Sparse hierarchical interaction learning with epigraphical projection

Mingyuan Jiu, Nelly Pustelnik, Stefan Janaqi, Mèriam Chèbre, Lin Qi, Philippe Ricoux

Sparse hierarchical interaction learning with epigraphical projection*

Mingyuan Jiu [‡†] Nelly Pustelnik [‡] Stefan Janaqi [§]
Mériam Chebre [§] Lin Qi [†] Philippe Ricoux [§]

26 August 2019

Abstract

This work focuses on learning optimization problems with quadratical interactions between variables, which go beyond the additive models of traditional linear learning. We investigate more specifically two different methods encountered in the literature to deal with this problem: "hierNet" and structured-sparsity regularization, and study their connections. We propose a primal-dual proximal algorithm based on an epigraphical projection to optimize a general formulation of these learning problems. The experimental setting first highlights the improvement of the proposed procedure compared to state-of-the-art methods based on fast iterative shrinkage-thresholding algorithm (i.e. FISTA) or alternating direction method of multipliers (i.e. ADMM), and then, using the proposed flexible optimization framework, we provide fair comparisons between the different hierarchical penalizations and their improvement over the standard ℓ_1-norm penalization. The experiments are conducted both on synthetic and real data, and they clearly show that the proposed primal-dual proximal algorithm based on epigraphical projection is efficient and effective to solve and investigate the problem of hierarchical interaction learning.

*This is a pre-print of an article published in Journal of Signal Processing Systems. The final authenticated version is available online at: https://doi.org/[insert DOI]. Corresponding author N. Pustelnik nelly.pustelnik@ens-lyon.fr.

[†]School of Information Engineering, Zhengzhou University, Zhengzhou, 450001, China. This work was supported by a grant from the National Natural Science Foundation of China (No. 61806180)

[‡]Univ Lyon, Ens de Lyon, Univ Lyon 1, CNRS, Laboratoire de Physique, Lyon, 69342, France

[§]TOTAL SA, Direction Scientifique, 92069 Paris La Défense, France

1 Introduction

Learning interactions between features is of great interest in data processing. In this work we focus on quadratic interaction effects between the features that extend linear models. Our work focuses both on the *regression problem* and *multiclass SVM*, which are still subjects of active research when small dataset are involved (especially in medical [38] or physics [30] applications).

Feature selection – Our work follows the line of feature selection methods [3, 8, 10, 16, 27, 43], which aim to remove the redundant features and only keep the informative ones. The relevant features are usually correlated and belong to the same group. In this work, additionally to perform feature selection, we learn the interactions between the selected features by making some hierarchy constraints, where the interactions may occur either between the selected features or if only one of both features is active.

Knowledge from the interactions – Features interactions can provide us a new knowledge about the correlation between the subjects. For instance, we can refer to gene interactions aiming to highlight relevant regulatory relationships for genes [31] or as mentioned in [7] "the co-occurrence of two symptoms may lead a doctor to be confident that a patient has a certain disease, whereas the presence of either symptom without the other would provide only a moderate indication of that disease".

Better discrimination – Dealing with interactions allows us to increase the feature space in order to provide a better discrimination in the learning process [2]. The integration of quadratic interactions in the learning problem is not new, for example, discriminant quadratic learning [2, 33, 42] learns the covariance matrix in the quadratic term to improve the discrimination ability. We can also refer to efficient multilevel procedures [1, 22] starting with fast learning algorithm focus on the linear model and then adding higher-order interaction features, possibly using the learned weights as a guide.

Dimensionality challenge and sparsity – One major challenge when dealing with quadratic interaction learning is that the interaction number quadratically increases with the feature size, for example, 1000 features will have about one million possible interactions, this therefore results in an overfitting problem due to insufficient observation samples in most of the regression problems and for some classification ones. To overcome this weakness, the linear weights and the quadratic interactions can be assumed to be sparse because most features would not contribute to the decision. Sparsity-based regularization is known to be mainly suitable when the feature size is larger than the training samples. Various sparsity regularizations have been proposed and extensively studied in the additive model context,

for instance, involving ℓ_0-pseudo-norm [41], ℓ_1-norm [39], ℓ_∞-norm [45], or structural sparsity norm [4]. See [3] for an exhaustive list of sparse-based regularization in learning and [21] to refer to structural sparsity. The formalism we consider in this work follows ideas proposed in [7, 23, 29, 37, 44]. A detailed comparison with these state-of-the-art methods is provided in next section.

Contributions – Quadratic interaction learning imposes a specific design of the penalization term (this will be formally described in Section 2). Several choices of penalizations have already been proposed in the literature, but each of them relies on a different algorithmic strategy, leading to unfair numerical comparisons. Indeed, do the numerical differences in the learning performance come from the penalization choice or from the algorithmic schemes ? For instance, it is well known that the number of inner iterations (as proposed in [24]) is often a parameter tricky to adjust, or that the algorithmic parameters such as the step-size may impact a lot the convergence speed [23]. For all these reasons, this work is dedicated to provide a common algorithmic scheme without inner iterations aiming to handle with the most recents state-of-the-art penalizations. Our detailed contributions are listed below:

- Propose a new algorithmic scheme relying of primal-dual algorithm and derive new closed form of epigraphical projections in order to solve the unifying minimization problem (cf. Problem 2.1) proposed in [37] when $q = +\infty$ and $z(\cdot) = \|\cdot\|_r$ with $r = \{1, +\infty\}$;
- Compare the proposed algorithms to the state-of-the-art methods (FISTA and ADMM), showing a unique and efficient algorithmic scheme for weak and strong formulation and for $r = \{1, +\infty\}$ allowing us to avoid inner iterations.
- Provide the counterpart of the regression minimization Problem 2.1 for classification task.
- Conduct extensive experiments in the regression and classification framework on the simulated data, on two disease applications (i.e. HIV and Parkinson) to analyze the correlations between the factors for the diseases, validating the effectiveness and efficiency of the proposed algorithms.

This work improves over our preliminary contribution [25] where we restricted our study to classification (cf. section 5 of this work). The algorithm was derived for $r = 1$, relying only on Proposition 4.3 and for which the proof was not provided. The experiments were conducted on face classification while in this work we focus on synthetic data and diseases applications.

Outline –This paper is organized as follows: Section 2 firstly describes the formal problem in the context of regression and refers to related works. Section 3 derives an epigraphical

writing of the objective function. Section 4 presents the primal-dual proximal algorithm iterations and the convergence guarantees. Its specification to our minimization problem with the hierarchical regularization term is specified and new epigraphical projections are provided. Section 5 provides the objective function and the associated algorithm in the context of multiclass SVM with hierarchical interactions. Section 6 evaluates the performance of the proposed strategy compared to FISTA and ADMM formulations proposed in [7, 24] both on synthetic data and real applications. Finally Section 7 gives our conclusion.

2 Regression model

The regression training set is denoted $\mathcal{R} = \{\big(y_\ell, \phi(\mathbf{x}_\ell)\big) \in \mathbb{R} \times \mathbb{R}^N \,|\, \ell \in \{1, \dots, L\}\}$, where $\mathbf{x}_\ell$. It can denote either data without structure or with structure such as a signal, an image or a graph of size M. Then $\phi(\mathbf{x}_\ell)$ denotes the coefficients associated with the data $\mathbf{x}_\ell$, e.g time-frequency coefficients [20], scattering coefficients [9] or CNN coefficients [28], where generally $N \gg M$.

Several minimization strategies have been provided in the literature to jointly select discriminating features among the features $\phi(\mathbf{x}_\ell)$ and identify meaningful interactions between these features. To clarify the state-of-the-art contributions in this context, we propose to recall a general minimization problem derived in [37]:

Problem 2.1 *Let $\mathcal{R} = \{\big(y_\ell, \phi(\mathbf{x}_\ell)\big) \in \mathbb{R} \times \mathbb{R}^N \,|\, \ell \in \{1, \dots, L\}\}$ be the training regression dataset. We aim to estimate the regression weight vector $\widehat{\mathbf{v}}$ and the interaction matrix $\widehat{\mathbf{\Theta}}$ solving:*

$$(\widehat{\mathbf{v}}, \widehat{\mathbf{\Theta}}) \in \underset{\substack{\mathbf{v}\in\mathbb{R}^N \\ \Theta\in\mathbb{R}^{N\times N}}}{\operatorname{arg\,min}} \frac{1}{2} \sum_{\ell=1}^{L} \big(y_\ell - \phi(\mathbf{x}_\ell)^\top \mathbf{v} - \phi(\mathbf{x}_\ell)^\top \mathbf{\Theta} \phi(\mathbf{x}_\ell)\big)^2 + \Omega(\mathbf{v}, \mathbf{\Theta}) \tag{1}$$

where, for every $\mathbf{v} = \big(v^{(i)}\big)_{1\leq i\leq N}$ and $\mathbf{\Theta} = \big(\Theta^{(i,j)}\big)_{1\leq i\leq N, 1\leq j\leq N}$,

$$\Omega(\mathbf{v}, \Theta) = \frac{\lambda}{2}\|\Theta\|_1 + \lambda \sum_{i=1}^{N} \|\big(v^{(i)}, z\big(\Theta^{(i,\cdot)}\big)\big)\|_q + \iota_C(\Theta) \tag{2}$$

with $z\colon \mathbb{R}^N \to \mathbb{R}$, $1 \leq q \leq +\infty$, and ι_C denotes the indicator function[1] of a closed convex set $C \subset \mathbb{R}^{N\times N}$.

[1]For every $x \in \mathcal{H}$, $\iota_C(x) = 0$ if $x \in C$ and $+\infty$ otherwise.

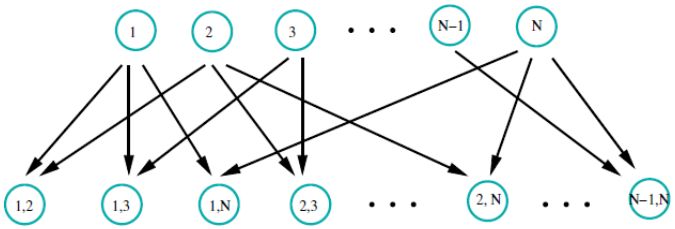

Figure 1: Hierarchical feature interaction. The first row is N additive features, and the second row is the interactions between pairs of features.

2.1 Choice for the constraint C: Weak/strong hierarchy

The hierarchy penalization Ω creates the relationship between $\mathbf{v}$ and $\boldsymbol{\Theta}$, as shown in Fig. 1. Two types of hierarchy constraints have been defined in "hierNet" [7]: (i) *weak hierarchy* where the interactions happen (i.e., $\Theta^{(i,j)} \neq 0$) only if one of the associated features in $\mathbf{v}$ is non-zero (i.e., $v^{(i)} \neq 0$ or $v^{(j)} \neq 0$) and (ii) *strong hierarchy* when both associated features are non-zero. These two types of hierarchy are imposed by means of the closed convex set $C \subset \mathbb{R}^{N\times N}$. The weak hierarchy is obtained with $C = \mathbb{R}^{N\times N}$ and strong hierarchy by imposing a symmetric structure for the matrix of interactions using $C = S = \{\boldsymbol{\Theta} \in \mathbb{R}^{N\times N} \,|\, \boldsymbol{\Theta} = \boldsymbol{\Theta}^\top\}$.

2.2 Positioning of Problem 2.1 w.r.t state-of-the-art

Most of the state-of-the-art penalizations may be interpreted as a specific case of Problem 2.1, for instance:

- [24] : $z(\cdot) = (\cdot)^\top$, $q = \{2, +\infty\}$,
- [7, 23] : $z(\cdot) = \|\cdot\|_1$ and $q = +\infty$,
- [37] : $z(\cdot) = (\cdot)^\top$ and $q = \{2\}$.

Note that [24] focuses on weak hierarchy using $C = \mathbb{R}^{N\times N}$. [7] is encountered under the name "hierNet", while [23] is named "FAMILY" and [37] is "Type-A GRESH". The latent overlapping group lasso formulation of [7] is known as "glinternet" [29].

From the algorithmic point of view several strategies can be encountered. In [24] and [37], the iterations are derived from FISTA [6]. This procedure appears to be very efficient when $C = \mathbb{R}^{N\times N}$ while for $C = S$ it requires inner iterations based on Dykstra's algorithm that significantly slows the convergence [14]. On the other hand, [7] and [23] resort to ADMM to deal with $C = S$.

Using recent developments of non-smooth convex optimization, the objective of this work is to provide a unified framework and an efficient algorithmic strategy based on primal-dual proximal algorithms and epigraphical projection to solve an epigraphical relaxation of (4) when $C = \mathbb{R}^{N\times N}$ or $C = S$, $z(\cdot) = \|\cdot\|_r$ with $r = \{1, +\infty\}$ and $q = +\infty$. The algorithmic procedure does not resort to inner iterations, thus leading to a faster implementation.[2]

3 Epigraphical formulation

In order to handle with complex optimization problems, the two major solutions encountered in the literature are either to formulate it in the dual or to increase the dimensionality of the problem. Dealing with an epigraphical formulation belongs to this second class and it is mainly adapted to nonlinear constraints. We recall that the epigraph of a function $f\colon \mathcal{H} \to]-\infty, +\infty]$ is defined as: $\text{epi}f = \{(x, \zeta) \in \mathcal{H} \times \mathbb{R} \,|\, f(x) \leq \zeta\}$, the projection of a point to this epigraph is called epigraphical projection. The utility of such epigraphical formulation can be illustrated when one wants to deal with a constraint of the form $\|x\|_1 \leq \eta$, for every $x = (x_i)_{i\in\mathbb{I}\subset\mathbb{N}}$ and $\eta > 0$ is known, whose projection does not have a closed form expression. An alternative is then to replace this constraint with these two constraints: $|x_i| \leq \zeta_i$ and $\sum_{i\in\mathbb{I}} \zeta_i = \eta$ where the first one denotes an epigraphical constraint and the second one denotes an hyperplane constraint, both having a closed form expression for the associated projection [15].

The resolution of Problem 2.1 is difficult, since it involves non-smooth convex functions and symmetry constraints. In [24], Jenatton et al. solve the problem for $r = \infty$ in the weak hierarchy formulation by means of a proximal algorithm. However, their algorithm is not well adapted to strong hierarchy. In [7], the authors reformulate Problem 2.1 when $r = 1$ under an epigraphical formulation which represents the problem as a set of hard constraints. This reformulation helps in the interpretation and it highlights a reduction in the shrinkage of certain main effects and an increase in the shrinkage of certain interactions. Its second advantage is to help in the design of the ADMM algorithmic scheme. The following proposition gives the epigraphical formulation of the minimization problem considered in this work.

Proposition 3.1 *The minimization problem*

$$\underset{\mathbf{v},\boldsymbol{\Theta}}{\text{minimize}}\ \frac{1}{2}\sum_{\ell=1}^{L}\left(y_\ell - \phi^\top(\mathbf{x}_\ell)\mathbf{v} - \phi^\top(\mathbf{x}_\ell)\boldsymbol{\Theta}\phi(\mathbf{x}_\ell)\right)^2 + \Omega(\mathbf{v}, \boldsymbol{\Theta}) \tag{3}$$

[2]Matlab codes will be made available at the time of the publication.

where, for every $\mathbf{v} = \left(v^{(i)}\right)_{1\leq i\leq N}$ *and* $\mathbf{\Theta} = \left(\Theta^{(i,j)}\right)_{1\leq i\leq N, 1\leq j\leq N}$,

$$\Omega(\mathbf{v}, \Theta) = \frac{\lambda}{2}\|\mathbf{\Theta}\|_1 + \lambda \sum_{i=1}^{N} \max\{v^{(i)}, \|\Theta^{(i,\cdot)}\|_r\} + \iota_C(\mathbf{\Theta}) \tag{4}$$

can be written in an epigraphical framework as

$$\underset{(\mathbf{v}^+, \mathbf{v}^-, \mathbf{\Theta})\in O\times O\times C}{minimize} \frac{1}{2}\sum_{\ell=1}^{L}\left(y_\ell - \phi^\top(\mathbf{x}_\ell)(\mathbf{v}^+ - \mathbf{v}^-) - \phi^\top(\mathbf{x}_\ell)\mathbf{\Theta}\phi(\mathbf{x}_\ell)\right)^2 + \frac{\lambda}{2}\|\mathbf{\Theta}\|_1 + \lambda 1^\top(\mathbf{v}^+ + \mathbf{v}^-) + \sum_{i=1}^{N} \iota_{E_r}\left(v^{+(i)}, v^{-(i)}, \Theta^{(i,\cdot)}\right) \tag{5}$$

where O *denotes the positive orthant,* $\mathbf{v} = \mathbf{v}^+ - \mathbf{v}^-$ *and* $E_r = \{(\omega^+, \omega^-, \mathbf{u}) \in \mathbb{R}\times\mathbb{R}\times \mathbb{R}^N \,|\, \|\mathbf{u}\|_r \leq \omega^+ + \omega^-\}$ *with* $r = \{1, +\infty\}$.

Proof 3.1 *The proof is given in Appendix A.*

The algorithmic strategy designed in next section will focus on this epigraphical formulation (5).

4 Primal-dual proximal algorithm

Proximal algorithms are derived from two main frameworks that are the forward-backward scheme and the Douglas-Rachford iterations (both being deduced from the Krasnoselskii-Mann scheme) [5]. FISTA can be presented as an accelerated version of forward-backward iterations [6, 11] while ADMM can be viewed as a Douglas-Rachford procedure in the dual [36]. In the literature dedicated to sparse regression the most encountered algorithm is FISTA when dealing with a sum of two convex functions where one is differentiable with a Lipschitz gradient. When the criterion involves more than two functions, typically an additional constraint such as the constraint S defined previously, is added and most of the works derive an ADMM procedure or compute the proximity operator by means of inner iterations which is often known to leading to an approximate solution even if global convergence can be obtained in specific cases [14]. Another class of algorithmic procedures allowing to minimize a criterion with more than two functions, possibly including a differentiable function with a Lipschitz gradient, is the class of primal-dual proximal approaches [12, 18, 26, 40].

Several primal-dual proximal schemes have been derived but one of the most popular is based on forward-backward iterations [18, 40] in order to estimate

$$\widehat{\mathbf{w}} \in \underset{\mathbf{w}\in\mathcal{H}}{\text{Argmin}}\; f(\mathbf{w}) + g(\mathbf{w}) + h(H\mathbf{w}), \tag{6}$$

where $H\colon \mathcal{H} \to \mathcal{G}$ denotes a bounded linear operator, $\mathcal{H}$ and $\mathcal{G}$ being two Hilbert spaces, $f\colon \mathcal{H} \to]-\infty, +\infty]$, $g\colon \mathcal{H} \to]-\infty, +\infty]$ and $h\colon \mathcal{G} \to]-\infty, +\infty]$ denote convex, l.s.c and proper functions. We additionally assume that f is differentiable and ∇f has a Lipschitz constant denoted as $\beta > 0$. Iterations are summarized in Algorithm 1 involving the proximity operator defined as

$$(\forall \mathbf{w} \in \mathcal{H}) \quad \operatorname{prox}_g(\mathbf{w}) = \arg\min_{\mathbf{u}\in\mathcal{H}} \frac{1}{2}\|\mathbf{u}-\mathbf{w}\|_2^2 + g(\mathbf{u})$$

and h^* denotes the Fenchel-Rockafellar conjugate of function h. The proximity operator of the conjugate can be computed according to the Moreau identity that is $\operatorname{prox}_{\sigma h^*}(\mathbf{w}) = \mathbf{w} - \sigma \operatorname{prox}_{h/\sigma}(\mathbf{w}/\sigma)$ for $\sigma > 0$. The sequence $(\mathbf{w}^{[k+1]})_{k\in\mathbb{N}}$ converges to a minimizer $\widehat{\mathbf{w}}$ of (6). Moreover, the sequence $(\mathbf{u}^{[k+1]})_{k\in\mathbb{N}}$ converge to a minimizer of the dual formulation of (6) that is

$$\widehat{\mathbf{u}} \in \operatorname*{Argmin}_{\mathbf{u}\in\mathcal{G}} (f+g)^*(-\mathbf{H}^*\mathbf{u}) + h^*(\mathbf{u}).$$

This algorithmic scheme can be extended for dealing with more than three functions by setting $h(Hx) = \sum_{k=1}^K h_k(H_k x)$ involving the computation of $\operatorname{prox}_{\sigma h_k^*}$. The interest of this scheme compared to ADMM is twofold: it first makes the possibility to deal with the gradient of the differentiable function and secondly it avoids to invert $\sum_k H_k^* H_k$.

Algorithm 1 Primal-dual splitting algorithm

Parameter settings: Set $\tau > 0$, $\sigma > 0$, such that $\frac{1}{\tau} - \sigma\|H\|^2 \geq \frac{\beta}{2}$
Initialization: $(\mathbf{w}^{[0]}, \mathbf{u}^{[0]}) \in \mathcal{H} \times \mathcal{G}$
For $k = 0, 1, \ldots$

$$\left\lfloor \begin{array}{l} \mathbf{w}^{[k+1]} = \operatorname{prox}_{\tau g}\big(\mathbf{w}^{[k]} - \tau\nabla f(\mathbf{w}^{[k]}) - \tau H^*\mathbf{u}^{[k]}\big) \\ \mathbf{u}^{[k+1]} = \operatorname{prox}_{\sigma h^*}\big(\mathbf{u}^{[k]} + \sigma H(2\mathbf{w}^{[k+1]} - \mathbf{w}^{[k]})\big) \end{array} \right.$$

4.1 Specificity for minimizing (5)

We first provide the iterations of Algorithm 1 specified to the minimization of (5). By setting $\mathbf{w} = (\mathbf{v}^+, \mathbf{v}^-, \boldsymbol{\Theta}) \in \mathbb{R}^N \times \mathbb{R}^N \times \mathbb{R}^{N\times N}$, for weak hierarchy, we can split it as follows:

$$\begin{cases} f(\mathbf{w}) = \frac{1}{2}\sum_{\ell=1}^L \big(y_\ell - \phi^\top(\mathbf{x}_\ell)(\mathbf{v}^+ - \mathbf{v}^-) - \phi^\top(\mathbf{x}_\ell)\boldsymbol{\Theta}\phi(\mathbf{x}_\ell)\big)^2 + \lambda 1^\top(\mathbf{v}^+ + \mathbf{v}^-), \\ g(\mathbf{w}) = \iota_O(\mathbf{v}^+) + \iota_O(\mathbf{v}^-) + \frac{\lambda}{2}\|\boldsymbol{\Theta}\|_1, \\ h(\mathbf{w}) = \sum_{i=1}^N \iota_{E_r}\big(v^{+(i)}, v^{-(i)}, \Theta^{(i,\cdot)}\big), \end{cases} \tag{7}$$

and for strong hierarchy:

$$\begin{cases} f(\mathbf{w}) = \frac{1}{2}\sum_{\ell=1}^{L}\left(y_\ell - \phi^\top(\mathbf{x}_\ell)(\mathbf{v}^+ - \mathbf{v}^-) - \phi^\top(\mathbf{x}_\ell)\boldsymbol{\Theta}\phi(\mathbf{x}_\ell)\right)^2 + \lambda 1^\top(\mathbf{v}^+ + \mathbf{v}^-), \\ g(\mathbf{w}) = \iota_O(\mathbf{v}^+) + \iota_O(\mathbf{v}^-) + \iota_S(\boldsymbol{\Theta}), \\ h_1(\mathbf{w}) = \frac{\lambda}{2}\|\boldsymbol{\Theta}\|_1, \\ h_2(\mathbf{w}) = \sum_{i=1}^{N} \iota_{E_r}\left(v^{+(i)}, v^{-(i)}, \Theta^{(i,\cdot)}\right). \end{cases} \tag{8}$$

The respective iterations are summarized in Algorithm 2 and 4. The projection onto the positive orthant and on S have well known closed form expressions [5] that are:

$$\begin{cases} P_O(\cdot) = \max\{0, \cdot\}, \\ P_S(\boldsymbol{\Theta}) = \frac{\boldsymbol{\Theta}+\boldsymbol{\Theta}^\top}{2}. \end{cases}$$

The difficulty comes from the computation of P_{E_r} whose expressions are provided in the next section.

Algorithm 2 – Weak-PD-ℓ_r – Primal-dual splitting algorithm to solve (5) when $C = \mathbb{R}^{N\times N}$

Parameter settings: Set $\beta = 2\sum_{\ell=1}^{L}|\phi(\mathbf{x}_\ell)|^2 + \sum_{\ell=1}^{L}|\phi(\mathbf{x}_\ell)^\top\phi(\mathbf{x}_\ell)|^2$, $\tau > 0$, $\sigma > 0$, such that $\frac{1}{\tau} - \sigma \geq \frac{\beta}{2}$.
Initialization: $(\mathbf{v}^{+[0]}, \mathbf{v}^{-[0]}, \boldsymbol{\Theta}^{[0]}, \mathbf{s}^{+[0]}, \mathbf{s}^{-[0]}, \boldsymbol{\Lambda}^{[0]}) \in \mathbb{R}^N \times \mathbb{R}^N \times \mathbb{R}^{N\times N} \times \mathbb{R}^N \times \mathbb{R}^N \times \mathbb{R}^{N\times N}$.
For $k = 0, 1, \ldots$

$$\begin{lfloor} b_\ell = y_\ell - \phi^\top(\mathbf{x}_\ell)(\mathbf{v}^{+[k]} - \mathbf{v}^{-[k]}) - \phi^\top(\mathbf{x}_\ell)\boldsymbol{\Theta}^{[k]}\phi(\mathbf{x}_\ell) \\ \mathbf{v}^{+[k+1]} = P_O\left(\mathbf{v}^{+[k]} + \tau\sum_\ell \phi(\mathbf{x}_\ell)b_\ell - \tau\lambda - \tau\mathbf{s}^{+[k]}\right) \\ \mathbf{v}^{-[k+1]} = P_O\left(\mathbf{v}^{-[k]} - \tau\sum_\ell \phi(\mathbf{x}_\ell)b_\ell - \tau\lambda - \tau\mathbf{s}^{-[k]}\right) \\ \boldsymbol{\Theta}^{[k+1]} = \operatorname{prox}_{\frac{\tau\lambda}{2}\|\cdot\|_1}\left(\boldsymbol{\Theta}^{[k]} + \tau\left(\sum_\ell \phi^{(i)}(\mathbf{x}_\ell)\phi^{(j)}(\mathbf{x}_\ell)b_\ell\right)_{1\leq i\leq N, 1\leq j\leq N} - \tau\boldsymbol{\Lambda}^{[k]}\right) \\ \text{For } i = 1, \ldots, N \\ \lfloor \quad (s^{+(i)[k+1]}, s^{-(i)[k+1]}, \Lambda^{(i,\cdot)[k+1]}) \\ \qquad = \operatorname{prox}_{\sigma\iota^*_{E_r}}\left(s^{+(i)[k]} + \sigma(2v^{+[k+1]} - v^{+[k]}), s^{-(i)[k]} + \sigma(2v^{-(i)[k+1]} - v^{-(i)[k]}),\right. \\ \qquad \left.\Lambda^{+(i,\cdot)[k]} + \sigma(2\Theta^{(i,\cdot)[k+1]} - \Theta^{(i,\cdot)[k]})\right) \end{lfloor}$$

4.2 New epigraphical projection

We first focus on the derivation of P_{E_∞}.

Algorithm 3 – Strong-PD-ℓ_r – Primal-dual splitting algorithm to solve (5) when $C = S$

Parameter settings: Set $\beta = 2\sum_{\ell=1}^{L} |\phi(\mathbf{x}_\ell)|^2 + \sum_{\ell=1}^{L} |\phi(\mathbf{x}_\ell)^\top \phi(\mathbf{x}_\ell)|^2$, $\tau > 0$, $\sigma > 0$, such that $\frac{1}{\tau} - 2\sigma \geq \frac{\beta}{2}$.

Initialization: $(\mathbf{v}^{+[0]}, \mathbf{v}^{-[0]}, \mathbf{\Theta}^{[0]}, \mathbf{s}^{+[0]}, \mathbf{s}^{-[0]}, \mathbf{\Lambda}_1^{[0]}, \mathbf{\Lambda}_2^{[0]}) \in \mathbb{R}^N \times \mathbb{R}^N \times \mathbb{R}^{N\times N} \times \mathbb{R}^N \times \mathbb{R}^N \times \mathbb{R}^{N\times N} \times \mathbb{R}^{N\times N}$.

For $k = 0, 1, \ldots$

$$
\begin{lfloor}
b_\ell = y_\ell - \phi^\top(\mathbf{x}_\ell)(\mathbf{v}^{+[k]} - \mathbf{v}^{-[k]}) - \phi^\top(\mathbf{x}_\ell)\mathbf{\Theta}^{[k]}\phi(\mathbf{x}_\ell) \\
\mathbf{v}^{+[k+1]} = P_O\big(\mathbf{v}^{+[k]} + \tau\textstyle\sum_\ell \phi(\mathbf{x}_\ell) b_\ell - \tau\lambda - \tau\mathbf{s}^{+[k]}\big) \\
\mathbf{v}^{-[k+1]} = P_O\big(\mathbf{v}^{-[k]} - \tau\textstyle\sum_\ell \phi(\mathbf{x}_\ell) b_\ell - \tau\lambda - \tau\mathbf{s}^{-[k]}\big) \\
\mathbf{\Theta}^{[k+1]} = P_S\big(\mathbf{\Theta}^{[k]} + \tau\big(\textstyle\sum_\ell \phi^{(i)}(\mathbf{x}_\ell)\phi^{(j)}(\mathbf{x}_\ell) b_\ell\big)_{1\leq i\leq N, 1\leq j\leq N} - \tau(\mathbf{\Lambda}_1^{[k]} + \mathbf{\Lambda}_2^{[k]})\big) \\
\mathbf{\Lambda}_1^{+[k+1]} = \mathrm{prox}_{\frac{\sigma\lambda}{2}\|\cdot\|_1^*}\big(\mathbf{\Lambda}_1^{+[k]} + \sigma(2\mathbf{\Theta}^{[k+1]} - \mathbf{\Theta}^{[k]})\big) \\
\text{For } i = 1, \ldots, N \\
\quad (s^{+(i)[k+1]}, s^{-(i)[k+1]}, \Lambda_2^{(i,\cdot)[k+1]}) \\
\qquad = \mathrm{prox}_{\sigma\iota_{E_r}^*}\big(s^{+(i)[k]} + \sigma(2v^{+[k+1]} - v^{+[k]}), s^{-(i)[k]} + \sigma(2v^{-(i)[k+1]} - v^{-(i)[k]}), \\
\qquad \Lambda_2^{+(i,\cdot)[k]} + \sigma(2\Theta^{(i,\cdot)[k+1]} - \Theta^{(i,\cdot)[k]})\big)
\end{lfloor}
$$

Proposition 4.1 *Let $\mathbf{u} = (u^{(i)})_{1\leq i\leq N}$. The projection of $(\omega^+, \omega^-, \mathbf{u}) \in \mathbb{R}\times\mathbb{R}\times\mathbb{R}^N$ on the epigraphic set E_∞ reads*

$$(\eta^+, \eta^-, \mathbf{p}) = P_{E_\infty}(\omega^+, \omega^-, \mathbf{u})$$

with

$$
\begin{cases}
\eta^- = \frac{\omega^- - (N-\bar{n}+1)(\omega^+ - \omega^-) + \sum_{i=\bar{n}}^{N} \nu^{(i)}}{1+2(N-\bar{n}+1)} \\
\eta^+ = \eta^- + \omega^+ - \omega^- \\
\mathbf{p} = P_{[-\eta^+ - \eta^-, \eta^+ + \eta^-]}(\mathbf{u})
\end{cases}
\tag{9}
$$

where $(\nu^{(1)}, \ldots, \nu^{(N)})$ is an ordered version of $(|u^{(i)}|)_{1\leq i\leq N}$ in an ascending order and with $\nu^{(0)} = -\infty$, and $\bar{n} \in \{1, \ldots, N\}$ such that $\nu^{(\bar{n}-1)} < \eta^+ + \eta^- \leq \nu^{(\bar{n})}$.

Proof 4.1 *The proof is given in Appendix B.*

The above result is obtained from arguments close to the ones derived in [15] [Proposition 5] for an epigraphical constraint of the form $\big\{(\omega, \mathbf{u}) \in \mathbb{R}\times\mathbb{R}^N \mid \max\{\tau_1|u^{(1)}|, \ldots, \tau_N|u^{(N)}|\} \leq \omega\big\}$ where $(\tau_i)_{1\leq i\leq N}$ denotes positive weights. The next proposition is a preliminary result to derive P_{E_1}.

Proposition 4.2 *Let $E_1^+ = \{(\omega^+, \omega^-, \mathbf{u}) \in \mathbb{R}\times\mathbb{R}\times\mathbb{R}^N \mid \|\mathbf{u}\|_1 = \omega^+ + \omega^-, \mathbf{u} \geq 0\}$. The projection of $(\omega^+, \omega^-, \mathbf{u}) \in \mathbb{R}\times\mathbb{R}\times\mathbb{R}^N$ on the epigraphic set E_1^+ reads*

$$(\eta^+, \eta^-, \mathbf{p}) = P_{E_1^+}(\omega^+, \omega^-, \mathbf{u})$$

with

$$\begin{cases} \eta^- = \frac{\sum_{i=1}^{\widetilde{n}} \mu^{(i)} - \omega^+ + (\widetilde{n}+1)\omega^-}{\widetilde{n}(1+2/\widetilde{n})} \\ \eta^+ = \eta^- + \omega^+ - \omega^- \\ \mathbf{p} = \mathbf{u} - \max\left\{0, \frac{\sum_{i=1}^{\widetilde{n}} \mu^{(i)} - (\eta^+ + \eta^-)}{\widetilde{n}}\right\} \end{cases} \tag{10}$$

with

$$\widetilde{n} = \max\{n \in \{1, \dots, N\} \,|,\ \mu^{(n)} - \frac{\sum_{i=1}^{n} \mu^{(i)} - (\eta^+ + \eta^-)}{n} > 0\} \tag{11}$$

and where $\boldsymbol{\mu} = (\mu^{(i)})_{1 \leq i \leq N}$ *is an ordered version of* $\mathbf{u} = (u^{(i)})_{1 \leq i \leq N}$ *in a descending order.*

Proof 4.2 *The proof is given in Appendix C.*

Next we get the solution of P_{E_1} from the ones of projection to E_1^+ according to [19][Lemma 3] and have the following proposition:

Proposition 4.3 *The projection of* $(\omega^+, \omega^-, \mathbf{u}) \in \mathbb{R} \times \mathbb{R} \times \mathbb{R}^N$ *on the epigraphic set* E_1 *reads*

$$(\eta^+, \eta^-, \mathbf{p}) = P_{E_1}(\omega^+, \omega^-, \mathbf{u})$$

with

$$\begin{cases} (v^+, v^-, \hat{\mathbf{p}}) = P_{E_1^+}(\omega^+, \omega^-, \mathrm{abs}(\mathbf{u})) \\ \mathbf{p} = \mathrm{sign}(\mathbf{u})\hat{\mathbf{p}} \end{cases} \tag{12}$$

where $\mathrm{abs}(\cdot)$ *and* $\mathrm{sign}(\cdot)$ *denote the componentwise absolute value and the componentwise sign.*

Integrating the projection derived in Proposition 4.1 into Algorithm 2 or Proposition 4.2 and Proposition 4.3 into Algorithm 4 with

$$\mathrm{prox}_{\sigma \iota^*_{E_r}}(\omega^+, \omega^-, \mathbf{u}) = (\omega^+, \omega^-, \mathbf{u}) - \sigma P_{E_r}\left(\frac{\omega^+}{\sigma}, \frac{\omega^-}{\sigma}, \frac{\mathbf{u}}{\sigma}\right)$$

ensures the convergence to a minimizer of (5) respectively for weak hierarchy and strong hierarchy.

4.3 Convergence

The proposed Algorithms 2 and 4 can be viewed as a particular case of the primal-dual algorithm [18, Algorithm 5.1] and thus the convergence guarantees can be derived from [18, Theorem 5.1]).

Theorem 4.4 *The sequences* $(\mathbf{v}^{+[k+1]}, \mathbf{v}^{-[k+1]}, \boldsymbol{\Theta}^{[k+1]})_{k\in\mathbb{N}})_{i\in\mathbb{N}}$ *generated by Algorithm 2 converges to a solution of* (5) *when* $C = \mathbb{R}^{N\times N}$ *(weak hierarchy).*

Theorem 4.5 *The sequences* $(\mathbf{v}^{+[k+1]}, \mathbf{v}^{-[k+1]}, \boldsymbol{\Theta}^{[k+1]})_{k\in\mathbb{N}})_{i\in\mathbb{N}}$ *generated by Algorithm 4 converges to a solution of* (5) *when* $C = S$ *(strong hierarchy).*

Remark 4.1 *In practice, the choice of* τ *and* σ *is made as follows:* $\tau = \frac{1.9}{\beta}$ *and* $\sigma = (1/\tau - \beta/2)$ *for Algorithm 2 (resp.* $\sigma = 1/2 * (1/\tau - \beta/2)$ *for Algorithm 4).*

5 Extension to multiclass SVM

5.1 Model

Formally, assuming that the training set is denoted

$$\mathcal{T} = \{(y_\ell, \phi(\mathbf{x}_\ell)) \in \{1,\dots,K\} \times \mathbb{R}^N \,|\, \ell \in \{1,\dots,L\}\}$$

with K classes and L training samples. $\mathbf{x}_\ell \in \mathbb{R}^M$ denotes the data (e.g. an image with M pixels) with the label y_ℓ. The transform $\phi\colon \mathbb{R}^M \to \mathbb{R}^N$ maps from the input space onto an arbitrary feature space, for instance, scattering features [9].

Our aim being to study quadratic interactions, for the k-th class, we consider a discrimination function taking the form:

$$f_k(\mathbf{x}_\ell) = \phi(\mathbf{x}_\ell)^\top \boldsymbol{\Theta}_k \phi(\mathbf{x}_\ell) + \mathbf{v}_k^\top \phi(\mathbf{x}_\ell) \tag{13}$$

where, $\mathbf{v}_k = \left(v_k^{(i)}\right)_{1\le i\le N}$ and $\boldsymbol{\Theta}_k = \left(\Theta_k^{(i,j)}\right)_{1\le i\le N, 1\le j\le N}$ model respectively the weights of main features and the matrix of interactions for the class $k \in \{1,\dots,K\}$.

5.2 Design of the objective function

A large panel of data term can be encountered in the multiclass SVM literature going from hinge loss to logistic data-term [8, 16]. In this work, we focus on the squared hinge loss data-term proposed in [8] leading to good classification performance and being differentiable with Lipschitz gradient $\beta > 0$, whose constant will be specified in Section 5.4. The proposed objective function is written as:

$$\underset{\substack{\mathbf{v}_k\in\mathbb{R}^N \\ \boldsymbol{\Theta}_k\in\mathbb{R}^{N\times N}}}{\text{minimize}} \sum_{\ell=1}^{L} \sum_{k\neq y_\ell} \max\left\{0, 1 - \left(f_{y_\ell}(\mathbf{x}_\ell) - f_k(\mathbf{x}_\ell)\right)\right\}^2 + \sum_{k=1}^{K} \Omega(\mathbf{v}_k, \boldsymbol{\Theta}_k) \tag{14}$$

where the first term denotes the squared hinge loss data-term and the second term penalizes the behavior of the discrimination function. Instead of ℓ_1-norm, making $\mathbf{v}_k$ and $\mathbf{\Theta}_k$ independent, we can force the hierarchy constraints between $\mathbf{v}_k$ and $\mathbf{\Theta}_k$ for k-th class to regulate them. In the classification setting the penalization term is written:

$$\Omega(\mathbf{v}_k, \mathbf{\Theta}_k) = \lambda_1 \sum_{i=1}^{N} \max\{|v_k^{(i)}|, \|\Theta_k^{(i,\cdot)}\|_r\} + \lambda_2 \|\mathbf{\Theta}_k\|_1 + \iota_C(\mathbf{\Theta}_k) \tag{15}$$

where $r = \{1, +\infty\}$, $\Theta_k^{(i,\cdot)} = \left(\Theta_k^{(i,j)}\right)_{1 \leq j \leq N}$ and ι_C is defined as in Section 2.

5.3 Epigraphical reformulation

The reformulation of problem (14) by means of epigraphical constraints [3] follows ideas derived in [7] (i.e. "hierNet") in the context of regression and for a different penalization Ω.

Proposition 5.1 *The minimization problem* (14)

with $\Omega(\mathbf{v}_k, \Theta_k)$ *defined in* (15) *can be reformulated as*

$$\begin{aligned} \underset{(\mathbf{v}_k^+, \mathbf{v}_k^-, \Theta_k) \in O \times O \times C}{\textit{minimize}} & \sum_{\ell=1}^{L} \sum_{k \neq y_\ell} \max\left\{0, 1 - \left(f_{y_\ell}(\mathbf{x}_\ell) - f_k(\mathbf{x}_\ell)\right)\right\}^2 \\ & + \sum_{k=1}^{K} \Big(\lambda_1 1^\top (\mathbf{v}_k^+ + \mathbf{v}_k^-) + \lambda_2 \sum_{i=1}^{N} \|\Theta_k^{(i,\cdot)}\|_1 \\ & + \sum_{i=1}^{N} \iota_{E_r}\left(v_k^{+(i)}, v_k^{-(i)}, \Theta_k^{(i,\cdot)}\right) + \iota_C(\mathbf{\Theta}_k) \Big) \end{aligned} \tag{16}$$

where O *denotes the positive orthant and* $E_r = \{(\omega^+, \omega^-, \mathbf{u}) \in \mathbb{R} \times \mathbb{R} \times \mathbb{R}^N \mid \|\mathbf{u}\|_r \leq \omega^+ + \omega^-\}$ *with* $r = \{1, +\infty\}$ *and* $f_k(\mathbf{x}_\ell) = \phi(\mathbf{x}_\ell)^\top \mathbf{\Theta}_k \phi(\mathbf{x}_\ell) + (\mathbf{v}_k^+ - \mathbf{v}_k^-)^\top \phi(\mathbf{x}_\ell)$.

The variable $\mathbf{v}_k$ is written as $\mathbf{v}_k = \mathbf{v}_k^+ - \mathbf{v}_k^-$ by variable splitting. The proof relies on similar arguments than for Proposition 3.1.

[3] The epigraph of a function $f \colon \mathcal{H} \to]-\infty, +\infty]$ is defined as: $\mathrm{epi} f = \{(v, \zeta) \in \mathcal{H} \times \mathbb{R} \mid f(v) \leq \zeta\}$ and the epigraphical projection onto $\mathrm{epi} f$ is denoted $P_{\mathrm{epi} f}$.

5.4 Primal-dual proximal algorithm based on epigraphical projection

The proposed objective function (5) is a specific case of (6) when $\mathbf{w} = (\mathbf{v}_k^+, \mathbf{v}_k^-, \mathbf{\Theta}_k) \in \mathbb{R}^N \times \mathbb{R}^N \times \mathbb{R}^{N\times N}$, $Q = 2$ and

$$\begin{cases} f(\mathbf{w}) = \sum_{\ell=1}^{L} \sum_{k\neq y_\ell} \max\Big\{0, 1 - \big(f_{y_\ell}(\mathbf{x}_\ell) - f_k(\mathbf{x}_\ell)\big)\Big\}^2 \\ \qquad\qquad + \sum_{k=1}^{K} \lambda_1 1^\top(\mathbf{v}_k^+ + \mathbf{v}_k^-) \\ g(\mathbf{w}) = \sum_{k=1}^{K} \iota_O(\mathbf{v}_k^+) + \iota_O(\mathbf{v}_k^-) + \iota_S(\Theta_k) \\ h_1(\mathbf{w}) = \sum_{k=1}^{K} \lambda_2 \|\mathbf{\Theta}_k\|_1 \\ h_2(\mathbf{w}) = \sum_{k=1}^{K} \sum_{i=1}^{N} \iota_{E_r}\big(v_k^{+(i)}, v_k^{-(i)}, \Theta_k^{(i,\cdot)}\big). \end{cases}$$

where f is differentiable with a Lipschitz constant $\beta = 4(K-1)(\sum_\ell \|\phi(\mathbf{x}_\ell)\|^2 + \sum_\ell \|\phi(\mathbf{x}_\ell)\phi(\mathbf{x}_\ell)^\top\|^2)$. The primal-dual proximal iterations are displayed in Algorithm 4. It involves four proximity operator computation: the projection onto O and S, the proximal operator of ℓ_1-norm and the epigraphical projection onto E_r, whose closed form expression are derived in Section 4.

Algorithm 4 Primal-dual proximal algorithm based on epigraphical projection for classification when $C = S$

Parameter settings: Set $\tau > 0$, $\sigma > 0$, such that $2(1/\tau - \sigma) \geq \beta$.
Initialization: $(\mathbf{v}_k^{+[0]}, \mathbf{v}_k^{-[0]}, \mathbf{\Theta}_k^{[0]}, \mathbf{s}_k^{+[0]}, \mathbf{s}_k^{-[0]}, \mathbf{\Lambda}_{k,1}^{[0]}, \mathbf{\Lambda}_{k,2}^{[0]}) \in \mathbb{R}^N \times \mathbb{R}^N \times \mathbb{R}^{N\times N} \times \mathbb{R}^N \times \mathbb{R}^N \times \mathbb{R}^{N\times N} \times \mathbb{R}^{N\times N} \quad \forall k = \{1, \ldots, K\}$.
For $t = 0, 1, \ldots$

$$\begin{aligned} & f(\mathbf{v}^+, \mathbf{v}^-, \mathbf{\Theta}) = \sum_{\ell=1}^{L} \sum_{k\neq y_\ell} \max\Big\{0, 1 - \big(\phi(\mathbf{x}_\ell)^\top \mathbf{\Theta}_{y_\ell} \phi(\mathbf{x}_\ell) + (\mathbf{v}_{y_\ell}^+ - \mathbf{v}_{y_\ell}^-)^\top \phi(\mathbf{x}_\ell) - \\ & \qquad\qquad \phi(\mathbf{x}_\ell)^\top \mathbf{\Theta}_k \phi(\mathbf{x}_\ell) - (\mathbf{v}_k^+ - \mathbf{v}_k^-)^\top \phi(\mathbf{x}_\ell)\big)\Big\}^2 + \sum_{k=1}^{K} \lambda_1 1^\top(\mathbf{v}_k^+ + \mathbf{v}_k^-); \\ & \mathbf{v}_k^{+[t+1]} = P_O\big(\mathbf{v}_k^{+[t]} - \tau \nabla_{\mathbf{v}_k^+} f(\mathbf{v}_k^{+[t]}, \mathbf{v}_k^{-[t]}, \mathbf{\Theta}_k^{[t]}) - \tau \mathbf{s}_k^{+[t]}\big) \quad \forall k = 1, \ldots, K \\ & \mathbf{v}_k^{-[t+1]} = P_O\big(\mathbf{v}_k^{-[t]} - \tau \nabla_{\mathbf{v}_k^-} f(\mathbf{v}_k^{+[t]}, \mathbf{v}_k^{-[t]}, \mathbf{\Theta}_k^{[t]}) - \tau \mathbf{s}_k^{-[t]}\big) \quad \forall k = 1, \ldots, K \\ & \mathbf{\Theta}_k^{[t+1]} = P_S\big(\mathbf{\Theta}_k^{[t]} - \tau \nabla_{\Theta_k} f(\mathbf{v}_k^{+[t]}, \mathbf{v}_k^{-[t]}, \mathbf{\Theta}_k^{[t]}) - \tau(\mathbf{\Lambda}_{k,1}^{[t]} + \mathbf{\Lambda}_{k,2}^{[t]})\big) \quad \forall k = 1, \ldots, K \\ & \mathbf{\Lambda}_{k,1}^{+[t+1]} = \mathrm{prox}_{\sigma\lambda_2 \|\cdot\|_1^*}\big(\mathbf{\Lambda}_{k,1}^{+[t]} + \sigma(2\mathbf{\Theta}_k^{[t+1]} - \mathbf{\Theta}_k^{[t]})\big) \quad \forall k = 1, \ldots, K \\ & \text{For } k = 1, \ldots, K, \quad i = 1, \ldots, N \\ & \lfloor \quad (s_k^{+(i)[t+1]}, s_k^{-(i)[t+1]}, \Lambda_{k,2}^{(i,\cdot)[t+1]}) \\ & \qquad = \mathrm{prox}_{\sigma\iota_{E_r}^*}\big(s_k^{+(i)[t]} + \sigma(2v_k^{+[t+1]} - v^{+[t]})_k, s_k^{-(i)[t]} + \sigma(2v_k^{-(i)[t+1]} - v_k^{-(i)[t]}), \\ & \qquad\qquad\qquad\qquad \Lambda_{k,2}^{+(i,\cdot)[t]} + \sigma(2\Theta_k^{(i,\cdot)[t+1]} - \Theta_k^{(i,\cdot)[t]})\big) \end{aligned}$$

6 Experiments

In this section, we firstly provide a comparison between ℓ_1 and ℓ_∞ proposed approaches with the two closest state-of-the-art procedures, that are "hierNet" [7] and Jenatton's framework [24], on a simulation dataset into regression framework. Comparisons are of two types, in terms of convergence behavior and in term of performing estimation.

Second, we apply the regression algorithms to HIV disease analysis, which is meaningful to the prevention of the disease. Third, classification experiments are conducted in the context of Parkinson disease classification and face classification. The proposed algorithms and the comparison approaches are implemented in Matlab on a computer with AMD AthlonX4 750 processor.

6.1 Simulated data

6.1.1 Dataset

The dataset is created according to [23]. It is initially composed of N main features and of $N(N-1)/2$ interactions (due to symmetry). We denote $(\overline{\mathbf{v}}, \overline{\mathbf{\Theta}}) \in \mathbb{R}^N \times \mathbb{R}^{N \times N}$ the ground truth generated according to strong hierarchy. $\overline{\mathbf{v}}$ denotes a sparse vector where the non-zero values are associated with randomly selected indexes. The value for the non-zero $\overline{v}^{(i)}$ is randomly selected from the set $\{-5, -4, \ldots, -1, 1, \ldots, 5\}$. Due to strong hierarchical constraint, $\overline{\mathbf{\Theta}}$ is only non-zero for those whose main effects are not zero. The values are randomly chosen from the set $\{-10, -8, \ldots, -2, 2, \ldots, 8, 10\}$.

The algorithmic performance is evaluated on two simulated datasets. The first one is composed of $N = 30$ features, the first 10 main features are non-zeros and the interaction ratio is $\rho = 3.45\%$ (Dataset30-345). The second dataset is of size $N = 100$, the first 30 features are non-zeros and the interaction ratio is $\rho = 0.30\%$ (Dataset100-030). The interaction ratio is calculated as the non-zero interaction number divided by the possible interaction number in $\overline{\mathbf{\Theta}}$ (i.e. $N(N-1)/2$).

The datasets are composed of a training, a validation and a testing set with 100 samples each. $\phi(\mathbf{x}_\ell)$ is randomly generated according to normal distribution $\mathcal{N}(0, \mathbb{I}_N)$. The observed value for each sample is set by $\mathbf{y}_\ell = \phi(\mathbf{x})_\ell^\top \overline{\mathbf{v}} + \phi(\mathbf{x}_\ell)^\top \overline{\mathbf{\Theta}} \phi(\mathbf{x}_\ell) + \varepsilon_\ell$, where ε_ℓ is independent Gaussian noise to make the signal-to-noise ratio approximately equal to 5dB.

6.1.2 Comparison with [24] – $r = \infty$ and weak hierarchy

In order to provide fair comparison with the work in [24], we first investigate the performance of proposed primal-dual proximal algorithm (Algorithm 2) when $r = \infty$ and in the configuration of weak hierarchy (no symmetry constraint is considered, i.e. $C = \mathbb{R}^{N\times N}$): "Weak-PD-$\ell_\infty$". The algorithmic procedure designed in [24] is based on "FISTA" and it is named "Weak-FISTA-ℓ_∞", where we use the proximity operator of tree-structured variables in "SPAMS" toolbox [24], it is worthy mentioning that the core function is written in C++, while the whole implementation of our proposed epigraphical primal-dual algorithm is implemented in MATLAB. Both are compared in two respects: the convergence of the objective function Eq. (5) and convergence of the iterates (i.e. $\|\mathbf{w}^{[k]} - \mathbf{w}^{[+\infty]}\|$).

The comparisons are displayed in Fig. 2 (for Dataset30-345) and in Fig. 3 (for Dataset100-030) for the optimal λ (cf. Section 6.1.4). The convergence behaviour are presented both w.r.t. iterations (first row) and time (second row). It is observed that:

i) "Weak-PD-ℓ_∞" and "Weak-FISTA-ℓ_∞" converge to the same value as shown in the first column of both Fig. 2 and 3;

ii) From the objective function convergence point-of-view, "Weak-FISTA-ℓ_∞" converges faster than 'Weak-PD-ℓ_∞" w.r.t. iteration numbers and also time;

iii) From the convergence of the iterates point of view (second column of Fig. 2 and 3), for Dataset30-345, "Weak-PD-ℓ_∞" converges much faster than "Weak-FISTA-ℓ_∞" to the optimal solution, especially from the comparison in terms of time (bottom figures). For Dataset100-030, although it seems that the convergence in iteration for "Weak-PD-ℓ_∞" slower, but it costs less time, this may because that the default values for τ and σ are not optimal for Dataset100-030. It shows that the proposed algorithm enables to be closest to the optimal solution with less time;

iv) Contrary to"Weak-FISTA-ℓ_∞", our proposed solution can provide a 'Strong-PD-ℓ_∞" counterpart of this 'Weak-PD-ℓ_∞" without inner iterations.

Further results, especially w.r.t the choice of λ and regression performance on the training/validation/test sets are provided in Section 6.1.4.

6.1.3 Comparison with [7] – $r = 1$ and strong hierarchy

Similar experiments are conducted for $r = 1$ and strong hierarchy (i.e., $C = S$). The proposed algorithm (Algorithm 4 with $r = 1$) is called "Strong-PD-ℓ_1" and it is compared with "hierNet" whose iterations are derived from an ADMM scheme, named"Strong-ADMM-ℓ_1",

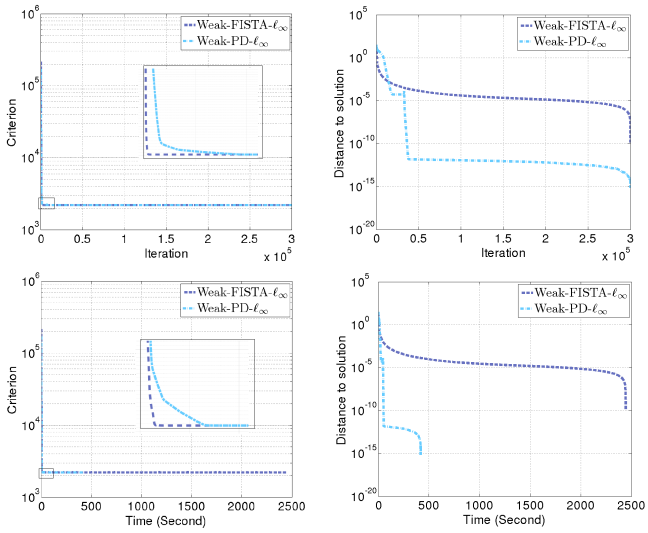

Figure 2: Comparison between the proposed "Weak-PD-ℓ_∞" and "Weak-FISTA-ℓ_∞" on Dataset30-345 for $\lambda = 14$. (top-left) Objective function in (3) w.r.t. iterations, (bottom-left) Objective function in w.r.t. time, (top-right) Distance $\|\mathbf{w}^{[k]} - \mathbf{w}^{[\infty]}\|$ w.r.t. iterations, (bottom-right) Distance $\|\mathbf{w}^{[k]} - \mathbf{w}^{[\infty]}\|$ w.r.t. time.

it is noted that the hyperparameter in the inner iteration for ADMM is set to be the default value (i.e. 1) in the following experiments.

The comparisons between these two schemes are displayed in Fig. 4 (for Dataset30-345) and Fig. 5 (for Dataset100-030) for the optimal λ (cf. Section 6.1.4). Both are compared in three respects: the convergence of the objective function Eq. (3), the convergence of the iterates (i.e. $\|\mathbf{w}^{[k]} - \mathbf{w}^{[+\infty]}\|$), and the distance to the set S. These quantities are displayed w.r.t. iteration number and time. It is observed that:

i) "Strong-PD-ℓ_1" and "Strong-ADMM-ℓ_1" converge to the same solution, but the convergence of the objective to the optimum solution is very sensitive to the trade-off hyper-parameter for "Strong-ADMM-ℓ_1";

ii) "Strong-PD-ℓ_1" is always faster either in iteration number or in time and either in terms of functional or in terms of iterations. The explanation mainly comes from the inner iterations required with "Strong-ADMM-ℓ_1" when dealing with $C = S$;

iii) Third column of Fig. 4 and 5 highlights that the constraint violations (i.e. distance to S) with the proposed method is always smaller than with ADMM;

Further results, especially w.r.t the choice of λ and regression performance on the training/validation/test sets are provided in Section 6.1.4.

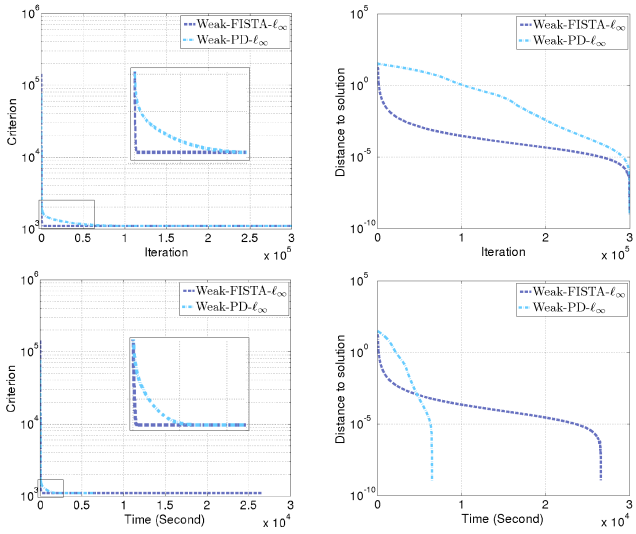

Figure 3: Comparison between the proposed "Weak-PD-ℓ_∞" and "Weak-FISTA-ℓ_∞ on Dataset100-030 for $\lambda = 8$. (top-left) Objective function (3) w.r.t. iterations, (bottom-left) Objective function (3) w.r.t. time, (top-right) Distance $\|\mathbf{w}^{[k]} - \mathbf{w}^{[\infty]}\|$ w.r.t. iterations, (bottom-right) Distance $\|\mathbf{w}^{[k]} - \mathbf{w}^{[\infty]}\|$ w.r.t. time.

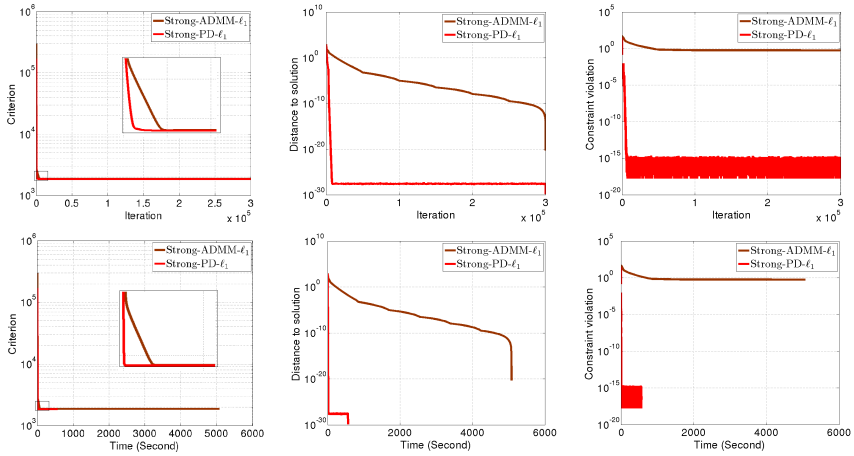

Figure 4: Comparison between the proposed "Strong-PD-ℓ_1" and "Strong-ADMM-ℓ_1" on Dataset30-345 for $\lambda = 8$. (top-left) Objective function (3) w.r.t. iterations, (bottom-left) Objective function (3) w.r.t. time, (top-middle) Distance $\|\mathbf{w}^{[k]} - \mathbf{w}^{[\infty]}\|$ w.r.t. iterations, (bottom-middle) Distance $\|\mathbf{w}^{[k]} - \mathbf{w}^{[\infty]}\|$ w.r.t. time, (top-right) Distance to the strong hierarchy constraint S w.r.t. iterations, (bottom-right) Distance to the strong hierarchy constraint S w.r.t. time.

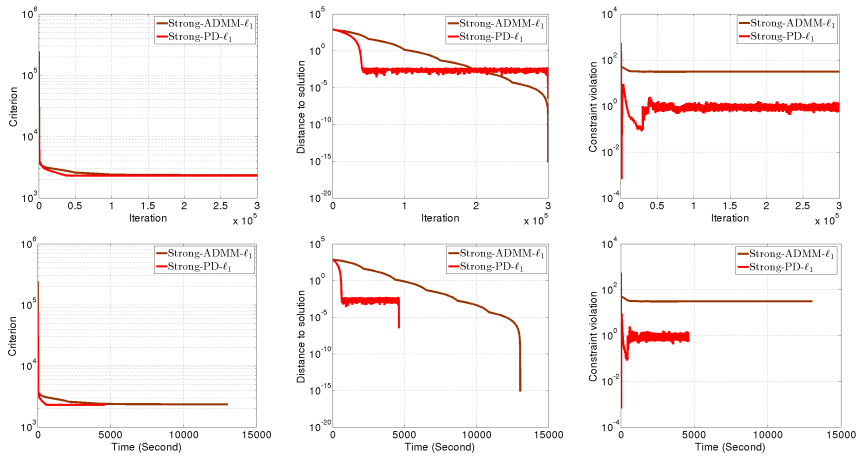

Figure 5: Comparison between the proposed "Strong-PD-ℓ_1" and "Strong-ADMM-ℓ_1" on Dataset100-030 for $\lambda = 10$. (top-left) Objective function (3) w.r.t. iterations, (bottom-left) Objective function (3) w.r.t. time, (top-middle) Distance $\|\mathbf{w}^{[k]} - \mathbf{w}^{[\infty]}\|$ w.r.t iterations, (bottom-middle) Distance $\|\mathbf{w}^{[k]} - \mathbf{w}^{[\infty]}\|$ w.r.t. time, (top-right) Distance to the strong hierarchy constraint S w.r.t. iterations, (bottom-right) Distance to the strong hierarchy constraint S w.r.t. time.

6.1.4 Choice of λ

In order to select the optimal λ, we perform 5 simulations and compute the average performance. Hold-out validation method is adopted for the selection of the trade-off parameter λ, where the models with different λ are trained on the training set, and then the best λ is selected based on the performance on the validation set, finally we give the performance of the model on the test set with the selected λ. We also compare the proposed algorithms to other two common methods:

- SVR-ℓ_2: it attempts to minimize the criteria with an empirical loss and a ℓ_2-norm on the weights, aiming to learn small weights, which is expressed as:

$$\underset{\mathbf{w}\in\mathbb{R}^{N+N\times N}}{\text{minimize}}\, \lambda \sum_{\ell=1}^{L} \Big(y_\ell - [\phi(\mathbf{x}_\ell),\ \phi(\mathbf{x}_\ell)^\top \phi(\mathbf{x}_\ell)]^\top \mathbf{w} \Big)^2 + \frac{1}{2}\|\mathbf{w}\|^2 \tag{17}$$

  We apply LIBSVM library [13] to solve this problem in the dual form with linear kernels on the training samples. In our experiments, λ is validated from 10 to 1000 with a space 10, and the results with the best selected λ are shown in Tab. 1.

- SVR-ℓ_1: it integrates the prior knowledge of sparsity to the criteria and aims to learn

a set of sparse weights, which can be written as:

$$\underset{\mathbf{w}\in\mathbb{R}^{N+N\times N}}{\text{minimize}} \sum_{\ell=1}^{L} \frac{1}{2}\Big(y_\ell - [\phi(\mathbf{x}_\ell),\ \phi(\mathbf{x}_\ell)^\top\phi(\mathbf{x}_\ell)]^\top\mathbf{w}\Big)^2 + \lambda\|\mathbf{w}\|_1 \tag{18}$$

We apply forward-backward proximal algorithm [17] to solve it. In our experiments, λ is validated in $[10^{-6}, 10^{-5}, \ldots, 1, 2, 4, \ldots, 2^5]$, and the results with the best selected λ are shown in Tab. 1.

The average performance in mean square errors (MSE) for different interaction models with different λ on validation set are shown in Fig. 6, and their performance with best selected λ on the three sets are given in Tab. 1. From the performance point of view in the Fig. 6 and Tab. 1, it can be seen that: i) The performance SVR-ℓ_2 are the worst, which is not surprise because the penalization term aims to learn small values, rather than sparse ones. SVR-ℓ_1 works better and sparse weights are learned, which meets the sparse property of interactions, but it is still inferior to strong hierarchy with $r = 1$; ii) "Weak-PD-ℓ_∞" and "Weak-FISTA-ℓ_∞" almost have the same performance for different λ on both simulation datasets; iii) For Dataset30-345, "Strong-ADMM-ℓ_1" have similar performance with "Strong-PD-ℓ_1", but due to sophisticated hyper-parameter selection in ADMM, fluctuations over different λ are observed, especially for large λ values; for Dataset100-030, there are some gaps between different λ, and the performance of "Strong-ADMM-ℓ_1" deteriorates seriously when λ becomes large. We conjecture that ADMM would become sensitive to the hyper-parameter selection when the number of features increases.

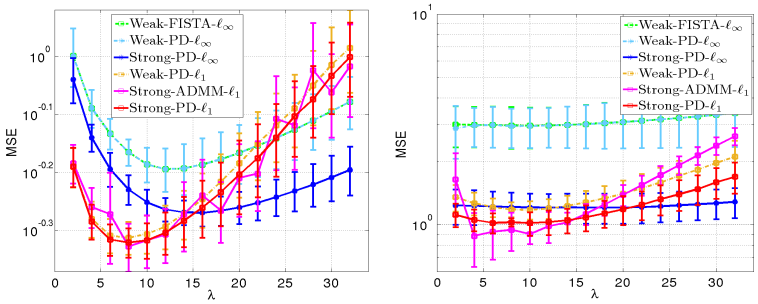

Figure 6: The performance in MSE on the validation set obtained with the six algorithms ("Weak-FISTA-ℓ_∞", "Weak-PD-ℓ_∞", "Strong-PD-ℓ_∞, "Weak-PD-ℓ_1","Strong-ADMM-ℓ_1 and "Strong-PD-ℓ_1) allowing to compare properly the different configurations of the regularization term (4). (left) Dataset30-345 (right) Dataset100-030).

Data	Approach	λ	TR	VAL	TE
Dataset30-347	SVR-ℓ_2	120	81.698±43.091	1046.957±135.988	1064.482±264.993
	SVR-ℓ_1	0.01	0.117±0.013	0.520±0.038	0.500±0.036
	Weak-FISTA-ℓ_∞	14	0.166±0.017	0.642±0.065	0.602±0.057
	Weak-PD-ℓ_∞	14	0.166±0.017	0.642±0.065	0.602±0.057
	Strong-PD-ℓ_∞	16	0.190±0.017	0.538±0.029	0.527±0.048
	weak-PD-ℓ_1	8	0.198±0.017	0.487±0.032	0.481±0.038
	Strong-ADMM-ℓ_1	8	0.184±0.023	0.471±0.045	0.471±0.061
	Strong-PD-ℓ_1	8	0.196±0.017	0.478±0.028	0.473±0.041
Dataset100-030	SVR-ℓ_2	90	96.217±61.627	1843.295±62.330	1870.404±174.944
	SVR-ℓ_1	0.1	1.496±0.191	2.918±1.596	2.831±1.255
	Weak-FISTA-ℓ_∞	10	0.032±0.001	2.954±0.649	3.049±0.749
	Weak-PD-ℓ_∞	2	0.001±0.0002	2.867±0.808	2.999±0.943
	Strong-PD-ℓ_∞	18	0.101±0.004	1.034±0.245	1.025±0.285
	weak-PD-ℓ_1	10	0.109±0.006	1.050±0.217	1.062±0.251
	Strong-ADMM-ℓ_1	4	0.025±0.001	0.884±0.252	0.814±0.212
	Strong-PD-ℓ_1	10	0.113±0.006	0.918±0.179	0.929±0.199

Table 1: The comparison results (MSE) on the train, validation and test set of different algorithms under best selected λ when both datasets are used.

6.1.5 Discussion regarding the choice of r

From the above algorithmic comparisons, we have observed that the proposed method, either for ℓ_1 or ℓ_∞, delivers an accurate solution (as other algorithmic strategies) but faster and with the possibility to include the strong-hierarchy constraint without inner iterations. In the following analysis, we will thus focus on comparisons between weak/strong ℓ_1 or ℓ_∞ using the proposed algorithm ("Weak-PD-ℓ_∞" , "Weak-PD-ℓ_1" , "Strong-PD-ℓ_∞" , "Strong-PD-ℓ_1"). The average performance with the associated variances are presented in Fig. 6 for different values of the regularization parameter λ. Moreover, the performance obtained by hold-out validation are presented in Tab. 1. It can be observed that:

i) The good behavior of the strong hierarchy constraint clearly appears for both datasets. Indeed, we recall that the data have been created with strong hierarchical structure and we can clearly observe that for both datasets the performances with the strong hierarchical constraint are always better either for $r = 1$ or $r = +\infty$;

ii) In our set of experiments, the regularization with $r = 1$ always leads to better performance than with $r = +\infty$;

iii) For "Weak-PD-ℓ_∞", MSE values associated with the Dataset100-030 are larger, probably due to overfitting. This assumption can be validated by the results obtained on the training dataset.

6.2 Application to HIV Data

Feature interaction learning is very meaningful to discover the intrinsic correlations between the variables, especially in the real-life applications, for example, genetic diseases and drug analysis. Since the effects usually depends on several genes or factors, the interaction of these associating factors is worth studying. In this section, we apply the proposed algorithms to drug analysis: HIV dataset on the susceptibility of the HIV-1 virus to six nucleoside reverse transcriptase inhibitors (NRTIs). This dataset[4] was collected by [32] and it was used to model HIV-1 susceptibility to the drugs because HIV-1 virus can become resistant through genome mutations for different subjects. In the dataset, there are 639 subjects and 240 gene locations. For each observation, the mutation status at each gene location is recorded and also give six drugs (log) susceptibility responses. In this work, we focus on the prediction model for 3TC drug.

The dataset is randomly split into two half sets for training and test, respectively as adopted by [7, 23]. The trade-off parameter λ is selected similarly by hold-out validation method. The performance are measured by the average MSE with the associating variance over 5 random splits on the test set for "Weak-PD-ℓ_∞", "Strong-PD-ℓ_∞", "Weak-PD-ℓ_1", "Strong-PD-ℓ_1", "Weak-FISTA-ℓ_∞" and "Strong-ADMM-ℓ_1".

6.2.1 Reduced dataset

Following [23], we first work on a reduced dataset over the bins of ten adjacent loci, rather than all 240 genes locations, resulting from the fact that nearby genomes have similar effects to the drug susceptibility, and also leading to less sparse data. So for the first set of experiments we have $N = 24$ features. The value for each bin is set to 1 if one of genes in that bin undergoes mutation.

Fig. 7 (left) shows the performance of six algorithms on the reduced dataset. It can be observed that: i) "Weak-FISTA-ℓ_∞ has the same performance with "Weak-PD-ℓ_∞ over different λ; ii) a slight better performance are obtained for "Strong-PD-ℓ_1" compared to "Strong-ADMM-ℓ_1"; iii) "Strong-PD-ℓ_∞" and "Strong-PD-ℓ_1" are slightly better than their weaker counterparts, and the best λ for "Strong-PD-ℓ_∞" and "Strong-PD-ℓ_1" are 12 and 10 respectively. The interactions over one split are visualized in the Fig. 8 (left). In order to better visualize the effect of the penalizations, we also plot the ℓ_∞-norm and ℓ_1-norm over each row of Θ for strong hierarchy, as shown in the middle and right of Fig. 8. Both have the strong effect for 19^{th} feature, which is consistent with the results observed in [23]. We also observe an interesting fact from the visualization of interactions, it can be seen that "Strong-PD-ℓ_1" is able to learn a sparser interaction matrix than "Strong-PD-ℓ_∞", however, the maximum strength (i.e. ℓ_∞-norm) and the sum of the strength (ℓ_1-norm) in Fig. 8 have

[4]It can be downloaded from https://hivdb.stanford.edu/pages/published_analysis/genophenoPNAS2006/

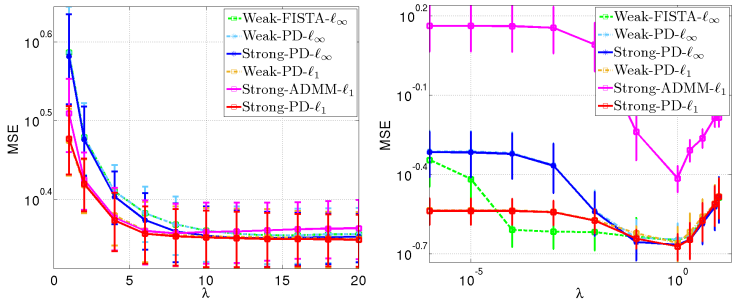

Figure 7: Performance comparisons between "Weak-FISTA-ℓ_∞", "Weak-PD-ℓ_∞", "Strong-PD-ℓ_∞, "Weak-PD-ℓ_1","Strong-ADMM-ℓ_1 and "Strong-PD-ℓ_1 for different λ values on the test set of HIV dataset when $N = 24$ (left) and $N = 240$ (right).

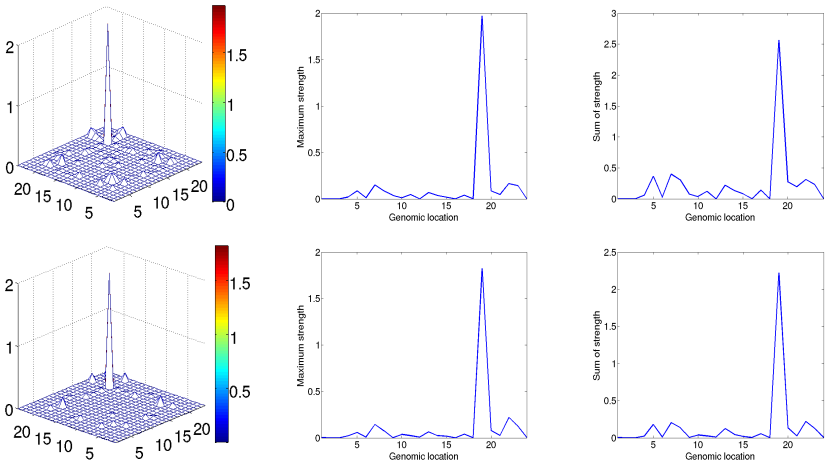

Figure 8: The figures in the first row shows the interaction (left column), ℓ_∞-norm over each row of interactions (middle column), ℓ_1-norm over each row of interaction (right column) when $N = 24$, $r = \infty$ and $\lambda = 12$. The figures in the second row shows the corresponding results when when $N = 24$, $r = 1$ and $\lambda = 10$.

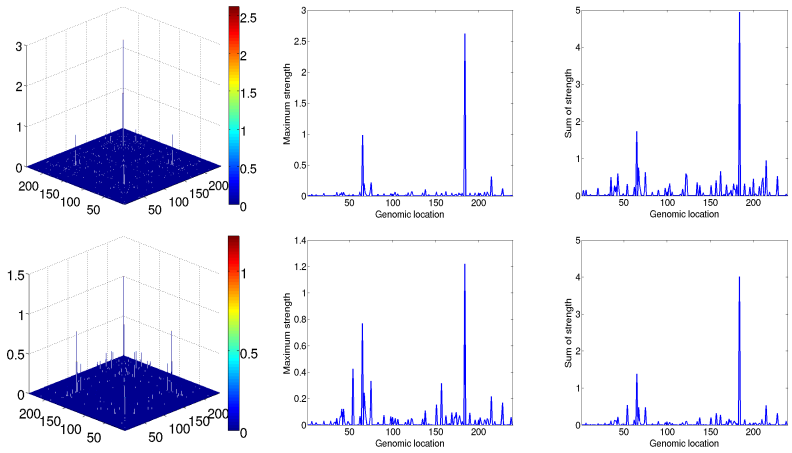

Figure 9: The figures in the first row shows the interaction (left column), ℓ_∞-norm over each row of interactions (middle column), ℓ_1-norm over each row of interactions (right column) when $N = 240$, $r = \infty$ and $\lambda = 1$. The figures in the second row shows the corresponding results when when $N = 240$, $r = 1$ and $\lambda = 1$.

similar distributions for both cases. The possible reason is that the extra interactions learned by "Strong-PD-ℓ_∞" are subtle, it can be also confirmed that both performances are very similar in the Fig. 7 (left).

6.2.2 Full dataset

We further work on the full data without binning operation leading to $N = 240$ features and each feature gives the indicator of mutation. The data splitting is the same as the one previously described. The average performance of MSE on the test set for six algorithms are shown in Fig. 7 (right). It is clear to demonstrate that: i) on large dataset, there are some performance gaps between FISTA, ADMM and their primal-dual counterparts for weak and strong hierarchy over different λ, especially ADMM deteriorates seriously and the performance becomes worse; ii) strong hierarchy still produces better performance than the weak one. The learned interactions and their feature effects from "Strong-PD-ℓ_∞" and "Strong-PD-ℓ_1" with best $\lambda = 1$ over one split are shown in Fig. 9. It is found that both algorithms can detect that 184^{th} genes location has the most strong effect. From the maximum strength and sum of strength for each gene in the Fig. 9, we can see the different properties for both algorithms. For "Strong-PD-ℓ_∞", maximum strength for each gene has a more sparse distribution than the sum of strength, which is consistent with its objective that the maximum of absolute values of $\Theta^{(i,:)}$ is not larger that $v^{(i)}$, whereas a more sparse distribution over the sum of strength is obtained for "Strong-PD-ℓ_1", whose objective is to ensure that the sum of absolute values of $\Theta^{(i,:)}$ is not larger that $v^{(i)}$. From Fig. 7(right), it also can be observed that "Strong-PD-ℓ_1" is slightly better than "Strong-PD-ℓ_∞" because

it imposes a stronger penalization for $\mathbf{\Theta}$ than "Strong-PD-ℓ_∞".

6.3 Multiclass learning

In [25], we performed preliminary validation of the proposed method compared to state-of-the-art in the context of face classification. In this work, we tackle Parkinson disease classification from the speech records of the persons. Thus, $\mathbf{x}_\ell$ models the speech data (1D) and $\phi(\mathbf{x}_\ell)$ the time-frequency based features. For all the experiments, we compare the performance in term of accuracy between:

- sparse multiclass SVM without interaction, i.e., $f_k(\mathbf{x}_\ell) = \mathbf{v}_k^\top \phi(\mathbf{x}_\ell)$ and $\Omega(\mathbf{v}_k) = \sum_k \|\mathbf{v}_k\|_1$,
- sparse multiclass SVM with full interactions: $f_k(\mathbf{x}_\ell) = \phi(\mathbf{x}_\ell)^\top \mathbf{\Theta}_k \phi(\mathbf{x}_\ell) + \mathbf{v}_k^\top \phi(\mathbf{x}_\ell)$ and $\Omega(\mathbf{v}_k, \mathbf{\Theta}_k) = \|\mathbf{v}_k\|_1 + \sum_i \|\mathbf{\Theta}_k^{(i,\cdot)}\|_1$,
- proposed approach that is sparse multiclass SVM with strong hierarchy.

The full interaction case differs from the proposed one in the sense that there is no coupling between $\mathbf{v}_k$ and Θ_k in Eq. (5).

Dataset – We first show the Parkinson disease classification by sparse multiclass SVM with strong hierarchy. The experiments are conducted on the Parkinson Speech dataset [34], which is used to study the underlying relationship between the Parkinson disease and the types of voices of the patients. This dataset consists of training set and test set. For the training set, the data are collected from 20 Parkinson patients (6 female, 14 male) and 20 healthy individuals (10 female, 10 male), for each subject, 26 sound samples including sustained vowels, words, and short sentences are recorded, containing in total 1040 training samples. 26 type of time-frequency based features (such as jitter, shimmer, median pitch and number of pulses etc.) are extracted for each sound samples. For the test set, the data are only collected from another 28 Parkinson patients. They are asked to say particular sustained vowels ('a' and 'o') three times and the same time-frequency based features are extracted, therefore it contains 168 samples. In our experiments, we only use the training set, because the test set does not contain negative samples. We randomly divide the training set to three subsets of equal size for model training, validation and test. The features are preprocessed by Gaussian normalization and 10^6 number of iterations are adopted to guarantee convergence in the learning stage.

Performance – We first show the results of multiclass SVM without interaction in Tab. 2. The accuracy on the training, validation and test set with different λ and also the non-zero

rate of $\mathbf{v}$ are given. From the results, it is seen that a sparse solution is obtained with $\lambda = 10^{-2}$, as well as it delivers the best test accuracy (63.98%) on the test set.

	lg(λ)					
	-1	-2	-3	-4	-5	-6
Train acc.	62.72	**65.90**	68.50	69.36	69.08	69.08
Val. acc.	58.79	**60.23**	59.94	58.79	58.50	58.50
Test acc.	62.54	**63.98**	62.82	61.96	61.96	61.96
NZ rate in $\mathbf{v}$	23.08	**65.38**	92.31	92.31	92.31	100

Table 2: The classification accuracy of multiclass SVM with ℓ_1-norm (in %) on Parkinson Speech dataset.

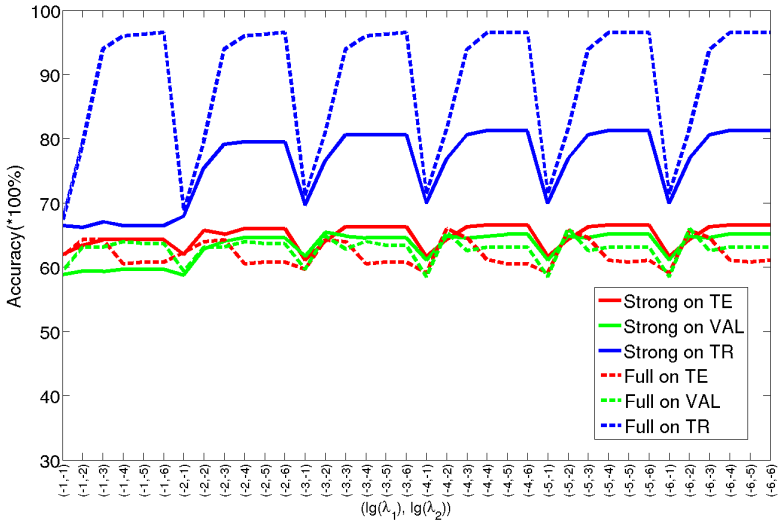

Figure 10: This figure shows classification accuracy comparison on the training, validation and test for multiclass SVMs with full interactions and strong hierarchy under different combinations of λ_1 and λ_2 on Parkinson Speech dataset.

Next we show the results of multiclass SVMs with full interaction and strong hierarchy, taking into account the interactions of features. Fig. 10 shows the performance comparison curves under different combinations of λ_1 and λ_2 for multiclass SVMs with full interactions and strong hierarchy. It is observed that: 1) the accuracy with full interaction on the training set are much higher than multiclass SVMs with ℓ_1-norm and strong hierarchy, but the accuracy on the validation and test set are not reached to as much as on the training set, the possible reason is that by combining interactions of features without further constraints, overfitting problem occurs when the learned weights (i.e. $\mathbf{v}$ and $\mathbf{\Theta}$) are increased; 2) when strong hierarchy is regularized to $\mathbf{v}$ and $\mathbf{\Theta}$, the overfitting is alleviated, and the training accuracy are decreased, while the accuracy on the validation and test set are improved compared to the one with full interactions; 3) From Tab. 2 and Fig. 10, the best accuracy on the test set for multiclass SVMs with ℓ_1-norm is 63.98% ($\lambda = 10^{-2}$), while the best

accuracy for full interaction and strong hierarchy are 65.99% ($\lambda_1 = 10^{-4}$, $\lambda_2 = 10^{-2}$) and 66.57% ($\lambda_1 = 10^{-4}$, $\lambda_2 = 10^{-4}$) respectively, which validates that strong hierarchy can help to improve the classification performance.

7 Conclusion

In this work, we propose a primal-dual proximal algorithm with epigraphical projection for regression and multiclass SVMs with strong hierarchy and apply it to several regression and classification task. Four algorithms are derived ("Weak-PD-ℓ_∞" , "Weak-PD-ℓ_1" , "Strong-PD-ℓ_∞" , "Strong-PD-ℓ_1") allowing to deal with and to properly compare different configurations of the regularization term (4). Compared to other state-of-the-art algorithms, for instance, FISTA or ADMM, the proposed algorithm is computationally more efficient (finding a solution belonging to S), and convergence in terms of iterates is faster. Compared to standard sparse learning strategies, we can integrate the main effects and interaction effects, and then obtain the underlying graphs between the features, as a result, the regularization with strong hierarchy make them more discriminative.

Appendix A: Proof of Proposition 3.1

The proof follows arguments derived in [7] for the specific case $r = 1$. Let $\mathbf{v} = \mathbf{v}^+ - \mathbf{v}^-$, (5) can be equivalently written as

$$\begin{aligned} &\underset{\mathbf{v},\mathbf{v}^+,\mathbf{v}^-,\boldsymbol{\Theta}}{\text{minimize}} \ \frac{1}{2}\sum_{\ell=1}^{L}\left(y_\ell - \phi^\top(\mathbf{x}_\ell)\mathbf{v} - \phi^\top(\mathbf{x}_\ell)\boldsymbol{\Theta}\phi(\mathbf{x}_\ell)\right)^2 + \frac{\lambda}{2}\|\boldsymbol{\Theta}\|_1 + \lambda 1^\top(2\mathbf{v}^+ - \mathbf{v}) \\ &\text{subj. to} \quad \begin{cases} \|\Theta^{(i,\cdot)}\|_r \leq 2v^{+(i)} - v^{(i)} \quad (\forall i \in \{1,\ldots,N\}) \\ \mathbf{v} = \mathbf{v}^+ - \mathbf{v}^- \\ (\mathbf{v}^+, \mathbf{v}^-, \Theta) \in O \times O \times C. \end{cases} \end{aligned} \tag{19}$$

The constraints involving $(\mathbf{v}, \mathbf{v}^+, \mathbf{v}^-)$ can be reformulated as

$$(\forall i \in \{1,\ldots,N\}) \quad \begin{cases} v^{+(i)} \geq \frac{1}{2}(\|\Theta^{(i,\cdot)}\|_r + v^{(i)}), \\ v^{+(i)} \geq v^{(i)} \\ v^{+(i)} \geq 0. \end{cases}$$

yielding $v^{+(i)} \geq \max\{\max\{0, v^{(i)}\},\ \frac{1}{2}(\|\Theta^{(i),\cdot}\|_r + v^{(i)})\}$. Using $|v^{(i)}| = 2\max\{0, v^{(i)}\} - v^{(i)}$, we get the expected result:

$$\begin{aligned}\underset{\mathbf{v},\boldsymbol{\Theta}}{\text{minimize}}\ & \frac{1}{2}\sum_{\ell=1}^{L}(y_\ell - \phi^\top(\mathbf{x}_\ell)\mathbf{v} - \phi^\top(\mathbf{x}_\ell)\boldsymbol{\Theta}\phi(\mathbf{x}_\ell))^2 \\ & + \frac{\lambda}{2}\|\boldsymbol{\Theta}\|_1 + \lambda\sum_{i=1}^{N}\max\{|v^{(i)}|, \|\Theta^{(i,\cdot)}\|_r\} + \iota_C(\boldsymbol{\Theta}). \end{aligned} \tag{20}$$

Appendix B: Proof of Proposition 4.1

According to the definition of the epigraphical projection, we solve the following minimization problem:

$$\begin{aligned} & P_{E_\infty}(\omega^+, \omega^-, \mathbf{u}) \\ & \qquad = \underset{\eta^+,\eta^-\geq 0}{\arg\min}\ (\eta^+ - \omega^+)^2 + (\eta^- - \omega^-)^2 + \min_{\substack{|p^{(1)}|<\eta^+ + \eta^- \\ \cdots \\ |p^{(N)}|<\eta^+ + \eta^-}} \|\mathbf{p} - \mathbf{u}\|^2 \end{aligned} \tag{21}$$

The inner minimization yields a simple projection for $\mathbf{u}$ to $[-\eta^+ - \eta^-, \eta^+ + \eta^-]$, i.e. $\mathbf{p} = P_{[-\eta^+-\eta^-,\eta^++\eta^-]}(\mathbf{u})$, thus Eq. (21) becomes:

$$\begin{aligned} P_{E_\infty}(\omega^+, \omega^-, \mathbf{u}) = \underset{\eta^+,\eta^-\geq 0}{\arg\min}\Bigg\{ & (\eta^+ - \omega^+)^2 + (\eta^- - \omega^-)^2 \\ & + \sum_{i=1}^{N}(\max\{|u^{(i)}| - \eta^+ - \eta^-, 0\})^2\Bigg\} \end{aligned} \tag{22}$$

and thus

$$P_{E_\infty}(\omega^+, \omega^-, \mathbf{u}) = \text{prox}_{\phi+\iota_O}(\omega^+, \omega^-) \tag{23}$$

where $\phi(\omega^+, \omega^-) = \sum_{i=1}^{N}(\max\{|u^{(i)}| - \eta^+ - \eta^-, 0\})^2$. We now focus on the computation of $\text{prox}_\phi(\omega^+, \omega^-)$.

First, we sort $(|u^{(i)}|)_{1\leq i\leq N}$ in ascending order to be $(\nu^{(1)}, \ldots, \nu^{(N)})$, such that $|\nu^{(1)}| \leq \ldots \leq |\nu^{(N)}|$, and also $\nu^{(0)} = -\infty$; and second $\bar{n} \in [1, N]$ can be found such that $\nu^{(\bar{n}-1)} \leq \eta^+ + \eta^- \leq \nu^{(\bar{n})}$, then $\phi(\omega^+, \omega^-) = \sum_{i=\bar{n}}^{M}(\eta^+ + \eta^- - \nu^{(\bar{n})})^2$, so the proximity operator of ϕ has:

$$\begin{cases} \nu^{(\bar{n}-1)} \leq \eta^+ + \eta^- \leq \nu^{(\bar{n})}, \\ \omega^+ - \eta^+ = (N - \bar{n} + 1)(\eta^+ + \eta^-) - \sum_{i=\bar{n}}^{N}\nu^{(i)}, \\ \omega^- - \eta^- = \omega^+ - \eta^+, \end{cases} \tag{24}$$

and that leads to

$$\begin{cases} \eta^- = \frac{\omega^- - (N-\bar{n}+1)(\omega^+ - \omega^-) + \sum_{i=\bar{n}}^N \nu^{(i)}}{1+2(N-\bar{n}+1)}, \\ \eta^+ = \eta^- + \omega^+ - \omega^-. \end{cases} \tag{25}$$

Appendix C: Proof of Proposition 4.2

The Lagrangian associated to the function involved in $P_{E_1^+}$ is:

$$\mathcal{L}(\eta^+, \eta^-, \mathbf{p}, \alpha, \boldsymbol{\xi}) = \frac{1}{2}\|\mathbf{p} - \mathbf{u}\|^2 + \frac{1}{2}(\eta^+ - \omega^+)^2 + \frac{1}{2}(\eta^- - \omega^-)^2 + \alpha\big(\sum_{i=1}^N p^{(i)} - \eta^+ - \eta^-\big) - \boldsymbol{\xi}^\top \mathbf{p}. \tag{26}$$

The KKT conditions are:

$$\begin{cases} (\forall i \in \{1, \dots, N\}) \quad p^{(i)} - u^{(i)} + \alpha - \xi_i = 0 \\ \eta^+ - \omega^+ - \alpha = 0 \\ \eta^- - \omega^- - \alpha = 0 \\ (\forall i \in \{1, \dots, N\}) \quad \xi^{(i)} p^{(i)} = 0 \\ \sum_{i=1}^N p^{(i)} - \eta^+ - \eta^- = 0. \end{cases} \tag{27}$$

From the fourth condition in Eq. (27), we know that if $p^{(i)} > 0$ then $\xi^{(i)} = 0$ and $p^{(i)} = u^{(i)} - \alpha$. Thus, if we denote $\boldsymbol{\pi} = (\pi^{(i)})_{1 \le i \le N}$ is an ordered version of $(p^{(i)})_{1 \le n \le N}$ in a decreasing order, we can write

$$\sum_{i=1}^N p^{(i)} = \sum_{i=1}^N \pi^{(i)} = \sum_{i=1}^{\widetilde{n}} \pi^{(i)} = \sum_{i=1}^{\widetilde{n}} (\mu^{(i)} - \alpha) = \eta^+ + \eta^-. \tag{28}$$

From similar arguments as in [19][Lemma 2] , $\widetilde{n}$ is computed as (11) and thus

$$\alpha = \frac{1}{\widetilde{n}}\Big(\sum_{i=1}^{\widetilde{n}} \mu^{(i)} - (\eta^+ + \eta^-)\Big). \tag{29}$$

From the second and third equality of the KKT conditions, we have $\eta^+ = \eta^- + \omega^+ - \omega^-$. Finally, combining (29) with $\eta^- = \omega^- + \alpha$, yields

$$\eta^- = \frac{1}{\widetilde{n}(1 + 2/\widetilde{n})}\Big(\sum_{i=1}^{\widetilde{n}} \mu^{(i)} - \omega^+ + (\widetilde{n}+1)\omega^-\Big). \tag{30}$$

Acknowledgements 7.1 *The authors would like to thank Marie-France Benassy, Mickael Bonnard, Laurent Grosset, Georges Oppenheim, Maxime Durot, Leire Oro-Urea from TOTAL for helpful discussion, Prof. Mark Asch from Université de Picardie Jules Verne for the comments.*

References

[1] Agarwal A, Beygelzimer A, Hsu D, Langford J, Telgarsky M (2014) Scalable nonlinear learning with adaptive polynomial expansions. Advances in Neural Information Processing Systems 3:2051–2059

[2] Anderson JA (1975) Quadratic logistic discrimination. Biometrika 62:149—154

[3] Bach F, Jenatton R, Mairal J, Obozinski G (2012) Optimization with sparsity-inducing penalties. Foundations and Trends in Machine Learning 4(1):1–106

[4] Bach F, Jenatton R, Mairal J, Obozinski G (2012) Structured sparsity through convex optimization. Statistical Science 27(4):450–468

[5] Bauschke H, Combettes P (2011) Convex analysis and monotone operator theory in hilbert spaces. Springer, New York

[6] Beck A, Teboulle M (2009) A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM Journal on Imaging Sciences 2(1):183–202

[7] Bien J, Taylor J, Tibshirani R (2013) A lasso for hierarchical interactions. Annals of Statistics 41(3):1111–1141

[8] Blondel M, Seki K, Uehara K (2013) Block coordinate descent algorithms for large-scale sparse multiclass classification. Machine Learning 93(1):31–52

[9] Bruna J, Mallat S (2013) Invariant scattering convolution networks. IEEE Transactions on Pattern Analysis and Machine Intelligence 35(8):1872–1886

[10] Chakraborty R, Pal NR (2015) Feature selection using a neural framework with controlled redundancy. IEEE Transactions on Neural Networks and Learning Systems 26(1):35–50

[11] Chambolle A, Dossal C (2015) On the convergence of the iterates of the "Fast Iterative Shrinkage/Thresholding Algorithm". Journal of Optimization Theory and Applications 166(3):968—982

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

[13] Chang CC, Lin CJ (2011) Libsvm: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology 2:1–27

[14] Chaux C, Pesquet JC, Pustelnik N (2009) Nested iterative algorithms for convex constrained image recovery problem. SIAM Journal on Imaging Sciences 2(2):730–762

[15] Chierchia G, Pustelnik N, Pesquet JC, Pesquet-Popescu B (2015) Epigraphical projection and proximal tools for solving constrained convex optimization problems. Signal, Image and Video Processing 9(8):1737–1749

[16] Chierchia G, Pustelnik N, Pesquet JC, Pesquet-Popescu B (2015) A proximal approach for sparse multiclass SVM. CoRR abs/1501.03669, URL `http://arxiv.org/abs/1501.03669`, `1501.03669`

[17] Combettes P, Pesquet JC (2011) Proximal splitting methods in signal processing. Fixed-Point Algorithms for Inverse Problems in Science and Engineering 49:185–212

[18] Condat L (2013) A primal-dual splitting method for convex optimization involving Lipschitzian, proximable and linear composite terms. Journal of Optimization Theory and Applications 158(2):460–479

[19] Duchi JC, S-S S Shai, Singer Y, Chandra T (2008) Efficient projections onto the l1-ball for learning in high dimensions. In: ICML, ACM International Conference Proceeding Series, vol 307, pp 272–279

[20] Flandrin P (1999) Time-frequency/time-scale analysis. Academic Press

[21] Gui J, Sun Z, Ji S, Member S, Tao D, Tan T (2016) Feature selection based on structured sparsity : a comprehensive study. IEEE Transactions on Neural Networks and Learning Systems 28(7):1–18

[22] Hao N, Feng Y, Zhang HH (2018) Model selection for high-dimensional quadratic regression via regularization. Journal of the American Statistical Association 113(522):615–625

[23] Haris A, Witten D, Simon N (2014) Convex modeling of interactions with strong heredity. Journal of Computational and Graphical Statistics pp 1–31

[24] Jenatton R, Mairal J, Obozinski G, Bach F (2011) Proximal methods for hierarchical sparse coding. Journal of Machine Learning Research 12:2297–2334

[25] Jiu M, Pustelnik N, Qi L (2018) Multiclass SVM with hierarchical interaction: application to face classification. 26th IEEE International Workshop on Machine Learning for Signal Processing pp 1–6

[26] Komodakis N, Pesquet JC (2015) Playing with duality: an overview of recent primal-dual approaches for solving large-scale optimization problems. IEEE Signal Processing Magazine 32(6):31–54

[27] Laporte L, Flamary R, Canu S, Dejean S, Mothe J (2014) Nonconvex regularizations for feature selection in ranking with sparse SVM. IEEE Transactions on Neural Networks and Learning Systems 25(6):1118–1130, DOI 10.1109/TNNLS.2013.2286696

[28] LeCun Y, Bengio Y (1995) Convolutional networks for images, speech, and time series. The handbook of brain theory and neural networks 3361(10)

[29] Lim M, Hastie T (2015) Learning interactions through hierarchical group-lasso regularization. Journal of Computational and Graphical Statistics 24:627–654

[30] Pascal B, Pustelnik N, Abry P, Serres M, Vidal V (2018) Joint Estimation Of Local Variance And Local Regularity For Texture Segmentation. Application To Multiphase Flow Characterization. In: 25th IEEE International Conference On Image Processing (ICIP), pp 2092–2096

[31] Pirayre A, Couprie C, Bidard F, Duval L, Pesquet JC (2015) BRANE Cut: Biologically-related apriori network enhancement with graph cuts for gene regulatory network inference. BMC Bioinformatics 16(1)

[32] Rhee SY, Taylor J, Wadhera G, B-H A, Brutlag DL, Shafer RW (2006) Genotypic predictors of human immunodeficiency virus type 1 drug resistance. Proceedings of the National Academy of Sciences 103(46):17,355–17,360

[33] Ronald HR, James DB, John SR, Robert VH (1978) Generalized linear and quadratic discriminant functions using robust estimates. Journal of the American Statistical Association 73:564—568

[34] Sakar E, Isenkul B, Sakar M, Sertbas C, Gurgen A, Delil F, Apaydin S, Kursun H (2013) Collection and analysis of a parkinson speech dataset with multiple types of sound recordings. IEEE Journal of Biomedical and Health Informatics 17(4):828–834

[35] Schmidt M, Roux NL, Bach F (2011) Convergence rates of inexact proximal-gradient methods for convex optimization. Advances in Neural Information Processing Systems 24:1458–1466

[36] Setzer S (2009) Split Bregman algorithm, Douglas-Rachford splitting and frame shrinkage, vol 5567, chap Scale Space and Variational Methods in Computer Vision. SSVM 2009. Lecture Notes in Computer Science, pp 464–476

[37] She Y, Jiang H (2016) Group regularized estimation under structural hierarchy. Journal of the American Statistical Association DOI 10.1080/01621459.2016.1260470

[38] Spilka J, Frecon J, Leonarduzzi R, Pustelnik N, Abry P, Doret M (2017) Sparse Support Vector Machine for Intrapartum Fetal Heart Rate Classification. IEEE Journal Of Biomedical And Health Informatics 21(3):664–671, DOI {10.1109/JBHI.2016.2546312}

[39] Tibshirani R (1994) Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society, Series B 58:267–288

[40] Vũ BC (2013) A splitting algorithm for dual monotone inclusions involving cocoercive operators. Advances in Computational Mathematics 38(3):667–681

[41] Weston J, Elisseeff A, Scholkopf B, Tipping M (2003) Use of the zero-norm with linear models and kernel methods. Journal of Machine Learning Research 3:1439–1461

[42] Witten DM, Tibshirani R (2009) Covariance-regularized regression and classification for high dimensional problems. Journal of the Royal Statistical Society Series B: Statistical Methodology 71(3):615–636

[43] Xu J, Tang B, He H, Man H (2017) Semisupervised feature selection based on relevance and redundancy criteria. IEEE Transactions on Neural Networks and Learning Systems 28(9):1974–1984, DOI 10.1109/TNNLS.2016.2562670

[44] Zhao P, Rocha G, Yu B (2009) The composite absolute penalties family for groupes and hierarchical variable selection. The Annals of Statistics 37(6A):3468—3497

[45] Zou H, Yuan M (2008) The F∞-norm support vector machine. Statistica Sinica 18:379–398