Accepted for publication ICPR 2020 Milan

Classification and feature selection using a primal-dual method and projection on structured constraints

Michel Barlaud
Fellow, IEEE
Côte d'Azur University
CNRS, I3S
Sophia Antipolis, France
michel.barlaud@i3s.unice.fr

Antonin Chambolle
École Polytechnique
CNRS, CMAP
Palaiseau, France
antonin.chambolle@cmap.polytechnique.fr

Jean-Baptiste Caillau
Côte d'Azur University
CNRS, Inria, LJAD
Nice, France
jean-baptiste.caillau@univ-cotedazur.fr

***Abstract*—This paper concerns feature selection using supervised classification on high dimensional datasets. The classical approach is to project data onto a low dimensional space and classify by minimizing an appropriate quadratic cost. We first introduced a matrix of centers in the definition of this cost. Moreover, as quadratic costs are not robust to outliers, we propose instead to use an ℓ_1 cost (or Huber loss to mitigate overfitting issues). While control on sparsity is commonly obtained by adding an ℓ_1 constraint on the vectorized matrix of weights used for projecting the data, we propose to enforce structured sparsity. To this end we used constraints that take into account the matrix structure of the data, based either on the nuclear norm, on the $\ell_{2,1}$ norm, or on the $\ell_{1,2}$ norm for which we provide a new projection algorithm. We optimize simultaneously the projection matrix and the matrix of centers with a new tailored constrained primal-dual method. The primal-dual framework is general enough to encompass the various robust losses and structured constraints we use, and allows a convergence analysis. We demonstrate the effectiveness of this approach on three biological datasets. Our primal-dual method with robust losses, adaptive centers and structured constraints does significantly better than classical methods, both in terms of accuracy and computational time.**

I. Introduction

We considered methods in which feature selection is embedded into a classification process, see [1], [2]. Sparse learning based methods have received a lot of attention in the last decade because of their high level of performance. The basic idea is to use a sparse regularizer that forces some coefficients to be zero. To achieve feature selection, the *Least Absolute Shrinkage and Selection Operator* (LASSO) formulation [3], [4], [5], [6], [7] adds an ℓ_1 penalty term to the classification cost, which can be interpreted as convexifying an ℓ_0 penalty [8], [9]. An issue concerns the use of the Frobenius norm (that is the ℓ_2 norm of the vectorized data) for the loss term is not robust to outliers. We propose a more drastic approach that uses an ℓ_1 norm on the regularization term and a Huber loss function. The idea is then to combine a splitting method [10] with a proximal approach. Proximal methods were introduced in [11] and have been used intensively in signal processing; see, *e.g.*, [12], [13], [14], [15], [16], [17]. The first step is the computation of the proximal operator for the loss term that involves the matrix of projection weights and the matrix of centers. We tackle this point by dualizing the norm computation. When using an ℓ_1 (or Huber loss) penalization to ensure sparsity, the computational time due to the treatment of the corresponding hyper-parameter is expensive (see [4], [18]). We propose alternatively a constrained approach that takes advantage of an available efficient projection on the ℓ_1 ball [19], [20] and projection on convex sets [21]. Regarding structured sparsity, the most common approaches are based on penalizations; see, *e.g.*, *group LASSO* [22], [23], [24], [25], [26], [27], [6], [7], *Exclusive LASSO* [28], [29], [30], or $\ell_{2,1}$ based penalties [31], [32], [33]. To the best of our knowledge, the only constrained approach was proposed in [32] for Group LASSO.

The paper is organized as follows. We first present our setting that combines dimension reduction, classification and feature selection. We provide in Section III a flexible primal-dual scheme for this constrained formulation of the classification problem. In Section IV, we lay the emphasis on structured sparsity and replace the ℓ_1 hard constraint by constraints based either on the nuclear norm, the $\ell_{2,1}$ norm (Group LASSO), or the $\ell_{1,2}$ norm (Exclusive LASSO). In Section V, we finally compare different methods experimentally. The tests involve four different bases: a synthetic dataset, and three biological datasets (two mass-spectrometric datasets and a single cell one).

II. A robust augmented variable modelling

Let X be the $m \times d$ data matrix made of m line samples $x_1, \ldots, x_m$ that belong to the d-dimensional space of features. Let $Y \in \{0,1\}^{m \times k}$ be the matrix of labels where $k \geqslant 2$ is the number of clusters. Each line of Y has exactly one nonzero element equal to one, $y_{ij} = 1$ indicating that the sample x_i belongs to the j-th cluster. Projecting the data in lower dimension is crucial to be able to separate them accurately. Let

W be the $d \times k$ projection matrix, where $k \ll d$.[1] The classical approach is to minimize the following squared Frobenius norm, see [7]:

$$\min_W \|Y - XW\|_F^2. \quad (1)$$

A more accurate criterion proposed in [18] based on the sum of square difference (and used in k-means [34]) can be cast as follows:

$$\|Y\mu - XW\|_F^2 = \sum_{j=1}^{k} \sum_{l \in C_j} \|(XW)(l,:) - \mu_j\|_2^2, \quad (2)$$

where $(XW)(l,:)$ denotes the l-th row of matrix (XW), $C_j \subset \{1, \ldots, m\}$ denotes the j-th class, and where the row vector μ_j is the centroid of this class. The matrix of centers μ is so a square matrix of order k. As the squared Frobenius loss is smooth, Fista algorithm [35] can be used. It is well known that the Frobenius norm is not robust to outliers, so we robustify the approach by replacing the Frobenius norm by the ℓ_1 norm of the loss term,

$$\|Y\mu - XW\|_1 = \sum_{j=1}^{k} \sum_{l \in C_j} \|(XW)(l,:) - \mu_j\|_1. \quad (3)$$

While for $\mu = I_k$ one has a robustified version of the loss (1), we will actually optimize jointly in (W, μ), adding some *ad hoc* penalty to break homogeneity and avoid the trivial solution $(W, \mu) = (0, 0)$. The advantage of optimising also the centers is illustrated in Section V (see Figure 4). Using both the projection W and the centers μ learnt during the training set, a new query x in the test set (a dimension d row vector) is classified according to the following rule: it belongs to class number j^* if and only if[2]

$$j^* \in \arg \min_{j \in \{1,\ldots,k\}} \|\mu_j - xW\|_1. \quad (4)$$

III. Primal-dual scheme, constrained formulation

A. Minimization of the ℓ_1 loss

We advocate the use of a convex constrained formulation of the supervised classification problem in order to reduce the computational cost due to the estimation of the hyper-parameter in penalty methods [32]. The number of selected features is a linear function of the constraint η (see Figure 2), thus the constraint η is easily tuned. Let us consider

$$\min_{(W,\mu)} \|Y\mu - XW\|_1 + \frac{\rho}{2}\|I_k - \mu\|_F^2 \text{ s.t. } \|W\|_1 \leqslant \eta \quad (5)$$

where I_k denotes the order k identity matrix. As previously stated, an ℓ_2 regularization has been added in order to avoid the trivial solution $(W, \mu) = (0, 0)$ while maintaining the matrix of centers μ not too far away for a rank k matrix spanning all directions in the low dimensional space used for projection.

[1]Note that the dimension of the projection space is equal to the number of clusters.

[2]In practice, there is one and only one such class.

So as to cope with the computation of proximal operator wrt. W, we dualize as:

$$\min_{(W,\mu)} \max_{\|Z\|_\infty \leqslant 1} \langle Z, Y\mu - XW \rangle + \frac{\rho}{2}\|I_k - \mu\|_F^2 \quad \text{s.t.} \|W\|_1 \leqslant \eta. \quad (6)$$

A possible primal-dual algorithm [14], [17] is then as follows:

$$\begin{aligned} W^{n+1} &= \arg \min_{\|W\|_1 \leqslant \eta} \frac{1}{2\tau}\|W - W^n\|_F^2 - \langle X^T Z^n, W \rangle, \\ \mu^{n+1} &= \arg\min_{\mu} \frac{1}{2\tau_\mu}\|\mu - \mu^n\|_F^2 + \frac{\rho}{2}\|\mu - I\|_F^2 \\ &\quad + \langle Y^T Z, \mu \rangle, \\ Z^{n+1} &= \text{proj}_{\{\|Z\|_\infty \leq 1\}} Z + \sigma(Y(2\mu^{n+1} - \mu^n) \\ &\quad - X(2W^{n+1} - W^n)). \end{aligned}$$

These proximal steps are computed as follows:

$$\begin{aligned} W^{n+1} &= \arg \min_{\|W\|_1 \leqslant \eta} \frac{1}{2\tau}\|W - (W^n + \tau X^T Z^n)\|^2, \\ &= \text{proj}_{\ell_1}(W^n + \tau X^T Z^n, \eta) \end{aligned}$$

where $\text{proj}_{\ell_1}(\cdot, \eta)$ is the projection on the ℓ_1 ball of radius η. Similarly,

$$\mu^{n+1} = \frac{1}{1 + \tau_\mu \rho}(\mu^n + \rho\tau_\mu I - \tau_\mu Y^T Z^n),$$

and the iteration on Z is standard. The resulting scheme is summarized by Algorithm 1. The convergence condition on

Algorithm 1 Primal-dual algorithm, ℓ_1 loss.

Input: $X, Y, N, \sigma, \tau, \tau_\mu, \eta, \rho, \mu_0, W_0, Z_0$
for $n = 1, \ldots, N$ **do**
  $W_{\text{old}} := W$; $\mu_{\text{old}} := \mu$
  $W := \text{proj}_{\ell_1}(W + \tau \cdot (X^T Z), \eta)$
  $\mu := \frac{1}{1+\tau_\mu \cdot \rho}(\mu_{\text{old}} + \rho \cdot \tau_\mu I_k - \tau_\mu \cdot (Y^T Z))$
  $Z := Z + \sigma \cdot (Y(2\mu - \mu_{\text{old}}) - X(2W - W_{\text{old}})))$
  $Z := \max(-1, \min(1, Z)))$
end for
Output: W, μ

the step-sizes τ, τ_μ and σ (see [36]) imposes that:

$$\sigma\left(\frac{\tau_\mu}{1 + \tau_\mu(\rho/4)}\|Y\|^2 + \tau\|X\|^2\right) < 1. \quad (7)$$

The norms involved in the previous expression are operator norms, that is, *e.g.*,

$$\|X\| = \sup_{\|W\|_F \leq 1} \|XW\|_F = \sup_{\|v\|_2 \leq 1} \|X(:)v\|_2. \quad (8)$$

Since the problem is strongly convex with respect to variable μ, then the descent step for the corresponding variable μ can be increased with respect to the choice in [14], [17].

B. A robust approach using Huber loss.

The drawback of the term $\|Y\mu - XW\|_1$ is that it enforces equality of the two matrices out of a sparse set, tuning the parameters to obtain a perfect matching of the training data. In order to soften this behaviour, we use the Huber function instead of the ℓ_1 norm. Letting $h_\delta(t) = t^2/(2\delta)$ for $|t| \leq \delta$, and $|t| - \delta/2$ for $|t| \geq \delta$, we replace the loss by

$$h_\delta(Y\mu - XW) := \sum_{i=1}^{m}\sum_{j=1}^{k} h_\delta((Y\mu - XW)_{i,j}), \qquad (9)$$

and consider the following updated problem:

$$\min_{(W,\mu)} h_\delta(Y\mu - XW) + \frac{\rho}{2}\|I_k - \mu\|_F^2 \text{ s.t. } \|W\|_1 \leqslant \eta. \quad (10)$$

This approach ensures that, up to a sparse set of outliers, the components of $Y\mu$ at optimality will lie at distance $\approx \delta$ of the components of XW. We can tune the primal-dual method to solve this problem, even with acceleration, *cf* [14], [17]. One has $h_\delta^*(s) = \delta s^2/2$ if $|s| \leq 1$, $+\infty$ else, hence we find the following saddle-point problem:

$$\min_{\mu,\ \|W\|_1 \leqslant \eta} \max_{\|Z\|_\infty \leqslant 1} \langle Z, Y\mu - XW\rangle + \frac{\rho}{2}\|I_k - \mu\|_F^2 - \frac{\delta}{2}\|Z\|_F^2. \qquad (11)$$

Compared to Algorithm 1, only the iteration on Z has to be updated using an appropriate re-weighting:

$$Z := \frac{1}{1 + \sigma\cdot\delta}(Z + \sigma\cdot(Y(2\mu - \mu_{\text{old}}) - X(2W - W_{\text{old}}))),$$

$$Z := \max(-1, \min(1, Z)).$$

IV. Structured sparsity

This section deals with the following structured constraint sparsity methods: nuclear constraint, Group LASSO and Exclusive LASSO methods. Our framework turns to be flexible enough to encompass these methods.

A. Nuclear norm

In applications, it is often crucial not to forget the matrix structure of the projection matrix W. To preserve this information, instead of the ℓ_1 norm of the vectorization of the matrix, one can consider the nuclear norm $\|W\|_*$, that is the sum of the singular values of W. This norm is very popular for matrix completion [37], *e.g.*. The projection on the nuclear ball can be computed according to the lemma below, and one can reuse the algorithms presented in the previous section after updating the projection for the iteration on W (see Algorithm 1).

Lemma 1 *If $W = V^T\Sigma U$ is the SVD decomposition of the matrix W, then the projection on the closed nuclear ball of radius η is $W^* = V^T\Sigma^* U$ where Σ^* is the diagonal matrix whose entries are the projection of the diagonal of Σ (that is of the singular values of W) on the ℓ_1 ball of radius η.*

B. Group LASSO

Group LASSO was first introduced in [22]. The main idea is to enforce parameters of different classes to share common features. Group sparsity reduces so complexity by eliminating entire features. It consists in using the $\ell_{2,1}$ norm for the constraint on W, which is defined as follows. The rowwise $\ell_{2,1}$ norm of a $d \times k$ matrix W (whose rows are denoted w_i, $i = 1, \ldots, d$) is

$$\|W\|_{2,1} := \sum_{i=1}^{d} \|w_i\|_2. \qquad (12)$$

We use the standard following approach to compute the projection W of a $d \times k$ matrix V (whose rows are denoted v_i, $i = 1, \ldots, d$) on the $\ell_{2,1}$ ball of radius η: compute the vector $t = (t_1, \ldots, t_d)$ which is the projection of the vector $(\|v_1\|_2, \ldots, \|v_d\|_2)$ on the ℓ_1 ball of $\mathbf{R}^d$ of radius η; then, each row of the projection is obtained by a series of projections on ℓ_2 balls of radii t_i, $i = 1, \ldots, d$, in $\mathbf{R}^k$:

$$w_i = \frac{t_i v_i}{\max\{t_i, \|v_i\|_2\}}, \quad i = 1, \ldots, d.$$

This last operation is denoted as $w_i := \text{proj}_{\ell_2}(v_i, t_i)$ in Algorithm 2. This is standard (and easy to derive); a variant is proposed in [38], however it requires to compute the roots of an equation using bisection, which is slow.

Algorithm 2 Projection on the $\ell_{2,1}$ ball.

Input: V, η
$t := \text{proj}_{\ell_1}((\|v_1\|_2, \ldots, \|v_d\|_2), \eta)$
for $i = 1, \ldots, d$ **do**
  $w_i := \text{proj}_{\ell_2}(v_i, t_i)$
end for
Output: W

C. Exclusive LASSO

Exclusive sparsity or exclusive LASSO was first introduced in [28]. The main idea is that if one feature in a class is selected (large weight), the method tends to assign small weights to the other features in the same class. This is enforced by employing the row-wise $\ell_{1,2}$ norm, defined for a $d \times k$ matrix with row vectors $w_1, \ldots, w_d$ as (compare (12))

$$\|W\|_{1,2} := \left(\sum_{i=1}^{d} \|w_i\|_1^2\right)^{1/2}.$$

So given a $d \times k$ matrix V, the projection on the corresponding balls consists in finding a matrix W which solves

$$\min_W \sum_{i,j} |w_{i,j} - v_{i,j}|^2 \text{ s.t.} \sum_i \left(\sum_j |w_{i,j}|\right)^2 \leq \eta^2. \qquad (13)$$

Our approach is to introduce a Lagrange multiplier for the constraint and then compute it by a variant of Newton's method. This is detailed in Algorithm 3 (see details in Appendix).

Algorithm 3 Projection on the $\ell_{1,2}$ ball.

Input: V, η, N
for $i = 1, \dots, d$ **do**
  Sort in decreasing order $|v(i,:)|$
  for $j = 1, \dots, k$ **do**
    $S_{i,j} := \sum_{l=1}^{j} |v_{i,j}|$
  end for
end for
$\lambda = \max_{p \in \{1,\dots,k\}} \frac{\frac{1}{\eta}\sqrt{\sum_i S_{i,p}^2} - 1}{p}$
$p_i = \arg\max_{p_i \in \{1,\dots,k\}} S_{i,p_i}/(1+\lambda p_i)$
if $\sum_{i=1}^{d} \left(\frac{S_{i,p_i}}{1+\lambda p_i}\right)^2 \leq \eta^2$, **terminate**
for $n = 1, \dots, N$ **do**
  $\lambda := \lambda + \frac{\sum_{i=1}^{d} \left(\frac{S_{i,p_i}}{1+\lambda p_i}\right)^2 - \eta^2}{2\sum_{i=1}^{d} p_i \frac{\left(S_{i,p_i}\right)^2}{(1+\lambda p_i)^3}}$
  for $i = 1, \dots, d$ **do**
    $p_i := \arg\max_{p_i \in \{1,\dots,k\}} \frac{S_{i,p_i}}{1+\lambda p_i}$.
  end for
end for
$\delta_i = \lambda \frac{S_{i,p_i}}{1+\lambda p_i}$
Output: $w_{i,j} = (|v_{i,j}| - \delta_i)_+ \mathbf{sign}(v_{i,j})$

V. Numerical experiments

A. Experimental settings

Our primal-dual method can be applied to classification problem with feature selection on a high dimensional dataset (stemming from computational biology, image recognition, social networks analysis, customer relationship management, *etc.*) We provide an experimental evaluation in computational biology on a single-cell sequencing dataset and two proteomic mass-spectrometric datasets. Feature selection is based on the sparsity induced by the ℓ_1 constraint. In class k, the gene j will be selected if $|w_{j,k}| > \varepsilon$. The set of non-zero column coefficients is interpreted as the signature of the corresponding class. We use [19] method to compute the projection on the ℓ_1-ball. We report the classical accuracy versus parameters using four folds for cross validation. Processing times are obtained on a laptop computer using an i7 processor (3.1 Ghz). In our experiments, we normalize the features according to $\sigma_{\max}(X) = 1$ (setting the largest singular value fixes the operator norm the matrix), and we set $\mu^0 = I_k$ and $\rho = 1$. We choose the value of the ℓ_1 bound η in connection with the desired number of genes. We set $\tau = 1$,

$$\tau_\mu = \frac{\beta}{2\sqrt{m}\|Y\| - (1/4)\beta\rho},$$

then we tune β and compute σ using equation (7).

Ovarian proteomic dataset [39]. The data available on UCI database consists of mass-spectra obtained with the SELDI technique. The samples include patients with cancer (ovarian or prostate cancer), and healthy or control patients. The dataset is composed of 216 samples and 15000 features.

Lung proteomic dataset [40]. The data were collected using unbiased liquid chromatography/mass spectrometry. The dataset is comprised of 1005 patients (469 among them with lung cancer and 536 control patients), and 2944 features.

Single cell dataset [41]. The dataset comes from a collection of mouse cells from the primary somatosensory cortex (S1) and the hippocampal CA1 region. This dataset is composed of 3005 cells, 7364 genes and $k = 9$ clusters. Note that class 8 and 9 have only 20 and 60 cells respectively. The set of selected features is currently evaluated by biologist partners.

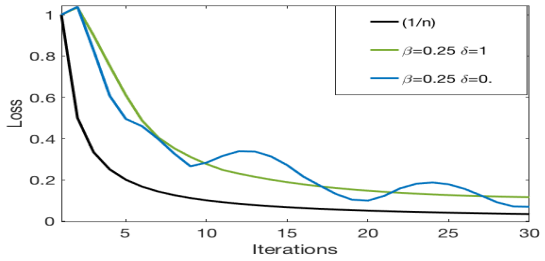

Fig. 1: This figure shows the benefit for the convergence of using Huber loss ($\delta = 1$) instead of ℓ_1 loss ($\delta = 0$) on Ovarian dataset.

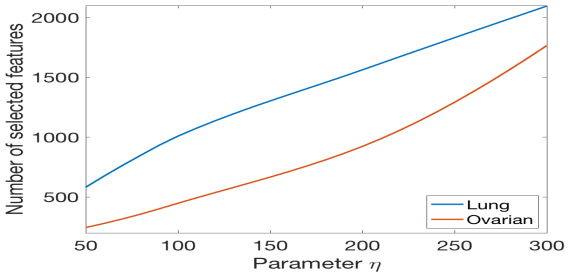

Fig. 2: Number of selected features versus constraint η for Lung and Ovarian databases. The number of genes grows linearly with the ℓ_1 bound.

Figure 1 shows the convergence of the ℓ_1 loss and Huber loss in the training set (normalized by the value of the first iterate). Note an oscillatory convergence of the ℓ_1 loss while convergence of Huber loss is perfectly smooth. The linear dependence of the number of selected genes on the ℓ_1 bound on the projection matrix is depicted Figure 2. Figure 3 shows a break in the slope of accuracy curve versus the number of selected genes for the three biological databases on which the tests wre performed. This drastic change of slope can be easily detected and used to determine the relevant (and small) number of genes to select for the analysis. For the experiments in next subsection, we chose the constraint such that the number of selected genes corresponds to this rapid change of slope.

B. Comparison of loss functions

Figure 4 and Table I show the improvement in accuracy on all biological datasets when using Huber loss instead of ℓ_1 or

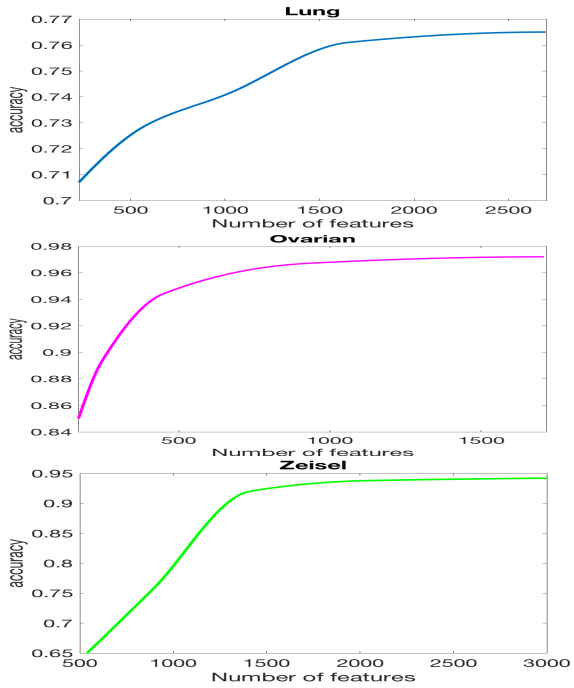

Fig. 3: Accuracy versus number of features for the three biological databases. The rapid change of slope allows to find the number of genes to select.

Methods	ℓ_1	Huber ($\mu = I$)	Huber	Froben.
Ovarian	90.%	95.83%	98.61%	90.7%
Lung	66%	72.1%	76.6%	70.2 %
Zeisel	79.6%	94.2%	95.5%	94.2 %

TABLE I: Accuracy test on the three biological datasets: using the primal-dual approach with Huber loss and optimizing on the matrix of centers (vs. fixed $\mu = I$) significantly improves accuracy over the other methods on all datasets.

Frobenius loss; ℓ_1 loss suffers from overfitting while Frobenius loss is not robust enough. Moreover, optimizing also wrt. the matrix of centers, μ, improved the accuracy by 2.78% on the Ovarian, 4.5% on Lung and by 1.5% on the Zeisel datasets respectively.

C. Comparison of computational times

Although it is not obvious to carry out a fair comparison between the different methods (because of the issue of between of implementation or choice of parameters issues), we propose the following numerical comparison. FISTA requires that one part of the objective is smooth and the other can be easily solved implicitly. This is the case for the non-robust Frobenius loss:

$$\min_{W,\mu} \|Y\mu - XW\|_F^2 + \frac{\rho}{2}\|I_k - \mu\|_F^2 \text{ s.t. } \|W\|_1 \leqslant \eta.$$

We report computational time in this case in Table II. However, with the squared Frobenius norm replaced by an ℓ_1 norm, this structure is lost (also in the dual, as the objective is strongly convex only in μ), and there is no way to implement

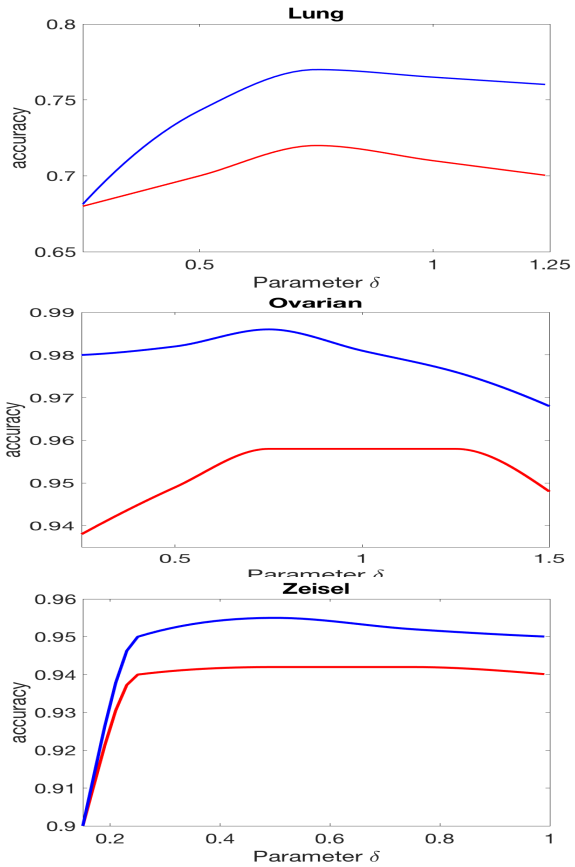

Fig. 4: Accuracy as a function of parameter δ. Blue curve is obtained using optimised μ centers, the red one fixing $\mu = I$. Top: Lung dataset, middle: Ovarian dataset, bottom: Zeisel dataset.

an accelerated method (while a subgradient method would be more expensive). The only reasonable alternative would be ADMM [42], [15], which makes sense as long as the matrix factorizations are not too hard to tackle (here it would be very computationally expensive when m and d are large and when X is a full rank matrix). For the numerical comparison, we generated $m \times d$ random data matrices, with random labels for $k = 2$ classes. Then we compared computational times of algorithms the primal-dual algorithm (with Huber loss and optimization on μ) vs. FISTA and ADMM [15]. Figure 5 and Table II show that computational time as a function of the number of features d of primal-dual is linear, while it is quadratic for FISTA. Figure 5 also shows that computational time as a function of the number of samples m is linear both for primal-dual and FISTA. Table II shows that the computational time of ADMM is one or order of magnitude greater than the others because of the linear algebra involved. Table III shows that the primal-dual method is 50 times faster than FISTA with the Ovarian, 3 times faster for the two other real datasets.

D. Comparison of projections on synthetic random data

We evaluate the cost of the different constraint projections using random matrices of size $d \times k$, keeping one of the two

TABLE II: Computational time of primal-dual vs. FISTA and ADMM on random data matrices of size $m \times d$ with $m = 1000$ and d ranging from 1000 to 15000. Time is in seconds. (NC = No Convergence)

d	1000	3000	5000	10000	15000
Primal-dual	0.025	0.12	0.205	0.42	0.63
Fista	0.075	0.481	1.16	4.62	10.6
ADMM	0.65	4.52	58	NC	NC

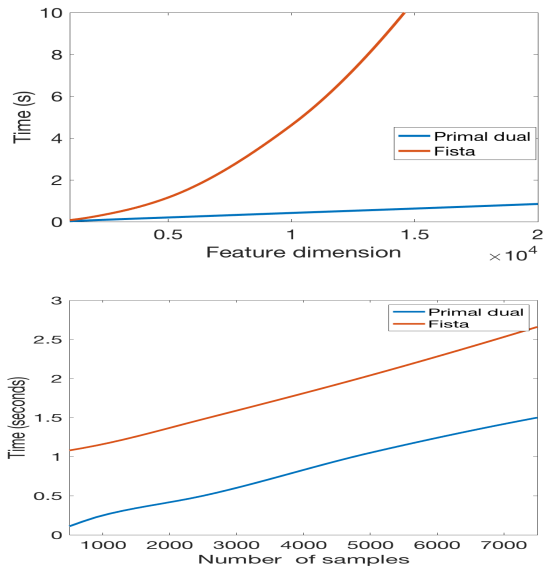

Fig. 5: Comparison of primal-dual and FISTA on a synthetic dataset. Top: computational time as function of the number of features d for $m = 1000$. Bottom: computational time as function of the number of samples m for $d = 5000$. While the two methods behave similarly wrt. the number of samples, primal-dual scales much more favorably than FISTA wrt. the number of features (a key issue for biological datasets for which the number of genes is large).

TABLE III: Comparison of computational time (seconds) for primal-dual and FISTA on biological datasets.

Dataset	primal-dual	FISTA
Ovarian ($m = 216$, $d = 15000$)	0.19	9.5
Lung ($m = 1005$, $d = 2944$)	0.12	0.4
Zeisel ($m = 3005$, $d = 7364$)	1.07	3.44

dimensions fixed. The discussion on the complexity of the various projection methods is delicate so we focus on a mere comparison of computation times. Figure 6 (top) shows that for small k the projection cost on the nuclear constraint is similar to the projection cost on the $\ell_{2,1}$ ball. The projection cost on the $\ell_{2,1}$ ball with our method outperforms the bisection method [38]. Figure 6 (bottom) shows that the cost of the projection on the $\ell_{1,2}$ ball grows linearly with d and k, and is slightly higher than for the projection on the ℓ_1 ball. Figure 7 shows that the $\ell_{1,2}$ constraint gives better results than other structured constraints. Our primal-dual algorithm provides accuracy for each cluster. In the case of the Single cell Zeisel dataset, we report accuracy for different constraints on class 8 and 9 that have only 20 and 60 cells, respectively. Figure 7 also shows

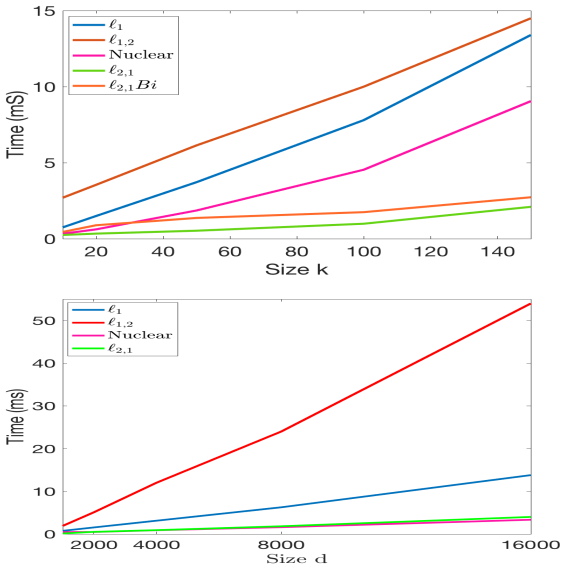

Fig. 6: Comparisons of projection numerical costs (performed on randomly generated $d \times k$ matrices). Top: time as function of k for $d = 1000$. Bottom: time as function of d for $k = 10$.

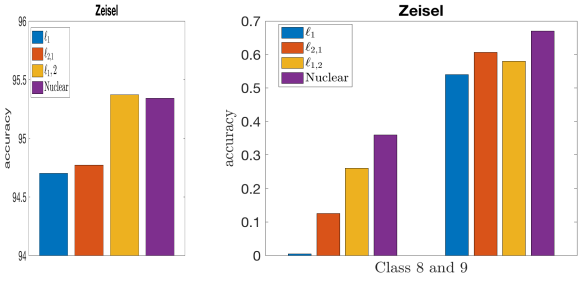

Fig. 7: Single cell Zeisel dataset: comparison of ℓ_1, $\ell_{2,1}$ and nuclear norm constraint. Left: global accuracy. Right: accuracy in classes 8 and 9 (classes with a small number of features).

that using the nuclear norm, the $\ell_{2,1}$ or $\ell_{1,2}$ norm to enforce structured sparsity improves accuracy on small classes. Figure 8 shows the fast convergence of the modified Newton inner loop in Algorithm 3, we typically use $N = 4$ or $N = 5$.

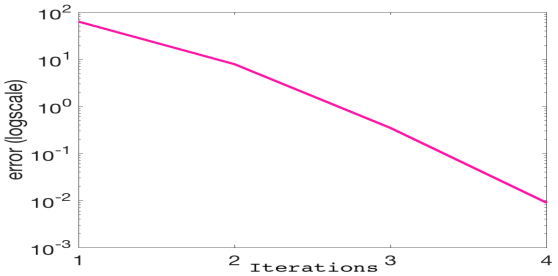

Fig. 8: Convergence of Algorithm 3 (modified Newton) for the $\ell_{1,2}$ projection.

VI. Conclusion

For supervised classification and feature selection, we advocate the use of a robust loss (ℓ_1, or Huber to deal with overfitting) together with joint optimization of the projection matrix (a crucial procedure for high-dimensional data) and of the matrix of centers of the clusters. We also put forward constrained formulations (instead of penalizations) to enforce sparsity for feature selection. The primal-dual framework we propose is flexible enough to encompass not only several kind of losses (smooth or nonsmooth), but also to cover various kind of constraints. This includes *ad hoc* constraints for structured sparsity such as Group or Exclusive LASSO. In the case of Exclusive LASSO, we provide original algorithms for the projection on the $\ell_{1,2}$ ball and illustrate its numerical efficiency. The merits of the primal-dual method are shown using several databases, synthetic and biological [39], [40], [41]. The method (the convergence of which is studied in [36]) can be extended to other criteria on condition that efficient projection (on the dual ball for the loss data term) algorithms are available.

Appendix

Projection on the $\ell_{1,2}$ ball. Given $(v_{i,j})_{i=1,\dots,n}^{j=1,\dots,m}$, our problem is to find $w=(w_{i,j})$ solving

$$\min_w \left\{ \sum_{i,j} |w_{i,j}-v_{i,j}|^2 : \sum_i \left(\sum_j |w_{i,j}| \right)^2 \leq \eta^2 \right\}. \quad (14)$$

The most direct approach is to compute the Lagrange multiplier associated with the $\ell_{1,2}$ constraint by a suitable adaption of Newton's method. For $\lambda>0$, we first consider

$$\min_w \sum_{i,j} |w_{i,j}-v_{i,j}|^2 + \lambda \sum_i \Big(\sum_j |w_{i,j}| \Big)^2. \quad (15)$$

This has the advantage to decouple into independent minimization problems as follows:

$$\sum_i \min_{w_{i,\cdot}} \sum_j |w_{i,j}-v_{i,j}|^2 + \lambda \Big(\sum_j |w_{i,j}| \Big)^2. \quad (16)$$

We then focus on the generic sub-problem (dropping index i):

$$\min_{w_j} \sum_j |w_j-v_j|^2 + \lambda \Big(\sum_j |w_j| \Big)^2. \quad (17)$$

Its solution is easily seen to satisfy:

$$w_j = \Big(|v_j| - \lambda \sum_{j'} |w_{j'}| \Big)^+ \mathrm{sgn} v_j. \quad (18)$$

Hence, letting $\delta = \lambda \sum |w_j|$, one sees that one needs to find δ such that $\delta = \lambda \sum_j (|v_j|-\delta)^+$, which has a unique solution in $[0, \max_j |v_j|]$. If $|v_j|$ are sorted in decreasing order, one must find p such that if

$$\delta = \frac{\lambda \sum_{j=1}^p |v_j|}{1+\lambda p} \quad (19)$$

one has $|v_p| \geq \delta$, $|v_{p+1}| \leq \delta$. If we interpret the RHS of (19) as an average of 0 with weight $1/\lambda$ and $|v_j|$: $((1/\lambda)\times 0 + \sum_{j=1}^p |v_j|)/(1/\lambda + p)$, we see that it increases as long as one adds terms above the average, and then decreases, so that:

$$\delta = \lambda \max_p \frac{\sum_{j=1}^p |v_j|}{1+\lambda p}. \quad (20)$$

Observe in addition that

$$\sum_j (|v_j|-\delta)^+ = \frac{\delta}{\lambda} = \max_p \frac{\sum_{j=1}^p |v_j|}{1+\lambda p}. \quad (21)$$

Returning to the original problem (14), we see that one needs to find $\lambda \geq 0$ such that (assuming all $|v_{i,\cdot}|$ are sorted in decreasing order and defining $S_{i,p} := \sum_{j=1}^p |v_{i,j}|$):

$$\sum_{i=1}^n \max_{p_i} \left(\frac{S_{i,p_i}}{1+\lambda p_i} \right)^2 = \eta^2. \quad (22)$$

This is found by Newton's method. The function in (22) is convex (as a max of convex functions), decreasing in λ. Starting from λ^0 and the corresponding values p_i^0 one should compute iteratively:

$$\lambda^{k+1} = \lambda^k + \frac{\sum_i \left(\frac{S_{i,p_i^k}}{1+\lambda^k p_i^k} \right)^2 - \eta^2}{2 \sum_i p_i^k \frac{(S_{i,p_i^k})^2}{(1+\lambda^k p_i^k)^3}} \quad (23)$$

and then update p_i^{k+1} by finding for each i:

$$\max_{p_i} \frac{S_{i,p_i}}{1+\lambda^{k+1} p_i}. \quad (24)$$

This process must converge as the function to invert in (22) is convex and decreasing, in particular if λ^0 is less than the optimal lambda it is easy to see that (λ^k) will converge monotonically, increasing towards the optimal value. It is not difficult to prove that this convergence is at least linear (with rate $1-f'(\lambda^*)/f'(\lambda^0)$ if $f(\lambda)$ denotes the left-hand side of (22) and λ^* the solution), and it is classical that it becomes quadratic when λ^k is close enough to the optimum (hence the importance of finding a good starting point). Once this has converged, one gets the thresholds δ_i by the formula

$$\delta_i = \lambda^k \frac{S_{i,p_i^k}}{1+\lambda^k p_i^k} \quad (25)$$

and then $w_{i,j} = (|v_{i,j}|-\delta_i)^+ \mathrm{sgn} v_{i,j}$ can be easily computed on the unsorted data. The process will converge faster if one can find a good estimate of the optimal λ as an initial guess. One has for the optimal λ^*:

$$\max_{\mathbf{p}=(p_1,\dots,p_n)} \sum_i \frac{S_{i,p_i}^2}{(1+\lambda^* p_i)^2} = \eta^2 \geq \max_p \frac{\sum_i S_{i,p}^2}{(1+\lambda^* p)^2}.$$

The idea here is that the max on arbitrary vectors $(p_1,\dots,p_n)$ is replaced with a (smaller) max over vectors $(p,\dots,p)$ with identical coordinates. It follows easily that:

$$\lambda^* \geq \max_p \frac{\frac{1}{\eta}\sqrt{\sum_i S_{i,p}^2} - 1}{p}. \quad (26)$$

In practice, we use the right-hand side of (26) as initial λ^0. This yields a good precision in a small number of iterations, typically $N \approx 4$ to 5, see Fig. 8.

REFERENCES

[1] T. S. Furey, N. Cristianini, N. Duffy, D. W. Bednarski, M. Schummer, and D. Haussler, "Support vector machine classification and validation of cancer tissue samples using microarray expression data," *Bioinformatics*, vol. 16, no. 10, pp. 906–914, 2000.

[2] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Machine learning*, vol. 46, no. 1-3, pp. 389–422, 2002.

[3] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.

[4] T. Hastie, S. Rosset, R. Tibshirani, and J. Zhu, "The entire regularization path for the support vector machine," *Journal of Machine Learning Research*, vol. 5, pp. 1391–1415, 2004.

[5] J. Friedman, T. Hastie, and R. Tibshirani, "Regularization path for generalized linear models via coordinate descent," *Journal of Statistical Software*, vol. 33, pp. 1–122, 2010.

[6] T. Hastie, R. Tibshirani, and M. Wainwright, "Statistcal learning with sparsity: The lasso and generalizations," *CRC Press*, 2015.

[7] J. Li, K. Cheng, S. Wang, F. Morstatter, R. P. Trevino, J. Tang, and H. Liu, "Feature selection: A data perspective," *ACM Computing Surveys*, vol. 50, 2016.

[8] D. Donoho, "Compressed sensing," *IEEE Trans. Inf. Theor. 52 (4)*, pp. 1289–1306, 2006.

[9] J. Candès, M. B. Wakin, and S. P. Boyd, "Enhancing sparsity by reweighted l1 minimization," *Journal of Fourier analysis and applications*, 2008.

[10] P.-L. Lions and B. Mercier, "Splitting algorithms for the sum of two nonlinear operators," *SIAM Journal on Numerical Analysis*, vol. 16, no. 6, pp. 964–979, 1979.

[11] J. Moreau, "Proximité et dualité dans un espace hilbertien," *Bull. Soc.Math. France., 93*, pp. 273–299, 1965.

[12] S. Mosci, L. Rosasco, M. Santoro, A. Verri, and S. Villa, "Solving structured sparsity regularization with proximal methods," in *Machine Learning and Knowledge Discovery in Databases*. Springer, 2010, pp. 418–433.

[13] P. L. Combettes and J.-C. Pesquet, "Proximal splitting methods in signal processing," in *Fixed-point algorithms for inverse problems in science and engineering*. Springer, 2011, pp. 185–212.

[14] A. Chambolle and T. Pock, "A first-order primal-dual algorithm for convex problems with applications to imaging," *Journal of Mathematical Imaging and Vision*, vol. 40, no. 1, pp. 120–145, May 2011.

[15] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Trends Machine Learning*, vol. 3, pp. 1–122, 2011.

[16] S. Sra, "Scalable nonconvex inexact proximal splitting," in *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012.*, 2012, pp. 539–547.

[17] A. Chambolle and T. Pock, "On the ergodic convergence rates of a first-order primal-dual algorithm," *Math. Program.*, vol. 159, no. 1-2, Ser. A, pp. 253–287, 2016. [Online]. Available: https://doi.org/10.1007/s10107-015-0957-3

[18] D. M. Witten and R. Tibshirani, "A framework for feature selection in clustering," *Journal of the American Statistical Association*, vol. 105, no. 490, pp. 713–726, 2010.

[19] L. Condat, "Fast projection onto the simplex and the l1 ball," *Mathematical Programming Series A*, vol. 158, no. 1, pp. 575–585, 2016.

[20] J. Duchi, S. Shalev-Shwartz, Y. Singer, and T. Chandra, "Efficient projections onto the l 1-ball for learning in high dimensions," in *Proceedings of the 25th international conference on Machine learning*. ACM, 2008, pp. 272–279.

[21] M. Barlaud, W. Belhajali, P. L. Combettes, and L. Fillatre, "Classification and regression using an outer approximation projection-gradient method," *IEEE Transactions on Signal Processing*, vol. 65, no. 17, pp. 4635–4644, 2017.

[22] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 68, no. 1, pp. 49–67, 2006.

[23] J. Friedman, T. Hastie, and R. Tibshirani, "A note on the group lasso and a sparse group lasso," *arXiv preprint arXiv:1001.0736*, 2010.

[24] H. Zou, T. Hastie, and R. Tibshirani, "Sparse principal component analysis," *Journal of computational and graphical statistics*, vol. 15, no. 2, pp. 265–286, 2006.

[25] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables." *J. R. Stat. Soc. Ser. B, 68(1)*, vol. 68(1), pp. 49–67, 2006.

[26] L. Jacob, G. Obozinski, and J.-P. Vert, "Group lasso with overlap and graph lasso," in *Proceedings of the 26th International Conference on Machine Learning (ICML-09)*, 2009, pp. 353–360.

[27] M. Liu and B. C. Vemuri, "A robust and efficient doubly regularized metric learning approach," in *Proceedings of the 12th European Conference on Computer Vision - Volume Part IV*, ser. ECCV'12, 2012.

[28] Y. Zhou, R. Jin, S. Chuâ, and H. Hoi, "Exclusive lasso for multi-task feature selection," in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, vol. 9. Chia Laguna Resort, Sardinia, Italy: PMLR, 13–15 May 2010, pp. 988–995.

[29] D. Kong, R. Fujimaki, J. Liu, F. Nie, and C. Ding, "Exclusive feature learning on arbitrary structures via $nell\_\{1,2\}$ -norm," in *Advances in Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2014, pp. 1655–1663.

[30] J. Yoon and S. J. Hwang, "Combined group and exclusive sparsity for deep neural networks," in *Proceedings of the 34th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 70, International Convention Centre, Sydney, Australia, 06–11 Aug 2017, pp. 3958–3966.

[31] A. Argyriou, T. Evgeniou, and M. Pontil, "Convex multi-task feature learning," *Machine Learning*, vol. 73, no. 3, pp. 243–272, Dec 2008.

[32] J. Liu, S. Ji, and J. Ye, "Multi-task feature learning via efficient l2, 1-norm minimization," in *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, ser. UAI '09. Arlington, Virginia, United States: AUAI Press, 2009, pp. 339–348.

[33] F. Nie, H. Huang, C. Xiao, and C. H. Ding, "Efficient and robust feature selection via joint l2,1-norms minimization," in *Advances in Neural Information Processing Systems 23*. Curran Associates, Inc., 2010, pp. 1813–1821.

[34] J.-B. McQueen, "Some methods for classification and analysis of multivariate observations," *Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1967.

[35] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM journal on imaging sciences*, vol. 2, no. 1, pp. 183–202, 2009.

[36] M. Barlaud, A. Chambolle, and J.-B. Caillau, "Robust supervised classification and feature selection using a primal-dual method," *arXiv cs.LG/1902.01600*, 2019.

[37] D. Zhang, Y. Hu, J. Ye, X. Li, and X. He, "Matrix completion by truncated nuclear norm regularization," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, June 2012.

[38] J. Liu and J. Ye, "Moreau-yosida regularization for grouped tree structure learning," in *Advances in Neural Information Processing Systems 23*. Curran Associates, Inc., 2010, pp. 1459–1467.

[39] I. Guyon, S. Gunn, M. Nikravesh, and L. . Zadeh, "Feature extraction, foundations and applications. studies in fuzziness and soft computing." *Physica-Verlag Springer*, 2017.

[40] E. Mathé *et al*, "Noninvasive urinary metabolomic profiling identifies diagnostic and prognostic markers in lung cancer," *Cancer research*, vol. 74, no. 12, p. 3259—3270, June 2014.

[41] A. Zeisel *et al*, "Cell types in the mouse cortex and hippocampus revealed by single-cell rna-seq," *Science*, vol. 347, pp. 1138–1142, 2015.

[42] D. Gabay and B. Mercier, "A dual algorithm for the solution of nonlinear variational problems via finite element approximation," *Computers & Mathematics with Applications*, vol. 2, no. 1, pp. 17–40, 1976.