Identifying groups of strongly correlated variables through Smoothed Ordered Weighted ℓ_1-norms

Raman Sankaran
Indian Institute of Science
ramans@csa.iisc.ernet.in

Francis Bach
INRIA - École Normale Supérieure, Paris
francis.bach@inria.fr

Chiranjib Bhattacharyya
Indian Institute of Science
chiru@csa.iisc.ernet.in

Abstract

The failure of LASSO to identify groups of correlated predictors in linear regression has sparked significant research interest. Recently, various norms [1, 2] were proposed, which can be best described as instances of ordered weighted ℓ_1 norms (OWL) [3], as an alternative to ℓ_1 regularization used in LASSO. OWL can identify groups of correlated variables but it forces the model to be constant within a group. This artifact induces unnecessary bias in the model estimation. In this paper we take a submodular perspective and show that OWL can be posed as the Lovász extension of a suitably defined submodular function. The submodular perspective not only explains the group-wise constant behavior of OWL, but also suggests alternatives. The main contribution of this paper is smoothed OWL (SOWL), a new family of norms, which not only identifies the groups but also allows the model to be flexible inside a group. We establish several algorithmic and theoretical properties of SOWL including group identification and model consistency. We also provide algorithmic tools to compute the SOWL norm and its proximal operator, whose computational complexity $O(d \log d)$ is significantly better than that of general purpose solvers in $O(d^2 \log d)$. In our experiments, SOWL compares favorably with respect to OWL in the regimes of interest.

Proceedings of the 20^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2017, Fort Lauderdale, Florida, USA. JMLR: W&CP volume 54.

1 Introduction

Parsimonious models have proven to be extremely successful in many applications such as computer vision [4], neuroimaging [5], bioinformatics [6, 7], etc. To promote parsimony one often resorts to regularization with appropriate norms. Consider the problem,

$$\min_{w \in \mathbb{R}^d} \frac{1}{2n} \|Xw - y\|_2^2 + \lambda \Omega(w), \tag{1}$$

where $X \in \mathbb{R}^{n \times d}, y \in \mathbb{R}^n$ with $\lambda \geq 0$. Lasso [8] is a classic approach which advocates $\Omega(w) = \|w\|_1$, the ℓ_1-norm of w, as a suitable regularizer for discovering sparse models; many coordinates of w are then encouraged to be set to zero. Since then, it has been extensively studied and there exists a large body of work analyzing the model consistency of Lasso [9, 10]. One of the major drawbacks of Lasso is that it tends to select only a small subset from a large group of strongly correlated co-variates [11, 1]. There has been recent interest in addressing this issue: [11] suggests a two-stage process where one uses clustering of the columns of X to identify the correlated variables and then apply Lasso-type penalties to learn the model. Recently there has been attempts to avoid the two-stage procedure by *simultaneously* discovering the correlated groups and learning the model by designing alternate norms for $\Omega(w)$. To achieve this, [1] devised an interesting norm which has been generalized in [12, 3, 2] and is called an ordered-weighted-ℓ_1 (OWL). When OWL is used as a regularizer it discovers groups of strongly correlated variables unlike Lasso. However OWL promotes models, where w is group-wise constant, i.e., it tends to set $|w_i| = |w_j|$ whenever i and j are in the same group. This is undesirable as it introduces unnecessary bias in the model.

The question of finding a norm $\Omega(w)$ which can simultaneously identify correlated groups and have the ability to learn w remains an open issue. In this paper we wish to address this problem and make the following contributions.

1. We study OWL from a submodular perspective and show that it can be interpreted as the Lovász extension of a cardinality based set-function of the support of w (see Proposition 3.1). Lovász extensions can be viewed as ℓ_∞-relaxations [13, 14] of cardinality-based submodular penalties, which explains the tendency of OWL to promote group-wise constant w. This can be restrictive, and we introduce the novel alternative smoothed-OWL norms (SOWL) which are ℓ_2 relaxations (in the sense of [13]) of cardinality-based combinatorial penalties.
2. We give an $O(d \log d)$ algorithm to evaluate the norm (Theorem 4.4) and compute its proximal operator (Corollary 4.6) which compares favourably with existing off-the-shelf algorithms which have $O(d^2 \log d)$ complexity.
3. We study the proximal denoising problem in Section 6 where we empirically show that the normalized-mean-squared-error (NMSE) of SOWL is better than that of OWL in the regimes of interest. This also shows that the proposed norms do not suffer from the group-wise constant property of OWL.
4. For orthogonal designs, we show that our norms control the false discovery rate (FDR), similar to SLOPE [2], a special case of OWL.
5. We show in Theorem 5.1, that correlated columns in the data matrix X for the problem (1) would lead to grouping of the corresponding model variables as the regularization parameter λ increases. A major difference of this result with that of OWL is that the model is not restricted to be piece-wise constant within each identified group. This leads to reduction in the bias of the model, which is empirically shown in our simulations.
6. In Theorem 5.3, we derive irrepresentability conditions for the problem (1), and show that the learnt models lead to consistent solutions.

Notations. Given $x \in \mathbb{R}^d$, we denote by $|x| \in \mathbb{R}^d$ the vector of item-wise absolute values of x, and denote by $|x|_{(i)}$ the i^{th} largest absolute entry of $|x|$. $\mathrm{D}(x)$ denotes a diagonal matrix with x as diagonal entries. $\mathrm{supp}(x)$ denotes the support set of x, which is $\{i | x_i \neq 0\}$. I_d denotes the identity matrix of size $d \times d$. Given $a \in \mathbb{R}$, we denote by a_d the d-dimensional vector of all a's. Given $x \in \mathbb{R}^d$ and $\mathcal{G} \subseteq \{1, \ldots, d\}$, $x_\mathcal{G}$ refers to the sub-vector of x restricted to $\mathcal{G}$. Given a matrix $X \in \mathbb{R}^{n \times d}$, $x_i \in \mathbb{R}^n$ denotes the i^{th} column of X; and for any $A \subseteq \{1, \ldots, d\}$, X_A denotes the $n \times |A|$-dimensional submatrix formed by extracting the columns in X given by A.

2 Related Work: OWL, OSCAR, and SLOPE

Given $c \in \mathbb{R}^d_+$ satisfying $c_1 \geq \cdots \geq c_d \geq 0$, the ordered-weighted-$\ell_1$ norm [3] is defined as

$$\Omega_\mathcal{O}(w) = \sum_{i=1}^{d} c_i |w|_{(i)}. \tag{2}$$

This reduces to the OSCAR penalty [1] for the choice of weights $c_i = \lambda_1 + \lambda_2(d-i)$ for $\lambda_1, \lambda_2 \geq 0$, which reduces to the ℓ_1-norm for $\lambda_2 = 0$. It is easy to see that one can generate all possible decreasing arithmetic progressions (AP) by varying λ_1 and λ_2. Throughout this paper, we assume WLOG that $c_1 = 1$ and hence we can parametrize the OSCAR penalty using the common difference $0 \leq \alpha \leq 1/(d-1)$ giving rise to

$$c = [1, 1-\alpha, \ldots, 1-(d-1)\alpha]^\top. \tag{OSCAR}$$

[3] gives efficient proximal operator in $O(d \log d)$ time for these norms, which make it easy to apply these norms using the proximal gradient algorithms like FISTA.

The other particular choice of weights studied by [2] is $c_i = \Phi^{-1}(1 - \frac{iq}{2d})$, where $q \in (0, 1)$ and Φ is the cumulative distribution function of the standard normal distribution. [2] studies these weights in the context of variable discovery, and for the case of orthogonal designs, i.e., $X^\top X = I_d$ in (1), they have shown that the false discovery rate (FDR) is upper bounded by qd_0/d, where $d_0 = d - |\mathrm{supp}(w^*)|$ and w^* is the true model. In the next section, we show how the OWL penalties are interpreted as convex relaxations of submodular penalties [13].

Other norms. Elastic nets [15] combine the ℓ_1 and ℓ_2 norm in a linear combination, which addresses the issue of ℓ_1 selecting only few of the correlated predictors, by selecting more variables in the model and leading to better predictive performance in correlated settings. The k-support norm [16] which is obtained as a tighter convex relaxation of ℓ_0 and ℓ_2 norms is also related to the elastic nets, but selects sparser solutions than elastic nets without losing the predictive accuracy. But both elastic nets and k-support norms do not explicitly provide results in identifying groups in the data. In this paper, we shall be interested in norms which simultaneously identify groupings within the predictors and learning the model.

3 The relationship of OWL with support: A submodular perspective

The relationship of OWL, $\Omega_\mathcal{O}(w)$, with $\mathrm{supp}(w)$, the support of w is not clear. In this section we take a sub-

modular perspective of OWL and establish that OWL can be understood as ℓ_∞ relaxations of a submodular function of its support set. Let $P : \{0,1\}^d \to \mathbb{R}$ be a submodular function [14], and $p : \mathbb{R}^d \to \mathbb{R}$ be the Lovász extension of P. In this paper, we consider only non-decreasing submodular penalties which depend only on the cardinality of input: that is, we assume that there exists a concave function $f : \mathbb{R}_+ \to \mathbb{R}_+$ such that $P(A) = f(|A|)$ where f satisfies $f(0) = 0, f(i) \geq f(i-1), i = 1, \ldots, d$. Popular choices include $P(A) = |A|, \forall A \subseteq V$, which leads to $p(w) = \|w\|_1$. The following proposition expresses the OWL penalty as the convex relaxation of P.

Proposition 3.1. *Let $w \in \mathbb{R}^d$ and $A = supp(w)$.*

1. *Consider $P(A) = f(|A|)$, a cardinality-based non-decreasing submodular function and let $p(w)$ denote its Lovász extension. Define $c_i = f(i) - f(i-1)$. Then $\Omega_{\mathcal{O}}(w) = p(|w|)$.*
2. *Conversely, given $c_1 \geq \cdots \geq c_d \geq 0$ and its associated OWL norm (2), we can derive a cardinality based non-decreasing submodular penalty $P(A) = f(|A|)$ as $f(0) = 0, f(i) = c_1 + \cdots + c_i$.*

Proof. Proven in the supplementary material as a corollary to [13]. □

The choice of regularizer based on submodular functions (without taking absolute values) has been shown by [17, Prop. 6] in the orthogonal case, $X = I$, to lead to piecewise constant weight vectors, with a set of allowed piecewise constant partitions which depend on the choice of P. We consider here an extension to groupings of absolute values.

3.1 Examples

Equating OWL (2) to submodular penalties thus opens up ways to design particular choices of c, which are not restricted to choices like OSCAR, SLOPE [1, 2]. We present below a few examples of the submodular penalties from [17] which promote piecewise constant patterns in the model.

1. $P(A) = f(|A|)$, where $f(x) = \tilde{\mu}x + \mu x(d-x)$. The Lovász extension is then such that $p(|w|) = \tilde{\mu}\|w\|_1 + \mu \sum_{i<j} |(|w_i| - |w_j|)| = \sum_{i=1}^d c_i |w|_{(i)}$, where $c_i = f(i) - f(i-1) = \mu(d - 2i + 1) + \tilde{\mu}$. We choose μ and $\tilde{\mu}$ such that $c_1 = 1$ (to make sure that the norm value of the unit vector along any axis equals 1) and $c_d \geq 0$. This is satisfied if $\mu \in [0, 1/(2(d-1))]$ and $\tilde{\mu} = 1 - \mu(d-1)$. This penalty leads to c which forms an arithmetic progression and hence equivalent to the OSCAR penalty [1].
2. $P(A) = f(|A|)$ where $f(x) = (1-\mu)x + \mu\hat{f}(x)$, where $\hat{f}(x) = 0$ if $x = 0$ or $x = d$ and $\hat{f}(x) = 1$ otherwise. The Lovász extension $p(w)$ satisfies $p(|w|) = (1-\mu)\|w\|_1 + \mu \max_{i,j} |(|w_i| - |w_j|)|$. Choosing $\mu \in [0, 0.5]$, we get $c = [1, \underbrace{1-\mu, \ldots, 1-\mu}_{d-2}, 1 - 2\mu]$. Because of the term $\max_{i,j} |(|w_i| - |w_j|)|$, this penalty will promote grouping of $|w|$ to piece-wise constant values more than the previous example does, which had a similar term, but summation of the differences $|(|w_i| - |w_j|)|$ instead of the max operator (See [17] for a comparison).

Discussion. From [13] we see that the Lovász extension of P can be reinterpreted as ℓ_∞ relaxations, which form the reason behind promoting piece-wise constant w. This can be understood from the norm balls of OWL, which have sharper corners (see Figure 1); see [14] for a further discussion. In the next section, we propose SOWL which are ℓ_2 relaxations (in the sense of [13]) of the corresponding penalty P which do not have sharp edges and hence allows for more variation of values within a group.

4 SOWL - Definition and Properties

Given a cardinality-based non-decreasing submodular penalty P, we consider its ℓ_2 relaxations [13, Lemma 8] leading to a smoothed version of OWL which we call smoothed ordered weighted ℓ_1 (SOWL) and is defined as follows:

Definition 4.1. *Let $w \in \mathbb{R}^d$ and $c \in \mathbb{R}^d_+$ satisfying $c_1 \geq \cdots \geq c_d > 0$. Define*

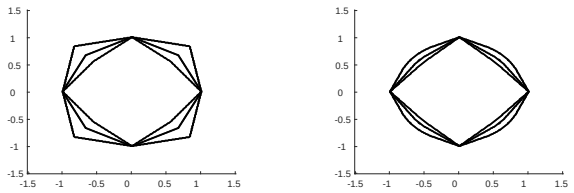
$$\Omega_{\mathcal{S}}(w) = \frac{1}{2} \min_{\eta \in \mathbb{R}^d_+} \underbrace{\sum_{i=1}^d \left(\frac{w_i^2}{\eta_i} + c_i \eta_{(i)} \right)}_{g(w,\eta)}. \qquad \text{(SOWL)}$$

Recall that, the penalty $P(A) = f(|A|)$ is related to c through $c_i = f(i) - f(i-1), \forall i = 1, \ldots, d$. See supplementary material for the proof that (SOWL) is a valid norm, and the conditions on c which guarantee that. We shall also discuss in the supplementary that the condition $c_d > 0$ can also be relaxed to $\sum_i c_i > 0$ without loss of generality.

The SOWL norms belong to the broad category of subquadratic norms where the term $\sum_{i=1}^d c_i \eta_{(i)}$ is generalized with any convex and positively homogeneous function Γ on η [18, 19, 13]. Norms of the form (SOWL) as well as more general ℓ_p-relaxations are studied by [13] for general submodular functions. They studied the algorithms to compute the norm Ω and its proximal operator using submodular function minimization, which

scales typically as $O(d^2 \log d)$, using the divide-and-conquer algorithm, which takes $O(d)$ steps of complexity $O(d \log d)$ [13].

SOWL vs OWL. It is easy to see that both OWL and SOWL coincide with the ℓ_1 norm for the choice $c = 1_d$. In the other extreme, for the choice $c = [1, 0_{d-1}] \in \mathbb{R}^d$, OWL equals the ℓ_∞ norm and SOWL coincides with the ℓ_2 norm. Hence we can interpret SOWL as a family of norms which span from the ℓ_1 to ℓ_2-norm. We plot the norm balls ($d = 2$) for ℓ_1, OWL and SOWL in Figure 1. While in OWL, the set $\{i \mid |w_i| = a\}$ for



(a) OWL (b) SOWL

Figure 1: Norm balls for OWL, SOWL. For each norm, plotted for the choices of c (from inner most to outer most ball) $c = [1, 0.8]^\top, c = [1, 0.5]^\top, c = [1, 0.2]^\top$.

some $a > 0$ is said to form a group, we build a notion of clusters through the proxy variable η: a group of indices $\mathcal{G} \subseteq \{1, \dots, d\}$ is said to form a cluster if $\eta_\mathcal{G}$ is constant-valued. Hence throughout the paper, we refer to η as a "*grouping variable*". This way we abstract out the notion of groups and the values of the model variable w.

4.1 Properties

Define $\eta_w \in \mathbb{R}^d$ such that $\Omega_\mathcal{S}(w) = g(w, \eta_w)$. We shall derive the necessary and sufficient conditions which guarantee the grouping of the variables in w through η. The following definition of a lattice helps us in establishing the forthcoming results.

Definition 4.2. *(1) For any $\eta \in \mathbb{R}^d_+$, we say that η belongs to the lattice $\mathcal{D}^{(k)}$ with a partition $\{\mathcal{G}_j\}_{j=1}^k$ of $\{1, \dots, d\}$ when $\eta_{\mathcal{G}_j} = \delta_j 1_{|\mathcal{G}_j|}$ with $\delta_1 > \cdots > \delta_k \geq 0$.*

(2) Given a lattice $\mathcal{D}^{(k)}$, we define the variables $\tau_1 = 0$, $\tau_j = |\mathcal{G}_1| + \cdots + |\mathcal{G}_{j-1}|$ for $j = 1, \dots, k-1$. And given the sequence $c_1 > \cdots > c_d$, we define the function $\mathcal{A}_j(i) = c_{\tau_j+1} + \cdots + c_{\tau_j+i}$, which is valid for $i = 1, \dots, |\mathcal{G}_j|$.

With these notations, we now characterize the minimizer η_w.

Proposition 4.3. *Let $w \in \mathbb{R}^d$, and let η_w be such that $\Omega_\mathcal{S}(w) = g(w, \eta_w)$. Let $\eta_w \in \mathcal{D}^{(k)}$ with unique values $\delta_1 > \cdots > \delta_k$. Then η_w is optimal if and only if the following conditions hold.*

1. *$\delta_j = \frac{\|w_{\mathcal{G}_j}\|_2}{\sqrt{\mathcal{A}_j(|\mathcal{G}_j|)}}, \forall j = 1, \dots, k$ and satisfies the ordering constraints $\delta_1 > \cdots > \delta_k \geq 0$.*
2. *$\forall j = 1, \dots, k$, $\forall i = 1, \dots, |\mathcal{G}_j|$, the following inequality holds for all $C_j \subseteq \mathcal{G}_j$.*

$$\frac{\|w_{C_j}\|_2^2}{\|w_{\mathcal{G}_j}\|_2^2} \leq \frac{\mathcal{A}_j(|C_j|)}{\mathcal{A}_j(|\mathcal{G}_j|)}. \tag{3}$$

Proof. Provided in the supplementary material. □

Discussion.

1. The first condition guarantees the ordering of η_w in the lattice $\mathcal{D}^{(k)}$. This ensures that the values $\|w_{\mathcal{G}_j}\|_2$ corresponding to the groups are well separated. The second condition (3) means that values within $w_{\mathcal{G}_j}$ are tight enough, and that the group can not be split into two, leading to a different partition of η_w.

2. We compare the grouping identified by (SOWL) with that of OWL. Given $w \in \mathbb{R}^d$, let $|w| \in \mathcal{D}^{(k)}$ with unique values $\bar{w}_1 > \cdots > \bar{w}_k$ in groups $\mathcal{G}_1, \dots, \mathcal{G}_k$ respectively. Note that SOWL identifies the same grouping in w if $\eta_w \in \mathcal{D}^{(k)}$, which happens if and only if the conditions of Proposition 4.3 are satisfied. When $|w| \in \mathcal{D}^{(k)}$, η_w satisfies condition (3) for all non-increasing sequences of c. Now, η satisfies the ordering property (Proposition (4.3), condition 1), if $\frac{\bar{w}_j^2}{\bar{w}_{j+1}^2} > \left(\frac{\sum_{i \in \mathcal{G}_j} c_i}{|\mathcal{G}_j|}\right) / \left(\frac{\sum_{i \in \mathcal{G}_{j+1}} c_i}{|\mathcal{G}_{j+1}|}\right), \forall j = 1, \dots, k-1$. This implies that SOWL requires $\bar{w}_j$ and $\bar{w}_{j+1}$ to be separated enough to be recognized as separate groups, otherwise which they will identified as part of the same group. See Figure 2 (bottom-left), in which we plot η_w, which illustrates this.

3. Let $\eta_w \in \mathcal{D}^{(k)}$, then $\Omega_\mathcal{S}(w) = \sum_{j=1}^k \sqrt{\mathcal{A}_j(|\mathcal{G}_j|)} \|w_{\mathcal{G}_j}\|_2$. This is equivalent to the grouped-Lasso [20] and Cluster-Group-Lasso [11] penalties with respect to the groups identified through η_w. In addition, if $|w| \in \mathcal{D}^{(k)}$ with unique values $\bar{w}_1 > \cdots > \bar{w}_k$ and $\eta_w \in \mathcal{D}^{(k)}$, $\Omega_\mathcal{S}(w) = \sum_{j=1}^k \sqrt{\mathcal{A}_j(|\mathcal{G}_j|)|\mathcal{G}_j|} |\bar{w}_j|$, which compares with the OWL norm $\Omega_\mathcal{O}(w) = \sum_{j=1}^k \mathcal{A}_j(|\mathcal{G}_j|)|\bar{w}_j|$. This means that for the piece-wise constant w, if both $|w|, \eta_w \in \mathcal{D}^{(k)}$, SOWL is equivalent to OWL (up to a difference in the weights).

Also, see Proposition 4.C in the supplementary material, which shows that $\Omega_\mathcal{S}$ is more robust in identifying groups within the model w even if w is perturbed away from a piece-wise constant partition.

4.2 Computation of SOWL, $\Omega_{\mathcal{S}}$

Algorithm 1 PAV Algorithm for computing $\text{prox}_{\Omega_{\mathcal{S}}}$

Require: $w, c \in \mathbb{R}^d$, such that $c_1 \geq \cdots \geq c_d$.
  Sort w to ensure $|w_1| \geq \cdots \geq |w_d|$;
  Set $\mathcal{I}$ - permutation to regenerate the original order of w.
  Initialize $j = 1$, $W_1 = w_1^2$, $C_1 = c_1$, $\mathcal{G}_1 = \{1\}, \eta_1 = \sqrt{W_1/C_1}$
  for $(i = 2, \ldots, d)$ **do**
    $j = j + 1$.
    $W_j = w_i^2$, $C_j = c_i$, $\mathcal{G}_j = \{i\}$, $\eta_j = \infty$
    if $C_j > 0$ **then**
      $\eta_j = \sqrt{W_j/C_j}$
    end if
    while $\eta_{j-1} < \eta_j$ **do**
      $W_{j-1} = W_{j-1} + W_j$, $C_{j-1} = C_{j-1} + C_j$
      $\mathcal{G}_{j-1} = \mathcal{G}_{j-1} \cup \mathcal{G}_j$, $j = j - 1$.
      $\eta_j = \sqrt{W_j/C_j}$
      if $j = 1$ **then**
        Break;
      end if
      $\eta_{j-1} = \sqrt{W_{j-1}/C_{j-1}}$
    end while
  end for
  $v = \sum_{k=1}^{j} \sqrt{C_k W_k}$.
  Permute v, η according to $\mathcal{I}$.
  return v, η.

We present in Algorithm 1, a procedure to compute $\Omega_{\mathcal{S}}$, which will also lead to an efficient algorithm to compute its proximal operator.

Theorem 4.4. *For any $w \in \mathbb{R}^d$, $\Omega_{\mathcal{S}}(w)$ can be computed in $O(d \log d)$ time using Algorithm 1.*

Proof. Provided in the supplementary material. □

Algorithm 1 is similar to the pool-adjacent-violator algorithm [21] in constructing the solution by merging violating pairs. Next, we show that Algorithm 1 can be used for computing the proximal operator for $\Omega_{\mathcal{S}}$.

4.3 Proximal operator for SOWL

The proximal operator is key to designing efficient algorithms for solving problems of the form (1). The proximal operator of any norm Ω is defined as follows:

$$\text{prox}_{\lambda}^{\Omega}(z) = \underset{w \in \mathbb{R}^d}{\text{argmin}} \frac{1}{2\lambda}\|w - z\|_2^2 + \Omega(w). \tag{4}$$

The following theorem, shows that the computational complexity for the prox operator is same that of computing the norm.

Theorem 4.5. ***(Proximal Problem).*** *Let $z \in \mathbb{R}^d$, $w^{(\lambda)} = prox_{\lambda}^{\Omega_{\mathcal{S}}}(z)$, $\eta_w^{(\lambda)}$ satisfy $\Omega_{\mathcal{S}}(w) = g\left(w^{(\lambda)}, \eta_w^{(\lambda)}\right)$. For a given $\lambda = \mu > 0$, let $\eta_w^{(\mu)} \in \mathcal{D}^{(k)}$ with unique values $\left(\delta_w^{(\mu)}\right)_1 > \cdots > \left(\delta_w^{(\mu)}\right)_k > 0$. Then,*

1. *For any given $\lambda > 0$, if $\left(\eta_w^{(\lambda)}\right)_i > 0, \forall i = 1, \ldots, d$, the solution $\eta_w^{(\lambda)} \in \mathcal{D}^{(k)}$.*
2. *Let j be smallest integer such that $\left(\delta_w{}^{(\lambda)}\right)_j = 0$ for any $\lambda > \mu$. Then the ordering $\left(\delta_w^{(\lambda)}\right)_1 > \cdots > \left(\delta_w^{(\lambda)}\right)_{j-1}$ is consistent with that of the lattice $\mathcal{D}^{(k)}$.*

Proof. Provided in the supplementary material. □

The above result along with Algorithm 1 gives us a simple algorithm to compute the proximal operator for $\Omega_{\mathcal{S}}$ in $O(d \log d)$ time.

Corollary 4.6. ***(Computing Prox$_{\Omega}$)*** *Given any $z \in \mathbb{R}^d$, let $w = prox_{\lambda}^{\Omega_{\mathcal{S}}}(z)$. Let η_z satisfy $\Omega_{\mathcal{S}}(z) = g(z, \eta_z)$, then η_w satisfying $\Omega_{\mathcal{S}}(w) = g(w, \eta_w)$ is given by $\max(\eta_z - \lambda, 0)$ and $w_i = z_i \left((\eta_w)_i / \left((\eta_w)_i + \lambda\right)\right)$.*

Proof. Provided in the supplementary material. □

This contrasts with the general algorithm with $O(d^2 \log d)$ provided in [13] to compute $\text{prox}_{\lambda}^{\Omega}$.

5 Regularization with SOWL

In this section we discuss the group identification and the consistency of learnt model obtained from SOWL. We will assume a fixed design setting, and assume that the columns x_i of X are normalized such that $\|x_i\|_2 = 1$. Let us consider the learning problem (1) with the norm $\Omega_{\mathcal{S}}$:

$$\min_{w \in \mathbb{R}^d, \eta \in \mathbb{R}^d_+} \frac{1}{2n}\|Xw - y\|_2^2 + \frac{\lambda}{2}\left(\sum_{i=1}^{d} \frac{w_i^2}{\eta_i} + c_i \eta_{(i)}\right). \tag{5}$$

5.1 SOWL regularization for Fixed Design

In this subsection we discuss the case for general X. First we shall discuss the identification of correlated groups, and then discuss recovery of the support of the true model.

5.1.1 Identification of correlated groups

We shall first show that the correlated predictors will be grouped together, through the grouping variable η assuming piece-wise constant values over the identified groups. In this discussion, we assume that c forms a

strictly decreasing sequence $c_1 > \cdots > c_d > 0$ and define $C = \min_{i=1}^{d-1}\{c_i - c_{i+1}\}$. The following result, shows that as the regularization parameter λ increases, values in η will be grouped together. This implies that, the more the correlation between any pair of variates, the more they will be clustered together very early in the regularization path.

Theorem 5.1. *Let $\hat{w}^{(\lambda)}, \hat{\eta}^{(\lambda)}$ denote the optimal solution of (5) for a given $\lambda > 0$. Given the indices i, j, let $\rho_{ij} = x_i^\top x_j > 0$. Let us assume that $\hat{\eta}_i^{(\lambda)} > 0$ and $\hat{\eta}_j^{(\lambda)} > 0$ are distinct from $\hat{\eta}_k^{(\lambda)}$ for $k \neq i, j$. Then there exists $0 < \lambda_0 \leq \frac{\|y\|_2}{\sqrt{C}}(4 - 4\rho_{ij}^2)^{\frac{1}{4}}$ such that for all $\lambda > \lambda_0$, $\hat{\eta}_i^{(\lambda)} = \hat{\eta}_j^{(\lambda)}$.*

Proof. Provided in the supplementary material. □

Remarks

1. The above statement is similar to Theorem 1 in [1]. Whereas in OSCAR, the predictors w get clustered into constant valued partitions. In our formulation, η are being clustered into constant valued partitions. From the optimality conditions for (5), we can show that when $\hat{\eta}_i > 0, \hat{w}_i^2 = c_i\hat{\eta}_i^2$. This implies that $\hat{w}_i$ and $\hat{w}_j$ will take different values even when $\hat{\eta}_i$ and $\hat{\eta}_j$ gets clustered, when the vector c takes unique values.

2. As discussed in [1], the assumption made in Theorem 5.1 that $\hat{\eta}_i$ and $\hat{\eta}_j$ to be distinct from the remaining $\hat{\eta}$ values is not very restrictive. When $\hat{\eta}$ has groups with k unique values $\delta_1 > \cdots > \delta_k$, we can redefine the problem (5) with new variables $\hat{\delta}$ replacing $\hat{\eta}$ and appropriately redefining c values. The statements of Theorem 5.1 extend to the new problem seamlessly for the newly defined variables δ.

5.1.2 Consistency and Support Recovery

We assume the following linear model for data $y = Xw^* + \varepsilon$, where $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$. Let us denote by $(\hat{w}, \hat{\eta})$ the optimal solution of (5).

Proposition 5.2. ***(Optimality Conditions for (5))*** *Consider the problem (5) and let $(\hat{w}, \hat{\eta})$ be the solution. Let $\hat{\eta} \in \mathcal{D}^{(k)}$ with unique values $\hat{\delta}_1 > \cdots > \hat{\delta}_k \geq 0$ over the partition $\{\mathcal{G}_1, \ldots, \mathcal{G}_k\}$. Then $(\hat{w}, \hat{\eta})$ is optimal if and only if the following conditions hold.*

1. $\forall \hat{\delta}_j > 0, X_{\mathcal{G}_j}^\top (X\hat{w} - y) + \lambda n\sqrt{\mathcal{A}_j(|\mathcal{G}_j|)}\frac{\hat{w}_{\mathcal{G}_j}}{\|\hat{w}_{\mathcal{G}_j}\|_2} = 0.$
2. *If* $\hat{\delta}_k = 0, \|X_{\mathcal{G}_k}^\top (X\hat{w} - y)\|_2 \leq \lambda n\sqrt{\mathcal{A}_j(|\mathcal{G}_k|)}.$
3. *$\hat{\delta}$ is optimal for $\hat{w}$ as per Proposition 4.3.*

Proof. Provided in the supplementary material. □

For the true model w^*, let η^* be the mimimizer of (SOWL). Let $\eta^* \in \mathcal{D}^{(k)}$ with unique values $\delta_1^* > \cdots > \delta_k^*$ in the sets $\mathcal{G}_1, \ldots, \mathcal{G}_k$. We denote by $\mathcal{J}$ the set $\{i | w_i \neq 0\}$ (and by $\mathcal{J}^c$ the set $\{i | w_i = 0\}$). Similarly we denote by $\hat{\mathcal{J}}$ the sparsity pattern of $\hat{w}$ the solution of (5). It would be necessary to derive conditions on which $\hat{\mathcal{J}}$ equals $\mathcal{J}$. In the theorem below, we denote by $D\left(\sqrt{\mathcal{A}_j(|\mathcal{G}_j|)}/\|w^*_{\mathcal{G}_j}\|_2\right)$, a block diagonal matrix with blocks sizes $|\mathcal{G}_j|$ in which each diagonal blocks is defined to be $\left(\sqrt{\mathcal{A}_j(|\mathcal{G}_j|)}/\|w^*_{\mathcal{G}_j}\|_2\right) I_{|\mathcal{G}_j|}$ for all $\delta_j^* > 0$.

Theorem 5.3. *Assume the model $y = Xw^* + \varepsilon$ with the rows in X sampled from a multivariate Gaussian with covariance matrix Σ. We assume that $\Sigma_{\mathcal{J},\mathcal{J}}$ is invertible. Consider the problem defined in (5) and let $(\hat{w}, \hat{\eta})$ be the solution. As $\lambda \to 0$, and $\lambda\sqrt{n} \to \infty$, the estimate $\hat{w}$ converges in probability to w^* and $\mathbb{P}(\hat{\mathcal{J}} = \mathcal{J}) \to 1$ if the following conditions hold.*

1. $\delta_k^* = 0$ *if* $|\mathcal{J}^c| \neq \emptyset$.
2. $\frac{\left\|\Sigma_{\mathcal{J}^c,\mathcal{J}}(\Sigma_{\mathcal{J},\mathcal{J}})^{-1} D\left(\sqrt{\mathcal{A}_j(|\mathcal{G}_j|)}/\|w^*_{\mathcal{G}_j}\|_2\right) w^*_{\mathcal{J}}\right\|_2}{\sqrt{\mathcal{A}_k(|\mathcal{G}_k|)}} < 1.$

Proof. Provided in the supplementary material. □

Discussion.

1. Note that the results from [13, Section 6.5] for model consistency and support recovery can be directly applied to (5), whose bounds are not dependent on the true model w^*, and the grouping information. Whereas, the irrepresentability conditions in Theorem 5.3 explicitly depends on w^* leading to better bounds than in [13].

2. Theorem 5.3 is analogous to that of grouped-Lasso by [20], which means that the regularizer has similar guarantees as the grouped-Lasso penalty, but without explicitly specifying the groups.

5.2 Orthogonal Design Case

In this subsection, we shall consider the special case of orthogonal design matrix : $X^\top X = I_d$. Minimizing with respect to w from (5) leads to the following program:

$$\min_{\eta \geq 0} \sum_{i=1}^{d} \left(\frac{z_i^2}{\eta_i + \lambda} + c_i \eta_{(i)} \right) \tag{6}$$

Theorem 5.4. *Condider problem (6), with c chosen as $c_i = \Phi^{-1}(1 - \frac{iq}{2d})$ where $q \in [0, 1]$ is the the level desired. Then the procedure (6) rejecting hypothesis for which $\eta_i \neq 0$ has FDR upper bounded by $q\frac{d_0}{d}$, where $d_0 = |\{i | \eta_i = 0\}|$.*

Proof. Provided in the supplementary material. □

The above result guarantees that we do not lose the FDR guarantees provided by SLOPE [2], by extending it to SOWL.

6 Simulations

We demonstrate the benefits of using SOWL through the following numerical simulations. First we consider the proximal problem, which illustrates the differences these norms have against the OWL penalties. Next we quantitatively measure the normalized mean square error for the proximal denoising problem and compare the performances of $\Omega_{\mathcal{O}}$ and $\Omega_{\mathcal{S}}$. Last, we illustrate the benefit of SOWL in in predictive experiments.

6.1 Effective group discovery using SOWL

Consider the problem (4) and let $\hat{w}$ be the minimizer. We intend to understand the evolution of solutions $\hat{w}$ for different norms as λ varies. We chose $z = [\underbrace{0,\ldots,0}_{10}, \underbrace{1,\ldots,2}_{5-\text{equispaced}}, \underbrace{2,\ldots,1}_{5-\text{equispaced}}]$ in (4) in experiments. We chose the first example given in Section 3.1 to generate the values of c. Figure 2 which shows the plots of the regularization paths for the norms ℓ_1, OWL and SOWL. we also plot the path of the grouping variable η in order to highlight the difference SOWL has over OWL. It is clear from Figure 2 that $\Omega_{\mathcal{S}}$ discovers the groups very early in the regularization path than $\Omega_{\mathcal{O}}$. This is consistent with Proposition 4.3 with respect to identification of groups. The plots for the other examples are similar and hence skipped.

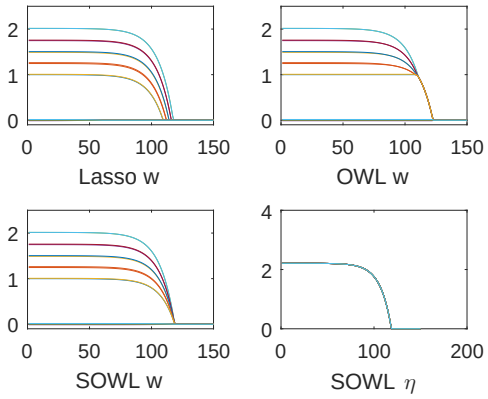


Figure 2: Regularization path for the proximal problem (4). In each plot, the x-axis refers to λ, the curves correspond to individual variables in $\hat{w}$.

6.2 Structure recovery using OWL and SOWL

Next, we quantitatively evaluate the differences between the norms with respect to recovering structured data. The mean square error (MSE) of the proximal denoising problem [22] is a popular measure used for this purpose. Formally, assuming $w^* \in \mathbb{R}^d$ as the true model, which is perturbed to $\tilde{w} = w^* + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$, the denoising problem computes $\hat{w} = \text{prox}^{\Omega}_{\sigma\lambda}(\tilde{w})$. The normalized-mean-squared-error (NMSE) defined as $\beta_\Omega(\sigma) = E[\|\hat{w} - w^*\|_2^2]/\sigma^2$ is used to evaluate the effectiveness of recovering the true model. In experiments, we use the signals (d = 400) plotted in Figure 3. We have included two different signals, Figure 3a would favour OWL more since the groups are piece-wise constant, whereas Figures 3b SOWL more. We evaluate the performance of

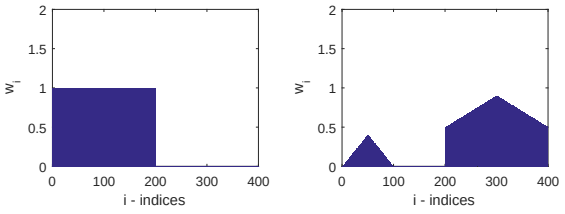


Figure 3: Signals used in the denoising experiment.

OWL and SOWL with reference to the ℓ_1-norm. Given $\sigma > 0$, let us denote by $\gamma_\Omega(\sigma) = \beta_\Omega(\sigma) - \beta_{\ell_1}(\sigma)$. For both OWL and SOWL, to maintain fairness, we experimented with c chosen from the examples given in Section 3.1 and chose the best performing one for each algorithm. The experiments were repeated 100 times and we plot the mean error values in Figure 4. We include the same plot with error bars in the supplementary material. We see from Figure 4a and 4b that

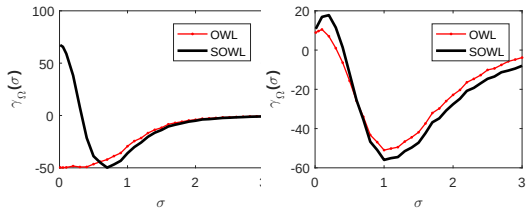


Figure 4: Proximal denoising plots (a), and (b) refer to examples in Figures 3a, 3b respectively.

OWL performs better[1] as expected for case in Figures 3a when the signal is piece-wise constant. SOWL is worse than ℓ_1 when the error variance σ is very small, and does better as σ increases. But overall OWL does better than SOWL in this case.

For the next case (Figure 3b), ℓ_1-norm performs better when σ is very small, and when σ crosses a thresh-

[1]Note that the performance is better if the curve goes negative.

Example		Med. MSE	MSE (10th Perc).	MSE (90th Perc)
1	LASSO	2.83 / 2.80 / 2.79	1.41 / 1.40 / 1.41	4.54 / 4.53 / 4.53
	OWL	**1.54** / **1.55** / 1.56	**0.26** / **0.27** / **0.27**	3.79 / 3.84 / 3.86
	El. Net	1.56 / 1.56 / 1.56	0.54 / 0.55 / 0.55	**3.73** / **3.73** / **3.73**
	$\Omega_{\mathcal{S}}$	1.59 / 1.57 / **1.55**	0.58 / 0.55 / 0.54	3.83 / 3.83 / 3.79
2	LASSO	46.1 / 45.2 / 45.5	32.8 / 32.7 / 33.2	60.0 / 61.5 / 61.4
	OWL	27.6 / 27.0 / 26.4	19.8 / 19.2 / 19.2	42.7 / 40.4 / 39.2
	El. Net	30.8 / 30.7 / 30.6	21.9 / 22.6 / 23.0	42.4 / 43.0 / 41.4
	$\Omega_{\mathcal{S}}$	**23.9** / **23.3** / **23.4**	**16.9** / **16.8** / **16.8**	**35.2** / **35.4** / **33.2**
3	LASSO	36.8 / 39.2 / 39.1	16.1 / 16.4 / 15.9	85.9 / 87.0 / 83.7
	OWL	35.6 / 35.7 / 35.3	14.0 / 15.0 / 14.2	82.4 / 86.4 / 85.0
	El. Net	28.9 / 31.4 / 30.1	10.8 / 9.4 / 10.7	73.6 / 76.4 / **80.8**
	$\Omega_{\mathcal{S}}$	**28.2** / **30.1** / **29.1**	**10.3** / **9.2** / **10.1**	**71.9** / **73.3** / 83.2

Table 1: Comparison of MSE of algorithms for the examples quoted in Section 6.3. The three subcolumns represent results for 3 levels of perturbation of the ground truth w^*.

old (typically around $\sigma = 0.75$), SOWL performs best with consistently lesser error than OWL. This may be understood as the prior information regarding the structure makes more sense in the presence of more noise than in low noise settings, and hence gives better performance in terms of NMSE. And when the noise is very low, both the structure enforcing norms do not play a part in this case. Figure 3b, in contrast to Figure 4a illustrates again the fact that OWL is so heavily dependent on the piece-wise constant assumption, which when violated, shows a sharp decrease in the performance.

6.3 Evaluation of Predictive Accuracy

Next, we compare the norms quantitatively in the least squares regression problem (1) similar to that of OSCAR [1]. We generate each sample $x \in \mathbb{R}^d$ as $x \sim \mathcal{N}(0, \Sigma)$. We generate the responses as $y = x^\top w^* + \varepsilon$, where $\varepsilon = \mathcal{N}(0, \sigma^2)$. The parameters for the algorithms are chosen through cross validation. Following [1] we calculate the mean square error (MSE), which is defined to be the expectation $E\left[\|x^\top(w - \hat{w})\|_2^2\right]$ where the expectation is defined over the distribution of x, which in our case equals $(w - \hat{w})^\top \Sigma (w - \hat{w})$. Each experiment is repeated for 300 datasets and we report the median MSE. Following OSCAR [1], the scenarios considered are:

1. We generate a non-sparse underlying model with $w^* = (0.85)\ 1_8$, with $n = 20$ samples, $\sigma = 3$. We set $\Sigma_{i,j} = 0.7^{|i-j|}$.

2. We set $n = 100$, $d = 40$ with $w^* = [0_{10}^\top, 2_{10}^\top, 0_{10}^\top, 2_{10}^\top]^\top$. We use $\sigma = 15$ and $\Sigma_{i,j} = 0.5$ if $i \neq j$ and 1 if $i = j$.

3. Here we set $n = 5$, $d = 40$ with the true model $w^* = [3_{15}^\top, 0_{25}^\top]^\top$ and $\sigma = 15$. We then generate i.i.d. $\mathcal{N}(0, 1)$ random variables Z_1, Z_2, Z_3. The i^{th} entry of Each sample x is generated as $x_i = \begin{cases} Z_1 + \epsilon_i^x, & i = 1, \ldots, 5 \\ Z_2 + \epsilon_i^x, & i = 6, \ldots, 10 \\ Z_3 + \epsilon_i^x, & i = 11, \ldots, 15 \\ \epsilon_i^x, & i = 16, \ldots, 40, \end{cases}$ where $\epsilon_i^x \sim \mathcal{N}(0, 0.16)$.

In all the cases we also study random perturbations of the vector w^* to see if the learnt models are robust. We used $\tilde{w}^* = w^* + \tilde{\epsilon}$, where $\tilde{\epsilon}_i$ was chosen in a uniformly random interval $[-\tau, \tau]$, where we set $\tau = 0, 0.2, 0.4$ in experiments. We report the results in Table 1. It is clear from the numbers that $\Omega_{\mathcal{S}}$ compares favourably against the compared norms ℓ_1, OWL, and elastic nets. This illustrates again the benefit the proposed SOWL provide for the predictive experiments.

7 Conclusions

We derived SOWL as ℓ_2 relaxations of cardinality based submodular functions, which are helpful in simultaneous grouping and learning the model variable w in a more robust way than the state-of-the-art procedures. SOWL is shown to be beneficial in the context of proximal denoising, and linear regression which leads to lesser MSE in the learnt model w than the existing group identification procedures. This opens up a large family of norms based on the submodular penalty appropriate and the choice of best submodular penalty for a particular domain is not well understood, and will be the scope of future work.

Acknowledgements

This work was supported by a generous grant from IFCAM, as well as the INRIA Associated Team "Bigfoks".

References

[1] H. D. Bondell and B. J. Reich. Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with oscar. *Biometrics*, 64:115–123, 2008.

[2] M. Bogdan, E. van den Berg, C. Sabatti, W. Su, and E. J. Cands. Statistical estimation and testing via the sorted l1 norm. *ArXiv preprint:1310.1969v2*, 2013.

[3] X. Zeng and M. Figueiredo. The ordered weighted ℓ_1 norm: Atomic formulation, projections, and algorithms. *ArXiv preprint:1409.4271v5*, 2015.

[4] R. G. Baraniuk, V. Cevher, M. F. Duarte, and C. Hegde. Model-based compressive sensing. *IEEE Transactions on Information Theory*, 56(4):1982–2001, 2010.

[5] R. Jenatton, A. Gramfort, V. Michel, G. Obozinski, F. Bach, and B. Thirion. Multi-scale mining of fmri data with hierarchical structured sparsity. *Pattern Recognition in NeuroImaging (PRNI)*, 2011.

[6] H. Zhou, M. E. Sehl, J. S. Sinsheimer, and K. Lange. Association screening of common and rare genetic variants by penalized regression. *Bioinformatics*, 26(19), 2010.

[7] F. Rapaport, E. Barillot, and J.P. Vert. Classification of arraychg data using fused svm. *Bioinformatics*, 24(13):i375–i382, 2008.

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

[9] M.J. Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (lasso). *IEEE Transactions On Information Theory*, 55(5):2183–2202, 2009.

[10] P. Zhao and B. Yu. On model selection consistency of lasso. *Journal of Machine Learning Research*, 7:2541–2563, 2006.

[11] P. Buhlmann, P. Rtimann, Sara van de Geer, and Cun-Hui Zhang. Correlated variables in regression : Clustering and sparse estimation. *Journal of Statistical Planning and Inference*, 43:1835–1858, 2013.

[12] M. A. T. Figueiredo and R. D. Nowak. Ordered weighted ℓ_1 regularized regression with strongly correlated covariates: Theoretical aspects. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2016.

[13] G. Obozinski and F. Bach. Convex relaxation for combinatorial penalties. Technical Report 00694765, HAL, 2012.

[14] F. Bach. Learning with submodular functions: A convex optimization perspective. *Foundations and Trends in Machine Learning*, 6-2-3:145–373, 2011.

[15] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. : Series B (Statistical Methodology)*, 67(2):301–320, 2005.

[16] A. Argyriou, R. Foygel, and N. Srebro. Sparse prediction with the k-support norm. In *Advances in Neural Information Processing Systems (NIPS)*, 2012.

[17] F. Bach. Shaping level sets with submodular functions. In *Advances in Neural Information Processing Systems (NIPS)*, 2011.

[18] F. Bach, R. Jenatton, J. Mairal, and G. Obozinski. Optimization with sparsity-inducing penalties. *Foundations and Trends in Machine Learning*, 4(1):1–106, 2011.

[19] C. Micchelli, J. Morales, and M. Pontil. Regularizers for structured sparsity. *Advances in Computational Mathematics*, 38(3):455–489, 2013.

[20] F. Bach. Consistency of the group lasso and multiple kernel learning. *Journal of Machine Learning Research*, 9:1179–1225, 2008.

[21] M. J. Best and N. Chakravarti. Active set algorithms for isotonic regression: a unifying framework. *Mathematical Programming*, 47(1):425–439, 1990.

[22] S. Oymak and B. Hassibi. Sharp MSE bounds for proximal denoising. *Found Comput Math.*, 16(4):965–1029, 2016.