HAL
open science

Robust supervised classification and feature selection using a primal-dual method

Michel Barlaud, Antonin Chambolle, Jean-Baptiste Caillau

To cite this version:

Michel Barlaud, Antonin Chambolle, Jean-Baptiste Caillau. Robust supervised classification and feature selection using a primal-dual method. 2019. hal-01992399v2

HAL Id: hal-01992399

https://inria.hal.science/hal-01992399v2

Preprint submitted on 14 Feb 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

arXiv:1902.01600v1 [cs.LG] 5 Feb 2019

Robust supervised classification and feature selection using a primal-dual method

Michel Barlaud [1] **Antonin Chambolle** [2] **Jean-Baptiste Caillau** [3]
(1) Université Côte d'Azur, CNRS, I3S, France
(2) École Polytechnique, CNRS, CMAP, France
(3) Université Côte d'Azur, CNRS, Inria, LJAD, France

Abstract

This paper deals with supervised classification and feature selection in high dimensional space. A classical approach is to project data on a low dimensional space and classify by minimizing an appropriate quadratic cost. A strict control on sparsity is moreover obtained by adding an ℓ_1 constraint, here on the matrix of weights used for projecting the data. It is well known that using a quadratic cost is not robust to outliers. We cope with this problem by using an ℓ_1 norm both for the constraint and for the loss function. Another second issue is that we optimize simultaneously the projection matrix and the centers used for classification. In this paper, we provide a novel tailored constrained primal-dual method to compute jointly selected features and classifiers. Extending our primal-dual method to other criteria is easy provided that efficient projection (on the dual ball for the loss data term) and prox (for the regularization term) algorithms are available. We illustrate such an extension in the case of a Frobenius norm for the loss term. We provide a convergence proof of our primal-dual method, and demonstrate its effectiveness on three datasets (one synthetic, two from biological data) on which we compare ℓ_1 and ℓ_2 costs.

Introduction

In this paper we consider methods where feature selection is embedded into a classification process, see (Furey et al., 2000; Guyon et al., 2002). However, classification in high dimension space suffers from the curse of dimensionality (Aggarwal, 2005; Radovanovic et al., 2010). In order to overcome this issue, the main idea of *Linear Discriminant Analysis* (LDA) (de la Torre & Kanade, 2006; Ding & Li, 2007) is to project data into a lower dimensional space. In parallel, sparse learning based methods have received a great attention in the last decade because of their high performance. The basic idea is to use a sparse regularizer which forces coefficients to be zero. To achieve feature selection, the *Least Absolute Shrinkage and Selection Operator* (LASSO) formulation (Tibshirani, 1996; Hastie et al., 2004; Ng, 2004; Friedman et al., 2010; Hastie et al., 2015) adds an ℓ_1 penalty term to the classification cost, which can be interpreted as convexifying an ℓ_0 penalty (Donoho & Elad, 2003; Donoho, 2006; Candès et al., 2008). However, an issue is that using the Frobenius norm $\|Y\mu - XW\|_F$ (that is the ℓ_2 norm of the vectorized matrix) for the data term is not robust to outliers. (In the previous expression, W is the projection matrix, μ the matrix of centers, and Y the binary matrix mapping each line to its class; see Section 1.) In order to overcome this issue, Nie *et al* proposed a feature selection based on $\ell_{2,1}$ norm minimization (Nie et al., 2010) on both loss function and the regularization. Note that $\ell_{2,1}$ norm regularization has strong connections with group LASSO methods (Zou & Hastie, 2005; Zou et al., 2006; Yuan & Lin; Li & Li, 2008; Jacob et al., 2009; Liu & Vemuri, 2012; Simon et al., 2013; Hastie et al., 2015; Li et al., 2016). In this paper, we propose a more drastic approach that uses an ℓ_1 norm both on the regularization term and on the loss function $\|Y\mu - XW\|_1$. In this case, the criterion is convex but not gradient Lipschitz. The basic idea is to use a splitting method (Lions & Mercier, 1979) together with proximal methods. Proximal methods were introduced in (Moreau, 1965) and have been intensively used in signal processing; see, *e.g.*, (Combettes & Wajs, 2005; Chaux et al., 2009; Mosci et al., 2010; Liu & Ye; Combettes & Pesquet, 2011; Chambolle & Pock, 2011; Boyd et al., 2011; Sra, 2012; O'Connor & Vandenberghe, 2014; Chambolle & Pock, 2016; Flammarion et al., 2017). The first issue is the computation of the proximal operator involving the affine transform $Y\mu - XW$ in the criterion. We tackle this point by dualizing the norm computation and use a primal-dual method. When one uses an ℓ_1 penalization to ensure sparsity, the computational time due to the treatment of the corresponding hyper-parameter has a worst case in $O(3^d)$, see (Mairal & Yu, 2012). We propose instead a constrained approach that takes advantage of an available ef-

ficient projection on the ℓ_1 ball (Condat, 2016; Duchi et al., 2008).

The paper is organized as follows. We first present our setting that combines dimension reduction, classification and feature selection. Then we propose a Lagrangian primal-dual scheme in Section 2. An efficient alternative is to replace the ℓ_1 penalization by a hard constraint: we provide in Section 3 an updated primal-dual scheme for this constrained formulation of the classification problem. Section 5 is devoted to analyze the convergence of this approach. In the last section, we give some experimental comparisons between ℓ_1 and Frobenius loss function. The tests involve three different bases: the first one is a synthetic dataset; the second base is a proteomics dataset on patients with ovarian or prostate cancer; the third and last database comes from single cell analysis.

1. Problem Statement

1.1. Projection of the data on a low dimensional space

Let X be the data $m \times d$ matrix made of m line samples $x_1, \dots, x_m$ 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. Projecting the data in lower dimension is crucial to be able to separate them accurately. Let $W \in \mathbb{R}^{d\times k}$ be the projection matrix, $k \ll d$. (Note that the dimension of the projection space is equal to the number of clusters.) One of the goals is to compute the projection matrix W.

1.2. Robust classification using ℓ_1 centers

In order to build a robust classifier for the projected data, XW, we compute an ℓ_1-center for each cluster. This is more robust to outliers than the outcome of the classical ℓ_2 modelling (Witten & Tibshirani, 2010). It is standard to define the ℓ_1-medoid μ_j of the j-th cluster as a vector that minimizes the average dissimilarity inside the class in the projected space: $\mu_j := (XW)(i^*, :)$ where

$$i^* = \arg \min_{i \text{ s.t. } y_{ij}=1} \sum_l \|(XW)(i,:) - (XW)(l,:)\|_1.$$

By analogy, we compute the matrix of centers, $\mu \in \mathbb{R}^{k\times k}$, by minimizing the following ℓ_1 norm:

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

2. Primal-dual scheme, Lagrangian formulation

We propose to minimize ℓ_1 loss cost with an ℓ_1 penalty term (a Lagrangian parameter λ is introduced) so as to promote sparsity and induce feature selection. So, given the matrix of labels, Y, and the matrix of data, X, we consider the following convex supervised classification problem where both μ and W are unknowns and I_k the identity matrix:

$$\min_{(W,\mu)} \quad \|Y\mu - XW\|_1 + \lambda\|W\|_1 + \frac{\rho}{2}\|I_k - \mu\|_F^2. \quad (1)$$

Note that an ℓ_2-regularization term has been added in order to avoid the trivial solution $\mu = 0$, $W = 0$. (An additional hyperparameter ρ is used.) We dualize the computation of the first ℓ_1 norm, and rely on a min-max / primal-dual approach to recast the problem as

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

where Z is an $m \times k$ matrix and $\|Z\|_\infty = \max_{i,j} |z_{ij}|$. We consider the following scheme, which consists in a proximal descent with respect to the primal variables (W, μ) followed by a proximal ascent with respect to the dual variable Z (see Section 5):

$$\begin{aligned}
W^{n+1} &:= \arg\min_W \frac{1}{2\tau}\|W - W^n\|_F^2 - \langle X^T Z^n, W\rangle + \lambda\|W\|_1 \\
\mu^{n+1} &:= \frac{1}{1+\tau_\mu\rho}(\mu^n + \rho\tau_\mu I_k - \tau_\mu Y^T Z^n) \\
Z^{n+1} &:= \mathrm{proj}_{B_\infty}(Z + \sigma(Y\mu - X(2W^{n+1} - W^n)))
\end{aligned}$$

where B_∞ is the closed unit in the space of $\mathbb{R}^{m\times k}$ matrices endowed with the ℓ_∞ norm. The update on W can be computed using a suitable soft thresholding operator:

$$\begin{aligned}
W_{i,j}^{n+1} &= W_{i,j}^n - \rho\cdot\tau_\mu I_k + \tau(X'Z^n)_{i,j} - \lambda\theta_{i,j}, \\
\theta_{i,j} &\leqslant 1 \quad \text{and} \quad \theta_{i,j} = \mathrm{sign}(W_{i,j}) \quad \text{if } W_{i,j} \neq 0
\end{aligned} \quad (3)$$

Writing explicitly the projection operator on B_∞, we eventually derive Algorithm 1. The convergence condition discussed in Section 5 imposes that

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

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

$$\|X\| = \sup_{\|W\|_F\leq 1} \|XW\|_F = \sup_{\|v\|_2\leq 1} \|X(:)v\|_2. \quad (5)$$

Algorithm 1 Primal-dual algorithm—$\mathrm{soft}(V, \lambda)$ is the standard soft thresholding , and μ is the ℓ_1 center of clusters.

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

Algorithm 2 Primal-dual algorithm, constrained case—$\mathrm{proj}(V, \eta)$ is the projection on the ℓ_1 ball of radius η (see (Condat, 2016)).

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

The main drawback of this penalty minimization is in the cost associated with the computation of the Lagrange multiplier λ using homotopy algorithms (Friedman et al., 2010; Hastie et al., 2004). The worst complexity case is $O(3^d)$ (Mairal & Yu, 2012), which is usually intractable on high-dimensional data sets such as genomics data sets. To overcome this computational issue, we introduce instead a constrained formulation of the problem.

3. Primal-dual scheme, constrained formulation

3.1. Constrained primal-dual approach

We consider the convex constrained supervised classification problem,

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

that we dualize as before:

$$\min_{(W,\mu)} \max_{\|Z\|_\infty \leqslant 1} \langle Z, Y\mu - XW\rangle + \frac{\rho}{2}\|I_k - \mu\|_F^2 \text{ s.t. } \|W\|_1 \leqslant \eta. \quad (7)$$

We adapt the update of W of Algorithm 1 by using a projected gradient step instead of thresholding, and devise Algorithm 2.

3.2. Constrained primal-dual approach with over-relaxation

An over-relaxed variant of the previous algorithm is presented below (Algorithm 3). In this case, condition (4) should be replaced with (see section 5.2)

$$\boxed{\sigma\left(\frac{\tau_\mu}{1 + \frac{\tau_\mu \rho}{4}\frac{1-2\gamma}{1-\gamma}}\|Y\|^2 + \tau\|X\|^2\right) < 1,} \quad (8)$$

at least if $\gamma < 1/2$. For $\gamma \geq 1/2$, the condition obtained for $\rho = 0$ is better and simply reads

$$\boxed{\sigma\left(\tau_\mu\|Y\|^2 + \tau\|X\|^2\right) < 1.} \quad (9)$$

Algorithm 3 Primal-dual algorithm, constrained case with over-relaxation.

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

4. Extension to other criteria: Frobenius loss

Our method can be extended straightforwardly to other criteria provided that (i) we can compute the projection on the dual ball for the loss data term, (ii) we can compute the prox for the regularization term. Such examples include combinations of ℓ_1 and ℓ_2 norms, as seen in group LASSO (Jacob et al., 2009). In this paper, we study an algorithm for the Frobenius norm. Note that our approach based on a dual computation of the norm allows us to use the norm itself, instead of the squared Frobenius norm (for with other approaches, taking care of the smoothness of the loss term, would be available; see, *e.g.*, (Combettes & Pesquet, 2011)).

4.1. Frobenius loss Minimization

We consider the following update of (6):

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

and dualize according to

$$\min_{(W,\mu)} \max_{\|Z\|_F \leqslant 1} \langle Z, Y\mu - XW\rangle + \frac{\rho}{2}\|I_k - \mu\|_F^2 \text{ s.t. } \|W\|_1 \leqslant \eta. \tag{11}$$

Obvious modifications of the previous scheme lead to Algorithm:

Algorithm 4 Primal-dual algorithm for Frobenius loss minimization: constrained case.

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

4.2. Frobenius loss minimization with over-relaxation

Obvious modifications of the previous scheme with over-relaxation lead to Algorithm 5. (Note that the algorithm requires computation of $\|Z\|_F$ at each iteration.)

Algorithm 5 Primal-dual algorithm for Frobenius loss minimization with over-relaxation.

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

A convergence analysis is drawn in Section 5. We provide in Section 6 a comparison between Algorithms 3 and 5.

5. Convergence Analysis

5.1. Convergence of primal-dual algorithms

The proof of convergence of the algorithms relies on Theorems 1 and 2 in (Chambolle & Pock, 2016) which we

slightly adapt for our setting. The algorithms we present here correspond to Alg. 1 and 2 in that reference, adapted to the particular case of problem (1) and its saddle-point formulation (2). In addition, here, the primal part of the objective is "partially strongly convex" (ρ-strongly convex with respect to the variable μ, thanks to the term $(\rho/2)\|\mu - I\|^2$. (We could exploit this to gain "partial acceleration" (Valkonen & Pock, 2017), however at the expense of a much more complex method and no clear gain for the variable W, while translated in the Euclidean setting, (Chambolle & Pock, 2016) remains simple and easy to improve.) For our setting we consider a general objective of the form:

$$\min_{x,x'} \max_{y} f(x) + g(x') + \langle Kx + K'x', y\rangle - h^*(y) \quad (12)$$

for f, g, h convex functions whose "prox" (see below) are easy to compute and K, K' linear operators, and we assume moreover f is ρ-strongly convex for some $\rho > 0$. We will show how this last property can be exploited to "boost" the convergence, allowing for larger steps than usually suggested by other authors.

When computing the "prox" $\hat{x}$ at point $\bar{x}$ of a ρ-strongly convex function $x \mapsto f(x)$, with parameter τ, that is, the minimizer

$$\hat{x} = \mathrm{prox}_{\tau f}(\bar{x}) := \arg\min_x f(x) + \frac{\|x - \bar{x}\|^2}{2\tau}, \quad (13)$$

one has for all test point x:

$$f(x) + \frac{\|x - \bar{x}\|^2}{2\tau} \geq f(\hat{x}) + \frac{\|\hat{x} - \bar{x}\|^2}{2\tau} + \frac{\|x - \hat{x}\|^2}{2\tau} + \frac{\rho}{2}\|x - \hat{x}\|^2.$$

However, combined with non-strongly convex iterates, the slight improvement given by the factor ρ is hard to exploit (whereas for simple gradient descent type iterates one obviously can derive linear convergence to the optimum), see for instance (Valkonen & Pock, 2017) for a possible strategy. We exploit here this improvement in a different way. We combine the parallelogram identity

$$\|x - \bar{x}\|^2 + \|x - \hat{x}\|^2 = \frac{1}{2}\|\bar{x} - \hat{x}\|^2 + 2\left\|x - \tfrac{\bar{x}+\hat{x}}{2}\right\|^2$$

with the previous inequality to obtain:

$$f(x) + (1 + \tau\tfrac{\rho}{2})\frac{\|x - \bar{x}\|^2}{2\tau} \geq f(\hat{x}) + (1 + \tau\tfrac{\rho}{4})\frac{\|\hat{x} - \bar{x}\|^2}{2\tau} + (1 + \tau\tfrac{\rho}{2})\frac{\|x - \hat{x}\|^2}{2\tau}. \quad (14)$$

The first type of algorithm we consider is Algorithm 1, which corresponds to Alg. 1 in (Chambolle & Pock, 2016) (see also (Pock et al., 2009; Esser et al., 2010; Chambolle & Pock, 2011)). It consists in tackling problem (12) by alternating a proximal descent step in x, x' followed by an ascent step in y:

$$\begin{aligned} x^{n+1} &= \mathrm{prox}_{\tau f}(x^n - \tau K^* y^n), \\ x'^{n+1} &= \mathrm{prox}_{\tau' g}(x'^n - \tau' K'^* y^n), \\ y^{n+1} &= \mathrm{prox}_{\sigma h^*}(y^n + \\ &\quad \sigma(K(2x^{n+1} - x^n) + K'(2x'^{n+1} - x'^n))). \end{aligned} \quad (15)$$

We then introduce the "ergodic" averages

$$X^N = \tfrac{1}{N}\sum_{n=1}^N x^n, \quad X'^N = \tfrac{1}{N}\sum_{n=1}^N x'^n, \quad Y^N = \tfrac{1}{N}\sum_{n=1}^N y^n.$$

Theorem 1 in (Chambolle & Pock, 2016), shows with an elementary proof the estimate, for any test point (x, x', y):

$$\mathcal{L}(X^N, X'^N, y) - \mathcal{L}(x, x', Y^N) \leq \frac{1}{2N}\left\|\begin{pmatrix} x \\ x' \\ y \end{pmatrix} - \begin{pmatrix} x^0 \\ x'^0 \\ y^0 \end{pmatrix}\right\|^2_{M_{\tau,\tau',\sigma}} \quad (16)$$

where $\mathcal{L}$ is the Lagrangian function in (12) and provided the matrix $M_{\tau,\tau',\sigma}$, given by

$$M_{\tau,\tau',\sigma} = \begin{pmatrix} \frac{I}{\tau} & 0 & -K^* \\ 0 & \frac{I}{\tau'} & -K'^* \\ -K & -K' & \frac{I}{\sigma} \end{pmatrix} \quad (17)$$

is positive-definite. Before exploiting the estimate (16), let us express the conditions on τ, τ', σ which ensure that this is true. We need that for any $(\xi, \xi', \eta) \neq 0$,

$$\frac{1}{\tau}\|\xi\|^2 + \frac{1}{\tau'}\|\xi'\|^2 + \frac{1}{\sigma}\|\eta^2\| > 2\langle K\xi, \eta\rangle + 2\langle K'\xi', \eta\rangle$$

and obviously, this is the same as requiring that for any a, a', b positive numbers,

$$\frac{a^2}{\tau} + \frac{a'^2}{\tau'} + \frac{b^2}{\sigma} > 2(\|K\|a + 2\|K'\|a')b.$$

The worst b in this inequality is $b = \sigma(\|K\|a + 2\|K'\|a')$, then one checks easily that the worse a, a' are of the form $\bar{a}\|K\|\tau$, $\bar{a}\|K'\|\tau'$ respectively, so that one should have for all $\bar{a} \neq 0$:

$$\bar{a}^2\left(\|K\|^2\tau + \|K'\|^2\tau'\right) > \sigma\bar{a}^2\left(\|K\|^2\tau + \|K'\|^2\tau'\right)^2,$$

yielding the condition

$$\sigma(\tau\|K\|^2 + \tau'\|K'\|^2) < 1.$$

We notice in addition that under such a condition, one also has

$$M_{\tau,\tau',\sigma} \leq 2\begin{pmatrix} \frac{I}{\tau} & & \\ & \frac{I}{\tau'} & \\ & & \frac{I}{\sigma} \end{pmatrix}$$

which allows to simplify a bit the expression in the right-hand side of (16) (at the expense of a factor 2 in front of the estimate).

We have not made use of the strong convexity up to now, and in particular, of (14). A quick look at the proof of Theorem 1 in (Chambolle & Pock, 2016) shows that it will improve slightly the latter condition, allowing to replace τ with the smaller effective step $\tau/(1+\tau\rho/4)$, yielding the new condition

$$\sigma\left(\frac{\tau}{1+\tau\frac{\rho}{4}}\|K\|^2+\tau'\|K'\|^2\right)<1. \tag{18}$$

This ensures now that (16) holds with $M_{\tau,\tau',\sigma}$ replaced with

$$M_{\tau,\tau',\sigma,\rho}=\begin{pmatrix}\left(\frac{1}{\tau}+\frac{\rho}{2}\right)I & 0 & -K^*\\ 0 & \frac{I}{\tau'} & -K'^*\\ -K & -K' & \frac{I}{\sigma}\end{pmatrix}\leq \begin{pmatrix}\left(\frac{2}{\tau}+\frac{3\rho}{4}\right)I & 0 & 0\\ 0 & \frac{2}{\tau'}I & 0\\ 0 & 0 & \frac{2}{\sigma}I\end{pmatrix} \tag{19}$$

where the last inequality follows from (18). Applied to problem (2), which is ρ-convex in μ, we find that (18) becomes the condition

$$\boxed{\sigma\left(\frac{\tau_\mu}{1+\tau_\mu\frac{\rho}{4}}\|Y\|^2+\tau\|X\|^2\right)<1.} \tag{20}$$

When (20) holds, then the ergodic iterates (here we denote W^n, etc, the value of W computed at the end of iteration n):

$$\bar{W}^N=\frac{1}{N}\sum_{n=1}^N W^n,\ \bar{\mu}^N=\frac{1}{N}\sum_{n=1}^N \mu^n,\ \bar{Z}^N=\frac{1}{N}\sum_{n=1}^N Z^n. \tag{21}$$

satisfy for all W,μ,Z:

$$\mathcal{L}(\bar{W}^N,\bar{\mu}^N,Z)-\mathcal{L}(W,\mu,\bar{Z}^N)\leq \frac{1}{N}\Big(\frac{3\rho}{8}\|\mu-\mu^0\|^2+\frac{\|\mu-\mu^0\|^2}{\tau_\mu}+\frac{\|W-W^0\|^2}{\tau}+\frac{\|Z-Z^0\|^2}{\sigma}\Big). \tag{22}$$

Here the Lagrangian $\mathcal{L}$ is:

$$\mathcal{L}(W,\mu,Z)=\langle Z,Y\mu-XW\rangle+\lambda\|W\|_1+\frac{\rho}{2}\|I-\mu\|^2.$$

We denote $\mathcal{E}(W,\mu)=\sup_Z\mathcal{L}(W,\mu,Z)$ the primal energy (which appears in (1)) and remark that in (2), Z is bounded ($|Z_{i,j}|\leq 1$ for all i,j) so that $\|Z-Z^0\|^2\leq 4mk$ in (22). Hence, taking the supremum on Z and choosing for (W,μ) a primal solution (minimizer of E) (W^*,μ^*), we deduce:

$$\mathcal{E}(\bar{W}^N,\bar{\mu}^N)-\mathcal{E}(W^*,\mu^*)\leq \frac{1}{N}\left(\frac{4mk}{\sigma}+\left(\frac{3\rho}{8}+\frac{1}{\tau_\mu}\right)\|\mu^*-\mu^0\|^2+\frac{\|W^*-W^0\|^2}{\tau}\right). \tag{23}$$

In general, if one can compute reasonable estimates Δ_μ, Δ_W for these quantities, one should take:

$$\tau=\frac{\Delta_W}{2\sqrt{mk}\|X\|},\quad \tau_\mu=\begin{cases}\frac{1}{\frac{2\sqrt{mk}\|Y\|}{\Delta_\mu}-\frac{\rho}{4}} & \text{if } \frac{8\sqrt{mk}\|Y\|}{\rho\Delta_\mu}>1\\ \tau_\mu>>1 & \text{else,}\end{cases}$$

$$\sigma=\frac{1}{\frac{\tau_\mu}{1+\tau_\mu\frac{\rho}{4}}\|Y\|^2+\tau\|X\|^2}.$$

to obtain (considering here only the case ρ small, that is when $\rho\Delta_\mu\leq 8\sqrt{mk}\|Y\|$):

$$\mathcal{E}(\bar{W}^N,\bar{\mu}^N)-\mathcal{E}(W^*,\mu^*)\leq\frac{\sqrt{mk}(5\Delta_\mu\|Y\|+4\Delta_W\|X\|)}{N},$$

There is no clear way how to estimate *a priori* the norm $\|W^*-W^0\|$ in the Lagrangian approach.

Remark 1 *Note that for the ℓ^1 constrained problem (6) Δ_W is bounded. Since $\|W\|_1\leq\eta$: $\|W^*-W^0\|\leq\|W^*-W^0\|_1\leq 2\eta$, we use the estimate $\Delta_W\leq 2\eta$. Using the initial value $\mu^0=I_k$, Δ_μ is also easily shown to be bounded (as W is). Empirically, we found that we can use the estimate $\Delta_\mu\lesssim\alpha\|I_k\|_F=\alpha\sqrt{k}$ where $\alpha\in[0,1]$ is a parameter to be tuned. Thus ρ being small we have $\frac{8\sqrt{m}\|Y\|}{\rho\alpha}>1$. Moreover, using $\|X\|=1$ (X can be normalized), we obtain the following reasonable parameter choice:*

$$\boxed{\tau=\frac{\Delta_W}{2\sqrt{mk}},\quad \tau_\mu=\frac{\alpha}{2\sqrt{m}\|Y\|-(1/4)\alpha\rho}\qquad \sigma=\frac{1}{\frac{\tau_\mu}{1+\tau_\mu\frac{\rho}{4}}\|Y\|^2+\tau}.} \tag{24}$$

In the case of Problem (11) (Sec. 4), Z is also bounded but then, one has simply $\|Z^*-Z^0\|^2\leq 4$, hence (23) must be replaced with

$$\mathcal{E}(\bar{W}^N,\mu^N)-\mathcal{E}(W^*,\mu^*)\leq \frac{1}{N}\left(\frac{4}{\sigma}+\left(\frac{3\rho}{8}+\frac{1}{\tau_\mu}\right)\|\mu^*-\mu^0\|^2+\frac{\|W^*-W^0\|^2}{\tau}\right). \tag{25}$$

(Obviously, now, the energy $\mathcal{E}$ is the primal energy in (10).) The same analysis as before remains valid, but now with mk replaced with 1.

5.2. Convergence with over-relaxation

For the over-relaxed variant (Algorithm 3), the adaption is a little bit more complicated, and one does not benefit much from taking into account the partial strong convexity. One approach is to rewrite the improved descent rule (14) as follows:

$$\begin{aligned} f(\hat{x}) \leq f(x) + \frac{1+\tau\rho/2}{2\tau}\big(\|x-\bar{x}\|^2 - \|x-\hat{x}\|^2 \\ - \|\hat{x}-\bar{x}\|^2\big) + \frac{\rho}{8}\|\hat{x}-\bar{x}\|^2 \\ = f(x) + \frac{1}{2\widetilde{\tau}}\langle x-\hat{x}, \hat{x}-\bar{x}\rangle + \frac{\rho}{8}\|\hat{x}-\bar{x}\|^2 \end{aligned} \quad (26)$$

where $\widetilde{\tau} = \tau/(1+\tau\rho/2)$ is an effective time-step.

As a result, we observe that the first (primal) update in (15) yields the same rule as an explicit-implicit primal update of a nonsmooth+smooth functions with effective step $\widetilde{\tau}$ and Lipschitz constant $\rho/4$, *cf* Eq. (9) in (Chambolle & Pock, 2016). Hence, the analysis of these authors (see Sec. 4.1 in the above reference) can be reproduced almost identically and will yield for the over-relaxed algorithm (3) similar convergence rates, cf. (16)-(23), now, with the factor $1/N$ replaced with $1/((1+\gamma)N)$. It requires that the matrix

$$\widetilde{M} = \begin{pmatrix} (\frac{1}{\widetilde{\tau}} - \frac{\rho/4}{1-\gamma})I & 0 & -K^* \\ 0 & \frac{I}{\tau'} & -K'^* \\ -K' & -K & \frac{I}{\sigma} \end{pmatrix} \quad (27)$$

be positive definite. Observe however that the estimates hold for the ergodic averages (cf (21)) of the variables obtained at the end of Step 12 of Algorithm 3 and Step 13 of Algorithm 5, rather than for the over-relaxed variables (which could not even be feasible). We derive that for this method, condition (20) should be replaced with

$$\sigma\left(\frac{\tau_\mu}{1+\frac{\tau_\mu\rho}{4}\frac{1-2\gamma}{1-\gamma}}\|Y\|^2 + \tau\|X\|^2\right) < 1, \quad (28)$$

at least if $\gamma < 1/2$.

As seen, for $\gamma \geq 1/2$, the condition obtained for $\rho = 0$ is better (hence, the partial strong convexity does not seem to yield any reasonable improvement for this algorithm). It simply reads

$$\sigma\left(\tau_\mu\|Y\|^2 + \tau\|X\|^2\right) < 1, \quad (29)$$

and one gets the estimate from (Chambolle & Pock, 2016) (Eq. (24), further simplified thanks to (28)):

$$\begin{aligned} &\mathcal{E}(\bar{W}^N, \bar{\mu}^N) - \mathcal{E}(W^*, \mu^*) \leq \\ &\frac{1}{(1+\gamma)N}\left(\frac{4mk}{\sigma} + \frac{\|\mu^* - \mu^0\|^2}{\tau_\mu} + \frac{\|W^* - W^0\|^2}{\tau}\right). \end{aligned} \quad (30)$$

6. Numerical experiments

6.1. Experimental settings

In this section, we compare the constrained primal-dual ℓ_1 approach with the one based on the Frobenius norm. Our primal-dual method can be applied to any classification problem with feature selection on 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 simulated and real single-cell sequencing dataset. There are two advantages of such biological dataset. First, many public data are now available for testing reproductibility; besides, these dataset suffer from outliers ("dropouts") with different levels of noise depending on sequencing experiments. Single-cell is a new technology which has been elected "method of the year" in 2013 by *Nature Methods* (Evanko, 2014). We provide also evaluation on proteomics mass-spectrometric dataset. A test query x (a dimension d row vector) is classified according to the following rule: it belongs to the unique class j^* such that

$$j^* \in \arg\min_{j=1,\ldots,k} \|\mu_j - xW\|_1. \quad (31)$$

Feature selection is based on the sparsity inducing ℓ_1 constraint. The projection on the ℓ_1 ball $Proj(V, \eta)$ aims at sparsifying the W matrix so that the gene j will be selected if $|W(j,:)\| > \varepsilon$. The set of non-zero column coefficients is interpreted as the signature of the corresponding cluster. We use the Condat method (Condat, 2016) to compute the projection on the ℓ_1 ball.

We report the classical *accuracy* versus η using cross validation (4 folds). We also define selectivity as the number of required genes to obtain a desired accuracy. Processing times are obtained on a laptop computer using an i7 processor (3.1 Ghz). In our experiments, we normalize the features according to $\|X\| = 1$, and we set $\mu^0 = I_k$ and $\rho = 0.0001$. We choose η in connection with the desired number of genes. As Δ_W and η are bounded, we can set a maximum value for τ, that we denote τ_0. Then we tune τ_μ and compute σ using equation (24). Based on results in Fig. 2, we set $N = 40$ for synthetic dataset and Tabula Muris dataset, while Ovarian data set requires $N = 60$ iterations. We report the ℓ_1 and Frobenius error (log-log plot) in the training set (normalized by the value of the first iterate) both for standard algorithm Fig. 4 and over-relaxed one in Fig. 5.

6.2. Synthetic dataset

The simulated dataset is a realistic simulation of single cell sequencing experiments (The dataset is provided in supplementary material). This dataset is composed of 600 samples 15,000 genes and $k = 4$ clusters.

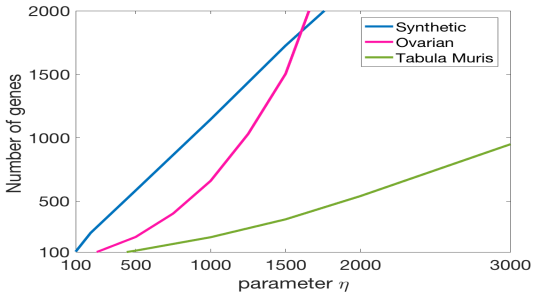

Figure 1. Synthetic dataset. The evolution of the number of selected genes versus the constraint η is a smooth monotonous function. The bound η for the ℓ^1 constraint is thus easily tuned.

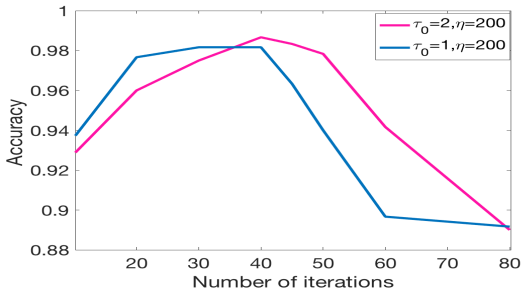

Figure 2. Synthetic dataset. This figure shows that roughly 40 iterations are optimal in order to avoid over-fitting.

6.3. Dataset: ovarian (Guyon et al., 2017)

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), and healthy or control patients. The dataset is composed of 216 samples and 15000 features.

6.4. Lung Tabula Muris single cell scRNA-seq dataset(Schaum, 2018)

Tabula Muris is a compendium of single cell transcriptome data from the model organism Mouse musculus, containing nearly 100,000 cells from 20 organs and tissues. The data allow comparison between gene expression in cell types. (The authors thank Pr Pascal Barbry team (IPMC Laboratory) for the group's biological expertise.) Lung Tabua Muris sub-dataset is a subset of Lung organ composed of 5,400 cells, 10,516 genes and k=14 clusters.

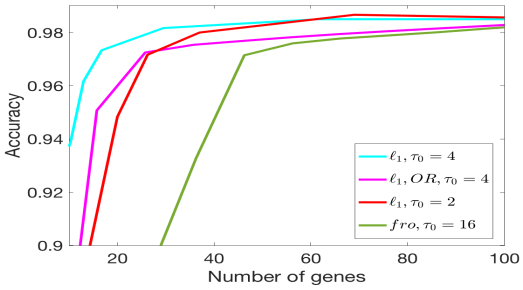

Figure 3. Synthetic dataset. The accuracy versus number of genes shows that a minimum number of genes is required to get the best possible classification followed by a large plateau. We get similar accuracy and selectivity for different values of parameter τ_0 and for over-relaxation. ℓ_1 norm improves selectivity over Frobenius norm minimization.

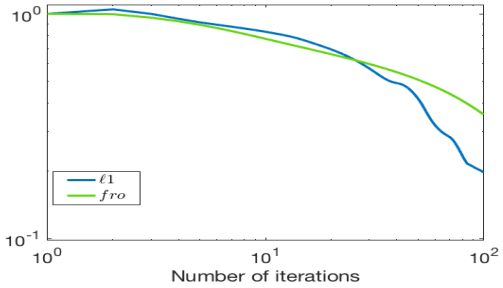

Figure 4. Synthetic dataset. Convergence rate is faster for ℓ_1 minimization.

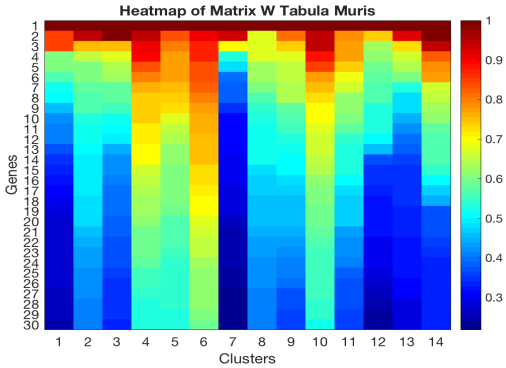

Figure 8. Tabula Muris: the signature is sparse and adapted to the clustering.

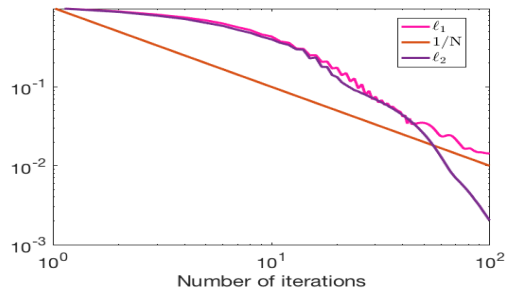

Figure 5. Synthetic dataset. Over-relaxation provides O(1/N) convergence rate for both ℓ_1 and Frobenius minimization.

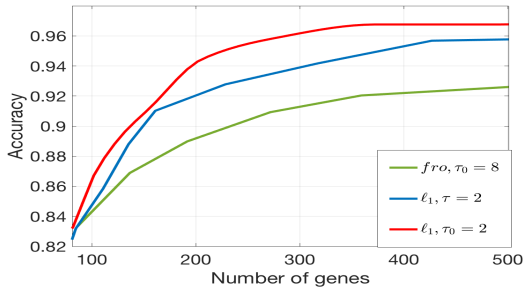

Figure 6. Ovarian dataset: using ℓ_1 norm minimization outperforms Frobenius minimization both for accuracy and feature selectivity.

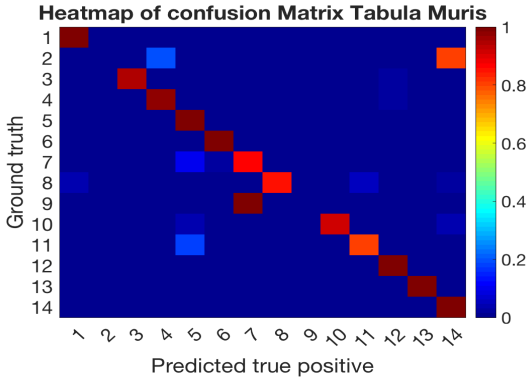

Figure 9. Tabula Muris Confusion Matrix: vertically Ground truth, Horizontally: Predicted True positive in %.

6.5. Accuracy, signature and scalability

Our numerical experiments demonstrate that our ℓ_1 norm minimization method is more robust to outliers (Fig. 6) and to dropouts on real large single cell datasets (Fig. 7). Fig. 3 shows that minimization of the ℓ_1 norm is robust to parameter τ_0. We report in Fig. 8 the sorted absolute value of the matrix W (divided by the maximum of each column). As expected by biologists, Fig. 8 shows that the sparsity of

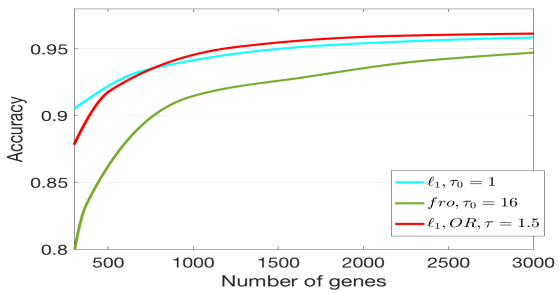

Figure 7. Tabula Muris dataset: accuracy: ℓ_1 norm minimization outperforms Frobenius norm minimization by 5 %. Moreover ℓ_1 minimization improves feature selectivity.

the signature of W columns is well adapted to each cluster. The accuracy for each cluster is given Fig. 7, illustrating the reliability of the signature. Fig. 4 shows that over relaxation improves convergence both for the ℓ_1 and Frobenius norm minimization. The complexity of our primal-dual algorithm is $O(d \times m)$ for primal iterates and $O(m \times k)$ for dual iterates. One must then add the cost of projection on the ℓ_1 ball; this cost is expected to be $O(d \times k)$ in average, in order that the our constrained Primal-dual method is scalable.

As pointed in Section (4) the squared Frobenius loss is smooth and we can use Fista algorithm (Beck & Teboulle, 2009; Chambolle & Dossal, 2015) combined with centroid classifier (Witten & Tibshirani, 2010) to minimize a squared Frobenius criterion. To do so, one performs a full run of Fista to retrieve W, then a single centroid computation to estimate μ. Table 1 shows that the primal-dual method outperforms Fista. (Note that the complexity of gradient based methods such as Fista is $O(d^2)$.)

Table 1. Time (seconds) according to methods: primal-dual is slightly faster for ℓ_1 norm minimization on large dataset since Frobenius norm minimization requires computation of Frobenius norm of $O(m \times k)$ matrix at each iteration. Primal-dual outperforms Fista primal algorithm.

Dataset	ℓ_1	ℓ_1 OR	Fro	Fista
Synthetic	**0.7**	**0.43**	**0.7**	**7.6**
Ovarian	**0.26**	**0.21**	**0.26**	**6.83**
Tumlung	**3.3**	**2.5**	**3.5**	**7.**

7. Conclusion

We have provided a new primal-dual method for supervised classification based on an ℓ_1 norm both for the data loss and the constraint (feature selection). Our algorithm computes jointly variables transform W and centers μ that are used to devise a classifier, and we establish convergence results. We demonstrate the effectiveness of our method on three

datasets (one synthetic, two from biological data), and provide a comparison between ℓ_1 and ℓ_2 costs. Both accuracy, selectivity and computational time are improved with the purely ℓ_1 approach. Extending the method to other criteria is easy on condition that efficient projection (on the dual ball for the loss data term) and prox (for the regularization term) algorithms are available. We have treated an example using the Frobenius norm of the loss. Current research extends our approach to other loss or regularization criteria such as sparse group LASSO. A code is available in supplementary material.

8. Appendix: minimization of the square Frobenius loss using Fista

Let us consider the classical feature selection criterion with the squared Frobenius loss:

$$\min_{(W,\mu)} \|Y - XW\|_F^2 \quad \text{s.t.} \quad \|W\|_1 \leqslant \eta. \tag{32}$$

We compute an estimator of W using Fista algorithm.

Algorithm 6 Feature selection algorithm with FISTA.

Input: $X, Y, \mu_0 = I, W_0, N, \gamma, \eta$
$W \leftarrow W_0$
$t \leftarrow 1$
for $n = 0, \ldots, N$ **do**
  $V \leftarrow W - \gamma X^T(XW - Y\mu)$
  $W_{\text{new}} \leftarrow P_\eta^1(V)$
  $t_{\text{new}} \leftarrow (n+5)/4$
  $\lambda \leftarrow 1 + (t-1)/t_{\text{new}}$
  $W \leftarrow (1-\lambda)W + \lambda W_{\text{new}}$
  $t \leftarrow t_{\text{new}}$
end for
Output: W

References

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

Beck, A. and Teboulle, M. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009.

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

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

Chambolle, A. and Dossal, C. On the convergence of the iterates of "fista". *Journal of Optimization Theory and Applications, Springer Verlag*, (166), 2015.

Chambolle, A. and Pock, T. 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.

Chambolle, A. and Pock, T. On the ergodic convergence rates of a first-order primal-dual algorithm. *Math. Program.*, 159(1-2, Ser. A):253–287, 2016. ISSN 0025-5610.

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

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

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

Condat, L. Fast projection onto the simplex and the l1 ball. *Mathematical Programming Series A*, 158(1):575–585, 2016.

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

Ding, C. and Li, T. Adaptive dimension reduction using discriminant analysis and k-means clustering. In *Proceedings of the 24th International Conference on Machine Learning*, pp. 521–528, 2007.

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

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

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

Esser, E., Zhang, X., and Chan, T. F. A general framework for a class of first order primal-dual algorithms for convex optimization in imaging science. *SIAM J. Imaging Sci.*, 3 (4):1015–1046, 2010.

Evanko, D. Method of the year 2013: Methods to sequence the dna and rna of single cells are poised to transform many areas of biology and medicine. *Nature Methods, Vol 11*, 2014.

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

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

Furey, T. S., Cristianini, N., Duffy, N., Bednarski, D. W., Schummer, M., and Haussler, D. Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, 16(10): 906–914, 2000.

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

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

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

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

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

Li, C. and Li, H. Network-constrained regularization and variable selection for analysis of genomic data. *Bioinformatics*, 24(9):1175–1182, 2008.

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

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

Liu, J. and Ye, J. Moreau-yosida regularization for grouped tree structure learning. In *Advances in Neural Information Processing Systems 23*.

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

Mairal, J. and Yu, B. Complexity analysis of the lasso regularization path. In *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, pp. 353–360, 2012.

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

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

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

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

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

Pock, T., Cremers, D., Bischof, H., and Chambolle, A. An algorithm for minimizing the mumford-shah functional. In *Computer Vision, 2009 IEEE 12th International Conference on*, pp. 1133–1140. IEEE, 2009.

Radovanovic, M., Nanopoulos, A., and Ivanovic, M. Hubs in space: Popular nearest neighbors in high-dimensional data. *Journal of Machine Learning Research*, 11:2487–2531, 2010.

Schaum, N. Single-cell transcriptomics of 20 mouse organs creates a tabula muris. *Nature*, 562(7727):367–372, 2018.

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. A sparse-group lasso. *Journal of Computational and Graphical Statistics*, 22(2):231–245, 2013.

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

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

Valkonen, T. and Pock, T. Acceleration of the PDHGM on partially strongly convex functions. *J. Math. Imaging Vision*, 59(3):394–414, 2017.

Witten, D. M. and Tibshirani, R. A framework for feature selection in clustering. *Journal of the American Statistical Association*, 105(490):713–726, 2010.

Yuan, M. and Lin, Y. Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B, 68(1)*, 68(1):49–67.

Zou, H. and Hastie, T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, 67:301–320, 2005.

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