Information Projection and Approximate Inference for Structured Sparse Variables

Rajiv Khanna UT Austin
Joydeep Ghosh UT Austin
Russell Poldrack Stanford University
Oluwasanmi Koyejo UIUC

Abstract

Approximate inference via information projection has been recently introduced as a general-purpose technique for efficient probabilistic inference given sparse variables. This manuscript goes beyond classical sparsity by proposing efficient algorithms for approximate inference via information projection that are applicable to any structure on the set of variables that admits enumeration using matroid or knapsack constraints. Further, leveraging recent advances in submodular optimization, we provide an efficient greedy algorithm with strong optimization-theoretic guarantees. The class of probabilistic models that can be expressed in this way is quite broad and, as we show, includes group sparse regression, group sparse principal components analysis and sparse collective matrix factorization, among others. Empirical results on simulated data and high dimensional neuroimaging data highlight the superior performance of the information projection approach as compared to established baselines for a range of probabilistic models.

1 Introduction

Parsimonious Bayesian models are being increasingly used for improving both robustness and generalization performance in applications involve large amounts of data and variables. They are especially well-suited to incorporating domain knowledge by attuning the prior design to apriori knowledge and constraints at hand. For instance, sparsity constraints and associated models have gained eminence in several fields where apriori knowledge corresponding to these constraints may be incorporated via the use of sparsity inducing priors.

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

A natural extension to the classical notion of sparsity is *structured* sparsity – where the sparse selection of variable dimensions includes additional information. Some examples of structured sparsity 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] and so on. While there is a significant body of literature on classically sparse probabilistic models, including [Archambeau and Bach, 2009, Koyejo et al., 2014, Wipf and Nagarajan, 2007, Sheikh et al., 2012, Khanna et al., 2015] probabilistic models for structured sparsity have been far less studied. Our work seeks to bridge this gap for a large family of information projection based techniques.

The information projection of a distribution to a constraint set is given by the argument that minimizes the Kullback-Leibler (KL) divergence while satisfying the constraints. The use of information projection for probabilistic inference with structured variables was recently proposed by Koyejo et al. [2014]. While information projection is a general approach, its application requires the design of efficient algorithms that are specific to pre-stipulated structural constraints of interest. Koyejo et al. [2014] focused on the case of sparsity, where the constraint structure is given by the union of sparse supports. For this case, they proposed approximate inference by information projection to the support that captures the largest probability mass. They showed that the resulting KL minimization can be reduced to a combinatorial submodular optimization problem, and then applied a greedy algorithm for efficient approximate inference. Subsequently, a similar mechanism was developed for sparse principal components analysis (sparse PCA) [Khanna et al., 2015].

This manuscript goes beyond sparsity by proposing efficient algorithms for approximate inference via information projection that are applicable to any *structured sparse* variable settings which admits enumeration using a *matroid*. The class of probabilistic models that can be expressed in this way is quite broad, and as we show, includes group sparse regression, group sparse principal components analysis and sparse collective matrix factorization, among others. The generalized framework introduced in this paper is not a sim-

ple extension, but rather involves discovery of non-trivial connections between sparse probabilistic inference and discrete submodular optimization.

Specifically, our main contributions are as follows:

- we present a framework for approximate inference via information projection for any constraint that can be enumerated as a matroid.
- we present an efficient scheme for this inference using a greedy algorithm. For general matroids, an approximation solution of $1/2$ to the best possible approximation is guaranteed. However, for some special cases such as cardinality constraints (classical sparsity), and group sparsity, stronger guarantees of $1-1/e$ are available.
- we show that the special cases of information projection under group sparsity and multi-view sparsity are submodular with knapsack constraint and partition matroid constraints respectively. These constraints are applied to develop new algorithms for group sparse regression, sparse principal components analysis (PCA), and sparse collective matrix factorization (CMF).

Further, we present empirical results on simulated data and real high dimensional neuroimaging data that highlight the performance of the information projection approach as compared to established baselines for a range of probabilistic models.

2 Notation and Background

We begin by outlining some 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 $(\cdot)^\top$. 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 $[d]$.

Relative Entropy: Let X be a measurable set, and $\mathrm{p}(\cdot)$ be a probability density defined on X. Let $\mathrm{E}_{X\sim p}[f]$ is the expectation of the function f with respect to p. The relative entropy, or Kullback-Leibler (KL) divergence between the density q and p is given by $\mathrm{KL}(\mathrm{q}\|\mathrm{p}) = \mathrm{E}_q[\log q - \log p]$. The relative entropy is jointly convex in both arguments.

Information Projection: 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 $p(\cdot)$ onto a constraint (measurable) set $\mathsf{S} \subset \mathsf{X}$ is defined as:

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

$\mathcal{F}_\mathsf{S}$ is closed and bounded for all cases of interest in this manuscript, so that q_* exists.

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 of special interest because greedy algorithm and its simple variants achieve provable approximation guarantees 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, E), where N is the *ground set*, and $\mathsf{E} \subset \mathfrak{p}(\mathsf{N})$ is a family of *independent* sets that satisfies: (i) $\mathsf{B} \in \mathsf{E}, \mathsf{A} \subset \mathsf{B} \implies \mathsf{A} \in \mathsf{E}$, and, (ii) $\mathsf{A} \in \mathsf{E}, \mathsf{B} \in \mathsf{E}, |\mathsf{A}| < |\mathsf{B}| \implies \exists x \in \mathsf{B} - \mathsf{A}$ s.t. $\mathsf{A} \cup x \in \mathsf{E}$. A *uniform* matroid has E as the set of all possible k and lesser sized subsets of N, and thus induces the k-cardinality constraint. A *partition* matroid partitions N into subsets $\{\mathsf{X}_1, \mathsf{X}_2, \ldots, \mathsf{X}_r\}$, with $\mathsf{E} = \{\mathsf{A} \,|\, \mathsf{A} \subset \mathsf{N}, |\mathsf{A}\cap\mathsf{X}_i| \leq k_i \forall i \in [r]\}$ for given $\{k_1, k_2, \ldots, k_r\}$.

Knapsack: A knapsack constraint also imposes a combinatorial structure – each candidate solution in E as a set of possible groups, each with an associated cost, such that the total cost of each candidate solution in E is less than or equal to the knapsack value. A knapsack constraint in general is not a matroid, unless all maximal groupings are of the same size.

2.1 Information Projection for Sparse Variables

A d dimensional variable $\mathbf{x}$ is k-sparse if it is non-zero on at most 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 density p is k-sparse if all random variables $\mathbf{x} \sim \mathrm{p}$ are k-sparse. Let A be the set of all $\frac{d!}{k!(d-k)!}$ k-sparse support sets. The information projection of p onto $\mathcal{F}_\mathsf{A}$ is equivalent to restriction of p onto A [Koyejo et al., 2014], which is a natural approach for constructing a sparse prior. Unfortunately, this information projection is generally intractable. Instead, Koyejo et al. [2014] propose the following approximation:

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

This information projection searches for the subset S that captures most of the mass of p as measured by $\min_{\mathrm{q}\in\mathcal{F}_\mathsf{S}} \mathrm{KL}(\mathrm{q}\|\mathrm{p})$. The inner optimization over $\mathcal{F}_\mathsf{S}$ can be solved in closed form as $\min_{\mathrm{q}\in\mathcal{F}_\mathsf{S}} \mathrm{KL}(\mathrm{q}\|\mathrm{p}) = -\log \mathrm{p}(\mathbf{x}_{\mathsf{S}^c} = 0)$. 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 equivalent to

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

Note that the cardinality constraint is a uniform matroid constraint. While (2) is combinatorial, the following theorem ensures a good approximation by greedy support selection.

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

Thus, a simple greedy algorithm achieves a $(1 - \frac{1}{e})$ approximate solution [Nemhauser et al., 1978].

3 Approximate Inference via Information Projection for Structured Sparse Variables

In this section, we generalize the cardinality constrained information projection in two ways. First, we consider information projection subject to *group* sparsity, and show that the resulting combinatorial problem of selecting the *most relevant* groups is monotone submodular subject to a knapsack constraint. Secondly, we consider general matroid constraints for structured sparsity, and present an algorithm that greedily selects from the enumeration of the matroid constraint. We consider the special case of partition matroid constraint where sets of variables are pre-grouped into *views*, and seek to select variables subject to constraints on the maximum number of variables selected from each view. We leverage the research in submodular optimization to present the respective variants of the greedy algorithm that provably guarantee constant factor approximations.

Approximate Inference: Let p_0 be the prior distribution and l be the likelihood, and let A represent a structured subset of the domain of p_0. Restriction of the prior p_0 to the subset A given by $p_{\mathsf{A},0} \propto p_0(X)1_{[X\in\mathsf{A}]}$is an effective approach for constructing a prior for the structured subset A. Unfortunately, this restriction is generally intractable. We consider approximate inference by fixing p as the posterior distribution $p(X) \propto p_0(X)l(X)$. Let $\mathsf{A} = \{\mathsf{S}_i\}$ be the set of subsets satisfying the constraint structure. e.g. for classical sparse modeling with d variables, each $\mathsf{S}_i \in [d]$ is a k-sparse subset, and A is the set of all such $\binom{d}{k}$ subsets. In this case, the information projection to a subset $\mathsf{S} \in \mathsf{A}$ is designed to approximate Bayesian inference with respect to intractable restricted prior as $p_{\mathsf{A}}(X) \propto p_{\mathsf{A},0}(X)l(X)$. We note that just as in standard variational inference, the posterior projection can be implemented using p_0 and $l(\cdot)$ without explicitly computing the unrestricted posterior p (useful when the p is itself intractable). The proposed approach for approximate inference differs from standard approaches such as mean field variational inference, as we take advantage of the combinatorial nature of the desired structure. As such, the proposed approximate inference is most accurate when the posterior mass is well captured by the optimal subset $\mathsf{S}^* \in \mathsf{A}$. The approximate posterior is given by the information projection of p onto $\mathcal{F}_{\mathsf{S}^*}$, and is in fact equivalent to the projection of the restricted posterior p_{A} onto $\mathcal{F}_{\mathsf{S}^*}$. We refer the interested reader to [Koyejo et al., 2014] for additional details.

3.1 General Structured Sparsity Constraints

Structured sparsity extends classic sparsity constraints with additional information on the sparse subsets. 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, E) where N are the base set of dimensions, and E represents the set of all possible candidate solutions under the given constraint. General structured sparsity is challenging to model using standard prior design techniques. Instead, one may consider Bayesian inference using the structured prior distribution recovered by restricting the base prior to the union of all possible structured subsets. As in the classic sparsity case, we consider an approximation of the resulting posterior based on the variable set which captures the maximum posterior mass. The resulting (N, E)-matroid constrained information projection of a density p is simply given by:

$$\min_{\mathsf{S}\in\mathsf{E}} \min_{supp(\mathrm{q})\in\mathsf{S}} \mathrm{KL}(\mathrm{q}\|\mathrm{p}) \tag{3}$$

A simple greedy algorithm on the enumeration of the matroid as outlined in Algorithm 1 can be used for support selection under general matroid constraints. Note that the greedy selection algorithm for the classic sparsity case [Koyejo et al., 2014] is a special case of Algorithm 1 with a uniform matroid. For the more general matroid constraints, greedy selection on the enumeration admits slightly weaker guarantees. Improved approximation guarantees can be achieved by randomized algorithms [Calinescu et al., 2011].

Theorem 2 (Calinescu et al. [2011])**.** *Algorithm 1 guarantees a constant factor approximation of $1/2$ for* (3)*.*

Multi view sparsity. A special case of structured sparsity is the multi view sparsity. The base set of dimensions are divided into v views/groups. Also given is a set of maximum number of allowed selections from each view $\{k_1, k_2, \ldots, k_v\}$. In other words, no more than k_i selections can be made from the i^{th} view/group. It should be straightforward to see that the multi view sparsity constraint induces a partition matroid structure defined in Section 2, and as such Algorithm 1 is applicable. Algorithm 1 can be easily re-written for the partition sparsity constraint to avoid exhaustive enumeration of the set E as Algorithm 2. The $1/2$ factor approximation guarantee carries over for Algorithm 2. We shall see in the sequel that this particular

algorithm leads to an efficient inference algorithm for sparse probabilistic CMF.

Algorithm 1: GreedyMatroid(N, E)

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

Algorithm 2: GreedyMultiView($k_1, k_2, \ldots, k_v, m(\cdot)$)

Input : N, Sparsities $\{k_1, k_2, \ldots, k_v\}$, mapping function $m : [d] \rightarrow [v]$.
$\mathsf{A} \leftarrow \emptyset$
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

3.2 Group Sparsity Constraints

Group sparsity involves selecting variables from r groups subject to the constraint that if a group is selected, all the variables within the group must be selected, but no more than k variables can be selected in all. Let $\mathsf{G} = \{\mathsf{G}_1, \mathsf{G}_2, \ldots, \mathsf{G}_r\}$ represent the set of r groups, so that $\forall i, \mathsf{G}_i \subset [d]$ and $\forall i \neq j, \mathsf{G}_i \cap \mathsf{G}_j = \emptyset$. As in the classic sparse case, information projection to the set of all group sparse subsets of $[d]$ is intractable in general. Instead, we propose approximate inference by seeking the projection to the set which maximizes the captured mass of p. The resulting group sparsity constrained information projection of a density p is given by:

$$\min_{\mathsf{S} \subset [r]} \min_{\left\{\mathsf{q} \mid supp(\mathsf{q}) \subset \bigcup_{i \in \mathsf{S}} \mathsf{G}_i, \sum_{i \in \mathsf{S}} |\mathsf{G}_i| \leq k\right\}} \mathrm{KL}(\mathsf{q} \| \mathsf{p}) \quad (4)$$

Theorem 3. *The group selection problem* (4) *is equivalent to a normalized monotone submodular maximization problem with a knapsack constraint.*

The proof is provided in the supplement. We present a re-weighted greedy algorithm with partial enumeration in Algorithm 3 to solve (4). The re-weighting is ensures that the greedy step chooses 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. We also note that Algorithm 3 is *not* a special case of Algorithm 1, as it exploits the special structure of group sparsity to construct a scheme with improved optimization-theoretic guarantees. The following theorem establishes the optimization guarantee of Algorithm 3.

Theorem 4 (Sviridenko [2004])**.** *Algorithm 3 with* $b = 3$ *guarantees a constant factor approximation of* $(1 - \frac{1}{e})$ *for* (4).

Algorithm 3: GreedyPartialEnum ($\mathsf{G}, k, c(\cdot)$)

Input: Set of groups G, Total max sparsity k, parameter b, cost function $c(\cdot)$
$\mathsf{S}_1 \leftarrow \arg\max_{\mathsf{s} \subset \mathsf{G}, |\mathsf{s}| < b, c(\mathsf{s}) \leq k} \tilde{J}(\mathsf{s})$
$\mathsf{S}_2 \leftarrow \emptyset$
for all $\mathsf{s} \subset \mathsf{G}, |\mathsf{s}| = b, c(\mathsf{s}) \leq k$ **do**
  $\mathsf{S}_3 \leftarrow$ ReweightedGreedy(G, $k - b - 1$, $c(\cdot)$, 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 4: 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

3.3 Other constraints

The submodular framework allows for constraints more general than what are induced by matroids. Note that by definition matroids are restricted to have all maximal sets of the same size. e.g. classic sparsity constraint (uniform matroid) of support size being less than equal to k has maximal size to be k. The *p-system* constraints allow generalization to constraint sets which do not have the same maximal sizes. The integer p refers to the ratio of biggest and smallest

maximal sets. The p-systems guarantee a $\frac{1}{p+1}$ approximation [Calinescu et al., 2011]. Matroids are special cases of p-systems with $p = 1$. Thus, greedy selection can be used for other constraints, more complicated constraints, but with weaker approximation guarantees.

4 Applications: Probabilistic Models with Matroid Constrained Variables

While the class of probabilistic models that admit a representation via matroid constraints is quite broad, we consider special cases in detail (i) group sparse regression (ii) group sparse principal components analysis, (iii) sparse collective matrix factorization. The framework developed in Section 3 readily yields efficient greedy solutions for feature selection for all three cases.

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 $\mathbf{C} \in \mathbb{R}^{d\times d}$. The noise ϵ is drawn from a Gaussian $\epsilon \sim \mathcal{N}(0, \sigma^2)$. The posterior distribution of β is 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 $\mathsf{G} = \{\mathsf{G}_1, \mathsf{G}_2, \ldots, \mathsf{G}_r\}$ be the given set of groups so that $\forall i \in [r], \mathsf{G}_i \subset [d]$, and $\forall i \neq j, \mathsf{G}_i \cap \mathsf{G}_j = \emptyset$. The optimization problem for sparse group selection is then given by (4). For the spacial case where p is Gaussian, the information projection to any structured subset remains in the Gaussian family [Koyejo and Ghosh, 2013]. Thus, the search for q in (4) 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 that (4) for group sparse linear regression is equivalent to the submodular maximization problem:

$$\max_{\left\{\mathsf{S}\subset[r],\, \mathsf{S}=\bigcup_{i\in\mathsf{S}}\mathsf{G}_i,\, |\mathsf{S}|\leq k\right\}} \mathbf{r}_\mathsf{s}^\top[\boldsymbol{\Sigma}^{-1}]_\mathsf{s}\mathbf{r}_\mathsf{s} - \log\det[\boldsymbol{\Sigma}^{-1}]_\mathsf{s}. \quad (5)$$

Once the support s is selected, the respective approximate posterior q^* can be obtained as the respective conditional $\mathrm{q}^*(\mathbf{x}) = \mathrm{p}(\mathbf{x}|\mathbf{x}_{\mathsf{S}^c} = 0)$.

4.2 Group Sparse Probabilistic Principal Components Analysis

Probabilistic PCA aims to factorize a 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. For simplicity, we only consider the rank 1 case i.e. where $\mathbf{x}, \mathbf{w}$ are vectors. The general matrix case follows using standard deflation techniques for multiple factors [Khanna et al., 2017], or by a joint estimation procedure within using our framework (details are in the supplement). The generative model for the observed data matrix is $\mathbf{T} = \mathbf{x}\mathbf{w}^\top + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \sigma^2)$. We consider the case where the prior $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{C})$, and in addition, $\mathbf{w}$ is assumed be sparse. Let $\theta = \{\mathbf{x}, \sigma\}$ represent the set of deterministic parameters.

The underlying $\mathbf{x}, \mathbf{w}$ may be estimated by maximizing the log likelihood using Expectation Maximization (EM), which 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 Neal and Hinton [1998] 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 is a search over the parameter space, keeping the latent random variable $\mathbf{w}$ fixed. Similarly, the E-step is the search over the space of distribution q of the latent variables $\mathbf{w}$, keeping the parameters θ fixed

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

This view of the EM algorithm provides the flexibility to design algorithms with any E and M steps that monotonically increase $\mathscr{F}$. When the search space of q in the E-step is unconstrained, E-step outputs the posterior $\mathrm{p}(\mathbf{w}|\mathbf{T};\theta)$. Constraining the search space of q leads to a *variational E-step*. In this section, we consider a restriction which approximates the combinatorial space of group sparse distributions using the framework developed in Section 3.2.

We now derive the explicit equations to apply Algorithm 3. 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 $\mathbf{r} := \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$. Expanding the KL divergence for information projection from (4) yields that the support selection requires the following submodular maximization problem:

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

The resulting approximate posterior is given by the respective conditional $\mathrm{q}^*(\mathbf{w}) = \mathrm{p}(\mathbf{w}|\mathbf{w}_{\mathsf{S}^c} = 0)$ (c.f. [Khanna et al., 2015]). Detailed derivations are presented in the supplement for the more general case of multiple factors.

4.3 Sparse Probabilistic Collective Matrix Factorization (Sparse PCMF)

Collective Matrix Factorization [Singh and Gordon, 2008, Klami et al., 2013a] is a multiview generalization of PCA. It is typically used to learn joint low rank factorizations with shared entities. The model is closely related to CCA [Witten et al., 2009], and its probabilistic counterpart [Bach and Jordan, 2005, Archambeau and Bach, 2008]. The models

are often used interchangeably, though there is a subtle difference. In their probabilistic counterparts CCA assumes full covariance matrix across dimensions while the CMF makes a simpler assumption of an isotropic Gaussian as noise [Klami et al., 2013b]. Both the models are used for studying cross relational effects. We chose to model sparse CMF over sparse CCA to illustrate the application of our framework under another matroidal constraint, namely the partition matroid. We note that sparse probabilistic CCA is within the purview of our framework, albeit it requires a bit more complicated constraint set and algorithm.

We describe the setup for CMF next. Instead of observing single view as an $n \times d$ matrix, or a single *view*, multiple views of the same entities are observed. Hence, we observe n samples of dimensions $d_1, d_2, \ldots, d_v$ as matrices $\mathbf{T}_1, \mathbf{T}_2, \ldots, \mathbf{T}_v$ each of which are one of the v views of the observed. The generative model assumes an underlying parameter $\mathbf{x} \in \mathbb{R}^n$ shared among all the views, and the random variables $\{\mathbf{w}_i \in \mathbb{R}^{d_i}, \forall i \in [v]\}$. As in Section 4.2, we note that $\mathbf{x}, \{\mathbf{w}_i\}$ can be matrices in general. To emphasize the proposed greedy information projection, we focus on modeling for the top-1 component. The random variables are drawn from Gaussian distribution as $\mathbf{w}_i \sim \mathcal{N}(\underline{0}, \mathbf{C}_i) \forall i \in [v]$, and each of the view is generated as $\mathbf{T}_i = \mathbf{x}\mathbf{w}_i + \epsilon$, where the noise is $\epsilon \sim \mathcal{N}(0, \sigma^2)$ For our experiments, $\mathbf{C}_i$ is set from domain knowledge (see Section 5), and σ^2 allows for additional isotropic variation to capture residuals from the cross correlation. We wish to infer sparse $\mathbf{w}_i$ so that $\forall i \in [v], |\text{supp}(\mathbf{w}_i)| \leq k_i$ for the supplied k_i. The parameters are optimized using an EM algorithm. The 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.

We now map the sparse PCMF 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 column-wise. 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 PCMF 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 Section 4.2, the submodular maximization problem can be written as:

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

Hence, Algorithm 1 or equivalently Algorithm 2 can be used for sparse inference. We focus on sparse PCMF for this manuscript. However, it should be easy to the see that further extension to group sparse PCMF is straightforward by modifying the constraining partition matroid appropriately.

5 Experiments

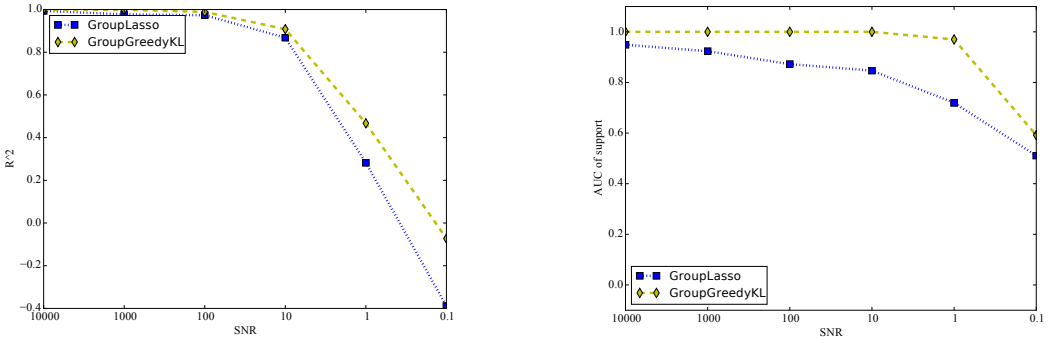
We now present empirical results comparing the proposed information projection based support selection technique to state of the art baselines. We begin by experiments on simulated data for group regression for model verification. We then present experiments for 2 applications for real world datasets, namely group sparse PCA, and sparse CMF. 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.

5.1 Experiment: Simulated data

We compare the proposed approach for group sparsity sparsity against the sparse-group lasso [Simon et al., 2013] implemented in the package SLEP [Liu et al., 2010] which is used in practice as state of the art. 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 separated 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. We use Bayes Factor to estimate k for GroupGreedyKL. For GroupLasso, we do a parameter sweep to get the best performing numbers. 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]



(a) R^2 performance on test data

(b) Recovery of true support

Figure 1: Group Sparse Regression performance on simulated data.

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 [Gorgolewski et al., 2015]. 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 $3mm^3$ voxels using the nilearn python package[1]. We then applied the standard brain mask, removing voxels outsize of the grey matter, resulting in d=65598 variables. We incorporate smoothness via spatial precision matrix $\mathbf{C}^{-1}$ on the prior on $\mathbf{W}$ which is generated by using the adjacency matrix of the three dimensional brain image voxels. This directly corresponds to the observation that nearby voxels tend to have similar functional behavior.

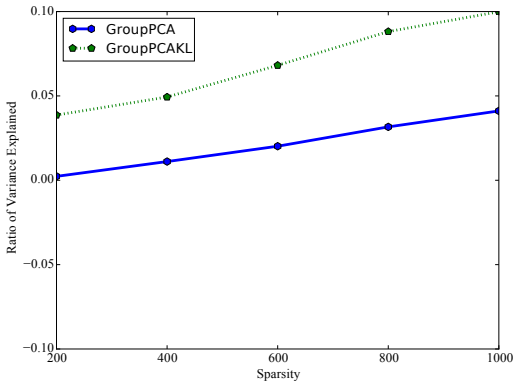
While our greedy algorithm can easily scale to dimensionality of size 65598, the matlab implementation of the baseline is not as scalable. We cluster the original set of dimensions to d =10000 dimensions using the spatially constrained Ward hierarchical clustering approach of Michel et al. [2012]. We 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], which is considered state of the art algorithm for group sparse PCA. 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 2a. The GroupPCAKL provides significant lifts (about times the variance explained) over groupPCA consistently across different sparsity levels.

Human Connectome 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 cross correlation based approaches (Section 4.3). We make use of the Human Connectome Project data (HCP) [Essen et al., 2013] for this purpose. It consists of large number of samples of high quality brain imaging and behavioral information collected from several healthy adults. We specifically use two datasets of different tasks - 2K (2 Back vs 0 Back contrast, measures working memory), and REL-match (REL vs MATCH contrast, measures relational processing)[2]. We download and extract brain statistical maps (a statistical map is a summary of each voxel in the brain in response to externally applied controlled stimulus) and respective behavioral variances from 497 adult subjects. Each subject has 380 behavioral variables, 27000 downsampled voxels. Further details on the task are available in the HCP documentation [Essen et al., 2013]. On the extracted maps, we perform the standard preprocessing for motion correction, and image registration to the MNI template for consistency of comparisons across subjects. The resulting maps we downsampled in the similar way as the Neurosynth data.

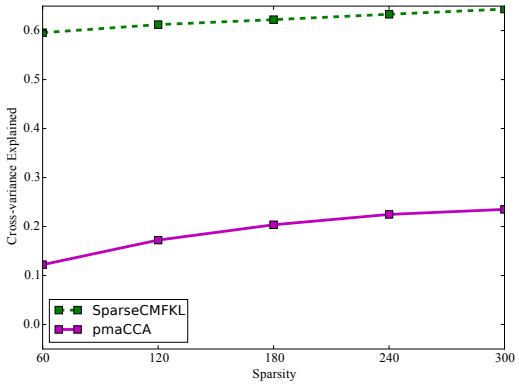
As before, to incorporate smoothness we use the spatial correlation matrix as the prior on the factors of view of statistical map. For the view of behavioral data, we use an identity matrix as the respective prior covariance matrix. We apply our Information Projection based Sparse CMF (SparseCMFKL) approach and compare it against the Sparse

[1]http://nilearn.github.io/

[2]https://wiki.humanconnectome.org/display/PublicData/Task+fMRI+Contrasts



(a) Group Sparse PCA performance on the Neurosynth data



(b) Cross-Correlation on n-back Human Connectome data

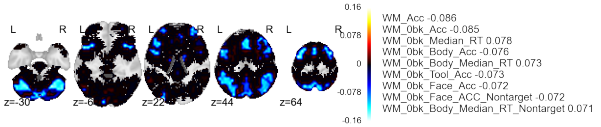


Figure 3: The first factor from 2-back task. Neural support is seen in a number of frontal and parietal regions and cerebellum, consistent with cognitive control systems usually engaged by the task. Behavioral correlates including both reaction time and accuracy on the task, showing greater neural engagement associated with slower and less accurate performance.

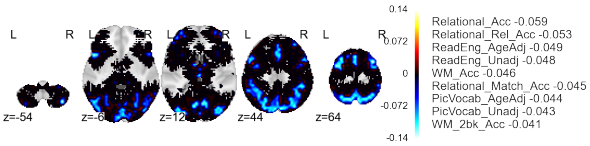


Figure 4: The first factor from relational reasoning task. Neural support is observed in frontal, parietal, and occipital cortex. Behavioral correlates captured both performance on this particular task, as well as independent measures related to higher cognitive functions including working memory capacity, vocabulary, and reading.

CCA algorithm developed by Witten et al. [2009] (pmdCCA) which is used in its original or slightly modified form as state of the art in many neuroscience and biomedical applications. For quantitative comparison, we use the n-back task dataset to 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 the respective (possible sparse) factors, the cross-variance is defined as : $\frac{\mathbf{u}^\top \mathbf{X}^\top \mathbf{Y}\mathbf{v}}{|\mathbf{u}^\top \mathbf{X}\mathbf{u}|^{1/2}|\mathbf{v}^\top \mathbf{Y}\mathbf{v}|^{1/2}}$. Note that the normalization ensures that the results are not driven by over estimating the within-view variance. We present strong performance of SparseCMFKL on the metric in Figure 2b. For lower sparsity (around 60-sparse) we obtain gains of the order of more than 4 times over pmdCCA. For higher sparsity levels, the order of the gap decreases a bit, but SparseCMFKL maintains a much stronger performance. We also show qualitative performance on the 2-back and relational task in Figure 3, 4 respectively. We note that applying pmdCCA on the same datasets yields inconsistent brainmaps.

6 Conclusion and Future Work

This manuscript proposes efficient algorithms for approximate inference via information projection that are applicable to any structure on the set of variables which admits enumeration using a matroid. The class of probabilistic models that can be expressed in this way is quite broad. In particular, we highlight the special cases of group sparse regression, group sparse principal components analysis and sparse canonical correlation analysis. We also presented empirical evidence of strong performance compared to established baselines of respective models on simulated and two real world fMRI datasets. Our strong results motivates us to further study the theoretical properties of the information projection framework, including sparsistency and robustness.

Acknowledgement

This work was supported by NSF grant IIS-1421729. fMRI data was provided by the Consortium for Neuropsychiatric Phenomics (NIH Roadmap for Medical Research grants UL1-DE019580, RL1MH083269, RL1DA024853, PL1MH083271).

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, *Advances in Neural Information*

Processing Systems 21, 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.

Krzysztof J. Gorgolewski, Gael Varoquaux, Gabriel Rivera, Yannick Schwarz, Satrajit S. Ghosh, Camille Maumet, Vanessa V. Sochat, Thomas E. Nichols, Russell A. Poldrack, Jean-Baptiste Poline, Tal Yarkoni, and Daniel S. Margulies. Neurovault.org: a web-based repository for collecting and sharing unthresholded statistical maps of the human brain. *Frontiers in Neuroinformatics*, 9:8, 2015. ISSN 1662-5196.

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.

Rajiv Khanna, Joydeep Ghosh, Russell A. Poldrack, and Oluwasanmi Koyejo. A deflation method for structured probabilistic PCA. In *SIAM International Conference on Data Mining (SDM)*, 2017.

Arto Klami, Guillaume Bouchard, and Abhishek Tripathi. Group-sparse embeddings in collective matrix factorization. *CoRR*, abs/1312.5921, 2013a.

Arto Klami, Seppo Virtanen, and Samuel Kaski. Bayesian canonical correlation analysis. *J. Mach. Learn. Res.*, 14 (1):965–1003, April 2013b. ISSN 1532-4435.

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.

A. Sheikh, J. A Shelton, and J. Lücke. A truncated variational EM approach for spike-and-slab sparse coding. *http://arxiv.org/abs/1211.3589*, 2012.

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

Ajit Paul Singh and Geoffrey J. Gordon. Relational learning via collective matrix factorization. In Ying Li, Bing Liu, and Sunita Sarawagi, editors, *KDD*, pages 650–658. ACM, 2008. ISBN 978-1-60558-193-4.

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.

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

David P Wipf and Srikantan S Nagarajan. A new view of automatic relevance determination. In *Advances in Neural Information Processing Systems*, pages 1625–1632, 2007.

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