Noname manuscript No.
(will be inserted by the editor)

Convex Relaxations of Penalties for Sparse Correlated Variables With Bounded Total Variation

Eugene Belilovsky · Andreas Argyriou · Gaël Varoquaux · Matthew Blaschko

Received: date / Accepted: date

Abstract We study the problem of statistical estimation with a signal known to be sparse, spatially contiguous, and containing many highly correlated variables. We take inspiration from the recently introduced k-support norm, which has been successfully applied to sparse prediction problems with correlated features, but lacks any explicit structural constraints commonly found in machine learning and image processing. We address this problem by incorporating a total variation penalty in the k-support framework. We introduce the (k, s) *support total variation norm* as the *tightest* convex relaxation of the intersection of a set of sparsity and total variation constraints. We show that this norm leads to an intractable combinatorial graph optimization problem, which we prove to be NP-hard. We then introduce a tractable relaxation with approximation guarantees that scale well for grid structured graphs. We devise several first-order optimization strategies for statistical parameter estimation with the described penalty. We demonstrate the effectiveness of this penalty on classification in the low-sample regime, classification with M/EEG neuroimaging data, and image recovery with synthetic and real data background subtracted image recovery tasks. We extensively analyse the application of our penalty on the complex task of identifying predictive regions from low-sample high-dimensional fMRI brain data, we show that our method is particularly useful compared to existing methods in terms of accuracy, interpretability, and stability.

Keywords Structured sparsity · Feature selection · Brain decoding · k-support · Total variation

This work is partially funded by ERC Grant 259112, FP7-MC-CIG 334380, FP7 2007-2013 Grant 246556, and DIGITEO 2013-0788D - SOPRANO.

Eugene Belilovsky
E-mail: eugene.belilovsky@inria.fr
Telephone: +33 141131098

1 Introduction

Regularization methods utilizing the ℓ_1 norm such as Lasso (Tibshirani, 1996) have been used widely for feature selection. They have been particularly successful at learning problems in which very sparse models are required. However, in many problems a better approach is to balance sparsity against an ℓ_2 constraint. One reason is that very often features are correlated and it may be better to combine several correlated features than to select fewer of them, in order to obtain a lower variance estimator and better interpretability. This has led to the method of *elastic net* in statistics (Zou and Hastie, 2005), which regularizes with a weighted sum of ℓ_1 and ℓ_2 penalties. More recently, it has been shown that the elastic net is not in fact the tightest convex penalty that approximates sparsity (ℓ_0) and ℓ_2 constraints at the same time (Argyriou et al, 2012). The tightest convex penalty is given by the k-support norm, which is parametrized by an integer k, and can be computed efficiently. This norm has been successfully applied to a variety of sparse vector prediction problems (Gkirtzou et al, 2013; McDonald et al, 2014; Misyrlis et al, 2014).

We study the problem of introducing structural constraints to sparsity and ℓ_2, from first principles. In particular, we seek to introduce a total variation smoothness prior in addition to sparsity and ℓ_2 constraints. Total variation is a popular regularizer used to enforce local smoothness in a signal (Michel et al, 2011; Rudin et al, 1992; Tibshirani et al, 2005). It has successfully been applied in image de-noising and has recently become of particular interest in the neural imaging community where it can be used to reconstruct sparse but locally smooth brain activation (Baldassarre et al, 2012b; Michel et al, 2011).Two kinds of total variation are commonly considered in the literature, isotropic $TV_I(w) = \|\nabla w\|_{2,1}$ and anisotropic $TV_A(w) = \|\nabla w\|_1$ (Beck and Teboulle, 2009). In our theoretical analysis we focus on the anisotropic penalty.

To derive a penalty incorporating these constraints we follow the approach of (Argyriou et al, 2012) by taking the convex hull of the intersection of our desired penalties and then recovering a norm by applying the gauge function. We then derive a formulation for the dual norm which leads us to a combinatorial optimization problem, which we prove to be NP-hard. We find an approximation to this penalty and prove a bound on the approximation error. Since the k-support norm is the tightest relaxation of sparsity and ℓ_2 constraints, we propose to use the intersection of the TV norm ball and the k-support norm ball. This leads to a convex optimization problem in which (sub)gradient computation can be achieved with a computational complexity no worse than that of the total variation. Furthermore, our approximation can be computed for variation on an arbitrary graph structure.

We discuss and utilize several first order optimization schemes including stochastic subgradient descent, iterative Nesterov-smoothing methods, and FISTA with an estimated proximal operator. We demonstrate the tractability and utility of the norm through applications of classification on MNIST with few samples, M/EEG classification, and background-subtracted image recovery. For the problem of identifying predictive regions in fMRI we show that we can get improved accuracy, stability, and interpretability along with providing the user with several potential tools and heuristics to visualize the resulting predictive models. This includes

several interesting properties that apply to the special case of k-support norm optimization as well.

2 Convex Relaxation of Sparsity, ℓ_2 and Total Variation

In this section we formulate the (k,s) *support total variation* norm, a tight convex relaxation of sparsity, ℓ_2, and total variation (TV) constraints. We derive its dual norm which results in an intractable optimization problem. Finally we describe a looser convex relaxation of these penalties which leads to a tractable optimization problem.

2.1 Derivation of the Norm

We start by defining the set of points corresponding to simultaneous sparsity, ℓ_2 and total variation (TV) constraints:

$$Q^2_{k,s} := \{w \in \mathbb{R}^d : \|w\|_0 \leq k, \|w\|_2 \leq 1, \|Dw\|_0 \leq s\}$$

where $k \in \{1,\dots,d\}, s \in \{1,\dots,m\}$ and $D \in \mathbb{R}^{m\times d}$ is a prescribed matrix. The bound of one on the ℓ_2 term is used for convenience since the cardinality constraints are invariant under scaling. D generally take the form of a discrete difference operator, but the discussion in the following sections is more general than that. It is easy to see that the set $Q^2_{k,s}$ is not convex due to the presence of the $\|\cdot\|_0$ terms. Hence using $Q^2_{k,s}$ in a regularization method is impractical. Thus we consider instead the convex hull of $Q^2_{k,s}$:

$$C^2_{k,s} := \mathrm{conv}(Q^2_{k,s}) = \Big\{w : w = \sum_{i=1}^{r} c_i z_i, \sum_{i=1}^{r} c_i = 1, c_i \geq 0, z_i \in \mathbb{R}^d, \\ \|z_i\|_0 \leq k, \|z_i\|_2 \leq 1, \|Dz_i\|_0 \leq s, r \in \mathbb{N}\Big\}.$$

For some values of D, k and s, this convex set may not span the entire $\mathbb{R}^d$, that is, it may be contained within a smaller subspace. In Section 2.2 we show a condition for which the set will span $\mathbb{R}^d$ (see Proposition 1). For a matrix D that is the transpose of an incidence matrix representing a graph with a maximum degree of l_{deg}, the value of s should be greater than or equal to l_{deg}.

Assuming some mild technical conditions on D,[1] the convex set $C^2_{k,s}$ is the unit ball of a certain norm. We call this norm the (k,s) *support total variation* norm. It equals the gauge function of $C^2_{k,s}$, that is,

$$\|x\|^{sptv}_{k,s} := \inf\Big\{\lambda \in \mathbb{R}_+ : x = \lambda \sum_{i=1}^{r} c_i z_i, \sum_{i=1}^{r} c_i = 1, \\ c_i \geq 0, z_i \in \mathbb{R}^d, \|z_i\|_0 \leq k, \|z_i\|_2 \leq 1, \|Dz_i\|_0 \leq s, r \in \mathbb{N}\Big\}. \quad (1)$$

Performing a variable substitution we define a set of components of x, $v_i = \lambda c_i z_i \Rightarrow \lambda = \frac{\sum_{i=1}^{r} \|v_i\|_2}{\sum_{i=1}^{r} c_i \|z_i\|_2}$. To maximize the denominator for fixed v_i, we note that

[1] The conditions are given in Proposition 1.

$\sum_{i=1}^{r} c_i \|z_i\|_2 \leq \left(\sum_{i=1}^{r} c_i\right) \max_{i=1}^{r} \|z_i\|_2 = 1$. The equality can be attained by applying the constraints in Equation (1). Substituting for λ and removing the constraints already applied above our norm now becomes

$$\|x\|_{k,s}^{sptv} = \inf\left\{\sum_{i=1}^{r} \|v_i\|_2 : \sum_{i=1}^{r} v_i = x, \|v_i\|_0 \leq k, \|Dv_i\|_0 \leq s, r \in \mathbb{N}\right\}. \quad (2)$$

The special case $s = m$ is simply the k-support norm (Argyriou et al, 2012), which trades off between the ℓ_1 norm ($k = 1, s = m$) and the ℓ_2 norm ($k = d, s = m$). Formula 2 is combinatorial in nature and hence is difficult to directly include in an optimization problem.

2.2 Derivation of the Dual Norm

A standard approach for analyzing structured norms is through analysis of the dual norm (Argyriou et al, 2012; Bach et al, 2012; Mairal and Yu, 2013). As such, it will be useful to derive an expression for the dual norm of $\|\cdot\|_{k,s}^{sptv}$. This will allow us to connect the norm with an optimization problem on a graph, use this to show the norm is NP-hard, and to derive an approximation bound (Proposition 2).

To obtain the dual of (k, s) support TV norm we first consider a more general class of norms. Each norm in this class is associated with a set of subspaces $S_1, \ldots, S_n$ and a set of norms $\|\cdot\|_{(1)}, \ldots, \|\cdot\|_{(n)}$. We assume that these subspaces span $\mathbb{R}^d$, that is, $\sum_{i=1}^{n} S_i = \mathbb{R}^d$, the summation here denotes addition of sets ($S_1 + S_2 = \{x : x = x_1 + x_2, x_1 \in S_1, x_2 \in S_2\}$). We may now define the following norm

$$\|w\| := \min\left\{\sum_{i=1}^{n} \|v_i\|_{(i)} \ : \ v_i \in S_i, \ \forall i \in \mathbb{N}_n, \ \sum_{i=1}^{n} v_i = w\right\} \ \forall w \in \mathbb{R}^d. \quad (3)$$

This is indeed a norm, since the subspaces span $\mathbb{R}^d$, and that the above minimum is attained. The (k, s) support TV norms can be written in the form (3) by specifying all n norms to be the ℓ_2 norm and the linear subspaces to correspond to the constraints on the supports.

We note that this definition is equivalent to an infimal convolution of n norms. Let δ_S denote the indicator function of a subspace S and the infimal convolution $(f_1 \,\square\, \ldots \,\square\, f_n)$ of n functions as $\square_{i=1}^{n} f_i$. Using this notation, the norm $\|\cdot\|$ can be written equivalently as $\|\cdot\| = \square_{i=1}^{n} \left(\|\cdot\|_{(i)} + \delta_{S_i}\right)$. We may derive the general form of the dual norm $\|\cdot\|^*$ of $\|\cdot\|$ by a direct application of standard duality results from convex analysis.

Lemma 1 *Let $\|\cdot\|_{(1)}, \ldots, \|\cdot\|_{(n)}$ be norms on $\mathbb{R}^d$ with duals $\|\cdot\|_{(1)*}, \ldots, \|\cdot\|_{(n)*}$, respectively, and let $S_1, \ldots, S_n$, be linear subspaces of $\mathbb{R}^d$ such that $\sum_{i=1}^{n} S_i = \mathbb{R}^d$. Then the dual norm of $\|\cdot\|$ defined in (3) is given by*

$$\|u\|^* = \max_{i=1}^{n} \min\left\{\|u - q\|_{(i)*} : q \in S_i^{\perp}\right\} = \max_{i=1}^{n} \left(\|\cdot\|_{(i)*} \,\square\, \delta_{S_i^{\perp}}\right)(u) \quad (4)$$

for all $u \in \mathbb{R}^d$. The unit ball of $\|\cdot\|^*$ equals $B_* = \bigcap_{i=1}^{n} \left(B_{i*} + S_i^{\perp}\right)$ where B_{i*} denotes the unit ball of $\|\cdot\|_{(i)*}$ for $i = 1, \ldots, n$.

Proof. Denote *convex conjugate* or *Fenchel conjugate* of a function $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ by f^* (Bauschke and Combettes, 2011). It is known that the convex conjugate of a norm equals the indicator function of its dual unit ball. Thus it holds that

$$\delta_{B_*} = \left(\square_{i=1}^{n} \left(\|\cdot\|_{(i)} + \delta_{S_i} \right) \right)^* .$$

Moreover, the conjugate of an infimal convolution equals the sum of conjugates (Bauschke and Combettes, 2011, Prop. 13.21). The converse duality also holds under Slater type conditions (Bauschke and Combettes, 2011, Thm. 15.3). Applying these facts successively, we obtain that

$$\delta_{B_*} = \sum_{i=1}^{n} \left(\|\cdot\|_{(i)} + \delta_{S_i} \right)^* = \sum_{i=1}^{n} \left(\|\cdot\|_{(i)}^* \,\square\, \delta_{S_i}^* \right) = \sum_{i=1}^{n} \left(\delta_{B_{i*}} \,\square\, \delta_{S_i}^* \right) .$$

We now use the facts that, for any subspace S, $\delta_S^* = \delta_{S^\perp}$ and that, for any nonempty sets $C, D \subseteq \mathbb{R}^d$, $\delta_C \,\square\, \delta_D = \delta_{C+D}$, obtaining that

$$\delta_{B_*} = \sum_{i=1}^{n} \left(\delta_{B_{i*}} \,\square\, \delta_{S_i^\perp} \right) = \sum_{i=1}^{n} \left(\delta_{B_{i*}+S_i^\perp} \right) .$$

It follows that $B_* = \bigcap_{i=1}^{n} \left(B_{i*} + S_i^\perp \right)$. The intersection of norm balls corresponds to maximum of the corresponding norms which gives the formula for $\|\cdot\|^*$. □

Equation (4) for the dual norm is interpreted as the maximum of the distances of x (with respect to the corresponding dual norms) from the orthogonal complements. We now specialize this formula to the case of (k,s) support TV norm.

Notation We define G_k as all subsets of $\{1,...,d\}$ of cardinality at most k and M_s as all subsets of $\{1,...,m\}$ of cardinality at most s. For every $I \in G_k$, we denote $I^c = \{1,...,d\} \backslash I$ and for every $J \in M_s$, $J^c = \{1,...,m\} \backslash J$. We denote D_{J^c} as the submatrix of D with only the rows indexed by J^c and for every $u \in \mathbb{R}^d$, u_I is the subvector of u with only the elements indexed by I.

It is the case that r in Equation (2) can be assumed to be at most $|G_k||M_s|$ (by grouping components with the same (I,J) pattern and applying the triangle inequality). We can now reduce the dual norm to

$$(\|x\|_{k,s}^{sptv})^* = \max_{(I,J) \in G_k \times M_s} \min\{\|x-q\|_2 : q \in S_{I,J}^\perp\} = \max_{(I,J) \in G_k \times M_s} E_{I,J}(x) \quad (5)$$

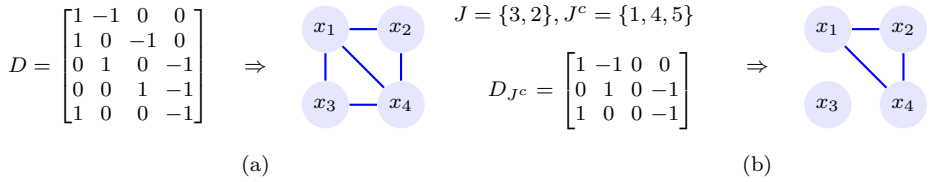
where $S_{I,J} = \{x \,|\, D_{J^c} x = 0 \text{ and } x_{I^c} = 0\}$, $S_{I,J}^\perp = \text{range}(D_{J^c}^\top) + \{x \,|\, x_I = 0\}$, and $E_{I,J}$ is an energy function we will derive (cf. Equation (6)). Before proceeding we use the described subspaces to note the conditions for which $\|x\|_{k,s}^{sptv}$ is a full fledged norm

Proposition 1 *If*

$$\sum_{\substack{I \subseteq \{1,\ldots,d\}, |I|=k \\ J \subseteq \{1,\ldots,m\}, |J|=s}} S_{I,J} = \mathbb{R}^d$$

then $\operatorname{span} C_{k,s}^2 = \mathbb{R}^d$.

This condition will depend on the choice of D, k and s. We choose D to be the transpose of the incidence matrix of a directed graph $G_d = (\mathcal{V}_d, \mathcal{E}_d)$, with the vertices corresponding to the elements of x. Furthermore $G = (\mathcal{V}, \mathcal{E})$ is an undirected graph with vertices $\mathcal{V} = \mathcal{V}_d$ and an unordered set of the same edges as $\mathcal{E}_d$. For a given J, we can consider the graph G_{J^c}, specified by the incidence matrix D_{J^c} as the original graph with $|J|$ edges removed. The notation presented is illustrated in Figure 1.



(a) (b)

Fig. 1: (a) Example of D matrix for a graph, and (b) an example D_{J^c} for a given instance of J. The graph in (b) has two subgraphs, one with nodes x_1, x_2, x_4 and the other the singleton, x_3

We consider the linear constraints specified by $D_{J^c}x = 0$. Each row of the transpose incidence matrix, D_{J^c}, represents an edge $\mathcal{E}_{dij} = (i, j)$. Coupled with the constraint each of these rows corresponds to a constraint $x_i = x_j$. We note that this constraint is independent of the ordering on the graph. For any two vertices a, b of the undirected graph G, if there exists a path between a and b then $x_a = x_b$. More formally, if we divide G_{J^c} into all of its disjoint subgraphs denoted by $G_\gamma = (\mathcal{V}_\gamma, \mathcal{E}_\gamma)$,

$$G_{J^c} = \bigcup_{\gamma \in \Gamma} G_\gamma, \qquad (a, b) \in \mathcal{V}_\gamma \times \mathcal{V}_\gamma \Rightarrow x_a = x_b \ .$$

Thus for any disjoint subgraph of G_{J^c} we can take any tree containing the vertices of the subgraph and the associated incidence matrix will be a representation of the subspace associated with the components represented by those vertices.

Since each disjoint subgraph will have an independent set of constraints on its associated variables we can subdivide the linear constraints specifying $S^\perp_{I,J}$. Divide the graph corresponding to D_{J^c} into all disjoint subgraphs enumerated by $\Gamma = \{1, ...p\}$. Let $D_{J^c_\gamma}$ be the incidence matrix corresponding to each subgraph. Then $S_{I,J} = \left\{ x \,|\, D_{J^c_\gamma} x = 0, \forall \gamma \in \Gamma, \text{ and } x_{I^c} = 0 \right\}$ and $S^\perp_{I,J} = \sum_{\gamma \in \Gamma} \text{range}(D^\top_{J^c_\gamma}) + \{x \,|\, x_I = 0\}$. A direct computation yields the projection on each subgraph V_γ as

$$P_\gamma = D_{J^c_\gamma}(D_{J^c_\gamma})^+ = \mathbf{I} - \tfrac{1}{n_\gamma}\mathbf{1}$$

if the subgraph has n_γ vertices. P_γ is exactly a centering matrix that projects orthogonal to the vector of all ones.

To compute the value of $E_{I,J}(x)$ we can split the parameters of x into independent groups, since the projection and thereby the residual of components corresponding to vertices in disjoint groups will have independent contribution. The components of $\text{Proj}_{S_{I,J}}(x)$ at I^c must be zero. Moreover, the members of

any group that contains a vertex from I^c will be zero. We can therefore compute $E_{I,J}(x)$ independently for each disjoint group, and only for the groups that do not contain a vertex in I^c. For each disjoint group the contribution to $E^2_{I,J}(x)$ is

$$E^2_\gamma(x) = \|(\mathbf{I} - P_\gamma)x\|^2 = \frac{1}{n_\gamma}\left(\sum_{i\in V_\gamma} x_i\right)^2 . \tag{6}$$

A graph based version of the combinatorial optimization problem is as follows. Given an undirected graph $G = (\mathcal{V}, \mathcal{E})$ and $I \subset \mathcal{V}$, $J \subset \mathcal{E}$, remove edges J and all disjoint subgraphs containing a vertex in I^c to obtain a graph G_{IJ}. The energy over this graph, $E^2_{I,J}$ can be computed as the sum of E^2_γ over all disjoint subgraphs in G_{IJ}. The dual norm is then given by Equation (5).

We can additionally show that we can limit M_s to maximum cardinality sets (cardinality s) and G_k to maximum cardinality sets (cardinality k). Indeed, adding indexes in I or J cannot decrease $S_{I,J}$ and hence cannot decrease the norm of the projection in Equation (5). Thus we can narrow the problem to removing s edges and $d-k$ nodes (with their associated subgraphs).

We have now reduced the computation of the dual norm to a graph partitioning problem. Graph partition problems are often NP-hard, and we show this to be the case here as well:

Theorem 1 *Computation of the (k,s) support total variation dual norm is NP-hard*

The proof of Theorem 1 is given in Appendix A.

Corollary 1 *Computation of the (k,s) support total variation norm is NP-hard.*

In light of this Theorem, we are unable to incorporate the (k,s) support total variation norm in a regularized risk setting. Instead in the sequel we examine a tractable approximation with bounds that scale well for the family of graphs of interest.

2.3 Approximating the Norm

Although special cases where s equals m or 1 are tractable, the general case for arbitrary values of s leads to an NP-hard graph partitioning problem for the dual norm, implying the norm itself is intractable. We thus relax the problem by taking instead the intersection of the k-support norm ball and the convex relaxation of total variation. This leads to the following penalty

$$\Omega_{sptv}(w) = \max\{\|w\|^{sp}_k, \tfrac{1}{\sqrt{s}\|D\|}\|Dw\|_1\} \tag{7}$$

where $\|\cdot\|$ denotes the spectral norm. We can bound the error of this approximation as follows:

Proposition 2 *For every $w \in \mathbb{R}^d$, it holds that*

$$\Omega_{sptv}(w) \leq \|w\|^{sptv}_{k,s} . \tag{8}$$

Moreover, suppose that $\text{range}(D^\top) = \mathbb{R}^d$ *and that for every* $I \in G_k$ *the submatrix* D_{*I} *has at least* $m - s$ *zero rows. Then it holds that*

$$\|w\|_{k,s}^{sptv} \leq \sqrt{1 + \frac{s\|D\|^2\|(D^\top)^+\|_\infty^2}{k}}\ \Omega_{sptv}(w) \tag{9}$$

where $\|\cdot\|_\infty$ *is the norm on* $\mathbb{R}^{m\times d}$ *induced by* ℓ_∞*, that is,* $\|A\|_\infty = \max_{i=1}^{m} \sum_{j=1}^{d} |A_{ij}|$.

Proof. First, note that $\|\cdot\|_k^{sp} \leq \|\cdot\|_{k,s}^{sptv}$. This follows directly from the definition of $\|\cdot\|_{k,s}^{sptv}$, since

$$\|w\|_k^{sp} = \left\|\sum_{i=1}^{r} v_i\right\|_k^{sp} \leq \sum_{i=1}^{r} \|v_i\|_k^{sp} = \sum_{i=1}^{r} \|v_i\|_2$$

for every $v_i \in \mathbb{R}^d$ such that $\|v_i\|_0 \leq k$, $i = 1, \ldots, r$, and $w = \sum_{i=1}^r v_i$. Now let $v_i \in \mathbb{R}^d$ such that $\|Dv_i\|_0 \leq s$, $i = 1, \ldots, r$, and $w = \sum_{i=1}^r v_i$. Then

$$\|Dw\|_1 = \left\|\sum_{i=1}^{r} Dv_i\right\|_1 \leq \sum_{i=1}^{r} \|Dv_i\|_1 \leq \sum_{i=1}^{r} \sqrt{s}\|Dv_i\|_2 \leq \sqrt{s}\|D\| \sum_{i=1}^{r} \|v_i\|_2\ .$$

The above two inequalities imply Equation (8).

For Equation (9), it suffices to show the dual inequality. Recall from Argyriou et al (2012) that the norm defined by $\|u\|_{(k)}^{(2)} := \left(\sum_{i=1}^{k} (|u|_i^\downarrow)^2\right)^{\frac{1}{2}}$ is the dual of $\|\cdot\|_k^{sp}$. This is the ℓ_2 norm of the largest k entries in $|u|$, and is known as the 2-k *symmetric gauge norm* (Bhatia, 1997). Thus, for every $a, w \in \mathbb{R}^d$, it holds that

$$\begin{aligned}
\langle x - D^\top a, w\rangle &\leq \|x - D^\top a\|_{(k)}^{(2)} \|w\|_k^{sp} \leq \|x - D^\top a\|_{(k)}^{(2)} \Omega_{sptv}(w) \\
\langle D^\top a, w\rangle &= \langle a, Dw\rangle \leq \|a\|_\infty \|Dw\|_1 \leq \sqrt{s}\|D\|\|a\|_\infty \Omega_{sptv}(w)
\end{aligned}$$

Adding up and taking the infima with respect to a, we obtain

$$\langle x, w\rangle \leq \inf_{a\in\mathbb{R}^d} \left\{\|x - D^\top a\|_{(k)}^{(2)} + \sqrt{s}\|D\|\|a\|_\infty\right\} \Omega_{sptv}(w).$$

and hence

$$\Omega_{sptv}^*(x) \leq \inf_{a\in\mathbb{R}^d} \left\{\|x - D^\top a\|_{(k)}^{(2)} + \sqrt{s}\|D\|\|a\|_\infty\right\}\ .$$

Next we pick I to be the set of indexes corresponding to the largest k elements of $|x|$. We also pick

$$a = (D^\top)^+ c, \qquad c_i = \begin{cases} \text{sgn}(x_i)\|x_{I^c}\|_\infty & \text{if } i \in I \\ x_i & \text{if } i \in I^c\ . \end{cases}$$

Since range$(D^\top) = \mathbb{R}^d$, it holds that $D^\top a = c$ and hence we obtain

$$\|x - D^\top a\|_{(k)}^{(2)} + \sqrt{s}\|D\|\|a\|_\infty = \sqrt{\sum_{i\in I}(|x_i| - \|x_{I^c}\|_\infty)^2} + \sqrt{s}\|D\|\|(D^\top)^+ c\|_\infty$$

$$\leq \sqrt{\sum_{i\in I}(x_i^2 - \|x_{I^c}\|_\infty^2)} + \sqrt{s}\|D\|\|(D^\top)^+\|_\infty\|x_{I^c}\|_\infty$$

$$= \sqrt{\sum_{i\in I} x_i^2 - k\|x_{I^c}\|_\infty^2} + \sqrt{s}\|D\|\|(D^\top)^+\|_\infty\|x_{I^c}\|_\infty \leq \sqrt{1 + \frac{s\|D\|^2\|(D^\top)^+\|_\infty^2}{k}}\,\|x_I\|_2.$$

By the hypothesis, we may choose $J \in M_s$ such that $D_{J^c I} = 0$. Then

$$\|x_I\|_2 = \max_{K\in M_s} \|\operatorname{Proj}_{\operatorname{null}(D_{K^c I})}(x_I)\|_2 \leq (\|x\|_{k,s}^{sptv})^*$$

□

We note that we can fulfil the technical condition on the range of D^T by augmenting the incidence matrix in a manner that does not change the result of the regularized risk minimization. The condition that the submatrix D_{*I} has at least $m - s$ zero rows has an intuitive interpretation when D is the transpose of an incidence matrix of a graph. It means that any group of k vertices in the graph involves at most s edges. This is true in many cases of interest, such as grid structured graphs if s is proportional to k. The term involving $\|(D^\top)^+\|_\infty^2$ is at most linear in the number of vertices. $\|D\|^2$ corresponding to the maximum eigenvalue of the graph Laplacian is bounded above by a constant for a given structure (e.g. 2-D with neighborhood of 4).

We have proposed a tractable approximation to the (k,s) support total variation norm, which was shown to be NP-hard. We showed that the error from this approximation has a bound that scales well for the case of grid graphs. We now discuss some optimization strategies for this approximate penalty and demonstrate several experiments showing its utility.

2.4 Optimization

Denoting $\hat{f}(w)$ as a loss function, $\Omega_{sptv}(w)$ as given by Equation (7), and $\lambda > 0$. It can be shown that, given appropriate parameter selection, the solution to a regularized risk minimization of $\hat{f}(w)$ constrained by $\Omega_{sptv}(w) \leq \lambda$ will be equivalent to optimizing any of the following objectives for some regularization parameters $\lambda_1, \lambda_2 > 0$. [2]

$$\min_w \ \hat{f}(w) + \lambda_1(\|w\|_k^{sp})^2 + \lambda_2 TV(w) \tag{10}$$

$$\min_w \ \hat{f}(w) + \lambda_1\|w\|_k^{sp} + \lambda_2 TV(w) \tag{11}$$

$$\min_w \ \hat{f}(w) + \lambda_2 TV(w) \ \ s.t.\ \|w\|_k \leq \lambda_1 \tag{12}$$

[2] The proof of this statement follows from the fact that optimization subject to the intersection of two constraints has a Lagrangian that is exactly a regularized risk minimization with the two corresponding penalties each with their own Lagrange multiplier.

We analyze several optimization strategies for optimizing the prescribed objectives: Iterated FISTA with a smoothed $TV(w)$, FISTA with an approximate computation of the $\|w\|_k^{sp} + TV(w)$, and the Excessive Gap Method. A common concern in TV related optimization is the convergence. The former two methods have previously shown good empirical and theoretical convergence (Dohmatob et al, 2014; Dubois et al, 2014) and we describe specifics of their implementation with our objective below. However,these approaches do not provide optimality guarantees on the solution. For solving Equation (12) we may apply the Excessive Gap Method, which has convergence guarantees on the duality gap. We describe the non-trivial analysis required for applying the excessive gap method on our objective, which also requires the newly derived k-support ball projection operator in Section 2.4.1. We note that this section constitutes a preliminary proposal demonstrating our objectives can be optimized with state-of-the-art convex optimization methods. A detailed analysis of the optimization is beyond the scope of this work, and we utilize a combination of the methods described throughout our experiments.

In Iterated FISTA, we may utilize the proximal operator for k-support along with Nesterov smoothing on the $TV(w)$ term to make it differentiable (Dohmatob et al, 2014; Nesterov, 2004). We can follow a strategy of repeatedly solving a FISTA problem with progressively decreasing smoothing parameter on the $TV(w)$ term as per (Dubois et al, 2014), who provide analysis of such an approach, which they call CONESTA. This technique can be used to solve any of Equations (11), (10), (12) given the relevant proximal mapping discussed in Section 2.4.1

We can estimate the proximal operator of $\lambda_1\|w\|_k^{sp} + \lambda_2 TV(w)$ using an accelerated proximal gradient method in the dual, as described in Beck and Teboulle (2009), and the projection operator onto the $\|w\|_k^{sp}$ dual ball given in Chatterjee et al (2014). This allows us another approach of directly applying FISTA, but with the inexact proximal operator in order to solve Equation (11).

To apply the Excessive Gap Method to k-support TV regularizations we note the primal and the dual of Equation (12) can be written as $\min_{\|w\|_{ksp} \leq \lambda_1} f(w) = \max_{\|u\|_\infty < 1} \phi(u)$ where the primal is given $f(w) = \hat{f}(w) + \max_{\|u\|_\infty < 1} \{\langle Dw, u\rangle\}$, and the dual is given by $\phi(u) = -\hat{\phi}(u) + \langle Dw_u^*, u\rangle + \hat{f}(w_u^*)$ with $w_u^* = \operatorname{arg\,min}_{\|w\|_k^{sp} \leq \lambda_1} \langle Dw, u\rangle + \hat{f}(x)$.

We can now smooth the primal function

$$f_\mu(w) = \hat{f}(w) + \max_{\|u\|_\infty < 1} \{\langle Dw, u\rangle - \mu\|u\|^2\} = \hat{f}(w) + \langle Dw, u_\mu(x)\rangle - \mu\|u_\mu(x)\|^2$$

The excessive gap method now allows us to take successive approximations of $f_\mu(x)$ with a decreasing sequence of μ while maintaining a bound on the duality gap proportional to μ. To apply the excessive gap method we need the smooth approximations $u_\mu(x)$ and the gradient mappings $T_\mu(x)$, defined in (Nesterov, 2005). We can obtain these using the simple projection of a vector, z, onto the ℓ_∞ ball,which we denote $P_{\|\cdot\|_\infty \leq 1}(z)$, obtained by truncating all values above magnitude 1. The relevant operations are then given by

$$u_\mu(w) = \operatorname*{arg\,min}_{\|u\|_\infty \leq 1} \{\langle Dw, u\rangle - \mu\|u\|^2\} = P_{\|\cdot\|_\infty \leq 1}\left(\frac{Dw}{2\mu}\right)$$

$$T_\mu(u) = \operatorname*{arg\,max}_{\|u\|_\infty \leq 1} \left\{\langle \nabla\phi, y - u\rangle - \frac{L_\phi}{2}\|y - u\|^2\right\} = P_{\|\cdot\|\infty \leq 1}\left(u + \frac{Dx(u)}{L_\phi}\right)$$

The sub-problem of finding $x(u)$ can be solved using an accelerated projected gradient method and the projection onto the k-support ball derived in Section 2.4.1.

2.4.1 Proximal Operators Associated With The k-support Norm

The proximal operator for $(\|w\|_k^{sp})^2$, associated with Equation (10) is given by McDonald et al (2014). The proximal operator for $\|w\|_k^{sp}$, associated with Equation (11), is given by Chatterjee et al (2014). In turn we can obtain the projection on the dual ball using Moreau decomposition (Parikh et al, 2014). The projection onto the $\|w\|_k^{sp}$ ball (proximal of the indicator function) is not yet addressed in the literature to the best of our knowledge and we show below how to obtain this projection. We define δ_{C_λ} as the indicator function on the k-support ball of size λ, C_λ. We note that k-support norm is given by

$$\|w\|_k^{sp} = \left(\sum_{i=1}^{k-r-1} (|w|_i^{\downarrow})^2 + \frac{1}{r+1} \left(\sum_{i=k-r}^{d} |w|_i^{\downarrow} \right)^2 \right)^{\frac{1}{2}} \tag{13}$$

where $|w|_i^{\downarrow}$ is the ith largest element of w. The projection onto $\|w\|_k^{sp}$ is given by:

Theorem 2 *Given $\lambda > 0$ and $x \in R^p$, if $\|x\|_k^{sp} < \lambda$, then the projection,$w^* = prox_{\delta_{C_\lambda}}(x)$,is simply x. If $\|x\|_k^{sp} > \lambda$, define $D_r = \sum_{i=1}^{k-r-1} (|x|^{\downarrow})^2, T_{r,l} = \sum_{i=k-r}^{l} |x|^{\downarrow}$, and $n = l - k + r + 1$, and construct the equation for $\beta_{r,l}$:*

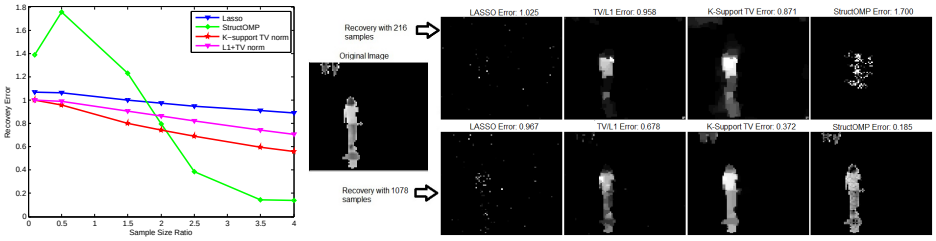
$$\beta^2 D_r + \left(\frac{(\beta+1)\beta(r+1)T_{r,l}}{n+\beta(r+1)} \right)^2 - \lambda^2(\beta+1)^2 = 0 \tag{14}$$

The projection onto the k-support ball is given by finding r, l which satisfy the conditions:

$$|x|_{k-r-1}^{\downarrow} > \frac{(\beta+1)T_{r,l}}{n+\beta(r+1)} \geq |x|_{k-r}^{\downarrow} \ , \ |x|_l^{\downarrow} > \frac{T_{r,l}}{n+\beta(r+1)} \geq |x|_{l+1}^{\downarrow}$$

Where β is a non-negative solution to Equation (14)*. Furthermore the binary search specified in Chatterjee et al (2014, Algorithm 2) with Equation* (14) *can be used to find the appropriate r and l in O(log(k) log(d-k)).*

Proof Sketch: Argyriou et al (2012, Algorithm 1) specifies conditions on the proximal map of $\frac{1}{2\beta}(\|w\|_k^{sp})^2$. For a given β there must be a corresponding λ such that $\|w\|_k^{sp} = \lambda$, and therefore $\|prox_{\frac{1}{2\beta}(\|w\|_k^{sp})^2}(x)\|_k^{sp} = \lambda$. Substituting Equation (13) and explicit form and constraints for $prox_{\frac{1}{2\beta}(\|w\|_k^{sp})^2}(x)$ in Argyriou et al (2012, Algorithm 1) we obtain Equation (14) when the constraints are satisfied. Chatterjee et al (2014, Theorem 3), holds since the constraints are the same □



(a) (b)

Fig. 2: (a) Average model error for background subtracted image reconstruction for various sample sizes. (b) Image example for different methods and sample sizes. k-support/TV regularization gives the best recovery error for 216 samples, and gives smoother recovery results than the other methods for both sample sizes.

3 Experimental Results

We evaluate the effectiveness of the introduced penalty on signal recovery and classification problems. We consider a sparse image recovery problem from compressed sensing, a small training sample classification task using MNIST, an M/EEG prediction task, and classification and recovery task for fMRI and synthetic data. We compare our regularizer against several common regularizers (ℓ_1 and ℓ_2) and popular structured regularizers for problems with similar structure. In recent work TV+ℓ_1, which adds the TV and ℓ_1 constraints, has been heavily utilized for data with similar spatial assumptions (Dohmatob et al, 2014; Gramfort et al, 2013) and is thus one of our main benchmarks. Source code for learning with the k-support/TV regularizer is available at `https://github.com/eugenium/StructuredSparsityRegularization`.

3.1 Background Subtracted Image Recovery

We apply k-support total variation regularization to a background subtracted image reconstruction problem frequently used in the structured sparsity literature (Baldassarre et al, 2012a; Huang et al, 2009). We use a similar setup to Baldassarre et al (2012a). Here we apply m random projections to a background-subtracted image along with Gaussian noise, and reconstruct the image using the projections and projection matrices. Our evaluation metric for the recovery is the mean squared pixel error. For this experiment we utilize the a squared loss function and the iterative FISTA with smoothed TV described in Section 2.4.

We selected 50 images from the background segmented dataset and converted them to grayscale. We use squared loss and k-support total variation to reconstruct the original images. We compute normalized recovery error for different number of samples m and compare our regularizer to LASSO, TV+ℓ_1, and StructOMP. The latter is a structured regularizer which performs best on this problem in Huang et al (2009). The average normalized recovery error is shown for different sample sizes in Figure 3.1(a). We used a separate set of images to set the parameters for each method.

In terms of recovery error we note that k-support total variation substantially outperforms LASSO and TV+ℓ_1, and outperforms StructOMP for low sample

sizes. Further examination of the images reveals other advantages of the k-support total variation regularizer. An example for one image recovery scenario is shown at 2 different sample sizes in Figure 3.1(b). Here we can see that at low sample sizes StructOMP and LASSO can completely fail in terms of creating a visually coherent reconstruction of the image. TV+ℓ_1 recovery at the low sample size improves upon the latter methods, producing smooth regions, but still not resembling the human shape pictured in the original image. k-support total variation has better visual quality at this low sample complexity, due to its ability to retain multiple groups of correlated variables in addition to the smoothness prior. For the case of a larger number of samples, illustrated by the bottom row of Figure 3.1(b), we note that although the recovery performance of StructOMP is better (lower error), the visual quality of the k-support total variation regualrizer produces smoother and more coherent image segments.

3.2 Low Sample Complexity MNIST Classification

We consider a simple classification problem using the MNIST data set (LeCun and Cortes, 2010). We select a very small subset of data to train with in order to demonstrate the effectiveness of our regularizer. We train a one versus all classifier for each digit. In the case of each digit we take 9 negative training samples, one from each other digit, and 9 positive training samples of the digit. We use a validation set consisting of 8000 examples to perform parameter selection. We use a regularized risk function consisting of the form (10) and logistic loss. Optimization for a single parameter setting took on the order of one second for a MatLab implementation on a 2.8 GHz core. We choose the best model parameters from $k \in \{1, 2^3, 2^5, 2^7, 2^9, d\}$, $\lambda_1 \in \{\frac{10^5}{N}, .., \frac{10^2}{N}\}$, and $\lambda_2 \in \{0, \frac{10^3}{N}, .., \frac{10^{-1}}{N}\}$, where N is the training set size. Here d corresponds to the image size (28×28) and the cases $k = 1$ and $k = d$ correspond the ℓ_1 and ℓ_2 norm, respectively, when $\lambda_2 = 0$. We test on the entire MNIST test set of 10000 images. We optimize a logistic loss function combined with our k-support total variation norm and compare to results from ℓ_1, ℓ_2, k-support norm, and TV/ℓ_1 penalties combined with logistic loss. We perform optimization using FISTA on the k-support norm (Argyriou et al, 2012; Nesterov, 2004) and a smoothing applied to the total variation. For the graph structure, specified by D, we use a grid graph with each pixel having a neighborhood consisting of the 4 adjacent pixels. We obtain surprisingly high classification accuracy using just 18 training examples. The results in Table 1 show classification accuracy for each one versus all classifier and the average of the classifiers. In all but two cases the k-support TV norm outperforms the other regularizers. We note that for the digit 9 classification the difference between the best classifier and k-support/TV is not statistically significant

3.3 M/EEG Prediction

We apply k-support total variation regularization to an M/EEG prediction problem from Backus et al (2011); Zaremba et al (2013), using the preprocessing from Zaremba et al (2013). This results in data samples with 60 channels, each consisting of a time-series presumed to be independent across channels. Following Zaremba et al (2013) we report results for subject 8 from this dataset. For the total variation graph structure, we impose constraints for adjacent samples within each channel, while values from different channels are not connected within the

Class.	ℓ_1	ℓ_2	KS	ℓ_1+TV	KS+TV
D0	$93.62 \pm .01$	$93.49 \pm .01$	$93.68 \pm .02$	$96.22 \pm .01$	$\mathbf{96.27 \pm .01}$
D1	$90.1 \pm .02$	$89.56 \pm .02$	$90.08 \pm .02$	$90.57 \pm .02$	$\mathbf{92.18 \pm .02}$
D2	$78.28 \pm .03$	$77.28 \pm .03$	$78.25 \pm .03$	$\mathbf{81.47 \pm .02}$	$81.39 \pm .03$
D3	$68.58 \pm .02$	$68.05 \pm .02$	$68.60 \pm .02$	$71.63 \pm .02$	$\mathbf{73.25 \pm .02}$
D4	$83.81 \pm .01$	$82.55 \pm .01$	$83.76 \pm .01$	$84.69 \pm .01$	$\mathbf{84.79 \pm .01}$
D5	$73.7 \pm .03$	$73.2 \pm .02$	$73.69 \pm .03$	$74.52 \pm .02$	$\mathbf{74.95 \pm .02}$
D6	$93.48 \pm .01$	$93.37 \pm .01$	$93.51 \pm .01$	$93.71 \pm .01$	$\mathbf{94.08 \pm .01}$
D7	$88.88 \pm .02$	$87.21 \pm .02$	$88.85 \pm .02$	$91.67 \pm .01$	$\mathbf{92.59 \pm .01}$
D8	$70.79 \pm .02$	$72.07 \pm .03$	$72.75 \pm .02$	$73.23 \pm .02$	$\mathbf{73.10 \pm .02}$
D9	$85.48 \pm .02$	$85.61 \pm .02$	$85.49 \pm .02$	$85.5 \pm .03$	$85.60 \pm .03$

Table 1: Accuracy for One versus All classifiers on MNIST using only 18 training examples and standard error computed on the test set. In all but two cases, k-support/TV regularization gives the best performance with significance. For digit '9' k-support/TV regularization is statistically tied for best performance.

Classifier	Mean Acc.	Acc std.
SVM (Zaremba et al, 2013)	65.44%	2.29%
ksp-TV SVM	66.84%	3.42%
TV-ℓ_1 SVM	60.70%	4.66%

Table 2: M/EEG accuracy for SVM, k-support total variation regularized SVM, and TV+ℓ_1 regularized SVM computed over 5 folds. k-support/TV regularization yields the best results on average.

graph. In the original work a latent variable SVM with delay parameter h is used to improve alignment of the samples. We consider only the case for $h = 0$, which reduces to the standard SVM. To directly compare our results we utilize hinge loss with a constant C of 2×10^4, the same regularization value used in Zaremba et al (2013). Thus we optimize the following objective

$$R(w) = \frac{C}{N} \sum_{i=1}^{N} \max\{0, 1 - y_i \langle w, x_i \rangle\} + (1 - \lambda)(\|w\|_k^{sp})^2 + \lambda \|Dw\|_1$$

Where λ allows us to easily trade off between k-support and total variation norms, while maintaining a fixed weight for our regularizer comparable to Zaremba et al (2013). We use $k = 2500$ (approximately 80% of the dimensions) and $\lambda = 0.1$. Table 2 shows the mean and standard deviation for the classification accuracy. We use the same partitioning of the data as described in (Zaremba et al, 2013), and on average obtain an improvement over the original results. We note that TV+ℓ_1 regularization has relatively poor performance. We hypothesize this is because the data used is very noisy and not very sparse.

3.4 Prediction and Identification in fMRI analysis

In this section we demonstrate the advantages of our sparse regularization method in the analysis of fMRI neuro-imaging data. Brain activation in response to stimuli is normally assumed to be sparse and locally contiguous, thus our proposed regularizer is ideal for describing our prior assumptions on this signal. An important aspect of analysing fMRI data is the ability to demonstrate how the predictive variables identified as important by an estimator correspond to relevant brain regions. Regularized risk minimization is one of few approaches which can handle the multivariate nature of this problem. However, in the presence of many highly

correlated variables, such as those in brain regions with many adjacent voxels being activated by a stimulus, using sparse regularization alone there may be many possible solutions with near equivalent predictive performance for small training sample size. Furthermore, from a practical standpoint, overly sparse solutions can be difficult to interpret when attempting to determine an implicated brain region. Thus regularization here allows us to not only converge to a good solution with lower sample complexity, but obtain more interpretable models from amongst the space of solutions with good prediction. Related to interpretability is solution stability, solutions which are more stable under different samples of training data, with regards to implicated voxels/regions allow the practitioner to make a more trustworthy interpretations of the model (Misyrlis et al, 2014; Yan et al, 2014). We evaluate our approach taking all these factors into account.

We first analyze our method using a synthetic simulation of a signal similar to brain activation patterns. This gives us the opportunity to assess the true support recovery performance, which we cannot obtain with real data. We then analyze a popular block-design fMRI dataset from a study on face and object representation in the human ventral temporal cortex (Dohmatob et al, 2014) and perform experiments on predicting and in turn utilizing the predictive models for identifying the relevant regions of interest. We attempt to classify scans taken when a user is shown a pair of scissor vs. when they observe scrambled pixels. We demonstrate that we can obtain improved accuracy, solution interpretability, and stability characteristics compared to previously applied sparse regularization methods incorporating spatial priors. For these experiments we use logistic loss and the $TV_I(w)$ penalty, which has been shown to work better in fMRI analysis. Optimization is done using FISTA and estimated proximal operator. As our baseline we focus on TV+ℓ_1 which has been recently popularized for fMRI applications as well as TV+ℓ_1+ℓ_2, which has been considered in structural MRI (Dubois et al, 2014).

We consider the estimation of an ideal weight vector with both spatial correlation and sparsity similar to brain activation patterns with spatial correlations between neurons which are active and not-active and the activated neurons often occuring in adjacent regions of the brain. We construct a 25x25 image with 84% of coefficients set to zero. The non-sparse portion of the image corresponds to Gaussian blobs. This image will serve as a set of parameters w we wish to recover. Figure 3 shows this ideal parameter vector. We construct data samples $X = Yw+\varepsilon$. Where Y is a sample from $\{-1, 1\}$ and ε is Gaussian noise. We take 150 training samples, 100 validation samples, and 1000 test samples. We consider a binary classification setting using only ℓ_1, ℓ_2, or k-support regularizers, Smooth-Lasso (Hebiri et al, 2011), TV+ℓ_1 regularizer, TV+ℓ_1+ℓ_2, and our k-support TV regularizer. For each of these scenarios we perform model selection using grid search and select the model with the highest accuracy on the validation set. We repeat this experiment with a new set of training, validation, and test samples 15 times so that we may obtain statistical significance results. The test set accuracy results for each method are shown in Table 4. For each competing method we perform a Wilcoxon signed-rank test against the k-support total variation results. In all listed cases the test rejects the null hypothesis (at a significance level of $p < 0.05$) that the samples come from the same distribution. We assess the support recovery of competing method by measuring the area under the precision-recall curve for different support thresholds. Finally we measure stability using Pearson correlation between weight vectors from different trials.

Description	Test Acc. (p-value)	Supp. Recovery	Stability
ℓ_2	67.8%(7E-4)	0.388	0.173
ℓ_1	68.4%(7E-4)	0.377	0.220
k-support	68.1%(7E-4)	0.398	0.217
Smooth-LASSO	77.0%(7E-4)	0.407	0.464
TV+ℓ_1	80.2%(9E-3)	0.739	0.620
TV+ℓ_1+ℓ_2	81.5%(2E-2)	0.796	0.688
k-support/TV	**82.2%**	**0.816**	**0.719**

Table 3: Average test accuracy, support recovery, and test accuracy results for 15 trials of synthetic data along with p-value for a Wilcoxon signed-rank test performed for each method against the k-support/TV result, below 0.05 for all cases. k-support/TV has both the highest accuracy, highest support recovery as well as the highest stability. Here stability is measured by average pairwise Pearson correlation between folds.

In Figure 3 we visualize the weight vector and precision-recall curve produced by the various regularization methods for one trial. We can see that in Figure 3 the k-support norm alone does a poor job at reconstructing a model with any of these local correlations in place. The Smooth-Lasso, TV+ℓ_1 and TV+ℓ_1+ℓ_2 regularizers do a substantially better job at indicating the areas of interest for this task but the k-support/TV regularizer produces more precise regions with fewer spurious patterns and substantially better classification accuracy and support recovery. We can see an additional advantage of the k-support/TV regularizer over the other methods in terms of stability of the results across trials. Figure 3(c) also shows the effectiveness of the k-support/TV regularizer for varying target weight vectors.

In the analysis of fMRI data we are often concerned with using the estimator to identify the predictive regions. Specifically the linear model is often mapped back to a brain volume and used for analysis. In this context regularization can not only improve predictive performance, but it can provide more interpretable brain maps. We prefer solutions which clearly indicate the areas of interest. Well converged TV+ℓ_1 solutions can overemphasize the sparsity. With the k variable we can encourage a less sparse solution, that may be more interpretable and include more highly correlated variables. Figure 4a shows this effect for maps of varying k values (note that $k = 1$ corresponds to $TV + \ell_1$).

We note that unlike the elastic-net penalty the k in k-support has an interpretable parameter setting for mixing sparsity and ℓ_2. We can interpret the k in our regularizer as an estimate of the number of voxel locations active in the brain. Thus we can set k based on prior knowledge. We fix the value of k to 500 representing approximately 2% sparsity, this allows us to directly compare to the state of the art method for sparse regularization in fMRI, $TV + \ell_1$, with an equal sized search space in model-selection. Below we show the accuracy and stability results for $TV + \ell_1$, $TV + \ell_1 + \ell_2$, and our TV+k-support.

Since the size of the data is small we often have equivalent average accuracies in model selection, we break ties based on intra-fold stability as measured by average pairwise Spearman correlations of the resulting weight vectors. Our result beats TV+ℓ_1 in terms of accuracy . Compared to TV+ℓ_1+ℓ_2 we have better classification accuracy, but not with a high statistical significance, however we obtain much more stable solutions and have more interpretable parameter settings. We describe another advantage of our approach compared to the competing methods below.

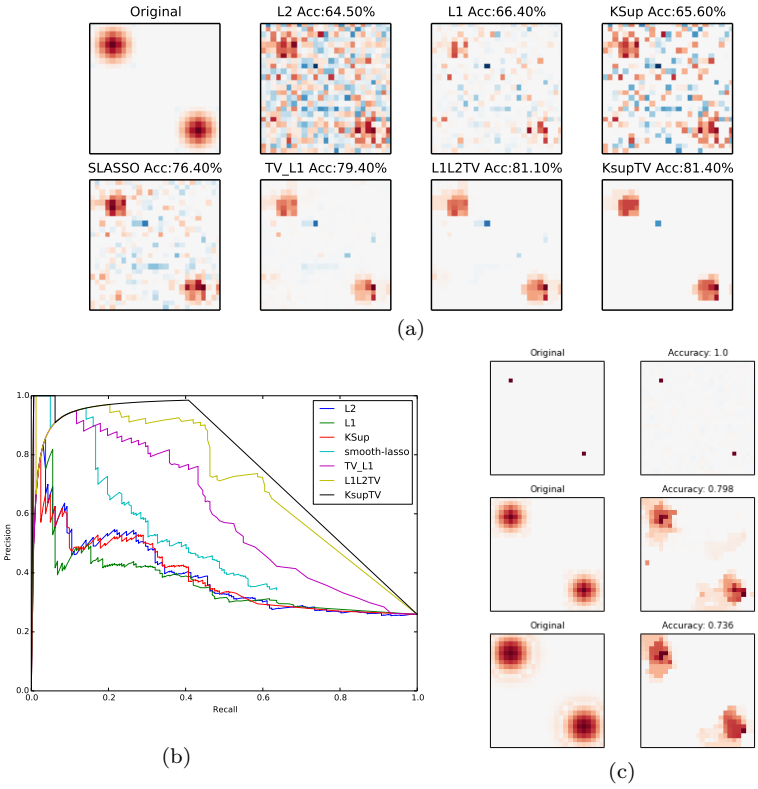
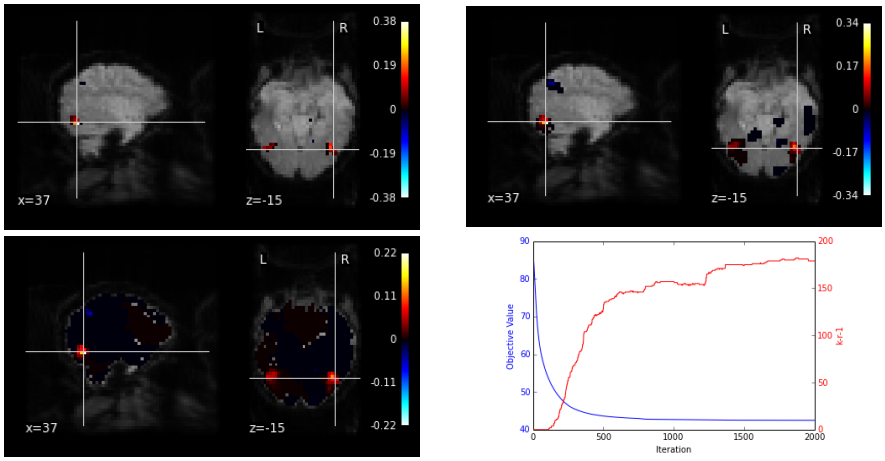


Fig. 3: (a) (*left top to bottom right*) ideal weight vector, weight vector obtained with ℓ_1, ℓ_2, k-support norm, TV+ℓ_1, and k-support/TV regularizer, and weight vector with combined total variation and k-support norm regularizer. The k-support/TV regularization gives the highest accuracy, support recovery, stability, and most closely approximates the target pattern. (b) Illustrates the improved precision-recall for k-support/TV versus the other methods on the support recovery for different thresholds. (c) Recovered support for varying ideal weight vector. This demonstrates that the k-support/TV regularization works well for a wide range of sparsity, correlation, and smoothness.

An additional issue in interpreting brain maps is where to threshold. Many sparse regularizers, even those such as ℓ_1 only have asymptotic guarantees for sparse solutions; in practice we threshold values at a specific value. This is particularly problematic when we add TV into the objective. Here we suggest a heuristic motivated by the properties of the k-support norm. As we can see in Equation (13) the k-support norm can be shown to a combination of ℓ_2 penalties on the highest magnitude $k-r-1$ terms and ℓ_1 penalty applied to the rest. Here r is the unique integer in $\{0, \ldots, k-1\}$ satisfying

$$|w|^{\downarrow}_{k-r-1} > \frac{1}{r+1} \sum_{i=k-r}^{d} |w|^{\downarrow}_i \geq |w|^{\downarrow}_{k-r}.$$



(a) (b)

Fig. 4: (a) Output map for k=1 (TV-ℓ_1), k=50, and k=500, in each case the Lateral Occipital Cortex is indicated. (b) Objective value of TV+k-support (k=500) and $k-r-1$ over iterations.

Description	**Test Acc. (p-value)**	**Stability**
TV+ℓ_1	84.72 (8E-4)	0.132
TV+ℓ_1+ℓ_2	86.06 (0.15)	0.186
k-support/TV	87.91	0.415

Table 4: Average test accuracy results for 20 trials along with p-value for a Wilcoxon signed-rank test performed for each method against the k-support/TV result. Solution stability is measured by averaging pairwise Spearman correlations between solutions from different folds of training data. We note that our accuracy is statistically significantly better than TV+ℓ_1 and we do much better in terms of solution stability.

Empirically we can show that the value of $k-r-1$ for the solution grows from 0 as the optimization progress as seen in Figure 4b. This can be loosely interpreted as the algorithm starting with ℓ_1 optimization,which attempts to push variables to zero, but as we progress we have flexibility to move onto parts of the k-support ball where specific key variables fall into the ℓ_2 term, while we still attempt to squash the remaining terms with ℓ_1. This property of the optimization of our penalty implies a visualization heuristic for the final solution of taking the top $k-r-1$ variables. Another view on this heuristic comes from the implicit delineation implied by Equation (3.4). For k much smaller than d and $k-r-1$ greater than 0 the definition of r implies the $k-r-1^{th}$ largest magnitude parameter will be a large factor $(\frac{d-k+r}{1+r})$ bigger than the mean of the rest of the parameters below it. Figure 3.4 illustrates thresholding based on a fixed threshold value and our heuristic of thresholding based on the final $k-r-1$ value in k-support TV optimization.

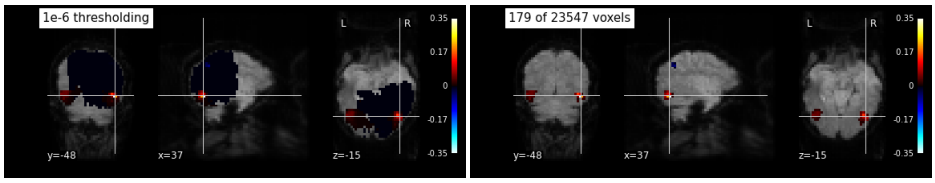


Fig. 5: Output map for fixed thresholding and thresholding based on converged $k - r - 1$ value

4 Conclusions

We have introduced a novel norm that incorporates spatial smoothness and correlated sparsity. This norm, called the (k, s) *support total variation norm*, extends both the total variation penalty which is a standard in image processing and the recently proposed k-support norm from machine learning. The (k, s) support TV norm is the *tightest convex penalty* that combines *sparsity*, ℓ_2 and *total variation* constraints jointly. We have derived a variational form for this norm for arbitrary graph structures. We have also expressed the dual norm as a combinatorial optimization problem on the graph. This graph problem is shown to be NP-hard motivating the use of a relaxation, which is shown to be equivalent to the weighted combination of a k-support norm and a total variation penalty. We have shown that this norm approximates the (k, s) support TV norm within a factor that depends on properties of the graph as well as on the parameters k and s, and that this bound scales well for grid structured graphs. Moreover, we have demonstrated that joint k support and TV regularization can be applied on a diverse variety of learning problems, such as classification with small samples, neural imaging and image recovery. These experiments have illustrated the utility of penalties combining k-support and total variation structure on problems where spatial structure, feature selection and correlations among features are all relevant. We have shown that this penalty has several unique properties that make it an excellent tool analysis of fMRI data. Some of our additional contributions include a generalized formulation of the dual norm of a norm which is the infimal convolution of norms, the first algorithm for projecting onto the k-support norm ball, and first analysis that notes interesting practical properties of the r variable of the k-support norm.

A Proof of Theorem 1

Proof. Computation of the (k, s)-support TV dual norm is an NP-hard problem, we do so by reduction from minimum weight multiway cut problem (Vazirani, 2001). Let the dual norm computation problem be denoted $P(z, D, s, k)$, where z is the input for the dual norm. We limit to the set of inputs where $k = d$, where d is the cardinality of z and D is an incidence matrix of a graph $G = (V, E)$ with vertex weights $z_i = w(v_i)$, $v_i \in V$. Additionally for simplifcation in later steps let $e = d - s$. This problem is refered to simply as $P1(G, e)$ and can be stated as follows: given a graph $G = (V, E)$ partition G, $G_p = (V, E_p)$ obtained by removing any e edges from E which maximizes

$$\max \sum_{V_i \in \{G_1, G_2, \ldots, G_k\}} \frac{1}{|G_i|} \left(\sum_{v_j \in V_i} w(v_j) \right)^2,$$

where $G_p = G_1 \cup G_2 \cup \cdots \cup G_k$ and elements of G are disjoint. V_i is the vertex set for G_i and $w(v_j)$ is the weight for vertex v_j.

The minimum 3-way cut problem (which we denote P_{M3}) is NP-hard. The problem can be stated as follows: given a graph $G = (V, E)$ and terminals $t_1, t_2, t_3 \in V$, find a minimum set of edges $E' \subseteq E$ such that the removal of E' from E disconnects each terminal t_i from the others. Furthermore the decision problem,denoted P_{M3D}, is to find out if its possible to disconnect t_1, t_2, t_3 by removing no more than e edges, where e is a part of the input. This problem is NP-complete. We show that any instances of P_{M3D} can be reduced in polynomial time to an instance of $P1$. Given an instance of P_{M3D} we have a graph G, terminals t_1, t_2, t_3 and integer e. We construct a new graph G_{aug} as follows

- We add weights to the vertices of the graph G, weighting non-terminal nodes 0 and the terminal nodes $1, 10, 100$ in any order.
- For each terminal we add N (the choice of N is described later on) more vertices to be its neighbor and weight these vertices 0. These augmented vertices are connected to the original graph only at the terminal vertices.

Figure 6 shows an example of an instance of P_{M3D} and the constructed augmented graph. We can now compute $P1(G_{aug}, e)$.

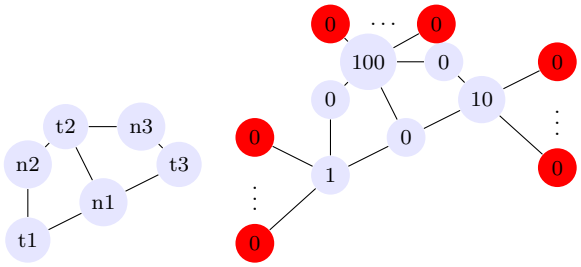


Fig. 6: Original unweighted graph of the 3-way mincut problem and the augmented weighted graph we construct as the input graph for problem P1.

If $N > |V|$ (the number of vertices in G) then none of the edges connected to the new vertices of G_{aug} will be removed by $P1$ since disconnecting two terminals from each other will always improve the result more than disconnecting one of the 0 nodes in the augmented vertices. Denoting the number of nodes in G (original graph) n, if $P1$ disconnects terminals $t1$,$t2$, and $t3$ the solution takes the form: $\frac{1}{N+n_1} + \frac{100}{N+n_2} + \frac{10000}{N+n_3}$ where $n_1 + n_2 + n_3 = n$. We can lower bound this value as $\frac{1}{N+n} + \frac{100}{N+n} + \frac{10000}{N+n} = \frac{10101}{N+n}$.

If $P1$ does not disconnect the terminals the solution can take on one of 4 forms, each of which can be upper bounded. For example,

$$\frac{11^2}{2N+n_1} + \frac{10000}{N+n_2} < \frac{11^2}{2N} + \frac{10000}{N} = \frac{10060.5}{N}.$$

Similarly for the other 3 cases, $\frac{1}{N+n_1} + \frac{101^2}{2N+n_2} < \frac{5101.5}{N}$, $\frac{1^2}{N+n_1} + \frac{110^2}{2N+n_2} < \frac{6051}{N}$ $\frac{111^2}{3N+n} < \frac{4107}{N}$ where $n_1 + n_2 = n$.
Examining the above inequalities we state that if terminals $t1, t2$, and $t3$ are not connected the solution will be at most $\frac{10060.5}{N}$. For $N > \frac{10060.5}{40.5}n$, the inequality $\frac{10101}{N+n} > \frac{10060.5}{N}$ always holds. Thus if it is possible to disconnect the terminals with e edges $P1$ will produce a value greater than $\frac{10060.5}{N}$ answering P_{M3D}. Since solutions of P_{M3D} are obtained in polynomial calls to $P1$, $P1$ is NP-hard. □

Acknowledgements
We would like to thank Tianren Liu for his help with showing that computation of the (k, s) support total variation norm is an NP-hard problem.

References

Argyriou A, Foygel R, Srebro N (2012) Sparse prediction with the k-support norm. In: NIPS, pp 1466–1474

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

Backus A, Jensen O, Meeuwissen E, van Gerven M, Dumoulin S (2011) Investigating the temporal dynamics of long term memory representation retrieval using multivariate pattern analyses on magnetoencephalography data. Tech. rep.

Baldassarre L, Morales J, Argyriou A, Pontil M (2012a) A general framework for structured sparsity via proximal optimization. In: AISTATS, pp 82–90

Baldassarre L, Mourao-Miranda J, Pontil M (2012b) Structured sparsity models for brain decoding from fMRI data. In: PRNI

Bauschke HH, Combettes PL (2011) Convex analysis and monotone operator theory in Hilbert spaces. CMS Books in mathematics, Springer

Beck A, Teboulle M (2009) Fast gradient-based algorithms for constrained total variation image denoising and deblurring problems. Image Processing, IEEE Transactions on 18(11):2419–2434

Bhatia R (1997) Matrix Analysis. Graduate Texts in Mathematics, Springer

Chatterjee S, Chen S, Banerjee A (2014) Generalized Dantzig selector: Application to the k-support norm. In: NIPS, pp 1934–1942

Dohmatob E, Gramfort A, Thirion B, Varoquaux G (2014) Benchmarking solvers for TV-l1 least-squares and logistic regression in brain imaging. In: PRNI

Dubois M, Hadj-Selem F, Lofstedt T, Perrot M, Fischer C, Frouin V, Duchesnay E (2014) Predictive support recovery with TV-elastic net penalty and logistic regression: An application to structural MRI. In: PRNI

Gkirtzou K, Honorio J, Samaras D, Goldstein RZ, Blaschko MB (2013) fMRI analysis of cocaine addiction using k-support sparsity. In: ISBI, pp 1078–1081

Gramfort A, Thirion B, Varoquaux G (2013) Identifying predictive regions from fMRI with TV-L1 prior. In: PRNI, pp 17–20

Hebiri M, Van De Geer S, et al (2011) The smooth-lasso and other 1+ 2-penalized methods. Electronic Journal of Statistics 5:1184–1226

Huang J, Zhang T, Metaxas D (2009) Learning with structured sparsity. In: Proceedings of the International Conference on Machine Learning, pp 417–424

LeCun Y, Cortes C (2010) MNIST handwritten digit database. http://yann.lecun.com/exdb/mnist/

Mairal J, Yu B (2013) Supervised feature selection in graphs with path coding penalties and network flows. JMLR 14(1):2449–2485

McDonald AM, Pontil M, Stamos D (2014) New perspectives on k-support and cluster norms. arXiv:14031481

Michel V, Gramfort A, Varoquaux G, Eger E, Thirion B (2011) Total variation regularization for fMRI-based prediction of behavior. IEEE Trans Med Imaging 30(7):1328–1340

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

Nesterov Y (2004) Introductory lectures on convex optimization. Springer

Nesterov Y (2005) Excessive gap technique in nonsmooth convex minimization. SIAM Journal on Optimization 16(1):235–249

Parikh N, Boyd S, et al (2014) Foundations and trends in optimization. Foundations and Trends in Theoretical Computer Science 8(1-2)

Rudin LI, Osher S, Fatemi E (1992) Nonlinear total variation based noise removal algorithms. Phys D 60(1-4):259–268

Tibshirani R (1996) Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B 58:267–288

Tibshirani R, Saunders M, Rosset S, Zhu J, Knight K (2005) Sparsity and smoothness via the fused lasso. Journal of the Royal Statistical Society Series B

Vazirani V (2001) Approximation Algorithms. Springer

Yan S, Yang X, Wu C, Zheng Z, Guo Y (2014) Balancing the stability and predictive performance for multivariate voxel selection in fMRI study. In: Brain Informatics and Health, pp 90–99

Zaremba W, Kumar MP, Gramfort A, Blaschko MB (2013) Learning from M/EEG data with variable brain activation delays. In: IPMI, pp 414–425
Zou H, Hastie T (2005) Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B 67(2):301–320