ESAIM: PROCEEDINGS AND SURVEYS, August 2021, Vol. 71, p. 11-20
Didier Auroux, Jean-Baptiste Caillau, Régis Duvigneau, Abderrahmane Habbal, Olivier Pantz, Luc Pronzato, Ludovic Rifford

ROBUST CLASSIFICATION WITH FEATURE SELECTION USING AN APPLICATION OF THE DOUGLAS-RACHFORD SPLITTING ALGORITHM

Michel Barlaud[1] and Marc Antonini[2]

Abstract. This paper deals with supervised classification and feature selection with application in the context of high dimensional features. A classical approach leads to an optimization problem minimizing the within sum of squares in the clusters (ℓ_2 norm) with an ℓ_1 penalty in order to promote sparsity. It has been known for decades that ℓ_1 norm is more robust than ℓ_2 norm to outliers. In this paper, we deal with this issue using a new proximal splitting method for the minimization of a criterion using ℓ_1 norm both for the constraint and the loss function. Since the ℓ_1 criterion is only convex and not gradient Lipschitz, we advocate the use of a Douglas-Rachford minimization solution. We take advantage of the particular form of the cost and, using a change of variable, we provide a new efficient tailored primal Douglas-Rachford splitting algorithm which is very effective on high dimensional dataset. We also provide an efficient classifier in the projected space based on medoid modeling. Experiments on two biological datasets and a computer vision dataset show that our method significantly improves the results compared to those obtained using a quadratic loss function.

Résumé. Ce document traite de la classification supervisée et de la sélection des descripteurs avec application dans le contexte des descripteurs de haute dimension. Une approche classique conduit à un problème d'optimisation minimisant la norme quadratique (norme ℓ_2) avec une pénalite de norme ℓ_1 afin de promouvoir la parcimonie. Il est bien connu que la norme ℓ_1 est plus robuste que la norme ℓ_2 aux valeurs aberrantes. Dans cet article, nous abordons cette question avec une nouvelle méthode proximale pour la minimisation d'un critère utilisant la norme ℓ_1 à la fois pour la contrainte et la fonction de perte. Comme le critère ℓ_1 n'est que convexe et non gradient de Lipschitz, nous proposons l'utilisation d'une solution de minimisation de type Douglas-Rachford. Nous tirons parti de la forme du critère et, en utilisant un changement de variable, nous fournissons un nouvel algorithme de type Douglas-Rachford dans le primal qui est très efficace dans le contexte des descripteurs de haute dimension. Nous fournissons également un classifieur efficace dans l'espace projeté basé sur la modélisation de medoïds. Des expériences sur des données biologiques et de vision par ordinateur montrent que notre méthode améliore les résultats par rapport à ceux obtenus avec une fonction de perte quadratique.

1. Introduction

In this paper we consider methods in which feature selection is embedded in a classification process [23, 25]. It is well-known that classification in a high dimension suffers from the curse of dimensionality: As dimensions increase, vectors become indiscernible and the predictive power of the aforementioned methods is drastically reduced [1, 36]. In order to overcome this issue, the main idea of the following methods is to project data onto a low dimensional space. A popular approach for high-dimensional data is to perform *Principal Component*

[1] I3S, Univ. Côte d'Azur & CNRS, F-06900 Sophia Antipolis.

[2] I3S, Univ. Côte d'Azur & CNRS, F-06900 Sophia Antipolis.

Article published online by EDP Sciences and available at https://www.esaim-proc.org or https://doi.org/10.1051/proc/202171102

Analysis (PCA) prior to classification. However in general this approach is not relevant [44]. The Partial least squares (PLS) method closely related to principal component regression is designed to deal with this issue with high dimensional correlated features [37, 45]. An alternative approach is to perform dimension reduction by means of *Linear Discriminant Analysis* (LDA) [14, 17]. The authors of [2] and [21] propose a convex relaxation of this approach in terms of a suitable semi-definite program (SDP) at the cost of an increased computational complexity. Nie *et al* proposed a feature selection based on $\ell_{2,1}$ norm minimization [34]. A popular approach for selecting sparse features in supervised classification or regression is the *Least Absolute Shrinkage and Selection Operator* (LASSO) formulation [22, 28, 33, 42, 46]. The LASSO formulation uses the ℓ_1 norm [8, 15, 18, 19] as an added penalty term instead of an ℓ_0 term. In this paper, we propose to minimize an ℓ_1 norm both on the penalty term and the loss function. In this case, the criterion is convex but not gradient Lipschitz. Thus, we propose to use the Douglas-Rachford splitting method for the minimization of our criterion. This splitting was successfully used in signal processing [4, 9, 10, 11, 20, 39, 41]. However, for classification, we cannot apply straightforwardly Douglas-Rachford like in [7] since the proximal operator for the affine transform involved in the criterion is not available (See equation 1). We take advantage of the particular form of the criterion: The sum of two ℓ_1 norms is equal to a single ℓ_1 norm after a change of variables, and the linear constraint can be integrated into the cost. We also provide an efficient classifier in the projected space based on medoid modelling. The paper is organized as follows. Section 2 deals with criterion and state of the art methods. In section 3 we develop the proposed solution for supervised classification. In section 4 experimental results are provided on real biological datasets and finally in section 5 we conclude the paper.

2. Robust supervised classification

2.1. A robust framework

Let X be the nonzero $m \times d$ matrix made of m line samples $x_1, \ldots, x_m$ belonging to the d-dimensional space of features. Let $Y \in \{0,1\}^{m\times k}$ be the label matrix 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, see [2, 21] . Let $W \in \mathbb{R}^{d\times k}$ be the projection matrix, $k \ll d$. The classical quadratic loss criterion in the projected space is:

$$\min_{W} \frac{1}{2}\|Y - XW\|_F^2 + \lambda\|W\|_1 \tag{1}$$

where $\|.\|_F$ stands for the Frobenius norm. Projecting the data in lower dimension is crucial for efficient feature selection [36]. Besides, it is well known that the quadratic Frobenius loss criterion is not robust to outliers [27, 29], so we propose to minimize instead the ℓ_1 loss cost, with an ℓ_1 penalty regularization to promote sparsity and induce feature selection. So, given the matrix of labels Y, we consider the following convex supervised classification problem:

$$\min_{W} \|Y - XW\|_1 + \lambda\|W\|_1. \tag{2}$$

Here, the ℓ_1 norm of an m by n matrix A denotes the ℓ_1 norm of its vectorization:

$$\|A\|_1 := \|A(:)\|_1 = \sum_{i=1}^{m}\sum_{j=1}^{n} |a_{ij}|.$$

2.2. State of the art methods based on ADMM

Problem (2) falls into the following class:

$$\min_{(x,y)} f(x) + g(y)$$

under the affine constraint $y = Ax$, where x and y are vectors of finite dimensional spaces. Then state of the art methods based on ADMM (Alternating-direction method of multipliers, see [6, 12, 24, 35]) considers the augmented Lagrangian ($\gamma > 0$ is a scalar parameter),

$$L_\gamma(x, y, z) := f(x) + g(y) + \frac{1}{\gamma}(z|Ax - y) + \frac{1}{2\gamma}\|Ax - y\|^2.$$

The Lagrangian is minimized over x, then over y, and a third step of the method updates the Lagrange multiplier z using a proximal maximization step. Note that ADMM algorithm can be derived from an application of the Douglas–Rachford algorithm to the dual.
In our case, we take advantage of the fact that the two functions f and g are simple convex functions (ℓ_1 norms) the sum of which is again a known convex function (yet another ℓ_1 norm). This eliminates one function and allows us to penalize the affine constraint to integrate it into the cost. The resulting proximal operator is a simple projection operator on a linear subspace, which can be computed (see, *e.g.*, Lemma 7). As explained in the next paragraph, it is thus possible to use a Douglas-Rachford algorithm to the primal.

3. Proposed method

3.1. An equivalent formulation

The cost in Problem (2) is the sum of two ℓ_1 norms, which suggests the use of a splitting algorithm [12]. Such methods are very efficient to minimize the sum of convex functions and do not require differentiability properties. Note indeed that, having replaced the squared Frobenius norm by the ℓ_1 norm of $Y - XW$, it is not possible to use forward-backward splitting [12] as none of the functions is differentiable. In order to use the more general Douglas-Rachford scheme (see next subsection) that is able to cope with mere convex functions, one still has to be able to compute their proximity operators. Now, while the prox of the ℓ_1 norm is well known and expressed in terms of soft thresholding, there is no explicit expression for the *prox* of the ℓ_1 norm of the affine transform $Y - XW$. We propose to introduce the auxiliary variable $\zeta := (Y - XW)/\lambda \in \mathbb{R}^{m\times k}$ and to minimize

$$\min_W \|W\|_1 + \|\zeta\|_1$$

with respect to W under the affine constraint

$$XW + \lambda\zeta = Y.$$

The sum of the two ℓ_1 norms is equal to the single ℓ_1 norm of the augmented variable $\tilde{W} := (W, \zeta) \in \mathbb{R}^{(d+m)\times k}$.
Let C be the affine subset of $(d + m) \times k$ matrices such that $\tilde{X}\tilde{W} = Y$ where

$$\tilde{X} := [X\ \lambda I_m] \in \mathbb{R}^{m\times(d+m)}.$$

where I_m is the identity matrix.
The problem can be recast as

$$\min_W \|\tilde{W}\|_1$$

under the affine constraint

$$\tilde{X}\tilde{W} = Y$$

Let moreover i_C be the indicator function of the set C, vanishing on C and equal to $+\infty$ outside C. Then problem (2) is equivalent to

$$\min_W \|\tilde{W}\|_1 + i_C(\tilde{W}). \tag{3}$$

This new problem is readily equivalent to the previous one as $\|\tilde{W}\|_1 = \|W\|_1 + \|\zeta\|_1$. (Note that the cost has been re-scaled by a factor λ.) Problem (3) involves two convex functions the proximity operators of which can be efficiently computed (see, *e.g.*, Lemma 7), which allows a Douglas-Rachford scheme to be used .

3.2. Douglas-Rachford splitting

The method was initially proposed by [20] to solve matrix equations and used in signal and image processing [7, 10, 11, 39, 41]. Problem (3) amounts to minimizing the sum of two convex functions $F(\tilde{W}) + G(\tilde{W})$, and the (constant-step) Douglas-Rachford scheme is the following [12]: Fix $\varepsilon \in]0, 1[$, $\tau > 0$, $\gamma \in [\varepsilon, 2 - \varepsilon]$, $V_0 \in \mathbb{R}^{(d+m)\times k}$, and define

$$\begin{aligned} \tilde{W}_n &:= prox_{\tau F}(V_n), && (4)\\ V_{n+1} &:= V_n + \gamma(prox_{\tau G}(2\tilde{W}_n - V_n) - \tilde{W}_n). && (5) \end{aligned}$$

Theorem
For $\varepsilon \in]0, 1[$, $\tau > 0$ and $\gamma \in [\varepsilon, 2 - \varepsilon]$, every sequence generated according to (4-5) converges to a solution to Problem 3, (see [12]).

We recall [12, 31] that the proximity operator "prox" $\hat{x}$ at point $\bar{x}$ of a lower semi-continuous convex function $x \mapsto f(x)$, with parameter τ, is the unique minimizer of:

$$\hat{x} = prox_{\tau f}(\bar{x}) := \arg\min_x f(x) + \frac{\|x - \bar{x}\|^2}{2\tau}, \quad (6)$$

In our case, F is the ℓ_1 norm, so $prox_{\tau F}$ is given by soft thresholding (parameterized by τ). Indeed, one can separate the variables and use the prox of the scaled absolute value dimension one:

$$\begin{aligned} soft(x, \tau) &= x + \tau \text{ if } x < -\tau,\\ &= 0 \text{ if } x \in [-\tau, \tau],\\ &= x - \tau \text{ otherwise.} \end{aligned}$$

Since G is an indicator function, whatever τ the associated prox simply is the projection operator on the affine subspace C of $(d+m)\times k$ matrices. For the sake of completeness, we recall the associated expression of this projection.

Lemma
Let A be an $m \times n$ matrix of rank $m < n$, and let b be a vector in $\mathbb{R}^m$. The orthogonal projection of $z \in \mathbb{R}^n$ on the affine subspace $\{x \in \mathbb{R}^n \mid Ax = b\}$ is

$$proj(z, A, b) = z - A^T(AA^T)^{-1}(Az - b). \quad (7)$$

This Lemma can readily be applied in $\mathbb{R}^{(d+m)\times k}$ to obtain the prox of G, independently of τ, and (4-5) can be translated into Algorithm 1.

Algorithm 1 Supervised classification Douglas-Rachford algorithm. $\tau > 0$ and $\gamma \in [\varepsilon, 2 - \varepsilon]$. The operators *soft* and *proj* denote the soft thresholding and projection operators, respectively.

```
Input: $X$, $\lambda$, $Y$, $\gamma$, $\tau$, $V$, $N$
$\tilde{X} \leftarrow [X\ \lambda I_m]$
for $n = 0, \ldots, N$ do
    $\tilde{W} \leftarrow soft(V, \tau)$
    $V \leftarrow V + \gamma(proj(2\tilde{W} - V, \tilde{X}, Y) - \tilde{W})$
end for
$(W, \zeta) \leftarrow \tilde{W}$
Output: $W$
```

3.3. Feature selection and scalability.

In applications, particularly in biological ones, the issue of feature selection may be even more important than classification itself. This selection is achieved by means of the ℓ_1 sparsity inducing penalty in Problem 2 cost. A feature $i \in \{1, \dots, m\}$ is selected if the corresponding line in the matrix of weights W is not zero ($\|W(i,:)\| \neq 0$). We precompute $X^T(XX^T)^{-1}$ only once. Note that the matrix (XX^T) is a small $(m \times m)$ matrix, so we can use fast adapted algorithm to compute the inverse. The complexity of the resulting iterations is $O(d \times k \times d)$ for the proximal part, plus $O(d \times k)$ for the projection.

3.4. Classification using medoid

In order to build a robust classifier, we compute a medoid for each cluster. In contrast to the k-means algorithm, k-medoids chooses actual data points as centers. For the j-th cluster, a medoid μ_j is any member of the class minimizing the average dissimilarity inside the class in the projected space. In general, the k-medoids problem is NP-hard to solve exactly. Heuristic solutions such as PAM (Partitioning Around Medoids) or Voronoi iteration methods have been proposed [38]. Let us define $\mu \in \mathbb{R}^{m \times k}$, the medoid matrix of the clusters. Computing medoids minimization can be reformulated as minimizing the following ℓ_1 norm:

$$\min_{\mu} \|Y\mu - XW\|_1. \tag{8}$$

Then, a new query x (a dimension d row vector) is classified according to the following rule: It belongs to the (supposedly unique) class $\bar{j}$ such that

$$\bar{j} = \min_{j=1\cdots k} \|xW - \mu_j\|_1. \tag{9}$$

The cost of medoid computation is expected to be $O(m \times k)$ in average [3].

3.5. Alternating Minimization

Let now define the following global problem

$$\min_{W,\mu} \|Y\mu - XW\|_1 + \lambda \|W\|_1. \tag{10}$$

We propose an alternating (or Gauss-Seidel) scheme to solve problem (10). Given the medoid matrix μ, the first subproblem in W is solved by using our primal Douglas-Rachford algorithm. Conversely, given the matrix of weights W, the second subproblem in μ is solved by medoid computation on the projected data as explained in the previous section. Algorithm 2 summarizes the resulting alternating minimization.

Algorithm 2 Alternating Douglas-Rachford splitting (ADRS) algorithm. $\tau > 0$ and $\gamma \in [\varepsilon, 2-\varepsilon]$. The operators *soft* and *proj* denote the soft thresholding and projection operators, respectively.

```
Input: $X$, $\lambda$, $Y$, $\gamma$, $\tau$, $V$, $N$,$L$
$\tilde{X} \leftarrow [X \;\; \lambda I_m]$
for $l = 0, \dots, L$ do
    for $n = 0, \dots, N$ do
        $\tilde{W} \leftarrow soft(V, \tau)$
        $V \leftarrow V + \gamma(proj(2\tilde{W} - V, \tilde{X}, Y) - \tilde{W})$
    end for
    $(W, \zeta) \leftarrow \tilde{W}$
    $\mu \leftarrow medoids(Y, XW)$
end for
Output: $W, \mu$
```

Proposition As each step of the alternating minimization scheme decreases the norm $\|Y - XW\|_1$, which is nonnegative, the following readily holds. The norm $\|Y - XW\|_1$ converges as the number of iterates L in Algorithm 2 goes to infinity.
Note that the criterion 8 is convex in (μ,W), so a primal dual method with convergence guarantees was proposed recently [5, 9].

4. Application to real datasets

Experimental settings. We compare the labels obtained from our classification with the true labels to compute the MCE (misclassification error, *i.e.* the number of misclassified observations divided by the number of observations) on the test set. We compare our ADRS method with four state of the art methods: the filter method using the Ttest (we rank the features according to their p-values using a PLS classifier (partial least squares) [45]), PLS with a sequential selection of features [25], and the Frobenius norm criterion using the classical forward-backward algorithm [12]. The different algorithms are evaluated and compared using three public datasets (Ovarian, FaceDrive and Cancer RNA-Seq) which are available at `https://archive.ics.uci.edu/ml/datasets`. In our experiments, we set $\gamma = 1$ and $N = 40$. Results are reported after a 4-fold cross-validation.

Dataset: Ovarian [26]. .
The data available on UCI data base were obtained from two sources: The National Cancer Institute (NCI) and the Eastern Virginia Medical School (EVMS). The samples include patients with cancer (ovarian or prostate cancer). Healthy or control patients form a set of 216 samples with $15,000$ features.

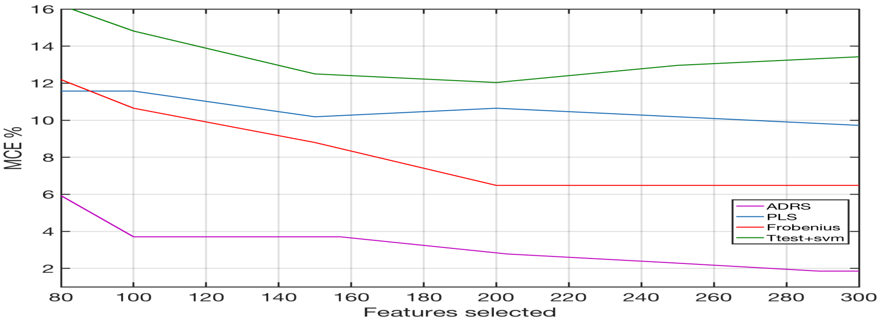

Figure 1. Ovarian dataset. The MCE versus the number of selected features selected from Ovarian shows that MCE decreases with the number of genes for our method, Frobenius and PLS. For 100 selected features, ADRS outperforms Ttest by 11.1% , Frobenius by 7% and PLS by 7.9%)

Dataset: FaceDrive [16]. .
The FaceDrive dataset contains image sequences of subjects while driving in real scenarios. It is composed of 606 samples with 4800 features each, acquired over different days from 4 drivers with several facial features like glasses and beards. A set of labels assigned each image into 3 possible clusters are given. Note that this dataset is highly unbalanced.

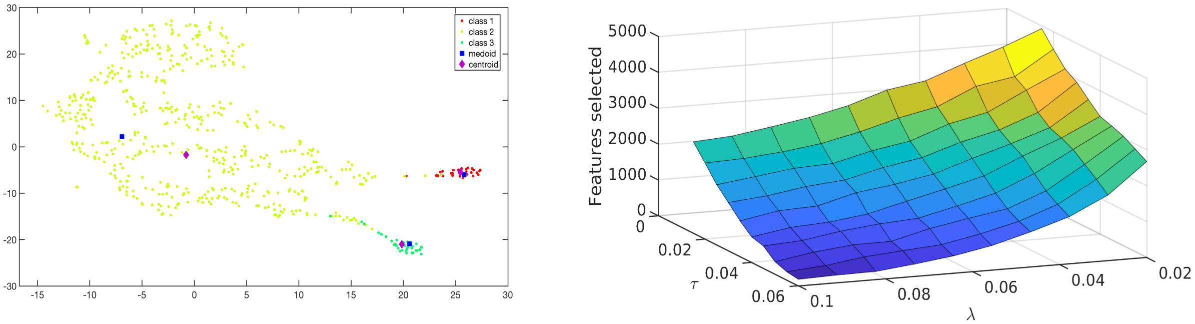

FIGURE 2. Left: The t-SNE [43] of the data in the projected space Xw shows that the distance between centroids and medoids depends on the clusters. Right: Features selected as a function of τ and λ

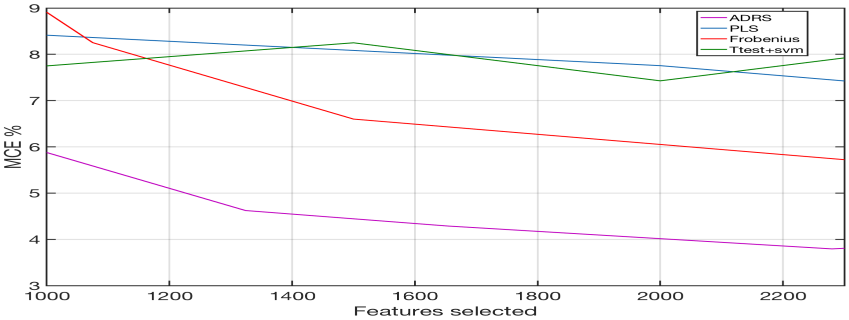

FIGURE 3. Facedrive dataset. The MCE and the number of selected features selected from Facedrive shows that 1500 features are necessary to obtain the best MCE. For 1500 selected features, Our method outperform Frobenius by 2.3%, and PLS by 3.6%)

Dataset: Gene expression cancer RNA-Seq Data Set. . This data is a subset of the RNA-Seq (HiSeq) PANCAN dataset, of patients with different types of tumor. This dataset is available at https://archive.ics.uci.edu/. It is composed of 801 cells with 20531 genes and $k = 5$ clusters.

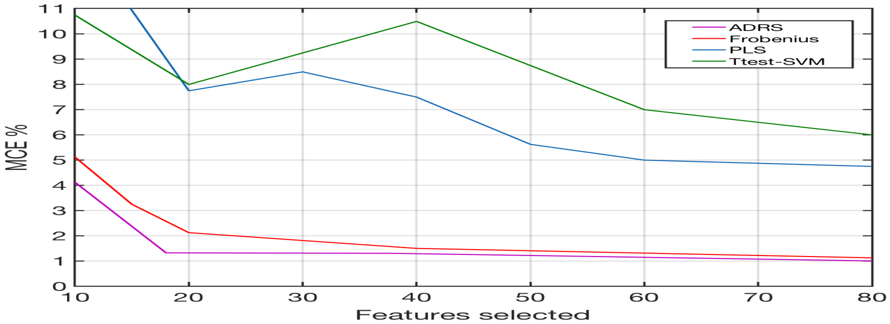

FIGURE 4. Cancer dataset. The MCE versus the number of selected features selected from the Cancer dataset shows that only 20 features are necessary to our method to obtain an accurate MCE. For 20 genes, our method outperforms the Frobenius minimization method by 0.8% , and PLS by 6.4%.

4.1. **Computational time**

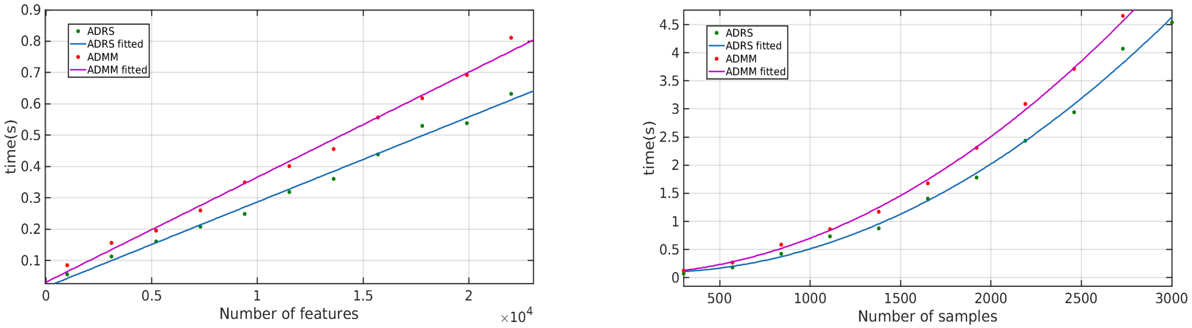

Figure 5. Time (using an intel Core i7 processor) Left: function of the number of features Right function of the number of sample

The only reasonable contender would be the popular ADMM algorithm first proposed by and Gabay & Mercier in 1976 [24] and extensively developed in [6, 12, 35]. Fig 5 shows computational as function of features d (left) and samples m (right). Computational times seem linear as a function of the features d and quadratic as a function of samples for both algorithms. Better performance is observed for ADRS.

5. Conclusion

In this paper, we propose to minimize the ℓ_1 norm related to the data term with an ℓ_1 regularization. Our contribution is twofold. First, we provide a new primal Douglas-Rachford splitting algorithm adapted to a high dimensional dataset. Second, we propose a new efficient classification algorithm using medoid and alternating minimization. Experiments on biological datasets and a computer vision dataset show that our method significantly improves the results of methods using a quadratic loss function.

The authors thank internship Yuxiang Zhou for processing simulations and Professor Jean-Baptiste Caillau for fruitful discussion.

6. Appendix: Minimize Frobenius norm using a gradient-projection splitting method

The criterion is given by:

$$\frac{1}{2}\|Y - XW\|_F^2 + \lambda\|W\|_1 \to \min$$

To solve this problem, we use a gradient-projection method. It belongs to the class of splitting methods ([12, 13, 30, 32, 40]). We use the following forward-backward scheme to generate a sequence of iterates. For any fixed step $\gamma \in (0, 2/\sigma^2_{\max}(X))$, the forward-backward scheme applied to the above criterion converges towards a solution.

References

[1] C. Aggarwal. On k-anonymity and the curse of dimensionality. *Proceedings of the 31st VLDB Conference, Trondheim, Norway*, 2005.

[2] F. R. Bach and Z. Harchaoui. Diffrac: a discriminative and flexible framework for clustering. In J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 49–56. Curran Associates, Inc., 2008.

Algorithm 3 Forward-Backward splitting

```
Input: $X, Y, W_0, N, \gamma, \tau$
for $n = 0, \dots, N$ do
    $V \leftarrow W - \gamma X^T(XW - Y)$
    $W \leftarrow soft(V, \tau)$
end for
Output: $W$
where soft$(V, \tau)$ is the soft thresholding.
```

[3] V. Bagaria, N. Kamath, and T. Zhang. Medoids in almost linear time via multi-armed bandits. *arXiv preprint*, 2017.

[4] 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*, 65(17):4635–4643, 2017.

[5] M. Barlaud, A. Chambolle, and J.-B. Caillau. Classification and feature selection using a primal-dual method and projection on structured constraints. *International Conference on Pattern Recognition, Milan*, pages 6538–6545, 2021.

[6] 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*, 3:1–122, 2011.

[7] L. Briceño-Arias, G. Chierchia, E. Chouzenoux, and J.-C. Pesquet. A Random Block-Coordinate Douglas-Rachford Splitting Method with Low Computational Complexity for Binary Logistic Regression. *Computational Optimization and Applications*, 72(3):707–726, 2019.

[8] E. J. Candès, M. B. Wakin, and S. P. Boyd. Enhancing sparsity by reweighted $\ell 1$ minimization. *Journal of Fourier analysis and applications*, 14(5-6):877–905, 2008.

[9] A. Chambolle and T. Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of Mathematical Imaging and Vision*, 40(1):120–145, May 2011.

[10] C. Chaux, J.-C. Pesquet, and N. Pustelnik. Nested iterative algorithms for convex constrained image recovery problems. *SIAM*, pages 730–762, 2009.

[11] J.-C. Combettes, P.L.and Pesquet. A douglas-rachford splitting approach to nonsmooth convex variational signal recovery. *IEEE J. Selected Topics Signal Process.*, pages 564–574, 2007.

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

[13] P. L. Combettes and V. R. Wajs. Signal recovery by proximal forward-backward splitting. *Multiscale Modeling & Simulation*, 4(4):1168–1200, 2005.

[14] F. de la Torre and T. Kanade. Discriminative cluster analysis. *ICML 06 Proceedings of the 23rd international conference on Machine learning, Pittsburgh, Pennsylvania, USA*, 2006.

[15] K. Degraux, G. Peyré, J. Fadili, and L. Jacques. Sparse support recovery with non-smooth loss functions. In *Advances in Neural Information Processing Systems*, volume 29, pages 4269–4277. Curran Associates, Inc., 2016.

[16] K. Diaz-Chito, A. Hernandez-Sabatie, and A. M. Lopez. A reduced feature set for driver head pose estimation, applied soft computing. *ISSN 1568-4946*, pages 98–107, 2016.

[17] C. Ding and T. Li. Adaptive dimension reduction using discriminant analysis and k-means clustering. In *Proceedings of the 24th International Conference on Machine Learning*, ICML '07, pages 521–528, 2007. ISBN 978-1-59593-793-3.

[18] D. L. Donoho and M. Elad. Optimally sparse representation in general (nonorthogonal) dictionaries via $\ell 1$ minimization. *Proceedings of the National Academy of Sciences*, 100(5):2197–2202, 2003.

[19] D. L. Donoho and B. F. Logan. Signal recovery and the large sieve. *SIAM Journal on Applied Mathematics*, 52(2): 577–591, 1992.

[20] J. Douglas and H. H. Rachford. On the numerical solution of heat conduction problems in two or three space variables. *Trans. Amer. Math. Soc.*, pages 421–439, 1956.

[21] N. Flammarion, B. Palaniappan, and F. R. Bach. Robust discriminative clustering with sparse regularizers. *Journal of Machine Learning Research,18(80):1-50*, 2017.

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

[23] 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*, 16(10):

906–914, 2000.

[24] D. Gabay and B. Mercier. A dual algorithm for the solution of nonlinear variational problems via finite element approximation. *Computers & Mathematics with Applications*, 2(1):17–40, 1976.

[25] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. Gene selection for cancer classification using support vector machines. *Machine learning*, 46(1-3):389–422, 2002.

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

[27] P. J. Huber. Robust statistics. 1981.

[28] 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)*, pages 353–360, 2009.

[29] S. Lambert-Lacroix and L. Zwald. Robust Regression through the Huber's criterion and adaptive lasso penalty. *Electronic journal of statistics* , 5:1015–1053, 2011.

[30] P.-L. Lions and B. Mercier. Splitting algorithms for the sum of two nonlinear operators. *SIAM Journal on Numerical Analysis*, 16(6):964–979, 1979.

[31] J. Moreau. Propriétés des applications "prox". *Comptes Rendus Acad Sciences Paris*, 256:1069–1071, 1963.

[32] 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*, pages 418–433. Springer, 2010.

[33] A. Y. Ng. Feature selection, l 1 vs. l 2 regularization, and rotational invariance. In *Proceedings of the twenty-first international conference on Machine learning*, page 78. ACM, 2004.

[34] F. Nie, H. Huang, X. Cai, and C. H. Ding. Efficient and robust feature selection via joint l2,1-norms minimization. In J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23*, pages 1813–1821. Curran Associates, Inc., 2010.

[35] D. O'Connor and L. Vandenberghe. Primal-dual decomposition by operator splitting and applications to image deblurring. *SIAM*, 7(3):1724–1754, 2014.

[36] M. Radovanovic, A. Nanopoulos, and M. Ivanovic. Hubs in space : Popular nearest neighbors in high-dimensional data. *Journal of Machine Learning Research. 11: 2487-2531.*, 2011.

[37] R. Rosipal and N. Kramer. Overview and recent advances in partial least squares. *Berlin, Germany: Springer-Verlag*, pages 34–51, 2006.

[38] E. Schubert and P. J. Rousseeuw. Faster k-medoids clustering: Improving the pam, clara, and clarans algorithms. *Lecture Notes in Computer Science*, page 171–187, 2019.

[39] S. Setzer. Split Bregman algorithm, Douglas-Rachford splitting and frame shrinkage. *Lecture Notes in Comput. Sci.*, page 464–476, 2009.

[40] S. Sra, S. Nowozin, and S. J. Wright. *Optimization for Machine Learning.* MIT Press, 2012.

[41] T. Steidl, G.and Teuber. Removing multiplicative noise by Douglas-Rachford splitting methods. *J. Math. Imaging*, page 168–184, 2010.

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

[43] L. van der Maaten and G. Hinton. Visualizing high-dimensional data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605, Nov 2008.

[44] C. Wei-Chien. On using principal components before separating a mixture of two multivariate normal distributions. *Journal of the Royal Statistical Society*, 32(3), 1983.

[45] S. Wold, M. Sjostrom, and L. Eriksson. PLS-regression: a basic tool of chemometrics. *Elsevier volume 58, issue 2*, pages 109–130, 2001.

[46] H. Zou, T. Hastie, and R. Tibshirani. Sparse principal component analysis. *Journal of computational and graphical statistics*, 15(2):265–286, 2006.