Sparse Submodular Probabilistic PCA

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

Abstract

We propose a novel approach for sparse probabilistic principal component analysis, that combines a low rank representation for the latent factors and loadings with a novel sparse variational inference approach for estimating distributions of latent variables subject to sparse support constraints. Inference and parameter estimation for the resulting model is achieved via expectation maximization with a novel variational inference method for the E-step that induces sparsity. We show that this inference problem can be reduced to discrete optimal support selection. The discrete optimization is submodular, hence, greedy selection is guaranteed to achieve 1-1/e fraction of the optimal. Empirical studies indicate effectiveness of the proposed approach for the recovery of a parsimonious decomposition as compared to established baseline methods. We also evaluate our method against state-of-the-art methods on high dimensional fMRI data, and show that the method performs as well as or better than other methods.

1 Introduction

Principal component analysis (PCA) [2] is a standard technique for representing data using low dimensional variables given by factors and loadings. The factors represent the basis shared by all examples, and the loadings are computed so that each example can be described as a linear combination of the shared basis. The factors and loadings are estimated to maximize the explained data variance, or equivalently, to minimize the data reconstruction error with respect to the Frobenius norm. PCA is often applied for exploratory data analysis which requires interpreting the estimated factors. Unfortunately, direct interpretation of the PCA factors is challenging as the recovered variables are not identifiable, i.e. the solutions are invariant to in-plane rotations. As a remedy, several methods have been devised for estimating rotations that lead to interpretable factors [12, 1]. In addition to interpretability, there is accumulating evidence that certain natural phenomena may be described by (approximately) sparse bases - motivating the development of sparse decomposition techniques.

Given the vast expanse of data being generated in various forms, and comparatively slower increase in computation power, models that succinctly explain data are desirable. While there are now many well studied computational models for sparse PCA [10, 29, 6, 15], probabilistic approaches are less developed. There are important reasons further motivating development of probabilistic methods. Probabilistic models capture uncertainty in the data and random variables and can be used to estimate "confidence bars" with respect to the model via posterior covariance estimates. Further, a probabilistic approach is able to handle missing data via marginalization, incremental learning using EM, and offer the possibility of automated hyper-parameter tuning using hierarchical priors. These properties motivate our development of a probabilistic approach for sparse PCA.

In developing the new approach, we combine the probabilistic PCA [24], with an approach for estimating the distribution of latent variables with sparse support [13]. The proposed sparse probabilistic PCA is optimized using Expectation Maximization (EM). The expectation-step (E-step) is modified to capture the sparsity constraints and results in factors supported on a sparse domain. The inference is reduced to support selection - a submodular discrete optimization problem. Hence, greedy selection is guaranteed to achieve $(1 - \frac{1}{e})$ fraction of the optimal. We emphasize that while we use the proposed variational inference method for sparse PCA, it is more general and can be used in any EM algorithm requiring sparse inference

Appearing in Proceedings of the 18^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2015, San Diego, CA, USA. JMLR: W&CP volume 38.

for the latent variables. e.g. another application could be sparse topic coding, which is a modified form of topic modelling in which each document is generated using only a few topics.

Main contributions of this paper are (a) a novel method for sparse variational inference to recover sparse factors by exploiting submodularity structure of the cost function (b) a novel approach for sparse probabilistic PCA as a special case of the above and its efficient parameter estimation. (c) evaluation on high dimensional simulated data generated by factors with sparse support and high dimensional resting state neuroimaging data. Our results show that the proposed approach recovers a parsimonious decomposition and outperforms established baseline methods.

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 †. 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.

The rest of the paper is organized as follows. Section 2 defines some concepts that will be used throughout the paper, and discusses some related work in the area. Section 3 summarizes recent progress in constructing sparse prior using constrained information projection. Section 4 uses the theory summarized in Section 3 to develop our method for inference under sparsity constraints, and Section 5 applies the new method for sparse PCA. Finally, Section 6 describes the experiments conducted.

2 Background and Related Work

KL Divergence: Let X be a measurable set, and $\mathrm{P}(\cdot)$ and $\mathrm{Q}(\cdot)$ be two probability densities defined on X, then the the Kullback-Liebler distance between $\mathrm{P}(\cdot)$ and $\mathrm{Q}(\cdot)$ is defined as

$$\mathrm{KL}(\mathrm{Q}\|\mathrm{P}) = \int_{x\in\mathsf{X}} \mathrm{Q}(x) \log \frac{\mathrm{Q}(x)}{\mathrm{P}(x)} dx.$$

It has an information theoretic interpretation: it measures expected number of additional bits needed to encode samples from $\mathrm{P}(\cdot)$, when using the code from $\mathrm{Q}(\cdot)$ rather than $\mathrm{P}(\cdot)$. For this reason, if $\mathrm{Q}(\cdot)$ is supported on a set $\mathsf{S} \subset \mathsf{X}$, while $\mathrm{P}(\cdot)$ is supported on X, then $\mathrm{Q}(\cdot)$ that minimizes the KL divergence, with some abuse of standard notation, is the information projection of $\mathrm{P}(\cdot)$ onto set S.

Domain restriction: Let P be a probability density defined on a measurable set X, and let $\mathsf{S} \subset \mathsf{X}$, then P_S is the S-restriction of P: $\mathrm{P}_\mathsf{S}(x) = 0$ if $x \notin \mathsf{S}$, $\mathrm{P}_\mathsf{S}(x) = \frac{\mathrm{P}(x)}{\int_{s\in\mathsf{S}} \mathrm{P}(s)ds}$ if $x \in \mathsf{S}$.

Sparse PCA: Principal component analysis aims to factorize a given matrix $\mathbf{T}$ of size $n \times d$ into a *loading* matrix $\mathbf{X}$ of size $n \times r$ and a latent variable matrix $\mathbf{W}$ of size $r \times d$, where r is known as the rank, or the number of factors. The decomposition is given by:

$$\mathbf{T} \approx \mathbf{X}\mathbf{W},$$

so that the reconstruction error $||\mathbf{T} - \mathbf{X}\mathbf{W}||_2$ is minimized. While PCA is non-convex, it can be solved using the singular value decomposition of $\mathbf{T}$ to estimate the largest p singular values and their corresponding singular vectors. Sparse PCA [10, 29, 6, 15] generally involves enforcing constraints that ensure $\mathbf{W}$ is sparse i.e. the entries $\mathbf{W}$ have only a few non-zero entries. Another way of looking at sparse PCA is from an optimization viewpoint. For the principal factor w,

$$w^* = \arg\max_{||w||_2=1, ||w||_0=k} w^T \mathbf{T} w.$$

Subsequent factors are learnt iteratively by projecting T onto orthogonal complement of w^*. The former approach is more suitable for developing the probabilistic PCA framework.

Sparse PCA has gained a lot of attention and there has been substantial amount of work in this area. The initial approaches used heuristics such as rotation[9], or thresholding [5]. Guan and Dy [8] proposed a sparse PCA model using the the low rank factorized representation. A Laplace prior was applied to these latent factors to encourage sparsity, based on the analogy to the l_1 norm that has been applied to much success in the computational sparse PCA literature. Similarly, Zou et al. [29] and Jolliffe et al. [10] also applied Lasso based techniques for obtaining sparse factors. In contrast, Archambeau and Bach [4] applied a variation of the relevance vector machine [23] and more sophisticated inverse Gamma priors to encourage sparsity in the latent factors. While not directly posed as probabilistic method, Sigg and Buhmann [22] modified the Expectation Maximization approach of probabilistic PCA to encourage sparsity or non-negativity in the recovered factors. Thus, it is possible to interpret the approach of [22] as a probabilistic model under an appropriate prior. [6] use a semidefinite relaxation to

compute a full regularized path using a greedy approach. [28] apply truncation to the power method to obtain sparsity in the factors. [21] use message passing to solve Bayesian PCA with two zero norm priors - spike-and-slab and microcanonical prior. Papailiopoulos et al. [19] introduce sparsity by eliminating vectors in low rank approximation of the given vectors. Sparse PCA models have been applied in a variety of applications, often with optimization guarantees. For instance, d'Aspremont et al. [6] applied sparse principal components to the task of gene ranking, while [15] used sparse coding techniques in image processing for in-painting.

3 Priors for Sparsity Constrained Variables

Constrained information projection is fundamental to our approach of introducing sparsity constraints. To motivate our choice of method of inducing sparsity in probabilistic PCA, we summarize some results from recently published work by Koyejo et al. [13] on the construction of sparse priors using information projection. They show that restricting the domain of a probability distribution from an ambient space to its subset is equivalent to information projection of the original distribution onto the constrained set of distributions defined on the subset.

Let P and Q represent probability densities. Throughout this paper, we assume that the base measure is supported over the entire domain set. If that is not the case, we simply redefine the domain as its support set. Moreover, the densities are such that the KL divergence is bounded. So the results presented here may not hold for degenerate distributions. We begin by characterizing information projection under equality constraints of expectation.

Lemma 1 (Altun and Smola [3])**.**

$$\text{[Primal]} \quad \min_{\mathrm{Q}} \mathrm{KL}(\mathrm{Q}\|\mathrm{P}) \text{ s.t. } \mathbb{E}_{\mathrm{Q}}\left[\boldsymbol{\beta}(x)\right] = \mathbf{c}$$

$$\text{[Dual]} \max_{\boldsymbol{\lambda}} \langle \boldsymbol{\lambda}, \mathbf{c} \rangle - \log \int_{x \in \mathsf{X}} \mathrm{P}(x) e^{\langle \boldsymbol{\lambda}, \boldsymbol{\beta}(x) \rangle} dx$$

and the unique solution is given by $\mathrm{Q}_*(x) = p(x)e^{\langle \boldsymbol{\lambda}_*, \boldsymbol{\beta}(x)\rangle - G(\boldsymbol{\lambda}_*)}$ *where* $\boldsymbol{\lambda}_*$ *is the dual solution and* $G(\boldsymbol{\lambda}_*)$ *ensures normalization.*

Let $\phi_{\mathsf{S}}(x) : \mathsf{X} \to \mathbb{R}$ be a function defined as $\phi_{\mathsf{S}}(x) = 0$ if $x \in \mathsf{S}$, and is strictly positive otherwise. Lemma 2 uses Lemma 1 to show the equivalence of domain restriction and an expectation constraint on ϕ.

Lemma 2. *Let* $\mathcal{P}_{\mathsf{X}}$ *be the set of all probability distributions defined on a measurable set* X*. Similarly,* $\mathcal{P}_{\mathsf{S}}$ *be the set of all probability distributions defined on* $\mathsf{S} \subset \mathsf{X}$*. Further, let* $\mathcal{P}_{\phi_{\mathsf{S}}} = \{\mathrm{P} \in \mathcal{P}_{\mathsf{X}} | \mathbb{E}_{\mathrm{P}}\left[\phi_{\mathsf{S}}\right] = 0\}$*. Then* $\mathcal{P}_{\phi_{\mathsf{S}}} = \mathcal{P}_{\mathsf{S}}$*.*

Proof. Follows from non-negativity of ϕ. □

We can now present the following result, that characterizes the relationship between restriction of densities and information projection subject to domain constraints.

Theorem 3. *Let* $\mathsf{X} \supset \mathsf{S}$ *be a measurable set. The information projection of a distribution* $\mathrm{P} \in \mathcal{P}_{\mathsf{X}}$ *to the constraint set* $\mathcal{P}_{\mathsf{S}}$ *is the restriction of P to the domain* S*.*

Proof. The information projection of P to $\mathcal{P}_{\mathsf{S}}$ is given by (using Lemma 1 and Lemma 2) $\mathrm{Q}_*(x) = \mathrm{P}(x)e^{\langle \lambda_*, \phi_{\mathsf{S}}(x)\rangle - G(\lambda_*)}$, where:

$$\lambda_* = \arg\min_{\lambda} \log \int_{x \in \mathbb{X}} \mathrm{P}(x) e^{\langle \lambda, \phi_{\mathsf{S}}(x)\rangle} dx,$$

which gives $\lambda_* = -\infty$, and thus $e^{\langle \lambda_*, \phi_{\mathsf{S}}(x)\rangle} \to \delta_{\mathsf{S}}(x)$, so $\mathrm{Q}_* = \mathrm{P}(x)\delta_{\mathsf{S}}(x)/\int_{\mathsf{S}} \mathrm{P}(x)dx$, where $\delta_{\mathsf{S}}(x)$ is the Dirac delta function. □

Thus, the information projection of a distribution P to the support constraints S is the conditional density that assigns 0 measure to S^c. Moreover, the KL distance between P and the projected density can be quantified. From Theorem 3, we find that $\mathrm{P}_{\mathsf{S}}(x) = \mathrm{P}(x)\delta_{\mathsf{S}}(x)/Z$, where Z is the normalization factor:

$$\begin{aligned} Z &= \int_{\mathsf{S}} \mathrm{P}(\boldsymbol{x})dx = \int_{\mathsf{X}} \mathrm{P}(\boldsymbol{x}_{\mathsf{S}}, \boldsymbol{x}_{\mathsf{S}^c})\delta_{\mathsf{S}}(\boldsymbol{x})dx \\ &= \int_{\mathsf{X}} \mathrm{P}(\boldsymbol{x}_{\mathsf{S}}|\boldsymbol{x}_{\mathsf{S}^c})\mathrm{P}(\boldsymbol{x}_{\mathsf{S}^c})\delta_{\mathsf{S}^c}(\boldsymbol{x})dx \\ &= \mathrm{P}(\boldsymbol{x}_{\mathsf{S}^c} = 0_{\mathsf{S}^c}) \end{aligned}$$

With this result, we may now compute the restriction explicitly:

$$\begin{aligned} \mathrm{P}_{\mathsf{S}}(\boldsymbol{x}) &= \mathrm{P}(\boldsymbol{x}_{\mathsf{S}}|\boldsymbol{x}_{\mathsf{S}^c})\mathrm{P}(\boldsymbol{x}_{\mathsf{S}^c})\delta_{\mathsf{S}}(\boldsymbol{x})/\mathrm{P}(\boldsymbol{x}_{\mathsf{S}^c} = 0_{\mathsf{S}^c}) \\ &= \mathrm{P}(\boldsymbol{x}_{\mathsf{S}}|\boldsymbol{x}_{\mathsf{S}^c} = 0_{\mathsf{S}^c})\delta_{\mathsf{S}}(\boldsymbol{x}). \end{aligned} \tag{1}$$

In other words, the information projection to a sparse support domain is the conditional distribution of $\boldsymbol{x} \in \mathsf{S}$ at $\boldsymbol{x}_{\mathsf{A}^c} = 0_{\mathsf{S}^c}$. The resulting gap is:

$$\begin{aligned} \mathrm{KL}(\mathrm{P}_{\mathsf{S}}\|\mathrm{P}) &= \int_{\mathsf{S}} \mathrm{P}_{\mathsf{S}}(\boldsymbol{x}) \log \frac{\mathrm{P}_{\mathsf{S}}(\boldsymbol{x})}{\mathrm{P}(\boldsymbol{x})} d\boldsymbol{x} \\ &= \int_{\mathsf{S}} \mathrm{P}_{\mathsf{S}}(\boldsymbol{x}) \log \frac{\mathrm{P}(\boldsymbol{x})}{\mathrm{P}(\boldsymbol{x})\mathrm{P}(\boldsymbol{x}_{\mathsf{S}^c} = 0_{\mathsf{S}^c})} d\boldsymbol{x} \\ &= -\log \mathrm{P}(\boldsymbol{x}_{\mathsf{S}^c} = \mathbf{0}_{\mathsf{S}^c}). \end{aligned} \tag{2}$$

Theorem 4. *[13] For a given support set* s*, define* $J(\mathsf{s})=\log \mathrm{P}(\boldsymbol{x}_{\mathsf{s}^c} = \mathbf{0}_{\mathsf{s}^c})$*.* $J(\mathsf{s})$ *is submodular.*

Proof sketch. Monotone: Let $\mathsf{c} \subset \mathsf{s}$, then:

$$\mathrm{P}(\boldsymbol{x}_{\mathsf{M}\setminus\mathsf{c}}) = \mathrm{P}(\boldsymbol{x}_{\mathsf{M}\setminus\mathsf{s}}, \boldsymbol{x}_{\mathsf{M}\setminus\mathsf{s}\cap\mathsf{c}}) \leq \mathrm{P}(\boldsymbol{x}_{\mathsf{M}\setminus\mathsf{s}})$$

Submodular: Consider $F(\mathsf{s}) = \log \mathrm{P}(\boldsymbol{x}_{\mathsf{s}} = \boldsymbol{c}_{\mathsf{s}})$, for a constant c_{s}. $F(\mathsf{s})$ is bounded above and below, and (assuming $\mathrm{P}(\cdot)$ has support on the set s) $F(\mathsf{s})$ does not take value $-\infty$. Since it can be written as $F(\mathsf{s}) = -\mathrm{KL}(\delta_{\mathsf{s}} \| p_{\mathsf{s}})$[27], so it is submodular [14]. Finally, we note that if $F(\mathsf{s})$ is submodular, so is its reflection $J(\mathsf{s}) = F(\mathsf{m}\setminus\mathsf{s})$. □

While the above results are valid for more general settings, we are particularly interested in the special case of sparsity as domain constraints.

A natural way of thinking about introducing sparsity in a distribution is restricting the support of the distribution. In this section, by showing equivalence of domain restriction and information projection, we also get a theoretical justification for the same.

4 Inference with Sparse Constraints

In this Section, we illustrate how priors constructed by restricting domain to sparse supports can be incorporated in practical algorithms for sparse inference.

Expectation Maximization can be described using the free energy interpretation [17]. Maximizing the negative log-likelihood can be shown to be equivalent to maximizing a free energy function $\mathscr{F}$ (see Equation 3). The E-step can be viewed as the search over the space of distributions Q(.) of the latent variables $\mathbf{W}$, keeping the parameters Θ fixed (Equation 4), and the M-step can be interpreted to be the search over the parameter space, keeping the latent variables $\mathbf{W}$ fixed (Equation 5). Let KL(.‖.) be the KL-divergence and $\mathbf{T}$ be the observed data, then the cost function for the EM is given by ([17]):

$$\mathscr{F}(\mathrm{Q}(\mathbf{W}), \Theta) = -\mathrm{KL}(\mathrm{Q}(\mathbf{W}) \| \mathrm{P}(\mathbf{W}|\mathbf{T}; \Theta)) + \log \mathrm{P}(\mathbf{T}; \Theta). \tag{3}$$

$$\text{E-step: } \max_{\mathrm{Q}} \mathscr{F}(\mathrm{Q}(\mathbf{W}), \Theta), \tag{4}$$

$$\text{M-step: } \max_{\Theta} \mathscr{F}(\mathrm{Q}(\mathbf{W}), \Theta). \tag{5}$$

This view of the EM algorithm provides the flexibility to design algorithms with any E and M steps that monotonically increase $\mathscr{F}$.

4.1 Variational E-step

An unconstrained optimization over Q in Equation 4 returns the posterior $\mathrm{P}(\mathbf{W}|\mathbf{T}; \Theta)$. Variational methods perform the search for best Q over a constrained set [26]. Let D be the set of distributions over $\mathbf{W}$ that fully factorize over individual rows of $\mathbf{W}$: $\mathrm{Q}(\mathbf{W}) = \prod_{i=1}^{r} \mathrm{Q}(\mathbf{W}_{i,\cdot})$. We restrict the search over Q to D. As a result of this restriction, we can optimize $\mathrm{Q}(W_{i,\cdot})$ for one i at a time in a co-ordinate descent fashion.

For introducing sparsity, we impose an additional constraint that $\forall i \in [r], \mathrm{Q}(\mathbf{W}_{i,\cdot})$ is k_i-sparse i.e. it has support only on at most k_i out of the ambient d dimensions. Let K_i be the set of all k_i-sparse supports. From Equations 4 and 3, it follows that the variational E-step is minimization of KL divergence over the sets K_i. As shown in Section 3, information projection to a set is equivalent to restricting domain to the respective set. So, minimizing the KL-divergence can be thought as searching for the sparse support set that loses the least amount of information by restricting the domain set of distributions to sparse sets. We get an optimization over the constrained space of distributions:

$$\min_{\mathsf{K}_1} \dots \min_{\mathsf{K}_r} \min_{\substack{\mathrm{Q}(\mathbf{W})\in\mathsf{D} \\ \forall i, \mathrm{Supp}(\mathrm{Q}(\mathbf{W}_{i,\cdot}))\in\mathsf{K}_i}} \mathrm{KL}(\mathrm{Q}(\mathbf{W}) \| \mathrm{P}(\mathbf{W}|\mathbf{T}; \Theta)). \tag{6}$$

For Gaussian P, Equation 6 can be re-written as an iterated information projection. For each row $i \in [r]$,

$$\min_{\mathsf{K}_i} \min_{\mathrm{Supp}(\mathrm{Q}(\mathbf{W}_{i,\cdot}))\in\mathsf{K}_i} \mathrm{KL}\Big(\mathrm{Q}(\mathbf{W}_{i,\cdot}) \| \hat{P}(\mathbf{W}_{i,\cdot})\Big), \tag{7}$$

where, with $\mathbf{W}_{\setminus i,\cdot}$ representing all rows of $\mathbf{W}$ except i, $\hat{P}_i$ depends on $\mathrm{Q}(\mathbf{W}_{\setminus i,\cdot})$ and $\log \mathrm{P}(\mathbf{W}|\mathbf{T}; \Theta)$.

Thus, the independence assumption on Q(.) allows for optimizing over an i, while holding the others fixed in a co-ordinate descent algorithm. Each information projection monotonically decreases the free energy function.

Recall that the KL-gap for the constraining support was characterized in Section 3. Using Equation 2, we can simplify Equation 7 for each i as

$$\text{For each row } i \in [r], \ \max_{\mathsf{K}_i} \log(\hat{P}_i([\mathbf{W}_{i,\cdot}]_{\mathsf{K}_i^c} = \mathbf{0}_{\mathsf{K}_i^c})) \tag{8}$$

Equation 8 is the resulting discrete optimization problem to be solved for variational E-step for each i. By Theorem 4, optimization problem over the set of dimensions is submodular, and hence instead of exhaustive search, each of the k_i dimensions can be selected

by a greedy algorithm which achieves at least a constant fraction $\left(1-\frac{1}{e}\right)$ of the objective value obtained by the optimal solution [18]. Moreover, no polynomial time algorithm can provide a better approximation guarantee unless P = NP [7].

To summarize, under the variational assumptions, the E-step can be solved iteratively over each i, and each optimization over i is a submodular discrete optimization problem with guarantees for using a greedy selection strategy. Moreover, since optimization over each i monotonically increases (or does not change) $\mathscr{F}$, updating the latent variable even for a single i suffices. As we shall see, this is particularly helpful for PCA.

5 Probabilistic PCA with Sparse Priors

We consider n observations of data vector valued variables in d dimensional ambient space, which are stacked in a matrix $\mathbf{T} \in \mathbb{R}^{n\times d}$. Drawing inspiration from traditional PCA, we seek a few sparse basis vectors whose linear combination generates the observation matrix with small error. The observation matrix is modelled as a product of a parameter $\mathbf{X} \in \mathbb{R}^{n\times r}$ and a *sparse* $\mathbf{W} \in \mathbb{R}^{r\times d}$. The sparse basis vectors are stacked as rows of $\mathbf{W}$, and their linear combination is modeled by $\mathbf{X}$. In cases which have $n \gg d$, the above factorization is useful for small r, which is set according to the domain. $\boldsymbol{\mu}$ is the matrix of column means generated as, $\boldsymbol{\mu} = \text{columnMeans}(\mathbf{T})^{\dagger} \otimes \mathbf{1}$, and Gaussian noise is represented by $\boldsymbol{\epsilon}_{ij} \sim \text{N}(0, \sigma^2), \forall i \in [n], \forall j \in [d]$.

The observation model is represented as:

$$\mathbf{T} = \mathbf{X}\mathbf{W} + \boldsymbol{\mu} + \boldsymbol{\epsilon}.$$

We will use a normal prior for each row of $\mathbf{W}$ i.e. $\mathbf{W}_{i,\cdot} \sim \text{N}(\mathbf{0}, \mathbf{C})\ \forall i \in [r]$, given a prior covariance matrix $\mathbf{C}$, while the rows are independent. The joint distribution can also be written as the matrix-variate normal $\mathbf{W} \sim \text{MVN}(0, \mathbf{C}, \mathbf{I})$. In the proposed model, the parameters are given by $\boldsymbol{\Theta} = \{\mathbf{X}, \sigma^2\}$, while $\mathbf{W}$ are the latent variables. Inference and learning can be performed using the EM algorithm applied on the log-likelihood $\log \text{P}(\mathbf{T}; \mathbf{X}, \sigma^2)$. As we shall see, the application of structural constraints of sparsity on the factors leads to a variational E-step.

Note that we can write the PCA equation as $\mathbf{T} = \sum_i \mathbf{X}_{\cdot,i}\mathbf{W}_{i,\cdot}$. Because of variational assumptions, the E-step can be solved by updating $\mathbf{W}_{i,\cdot}$ for a single i while keeping the others fixed. The M-step inherently has the same property for any individual $\mathbf{X}_{\cdot,i}$.

5.1 E-Step

While the guarantees for Equation 8 for inference over sparse support hold for general distributions, for sparse probabilistic PCA, we apply it for Gaussian P and Q.

With $\mathbf{Z}_i = \mathbf{T} - \sum_{j\neq i} \mathbf{X}_{\cdot,j}\mathbb{E}\left[\mathbf{W}_{j,\cdot}\right]$, using some basic algebra and properties of the Gaussian distribution, we can re-write Equation 8 for the sparse PCA setup as:

For each row $i \in [r]$,

$$\max_{\mathsf{K}_i} \log(\text{P}([\mathbf{W}_{i,\cdot}]_{\mathsf{K}_i^c} = \mathbf{0}_{\mathsf{K}_i^c} | \mathbf{Z}_i; \mathbf{X}, \sigma^2)) \tag{9}$$

with,

$$\text{P}(\mathbf{W}_{i,\cdot}|\mathbf{Z}_i; \mathbf{X}_{\cdot,i}, \sigma^2) \sim \mathcal{N}(\boldsymbol{m}^i, \boldsymbol{\Sigma}^i)$$

where,

$$[\boldsymbol{\Sigma}^i]^{-1} = \frac{1}{\sigma^2}(\mathbf{X}_{\cdot,i}^{\dagger}\mathbf{X}_{\cdot,i}) + \mathbf{C}^{-1},$$
$$\mathbf{m}^i = \boldsymbol{\Sigma}^i\ \frac{1}{\sigma^2}(\mathbf{X}_{\cdot,i}^{\dagger})(\mathbf{Z}_i).$$

We can now expand Equation 9 (for a given i):

$$\max_{\mathsf{K}_i}\ \mathbf{m}^i_{\mathsf{K}_i^c}{}^{\dagger}[\boldsymbol{\Sigma}^i_{\mathsf{K}_i^c}]^{-1}\mathbf{m}^i_{\mathsf{K}_i^c} - \log\det \boldsymbol{\Sigma}^i_{\mathsf{K}_i^c}$$

which can now be solved as a combinatorial problem to obtain the optimal support set. However, since the complement set is usually much larger than the intended sparse support set, it is more convenient to transform the objective from being in form of K_i^c to one with K_i. It turns out this is easy, again, because of standard properties of the Gaussian distribution. The equivalent objective function with K_i as the optimization variable is:

$$\max_{\mathsf{K}_i}\ \mathbf{r}^i_{\mathsf{K}_i}{}^{\dagger}[[\boldsymbol{\Sigma}^{i^{-1}}]_{\mathsf{K}_i}]^{-1}\mathbf{r}^i_{\mathsf{K}_i} - \log\det[\boldsymbol{\Sigma}^{i^{-1}}]_{\mathsf{K}_i}, \quad \text{where } \mathbf{r}^i = \boldsymbol{\Sigma}^{i^{-1}}\mathbf{m}^i \tag{10}$$

Recall that $\boldsymbol{\Sigma}_{\mathsf{K}}$ is the submatrix of $\boldsymbol{\Sigma}$ supported on K, similarly for $[\boldsymbol{\Sigma}^{-1}]_{\mathsf{K}}$. After solving the constrained optimization problem specified by Equation 10 by a greedy selection for K_i^*, the resulting solution density, q_i^*, known to be the conditional by properties of the Gaussian, is given by:

$$q_i^* \sim \mathcal{N}(\mathbf{c}^i, \mathbf{D}^i)$$

where,

$$[\mathbf{D}^i]^{-1} = [\boldsymbol{\Sigma}^{i^{-1}}]_{\mathsf{K}_i^*}, \quad \mathbf{c}^i = \mathbf{D}^i\mathbf{r}^i_{\mathsf{K}_i^*} \tag{11}$$

Recall, q_i^* has support only on K_i^*, so in Equation 11, $\mathbf{c}^i \in \mathbb{R}^{|\mathsf{K}_i^*|}, \mathbf{D}^i \in \mathbb{R}^{|\mathsf{K}_i^*|\times|\mathsf{K}_i^*|}$.

5.2 M-step

Since the free energy view of the EM shows that any M-step that increases $\mathscr{F}$ suffices, we maximize the log likelihood portion of $\mathscr{F}$ for the M-step. It turns out solving for $\{\mathbf{X}, \sigma^2\}$ over $\mathscr{F}$ directly is computationally hard, so M-step is done for one column of $\mathbf{X}$ at a time, corresponding to row-wise E-step. If q^* is the distribution on $\mathbf{W}$ obtained from the E-step, the effective M-step for column i of $\mathbf{X}$ is:

$$\max_{\mathbf{X}, \sigma^2} \mathbb{E}_{q^*} \left[\log \mathrm{P}(\mathbf{Z}_i | \mathbf{W}_{i,\cdot}; \mathbf{X}_{\cdot,i}, \sigma^2)\right] \tag{12}$$

For, any particular $i \in [d]$, let $\widehat{\mathbf{c}^i}$ represent the mean vector $\mathbf{c}^i$ expanded from $|\mathsf{K}_i^*|$ to ambient dimension d, with zeroes padded as needed. Equation 12 can be written as:

$$\begin{aligned}
&\max_{\mathbf{X}_{\cdot,i}, \sigma^2} \mathbb{E}_{q^*}[\frac{-1}{2\sigma^2}(\mathbf{Z}_i - \mathbf{X}_{\cdot,i}\mathbf{W}_{i,\cdot})^\dagger(\mathbf{Z}_i - \mathbf{X}_{\cdot,i}\mathbf{W}_{i,\cdot}) \\
&\qquad -nd\log\sigma^2] \\
\equiv &\max_{\mathbf{X}_{\cdot,i}, \sigma^2} \frac{-1}{2\sigma^2}\mathcal{V}(\mathbf{X}_{\cdot,i}) - nd\log\sigma^2, \\
&\text{where,} \\
&\mathcal{V}(\mathbf{X}_{\cdot,i}) = \mathbf{X}_{\cdot,i}^\dagger \mathbf{X}_{\cdot,i} \mathrm{tr}\Big((\mathbf{c}^i\mathbf{c}^{i\dagger} + \mathbf{D}^i)\Big) - 2\widehat{\mathbf{c}^i}^\dagger(\mathbf{X}_{\cdot,i}^\dagger \mathbf{Z}_i)
\end{aligned}$$

Clearly , $\mathbf{X}$ and σ^2 can be updated separately, and in closed form by taking the gradient and setting to 0.

$$\mathbf{X}^*_{\cdot,i} = \frac{\widehat{\mathbf{c}^i}^\dagger \mathbf{Z}_i}{\mathrm{tr}\Big((\mathbf{c}^i\mathbf{c}^{i\dagger} + \mathbf{D}^i)\Big)} \tag{13}$$

$$\sigma^{*2} = \frac{\mathcal{V}(\mathbf{X}_{\cdot,i})}{2nd} \tag{14}$$

Algorithm 1 delineates the entire algorithm stepwise.

Algorithm 1: EM Algorithm for SparsePCA

Input: $k, r, \mathbf{C}, \mathbf{T}$
Initialize $\forall j \neq 1, \mathbf{X}_{\cdot,j} = 0$, $\mathbf{X}_{\cdot,1}$ randomly

while not converged **do**
  for $i = 1 \ldots r$ **do**
    $\mathbf{Z}_i = \mathbf{T} - \sum_{j \neq i} \mathbf{X}_{\cdot,j}\mathbf{c}^j$
    E-Step
    Init: $\mathsf{K}_i^* = \{\}$
    for $j = 1 \ldots k$ **do**
      Update K_i^*:
      $\mathsf{K}_i^* = \arg\max$ Eq. 10 over $\mathsf{K}_i^* \bigcup \{t\}$, $\forall t \in [d], t \notin \mathsf{K}_i^*$
    end for
    Use Equation 11 to update $\mathbf{c}^i$ and $\mathbf{D}^i$ for q_i^*
    M-Step
    Update $\mathbf{X}_{\cdot,i}$ using Equation 13
    Update σ^2 using Equation 14
  end for
end while
return$(\mathbf{q}^*, \mathbf{X}, \sigma^2)$

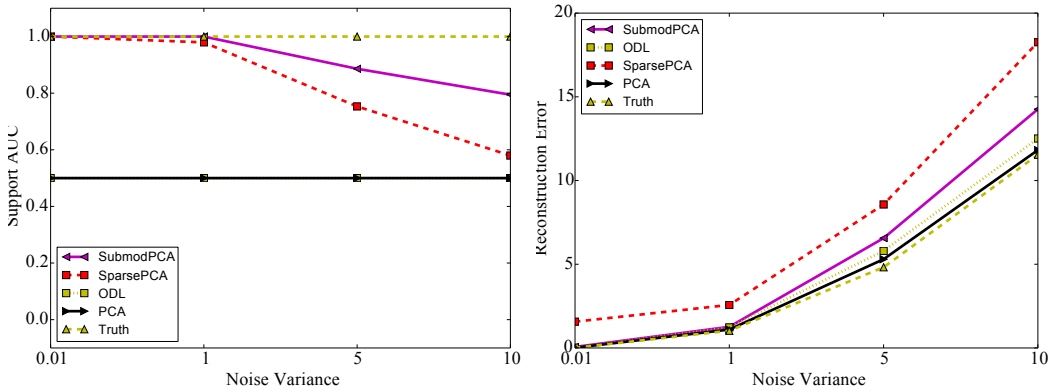
6 Experiments

We implement our method in Python using Numpy and Scipy libraries as required. The inference is efficiently implemented by building the variance of the sparsely supported posterior incrementally using block matrix inversion formula while employing the greedy search. This helps us avoid taking explicit inverses that can lead to numeric inconsistencies. We compare performance of submodular sparse probabilistic PCA with other state of the art methods that are used in practice to obtain lower dimensional representations of the data. We also make use of the fact that greedy search in the E-step is trivially parallelizable - since computations can be performed in parallel over all candidate dimensions, only to compare the final objective value amongst them.

Selecting sparsity $|\mathsf{K}^*|$ is dependent on the data, domain and the problem at hand. We employ a Bayes Factor approach as the criterion for selecting how many dimensions to choose in the greedy E-step. We observe the decay in increase in likelihood as we add dimensions greedily, and stop when the increase is not significant anymore.

6.1 Simulated Data

In real world datasets, even if the data conforms to the assumptions made by a model, the underlying truth is seldom known. For the case of sparse PCA, we would not know the true underlying support for the principal components in the real datasets. Hence, we first validate our model on simulated datasets. We tested the proposed algorithm and some competing baselines on several instances of toy data generated as follows. We fix the number of data points $n = 100$, the ambient dimension size $d = 1000$, the rank $r = 5$, and the sparsity at $k = 20$. We draw r principal components from a Gaussian distribution in the ambient space and zero out all but k randomly chosen dimensions in each of them to form $\mathbf{W}$. We draw $\mathbf{X}$, the linear weights, independently from $\mathcal{N}(0, 1)$. Finally, for each entry of $\mathbf{T} = \mathbf{XW}$, a noise value is added that is drawn from $\mathcal{N}(0, \sigma^2)$ for various values of σ^2. For metrics,



(a) Support Recovery as a function of noise variance

(b) Reconstruction error as a function of noise variance

Figure 1: Performance on Simulated Data

firstly we report the Receiver Operator Characteristic Area Under the Curve (ROC-AUC) on the support recovery of sparse matrix. Secondly, we look at the reconstruction error as the average $||\mathbf{T} - \hat{\mathbf{X}}\hat{\mathbf{W}}||$, where $\hat{\mathbf{X}}$ and $\hat{\mathbf{W}}$ indicate the respective fitted matrices for each of the methods. We compare against Online Dictionary Learning (ODL) [15], and scikit's sparsePCA and standard PCA. We use scikit's implementation for these[20]. 'Truth' in both the graphs is the value obtained using the correct generating parameters. For clarity, we have not included results from methods such as GPower. The presented models are the ones that are usually considered for recovering underlying bases, or for sparse reconstruction.

The results are summarized in Figure 1. We simulate 10 different datasets and graph the average values of AUC and reconstruction error. The figure shows how methods such as PCA perform well on reconstruction error but are not sparse as they overfit by including the noise dimensions as well. On the other hand, scikit's sparsePCA does reasonably well on capturing the underlying support but does not reconstruct well. Our method (SubmodPCA) does well on reconstruction error while also consistently recovering support, and degrades more gracefully as the noise increases.

6.2 fMRI data

Resting state (Functional Magnetic Resonance Imaging) fMRI data are commonly analyzed in order to identify coherently modulated brain networks that reflect intrinsic brain connectivity, which can vary in association with disease and phenotypic variables. We examined the performance of the present method on a resting-state fMRI scan lasting 10 minutes (3T whole-brain multiband EPI, TR=1.16 secs, 2.4 mm resolution), obtained from a healthy adult subject. Data were processed using a standard processing stream including motion correction and brain extraction (FSL).

The data originally captured has 518 data points, and over 100,000 dimensions. The ambient set of dimensions are clustered to fewer dimensions using the spatially constrained Ward hierarchical clustering approach of [16], to produce three smaller dimensional datasets with 100, 1000, 10000 dimensions. This makes the dataset challenging to deal with because we have cases where the dimensionality exceeds the number of datapoints.

We examined the support recovered from these data after estimating four components using our method. The first three components were largely restricted to regions reflecting motion artifacts, which suggests that this method may have utility in the detection and removal of artifacts from fMRI data (similar to previous use of ICA by Tohka et al. [25]). Figure 2 shows the brain map generated using the first principal component extracted using our algorithm. For the three datasets, we compare the ratio of variance explained by the k-sparse first principal component vector (i.e. number of non-zero entries is k) to the total variance in the dataset, for varying values of k. We compare against methods: Generalized Power Method [11] (Gpower), PCA via Low rank [19] (LR-PCA), Truncated Power Method [28] (Tpower), Online Dictionary Learning [15] (ODL) and Full Regularized Path Sparse PCA [6] (PathSPCA). For comparison, we run the standard PCA (non-sparse), and plot the ratio of explained variance along with all the above mentioned methods. Figure 3 shows the plots for all the three datasets. Note that Gpower and ODL take a regularization parameter rather than sparsity level directly. For both of them, the regularization parameter was adjusted to reach the intended sparsity level and those results are reported. For LRPCA, the

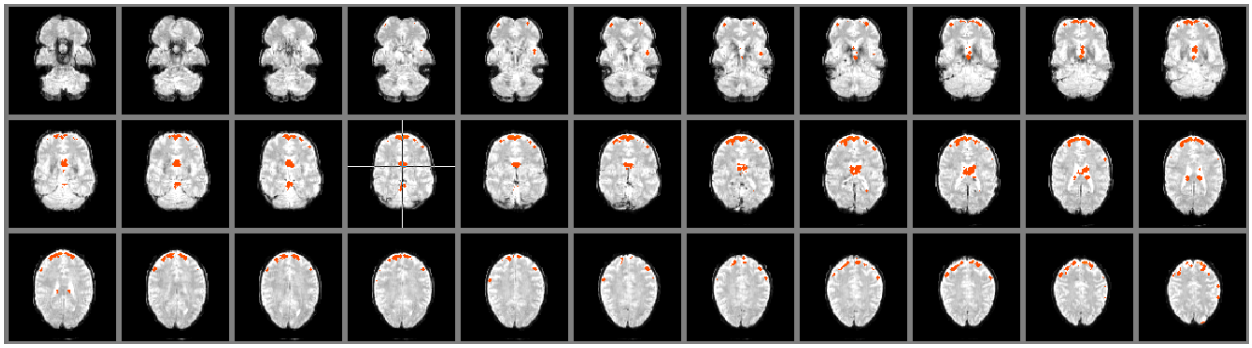


Figure 2: A projection of the first sparse component (shown in red) onto the mean fMRI image. The component is seen primarily in regions at the frontal surface as well as in the ventricles, consistent with motion artifact.

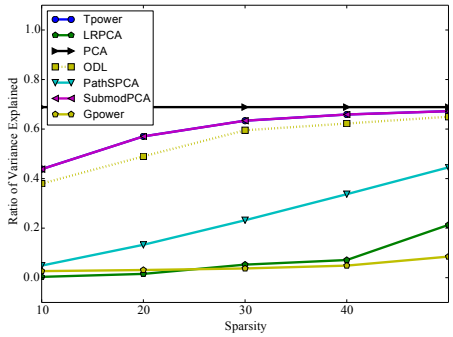
authors had implementation for rank =1 and higher ranks. The numbers we report are for rank=2 for d=100 and d=1000 and rank=1 for d=10000. This is because rank=2 for d=10000 was too slow and did not finish after 2 days. We did not notice significant difference in numbers between rank=1 and rank=2 for lower d. The plots clearly show that our method (SubmodPCA) performs consistently at least as well as any of the other sparse methods.

7 Conclusion

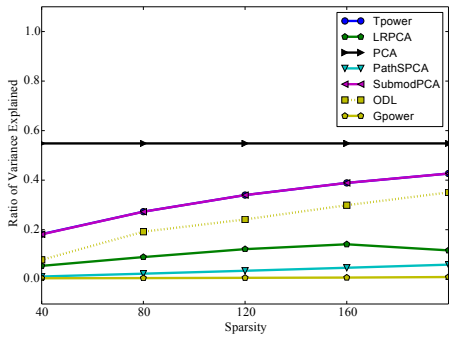
We have presented a novel method for variational inference under sparsity constraints by information projection, and applied it with the probabilistic PCA framework for sparse PCA. We also showed consistent performance vis-a-vis various baselines. The sparse inference method is general enough to be applied to other algorithms requiring sparsity in latent variables, such as sparse topic coding, and it is a natural future direction to take. This would be an interesting direction, since it would require exploration of distributions other than the Gaussian, so a lot of the algebra that worked out nicely may not hold. The sparse probabilistic framework can also be applied for problems like sparse inverse covariance estimation, and we plan to explore this direction in near future too.

Acknowledgement

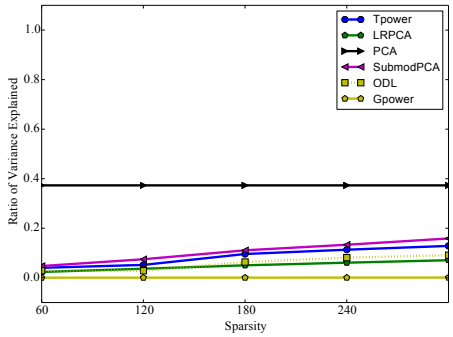
This work was supported by NSF grant IIS-1421729. fMRI data was provided by the Consortium for Neuropsychiatric Phenomics (NIH Roadmap for Medical Research grants UL1-DE019580, RL1MH083269, RL1DA024853, PL1MH083271).



(a) d=100



(b) d=1000



(c) d=10000

Figure 3: Performance on fMRI data

References

[1] Hervé Abdi. Factor rotations in factor analyses. *Encyclopedia for Research Methods for the Social Sciences. Sage: Thousand Oaks, CA*, pages 792–795, 2003.

[2] Hervé Abdi and Lynne J Williams. Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(4):433–459, 2010.

[3] Yasemin Altun and Alexander J. Smola. Unifying divergence minimization and statistical inference via convex duality. In *COLT*, 2006.

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

[5] Jorge Cadima and Ian T. Jolliffe. Loading and correlations in the interpretation of principle compenents. *Journal of Applied Statistics*, 22(2):203–214, 1995.

[6] Alexandre d'Aspremont, Francis R. Bach, and Laurent El Ghaoui. Full regularization path for sparse principal component analysis. In *Proceedings of the 24th International Conference on Machine Learning*, ICML '07, pages 177–184, New York, NY, USA, 2007. ACM. ISBN 978-1-59593-793-3.

[7] Uriel Feige. A threshold of ln n for approximating set cover. *Journal of the ACM (JACM)*, 45(4):634–652, 1998.

[8] Yue Guan and Jennifer G Dy. Sparse probabilistic principal component analysis. In *International Conference on Artificial Intelligence and Statistics*, pages 185–192, 2009.

[9] Ian T. Jolliffe. Rotation of principal components: choice of normalization constraints. *Journal of Applied Statistics*, 22(1):29–35, 1995.

[10] Ian T Jolliffe, Nickolay T Trendafilov, and Mudassir Uddin. A modified principal component technique based on the lasso. *Journal of Computational and Graphical Statistics*, 12(3):531–547, 2003.

[11] Michel Journée, Yurii Nesterov, Peter Richtárik, and Rodolphe Sepulchre. Generalized power method for sparse principal component analysis. *J. Mach. Learn. Res.*, 11:517–553, March 2010. ISSN 1532-4435. URL `http://dl.acm.org/citation.cfm?id=1756006.1756021`.

[12] Henry F Kaiser. The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, 23(3):187–200, 1958.

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

[14] Mokshay Madiman and Prasad Tetali. Information inequalities for joint distributions, with interpretations and applications. *Information Theory, IEEE Transactions on*, 56(6):2699–2713, 2010.

[15] Julien Mairal, Francis Bach, Jean Ponce, and Guillermo Sapiro. Online dictionary learning for sparse coding. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 689–696, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-516-1.

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

[17] 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.

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

[19] Dimitris S. Papailiopoulos, Alexandros G. Dimakis, and Stavros Korokythakis. Sparse pca through low-rank approximations. *ICML*, 2013.

[20] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[21] Kevin Sharp and Magnus Rattray. Dense message passing for sparse principal component analysis, 2010.

[22] Christian D. Sigg and Joachim M. Buhmann. Expectation-maximization for sparse and non-negative pca. In *ICML*, pages 960–967, New York, NY, USA, 2008. ACM. ISBN 978-1-60558-205-4.

[23] Michael E. Tipping. Sparse bayesian learning and the relevance vector machine. *JMLR*, 1:211–244, September 2001.

[24] 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.

[25] Jussi Tohka, Karin Foerde, Adam R Aron, Sabrina M Tom, Arthur W Toga, and Russell A Poldrack. Automatic independent component labeling for artifact removal in fmri. *Neuroimage*, 39(3):1227–1245, 2008.

[26] Dimitris G Tzikas, CL Likas, and Nikolaos P Galatsanos. The variational approximation for bayesian inference. *Signal Processing Magazine, IEEE*, 25(6): 131–146, 2008.

[27] Peter M Williams. Bayesian conditionalisation and the principle of minimum information. *The British Journal for the Philosophy of Science*, 31(2):131–144, 1980.

[28] Xiao-Tong Yuan and Tong Zhang. Truncated power method for sparse eigenvalue problems. *J. Mach. Learn. Res.*, 14(1):899–925, April 2013. ISSN 1532-4435.

[29] Hui Zou, Trevor Hastie, and Robert Tibshirani. Sparse principal component analysis. *Journal of Computational and Graphical Statistics*, 15, 2006.