Greedy Selections for Structured Sparse Probabilistic Projections

Anonymous Author(s)
Affiliation
Address
email

Abstract

We introduce a framework for designing structured sparsity inducing priors in Probabilistic models using Information Projection. Our approach is flexible and can be used for any structure that allows an enumeration using a *matroid*. By leveraging advancements in submodular optimization, our framework makes greedy selections on the said matroid for efficient inference and guarantees provable approximations to the best possible projections. We complement the algorithmic development and theoretical guarantees with strong empirical performance on simulated and real world fMRI datasets compared to established baselines for three special cases of our framework - Group Sparse Regression, Group Sparse PCA, and Sparse CCA.

1 Introduction

With information being generated at a rate that greatly overshadows the advancement in available computing prowess, models that encourage parsimony become important in more ways than one. There is growing evidence that suggests that the data in many scientific and commercial fields inherently incorporates some latent parsimonious structure that in principle could be exploited for better generalization and robustness. However, in many such disciplines, obtaining actual samples of the data can be expensive. For such models, a-priori domain knowledge obtained from expertise and experience becomes vital to recovering meaningful models. Bayesian approaches are especially suited for incorporating the said knowledge by attuning the prior design to the a-prior knowledge and constraints at hand. Specifically, for sparse structures, such domain knowledge can be incorporated as sparsity inducing priors.

Over the past few years, sparsity has gained eminence in several fields. A natural extension to the classical notion of sparsity is *structured* sparsity - that allows additional information to be captured within the prior to be designed. Some examples of structured sparsity could include *smoothness* [Koyejo et al., 2014, Khanna et al., 2015], group sparsity [Witten et al., 2009, Jenatton et al., 2010, Liu et al., 2010, Simon et al., 2013], tree/graph sparsity [Hegde et al., 2015] etc. For sparse probabilistic models, there is a significant body of literature that incorporates the classical notion of sparsity [Archambeau and Bach, 2009, Koyejo et al., 2014, Khanna et al., 2015] and references therein. However, there is little work for structured sparsity. This is because what probabilistic models gain by allowing for greater flexibility, they lose on efficient inference techniques that are faithful to the desired structure.

In this work, we propose a framework for incorporating structured sparsity in probabilistic models by leveraging the recent advancements in research on efficient approximate inference for restriction to sparse supports by Information Projection. More specifically, Koyejo et al. [2014] showed that restriction of a density to a parsimonious support structure can be posed as a variational optimization problem of finding the Information Projection of the said density onto the set of all densities supported on the structured support. .In other words, the problem of restriction of a density can be reduced

Submitted to 29th Conference on Neural Information Processing Systems (NIPS 2016). Do not distribute.

to solving a KL minimization problem. They also prove that when the constraint set is of k sparse supports, the KL minimization can be reduced to a submodular optimization and a greedy algorithm can be used efficient approximate inference. Khanna et al. [2015] applied the Information Projection based restriction to Sparse PCA.

We extend the works of Koyejo et al. [2014], Khanna et al. [2015] by showing that the submodularity property can be exploited to give efficient inference schemes for *any* structured sparsity constraint as long as it can be enumerated by a matroid. For the general matroids, an approximation of $1/2$ to the best possible approximation is guaranteed by the theory of submodular optimization. However, for some special cases such as cardinality constraints (classical sparsity), and group sparsity, stronger guarantees of $1 - 1/e$ are available.

Our contributions are as follows: (1) we present a framework for designing structured sparsity priors under any matroid constraint; (2) we present an inference scheme that is both efficient while incorporating the desired structure (3) we show information projection under group sparsity and multi-view sparsity are submodular with knapsack constraint and partition matroid constraints respectively, leading to direct applications of group sparse regression and PCA, and Sparse CCA; (4) we present strong empirical performance of application of our suggested techniques to simulated data and real world fMRI datasets.

Notation. We represent vectors as small letter bolds e.g. $\mathbf{u}$. Matrices are represented by capital bolds e.g. $\mathbf{X}, \mathbf{T}$. Matrix transposes are represented by superscript $\top$. Identity matrices of size s are represented by $\mathbf{I}_s$. $\mathbf{1}(\mathbf{0})$ is a column vector of all ones (zeroes). The i^{th} row of a matrix $\mathbf{M}$ is indexed as $\mathbf{M}_{i,\cdot}$, while j^{th} column is $\mathbf{M}_{\cdot,j}$. We use $\mathrm{p}(\cdot), \mathrm{q}(\cdot)$ to represent probability densities over random variables which may be scalar, vector, or matrix valued which shall be clear from context. Sets are represented by sans serif fonts e.g. S, complement of a set S is S^c. For a vector $\mathbf{u} \in \mathbb{R}^d$, and a set S of support dimensions with $|\mathsf{S}| = k, k \leq d$, $\mathbf{u}_\mathsf{S} \in \mathbb{R}^k$ denotes subvector of $\mathbf{u}$ supported on S. Similarly, for a matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$, $\mathbf{X}_\mathsf{S} \in \mathbb{R}^{k \times k}$ denotes the submatrix supported on S. We denote $\{1, 2, \ldots, d\}$ as $[d]$. Let $\mathfrak{p}(d)$ be the power set of $[d]$.

The rest of the paper is as follows. We present some relevant required definitions in Section 2. The relevant details of previous work that we build upon are also included in this section for completeness. In Section 3 we present our framework of general structured sparsity, and an applications to group sparsity. We apply the developed framework on Section 4 on three problems that can make use of the sparse structured priors. Finally, we provide the experimental results in Section 5.

2 Background

PCA and CCA: The deterministic PCA problem is to find the top-eigen vector of a psd matrix $\mathbf{T}$:

$$\max_{\mathbf{w} \in \mathsf{S}} \mathbf{w}^\top \mathbf{T} \mathbf{w}.$$

The set S refers to the constraint set. When no structure is desired in $\mathbf{w}$, S is the unit norm-2 ball. If $\mathbf{w}$ is required to be k-sparse, $\mathsf{S} := \{\mathbf{x} : \|\mathbf{x}\|_0 \leq k, \|\mathbf{x}\|_2 = 1\}$. The deterministic CCA is extension of PCA to multiple views. For 2 views $\mathbf{X}, \mathbf{Y}$, the constrained CCA problem can be written:

$$\max_{\mathbf{x} \in \mathsf{X}, \mathbf{y} \in \mathsf{Y}} \frac{\mathbf{x}^\top \mathbf{X}^\top \mathbf{Y} \mathbf{y}}{|\mathbf{x}^\top \mathbf{X} \mathbf{x}||\mathbf{y}^\top \mathbf{Y} \mathbf{y}|}$$

Probabilistic and Bayesian models employ priors on $\mathbf{w}, \mathbf{x}, \mathbf{y}$ rather than enforcing constraints using sets $\mathsf{S}, \mathsf{X}, \mathsf{Y}$ respectively. The general models for probabilistic PCA and CCA are discussed in Section 4.

Information Projection: Let X be a measurable set, and $\mathrm{p}(\cdot)$ be a probability density defined on X. Let $\mathcal{F}_\mathsf{S}$ be the set of all densities supported on $\mathsf{S} \subset \mathsf{X}$. The information projection of a base density $\mathrm{p}(\cdot)$ supported on the ambient set X onto a constraint (measurable) set $\mathsf{S} \subset \mathsf{X}$ is defined as:

$$\operatorname*{argmin}_{\mathrm{q} \in \mathcal{F}_\mathsf{S}} \mathrm{KL}(\mathrm{q} \| \mathrm{p})$$

For the purpose of our discussion, we assume $\mathcal{F}_\mathsf{S}$ to be a convex set so that the projection is unique.

Submodular functions: Let $f : \mathfrak{p}(d) \to \mathbb{R}$ be a set function. f is a *submodular* function if for all sets x, y in its domain $f(\mathsf{x} \cup \mathsf{y}) + f(\mathsf{x} \cap \mathsf{y}) \leq f(\mathsf{x}) + f(\mathsf{y})$. Further, f is *normalized* if $f(\emptyset) = 0$. f is monotone if for $\mathsf{x} \subset \mathsf{y}$, $f(\mathsf{x}) \leq f(\mathsf{y})$.

Submodular functions are specially interesting because they allow for provable approximation guarantees by using the greedy algorithm and its simple variants for several otherwise NP-Hard combinatorial optimization problems [Nemhauser et al., 1978, Sviridenko, 2004, Calinescu et al., 2011].

Matroids: A matroid is a structure (N, I), where N is the *ground set*, and $\mathsf{I} \subset \mathfrak{p}(\mathsf{N})$ is the family of *independent* sets that satisfy: (1) $\mathsf{B} \in \mathsf{I}, \mathsf{A} \subset \mathsf{B} \implies \mathsf{A} \in \mathsf{I}$, and, (2) $\mathsf{A} \in \mathsf{I}, \mathsf{B} \in \mathsf{I}, |\mathsf{A}| < |\mathsf{B}| \implies \exists x \in \mathsf{B} - \mathsf{A}$ s.t. $\mathsf{A} \cup x \in \mathsf{I}$.

Matroids can be viewed as generalization of independent bases sets in vector spaces, and can be used to encode constraints for combinatorial problems where N is the base set of variables, and I is the set of enumerated candidate solution sets. We consider specific examples here. A *uniform* matroid has I to be set of all possible k and lesser sized subsets of N, and thus induces the k-cardinality constraint. Similarly, a knapsack constraint can be encoded by a matroid which has each candidate solution in I as set of possible groups each with an associated cost so that the total cost of each candidate solution is less than or equal to the knapsack value. Further, a *partition* matroid partitions N into subsets $\{\mathsf{X}_1, \mathsf{X}_2, \ldots, \mathsf{X}_r\}$, with $\mathsf{I} = \{\mathsf{A} \mid \mathsf{A} \subset \mathsf{N}, |\mathsf{A} \cap \mathsf{X}_i| \leq k_i \forall i \in [r]\}$ for given $\{k_1, k_2, \ldots, k_r\}$.

2.1 Priors for Sparsity

The constrained information projection approach to introducing sparsity was introduced by Koyejo et al. [2014]. In this section, we review the relevant results that are fundamental to our development of priors for structured sparsity.

A d dimensional variable $\mathbf{x}$ is k-sparse if it is non-zero on atmost k dimensions. The support of the variable $\mathbf{x} \in \mathbb{R}^d$ is defined as $supp(\mathbf{x}) := \{i \in [d] | \mathbf{x}_i \neq 0\}$. Similarly, a d dimensional probability distribution $\boldsymbol{P}$ is k-sparse if all random variables $\mathbf{x} \sim \boldsymbol{P}$ are k-sparse. Let A be the set of all $\frac{d!}{k!(d-k)!}$ k-sparse support sets. The information projection of a given density p onto A is a natural way of introducing sparsity since it is equivalent to restriction of p onto A [Koyejo et al., 2014]. However, the set A is non-convex and the information projection of a given density onto this set is generally intractable. Koyejo et al. [2014] suggest an approximation by proposing the combinatorial problem:

$$\min_{\mathsf{S} \subset \mathsf{A}} \min_{q \in \mathcal{F}_{\mathsf{S}}} \mathrm{KL}(q \| p) \tag{1}$$

The inner optimization over $\mathcal{F}_{\mathsf{S}}$ is a conditional and the solution can be written in closed form as $\min_{q \in \mathcal{F}_{\mathsf{S}}} \mathrm{KL}(q \| p) = -\log \mathrm{p}(\mathbf{x}_{\mathsf{S}^c} = 0)$ [Koyejo et al., 2014].

Define the function $J : \mathfrak{p}(d) \to \mathbb{R}$ as $J(\mathsf{S}) := \log \mathrm{p}(\mathbf{x}_{\mathsf{S}^c} = 0)$, and the function $\tilde{J} : \mathfrak{p}(d) \to \mathbb{R}$ as $\tilde{J}(\mathsf{S}) := J(\mathsf{S}) - J(\emptyset)$. The optimization problem (1) is then equivalent to

$$\max_{|\mathsf{S}| = k} \tilde{J}(\mathsf{S}) \tag{2}$$

While (2) is combinatorial, the following theorem states a simple greedy solution comes provably close to the true optimum.

Theorem 1 ([Koyejo et al., 2014])**.** *$\tilde{J}(\mathsf{S})$ is normalized monotone submodular.*

Theorem 1 guarantees a $(1 - \frac{1}{e})$ approximate solution by a greedy algorithm [Nemhauser et al., 1978]. This result is based off of a uniform matroid constraint. For other structured sparsity structures, we shall see that generalizations to other matroid constraints result in similar guarantees for simple greedy algorithm variants. In the next section, we make these generalizations, and present the respective algorithms.

3 Priors for Structured Sparsity

In this section, we generalize the cardinality constrained variable selection for designing sparsity inducing priors in two ways. Firstly, we use the information projection framework for designing priors that induce *group* sparsity, and show that the resulting combinatorial problem of selecting the *most relevant* groups is monotone submodular under a knapsack constraint. Secondly, we consider the constraint of partition matroid in which the set of dimensions are pre-grouped into *views* and possible number of selections from each view is capped. We leverage the research in submodular optimization to present the respective variants of the greedy algorithm that provably guarantee constant factor approximations. Finally we present the application of these two extensions for group sparse regression, group sparse probabilistic PCA and probabilistic CCA.

3.1 Group sparsity

Let p be the ambient density in d dimensions. Let $\mathsf{G} = \{\mathsf{G}_1, \mathsf{G}_2, \ldots, \mathsf{G}_r\}$ so that $\forall i, \mathsf{G}_i \subset [d]$ and $\forall i \neq j, \mathsf{G}_i \cap \mathsf{G}_j = \emptyset$. The set G represents the groups of dimensions provided for group sparsity. For the design of the group sparsity inducing prior, we need to solve:

$$\min_{\mathsf{S} \subset [r]} \min_{\substack{\sum_{i \in \mathsf{S}} |\mathsf{G}_i| \leq k \\ supp(\mathrm{q}) \subset \bigcup_{i \in \mathsf{S}} \mathsf{G}_i}} \mathrm{KL}(\mathrm{q} \| \mathrm{p}) \tag{3}$$

Theorem 2. *The group selection problem* (3) *is equivalent to a normalized monotone submodular maximization problem with a knapsack constraint.*

Proof. We prove by mapping (3) to an equivalent problem by performing a variable change.

Let $\mathsf{G}_\mathsf{S} := \bigcup_{i \in \mathsf{S}} \mathsf{G}_i$. Note that the inner optimization $\min_{|\mathsf{G}_\mathsf{S}| \leq k, q \in \mathcal{F}_{\mathsf{G}_\mathsf{S}}} \mathrm{KL}(\mathrm{q} \| \mathrm{p}) = -\log p(\mathbf{x}_{\mathsf{G} \setminus \mathsf{G}_\mathsf{S}})$ [Koyejo et al., 2014].

Define the function $J : \mathfrak{p}(r) \to \mathbb{R}$ as $J(\mathsf{S}) := \log \mathrm{p}(\mathbf{x}_{\mathsf{G} \setminus \mathsf{G}_\mathsf{S}} = 0)$, and the function $\tilde{J} : \mathfrak{p}(r) \to \mathbb{R}$ as $\tilde{J}(\mathsf{S}) := J(\mathsf{S}) - J(\emptyset)$.

Define the costs associated with picking G_i as $c_i = |\mathsf{G}_i| \forall i \in [r]$. The cost function of a set $\mathsf{s} \subset \mathsf{G}$ can thus be written as $c(\mathsf{s}) := \sum_{\forall i \text{ s.t. } \mathsf{G}_i \in \mathsf{s}} c_i$ The optimization problem 3 is then equivalent to $\max_{\sum_{i \in \mathsf{S}} c_i \leq k} \tilde{J}(\mathsf{S})$.

The result follows from Theorem 1.

$\square$

A variant of the greedy algorithm guarantees a constant factor approximation of the chosen support set. Similar to the results of Koyejo et al. [2014], the constant factor approximation is obtained for mapped function $\tilde{J}(\cdot)$, which can be easily backtracked to obtain data dependent bounds on the original combinatorial optimization problem 3.

The re-weighted greedy algorithm with partial enumeration is presented in Algorithm 1. The re-weighting is to make sure that the greedy step choses the best possible myopic marginal gain. However, with the re-weighting alone the approximation factor can be arbitrarily bad. To bound it to a constant factor, partial enumeration is required. Further details are available in the work by Sviridenko [2004].

Theorem 3 ([Sviridenko, 2004])**.** *For* $m = 3$*, Algorithm 1 guarantees a constant factor approximation of* $(1 - \frac{1}{e})$ *for* $\tilde{J}(\cdot)$.

3.2 General Structured Sparsity

Structured sparsity in a random variable could be required to be dictated by constraints other than just cardinality or group sparsity depending on the domain. For example, the sparsity could be constrained by a tree structure so that selection of a parent node implicitly selects all its children as well. The structural constraint can be encoded as a matroid N, I where N are the base set of dimensions, and I represents the set of all possible candidate solutions under the given constraint.

As an example, say the ambient density is four dimensional, and we wish to find the best possible two dimensional density under some cost function (KL divergence in this manuscript) with the sparsity constrained by a tree structure as follows. The tree has $\{1\}$ as the root node, $\{2\}$ as the left child,

Algorithm 1: GreedyPartialEnum ($\mathsf{G}, k, c(\cdot)$)

input Set of groups G, Total max sparsity k, parameter m, cost function $c(\cdot)$

$\mathsf{S}_1 \leftarrow \arg\max_{\mathsf{s} \subset \mathsf{G}, |\mathsf{s}| < m, c(\mathsf{s}) \leq k} \tilde{J}(\mathsf{s})$
$\mathsf{S}_2 \leftarrow \emptyset$
for all $\mathsf{s} \subset \mathsf{G}, |\mathsf{s}| = m, c(\mathsf{s}) \leq k$ **do**
  $\mathsf{S}_3 \leftarrow$ ReweightedGreedy(G, $k - m - 1$, $c(\cdot)$, s)
  if $\tilde{J}(\mathsf{S}_2) \leq \tilde{J}(\mathsf{S}_3)$ **then**
    $\mathsf{S}_2 \leftarrow \mathsf{S}_3$
  end if
end for
Return $\arg\max\{\tilde{J}(\mathsf{S}_1), \tilde{J}(\mathsf{S}_2)\}$

Algorithm 2: ReweightedGreedy ($\bar{\mathsf{G}}$, $\bar{k}$,$c(\cdot)$, $\bar{\mathsf{S}}_2$)

input Set of groups $\bar{\mathsf{G}}$, Total max sparsity $\bar{k}$, cost function $c(\cdot)$, Init groups $\bar{\mathsf{S}}_2$

$\mathsf{A} \leftarrow \bar{\mathsf{S}}_2$
while $\bar{\mathsf{G}} \backslash \mathsf{A} \neq \emptyset$ **do**
  $\mathsf{s}^* \leftarrow \max_{\mathsf{s} \in \bar{\mathsf{G}} \backslash \mathsf{A}} \frac{J(\mathsf{A} \cup \mathsf{s}) - J(\mathsf{A})}{c(\mathsf{s})}$
  if $c(\mathsf{A} \cup \mathsf{s}^*) \leq k$ **then**
    $\mathsf{A} = \mathsf{A} \cup \mathsf{s}^*$
  end if
  $\bar{\mathsf{G}} = \bar{\mathsf{G}} - \mathsf{s}^*$
end while
Return A

$\{3\}$ as its right child. Further $\{4\}$ is the sole child of $\{3\}$. Thus, the set of possible sparse densities is restricted. e.g. $\{1\}$ can never be selected in a resulting 2-sparse density as selecting it implicitly requires selecting all the 4 dimensions. The respective matroid that encodes the structured sparsity is written as as $\mathsf{N} = [4], \mathsf{I} = \{\{2\}, \{3\}, \{4\}, \{3, 4\}\}$.

Special structures for structured sparse projections, such as cardinality constraints [Koyejo et al., 2014, Khanna et al., 2015] and group sparsity (Section 3.1) yield provable constant factor $(1 - 1/e)$ approximation guarantees on the cost function $J(\cdot)$ using simple greedy algorithm variants. A simple greedy algorithm can also be used for support selection under general matroid constraints. The algorithm is outlined in Algorithm 4. Note that the greedy selection algorithms used by Koyejo et al. [2014], Khanna et al. [2015] are special cases of Algorithm 4 with a uniform matroid. Algorithm 1 is *not* a special case, as it exploits the special structure of group sparsity to modify the simple greedy scheme for better than the general guarantees.

For the more general matroid constraints, simple greedy selection admits slightly weaker guarantees.

Theorem 4. *[Calinescu et al., 2011] Algorithm 4 guarantees a $1/2$ factor approximation guarantee on $\tilde{J}(\cdot)$*

We note that better approximation guarantees can be achieved by randomized algorithms [Calinescu et al., 2011] which can require significantly more effort to implement, specially for our application of designing structured sparse priors. Hence, in this manuscript we restrict our attention to simpler greedy selection schemes.

4 Applications

4.1 Group Sparse Linear Regression

Consider a generative model for n samples given by a linear model and an additive Gaussian noise: $\mathbf{y} = \mathbf{Z}\boldsymbol{\beta} + \epsilon,$, where $y \in \mathbb{R}^n$ is the response, $\mathbf{Z} \in \mathbb{R}^{n \times d}$ is the feature matrix, and $\boldsymbol{\beta} \in \mathbb{R}^d$ is the vector of regression weights. The weights have an associated normal prior, $\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{0}, \mathbf{C})$ for a known $\mathbf{C}$. The noise ϵ is drawn from a Gaussian $\epsilon \sim \mathcal{N}(0, \sigma^2)$. The posterior distribution of β is

Algorithm 3: GreedyMatroid(N, I)

input Matroid (N, I)

$\mathsf{A} \leftarrow \emptyset$
while N is not empty **do**
  $s^* \leftarrow \arg\max_{s \in \mathsf{N}} J(\mathsf{A} \cup \{s\}) - J(\mathsf{A})$
  if $\mathsf{A} \cup \{s^*\} \in \mathsf{I}$ **then**
    $\mathsf{A} = \mathsf{A} \cup \{s^*\}$
  end if
  $\mathsf{N} = \mathsf{N} - \{s^*\}$
end while
Return A

Algorithm 4: GreedyMultiView($k_1, k_2, \dots, k_v, m(\cdot)$)

input N, Sparsities $\{k_1, k_2, \dots, k_v\}$, mapping function $m : [d] \to [v]$.

$\mathsf{A} \leftarrow \emptyset$
selected[i]=0, $\forall i \in [v]$
while N is not empty **do**
  $s^* \leftarrow \arg\max_{s \in \mathsf{N}} J(\mathsf{A} \cup \{s\}) - J(\mathsf{A})$
  if selected$[m(s^*)] < k_i$ **then**
    $\mathsf{A} = \mathsf{A} \cup \{s^*\}$
    selected$[m(s^*)]$ +=1
  end if
  $\mathsf{N} = \mathsf{N} - \{s^*\}$
end while
Return A

also a Gaussian, $\mathrm{p}(\boldsymbol{\beta}|\mathbf{y}) \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and can be written in closed form by standard Bayes theorem with $\boldsymbol{\Sigma}^{-1} = \mathbf{C}^{-1} + \frac{1}{\sigma^2}\mathbf{Z}^\top\mathbf{Z}$, and, $\boldsymbol{\mu} = \frac{1}{\sigma^2}\boldsymbol{\Sigma}\mathbf{Z}^\top\mathbf{y}$.

Let $\mathsf{G} = \{\mathsf{G}_1, \mathsf{G}_2, \dots, \mathsf{G}_r\}$ be the given set of groups so that $\forall i \in [r], \mathsf{G}_i \subset [d]$, and $\forall i \neq j, \mathsf{G}_i \cap \mathsf{G}_j = \emptyset$. For sparse group selection, the optimization problem that needs to be solved is then given by (3). It is known that if p in the constrained optimization problem (3) is Gaussian, thenthe solution of (3) is also Gaussian as observed by Koyejo and Ghosh [2013]. Thus, the search for q in (3) can be restricted to Gaussians. Define $\mathbf{r} = \frac{1}{\sigma^2}\mathbf{Z}^\top\mathbf{y}$. It is easy to show by expanding the KL-gap that (3) for group sparse linear regression is equivalent to the submodular maximization problem:

$$\max_{\substack{\mathsf{S} \subset [r] \\ \mathsf{s} = \bigcup_{i \in \mathsf{S}} \mathsf{G}_i \\ |\mathsf{s}| \leq k}} \quad \mathbf{r}_\mathsf{s}^\top [\boldsymbol{\Sigma}^{-1}]_\mathsf{s} \mathbf{r}_\mathsf{s} - \log\det[\boldsymbol{\Sigma}^{-1}]_\mathsf{s}. \tag{4}$$

Once the support s is selected by solving (4), the respective q^* can be obtained as the respective conditional $\mathrm{q}^*(\mathbf{x}) = \mathrm{p}(\mathbf{x}|\mathbf{x}_{\mathsf{s}^c} = 0)$ [Koyejo et al., 2014].

4.2 Group Sparse Probabilistic PCA

Probabilistic PCA aims to factorize a given matrix $\mathbf{T} \in \mathbb{R}^{n\times}$ as $\mathbf{T} \approx \mathbf{x}\mathbf{w}^\top$, where $\mathbf{x} \in \mathbb{R}^n$ is a deterministic vector, and $\mathbf{w} \in \mathbb{R}^d$ is a random variable. Similar to the linear regression (Section 4.1), the generative equation for observed data matrix is $\mathbf{T} = \mathbf{x}\mathbf{w}^\top + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \sigma^2)$. Note that we focus our attention to $\mathbf{x}, \mathbf{w}$ being vectors for simplicity, and in general they can be matrices [Khanna et al., 2015, Tipping and Bishop, 1999]. For group sparse probabilistic PCA, we consider the case where $\mathbf{w}$ is required to be sparse, and has a Gaussian smoothness prior associated with it $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{C})$.

Having observed $\mathbf{T}$, one way to optimize for the underlying $\mathbf{x}, \mathbf{w}$ is by maximizing the log likelihood using an Expectation Maximization algorithm. The EM algorithm optimizes for $\mathbf{x}$ and $\mathbf{w}$ in an alternating manner in the M-step and the E-step respectively. The algorithm can be interpreted as minimizing *free energy* cost function $\mathscr{F}$ Neal and Hinton [1998]. Let $\theta = \{\mathbf{x}, \sigma\}$ represent the set of deterministic parameters of the system. The function is given by:

$$\mathscr{F}(\mathrm{q}(\mathbf{w}), \theta) = -\mathrm{KL}(\mathrm{q}(\mathbf{w}) \| \mathrm{p}(\mathbf{w}|\mathbf{T}; \theta)) + \log \mathrm{p}(\mathbf{T}; \theta),$$

where $\log \mathrm{p}(\mathbf{T}; \theta)$ is the marginal log-likelihood.

The M-step can be interpreted to be the search over the parameter space, keeping the latent random variable $\mathbf{w}$ fixed:

$$\text{M-step: } \max_{\theta} \mathscr{F}(\mathrm{q}(\mathbf{w}), \theta).$$

Similarly, the E-step is the search over the space of distribution q(.) of the latent variables $\mathbf{w}$, keeping the parameters θ fixed:

$$\text{E-step: } \max_{\mathrm{q}} \mathscr{F}(\mathrm{q}(\mathbf{w}), \theta).$$

This view of the EM algorithm provides the flexibility to design algorithms with any E and M steps that monotonically increase $\mathscr{F}$. Note that if the search space of q in the E-step is unconstrained, E-step outputs the true posterior $\mathrm{p}(\mathbf{w}|\mathbf{T}; \theta)$ (so that the KL distance is 0 in $\mathscr{F}$. Constraining the search space of q leads to *variational E-step*. In this section, we restrict the search space for q to sparse supports dictated by the given groups. Since the restriction is imposed by the minimizing the KL distance, the framework developed in Section 3.1 is applicable.

We now derive the explicit equations to apply Algorithm 1. Again, say $\mathsf{G} = \{\mathsf{G}_1, \mathsf{G}_2, \ldots, \mathsf{G}_r\}$ be the given set of groups. The posterior $\mathrm{p}(\mathbf{w}|\mathbf{T}; \theta)$ is Gaussian with $\mathrm{p} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}^{-1} = \mathbf{C}^{-1} + \frac{\|\mathbf{x}\|_2^2}{\sigma^2}$, and $\boldsymbol{\mu} = \frac{1}{\sigma^2}\boldsymbol{\Sigma}\mathbf{T}^\top\mathbf{x}$. Define $\mathbf{r} := \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$. Proceeding as in Section 4.1, the support selection requires the following submodular maximization problem:

$$\max_{\substack{\mathsf{S} \subset [r] \\ \mathsf{s} = \bigcup_{i \in \mathsf{S}} \mathsf{G}_i \\ |\mathsf{s}| \leq k}} \mathbf{r}_{\mathsf{s}}^\top [\boldsymbol{\Sigma}^{-1}]_{\mathsf{s}} \mathbf{r}_{\mathsf{s}} - \log\det[\boldsymbol{\Sigma}^{-1}]_{\mathsf{s}}.$$

Note that the E-step is identical to the group sparse regression optimization with just one feature (the vector $\mathbf{x}$). to The M-step equations for $\mathbf{x}$ and σ^2 are also easily obtained as closed form updates Khanna et al. [2015].

4.3 Sparse Probabilistic CCA

Probabilistic CCA Bach and Jordan [2005], Klami and Kaski [2007], Archambeau and Bach [2008] is a multi-view generalization of Probabilistic PCA. In the PCA setup, n samples of a d dimensional variable are observed as the data matrix $\mathbf{T}$. However, in many applications, multiple *views* of the data are observed that would make little sense to be concatenated as feature vectors in the same space. Hence, we observe n samples of $d_1, d_2, \ldots, d_v$ as matrices $\mathbf{T}_1, \mathbf{T}_2, \ldots, \mathbf{T}_v$ each of which are one of the v views of the observee.

The generative model assumes an underlying parameter $\mathbf{x} \in \mathbb{R}^n$ shared among all the views, and the random variables $\{\mathbf{w}_i \in \mathbb{R}^{d_i}, \forall i \in [v]\}$. Again, as in Section 4.2, we $\mathbf{x}, \mathbf{w}_i$ can be matrices in general but for clarity, we focus on modeling for the top-1 components. The random variables are drawn from Gaussian distributions $\forall i \in [v], \mathbf{w}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{C}_i)$. Each of the view is generated as $\forall i \in [v], \mathbf{T}_i = \mathbf{x}\mathbf{w}^\top + \epsilon$, where the noise $\epsilon \sim \mathcal{N}(0, \sigma^2)$. Further, we wish to infer sparse $\mathbf{w}_i$ so that $\forall i \in [v], |\mathrm{supp}(\mathbf{w}_i)| \leq k_i$ for the supplied k_i.

The underlying parameters can be optimized for using an EM algorithm. Similar to the construction in Section 4.2, a variational E-step can be formulated to honor the sparsity constraints on the random variables. We next that show that the variational E-step solves a submodular maximization problem subject to a partition matroidal constraint, and so Algorithm 4 can be used for efficient $1/2$ order approximation.

Partition matroid: Let N be the base set, and $\{\mathsf{A}_1, \mathsf{A}_2, \ldots, \mathsf{A}_v\}$ be a partition of N. Let $\mathsf{I} = \{\mathsf{S} \text{ s. t. } |\mathsf{S} \cap \mathsf{A}_i| \leq k_i\}$. (N, I) is called a partition matroid.

We now map the sparse PCCA problem to the partition matroidal constrained optimization. Let $\mathbf{T} = [\mathbf{T}_1, \mathbf{T}_2, \ldots, \mathbf{T}_v]$ be the matrix of size $n \times (\sum_i d_i)$ constructed by stacking all the observed views columnwise. Similarly, $\mathbf{w} = [\mathbf{w}_1; \mathbf{w}_2; \ldots; \mathbf{w}_v$ be the vector obtained by end-to-end concatenation of random variable vectors of all views. Define $\mathbf{C} \in \mathbb{R}^{(\sum_i d_i)\times(\sum_i d_i)}$ as the block diagonal matrix with $\mathbf{C}_i$ as its block. The generative model of PCCA can now be equivalently and succinctly encoded as $\mathbf{T} = \mathbf{x}\mathbf{w}^\top + \epsilon$ where $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{C})$, and, $\epsilon \sim \mathcal{N}(0, \sigma^2)$. Further, the partition matroid is easy to construct with $\mathsf{N} = [\sum_i d_i]$, and A_i to to be the respective index set of $\mathbf{w}_i$ in $\mathbf{w}$. Again, proceeding as in Sections 4.2, 4.1, the submodular maximization problem can be written as:

$$\max_{\substack{\mathsf{s}\in\mathsf{I} \\ \text{Matroid}(\mathsf{N},\mathsf{I})}} \quad \mathbf{r}_{\mathsf{s}}^\top [\mathbf{\Sigma}^{-1}]_{\mathsf{s}} \mathbf{r}_{\mathsf{s}} - \log\det[\mathbf{\Sigma}^{-1}]_{\mathsf{s}}.$$

It should be easy to the see that further extension to group sparse CCA is straightforward by tweaking the constraining partition matroid appropriately.

5 Experiments

We now present empirical results comparing the proposed information projection based support selection technique to established baselines for 3 applications, namely group sparse linear regression, group sparse PCA, and sparse CCA. For model verification, we present the group regression results on simulated data, and present group sparse PCA and sparse CCA results on real world fMRI datasets. We implement our method in Python using numpy and scipy libraries. The greedy selection is parallelized by Message Passing Interface using `mpi4py`. We make use of Woodbury matrix inversion identity in the cost function to greedily build up the cost function. This avoids taking explicit inverses that can lead to inconsistencies.

5.1 Simulated data

Most models are built based on certain assumptions that are seldom true in the real world datasets. As such, it is important to verify the built model on simulated data that conforms to the underlying assumptions. In this section, we compare our method of imposing group-based sparsity against the sparse-group lasso [Simon et al., 2013] implemented in the package SLEP [Liu et al., 2010].

We fix the ambient dimension to be $d = 1000$. We generate an arbitrary fixed weight vector $\boldsymbol{\beta} \in \mathbb{R}^d$ with all but $k = 20$ dimensions zeroed out, arbitrarily made into 5 groups of 4 each. We sample from the d-variate normal distribution with identity covariance $n = 1000$ times to get the feature matrix $\mathbf{X} \in \mathbb{R}^{n\times d}$. Finally we obtain the response vector $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}+\epsilon$, where $\epsilon \sim \mathcal{N}(0, \sigma^2)$ with σ^2 being set with varying values of the Signal-to-Noise ratio (SNR) so that SNR=$\{10000, 1000, 100, 10, 1, 0.1\}$ to generate 6 datasets. Note that SNR < 1 implies variance of the noise is more than that of the signal. We split the data $50 - 10 - 40$ into training, validation and test sets. We compare performance of GroupGreedyKL (group selection based on KL projection) and GroupLasso [Simon et al., 2013] on two metrics - the AUC of the support recovered, and R^2 on test data. For both the methods, we assume it is known that $k = 20$. For GroupLasso, we do a parameter sweep to get the best performing numbers while making sure that the sparsity is 20. For each of the 6 different SNRs, data is generated 10 different times randomly and the average results are reported. The results are presented in Figure 1. GroupGreedyKL performs consistently better than GroupLasso, and degrades more gracefully as SNR decreases.

5.2 fMRI data

Neurovault data A key question in functional neuroimaging is the extent to which task brain measurements incorporate distributed regions in the brain. One way to tackle this hypothesis is to decompose a collection of task statistical maps and examine the shared factors. Smith et al. [2009] considered a similar question using the brain map database decomposed via ICA, showing correspondence between task activation factors and resting state factors. Following their approach, we downloaded 1669 fMRI task statistical maps from neurovault (http://neurovault.org/). Each image in the collection represents a standardized statistical map of univariate brain voxel activation in response to an experimental manipulation. The statistical maps were downsampled from 2mm∧3 voxels to

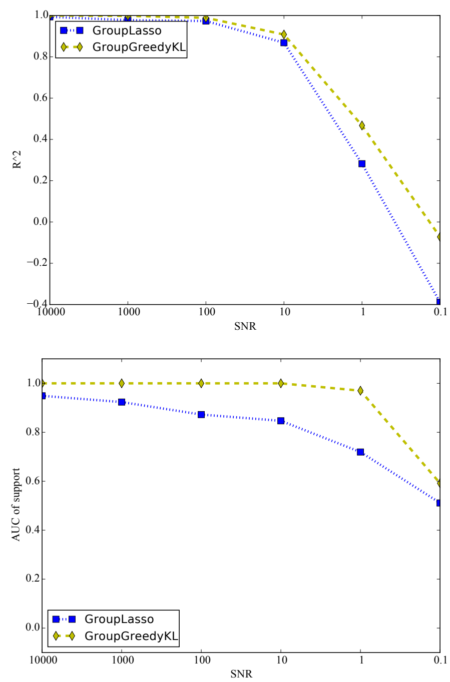

Figure 1: Group Sparse Regression performance on simulated data

3mm∧3 voxels using the nilearn python package (`http://nilearn.github.io/`). We then applied the standard brain mask, removing voxels outsize of the grey matter, resulting in d=65598 variables (dimensions). We incorporate smoothness via spatial correlation matrix $\mathbf{C}$ on the prior on $\mathbf{W}$.

While our greedy algorithm can easily scale to dimensionality of size 65598, the matlab implementation of the baseline is not as scalable. We cluster the original set of dimensions to d =10000 dimensions using the spatially constrained Ward hierarchical clustering approach of Michel et al. [2012]. We further apply the same hierarchical clustering to group the dimensions into 500 groups, with group sizes ranging from 1 to 1500 with average group size close to 20. We apply our information projection based Group Sparse PCA algorithm (GroupPCAKL) developed in Section 4.2. The group sparse constraint specifies that each group can be either wholly included or completely discarded from the model. Our algorithm adheres to this specification. It is possible to have a soft version of the constraint which allows for sparsity within each chosen group. This is typically imposed as a regularization trade-off between sparsity across and within groups. We compare against the Structured Sparse PCA algorithm (GroupPCA) of Jenatton et al. [2010]. We report the ratio of variance explained by the top k-sparse eigenvector at different values of k and show superior performance of GroupPCAKL in Figure 2.

Human Connectonome Project Another interesting question that the neuroscientists are interested to address is about the association of human brain function to human behavior. The brain function and the human behavior can be thought of as two *views* of underlying latent traits. This intuition suggests possible application of the CCA based approaches (Section 4.3). We make use of the Human Connectonome Project data (HCP) [Essen et al., 2013] for this purpose. It consists of a large number of samples of high quality brain imaging and behavioral information collected from several healthy adults. We download and extract brain statistical maps and respective behavioral variances from 497 adult subjects. For behavioral variables, we select the same subset as done by Smith et al. [2015] including those of scores from physiological measurements and behavior questions. For statistical maps, we extract the ones corresponding to the task of n-back. A statistical map is a summary of each voxel in the brain in response to externally applied controlled stimulus. The “n-back” task is designed for the working memory. Items are presented one at a time for the subjects to identify whether an item was item n items ago. Further details on the task are available in the HCP documentation [Essen et al., 2013]. One the extracted maps, we perform the standard preprocessing for motion correction,

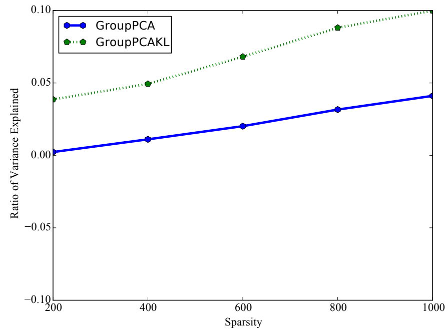

Figure 2: Group Sparse PCA performance on the Neurosynth data

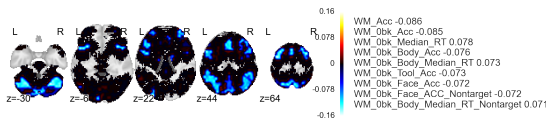

Figure 3: The first factor from 2K. TODO: Russ's comments here.

and image registration to the MNI template for consistency of comparisons across subjects. The resulting maps were downsampled in the similar way as the neurosynth data.

As before, to incorporate smoothness we use the spatial correlation matrix as the prior on the factors of view of statistical map. For the view of behavioral data, we use an identity matrix as the respective prior covariance matrix. We apply our Information Projection based Sparse CCA (SparseCCAKL) approach and compare it against the Sparse CCA algorithm developed by Witten et al. [2009]. We report the cross-variance explained which is defined as follows. If $\mathbf{X}, \mathbf{Y}$ are the two views, and $\mathbf{u}, \mathbf{v}$ are respective CCA (possible sparse) factors, the cross-variance is defined as : $\frac{\mathbf{u}^\top \mathbf{X}^\top \mathbf{Y} \mathbf{v}}{|\mathbf{u}^\top \mathbf{X} \mathbf{u}| \|\mathbf{v}^\top \mathbf{Y} \mathbf{v}\|}$. We show strong performance of SparseCCAKL on the metric in Figure 5.

6 Conclusion and Future Work

We presented a framework for designing sparsity inducing priors and showed its applicability on a wide variety of models and structures. By leveraging the advancements made in research on submodular combinatorial optimization, the framework allows for the flexibility of probabilistic modeling with varied structures while maintaining tractable inference by greedy strategies that are provably close to optimum. We also presented empirical evidence of strong performance compared to established baselines of respective models on simulated and two real world fMRI datasets for three special cases of our framework, namely Group Sparse Linear regression, Group Sparse PCA, and Sparse CCA. For future work, we wish to study qualitative interpretations of our results on the Neurosynth and the Human Connectonome fMRI datasets. Given our strong results, we plan to further study additional theoretical properties of the information projection framework including that of provable sparsistency and robustness.

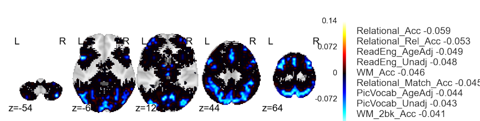

Figure 4: The first factor from REL. TODO: Russ's comments here.

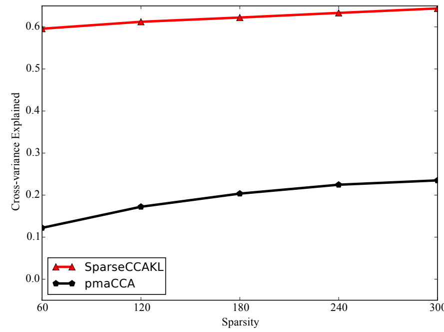

Figure 5: Sparse CCA performance on Human Connectonome Project data

References

Cédric Archambeau and Francis Bach. Sparse probabilistic projections. In *NIPS*, pages 73–80, 2008.

Cédric Archambeau and Francis R. Bach. Sparse probabilistic projections. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *NIPS*, pages 73–80. Curran Associates, Inc., 2009.

Francis R. Bach and Michael I. Jordan. A probabilistic interpretation of canonical correlation analysis. Technical report, UC Berkeley, 2005.

Gruia Calinescu, Chandra Chekuri, Martin Pál, and Jan Vondrák. Maximizing a monotone submodular function subject to a matroid constraint. *SIAM Journal on Computing*, 40(6):1740–1766, 2011.

David C. Van Essen, Stephen M. Smith, Deanna M. Barch, Timothy E.J. Behrens, Essa Yacoub, and Kamil Ugurbil. The wu-minn human connectome project: An overview. *NeuroImage*, 80:62 – 79, 2013. ISSN 1053-8119. Mapping the Connectome.

Chinmay Hegde, Piotr Indyk, and Ludwig Schmidt. A nearly-linear time framework for graph-structured sparsity. In Francis R. Bach and David M. Blei, editors, *ICML*, volume 37 of *JMLR Proceedings*, pages 928–937. JMLR.org, 2015.

R. Jenatton, G. Obozinski, and F. Bach. Structured sparse principal component analysis. In *AISTATS*, 2010.

Rajiv Khanna, Joydeep Ghosh, Russell A. Poldrack, and Oluwasanmi Koyejo. Sparse submodular probabilistic PCA. In *AISTATS*, 2015.

Arto Klami and Samuel Kaski. Local dependent components. In *Proceedings of the 24th International Conference on Machine Learning*, ICML '07, pages 425–432, New York, NY, USA, 2007. ACM.

Oluwasanmi Koyejo and Joydeep Ghosh. Constrained Bayesian inference for low rank multitask learning. *UAI*, 2013.

Oluwasanmi Koyejo, Rajiv Khanna, Joydeep Ghosh, and Poldrack Russell. On prior distributions and approximate inference for structured variables. In *NIPS*, 2014.

Jun Liu, Shuiwang Ji, and Jieping Ye. Slep: Sparse learning with efficient projections, 2010.

Vincent Michel, Alexandre Gramfort, Gaël Varoquaux, Evelyn Eger, Christine Keribin, and Bertrand Thirion. A supervised clustering approach for fmri-based inference of brain states. *Pattern Recognition*, 45(6):2041–2049, 2012.

Radford Neal and Geoffrey E. Hinton. A view of the EM algorithm that justifies incremental, sparse, and other variants. In *Learning in Graphical Models*, pages 355–368. Kluwer Academic Publishers, 1998.

George L Nemhauser, Laurence A Wolsey, and Marshall L Fisher. An analysis of approximations for maximizing submodular set functions—i. *Mathematical Programming*, 14(1):265–294, 1978.

Noah Simon, Jerome Friedman, Trevor Hastie, and Rob Tibshirani. A sparse-group lasso. *Journal of Computational and Graphical Statistics*, 2013.

Stephen M Smith, Peter T Fox, Karla L Miller, David C Glahn, P Mickle Fox, Clare E Mackay, Nicola Filippini, Kate E Watkins, Roberto Toro, Angela R Laird, et al. Correspondence of the brain's functional architecture during activation and rest. *Proceedings of the National Academy of Sciences*, 106(31):13040–13045, 2009.

Stephen M Smith, Thomas E Nichols, Diego Vidaurre, Anderson M Winkler, Timothy E J Behrens, Matthew F Glasser, Kamil Ugurbil, Deanna M Barch, David C Van Essen, and Karla L Miller. A positive-negative mode of population covariation links brain connectivity, demographics and behavior. *Nature NeuroScience*, pages 1565–1567, 2015.

Maxim Sviridenko. A note on maximizing a submodular set function subject to a knapsack constraint. *Operations Research Letters*, 2004.

Michael E Tipping and Christopher M Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):611–622, 1999.

Daniela M. Witten, Trevor Hastie, and Robert Tibshirani. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, 2009.