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

Michel Barlaud,[1] **Antonin Chambolle**[2] **and Jean-Baptiste Caillau**[3]

(1) Université Côte d'Azur, CNRS, I3S (2) CEREMADE, CNRS & Paris-Dauphine PSL (3) Université Côte d'Azur, CNRS, Inria, LJAD

Abstract

This work 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 proposed 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 tailored constrained primal-dual method. The primal-dual framework is general enough to encompass the various robust losses and structured constraints and allows for 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.

Problem statement: Minimization of the ℓ_1 loss

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 $k \geq 2$ is the number of clusters. 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 $k \ll d$.

$$\min_{(W,\mu)} \|Y\mu - XW\|_1 + \frac{\rho}{2}\|I_k - \mu\|_F^2 \text{ s.t. } \|W\|_1 \leq \eta \tag{1}$$

where I_k denotes the order k identity matrix. An ℓ_2 regularization has been added in order to avoid the trivial solution $(W, \mu) = (0, 0)$. We use the algorithmic setting described in [2].

A primal dual algorithm

Problem (1) is rewritten in the form of the saddle point problem:

$$\min_{(W,\mu)} \sup_{Z_{i,j} \in [-1,1]} Z \cdot (Y\mu - XW) + \frac{\rho}{2}\|I_k - \mu\|_F^2 \text{ s.t. } \|W\|_1 \leq \eta \tag{2}$$

which can be solved by a primal-dual algorithm as studied in [2] (Algorithm 1)

Algorithm 1 Primal-dual algorithm, ℓ_1 loss.

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

The convergence condition on the step-sizes τ, τ_μ and σ are given in [1]). The drawback of the term $\|Y\mu - XW\|_1$ is that it enforces equality of the two matrices out of a sparse set. In order to soften this behaviour, we use the Huber function instead of the ℓ_1 norm. In the algorithm, it simply consists in replacing line 7 with the appropriate "prox", in this case we divide Z in the expression at line 7 with $1 + \sigma\epsilon$ for a small parameter $\epsilon > 0$ before truncating at -1 and 1.

Group LASSO constraint

Group LASSO was first introduced in [3]. 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

$$\|W\|_{2,1} := \sum_{i=1}^{d} \|w_i\|_2. \tag{3}$$

Projecting a matrix w on this ball is easy as it amount to project first the norms of the rows $(\|w_i\|)_i$ on the ℓ_1 ball of radius η to obtain radii (r_i) with $\sum_i r_i \leq \eta$, and then each row w_i of w on the ℓ_2 ball of radius r_i. In order to solve Problem (1) with this new constraint, we simply replace at line 4 of Algorithm 1 the projection onto the ℓ_1 ball with the appropriate modified projection.

Exclusive LASSO constraint

Exclusive sparsity or exclusive LASSO was first introduced in [4]. 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. So given a $d \times k$ matrix V, the projection on the corresponding balls consists in finding a matrix W which solves

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

Our approach is to introduce a Lagrange multiplier for the constraint and then compute it by a variant of Newton's method.

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

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

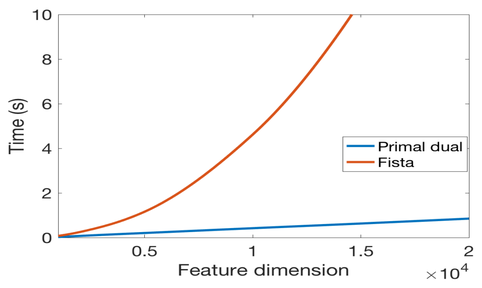
Results and comparison of methods

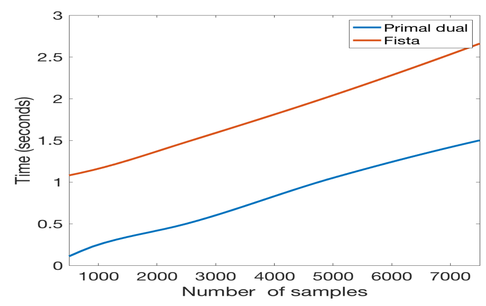
Ovarian proteomic dataset is available on UCI database consists of mass-spectra obtained with the SELDI technique. The dataset is composed of 216 samples, 15000 features and two clusters. Lung proteomic dataset were collected using unbiased liquid chromatography/mass spectrometry. The dataset is comprised of 1005 patients (469 among them with lung cancer and 536 control patients), and 2944 features and and two clusters. Zeiseil is a Single cell dataset composed of 3005 cells, 7364 genes and $k = 9$ clusters.

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 %

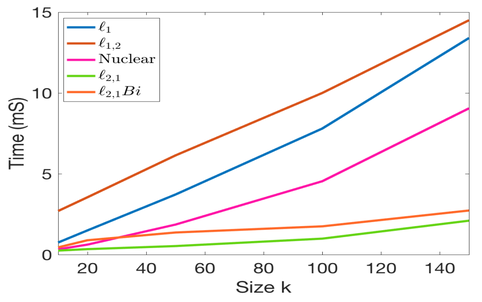
The table above shows the improvement in accuracy on all biological datasets when using Huber loss instead of ℓ_1 or Frobenius loss; ℓ_1 loss suffers from overfitting while Frobenius loss is not robust enough. Optimizing the matrix of centers (*v.s.* fixing $\mu = I$) also improves accuracy on the three datasets.

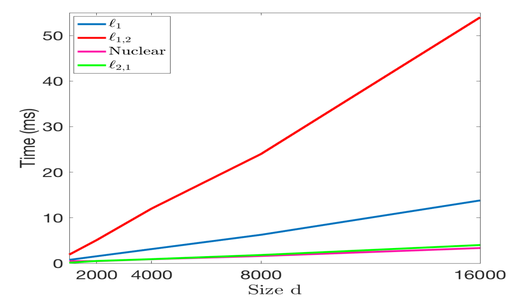
d	1000	3000	5000	10000	15000
Primal-dual	0.025	0.12	0.205	0.42	0.63
Fista	0.075	0.481	1.16	4.62	10.6
ADMM	0.65	4.52	58	NC	NC





The second table and the leftmost figure above shows that computational time as a function of the number of samples m is linear both for primal-dual and FISTA and that the computational time of ADMM is one or order of magnitude greater than the others because of the linear algebra involved. The figure on the rightshows computational time as a function of the number of features: 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)





The figure on the left 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. On the right, the cost of the projection on the $\ell_{1,2}$ ball is shown to grow linearly with d and k, and is slightly higher than for the projection on the ℓ_1 ball.

References

[1] Michel Barlaud, Antonin Chambolle, and Jean-Baptiste Caillau. Robust supervised classification and feature selection using a primal-dual method. *arXiv cs.LG/1902.01600*, 2019.

[2] Antonin Chambolle and Thomas Pock. On the ergodic convergence rates of a first-order primal-dual algorithm. *Math. Program.*, 159(1-2, Ser. A):253–287, 2016.

[3] Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006.

[4] Yang Zhou, Rong Jin, Steven Chuâ, and Hong Hoi. Exclusive lasso for multi-task feature selection. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 988–995, 2010.