THE GROUP k-SUPPORT NORM FOR LEARNING WITH STRUCTURED SPARSITY

Nikhil Rao, Miroslav Dudík, Zaid Harchaoui

Technicolor R&I, Microsoft Research, University of Washington

ABSTRACT

Several high-dimensional learning applications require the parameters to satisfy a "group sparsity" constraint, where clusters of coefficients are required to be simultaneously selected or rejected. The group lasso and its variants are common methods to solve problems of this form. Recently, in the standard sparse setting, it has been noted that tighter convex relaxations than the ℓ_1 norm can be obtained by using other regularizers, leading to better predictive performance. Motivated by these discoveries, we develop the group k-support norm to achieve group sparsity with a better predictive accuracy. We develop an efficient algorithm to solve the resulting optimization problems, and show that our approach outperforms traditional methods for prediction and recovery under group sparsity.

1. INTRODUCTION

High-dimensional parameter estimation and prediction is a central task in machine learning and signal processing. Given data or a sensing matrix $\mathbf{\Phi} \in \mathbb{R}^{n \times p}$ for n samples in p dimensions, $p \gg n$, and the corresponding vector of observations $\mathbf{y} \in \mathbb{R}^n$, a common way to solve for parameters $\mathbf{x} \in \mathbb{R}^p$ is by the following convex program:

$$\hat{\mathbf{x}} = \underset{\mathbf{x}}{\operatorname{argmin}}\, f(\mathbf{\Phi x}, \mathbf{y}) \quad \text{s.t.} \quad \|\mathbf{x}\|_1 \le \tau, \tag{1}$$

where f is a loss function convex in its first argument, such as the squared or the logistic loss, and the ℓ_1 norm encourages sparsity among the parameters.

While ℓ_1-constrained estimation and its regularized variants have many favorable properties, their accuracy can suffer in the presence of correlations among features (columns of $\mathbf{\Phi}$) [4]. Indeed, if two columns of $\mathbf{\Phi}$ are nearly identical, programs of the form (1) will select at most one of the features, thus potentially discarding useful information and obtaining less accurate predictions. A simple alternative is to use the elastic net [22], where one adds an additional ℓ_2-norm-squared penalty into the objective of (1). Elastic nets better leverage correlated information, by selecting and weighting relevant correlated columns, and also yield more interpretable models [22].

The ℓ_1 constraint in (1) is frequently motivated as the tightest convex relaxation of the sparsity constraint over an ℓ_∞ ball [1]. Elastic nets effectively impose an additional ℓ_2 constraint. Recently, Argyriou et al. [1] argued that instead of elastic nets, we should use the tightest convex relaxation of the sparsity constraint over an ℓ_2 ball, i.e., the tightest convex relaxation of the set

$$\mathcal{S}_k := \{\mathbf{x} : \|\mathbf{x}\|_0 \le k, \;\; \|\mathbf{x}\| \le 1\}, \tag{2}$$

where $\|\cdot\|_0$ is the number of nonzero entries and $\|\cdot\|$ is the ℓ_2 norm. The convex hull of $\mathcal{S}_k$ is taken as a norm-one ball, defining the norm called the *k-support norm*. This norm outperforms both the lasso and the elastic net in several domains [1, 16, 3], and has been successfully generalized to clustered multi-task learning and low-rank optimization [14, 15].

In this paper, we propose a generalization of the k-support norm to *group sparsity* settings. Group sparsity has many applications, including multi-task learning [13], image processing [19], and computational biology [11]. In these settings, entire groups (or clusters) of variables are required to be selected or discarded. Note that the standard sparse regression is a special case of group sparsity, with singleton groups. Furthermore, group sparsity can be thought of as a highly general structural assumption, since the groups can be arbitrarily defined. Given such a set of possibly overlapping groups $\mathcal{G} = \{G_1, \ldots, G_M\}$, $G_i \subseteq \{1, \ldots, p\}$, it is standard to apply the *group lasso* constraint or regularization [11, 21]:

$$\hat{\mathbf{x}} = \underset{\mathbf{x}}{\operatorname{argmin}}\, f(\mathbf{\Phi x}, \mathbf{y}) \quad \text{s.t.} \quad \|\mathbf{x}\|_{\mathcal{G}} \le \tau, \tag{3}$$

where $\|\cdot\|_{\mathcal{G}}$ is the group lasso[1] norm, defined as

$$\|\mathbf{x}\|_{\mathcal{G}} := \inf\Big\{\sum_{G \in \mathcal{G}} \|\mathbf{z}_G\| : \;\; \mathbf{x} = \sum_{G \in \mathcal{G}} \mathbf{z}_G\Big\}, \tag{4}$$

where $\operatorname{supp}(\mathbf{z}_G) \subseteq G$.

Similar to lasso, it can be shown that the group lasso constraint is the tightest convex relaxation of the "group sparsity" over an ℓ_∞ ball. To improve the accuracy in the presence of group sparsity, we propose an analogous approach as we saw in the case of the k-support norm. We consider the tightest relaxation of the group sparsity constraint over the ℓ_2 ball, namely, the relaxation of the set

$$\mathcal{S}_{k,\mathcal{G}} := \{\mathbf{x} : \|\mathbf{x}\|_{0,\mathcal{G}} \le k, \;\; \|\mathbf{x}\| \le 1\}, \tag{5}$$

[1] We refer to the overlapping case as the group lasso since it subsumes the standard non-overlapping case.

978-1-5090-4117-6/17/$31.00 ©2017 IEEE

where $\|\mathbf{x}\|_{0,\mathcal{G}}$ measures the smallest number of groups required to cover $\mathrm{supp}(\mathbf{x})$. We call the norm whose unit ball is the convex hull of (5) the *group k-support* norm.

Similar to the k-support norm, the group k-support norm can be written as an overlapping group lasso norm [1], with the number of groups exponential in k. While this seems intractable at the first sight, we can exploit the structure of the problem and solve the resulting optimization via the conditional gradient approach [17]. We experimentally verify that the group k-support norm outperforms both the group lasso and the group elastic net in several problems: multilabel learning, compressed sensing and computational biology.

2. THE GROUP k-SUPPORT NORM

In this section, we formally define and analyze the group k-support norm. Assume we are given an arbitrary set of groups $\mathcal{G}$, and let $\mathcal{G}_k$ denote the set consisting of k-unions of groups in $\mathcal{G}$, i.e.,

$$H \in \mathcal{G}_k \;\text{ iff }\; H = \bigcup_{i=1}^{k} G_i \;\text{ for some }\; G_i \in \mathcal{G}. \qquad (6)$$

We shall refer to every $H \in \mathcal{G}_k$ as a supergroup, since it consists of multiple groups. We define the group k-support norm $\|\cdot\|_{k,\mathcal{G}}$ as the group lasso norm for $\mathcal{G}_k$:

$$\|\mathbf{x}\|_{k,\mathcal{G}} := \|\mathbf{x}\|_{\mathcal{G}_k} = \inf\Big\{ \sum_{H\in\mathcal{G}_k} \|\mathbf{z}_H\| : \sum_{H\in\mathcal{G}_k} \mathbf{z}_H = \mathbf{x} \Big\}, \quad (7)$$

where $\mathbf{z}_H$ is a vector whose support is restricted to the supergroup H, which in turn is the union of k groups in $\mathcal{G}$. We will show that the unit ball of (7) is the convex hull of $\mathcal{S}_{k,\mathcal{G}}$ defined in (5).

Since (7) is a special case of the overlapping group lasso penalty [11], it immediately follows that it is a norm. We are interested in solving problems of the form

$$\hat{\mathbf{x}} = \operatorname*{argmin}_{\mathbf{x}} f(\boldsymbol{\Phi}\mathbf{x},\, \mathbf{y}) \;\text{ s.t. }\; \|\mathbf{x}\|_{k,\mathcal{G}} \le \tau. \qquad (8)$$

We next analyze the group k-support norm in more detail.

2.1. Formulation as an Atomic Norm

Our algorithm exploits the fact that the group k-support norm can be written as an *atomic norm*. This has been previously noted for an arbitrary overlapping group lasso norm and the standard k-support norm in particular [20, 7]. The definition of an atomic norm begins with the set of atoms, in our case,

$$\mathscr{A} = \{\mathbf{z} \in \mathbb{R}^p : \; \|\mathbf{z}\| = 1 \text{ and } \mathrm{supp}(\mathbf{z}) \subseteq H \in \mathcal{G}_k\}. \quad (9)$$

The atomic norm is then defined as the gauge function of the convex hull of $\mathscr{A}$ [6]:

$$\|\mathbf{x}\|_{\mathscr{A}} := \inf\Big\{ \sum_{\mathbf{a}\in\mathscr{A}} c_{\mathbf{a}} : \mathbf{x} = \sum_{\mathbf{a}\in\mathscr{A}} c_{\mathbf{a}}\mathbf{a}, \text{ and } c_{\mathbf{a}} \ge 0 \text{ for all } \mathbf{a} \Big\}.$$

The group k-support norm can be characterized as follows:

Proposition 2.1. *The group k-support norm is the gauge function of the convex hull of the set $\mathscr{A}$. Furthermore, the unit ball of the group k-support norm corresponds to the set conv($\mathcal{S}_{k,\mathcal{G}}$), with $\mathcal{S}_{k,\mathcal{G}}$ defined in* (5).

Proof. The first part of the proposition is an immediate consequence of properties of the atomic norm [6, 1]. To prove the second part, first consider a vector $\mathbf{x}$ such that $\|\mathbf{x}\|_{k,\mathcal{G}} \le 1$. Let $\mathbf{z}_H$ for $H \in \mathcal{G}_k$ be the set of vectors that attain the inf in the definition (7). The infimum is always attained because the minimized objective has compact level sets. Furthermore, let $\mathbf{z}_H = c_H\mathbf{u}_H$, where $c_H = \|\mathbf{z}_H\|$ and $\mathbf{u}_H$ is a unit vector in the same direction and hence with the same support as $\mathbf{z}_H$ (if $\mathbf{z}_H = \mathbf{0}$, we set $\mathbf{u}_H$ to an arbitrary unit vector with support H). Hence $\mathbf{u}_H \in \mathcal{S}_{k,\mathcal{G}}$. Then,

$$1 \ge \sum_{H\in\mathcal{G}_k} \|\mathbf{z}_H\| = \sum_{H\in\mathcal{G}_k} c_H\|\mathbf{u}_H\| = \sum_{H\in\mathcal{G}_k} c_H \;.$$

Since $\mathbf{0} \in \mathcal{S}_{k,\mathcal{G}}$, this means that we can write $\mathbf{x}$ as a convex combination of vectors in $\mathcal{S}_{k,\mathcal{G}}$. Conversely, consider a vector $\mathbf{x}$ which can be written as a convex combination of N vectors $\mathbf{v}_i \in \mathcal{S}_{k,\mathcal{G}}$

$$\mathbf{x} = \sum_{i=1}^{N} c_i\mathbf{v}_i \;.$$

Since $\mathbf{v}_i \in \mathcal{S}_{k,\mathcal{G}}$, its support can be covered by at most k groups from $\mathcal{G}$, which means there is a supergroup $H_i \in \mathcal{G}_k$ such that $\mathrm{supp}(\mathbf{v}_i) \subseteq H_i$. From (7), we have

$$\begin{aligned}\|\mathbf{x}\|_{k,\mathcal{G}} &\le \sum_{H\in\mathcal{G}_k} \Big\| \sum_{i:\, H_i=H} c_i\mathbf{v}_i \Big\| \\ &\le \sum_{H\in\mathcal{G}_k} \sum_{i:\, H_i=H} c_i\|\mathbf{v}_i\| = \sum_{i=1}^{N} c_i\|\mathbf{v}_i\| \le \sum_{i=1}^{N} c_i = 1,\end{aligned}$$

completing the proof. □

A key consequence of viewing $\|\cdot\|_{k,\mathcal{G}}$ as an atomic norm is that its dual can also be defined in terms of atoms:

$$\|\mathbf{x}\|_{k,\mathcal{G}}^* = \sup_{\mathbf{u}:\|\mathbf{u}\|_{k,\mathcal{G}}=1} \mathbf{u}^T\mathbf{x} = \sup_{\mathbf{a}\in\mathscr{A}} \mathbf{a}^T\mathbf{x} = \sup_{H\in\mathcal{G}_k} \|\mathbf{x}_H\|. \quad (10)$$

Here, $\mathbf{x}_H$ denotes the block of $\mathbf{x}$ consisting of the coordinates listed in H. The formulation and computation of the dual plays a central role in the algorithms for solving (8).

3. OPTIMIZATION ALGORITHMS

We solve (8) using the conditional gradient (CG) approach [9], outlined in Algorithm 1. CG methods have recently gained popularity due to their ease of implementation and applicability to a wide range of structurally constrained optimization problems [18, 8, 10, 12].

Algorithm 1 Conditional Gradient Method

```
1: Inputs: data $\mathbf{\Phi}$, $\mathbf{y}$; groups $\mathcal{G}$, sparsity $k$, constraint $\tau$;
     step size sequence $\{\gamma^t\}_{t=1}^{\infty}$
2: let $F(\mathbf{x}) = f(\mathbf{\Phi x},\, \mathbf{y})$ denote the objective
3: set initial iterate $\mathbf{x}^0 = \mathbf{0}$
4: for $t = 1, 2, \ldots$ do
5:     evaluate gradient $\nabla F(\mathbf{x}^{t-1})$
6:     find $\mathbf{a}^t = \operatorname{argmax}_{\mathbf{a}\in\mathscr{A}} \mathbf{a}^T \nabla F(\mathbf{x}^{t-1})$
7:     update $\mathbf{x}^t = (1-\gamma^t)\mathbf{x}^{t-1} + \tau\gamma^t\mathbf{a}$
8: end for
```

A key step in Algorithm 1 is step 6, which involves optimizing a linear objective over $\mathscr{A}$ to find the next atom to add to the current representation $\mathbf{x}^{t-1}$. In fact, it is the same optimization as in the definition of the dual norm (10), and hence solving step 6 is equivalent to computing the dual norm of the gradient of the objective F at $\mathbf{x}^{t-1}$. While the computation of the dual norm of overlapping group lasso is NP hard in general [2], our supergroups have a specific structure which allows us to solve the optimization problem efficiently. We first show how to efficiently implement step 6 when the groups in $\mathcal{G}$ do not overlap, and then show how the problem with the overlapping groups can be re-cast as the problem with disjoint groups without any loss of computational efficiency.

3.1. Computing the Dual Norm for Disjoint Groups

According to (10), the dual norm can be computed by considering all supergroups $H \in \mathcal{G}_k$ and picking the one which maximizes the ℓ_2 norm of its block $\mathbf{x}_H$. If the number of original groups is M, the number of supergroups is $M^{\Omega(k)}$, so a brute-force search would be prohibitive. However, since the original groups are disjoint, we have the following additive decomposition of the squared norm of the block $H = \cup_{i=1}^{k} G_i$:

$$\|\mathbf{x}_H\|^2 = \sum_{i=1}^{k} \|\mathbf{x}_{G_i}\|^2 \ .$$

So to maximize the norm across all k-unions of groups in $\mathcal{G}$, it suffices to pick the union of the k groups with the largest values of $\|\mathbf{x}_G\|$, which is implemented in Algorithm 2.

Algorithm 2 Atom Selection for Disjoint Groups

```
1: Inputs: vector $\mathbf{x}$; disjoint groups $\mathcal{G}$, sparsity $k$
2: let $G_1, \ldots, G_k$ be the $k$ groups with the largest $\|\mathbf{x}_G\|$
3: let $H = \cup_{i=1}^{k} G_i$
4: return $\begin{cases} \mathbf{x}_H/\|\mathbf{x}_H\| & \text{if } \mathbf{x}_H \neq \mathbf{0}, \\ \text{an arbitrary atom} & \text{if } \mathbf{x}_H = \mathbf{0}. \end{cases}$
```

Algorithm 2 is highly efficient. It runs in time $O(p)$ since the dominant cost is the initial calculation of norms $\|\mathbf{x}_G\|$ for the groups $G \in \mathcal{G}$ and the groups are disjoint. Determining the top k groups can then be done in time $O(M)$, and thanks to the disjointness, we have $M \leq p$.

3.2. Solving the Problem for Overlapping Groups

For overlapping groups, we do not aim to compute the dual norm, but instead we express the original overlapping problem as a non-overlapping problem by replicating the variables (in the same spirit as [11]), and only then apply the conditional gradient. Specifically, given groups $\mathcal{G} = \{G_1, \ldots, G_M\}$, and data $\mathbf{\Phi} \in \mathbb{R}^{n\times p}$, we define a new data matrix

$$\mathbf{\Phi}' = [\mathbf{\Phi}_{G_1},\, \mathbf{\Phi}_{G_2}, \ldots, \mathbf{\Phi}_{G_M}] \in \mathbb{R}^{n\times\sum_{i=1}^{M}|G_i|}$$

where $\mathbf{\Phi}_G$ is the submatrix of $\mathbf{\Phi}$ whose columns are indexed by group G. Letting $\mathbf{x}'$ be the new parameter vector and $\mathcal{G}'$ be the new groups, corresponding to the blocks $\mathbf{\Phi}_{G_i}$ within $\mathbf{\Phi}'$, the original problem is equivalent to

$$\hat{\mathbf{x}}' = \operatorname*{argmin}_{\mathbf{x}'} f(\mathbf{\Phi}'\mathbf{x}',\, \mathbf{y}) \ \text{ s.t. } \ \|\mathbf{x}'\|_{k,\mathcal{G}'} \leq \tau. \qquad (11)$$

However, since the groups in $\mathcal{G}'$ do not overlap, we can solve (11) by Algorithm 2. The resulting $\hat{\mathbf{x}}'$ will be a vector of length $p' = \sum_{i=1}^{M}|G_i|$, and one can recover $\hat{\mathbf{x}} \in \mathbb{R}^p$ by adding up the replicated variables.

Note that this approach can be implemented without explicitly constructing $\mathbf{\Phi}'$. Let $F(\mathbf{x}) = f(\mathbf{\Phi x},\, \mathbf{y})$, $F'(\mathbf{x}') = f(\mathbf{\Phi}'\mathbf{x}',\, \mathbf{y})$, and let $\mathbf{M} \in \mathbb{R}^{p\times p'}$ be the 0-1 matrix where each column contains exactly a single 1, indicating which old coordinate corresponds to the given new coordinate. Thus, $\mathbf{\Phi}' = \mathbf{\Phi M}$, so we can write $F'(\mathbf{x}') = F(\mathbf{Mx}')$. The gradient of F' is then $\nabla F'(\mathbf{x}') = \mathbf{M}^T(\nabla F(\mathbf{Mx}'))$. To calculate the gradient $\nabla F'$ in step 5 of Algorithm 1, it suffices to first determine $\mathbf{x} = \mathbf{Mx}'$ in time $O(p')$; then calculate $\nabla F(x)$, which will be the dominant cost, typically on the order $O(np)$ or the number of non-zeros in $\mathbf{\Phi}$; then multiply by $\mathbf{M}^T$ in time $O(p')$; and finally calculate the dual norm by Algorithm 2 in time $O(p')$.

It is important to remark that the replication strategy will increase the variable size to p' and not $\sum_{H\in\mathcal{G}_k}|H|$, which would be prohibitively large and in fact infeasible. In most problems of interest, $p' = O(p)$ or possibly a small polynomial in p, and the additional price we pay in per-iteration running time is negligible, as it tends to be dominated by the calculation of the gradient ∇F.

3.3. Theoretical Considerations

In the case of the k-support norm, Argyriou et al. [1] showed that one can hope to obtain at most a constant-factor improvement over the sample complexity of the lasso. A similar argument holds in the case of the group k-support norm, and we omit details here due to space constraints.

4. EXPERIMENTS

We compare the group lasso (GL) with the group elastic net (GEN) and the group k-support norm (GKS). We first consider a multilabel learning problem where the groups do not overlap. We then consider two problems with overlapping groups: compressed sensing for signal recovery and gene selection in computational biology. In all the experiments, we use the least-squares loss function. All methods have the regularization parameter τ, whereas GKS requires the additional parameter k, and GEN requires the additional parameter λ (the cooefficient for the ℓ_2 penalty).

4.1. Non-Overlapping Groups: Multilabel Learning

In multilabel learning, the goal is to predict a vector in $\{0, 1\}^\ell$, where ℓ is the number of labels, based on a d-dimensional covariate vector. We predict each label by a separate linear model, so the dimensionality of the full parameter vector is $p = \ell d$. If the original number of multilabel examples is n_0, we effectively create $n = n_0 \ell$ binary classification examples. We use the bibtex database[2] and apply the traditional multilabel group lasso regularization, with d groups, each containing the parameters corresponding to the same covariate. The original data consists of $n_0 = 4880$ examples in $d = 1836$ dimensions, with $\ell = 159$ labels. A test set with 2515 examples is also provided. We compare the group k-support framework to the traditional group lasso [5]. We vary the regularization parameter $\tau \in \{2^5, 2^6, \ldots, 2^{10}\}$, $k \in \{1, 2, \ldots, 10\}$ and $\lambda \in \{2^{-3}, 2^{-2}, \ldots, 2^1\}$. Table 1 shows that the group k-support norm yields superior performance. Since each example will only have a small number of positive labels, we display the area under the ROC curve, which is not sensitive to this "class imbalance".

Method	**Train AUC**	**Test AUC**
GL	0.5437	0.5433
GEN	0.6261	0.6254
GKS	**0.7159**	**0.7087**

Table 1. Multilabel learning on the BIBTEX dataset. GKS yields significantly better results than GL and GEN.

4.2. Overlapping Groups: Wavelet Signal Recovery

For the overlapping groups case, we use our method to recover wavelet coefficients of signals, utilizing the same group structure as Rao et al. [19]. The task is the standard (least squares) compressed sensing. Table 2 shows the mean square error (MSE) obtained as a result of solving the compressed sensing problem with the group k-support and the group lasso regularizers. We use one-dimensional test signals popular in the signal processing community, keeping the signal length at $p = 4096$ and obtaining $n = 820$ i.i.d. Gaussian measurements, corrupted by Gaussian noise with standard deviation $\sigma = 0.05$. We clairvoyantly selected the best regularization parameters τ, k, λ for all the methods, since we had access to the ground truth.

Signal	**GL**	**GEN**	**GKS**
HeaviSine	0.0114	0.0013	**0.0005**
Blocks	0.0127	0.0007	**0.0003**
Piece-Reg	0.0072	0.0011	**0.0004**
Piece-Poly	0.0103	0.0012	**0.0010**

Table 2. Performance of the group k-support norm for signal recovery. GKS consistently outperforms the other methods.

4.3. Overlapping Groups: Gene Selection

Finally, we consider a classification problem in computational biology. We use the breast cancer dataset previously studied by Jacob et al. [11],[3] where the task is to predict whether a particular patient's tumor will undergo metastasis. The groups correspond to overlapping gene pathways.

The data contains 295 examples in 3510 dimensions (after retaining only the dimensions that appear in at least one pathway). The dataset is unbalanced, with 3 times as many negative examples as the positive ones. We first randomly retain 20% of the positive and negative examples in a test set. To balance the remainder of the data (to be used for training), we replicate the positive examples two more times, to yield the final training set of size 363×3510.

We vary $k \in \{1, 2, \ldots, 15\}$, $\tau \in \{2^5, 2^6, \ldots, 2^{15}\}$, and $\lambda \in \{2^{-3}, 2^{-2}, \ldots, 2^1\}$. We solve the least squares problem, then set $\hat{y} = \text{sign}(\boldsymbol{\phi}^T \hat{\mathbf{x}})$, and report the misclassification error on the test set in Table 3. GKS again outperforms both the standard group lasso and the group elastic net.

Method	GL	GEN	GKS
Error	0.3448	0.3310	**0.3276**

Table 3. Comparison on a Breast Cancer dataset

5. CONCLUSIONS AND DISCUSSION

We introduced and analyzed the group k-support norm, for both overlapping and non-overlapping groups. We derived an efficient conditional gradient algorithm to solve the group k-support-norm-constrained convex program. Our experiments on a number of data sets show that the group k-support norm outperforms both the standard (overlapping) group lasso and the group elastic net.

[2] http://mulan.sourceforge.net/datasets-mlc.html

[3] http://cbio.ensmp.fr/~ljacob/documents/overlasso-package.tgz

6. REFERENCES

[1] A. Argyriou, R. Foygel, and N. Srebro. Sparse prediction with the k-support norm. In *Advances in Neural Information Processing Systems*, pages 1457–1465, 2012.

[2] N. Bhan, L. Baldassarre, and V. Cevher. Tractability of interpretability via selection of group-sparse models. In *Information Theory Proceedings (ISIT), 2013 IEEE International Symposium on*, pages 1037–1041. IEEE, 2013.

[3] M. Blaschko. A note on k-support norm regularized risk minimization. *arXiv preprint arXiv:1303.6390*, 2013.

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

[5] S. S. Bucak, R. Jin, and A. K. Jain. Multi-label learning with incomplete class assignments. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 2801–2808. IEEE, 2011.

[6] V. Chandrasekaran, B. Recht, P. A. Parrilo, and A. S. Willsky. The convex geometry of linear inverse problems. *Foundations of Computational mathematics*, 12(6):805–849, 2012.

[7] S. Chatterjee, S. Chen, and A. Banerjee. Generalized dantzig selector: Application to the k-support norm. In *Advances in Neural Information Processing Systems*, pages 1934–1942, 2014.

[8] M. Dudik, Z. Harchaoui, and J. Malick. Lifted coordinate descent for learning with trace-norm regularization. In *AISTATS-Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics-2012*, volume 22, pages 327–336, 2012.

[9] M. Frank and P. Wolfe. An algorithm for quadratic programming. *Naval research logistics quarterly*, 3(1-2):95–110, 1956.

[10] Z. Harchaoui, A. Juditsky, and A. Nemirovski. Conditional gradient algorithms for norm-regularized smooth convex optimization. *Mathematical Programming*, pages 1–38, 2014.

[11] L. Jacob, G. Obozinski, and J.-P. Vert. Group lasso with overlap and graph lasso. In *Proceedings of the 26th annual international conference on machine learning*, pages 433–440. ACM, 2009.

[12] M. Jaggi. Revisiting frank-wolfe: Projection-free sparse convex optimization. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, pages 427–435, 2013.

[13] K. Lounici, M. Pontil, A. B. Tsybakov, and S. Van De Geer. Taking advantage of sparsity in multi-task learning. *COLT*, 2009.

[14] A. McDonald, M. Pontil, and D. Stamos. Spectral k-support regularization. *Advances in Neural Information Processing Systems*, 28:3644–3652, 2014.

[15] A. M. McDonald, M. Pontil, and D. Stamos. New perspectives on k-support and cluster norms. *arXiv preprint arXiv:1512.08204*, 2015.

[16] M. Misyrlis, A. Konova, M. Blaschko, J. Honorio, N. Alia-Klein, R. Goldstein, and D. Samaras. Predicting cross-task behavioral variables from fMRI data using the k-support norm. In *Sparsity Techniques in Medical Imaging (STMI)*, 2014.

[17] N. Rao, P. Shah, and S. Wright. Forward–backward greedy algorithms for atomic norm regularization. *Signal Processing, IEEE Transactions on*, 63(21):5798–5811, 2015.

[18] N. Rao, P. Shah, S. Wright, and R. Nowak. A greedy forward-backward algorithm for atomic norm constrained minimization. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 5885–5889. IEEE, 2013.

[19] N. S. Rao, R. D. Nowak, S. J. Wright, and N. G. Kingsbury. Convex approaches to model wavelet sparsity patterns. In *Image Processing (ICIP), 2011 18th IEEE International Conference on*, pages 1917–1920. IEEE, 2011.

[20] N. S. Rao, B. Recht, and R. D. Nowak. Universal measurement bounds for structured sparse signal recovery. In *International Conference on Artificial Intelligence and Statistics*, pages 942–950, 2012.

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

[22] 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.