Robust classification with feature selection using alternating minimization and Douglas-Rachford splitting method

Yuxiang Zhou, Jean-Baptiste Caillau, Marc Antonini, Michel Barlaud

Robust classification with feature selection using alternating minimization and Douglas-Rachford splitting method

Yuxiang Zhou, Jean-Baptiste Caillau, Marc Antonini and Michel Barlaud
Michel Barlaud, Yuxiang Zhou and Marc Antonini
are with I3S, Univ. Côte d'Azur & CNRS, F-06900 Sophia Antipolis.
J.-B. Caillau is with LJAD, Univ. Côte d'Azur & CNRS/INRIA, F-06108 Nice.
Contact: barlaud@i3s.unice.fr

Abstract

This paper deals with supervised classification and feature selection. A classical approach is to project data on a low dimensional space with a strict control on sparsity. This results in an optimization problem minimizing the within sum of squares in the clusters (Frobenius norm) with an ℓ_1 penalty in order to promote sparsity. It is well known though that the Frobenius norm is not robust to outliers. In this paper, we propose an alternative approach with an ℓ_1 norm minimization both for the constraint and the loss function. Since the ℓ_1 criterion is only convex and not gradient Lipschitz, we advocate the use a Douglas-Rachford approach. 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. We also provide an efficient classifier in the projected space based on medoid modeling. The resulting algorithm, based on alternating minimization and primal Douglas-Rachford splitting, is coined ADRS. Experiments on biological data sets and computer vision dataset show that our method significantly improves the results obtained with a quadratic loss function.

1 Introduction

This paper deals with supervised classification with feature selection in high dimensional spaces. In this paper we consider methods where feature selection is embedded in a classification process [30]. However, classification in 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, 28]. In order to overcome this issue, the main idea of the following methods is to project data in a low dimensional space. A popular approach for high-dimensional data is to perform *Principal Component Analysis* (PCA) prior to classification. This approach is however not relevant in general [36]. Partial least squares method closely related to principal component regression is designed to cope with this issue with high dimensional correlated features [37, 29]. An alternative approach is to perform dimension reduction by means of *Linear Discriminant Analysis* (LDA) [12, 14]. The authors of [2] and [18] 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 [26]. A popular approach for selecting sparse features in supervised classification or regression is the *Least Absolute Shrinkage and Selection Operator* (LASSO) formulation [34, 19, 22, 25, 38]. The LASSO formulation uses the ℓ_1 norm [5, 6, 15, 16] as an added penalty term instead of an ℓ_0 term. However, an issue is that using an ℓ_2 norm for the data is not robust to outliers. In this paper, we cope with this problem by minimizing an ℓ_1 norm both on the penalty term and the loss function. In this case, the criterion is convex but

Preprint. Work in progress.

not gradient Lipschitz. We propose so to use a Douglas-Rachford splitting method. This splitting was successfully used in signal processing [8, 9, 17, 7, 31, 33, 4]. However, for classification, we cannot apply straightforwardly Douglas-Rachford since the proximal operator for the affine transform $Y - XW$ involved in the criterion is not available. We take advantage of the particular form of the original cost: 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.

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, 18] . 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 to be able to separate them. Besides, it is well known that the quadratic Frobenius loss criterion is not robust to outliers, 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:

Problem 1

$$\min_W \|Y - XW\|_1 + \lambda \|W\|_1.$$

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 An equivalent formulation

The cost in Problem 1 is the sum of two ℓ_1 norms, which suggests to use a splitting algorithm [10]. 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 [10] 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 $\mathbf{R}^{m \times k}$ and to minimize

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

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 $\widetilde{W} := (W, \zeta) \in \mathbf{R}^{(d+m) \times k}$. Let C be the affine subset of $(d+m) \times k$ matrices such that $\widetilde{X}\widetilde{W} = Y$ where

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

The problem can be recast as

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

under the affine constraint

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

Let moreover i_C be the indicator function of the set C, vanishing on C and equal to $+\infty$ outside C. Then problem 1 is equivalent to

Problem 2

$$\min_W \|\widetilde{W}\|_1 + i_C(\widetilde{W}).$$

This new problem is readily equivalent to the previous one as $\|\widetilde{W}\|_1 = \|W\|_1 + \|\zeta\|_1$. (Note that the cost has been rescaled by a factor λ.) Problem 2 involves two convex functions whose proximity operators can be efficiently computed (see, *e.g.*, Lemma 1), which allows to use a Douglas-Rachford scheme.

Remark 1 *Our approach can be seen as a special case of the alternating-direction method of multipliers (ADMM, see [10]). Indeed, Problem 1 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 ADMM 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. See also [7] for a similar approach. In our case, we take advantage of the fact that the two functions f and g are simple convex functions (ℓ_1 norms) whose sum is again a known convex function (yet another ℓ_1 norm). This eliminates one function and allows 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 is easily computed (see, e.g.*, Lemma 1).*

2.3 Douglas-Rachford splitting

The method was initially proposed by [17] for solving matrix equations and used in signal and image processing [8, 9, 31, 33]. Problem 2 amounts to minimizing the sum of two convex functions $F(\widetilde{W}) + G(\widetilde{W})$, and the (constant-step) Douglas-Rachford scheme is the following [10]: Fix $\varepsilon > 0$, $\tau > 0$, $\gamma \in (\varepsilon, 2 - \varepsilon)$, $V_0 \in \mathbf{R}^{(d+m)\times k}$, and define

$$\widetilde{W}_n := \mathrm{prox}_{\tau F}(V_n), \tag{2}$$

$$V_{n+1} := V_n + \gamma(\mathrm{prox}_{\tau G}(2\widetilde{W}_n - V_n) - \widetilde{W}_n). \tag{3}$$

Theorem 1 *[10] For $\varepsilon \in (0, 1)$, $\tau > 0$ and $\gamma \in [\varepsilon, 2 - \varepsilon]$, every sequence generated according to (2-3) converges to a solution to Problem 2.*

We recall that, given a lower semicontinuous convex function f from a vector space to $\mathbf{R} \cup \{+\infty\}$, the proximity operator prox_f of f maps a given x to the unique minimizer of $f(y) + (1/2)\|x - y\|^2$. In our case, F is the ℓ_1 norm, so $\mathrm{prox}_{\tau F}$ 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:

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

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

$$\mathrm{proj}(z, A, b) = z - A^T(AA^T)^{-1}(Az - b). \tag{4}$$

Proof. The projection $\bar{z}$ is characterized by the fact that $z - \bar{z}$ belongs to $(\ker A)^\perp = \mathrm{Im}\, A^T$, plus the fact that $A\bar{z} = b$. So there is some $\bar{y}$ in $\mathbf{R}^m$ such that $\bar{z} = z - A^T\bar{y}$, and

$$y = Az - AA^T\bar{y}.$$

Since AA^T is invertible, $\bar{y} = (AA^T)^{-1}(Az - y)$ which gives the desired expression. □

The lemma can readily be applied in $\mathbf{R}^{(d+m)\times k}$ to obtain the prox of G, independently of τ, and (2-3) can be translated into Algorithm 1.

Algorithm 1 Supervised classification Douglas-Rachford algorithm. The operators soft and proj denote the soft thresholding and projection operators, respectively.

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

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 1 cost. A feature $i \in \{1, \ldots, m\}$ is then selected if the corresponding line in the matrix of weights W is not zero ($\|W(i,:)\| \neq 0$). The set of nonzero columns is interpreted as the signature of the corresponding cluster. In order to evaluate this signature, we compute the projected matrix XW. Each block shows the correlation between the data and the signature. We precompute $X^T(XX^T)^{-1})$ only once. Note that 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.

2.4 Classification using medoid

In order to build a robust classifier, we compute a medoid for each cluster. 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: $\mu_j := (XW)(\bar{i}, :)$ where

$$\bar{i} \in \arg \min_{i \text{ s.t. } y_{ij}=1} \sum_{\ell \text{ s.t. } y_{\ell j}=1,\ \ell \neq i} \|(XW)(i,:) - (XW)(\ell,:)\|_1.$$

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:

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

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

$$\bar{j} \in \arg \min_{j=1,k} \|xW - \mu_j\|_1.$$

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

2.5 Alternating Minimization

In this section, we iterate the previous procedure. Let define now the within cluster sum of absolute difference criterion:

Problem 3

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

We propose an alternating (or Gauss-Seidel) scheme. 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 Supervised classification with alternating minimization and Douglas-Rachford splitting algorithm.

Input: X, $\mu = I_m$, λ, Y, $\gamma = 1$, τ, V, N, L
$\widetilde{X} \leftarrow [X\ \lambda I_m]$
for $l = 0, \ldots, L$ **do**
  for $n = 0, \ldots, N$ **do**
    $\widetilde{W} \leftarrow \mathrm{soft}(V, \tau)$
    $V \leftarrow V + \gamma(\mathrm{proj}(2\widetilde{W} - V, \widetilde{X}, Y\mu) - \widetilde{W})$
  end for
  $(W, \zeta) \leftarrow \widetilde{W}$
  $\mu \leftarrow \texttt{medoids}(Y, XW)$
end for
Output: W, μ

Proposition 1 *As each step of the alternating minimization scheme decreases the $\ell_1 norm$ $\|Y\mu - XW\|_1$, which is nonnegative, the following readily holds.*
The $\ell_1 norm$ $\|Y\mu - XW\|_1$ converges as the number of iterates L in Algorithm 2 goes to infinity. This property is illustrated in the next section on real data.

3 Application to real datasets

3.1 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 provide TSNE results for visual evaluation [35]. In order to evaluate the signature, we plot the heat map of matrix XW in order to visualize the correlation between the data and the signature. We compare our ADRS method with four state of the art methods: Filter method using Ttest (we rank the features according to their p-values using a PLS classifier (Partial least squares) [37]), PLS with a sequential selection of features [20], and the Frobenius norm criterion (1) using the classical forward-backward algorithm. We compare algorithms on three public datasets which are available at `https://archive.ics.uci.edu/ml/datasets`. In our experiments, we set $\gamma = 1$ and $N = 10$ (see Figure 1 for preliminary observations on convergence). Results are reported after a 4-fold cross-validation.

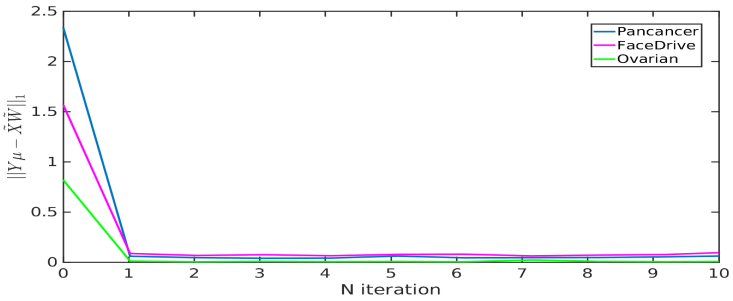

Figure 1: Convergence of our primal Douglas-Rachford algorithm: The decay of $\|Y\mu - \widetilde{X}\widetilde{W}\|_1$ versus the number of iterations N emphasizes the fast convergence.

3.2 Dataset: Ovarian [21]

The data available on UCI data base were obtained from two sources: The National Cancer Institute (NCI) and the Eastern Virginia Medical School (EVMS). All the data consist of mass-spectra obtained with the SELDI technique. The samples include patients with cancer (ovarian or prostate cancer). Healthy or control patients form a set of 216 samples with 15000 features. Applying classical Ttest, there are 40% p-values smaller than 0.05, thus 6000 features that have strong potential discrimination.

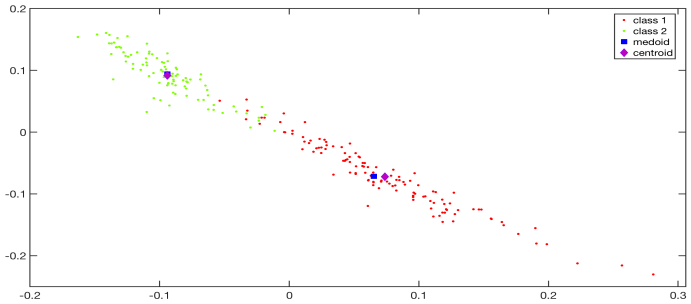

Figure 2: Ovarian: The data in the projected space XW show that the centroids and the medoids are very close.

Table 1: MCE on Ovarian dataset (216 samples, 15000 features, $k = 2$ clusters which have 121 and 95 samples, respectively): For 100 selected features, ADRS outperforms Ttest by 11.1% , Frobenius by 7% and PLS by 7.9%).

Ovarian	Ttest	Frobenius	PLS	ADRS
MCE % 100 features	14.82	10.65	11.57	**3.70**
MCE % 200 features	12.03	6.48	10.65	**2.78**
MCE % 300 features	13.43	6.48	9.72	**1.85**

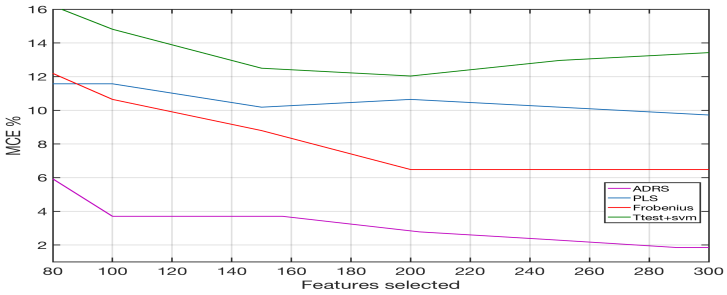

Figure 3: The MCE versus the number of selected features selected on Ovarian shows that MCE decreases with the number of genes for ADRS, Frobenius and PLS. Note that Ttest assume that data are independent, which is not the case for proteomics dataset.

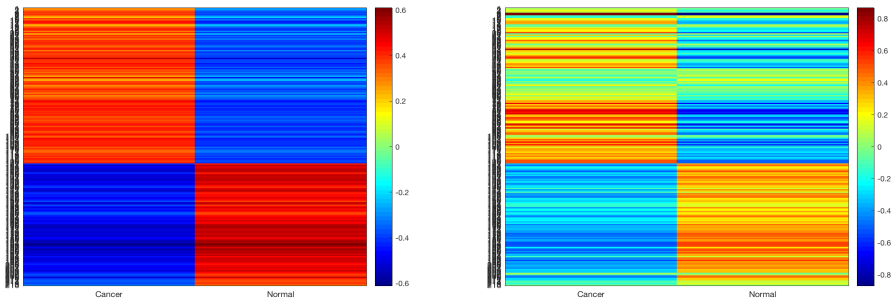

Figure 4: The heatmap of correlation between data and signature, left with ADRS, right with a quadratic (Frobenius) cost.

Table 2: This table shows the mean of the correlation on the diagonal of the previous heatmap. The signature obtains with ADRS readily improves the one obtained using a Frobenius norm.

Ovarian	Cancer	Normal
ADRS	0.43	0.55
Frobenius	0.12	0.16

3.3 Dataset:FaceDrive [13].

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 beard. A set of labels assigning each image into 3 possible clusters are given. Note that this dataset is highly unbalanced.

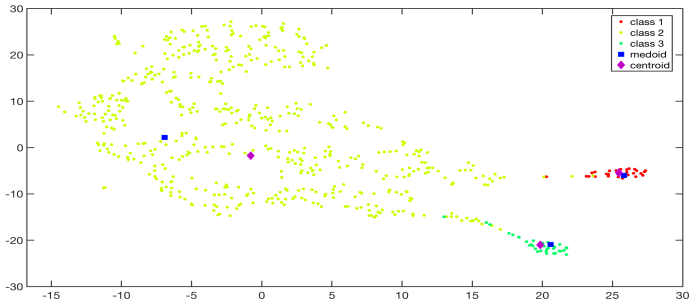

Figure 5: The Tsne of the the data in the projected space Xw shows that the centroids and the medoids are very close in the minority class while they are far in the majority class.

Table 3: MCE on Facedrive (606 samples, 6400 features, $k = 3$ clusters which have 27, 546 and 33 samples, respectively): For 1500 selected features, ADRS outperform Frobenius by 2.3%, and PLS by 3.6%).

Facedrive	Ttest+svm	Frobenius	PLS	ADRS
MCE % $1000 features$	7.75	8.89	8.41	**5.87**
MCE % $1500 features$	8.25	6.6	7.92	**4.29**
MCE % 2000 features	7.43	6.08	7.76	**4.01**

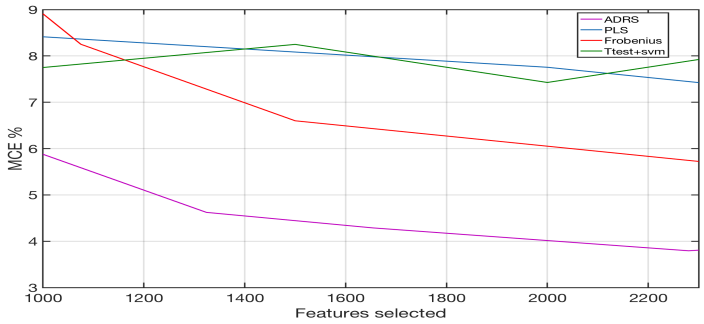

Figure 6: The MCE and the number of selected features selected on Facedrive shows that 1500 features are necessary to obtain the best MCE.

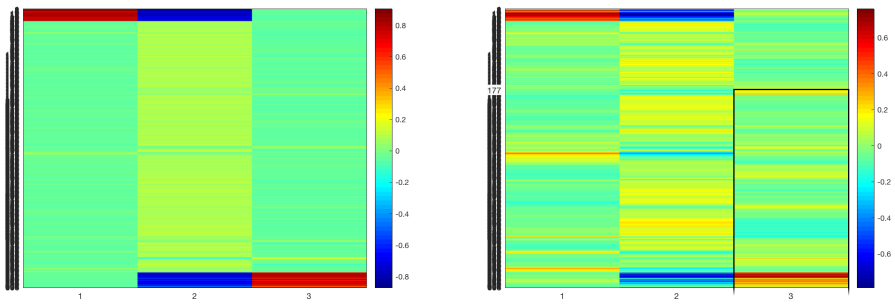

Figure 7: heatmap of correlation, Left using ADRS , Right using Frobenius

Table 4: This table shows the mean of the correlation on the diagonal of the previous heatmap. Again, ADRS signature improves the signature obtained with a Frobenius cost.

FaceDrive	Cluster1	Cluster2	Cluster3
ADRS	0.42	0.04	0.32
Frobenius	0.17	0.02	0.10

3.4 Dataset: Gene expression cancer RNA-Seq Data Set.

This data is a subset of the RNA-Seq (HiSeq) PANCAN dataset, of patients having different types of tumor: BRCA, KIRC, COAD, LUAD and PRAD. It is available at `https://archive.ics.uci.edu/ml/datasets` and `https://www.synapse.org`. This dataset is composed of 801 cells with 20531 genes. There are $k = 5$ clusters which have 300, 78, 146, 141, and 136 samples, respectively.

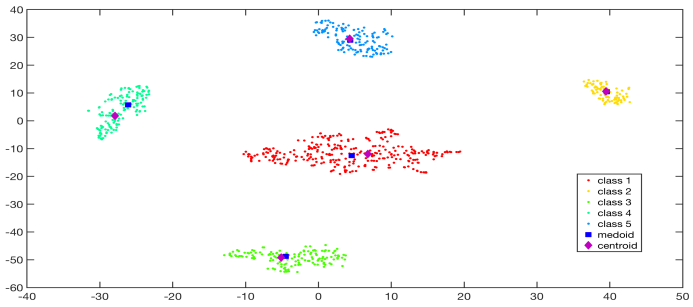

Figure 8: The Tsne of the data in the projected space Xw shows that the distance between centroids and medoids depends on the clusters.

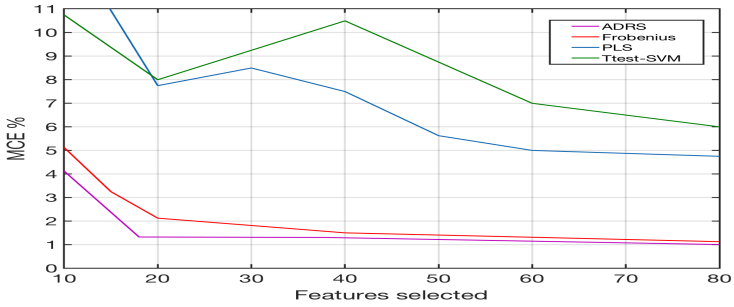

Figure 9: The MCE versus the number of selected features selected on Cancer dataset shows that only 20 features are necessary to ADRS to obtain an accurate MCE.

Table 5: MCE on Cancer dataset (801 cells and 20531 genes with $k = 5$ clusters): For 20 genes, ADRS outperforms Frobenius by 0.8% , and PLS by 6.4%.

Cancer	Ttest+svm	Frobenius	PLS	ADRS
MCE % 10 genes	10.74	5.12	14.11	**4.12**
MCE % 20 genes	7.99	2.11	7.74	**1.34**
MCE % 40 genes	10.49	1.5	7.49	**1.3**
MCE % 80 genes	5.99	1.1	4.74	**1.0**

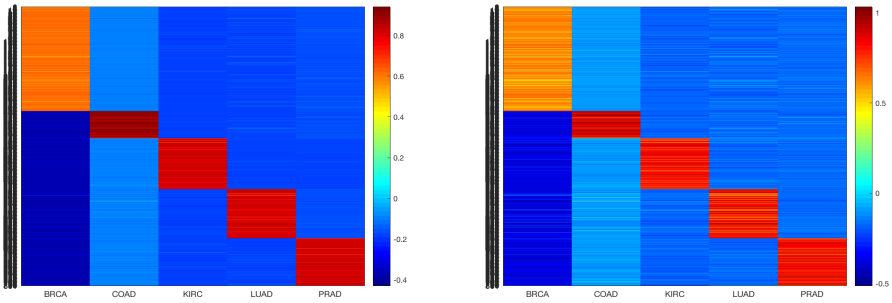

Figure 10: heatmap of correlation, Left using ADRS , Right using Frobenius

Table 6: This table shows the mean of the correlation on the diagonal of the previous heatmap, it shows that ADRS signature outperforms Frobenius signature.

Pancancer	BRCA	COAD	KIRC	LUAD	PRAD
ADRS	0.63	0.90	0.82	0.82	0.83
Frobenius	0.55	0.75	0.73	0.67	0.74

4 Conclusion

In this paper, we propose to minimize the ℓ_1 norm related to the data term with an ℓ_1 regularization. This problem can be solved by classical ADMM algorithm [10, 27, 7]. Our contribution is twofold. First, we provide a new primal Douglas-Rachford splitting algorithm adapted to the criterion. Second, we propose a new efficient classification algorithm (ADRS) 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.

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

The criterion is given by:

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

To solve this problem, we use a gradient-projection method. It belongs to the class of splitting methods ([10, 11, 23, 24, 32]) using separately the convexity properties of the Frobenius cost on one hand, and of the convexity of the set C on the other. 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.

Algorithm 3 Forward-Backward splitting

Input: $X, Y, W_0, N, \gamma, \tau$
for $n = 0, \ldots, N$ **do**
  $V \leftarrow W - \gamma X^T(XW - Y)$
  $W \leftarrow \text{prox}(V, \tau)$
end for
Output: W

where $\text{prox}(V, \tau)$ denotes the proximity operator of the ℓ_1 norm (soft thresholding) already define in section 2.

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. URL `http://papers.nips.cc/paper/3269-diffrac-a-discriminative-and-flexible-framework-for-clustering.pdf`.

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

[4] 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.

[5] E. J. Candès. The restricted isometry property and its implications for compressed sensing. *Comptes Rendus Acad Sciences Paris*, 346(1):589–592, 2008.

[6] 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.

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

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

[9] 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.

[10] 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.

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

[12] 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.

[13] 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.

[14] 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.

[15] 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.

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

[17] 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.

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

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

[20] 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.

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

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

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

[28] 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.*

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

[30] Y. Saeys1, I. Inza, and P. Larranaga. A review of feature selection techniques in bioinformatics. *Bioinformatics*, 2007.

[31] S. Setzer. Split bregman algorithm, douglas-rachford splitting and frame shrinkage. *Lecture Notes in Comput. Sci.*, page 464–476, 2009.

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

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

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

[35] L. J. P. Van der Maaten and G. E. Hinton. Visualizing high-dimensional data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

[36] 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.

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

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