Learning with Structured Sparsity

Junzhou Huang JZHUANG@CS.RUTGERS.EDU
Department of Computer Science, Rutgers University, NJ, USA

Tong Zhang TZHANG@STAT.RUTGERS.EDU
Department of Statistics, Rutgers University, NJ, USA

Dimitris Metaxas DNM@CS.RUTGERS.EDU
Department of Computer Science, Rutgers University, NJ, USA

Abstract

This paper investigates a new learning formulation called *structured sparsity*, which is a natural extension of the standard sparsity concept in statistical learning and compressive sensing. By allowing arbitrary structures on the feature set, this concept generalizes the group sparsity idea. A general theory is developed for learning with structured sparsity, based on the notion of coding complexity associated with the structure. Moreover, a structured greedy algorithm is proposed to efficiently solve the structured sparsity problem. Experiments demonstrate the advantage of structured sparsity over standard sparsity.

1. Introduction

We are interested in the sparse learning problem under the fixed design condition. Consider a fixed set of p basis vectors $\{\mathbf{x}_1, \ldots, \mathbf{x}_p\}$ where $\mathbf{x}_j \in \mathbb{R}^n$ for each j. Here, n is the sample size. Denote by X the $n \times p$ data matrix, with column j of X being $\mathbf{x}_j$. Given a random observation $\mathbf{y} = [\mathbf{y}_1, \ldots, \mathbf{y}_n] \in \mathbb{R}^n$ that depends on an underlying coefficient vector $\bar{\beta} \in \mathbb{R}^p$, we are interested in the problem of estimating $\bar{\beta}$ under the assumption that the target coefficient $\bar{\beta}$ is sparse. Throughout the paper, we consider fixed design only. That is, we assume X is fixed, and randomization is with respect to the noise in the observation $\mathbf{y}$.

We consider the situation that $\mathbb{E}\mathbf{y}$ can be approximated by a sparse linear combination of the basis vectors: $\mathbb{E}\mathbf{y} \approx X\bar{\beta}$, where we assume that $\bar{\beta}$ is sparse. Define the support of a vector $\beta \in \mathbb{R}^p$ as $\mathrm{supp}(\beta) = \{j : \bar{\beta}_j \neq 0\}$ and $\|\beta\|_0 = |\mathrm{supp}(\beta)|$. A natural method for sparse learning is L_0 regularization for desired sparsity s:

$$\hat{\beta}_{L0} = \arg\min_{\beta \in \mathbb{R}^p} \hat{Q}(\beta) \quad \text{subject to } \|\beta\|_0 \leq s,$$

For simplicity, we only consider the least squares loss $\hat{Q}(\beta) = \|X\beta - \mathbf{y}\|_2^2$ in this paper. Since this optimization problem is generally NP-hard, in practice, one often considers approximate solutions. A standard approach is convex relaxation of L_0 regularization to L_1 regularization, often referred to as Lasso (Tibshirani, 1996). Another commonly used approach is greedy algorithms, such as the orthogonal matching pursuit (OMP) (Tropp & Gilbert, 2007).

In practical applications, one often knows a structure on the coefficient vector $\bar{\beta}$ in addition to sparsity. For example, in group sparsity (Yuan & Lin, 2006; Bach, 2008; Stojnic et al., 2008; Huang & Zhang, 2009), one assumes that variables in the same group tend to be zero or nonzero simultaneously. However, the groups are assumed to be static and fixed a priori. Moreover, algorithms such as group Lasso do not correctly handle overlapping groups (in that overlapping components are over-counted); that is, a given coefficient should not belong to different groups. This requirement is too rigid for many practical applications. To address this issue, a method called composite absolute penalty (CAP) is proposed in (Zhao et al.,) which can handle overlapping groups. Unfortunately, no theory is established to demonstrate the effectiveness of the approach. Other structures have also been explored in the literature. For example, so-called tonal and transient structures were considered for sparse decomposition of audio signals in (Daudet, 2004), but again without any theory. Grimm et al. (Grimm et al., 2007) investigated positive polynomials with structured sparsity from an optimization perspective. The theoretical result there did not address the effectiveness of such methods in comparison to standard sparsity. The closest work to ours is a recent paper (Baraniuk et al., 2008) which was pointed out to us

Appearing in *Proceedings of the 26^{th} International Conference on Machine Learning*, Montreal, Canada, 2009.

by an anonymous reviewer. In that paper, a specific case of structured sparsity, referred to as model based sparsity, was considered. It is important to note that some theoretical results were obtained there to show the effectiveness of their method in compressive sensing. However, their setting is more restrictive than the structured sparsity framework which we shall establish here.

The purpose of this paper is to present a framework for structured sparsity, and to study the more general estimation problem under this framework. If meaningful structures exist, we show that one can take advantage of such structures to improve the standard sparse learning.

2. Structured Sparsity

In structured sparsity, not all sparse patterns are equally likely. For example, in group sparsity, coefficients within the same group are more likely to be zeros or nonzeros simultaneously. This means that if a sparse coefficient's support set is consistent with the underlying group structure, then it is more likely to occur, and hence incurs a smaller penalty in learning. One contribution of this work is to formulate how to define structure on top of sparsity, and how to penalize each sparsity pattern.

In order to formalize the idea, we denote by $\mathcal{I} = \{1, \ldots, p\}$ the index set of the coefficients. We assign a cost $\mathrm{cl}(F)$ to any sparse subset $F \subset \{1, \ldots, p\}$. In structured sparsity, $\mathrm{cl}(F)$ is an upper bound of the coding length of F (number of bits needed to represent F by a computer program) in a pre-chosen prefix coding scheme. It is a well-known fact in information theory that mathematically, the existence of such a coding scheme is equivalent to $\sum_{F \subset \mathcal{I}} 2^{-\mathrm{cl}(F)} \leq 1$. From the Bayesian statistics point of view, $2^{-\mathrm{cl}(F)}$ can be regarded as a lower bound of the probability of F. The probability model of structured sparse learning is thus: first generate the sparsity pattern F according to probability $2^{-\mathrm{cl}(F)}$; then generate the coefficients in F.

Definition 2.1 *A cost function $\mathrm{cl}(F)$ defined on subsets of $\mathcal{I}$ is called a coding length (in base-2) if*

$$\sum_{F \subset \mathcal{I}, F \neq \emptyset} 2^{-\mathrm{cl}(F)} \leq 1.$$

We give $\emptyset$ a coding length 0. The corresponding structured sparse coding complexity of F is defined as

$$c(F) = |F| + \mathrm{cl}(F).$$

A coding length $\mathrm{cl}(F)$ is sub-additive if $\mathrm{cl}(F \cup F') \leq \mathrm{cl}(F) + \mathrm{cl}(F')$, and a coding complexity $c(F)$ is sub-additive if $c(F \cup F') \leq c(F) + c(F')$.

Clearly if $\mathrm{cl}(F)$ is sub-additive, then the corresponding coding complexity $c(F)$ is also sub-additive. Based on the structured coding complexity of subsets of $\mathcal{I}$, we can now define the structured coding complexity of a sparse coefficient vector $\bar{\beta} \in \mathbb{R}^p$.

Definition 2.2 *Giving a coding complexity $c(F)$, the structured sparse coding complexity of a coefficient vector $\bar{\beta} \in \mathbb{R}^p$ is*

$$c(\bar{\beta}) = \min\{c(F) : \mathrm{supp}(\bar{\beta}) \subset F\}.$$

Later in the paper, we will show that if a coefficient vector $\bar{\beta}$ has a small coding complexity $c(\bar{\beta})$, then $\bar{\beta}$ can be effectively learned, with good in-sample prediction performance (in statistical learning) and reconstruction performance (in compressive sensing). In order to see why the definition requires adding $|F|$ to $\mathrm{cl}(F)$, we consider the generative model for structured sparsity mentioned earlier. In this model, the number of bits to encode a sparse coefficient vector is the sum of the number of bits to encode F (which is $\mathrm{cl}(F)$) and the number of bits to encode nonzero coefficients in F (this requires $O(|F|)$ bits up to a fixed precision). Therefore the total number of bits required is $\mathrm{cl}(F) + O(|F|)$. This information theoretical result translates into a statistical estimation result: without additional regularization, the learning complexity for least squares regression within any fixed support set F is $O(|F|)$. By adding the model selection complexity $\mathrm{cl}(F)$ for each support set F, we obtain an overall statistical estimation complexity of $O(\mathrm{cl}(F) + |F|)$. While the idea of using coding based penalization resembles minimum description length (MDL), the actual penalty we obtain for structured sparsity problems is different from the standard MDL penalty for model selection. This difference is important, and thus in order to prevent confusion, we avoid using MDL in our terminology.

3. General Coding Scheme

We introduce a general coding scheme called *block coding*. The basic idea of block coding is to define a coding scheme on a small number of base blocks (a block is a subset of $\mathcal{I}$), and then define a coding scheme on all subsets of $\mathcal{I}$ using these base blocks.

Consider a subset $\mathcal{B} \subset 2^{\mathcal{I}}$. That is, each element (a block) of $\mathcal{B}$ is a subset of $\mathcal{I}$. We call $\mathcal{B}$ a block set if $\mathcal{I} = \cup_{B \in \mathcal{B}} B$ and all single element sets $\{j\}$ belong to $\mathcal{B}$ ($j \in \mathcal{I}$). Note that $\mathcal{B}$ may contain additional non single-element blocks. The requirement of $\mathcal{B}$ containing all single element sets is for convenience, as it implies that every subset $F \subset \mathcal{I}$ can be expressed as the union of blocks in $\mathcal{B}$.

Let cl_0 be a code length on $\mathcal{B}$:

$$\sum_{B \in \mathcal{B}} 2^{-\mathrm{cl}_0(B)} \leq 1,$$

we define $\mathrm{cl}(B) = \mathrm{cl}_0(B) + 1$ for $B \in \mathcal{B}$. It not difficult to show that the following cost function on $F \subset \mathcal{I}$ is a code-length

$$\mathrm{cl}(F) = \min \left\{ \sum_{j=1}^{b} \mathrm{cl}(B_j) : F = \bigcup_{j=1}^{b} B_j \quad (B_j \in \mathcal{B}) \right\}.$$

This is a coding length because

$$\sum_{F \subset \mathcal{I}, F \neq \emptyset} 2^{-\mathrm{cl}(F)} \leq \sum_{b \geq 1} \sum_{\{B_\ell\} \in \mathcal{B}^b} 2^{-\sum_{\ell=1}^{k} \mathrm{cl}(B_\ell)}$$
$$\leq \sum_{b \geq 1} \prod_{\ell=1}^{b} \sum_{B_\ell \in \mathcal{B}} 2^{-\mathrm{cl}(B_\ell)} \leq \sum_{b \geq 1} 2^{-b} = 1.$$

It is obvious that block coding is sub-additive.

The main purpose of introducing block coding is to design computational efficient algorithms based on the block structure. In particular, we consider a structured greedy algorithm that can take advantage of block structures. In the structured greedy algorithm, instead of searching over all subsets of $\mathcal{I}$ up to a fixed coding complexity s (exponential in s number of such subsets), we greedily add blocks from $\mathcal{B}$ one at a time. Each search problem over $\mathcal{B}$ can be efficiently performed because $\mathcal{B}$ is supposed to contain only a computationally manageable number of base blocks. Therefore the algorithm is computationally efficient. Concrete structured sparse coding examples described below can be efficiently approximated by block coding.

Standard sparsity

A simple coding scheme is to code each subset $F \subset \mathcal{I}$ of cardinality k using $k \log_2(2p)$ bits, which corresponds to block coding with $\mathcal{B}$ consisted only of single element sets, and each base block has a coding length $\log_2 p$. This corresponds to the complexity for the standard sparse learning.

Group sparsity

The concept of group sparsity has been appeared in various recent work, such as the group Lasso in (Yuan & Lin, 2006). Consider a partition of $\mathcal{I} = \cup_{j=1}^{m} G_j$ to m disjoint groups. Let $\mathcal{B}_G$ contain the m groups G_j, and $\mathcal{B}_1$ contain p single element blocks. The strong group sparsity coding scheme is to give each element in $\mathcal{B}_1$ a code-length cl_0 of ∞, and each element in $\mathcal{B}_G$ a code-length cl_0 of $\log_2 m$. Then the block coding scheme with blocks $\mathcal{B} = \mathcal{B}_G \cup \mathcal{B}_1$ leads to group sparsity, which only looks for signals consisted of the groups. The resulting coding length is: $\mathrm{cl}(B) = g \log_2(2m)$ if B can be represented as the union of g disjoint groups G_j; and $\mathrm{cl}(B) = \infty$ otherwise. Note that if the signal can be expressed as the union of g groups, and each group size is k_0, then the group coding length $g \log_2(2m)$ can be significantly smaller than the standard sparsity coding length of $g k_0 \log_2(p)$. As we shall see later, the smaller coding complexity implies better learning behavior, which is essentially the advantage of using group sparse structure.

Graph sparsity

We consider a generalization of the group sparsity idea that employs a directed graph structure G on $\mathcal{I}$. Each element of $\mathcal{I}$ is a node of G but G may contain additional nodes. For simplicity, we assume G contains a starting node not in $\mathcal{I}$. At each node $v \in G$, we define coding length $\mathrm{cl}_v(S)$ on the neighborhood N_v of v (that contains the empty set), as well as any other single node $u \in G$ with $\mathrm{cl}_v(u)$, such that $\sum_{S \subset N_v} 2^{-\mathrm{cl}_v(S)} + \sum_{u \in G} 2^{-\mathrm{cl}_v(u)} \leq 1$. To encode $F \subset G$, we start with the active set containing only the starting node, and finish when the set becomes empty. At each node v before termination, we may either pick a subset $S \subset N_v$, with coding length $\mathrm{cl}_v(S)$, or a node in $u \in G$, with coding length $\mathrm{cl}_v(u)$, and then put the selection into the active set. We then remove v from the active set (once v is removed, it does not return to the active set anymore). This process is continued until the active set becomes empty.

The wavelet coefficients of a signal are well known to have a tree-graph structure, which has been widely used for compressing natural images and is a special case of graph sparsity. Each wavelet coefficient of the signal is connected to its parent coefficient and its child coefficients. The wavelet coefficients of 1D signals have a binary tree connected graph structure while the wavelet coefficients of 2D images have a quad-tree connected graph structure.

As a concrete example, we consider image processing problem, where each image is a rectangle of pixels (nodes); each pixel is corrected to four adjacent pixels, which forms the underlying graph structure. At each pixel, the number of subsets in its neighborhood is $2^4 = 16$ (including the empty set), with a coding length $\mathrm{cl}_v(S) = 5$ each; we also encode all other pixels in the image with random jumping, each with a coding length $1 + \log_2 p$. Using this scheme, we can encode each connected region F by no more than $\log_2 p + 5|F|$ bits by growing the region from a single point in the region. Therefore if F is composed of g connected regions, then the coding length is $g \log_2 p + 5|F|$, which can be significantly better than standard sparse coding length of $|F| \log_2 p$. This example shows that the general graph coding scheme presented here favors connected regions (that is, nodes that are grouped together with respect to the graph structure). This scheme can be efficiently approximated with block coding as follows: we consider relatively small sized base blocks consisted of nodes that are close together with respect to the graph structure, and then use the induced block coding scheme to approximate the graph coding.

4. Algorithms for Structured Sparsity

The following algorithm is a natural extension of L_0 regularization to structured sparsity problems. It penalizes the coding complexity instead of the cardinality (sparsity) of the feature set.

$$\hat{\beta}_{constr} = \arg\min_{\beta\in\mathbb{R}^p} \hat{Q}(\beta) \quad \text{subject to } c(\beta) \le s. \qquad (1)$$

The optimization of (1) is generally hard. There are two common approaches to alleviate this problem. One is convex relaxation (L_1 regularization to replace L_0 regularization for standard sparsity); the other is forward greedy algorithm. We do not know any extensions of L_1 regularization like convex relaxation that can handle general structured sparsity formulations. However, one can extend greedy algorithm by using a block structure. We call the resulting procedure structured greedy algorithm (see Algorithm 1), which approximately solves (1).

It is important to understand that the block structure is only used to limit the search space in the greedy algorithm. The actual coding scheme does not have to be the corresponding block coding. It is also useful to understand that our result does not imply that the algorithm won't be effective if the actual coding scheme cannot be approximated by block coding.

Algorithm 1 Structured Greedy Algorithm (StructOMP)

Input: $(X, \mathbf{y})$, $\mathcal{B} \subset 2^{\mathcal{I}}$, $s > 0$
Output: $F^{(k)}$ and $\beta^{(k)}$
let $F^{(0)} = \emptyset$ and $\beta^{(0)} = 0$
for all $K = 1, ...$ **do**
  select $B^{(k)} \in \mathcal{B}$ to maximize progress $(*)$
  let $F^{(k)} = B^{(k)} \cup F^{(k-1)}$
  let $\beta^{(k)} = \arg\min_{\beta\in\mathbb{R}^p} \hat{Q}(\beta)$ subject to $\mathrm{supp}(\beta) \subset F^{(k)}$
  if $(c(\beta^{(k)}) > s)$ **break**
end for

In Algorithm 1, we are given a set of blocks $\mathcal{B}$ that contains subsets of $\mathcal{I}$. Instead of searching all subsets $F \subset \mathcal{I}$ up to a certain complexity $|F| + c(F)$, which is computationally infeasible, we search only the blocks restricted to $\mathcal{B}$. It is assumed that searching over $\mathcal{B}$ is computationally manageable. At each step $(*)$, we try to find a block from $\mathcal{B}$ to maximize progress. It is thus necessary to define a quantity that measures progress. Our idea is to approximately maximize the gain ratio:

$$\lambda^{(k)} = \frac{\hat{Q}(\beta^{(k-1)}) - \hat{Q}(\beta^{(k)})}{c(\beta^{(k)}) - c(\beta^{k-1})},$$

which measures the reduction of objective function per unit increase of coding complexity. This greedy criterion is a natural generalization of the standard greedy algorithm, and essential in our analysis. For least squares regression, we can approximate $\lambda^{(k)}$ using the following definition

$$\phi(B) = \frac{\|P_{B-F^{(k-1)}}(X\beta^{(k-1)} - \mathbf{y})\|_2^2}{c(B \cup F^{(k-1)}) - c(F^{(k-1)})}, \qquad (2)$$

where $P_F = X_F(X_F^\top X_F)^{-1}X_F^\top$ is the projection matrix to the subspaces generated by columns of X_F. We then select $B^{(k)}$ so that

$$\phi(B^{(k)}) \ge \gamma \max_{B\in\mathcal{B}} \phi(B),$$

where $\gamma \in (0, 1]$ is a fixed approximation ratio that specifies the quality of approximate optimization.

5. Theory of Structured Sparsity

Due to the space limitation, the proofs of the theorems are detailed in (Huang et al., 2009).

5.1. Assumptions

We assume sub-Gaussian noise as follows.

Assumption 5.1 *Assume that $\{\mathbf{y}_i\}_{i=1,...,n}$ are independent (but not necessarily identically distributed) sub-Gaussians: there exists a constant $\sigma \ge 0$ such that $\forall i$ and $\forall t \in R$, $\mathbb{E}_{\mathbf{y}_i}\, e^{t(\mathbf{y}_i - \mathbb{E}\mathbf{y}_i)} \le e^{\sigma^2 t^2/2}$.*

We also need to generalize sparse eigenvalue condition, used in the modern sparsity analysis. It is related to (and weaker than) the RIP (restricted isometry property) assumption (Candes & Tao, 2005) in the compressive sensing literature. This definition takes advantage of coding complexity, and can be also considered as (a weaker version of) structured RIP. We introduce a definition.

Definition 5.1 *For all $F \subset \{1, \ldots, p\}$, define*

$$\rho_-(F) = \inf\left\{\frac{1}{n}\|X\beta\|_2^2/\|\beta\|_2^2 : \mathrm{supp}(\beta) \subset F\right\},$$
$$\rho_+(F) = \sup\left\{\frac{1}{n}\|X\beta\|_2^2/\|\beta\|_2^2 : \mathrm{supp}(\beta) \subset F\right\}.$$

Moreover, for all $s > 0$, define

$$\rho_-(s) = \inf\{\rho_-(F) : F \subset \mathcal{I}, c(F) \le s\},$$
$$\rho_+(s) = \sup\{\rho_+(F) : F \subset \mathcal{I}, c(F) \le s\}.$$

In the theoretical analysis, we need to assume that $\rho_-(s)$ is not too small for some s that is larger than the signal complexity. Since we only consider eigenvalues for submatrices with small cost $c(\bar{\beta})$, the sparse eigenvalue $\rho_-(s)$ can be significantly larger than the corresponding

ratio for standard sparsity (which will consider all subsets of $\{1,\ldots,p\}$ up to size s). For example, for random projections used in compressive sensing applications, the coding length $c(\mathrm{supp}(\bar\beta))$ is $O(k\ln p)$ in standard sparsity, but can be as low as $c(\mathrm{supp}(\bar\beta)) = O(k)$ in structured sparsity (if we can guess $\mathrm{supp}(\bar\beta)$ approximately correctly. Therefore instead of requiring $n = O(k\ln p)$ samples, we requires only $O(k+\mathrm{cl}(\mathrm{supp}(\bar\beta)))$. The difference can be significant when p is large and the coding length $\mathrm{cl}(\mathrm{supp}(\bar\beta)) \ll k\ln p$.

More precisely, we have the following random projection sample complexity bound for the structured sparse eigenvalue condition. The theorem implies that the structured RIP condition is satisfied with sample size $n = O((k/k_0)\ln(p/k_0))$ in group sparsity rather than $n = O(k\ln(p))$ in standard sparsity.

Theorem 5.1 (Structured-RIP) *Suppose that elements in X are iid standard Gaussian random variables $N(0,1)$. For any $t>0$ and $\delta\in(0,1)$, let*

$$n \geq \frac{8}{\delta^2}[\ln 3 + t + s\ln(1+8/\delta)].$$

Then with probability at least $1-e^{-t}$, the random matrix $X\in\mathbb{R}^{n\times p}$ satisfies the following structured-RIP inequality for all vector $\bar\beta\in\mathbb{R}^p$ with coding complexity no more than s:

$$(1-\delta)\|\bar\beta\|_2 \leq \frac{1}{\sqrt{n}}\|X\bar\beta\|_2 \leq (1+\delta)\|\bar\beta\|_2.$$

5.2. Coding complexity regularization

Theorem 5.2 *Suppose that Assumption 5.1 is valid. Consider any fixed target $\bar\beta\in\mathbb{R}^p$. Then with probability exceeding $1-\eta$, for all $\lambda\geq 0, \epsilon\geq 0, \hat\beta\in\mathbb{R}^p$ such that: $\hat Q(\hat\beta)\leq \hat Q(\bar\beta)+\epsilon$, we have*

$$\|X\hat\beta - \mathbb{E}\mathbf{y}\|_2 \leq \|X\bar\beta - \mathbb{E}\mathbf{y}\|_2 + \sigma\sqrt{2\ln(6/\eta)} + 2\Gamma,$$
$$\Gamma = (7.4\sigma^2 c(\hat\beta) + 2.4\sigma^2\ln(6/\eta)+\epsilon)^{1/2}.$$

Moreover, if the coding scheme $c(\cdot)$ is sub-additive, then

$$n\rho_-(c(\hat\beta)+c(\bar\beta))\|\hat\beta-\bar\beta\|_2^2 \leq 10\|X\bar\beta-\mathbb{E}\mathbf{y}\|_2^2 + \Delta,$$
$$\Delta = 37\sigma^2 c(\hat\beta) + 29\sigma^2\ln(6/\eta) + 2.5\epsilon.$$

This theorem immediately implies the following result for (1): $\forall\bar\beta$ such that $c(\bar\beta)\leq s$,

$$\frac{1}{\sqrt{n}}\|X\hat\beta_{constr} - \mathbb{E}\mathbf{y}\|_2 \leq \frac{1}{\sqrt{n}}\|X\bar\beta-\mathbb{E}\mathbf{y}\|_2 + \Lambda,$$
$$\Lambda = \frac{\sigma}{\sqrt{n}}\sqrt{2\ln(6/\eta)} + \frac{2\sigma}{\sqrt{n}}(7.4s+4.7\ln(6/\eta))^{1/2},$$
$$\|\hat\beta_{constr}-\bar\beta\|_2^2 \leq \frac{1}{\rho_-(s+c(\bar\beta))n}\left[10\|X\bar\beta-\mathbb{E}\mathbf{y}\|_2^2+\Pi\right],$$
$$\Pi = 37\sigma^2 s + 29\sigma^2\ln(6/\eta).$$

In compressive sensing applications, we take $\sigma=0$, and we are interested in recovering $\bar\beta$ from random projections. For simplicity, we let $X\bar\beta = \mathbb{E}\mathbf{y} = \mathbf{y}$, and our result shows that the constrained coding complexity penalization method achieves exact reconstruction $\hat\beta_{constr}=\bar\beta$ as long as $\rho_-(2c(\bar\beta))>0$ (by setting $s=c(\bar\beta)$). According to Theorem 5.1, this is possible when the number of random projections (sample size) reaches $n = O(2c(\bar\beta))$. This is a generalization of corresponding results in compressive sensing (Candes & Tao, 2005). As we have pointed out earlier, this number can be significantly smaller than the standard sparsity requirement of $n = O(\|\bar\beta\|_0\ln p)$, when the structure imposed is meaningful.

5.3. Structured greedy algorithm

Definition 5.2 *Given $\mathcal{B}\subset 2^{\mathcal{I}}$, define*

$$\rho_0(\mathcal{B}) = \max_{B\in\mathcal{B}}\rho_+(B),\ c_0(\mathcal{B}) = \max_{B\in\mathcal{B}} c(B)$$

and

$$c(\bar\beta,\mathcal{B}) = \min\left\{\sum_{j=1}^b c(\bar B_j) : \mathrm{supp}(\bar\beta)\subset\bigcup_{j=1}^b \bar B_j\ (\bar B_j\in\mathcal{B})\right\}.$$

The following theorem shows that if $c(\bar\beta,\mathcal{B})$ is small, then one can use the structured greedy algorithm to find a coefficient vector $\beta^{(k)}$ that is competitive to $\bar\beta$, and the coding complexity $c(\beta^{(k)})$ is not much worse than that of $c(\bar\beta,\mathcal{B})$. This implies that if the original coding complexity $c(\bar\beta)$ can be approximated by block complexity $c(\bar\beta,\mathcal{B})$, then we can approximately solve (1).

Theorem 5.3 *Suppose the coding scheme is sub-additive. Consider $\bar\beta$ and ϵ such that $\epsilon\in(0,\|\mathbf{y}\|_2^2-\|X\bar\beta-\mathbf{y}\|_2^2]$ and*

$$s \geq \frac{\rho_0(\mathcal{B})c(\bar\beta,\mathcal{B})}{\gamma\rho_-(s+c(\bar\beta))}\ln\frac{\|\mathbf{y}\|_2^2-\|X\bar\beta-\mathbf{y}\|_2^2}{\epsilon}.$$

Then at the stopping time k, we have

$$\hat Q(\beta^{(k)}) \leq \hat Q(\bar\beta)+\epsilon.$$

By Theorem 5.2, the result in Theorem 5.3 implies that

$$\|X\beta^{(k)} - \mathbb{E}\mathbf{y}\|_2 \leq \|X\bar\beta-\mathbb{E}\mathbf{y}\|_2 + \sigma\sqrt{2\ln(6/\eta)} + \Lambda,$$
$$\Lambda = 2\sigma(7.4(s+c_0(\mathcal{B})) + 4.7\ln(6/\eta) + \epsilon/\sigma^2)^{1/2},$$
$$\|\beta^{(k)}-\bar\beta\|_2^2 \leq \frac{\left[10\|X\bar\beta-\mathbb{E}\mathbf{y}\|_2^2+\Pi\right]}{\rho_-(s+c_0(\mathcal{B})+c(\bar\beta))n},$$
$$\Pi = 37\sigma^2(s+c_0(\mathcal{B})) + 29\sigma^2\ln(6/\eta) + 2.5\epsilon.$$

The result shows that in order to approximate a signal $\bar\beta$ up to ϵ, one needs to use coding complexity

$O(\ln(1/\epsilon))c(\bar{\beta}, \mathcal{B})$. If $\mathcal{B}$ contains small blocks and their sub-blocks with equal coding length, and the coding scheme is block coding generated by $\mathcal{B}$, then $c(\bar{\beta}, \mathcal{B}) = c(\bar{\beta})$. In this case we need $O(s \ln(1/\epsilon))$ to approximate a signal with coding complexity s.

In order to get rid of the $O(\ln(1/\epsilon))$ factor, backward greedy strategies can be employed, as shown in various recent work such as (Zhang, 2008). For simplicity, we will not analyze such strategies in this paper. However, in the following, we present an additional convergence result for structured greedy algorithm that can be applied to weakly sparse p-compressible signals common in practice. It is shown that the $\ln(1/\epsilon)$ can be removed for such weakly sparse signals. More precisely, we introduce the following concept of weakly sparse compressible target that generalizes the corresponding concept of compressible signal in standard sparsity from the compressive sensing literature (Donoho, 2006).

Definition 5.3 *The target $\mathbb{E}\mathbf{y}$ is (a, q)-compressible with respect to block $\mathcal{B}$ if there exist constants $a, q > 0$ such that for each $s > 0$, $\exists \bar{\beta}(s)$ such that $c(\bar{\beta}(s), \mathcal{B}) \leq s$ and*

$$\frac{1}{n}\|X\bar{\beta}(s) - \mathbb{E}\mathbf{y}\|_2^2 \leq as^{-q}.$$

Theorem 5.4 *Suppose that the target is (a, q)-compressible with respect to $\mathcal{B}$. Then with probability $1 - \eta$, at the stopping time k, we have*

$$\hat{Q}(\beta^{(k)}) \leq \hat{Q}(\bar{\beta}(s')) + 2na/s'^q + 2\sigma^2[\ln(2/\eta) + 1],$$

where

$$s' \leq \frac{s\,\gamma}{(10 + 3q)\rho_0(\mathcal{B})} \min_{u \leq s'} \rho_-(s + c(\bar{\beta}(u))).$$

This result shows that we can approximate a compressible signal of complexity s' with complexity $s = O(qs')$ using greedy algorithm. This means the greedy algorithm obtains optimal rate for weakly-sparse compressible signals. The sample complexity suffers only a constant factor $O(q)$. Combine this result with Theorem 5.2, and take union bound, we have with probability $1 - 2\eta$, at stopping time k:

$$\frac{1}{\sqrt{n}}\|X\beta^{(k)} - \mathbb{E}\mathbf{y}\|_2 \leq \sqrt{\frac{a}{s'^q}} + \sigma\sqrt{\frac{2\ln(6/\eta)}{n}} + 2\sigma\sqrt{\Lambda},$$

$$\Lambda = \frac{7.4(s + c_0(\mathcal{B})) + 6.7\ln(6/\eta)}{n} + \frac{2a}{\sigma^2 s'^q},$$

$$\|\beta^{(k)} - \bar{\beta}\|_2^2 \leq \frac{1}{\rho_-(s + s' + c_0(\mathcal{B}))}\left[\frac{15a}{s'^q} + \frac{\Pi}{n}\right],$$

$$\Pi = 37\sigma^2(s + c_0(\mathcal{B})) + 34\sigma^2\ln(6/\eta).$$

Given a fixed n, we can obtain a convergence result by choosing s (and thus s') to optimize the right hand side. The resulting rate is optimal for the special case of standard sparsity, which implies that the bound has the optimal form for structured q-compressible targets. In particular, in compressive sensing applications where $\sigma = 0$, we obtain when samples size reaches $n = O(qs')$, the reconstruction performance is

$$\|\bar{\beta}^{(k)} - \bar{\beta}\|_2^2 = O(a/s'^q),$$

which matches that of the constrained coding complexity regularization method in (1) up to a constant $O(q)$.

6. Experiments

The purpose of these experiments is to demonstrate the advantage of structured sparsity over standard sparsity. We compare the proposed StructOMP to OMP and Lasso, which are standard algorithms to achieve sparsity but without considering structure. In our experiments, we use Lasso-modified least angle regression (LAS/Lasso) as the solver of Lasso (Bradley Efron & Tibshirani, 2004). In order to quantitatively compare performance of different algorithms, we use recovery error, defined as the relative difference in 2-norm between the estimated sparse coefficient vector $\hat{\beta}_{est}$ and the ground-truth sparse coefficient $\bar{\beta}$: $\|\hat{\beta}_{est} - \bar{\beta}\|_2/\|\bar{\beta}\|_2$. Our experiments focus on graph sparsity that is more general than group sparsity. In fact, connected regions may be regarded as dynamic groups that are not pre-defined. For this reason, we do not compare to group-Lasso which requires pre-defined groups.

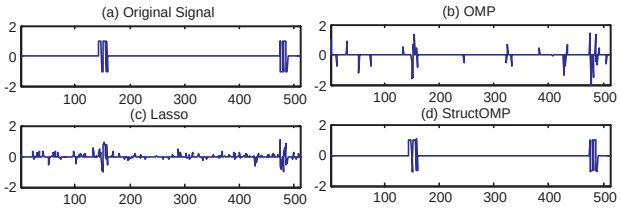
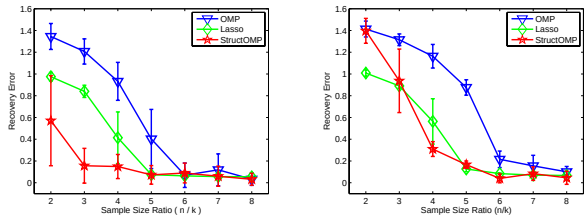


Figure 1. Recovery results of 1D signal with graph-structured sparsity. (a) original data; (b) recovered results with OMP (error is 0.9921); (c) recovered results with Lasso (error is 0.6660); (d) recovered results with StructOMP (error is 0.0993).

6.1. 1D Signals with Line-Structured Sparsity

In the first experiment, we randomly generate a $1D$ structured sparse signal with values ± 1, where $p = 512$, $k = 32$ and $g = 2$. The support set of these signals is composed of g connected regions. Here, each element of the sparse coefficient is connected to two of its adjacent elements, which forms the underlying graph structure. The graph sparsity concept introduced earlier is used to compute the coding length of sparsity patterns in StructOMP. The projection matrix X is generated by creating an $n \times p$ matrix with i.i.d. draws from a standard Gaussian distribution $N(0, 1)$.

For simplicity, the rows of X are normalized to unit magnitude. Zero-mean Gaussian noise with standard deviation $\sigma = 0.01$ is added to the measurements. Figure 1 shows one generated signal and its recovered results by different algorithms when $n = 4k = 128$. To study how the sample size n effects the recovery performance, we change the sample size and record the recovery results by different algorithms. Figure 2(a) shows the recovery performance of the three algorithms, averaged over 100 random runs for each sample size. As expected, StructOMP is better than the OMP and Lasso and can achieve better recovery performance for structured sparsity signals with less samples.



(a) (b)

Figure 2. Recovery error vs. Sample size ratio (n/k): a) 1D signals; (b) 2D gray images

6.2. 2D Images with Graph-structured Sparsity

To demonstrate the structure sparsity concept on 2D images, we randomly generate a $2D$ structured sparsity image by putting four letters in random locations, where $p = H * W = 48 * 48$, $k = 160$ and $g = 4$. The support set of these signals is thus composed of g connected regions. Here, each pixel of the 2D gray image is connected to four of its adjacent pixels, which forms the underlying graph structure. The graph sparsity coding scheme discussed earlier is applied to calculate coding length of a sparsity pattern. Figure 3 shows one example of 2D gray images and the recovered results by different algorithms when $m = 4k = 640$. We also record the recovery results by different algorithms with increasing sample sizes. Figure 2(b) shows the recovery performance of the three algorithms, averaged over 100 random runs for each sample size. The recovery results of StructOMP are always better than those of OMP. Comparing to Lasso, however, the difference is not always clear cut. This result is reasonable, considering that this artificial signal is strongly sparse, and our theory says that OMP works best for weakly sparse signals. For strongly sparse signals, recovery bounds for Lasso are known to be better than that of OMP. However, as shown in the next two examples, real data are often not strongly sparse, and StructOMP can significantly outperform Lasso. We shall mention that a few recent works have shown that the backward greedy strategies can be added to further improve the forward greedy methods and obtain similarly results as those of L_1 regularization based methods (Needell & Tropp, 2008)(Zhang, 2008). It will be a future work to include such modifications into StructOMP.

(a) (b) (c) (d)

Figure 3. Recovery results of a 2D gray image: (a) original gray image, (b) recovered image with OMP (error is 0.9012), (c) recovered image with Lasso (error is 0.4556) and (d) recovered image with StructOMP (error is 0.1528)

6.3. 2D Images with Tree-structured Sparsity

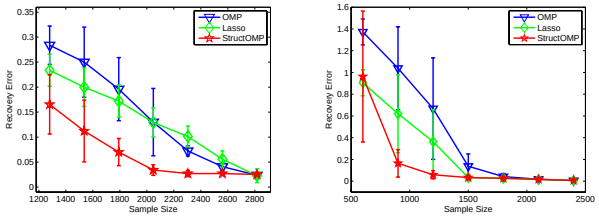
It is well known that the 2D natural images are sparse in a wavelet basis. Their wavelet coefficients have a hierarchical tree structure (Mallat,). Figure 4(a) shows a widely used example image with size 64×64: *cameraman*. Each 2D wavelet coefficient of this image is connected to its parent coefficient and child coefficients, which forms the underlying hierarchical tree structure (which is a special case of graph sparsity). In our experiment, we choose Haar-wavelet to obtain its tree-structured sparsity wavelet coefficients. The projection matrix X and noises are generated with the same method as that for 1D structured sparsity signals. OMP, Lasso and StructOMP are used to recover the wavelet coefficients from the random projection samples respectively. Then, the inverse wavelet transform is used to reconstruct the images with these recovered wavelet coefficients. Our task is to compare the recovery performance of the StructOMP to those of OMP and Lasso. Figure 4 shows one example of the recovered results by different algorithms. It shows that StructOMP obtains the best recovered result. Figure 5(a) shows the recovery performance of the three algorithms, averaged over 100 random runs for each sample size. The StructOMP algorithm is better than both Lasso and OMP in this case. The difference of this example from the previous example is that real image data are only weakly sparse, for which even the standard OMP (without structured sparsity) bound obtained in this paper matches that of Lasso. It is thus consistent with our theory that StructOMP should outperform unstructured Lasso in this case.

6.4. Background Subtracted Images

Background subtracted images are typical structure sparsity data in static video surveillance applications. They generally correspond to the foreground objects of interest. These images are not only spatially sparse but also inclined to cluster into groups, which correspond to different foreground objects. In this experiment, the testing video is downloaded from

(a) (b) (c) (d)

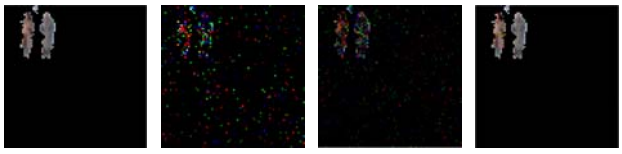
Figure 4. Recovery results with sample size $n = 2048$: (a) the background subtracted image, (b) recovered image with OMP (error is 0.21986), (c) recovered image with Lasso (error is 0.1670) and (d) recovered image with StructOMP (error is 0.0375)



(a) (b)

Figure 5. Recovery error vs. Sample size: a) 2D image with tree-structured sparsity in wavelet basis; (b) background subtracted images with structured sparsity

http://homepages.inf.ed.ac.uk/rbf/CAVIARDATA1/. One sample image frame is shown in Figure 6(a). Each pixel of the 2D background subtracted image is connected to four of its adjacent pixels, forming the underlying graph structure. We randomly choose 100 background subtracted images as test images. The recovery performance is recorded as a function of increasing sample sizes. Figure 6 and Figure 5(b) demonstrate that StructOMP significantly outperforms OMP and Lasso in recovery performance on video data.



(a) (b) (c) (d)

Figure 6. Recovery results with sample size $n = 900$: (a) the background subtracted image, (b) recovered image with OMP (error is 1.1833), (c) recovered image with Lasso (error is 0.7075) and (d) recovered image with StructOMP (error is 0.1203)

7. Conclusion

This paper develops a theory for structured sparsity where prior knowledge allows us to prefer certain sparsity patterns to others. A general framework is established based on a coding scheme, which includes the group sparsity idea as a special case. The proposed structured greedy algorithm is the first efficient algorithm to handle the general structured sparsity learning. Experimental results demonstrate that significant improvements can be obtained on some real problems that have natural structures, and the results are consistent with our theory. Future work include additional computationally efficient methods such as convex relaxation methods and backward greedy strategies.

References

Bach, F. R. (2008). Consistency of the group lasso and multiple kernel learning. *Journal of Machine Learning Research*, *9*, 1179–1225.

Baraniuk, R., Cevher, V., Duarte, M., & Hegde, C. (2008). Model based compressive sensing. preprint.

Bradley Efron, Trevor Hastie, I. J., & Tibshirani, R. (2004). Least angle regression. *Annals of Statistics*, *32*, 407–499.

Candes, E. J., & Tao, T. (2005). Decoding by linear programming. *IEEE Trans. on Information Theory*, *51*, 4203–4215.

Daudet, L. (2004). Sparse and structured decomposition of audio signals in overcomplete spaces. *International Conference on Digital Audio Effects*.

Donoho, D. (2006). Compressed sensing. *IEEE Transactions on Information Theory*, *52*, 1289–1306.

Grimm, D., Netzer, T., & Schweighofer, M. (2007). A note on the representation of positive polynomials with structured sparsity. *Arch. Math.*, *89*, 399–403.

Huang, J., & Zhang, T. (2009). *The benefit of group sparsity* (Technical Report). Rutgers University.

Huang, J., Zhang, T., & Metaxas, D. (2009). *Learning with structured sparsity* (Technical Report). Rutgers University. available from http://arxiv.org/abs/0903.3002.

Mallat, S. *A Wavelet Tour of Signal Processing*. Academic Press.

Needell, D., & Tropp, J. (2008). Cosamp: Iterative signal recovery from incomplete and inaccurate samples. *Applied and Computational Harmonic Analysis*. Accepted.

Stojnic, M., Parvaresh, F., & Hassibi, B. (2008). On the reconstruction of block-sparse signals with an optimal number of measurements. Preprint.

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

Tropp, J., & Gilbert, A. (2007). Signal recovery from random measurements via orthogonal matching pursuit. *IEEE Transactions on Information Theory*, *53*, 4655–4666.

Yuan, M., & Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of The Royal Statistical Society Series B*, *68*, 49–67.

Zhang, T. (2008). Adaptive forward-backward greedy algorithm for learning sparse representations. *Proceedings of NIPS*.

Zhao, P., Rocha, G., & Yu, B. Grouped and hierarchical model selection through composite absolute penalties. *The Annals of Statistics*. to appear.