters when the coefficient associated to a selected variable does not vanish. For small values of t, $1/\lambda_1$ and $1/\lambda_3$, the variables are mainly picked in the range $[6600, 6700]$ cm^{-1} where broad and intense spectral peaks are observed (see Fig. 2). By increasing the value of these parameters, more and more variables are selected. In a given application, the practitioner can stop the selection when the desired number of variables is reached. In our case, out of about fourty variables (with $t = 2$, $\lambda_1 = 0.045$ and $\lambda_3 = 0.35$), SVS shares 28 common variables with FSL and FSGL. The latter algorithms share 38 common variables. It can also be seen that some wavenumbers actually have a chemical interpretation. For instance, the variables located in the ranges 4000-4500 cm^{-1} and 5800-8200 cm^{-1} are related to the main components of wood including cellulose, hemicellulose and and lignin [43, 44].

The computational time required by each algorithm to perform variable selection is reported in Table 2. The results are obtained using a 2.4 Ghz Intel Core i5 processor with 8 Gigabytes of RAM. We note that FSL is generally faster than all other approaches. FSGL algorithm is a bit slower. The additional loop with hard thresholding operator makes NN-FSGL about ten times time demanding than its unconstrained version.



(a) SVS [22] (b) FSL ($\lambda_2 = 0.5$) (c) FSGL ($\lambda_1 = 0, \lambda_2 = 0.5$)

Figure 3: Selected variables *versus* tuning parameters.

Table 2: Computational time (in seconds) of the different approaches for variable selection

# Variables	FSL	FSGL	NN-FSGL	G-SVM
25	9	17	110	27
32	9	17	101	27
40	10	15	112	29
50	9	12	105	31

For SVS we did not try all the configurations because we found that this algorithm is much more slower and needs about four hours to select 32 variables. Finally, FSL and FSGL algorithms are not only numerically efficient but also provide good classification rates as will be shown in the next paragraph.

4.5. Classification of wood wastes using NIR spectra

Here we perform classification of recyclable and non-recyclable wood samples using SVM with quadratic kernel function. All variable selection algorithms are tuned to produce 32 spectral bands. Classification is then performed using matrix $\boldsymbol{X}$ corresponding to the unconstrained least-squares solution of equation (2a) where dictionary $\boldsymbol{\Phi}$ is restricted to the 32 active atoms. The results are compared to SVM classification without variables selection (*i.e.* using the original data in $\boldsymbol{Y}$) and G-SVM [45]. The latter algorithm consists in solving an augmented SVM criterion where the sparsity constraint is imposed on the support vectors. The solution is computed using a projected gradient method. In G-SVM, the sparsity of the support vectors is controlled by parameter[2] C which is set to $C = 150$ making the decision rule made on 32 coefficients of the support vectors.

The classification results for 10 cross validation runs are shown in Table 3. For each method we report the overall rate of success, the true positive rate TPR (rate of recyclable samples correctly identified), and the true negative rate TNR (rate of non-recyclable samples correctly rejected). In terms of total accuracy, FSL, FSGL ad NN-FSGL clearly outperform SVM, SVS and G-SVM. The best result is about 88% obtained with FSGL. As in our application it is also important to reject the maximum number of polluted samples from the recycling process, the best parameters for FSGL are $\lambda_1 = 0$, $\lambda_2 = 0.5$ and $\lambda_3 = 0.625$ (see also section 4.6). Figure 4 shows an example of error rates resulting in each group of wood wastes when the spectra are randomly split into training samples (203 spectra) and test ones (87 spectra). The results obtained using the variables selected by FSGL are: TPR = 87.3%, TNR = 94.4%, and an overall rate of success of 89.7%. One can see that the two pieces of painted particle boards are misclassified. This is mainly due to *(i)* the small number of samples in the corresponding group: only 4 samples are used for training and 2 for the test; *(ii)* the presence of painted wood samples in both classes.

4.6. Parameter adjustment

The performances of all the algorithms considered here depend on the choice of tuning parameters. For instance, the set

[2] `http://remi.flamary.com/soft/soft-gsvm.html`

Table 3: Wood wastes classification accuracy using SVM

Variable selection algorithm	Classifier	Number of variables	Parameters			Accuracy		
			λ_1	λ_2	λ_3	Success	TPR	TNR
–	SVM	1647	–			82.2%	80.4%	84.8%
G-SVM		32	$C = 150$			76.9%	81.4%	71.3%
SVS	SVM	32	$t = 1.75$			84.1%	83.1%	85.6%
FSL	SVM	32	0.075	0.5	–	85.9%	85.6%	86.3%
FSGL	SVM	32	0	0.5	0.625	**87.8**%	86.1%	**90.3**%
		32	0.04	0.2	0.295	86.5%	**86.4**%	86.7%
NN-FSGL	SVM	32	0	0.5	0.6	86.9%	85.5%	88.8%

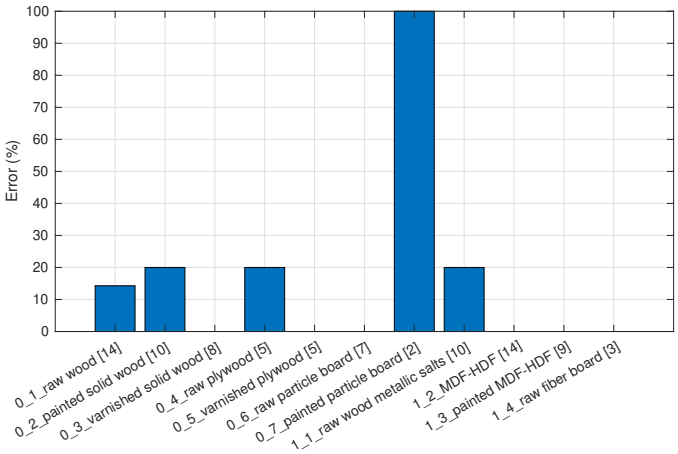


Figure 4: An example of classification error rate in each group. The dataset is randomly split into training samples (70%) and test samples (30%).

of selected variables (and thus, the overall classification performance) with FSGL depend on λ_1, λ_2 and λ_3. Our aim now is to evaluate the impact of each parameter on the number of selected variables and classification rates using 10-fold cross validation. For $\lambda_2 = 0.2$ and without the group penalty ($\lambda_3 = 0$), the results in terms of total classification error and cardinality of the support are reported in Figure 5(a), for several values of λ_1. We note that the classification error rate decreases from 35% ($\lambda_1 = 10^{-2}$) to 14% ($\lambda_1 \in [0.18, 0.28]$). Naturally, the performances degrade drastically for values of λ_1 beyond 0.3 which correspond to less than 20 variables. For $\lambda_2 = 0.5$ and without the sparsity term ($\lambda_1 = 0$), the results are shown on Figure 5(b), for different values of the grouping parameter λ_3. We observe that the total classification error rate is under 15% for $\lambda_3 \in [0.1, 1.5]$. In particular the value of λ_3 yielding the lowest error rate (12.2%) is shown in Table 3 with 32 spectral bands. It is worth to notice from these two experiments that both sparsity and grouping parameters act directly on the number of selected variables but not with the same intensity: the sparsity parameter has stronger influence than the grouping parameter. For instance, to obtain less than 80 variables, the grouping parameter should be set to 0.2 (and $\lambda_1 = 0$) while the same number of variables is obtained for $\lambda_1 \approx 0.06$ (and $\lambda_3 = 0$). To analyse the effect of the fusion parameter on the general classification performances, we set the sparsity parameter λ_1 to 0 and vary both the grouping and the fusion parameters such that 40 variables are retained. The results are reported on Figure 5(c). The error rate is less than 15% in the range $\lambda_2 \in [0.1, 0.6]$. The minimum value of the classification error rate is 11.6%; it corresponds to $\lambda_2 = 0.55$.

5. Conclusion

In this paper, simultaneous regularized sparse approximation algorithms for variable selection are proposed. The first idea of this work is to reduce data dimensionality of NIR spectra using sparse decomposition. Moreover, to improve the classification performances, we incorporate a regularization constraint along the rows of the coefficient matrix to enforce a piecewise constant form. This is done by applying a ℓ_1-norm penalty on the difference between successive coefficients. The corresponding algorithm is the sparse fused Lasso. Using a FISTA iteration, we have shown that the criterion may be solved efficiently thanks to the fused Lasso signal approximator (FLSA) applied on each row of the coefficient matrix. Additional penalties have also been incorporated to the criterion to enforce simultaneous selection and non-negativity of the solution. The resulting algorithms have a low computational cost suitable for large-scale problems. The effectiveness of the algorithms is demonstrated on real NIR spectra both in terms of variable selection and classification performance.

References

[1] B. Stuart, Infrared spectroscopy, Wiley Online Library, New York, USA, 2005.

[2] H. W. Siesler, Y. Ozaki, S. Kawata, H. M. Heise, Near-infrared spectroscopy: Principles, instruments, applications, Wiley-VCH, Weinheim, Germany, 2002.

[3] M. J. Adams, Chemometrics in analytical spectroscopy, Royal Society of Chemistry, Cambridge, UK, 1995.

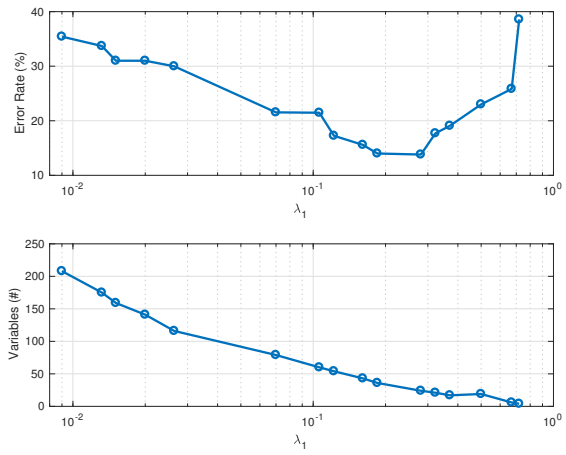
[4] P. Jonsson, S. J. Bruce, T. Moritz, J. Trygg, M. Sjöström, R. Plumb, J. Granger, E. Maibaum, J. K. Nicholson, E. Holmes, H. Antti, Extraction, interpretation and validation of information for comparing samples in metabolic LC/MS data sets, Analyst 130 (5) (2005) 701–707.

[5] G. M. Furnival, R. W. Wilson, Regressions by leaps and bounds, Technometrics 42 (1) (2000) 69–79.

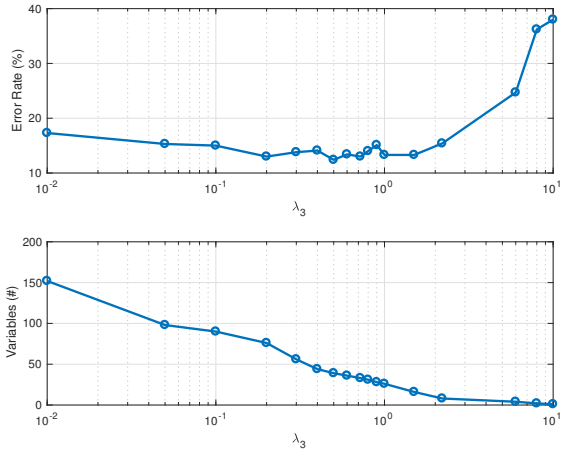
[6] A. J. Miller, Subset selection in regression, Chapman and Hall, London, London, England, 1990.

[7] T. Marill, D. Green, On the effectiveness of receptors in recognition systems, IEEE transactions on Information Theory 9 (1) (1963) 11–17.

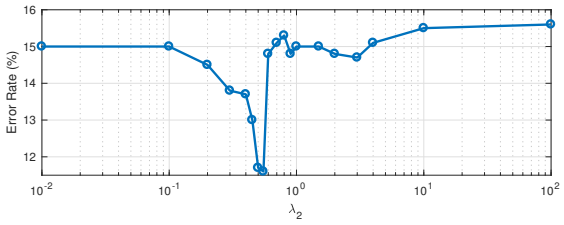
[8] Z. Xiaobo, Z. Jiewen, M. J. Povey, M. Holmes, M. Hanpin, Variables selection methods in near-infrared spectroscopy, Analytica Chimica Acta 667 (1) (2010) 14–32.



(a) Sparsity parameter ($\lambda_2 = 0.2, \lambda_3 = 0$)



(b) Grouping parameter ($\lambda_1 = 0, \lambda_2 = 0.5$)



(c) Fusion parameter ($\lambda_1 = 0$ and λ_3 is adjusted to yield 40 variables)

Figure 5: Evolution of the total classification error rate as a function of the regularization parameters.

[9] R. Tibshirani, Regression shrinkage and selection via the Lasso, Journal of the Royal Statistical Society. Series B (Methodological) (1996) 267–288.

[10] D. Donoho, Compressed sensing, IEEE Transactions on Information Theory 52 (2006) 1289–1306.

[11] E. J. Candès, J. Romberg, T. Tao, Stable signal recovery for incomplete and inaccurate measurements, Communication on Pure and Applied Mathematics 59 (2006) 1207–1223.

[12] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, Y. Ma, Robust face recognition via sparse representation, IEEE Transactions on Pattern Analysis and Machine Intelligence 31 (2) (2009) 210–227.

[13] K. Huang, S. Aviyente, Sparse representation for signal classification, in: Advances in Neural Information Processing Systems, 2006, pp. 609–616.

[14] J. Kim, H. Park, Sparse nonnegative matrix factorization for clustering, Tech. rep., Georgia Institute of Technology (2008).

[15] M. D. Iordache, J. Bioucas-Dias, A. Plaza, Sparse unmixing of hyperspectral data, IEEE Transactions on Geoscience and Remote Sensing 49 (2011) 2014–2039.

[16] Y. Chen, N. M. Nasrabadi, T. D. Tran, Hyperspectral image classification using dictionary-based sparse representation, IEEE Transactions on Geoscience and Remote Sensing 49 (10) (2011) 3973–3985.

[17] Z. Wang, R. Zhu, K. Fukui, J.-H. Xue, Cone-based joint sparse modelling for hyperspectral image classification, Signal Processing 144 (2018) 417–429.

[18] J. A. Tropp, A. C. Gilbert, M. J. Strauss, Algorithms for simultaneous sparse approximation. Part I: Greedy pursuit, Signal Processing 86 (2006) 572–588.

[19] L. Belmerhnia, E.-H. Djermoune, D. Brie, Greedy methods for simultaneous sparse approximation, in: 22nd European Signal Processing Conference (EUSIPCO), 2014, pp. 1851–1855.

[20] D. Kim, J. P. Haldar, Greedy algorithms for nonnegativity-constrained simultaneous sparse recovery, Signal Processing 125 (2016) 274–289.

[21] J. A. Tropp, Algorithms for simultaneous sparse approximation. Part II: Convex relaxation, Signal Processing 86 (2006) 589–602.

[22] B. A. Turlach, W. N. Venables, S. J. Wright, Simultaneous variable selection, Technometrics 47 (3) (2005) 349–363.

[23] B. Malli, T. Natschläger, Fused stagewise regression – A waveband selection algorithm for spectroscopy, Chemometrics and Intelligent Laboratory Systems 149 (2015) 53–65.

[24] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, K. Knight, Sparsity and smoothness via the fused lasso, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67 (1) (2005) 91–108.

[25] S. F. Cotter, B. D. Rao, K. Engan, K. Kreutz-Delgado, Sparse solutions to linear inverse problems with multiple measurement vectors, IEEE Transactions on Signal Processing 53 (7) (2005) 2477–2488.

[26] J. Chen, X. Huo, Theoretical results on sparse representations of multiple-measurement vectors, IEEE Transactions on Signal Processing 54 (12) (2006) 4634–4643.

[27] L. Belmerhnia, E.-H. Djermoune, C. Carteret, D. Brie, Simultaneous regularized sparse approximation for wood wastes NIR spectra features selection, in: 2015 IEEE 6th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP), IEEE, 2015, pp. 437–440.

[28] P. E. Gill, W. Murray, M. A. Saunders, User's guide for SNOPT 5.3: A Fortran package for large-scale nonlinear programming, Department of Mathematics, University of California, San Diego, USA, 1998.

[29] R. J. Tibshirani, J. E. Taylor, E. J. Candès, T. Hastie, The solution path of the generalized lasso, Ph.D. thesis, Stanford University (2011).

[30] J. Friedman, T. Hastie, H. Höfling, R. Tibshirani, Pathwise coordinate optimization, The Annals of Applied Statistics 1 (2) (2007) 302–332.

[31] B. Xin, Y. Kawahara, Y. Wang, W. Gao, Efficient generalized fused lasso and its application to the diagnosis of Alzheimer's disease, in: Twenty-Eighth AAAI Conference on Artificial Intelligence, 2014, pp. 2163–2169.

[32] A. Beck, M. Teboulle, A fast iterative shrinkage-thresholding algorithm for linear inverse problems, SIAM Journal on Imaging Sciences 2 (1) (2009) 183–202.

[33] H. Hoefling, A path algorithm for the fused lasso signal approximator, Journal of Computational and Graphical Statistics 19 (4) (2010) 984–1006.

[34] J. Liu, L. Yuan, J. Ye, An efficient algorithm for a class of fused lasso

problems, in: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2010, pp. 323–332.

[35] J. Zhou, J. Liu, V. A. Narayan, J. Ye, Modeling disease progression via fused sparse group lasso, in: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2012, pp. 1095–1103.

[36] J. Nocedal, S. J. Wright, Numerical optimization, 2nd Edition, Springer Series on Operation Research and Financial Engineering, New York, USA, 2006.

[37] P. Tatzer, M. Wolf, T. Panner, Industrial application for inline material sorting using hyperspectral imaging in the NIR range, Real-Time Imaging 11 (2) (2005) 99–107.

[38] S. C. Yoon, B. Park, K. C. Lawrence, W. R. Windham, G. W. Heitschmidt, Line-scan hyperspectral imaging system for real-time inspection of poultry carcasses with fecal material and ingesta, Computers and Electronics in Agriculture 79 (2) (2011) 159–168.

[39] S. Serranti, A. Gargiulo, G. Bonifazi, Classification of polyolefins from building and construction waste using NIR hyperspectral imaging system, Resources, Conservation and Recycling 61 (2012) 52–58.

[40] G. Elmasry, M. Kamruzzaman, D. W. Sun, P. Allen, Principles and applications of hyperspectral imaging in quality evaluation of agro-food products: a review, Critical Reviews in Food Science and Nutrition 52 (11) (2012) 999–1023.

[41] D. Lorente, N. Aleixos, J. Gómez-Sanchis, S. Cubero, O. L. García-Navarrete, J. Blasco, Recent advances and applications of hyperspectral imaging for fruit and vegetable quality assessment, Food and Bioprocess Technology 5 (4) (2012) 1121–1142.

[42] V. Mazet, C. Carteret, D. Brie, J. Idier, B. Humbert, Background removal from spectra by designing and minimising a non-quadratic cost function, Chemometrics and Intelligent Laboratory Systems 76 (2) (2005) 121–133.

[43] C. Krongtaew, K. Messner, T. Ters, K. Fackler, Characterization of key parameters for biotechnological lignocellulose conversion assessed by FT-NIR spectroscopy. Part I: Qualitative analysis of pretreated straw, BioResources 5 (4) (2010) 2063–2080.

[44] M. Schwanninger, J. Rodrigues, K. Fackler, A review of band assignments in near infrared spectra of wood and wood components, Journal of Near Infrared Spectroscopy 19 (2011) 287?308.

[45] R. Flamary, N. Jrad, R. Phlypo, M. Congedo, A. Rakotomamonjy, Mixed-norm regularization for brain decoding, Computational and Mathematical Methods in Medicine 2014 (2014) ID 317056.