A Deflation Method for Structured Probabilistic PCA

Rajiv Khanna* Joydeep Ghosh † Russell Poldrack ‡ Oluwasanmi Koyejo §

Abstract

Modern treatments of structured Principal Component Analysis often focus on the estimation of a single component under various assumptions or priors, such as sparsity and smoothness, and then the procedure is extended to multiple components by sequential estimation interleaved with deflation. While prior work has highlighted the importance of proper deflation for ensuring the quality of the estimated components, to our knowledge, proposed techniques have only been developed and applied to non-probabilistic principal component analyses, and are not trivially extended to probabilistic analyses. This work introduces a novel, robust and efficient deflation method for Probabilistic Principal Component Analysis using tools recently developed for constrained probabilistic estimation via information projection. The components estimated using the proposed deflation regain some of the interpretability of classic PCA such as straightforward estimates of variance explained, while retaining the ability to incorporate rich prior structure. Moreover, sequential estimation allows for scaling probabilistic techniques to be at par with their deterministic counterparts. Experimental results on simulated data demonstrate the utility of the proposed deflation in terms of component recovery, and evaluation on neuroimaging data show both qualitative and quantitative improvements in the quality of the estimated components. We also present timing experiments on real data to illustrate the importance of sequential estimation with proper deflation for scalability.

1 Introduction

Principal Component Analysis (PCA) is a well known technique for data exploration and dimensionality reduction [5]. The goal of PCA is to represent a centered data matrix as a linear combination of a few basis vectors (known as components) combined using weights. In the classical deterministic setting, the components are extracted as orthonormal vectors that maximize the explained variance in the data matrix. Beyond classic PCA, various extensions have been proposed that incorporate sparsity and/or other domain structure, or are designed to incorporate useful statistical properties such as noise tolerance in high dimensions [6, 26, 1, 7, 16]. Most modern treatments of Principal Component Analysis and its extensions focus on the estimation of a single component, leaving multi-component extensions to sequential estimation interleaved with deflation. This is more than a mere convenience, as sequential estimation may be necessary to enable scalability for modern "big data" problems, where one may be interested in computing only the top few principal components given a high-dimensional ambient feature space. Further, selecting the appropriate number of components (the rank) via sequential estimation avoids the significant computational overhead of re-estimating all the components each time the rank is changed, which is required without proper deflation.

Informally, the purpose of deflation is to minimize the influence of previously computed components on subsequent components, most often by assuming that subsequent components are mutually orthogonal. In the classic setting, orthogonal deflation can be accomplished simply by computing the residual[1], as the orthogonality of the components ensures that the result is equivalent to the residual of the data matrix row-projected onto the subspace of the first component [4, 2]. More sophisticated, but equivalent approaches include Hotelling's deflation [4] and Schur's complement deflation [25], applied to the data covariance matrix. Based on this intuition, much of the early work applied classic PCA deflation to more structured settings without justification. In response, Mackey [13] investigated the effect of deflation choices on the quality of inferences from sparse PCA, showing that many classic PCA deflation techniques may no longer be appropriate or equivalent. In particular, Mackey [13] showed that careless deflation does not preserve orthogonality and could lead to pathological results such as estimating the same component multiple times without explaining additional variance.

Probabilistic models remain one of the most popular approaches for data analysis. The generative view builds notions of parameter uncertainty into inferential procedure. Probabilistic models also simplify the process of incorporating rich domain knowledge via intelligent construction of prior distributions. Our work follows research by several authors who have explored probabilistic variants of principal components

*rajivak@utexas.edu, UT Austin
†jghosh@utexas.edu, UT Austin
‡poldrack@stanford.edu, Stanford University
§sanmi@illinois.edu, University of Illinois Urbana-Champaign

[1] given by the difference between the data and the rank one estimate i.e. the outer product of the component and the weight vectors

analysis [3, 18, 8]. Despite the rich prior literature on PCA, research has primarily focused on batch inferences and does not incorporate notions of sequential component estimation or deflation. Further, proposed techniques for deflation have only focused non-probabilistic Principal Component Analyses, and are not trivially extensible to probabilistic analyses.

In this manuscript, we seek to bridge this gap in the literature by first highlighting issues that may occur with improper deflation, and then presenting a robust and efficient deflation approach for probabilistic PCA (PPCA) that solves for each component sequentially, such that the posterior distribution of subsequent components are supported only on the subspace orthogonal to the subspace spanned by means of previously estimated component distributions. This approach guarantees orthogonality in the collected matrix component means. We also apply the framework to sparse PPCA and show that the means of components so obtained correspond to those obtained by known deterministic techniques in a special case. To the best of our knowledge, the correspondence between a sparse PPCA algorithm and a sparse deterministic PCA algorithm has not been established before. The components estimated using the proposed deflation regain some of the interpretability of classic principal component analysis such as straightforward estimates of variance explained, while retaining the ability to incorporate rich prior structure. Such simple interpretability if lost when doing joint estimation, since joint estimation only recovers the components upto a rotation. Furthermore, the issue of scalability from sequential estimation can not be overstated. We refer to prior work in the Bayesian literature evaluated on data with dimensionality $< O(10^3)$ e.g. [19] used data of dimensionality 35, while deterministic approaches are routinely applied to $> O(10^4)$ data. Our proposal pushes the scalability of probabilistic methods on par with their deterministic counterparts.

Our key contributions are as follows:

- we propose a novel deflation technique for probabilistic PCA via information projection of the posterior of each subsequent component onto the subspace orthogonal to the means of previously estimated component distributions.
- we explore an application of the proposed deflation approach to *sparse* probabilistic PCA
- we establish a correspondence of the proposed (sparse and non-sparse) PPCA algorithms to known deterministic techniques under special conditions, which may be of independent interest.

Experimental evaluation on neuroimaging data shows that deflation leads to improved interpretability (qualitative evaluation) and can improve variance explained by each component (quantitative evaluation). Furthermore, we also present empirical evidence of scalability by reporting the timing of sequential and joint estimation of components.

Notation: We represent vectors as small letter bolds e.g. $\mathbf{u}$. Matrices are represented by capital bolds e.g. $\mathbf{X}, \mathbf{T}$. Vector/matrix transposes are represented by superscript $\dagger$. Identity matrix is represented as $\mathbf{I}$. The i^{th} row of a matrix $\mathbf{M}$ is indexed as $\mathbf{M}_{i,.}$, while j^{th} column is $\mathbf{M}_{.,j}$. Sets are represented by sans serif fonts e.g. S. 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. Let $\mathrm{tr}(\mathbf{X})$ denote the trace of the matrix $\mathbf{X}$.

2 Background and Related Work

Let $\mathbf{T} \in \mathbb{R}^{n \times d}$ represent the data matrix, with n samples and dimensionality d. Without loss of generality, we assume that the data matrix is mean centered in each dimension. Given a desired rank r, PCA decomposes a centered data matrix into components $\mathbf{W} \in \mathbb{R}^{r \times d}$ and weights $\mathbf{X} \in \mathbb{R}^{n \times r}$.

In the classical deterministic setting, components are extracted as orthonormal vectors that maximize the explained variance in the data matrix. The *first* principal component $\mathbf{w} \in \mathbb{R}^d$ may be computed as:

$$\max_{||\mathbf{w}||_2=1} \mathbf{w}^\dagger \mathbf{\Sigma} \mathbf{w}, \tag{2.1}$$

where $\mathbf{\Sigma} = \mathbf{T}^\dagger \mathbf{T} \in \mathbb{R}^{d \times d}$ is the data covariance matrix. The solution is the eigenvector of the covariance matrix which is associated with the largest eigenvalue. The associated *variance explained* is simply the value of the cost function (2.1) at the solution. To obtain the next component, the covariance matrix is *deflated* to remove the variance component explained, then (2.1) is re-solved with the deflated covariance. Using Hotelling's deflation [4], the subsequent covariance matrix at step $i+1$ is computed from the i^{th} covariance matrix as:

$$\mathbf{\Sigma}_{i+1} = \mathbf{\Sigma}_i - \mathbf{w}_i \mathbf{w}_i^\dagger \mathbf{\Sigma}_i \mathbf{w}_i \mathbf{w}_i^\dagger, \tag{2.2}$$

where $\mathbf{\Sigma}_0 = \mathbf{\Sigma}$ and $\mathbf{w}_0$ is the first component. Alternatives to Hotelling's deflation include Schur complement deflation and orthogonal projection (see Mackey [13] for more details).

While the covariance approach is perhaps the most popular, an alternative and equivalent approach is to estimate both the components and the weights to minimize the reconstruction error [17] as:

$$\min_{\mathbf{x}, ||\mathbf{w}||_2=1} ||\mathbf{T} - \mathbf{x}\mathbf{w}^\dagger||_F^2. \tag{2.3}$$

It is clear that the reconstruction error view of classic PCA is equivalent to modeling using the Gaussian likelihood. The optimal $\mathbf{w}$ is given by the right singular vector of the data matrix which is associated with the largest singular value and $\mathbf{x}$ is the corresponding left singular vector multiplied

by the singular value. The associated *variance explained* can be computed using the same equation as the covariance approach (2.1) at the solution. The reconstruction error view suggests the naïve deflation for the subsequent components by replacing the data matrix with the residual in (2.3), given by:

$$\mathbf{T}_{i+1} = \mathbf{T}_i - \mathbf{x}_i \mathbf{w}_i^\dagger, \tag{2.4}$$

where $\mathbf{T}_0 = \mathbf{T}$ and $\mathbf{x}_0$ is the *first* weight vector. The naïve deflation is equivalent to other deflation techniques in the classic setting.

Probabilistic PCA (PPCA) is the probabilistic extension of the deterministic PCA. The likelihood is chosen to match the reconstruction error view of the classic PCA. The components ($\mathbf{w}_i$) are random variables with a prior that is designed by domain knowledge. The prior can be used to incorporate structural properties such as smoothness, sparsity, and non-negativity into the components as suggested by the domain. The factorization can then be obtained by maximizing the log likelihood, typically by EM algorithm to solve for all r components at the same time. See [21] for details.

Related Work on PPCA: Probabilistic PCA was first proposed by Tipping and Bishop [21] based on an extension of the well established component models in statistics. Tipping and Bishop [21] showed that the result was equivalent to standard PCA under certain choices of hyperparameters, and generalized PPCA to incorporate priors on the weights. Šmídl and Quinn [19] extended this work to a full Bayesian treatment which included priors on both components and weights, and considered the use of appropriate priors on the components to enforce orthogonality. Beyond standard PCA, several authors have proposed additional priors to encourage sparsity or non-negativity on the components [3, 18]. Recently Khanna et al. [8] proposed a submodular formulation for sparse probabilistic PCA but, like other prior work, focused on single component estimation rather than on (sequential) deflation.

2.1 Constrained Probabilistic Inference via Information Projection In the interest of a self contained discussion, this section outlines relevant background on constrained probabilistic inference via information projection, which will be useful for the development of our proposed deflation technique. We begin with the definitions of the Kullback-Leibler divergence and information projection. Let X represent the sample space of interest. Let $\mathcal{P}$ represent the set of bounded densities supported on X.

Assumptions: The sample space X is either a subset of the Euclidean space with associated Lebesgue measure as the dominating measure, or a countable set with associated counting measure. All probability distributions considered are absolutely continuous w.r.t the respective dominating measures, have bounded densities. These assumptions allow us to define consistent conditional densities on subsets $\mathsf{S} \subset \mathsf{X}$ of measure 0 by disintegration, and allow for well-defined information projection onto S. Furthermore, the notion that such restriction can be posed as a variational optimization problem allows us to incorporate sparsity constraints. See [11] for more details.

Information Projection: Let $p \in \mathcal{P}$ and $q \in \mathcal{P}$. We use $\mathrm{KL}(q\|p)$ to denote the Kullback-Leibler divergence [12] between p and q. Given a set $\mathcal{Q} \subseteq \mathcal{P}$, the *information projection* of $p \in \mathcal{P}$ to the set $\mathcal{Q}$ is given by:

$$\inf_{q \in \mathcal{Q}} \mathrm{KL}(q\|p).$$

As we only consider closed subsets, *inf* above can be replaced by *min*. Let $\mathsf{S} \subset \mathsf{X}$ represent a closed subset of X, so $\mathcal{P}_\mathsf{S}$ is the set of all probability densities supported on S. Our analysis will focus on the information projection of p onto $\mathcal{P}_\mathsf{S}$. We will sometimes refer to this as the information of p to the set S.

Domain restriction: Let $p \in \mathcal{P}$ be a probability density on X, and let $\mathsf{S} \subset \mathsf{X}$, then p_S is the S-restriction of p: $p_\mathsf{S}(\mathbf{x}) = 0$ if $x \notin \mathsf{S}$, $p_\mathsf{S}(\mathbf{x}) = \frac{p(\mathbf{x})}{\int_{\mathbf{s} \in \mathsf{S}} p(\mathbf{s}) d\mathbf{s}}$ if $\mathbf{x} \in \mathsf{S}$.

The following Lemma establishes the equivalence of domain restriction and a certain information projection. As a result, domain restriction may be solved as a variational optimization problem.

LEMMA 2.1. (KOYEJO ET AL. [11]) *Let p be a probability density defined on a measurable set X, $\mathsf{S} \subset \mathsf{X}$ be a closed set, p_S be the S-restriction of p, $\mathcal{P}_\mathsf{S}$ be the set of all probability distributions supported on S then $p_\mathsf{S} = \min_{q \in \mathcal{P}_\mathsf{S}} \mathrm{KL}(q\|p)$.*

2.2 Information Projection onto Subspaces The results in this section hold for general subsets of the Euclidean space, but for our purposes we will restrict our attention to subspaces. We overload the nomenclature a little and use "information projection onto the set of distributions supported on a subspace" and "information projection onto the subspace" interchangeably. Let $\mathcal{M}$ be the target subspace onto which we aim to restrict a probability distribution p. We will apply Lemma 2.1 to pose it as a variational optimization problem. To encode the constraint set, we make use of the setup of Bayesian optimization under expectation constraints which has been well studied in the literature [9]. The following proposition characterizes information projection onto subspaces.

PROPOSITION 2.1. *Consider a function $\phi_\mathcal{M} : \mathbb{R}^d \to \mathbb{R}$ which satisfies $\phi_\mathcal{M}(\mathbf{x}) = 0$ if $x \in \mathcal{M}$, and $\phi_\mathcal{M}(\mathbf{x}) > 0$ if $x \notin \mathcal{M}$. The restriction of the density p to a subspace $\mathcal{M}$ can be obtained as:*

$$\underset{q}{\textit{argmin}}\ \mathrm{KL}(q\|p) \textit{ s.t. } \mathbb{E}_q[\phi_\mathcal{M}(\mathbf{x})] = 0. \tag{2.5}$$

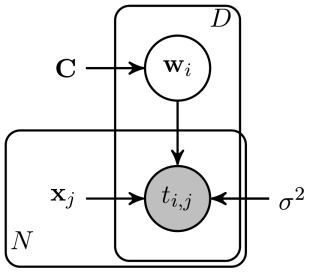


Figure 1: Plate model for Probabilistic PCA. The matrix $\mathbf{C}$ is the prior design matrix. In classical PPCA [21], $\mathbf{C}$ is identity

2.2.1 Information Projection of Gaussian Distributions

The special case where p is a Gaussian is of particular interest to our development of deflation for sparse PPCA. Let $\mathcal{N}(\boldsymbol{\mu}, \mathbf{S})$ represent a multivariate Gaussian distribution with mean $\boldsymbol{\mu} \in \mathbb{R}^d$ and covariance $\mathbf{S} \in \mathbb{R}^{d \times d}$. Let $\mathcal{M}_\perp$ represent the orthogonal complement of a subspace $\mathcal{M}$. We denote the projection matrix associated with the subspace $\mathcal{M}$ by $\mathrm{P}_\mathcal{M}$. We can write $\phi_\mathcal{M}(\mathbf{x}) := x^\dagger \mathrm{P}_{\mathcal{M}_\perp} x$. It is clear that $x \in \mathcal{M} \implies \phi_\mathcal{M}(\mathbf{x}) = 0$ and $x \notin \mathcal{M} \implies \phi_\mathcal{M}(\mathbf{x}) > 0$.

When p is Gaussian, it is known that the information projection onto $\mathrm{P}_\mathcal{M}$ is also a Gaussian distribution [10, 9]. We emphasize that this is not an assumption, but rather a property of the KL divergence. Thus, the search for the projection may be restricted to optimization over the members of the family $q \sim \mathcal{N}(\mathbf{a}, \mathbf{B})$ identified by the mean and covariance $\{\mathbf{a}, \mathbf{B}\}$. The constraint in (2.5) can be expanded as $\mathbb{E}_q[\phi_\mathcal{M}(\mathbf{x})] = 0 \implies \mathrm{tr}\big(\mathrm{P}_{\mathcal{M}_\perp} \mathbf{a}\mathbf{a}^\dagger + \mathrm{P}_{\mathcal{M}_\perp} \mathbf{B}\big) = 0 \implies \mathrm{tr}\big(\mathrm{P}_{\mathcal{M}_\perp} \mathbf{a}\mathbf{a}^\dagger\big) = 0$ and $\mathrm{tr}(\mathrm{P}_{\mathcal{M}_\perp} \mathbf{B}) = 0$ (since all projection matrices are positive semidefinite). Expanding Equation 2.5 using definition of KL between two multivariate Gaussian distributions results in the decoupled optimization problems:

$$\min_{\mathbf{B}} \mathrm{tr}\big(\mathbf{S}^{-1}\mathbf{B}\big) - \ln\det \mathbf{B} \text{ s.t. } \mathrm{tr}(\mathrm{P}_{\mathcal{M}_\perp}\mathbf{B}) = 0,$$
$$\min_{\mathbf{a}} (\mathbf{a} - \boldsymbol{\mu})^\dagger \mathbf{S}^{-1} (\mathbf{a} - \boldsymbol{\mu}) \text{ s.t. } \mathrm{tr}\big(\mathrm{P}_{\mathcal{M}_\perp}\mathbf{a}\mathbf{a}^\dagger\big) = 0.$$

These are solved by $(\mathbf{B}^*)^{-1} = \mathrm{P}_\mathcal{M} \mathbf{S}^{-1} \mathrm{P}_\mathcal{M}$ and $\mathbf{a}^* = \mathbf{B}^* \mathrm{P}_\mathcal{M} \mathbf{S}^{-1} \boldsymbol{\mu}$. Thus, the information projection of $p \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ to the subspace $\mathcal{M}$ is given by $q^* \sim \mathcal{N}(\mathbf{a}^*, \mathbf{B}^*)$.

3 Deflation for Probabilistic PCA

We consider n observations of d dimensional vectors stacked as the data matrix $\mathbf{T} \in \mathbb{R}^{n \times d}$. Without loss of generality, we assume that the matrix is centered i.e. each column has mean 0. The data matrix is modeled as a product of parameter $\mathbf{X}$ and latent variable $\mathbf{W}$ which has the matrix-variate normal prior MVN(0, $\mathbf{C}$, $\mathbf{I}$) where, $\mathbf{C}$ is the column covariance matrix, and $\mathbf{I}$ (identity matrix) is the row covariance matrix. The observation model is $\mathbf{T} = \mathbf{X}\mathbf{W} + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon}$ is the noise with prior $\boldsymbol{\epsilon}_{ij} \sim \mathcal{N}(0, \sigma^2)$. See Figure 1.

Motivating Example: Consider the following example showing the a potential failure of probabilistic PCA with naïve deflation. We selected the components and sample the weights and noise as: $\mathbf{W} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix}$, $\mathbf{x}_k \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, $\mathbf{e}_n \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, where $n = 100,000$.Note that this generative scheme adheres to the specification of the PPCA above. We applied probabilistic PCA of Tipping and Bishop [21] sequentially using the naïve deflation of (2.4) based on the estimated expected component. As shown in Figure 2, the procedure estimated degenerate components with expected value: $\begin{bmatrix} 1 & 1 \\ 0.1 & -0.1 \\ 0 & 0 \end{bmatrix}$ rounded to two significant digits. This is partially due to the noise and the effects of prior regularization. Such a degenerate result was not observed for the full model fit or with the proposed deflation that enforces orthogonality, where the correct components were recovered, and this seems to be specific to the naïve deflation. Fitting a full model however is less scalable, and loses on the opportunity of run time model selection. Moreover, for a full model fit, the interpretation of individual components as directions maximizing explained variance sequentially is also no longer valid. An alternative to retain such an interpretation would be to add orthogonality in the full model which may not be easy as it requires handling distributions on the Grassmanian [24].

3.1 Inference for Probabilistic PCA via Variational EM

Probabilistic PCA is typically solved by an EM algorithm which obviates construction of the full covariance matrix, and instead enables working with the data matrix while returning both the weights and the components. Maximizing the negative log-likelihood can be shown to be equivalent to maximizing a free energy function $\mathscr{F}$ (3.6). 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 (3.7), and the M-step can be interpreted to be the search over the parameter space, keeping the latent variables distribution q fixed (3.8). The cost function for the EM is given by [15]:

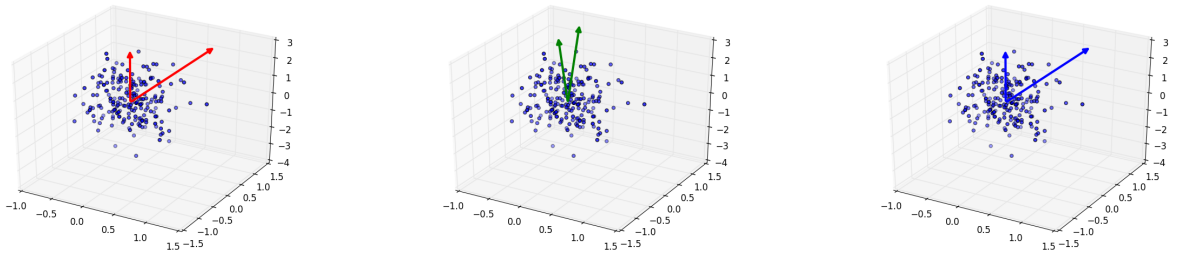
(3.6)
$$\mathscr{F}(q(\mathbf{W}), \Theta) = -\mathrm{KL}(q(\mathbf{W}) \| p(\mathbf{W}|\mathbf{T}; \Theta)) + \log p(\mathbf{T}; \Theta).$$

with the E-step and M-step given by:

(3.7) E-step: $\max_q \mathscr{F}(q(\mathbf{W}), \Theta)$,

(3.8) M-step: $\max_\Theta \mathscr{F}(q(\mathbf{W}), \Theta)$.

This view of the EM algorithm provides the flexibility to design algorithms with any E and M steps that monotonically increase $\mathscr{F}$. An unconstrained optimization over q in (3.7) returns the posterior $p(\mathbf{W}|\mathbf{T}; \Theta)$. Variational methods perform the search for best q over a constrained set [22] using constrained KL minimization. Let D be the set of distributions over $\mathbf{W}$ that fully factorize over individual rows



(a) Ground Truth (b) Estimation via naive deflation (c) Estimation via proposed deflation

Figure 2: Simulated data example showing incorrect estimates using the naïve deflation, while the proposed deflation technique recovers close to the ground truth.

of $\mathbf{W}$: $q(\mathbf{W}) = \prod_{i=1}^{r} q_i(\mathbf{W}_{i,.})$, and $\forall i, q_i$ is Gaussian. We restrict the search over q to D. This restriction is known as the mean field variational approximation. Based on the factorization assumption, the KL minimization separates out for each i and can be solved for each q_i iteratively.

Naïve Deflation: As generative models do not modify the data matrix directly, deflation is achieved implicitly by fixing the distributions of the estimated components $q(\mathbf{W}_{\setminus i})$ when estimating the distribution of the new component $q(\mathbf{w}_i)$. Following the E-step (3.7), the effect on the model is straightforward to compute as (up to additive and multiplicative constants):

$$\begin{aligned}&\mathrm{E}_{q(\mathbf{W}_{\setminus i})}\left[\log P(\mathbf{T}|\mathbf{X}, \mathbf{W}_{\setminus i}, \mathbf{x}_i, \mathbf{w}_i)\right] \\ &\propto \left|\left|\mathbf{T} - \mathbf{X}\mathrm{E}_{q(\mathbf{W}_{\setminus i})}\left[\mathbf{W}_{\setminus i}\right] - \mathbf{x}_i\mathbf{w}_i^{\dagger}\right|\right|_F^2.\end{aligned}$$

With components $j < i$ fixed, it is clear that this is equivalent to the naïve deflation of (2.4) using the estimated posterior mean.

Deflation is well-studied for the deterministic PCA, but we are not aware of any previous work that address proper deflation while solving for probabilistic components incrementally. While the EM algorithm returns full distribution of components, typically normalized expected values are used as point estimates. The standard approach for deflation is the naïve approach outlined in (2.4), using the expected value of the component with respect to the posterior. This is given by $\mathbf{Z}_i = \mathbf{T} - \Sigma_{j<i}\mathbf{X}_{.,j}^{\dagger}\mathbf{m}_j$, and is a direct consequence of the variational estimation [8].

3.2 Orthogonal Deflation We propose deflation following the classic definition of orthogonality. Specifically, we consider orthogonality between the posterior means of the estimated subspaces. This is implemented using the information projection approach outlined in Section 2.2. Let $\mathcal{M}^i$ be subspace spanned by means of first i components i.e. the subspace spanned by $\bigcup_{j=1}^{i} \mathbb{E}[\mathbf{W}_{j.,}]$. We restrict the support of component $(i+1)$ to be $\mathcal{M}_{\perp}^{i}$.

Let $\mathbf{Z}_i = \mathbf{T} - \Sigma_{j<i}\mathbf{X}_{.,j}\mathbb{E}[\mathbf{W}_{j,.}]$. Following the formula for information projection for Gaussians derived in Section 2.2.1, the E-step update for the i^{th} component is given by $q_i(\mathbf{W}_{i,.}) \sim \mathcal{N}(\mathbf{m}_i, \mathbf{\Sigma}_i)$ where:

$$\begin{aligned}\mathbf{\Sigma}_i^{-1} &= \mathrm{P}_{\mathcal{M}_{\perp}^{(i-1)}}\left(\frac{1}{\sigma^2}(\mathbf{X}_{.,i}^{\dagger}\mathbf{X}_{.,i})\mathbf{I} + \mathbf{C}^{-1}\right)\mathrm{P}_{\mathcal{M}_{\perp}^{(i-1)}}, \\ \mathbf{m}_i &= \frac{1}{\sigma^2}\mathbf{\Sigma}_i\mathrm{P}_{\mathcal{M}_{\perp}^{(i-1)}}\mathbf{Z}_i^{\dagger}\mathbf{X}_{.,i}.\end{aligned} \tag{3.9}$$

Note that (3.9) without the projection operator $\mathrm{P}_{\mathcal{M}_{\perp}}$ would be the posterior under the standard mean field assumption, the projection operator is added as a result of the variational E-step that performs the information projection.

The M-step is also straightforward to derive as:

$$\begin{aligned}\mathbf{X}_{.,i} &= \frac{\mathbf{Z}_i\mathbf{m}_i}{\mathrm{tr}\left(\mathbf{m}_i\mathbf{m}_i^{\dagger} + \mathbf{\Sigma}_i\right)}, \\ \sigma^2 &= \frac{\mathrm{tr}\left(\mathbf{Z}_i^{\dagger}\mathbf{Z}_i\right) + \mathbf{X}_{.,i}^{\dagger}\mathbf{X}_{.,i}\mathrm{tr}\left(\mathbf{m}_i\mathbf{m}_i^{\dagger} + \mathbf{\Sigma}_i\right) - 2\mathbf{m}_i^{\dagger}\mathbf{Z}_i^{\dagger}\mathbf{X}_{.,i}}{nd}.\end{aligned} \tag{3.10}$$

We term the resulting procedure of sequential estimation and deflation Orthogonal Probabilistic PCA (oPPCA).

3.3 Reduction to PCA The orthogonal deflation is reminiscent of the Hotelling deflation on the data matrix. Indeed, if $\mathbf{C} = \mathbf{I}$ and $\mathbf{m}_i$ are normalized, by substituting the value of $\mathbf{X}_{.,i}$ from the M-step into (3.9), we compute:

$$\mathbf{Z}_i = \mathbf{T}(\mathbf{I} - \Sigma_{j<i}\alpha_j\mathbf{m}_j\mathbf{m}_j^{\dagger}) \tag{3.11}$$

for constants α_i (which represent the explained variance by component $\mathbf{m}_i$ while like in PCA and PPCA, σ^2 measures the noise or unexplained variance).

PROPOSITION 3.1. *If* $\mathbf{C} = \mathbf{I}$ *the means of components estimated by oPPCA correspond to the components estimated by deterministic PCA.*

Proof. Substitute the value of $\mathbf{X}_{.,1}$ from the M-step into the update equation of $\mathbf{m}_1$ in the E-step equation to get $\mathbf{m}_1 \propto \mathbf{T}^\dagger\mathbf{T}\mathbf{m}_1$ which shows that solving for the first component $\mathbf{m}_1$ is equivalent to performing power iterations on $\mathbf{T}^\dagger\mathbf{T}$. From (3.11), for the subsequent components, solving for $\mathbf{m}_i$ is equivalent to performing power iterations on the deflated matrix $(\mathbf{I} - \Sigma_{j<i}\alpha_j m_j m_j^\dagger)\mathbf{T}^\dagger\mathbf{T}(\mathbf{I} - \Sigma_{j<i}\alpha_j m_j m_j^\dagger)$.

4 Deflation for Sparse Probabilistic PCA (soPPCA)

The proposed deflation using the framework may also be extended to sparse Probabilistic PCA, where the support of components is to be restricted to a few dimensions. We focus on the approach proposed by Khanna et al. [8] as it directly utilizes information projection to impose sparsity on the components. Thus, for component i and given $k_i < d$, we can directly extend the variational E-step to restrict the support to the *best* k_i dimensions in terms of the minimum KL divergence. minimized.

Let $\mathcal{S}_{k_i}$ be the set of all subspaces of dimension k_i spanned by k_i-sized subsets of the power set of set of standard bases $\{e_j, j \in [1..d]\}$. Also, let $\bar{p_i}$ be the full posterior for the i^{th} component (before the information projection). The variational E-step for sparse component $\mathbf{W}_{i,.}$ is given by:

$$\min_{\substack{\text{Supp}(\bar{q}_i(\mathbf{W}_{i,.}))\in(\mathrm{P}_{M_\perp^{(i-1)}}\cap\mathcal{S}) \\ \mathcal{S}\in\mathcal{S}_{k_i}}} \mathrm{KL}(\bar{q}_i(\mathbf{W}_{i,.})\|\bar{p_i}(\mathbf{W}_{i,.}|\mathbf{Z}_i)). \tag{4.12}$$

The support constraint on $\bar{q}$ requires information projection onto an intersection of two sets. It can be shown that it is equivalent to minimizing the constrained KL divergence by enforcing the support constraints of each set one after the other. This equivalence is due to a property of iterated information projections (see Koyejo et al. [11] for details). Combined with the proposed orthogonal deflation, and with q_i as defined in 3.9, the resulting E-step is solved via:

$$\min_{\text{Supp}(\bar{q}_i(\mathbf{W}_{i,.}))\in\mathcal{S}_{k_i}} \mathrm{KL}(\bar{q}_i(\mathbf{W}_{i,.})\|q_i(\mathbf{W}_{i,.})), \tag{4.13}$$

Expanding the KL in the optimization problem (4.13), we obtain the equivalent combinatorial problem:

$$\max_{\mathcal{S}\in\mathcal{S}_{k_i}} (\mathrm{P}_\mathcal{S}\mathbf{r}_i)^\dagger(\mathrm{P}_\mathcal{S}\mathbf{\Sigma}_i^{-1}\mathrm{P}_\mathcal{S})^{-1}(\mathrm{P}_\mathcal{S}\mathbf{r}_i) - \log\det \mathrm{P}_\mathcal{S}\mathbf{\Sigma}_i^{-1}\mathrm{P}_\mathcal{S}, \tag{4.14}$$

where $\mathbf{r}_i = \mathbf{\Sigma}_i^{-1}\mathbf{m}_i$. Koyejo et al. [11] showed that the resulting optimization problem is submodular, so a greedy search is guaranteed to find a solution close to the global optimal. This greedy approach has also be shown to be effective in practice for linear regression [11] and sparse PPCA [8]. We also effectively employ greedy search, and following optimization of $\mathcal{S}_i^*$, estimate the approximate posterior $\bar{q}_i \sim \mathcal{N}(\mathbf{c}_i, \mathbf{D}_i)$ where

$$(\mathbf{D})^{-1} = \mathrm{P}_{\mathcal{S}_i^*}\mathbf{\Sigma}_i^{-1}\mathrm{P}_{\mathcal{S}_i^*},\ \mathbf{c}_i = \mathbf{D}\mathrm{P}_{\mathcal{S}_i^*}\mathbf{r}_i \tag{4.15}$$

The M-step equations are again solved by (3.10) where $\bar{q}_i$ is substituted for q_i. We term the overall procedure sparse orthogonal probabilistic PCA (soPPCA).

The method derived above reduces to the Truncated Power Method (TPower[7]) with orthogonal projection deflation (see the supplement for details) when $\mathbf{C}$ is identity. soPPCA is thus, a generalization of TPower as a non-identity $\mathbf{C}$ helps in incorporating domain knowledge which can be useful as we see in Section 5.

4.1 Reduction of soPPCA to the Truncated Power Method Truncated power method is a simple algorithm to evaluate $k-$sparse principal eigenvector of a positive semidefinite matrix. It is similar to the standard power method, except that at every iteration it truncates the iterating vector to top-k absolute values and zeros out the rest of the vector before normalizing (see Yuan and Zhang [23] for details and recovery guarantees). The following proposition shows an equivalence between a single component from soPPCA and the truncated power method.

PROPOSITION 4.1. *If* $\mathbf{C} = \mathbf{I}$*, the normalized mean of the component* $\mathbf{m}_1$ *is equal to the principal sparse eigenvector obtained by the truncated power method on the covariance matrix of* $\mathbf{T}$.

Proof. If $\mathbf{C} = \mathbf{I}$, the optimization problem 4.14 reduces to (by combining E-step and M-step, and ignoring scaling constants since they vanish when normalizing):

$$\begin{aligned}
\max_{\mathcal{S}\in\mathcal{S}_{k_1}} (\mathrm{P}_\mathcal{S}\mathbf{r}_1)^\dagger(\mathrm{P}_\mathcal{S}\mathbf{r}_1) &\equiv \max_{\substack{\mathsf{K}\subset[d] \\ |\mathsf{K}|=k_i}} \mathbf{r}_\mathsf{K}^\dagger\mathbf{r}_\mathsf{K} \\
&\equiv \max_{\substack{\mathsf{K}\subset[d] \\ |\mathsf{K}|=k_i}} \mathrm{abs}(\mathbf{r}_\mathsf{K}) \\
&\equiv \max_{\substack{\mathsf{K}\subset[d] \\ |\mathsf{K}|=k_i}} \mathrm{abs}(\mathbf{T}^\dagger\mathbf{T}\mathbf{m}_1)
\end{aligned}$$

Orthogonal projection deflation of the covariance matrix involves a Gram-Schmidt procedure to build orthogonal set of components from possibly non-orthogonal ones obtained after projection deflation [13]. The following corollary shows an equivalence between soPPCA and the truncated power method with orthogonal projection covariance deflation.

COROLLARY 4.1. *If* $\mathbf{C} = \mathbf{I}$*, the means of the components estimated by soPPCA recover the sparse eigenvectors obtained by the truncated power method with orthogonal projection deflation.*

Proof. Follows from Proposition 4.1, the projection deflation formula 3.11 and the fact that projection deflation with truncated power method is equivalent to orthogonal projection deflation.

Thus, for the special case of $\mathbf{C} = \mathbf{I}$, the recovery and performance guarantees obtained by Truncated Power method are also inherited by soPPCA. Note that the covariance matrix $\mathbf{C}$ can provide important information about the domain and can aid in extracting out more meaningful components as compared to using $\mathbf{I}$. We shall see the performance gains in Section 5.

5 Experiments

In this section we present empirical results to illustrate the utility of orthogonality in probabilistic PCA models in practice. We perform quantitative analysis on resting state fMRI dataset with state of the art sparse PCA methods. To illustrate the practical applicability, we also provide qualitative analysis regarding the interpretation of the extracted components by a domain expert.

One of the key questions 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 components. Smith et al. [20] considered a similar question using the brain map database decomposed via ICA, showing correspondence between task activation components and resting state components. Following their aproach, we downloaded 1669 fMRI task statistical maps from neurovault (http://neurovault.org/). Each image in the collection represents a standardized statistical map of univariate brain voxel activation in response to an experimental manipulation. The statistical maps were downsampled from 2mm $\wedge$ 3 voxels to 3mm $\wedge$ 3 voxels using the nilearn python package (http://nilearn.github.io/). We then applied the standard brain mask, removing voxels outside of the grey matter, resulting in D=65598 variables (dimensions).

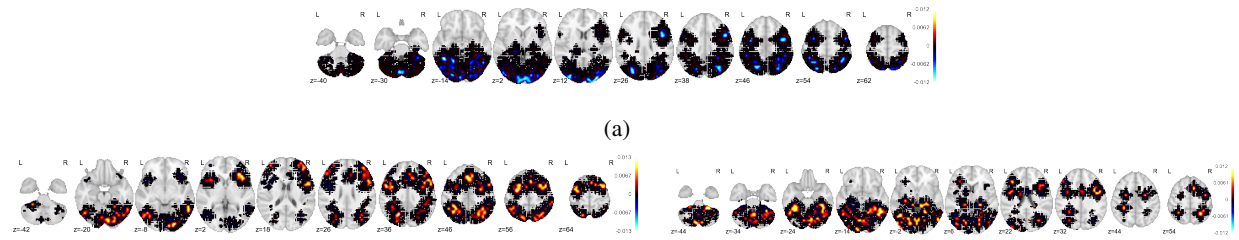
We cluster the original set of dimensions to fewer dimensions using the spatially constrained Ward hierarchical clustering approach of [14], to produce three smaller dimensional datasets with 100, 1000, 10000 dimensions. This makes the dataset challenging to deal with because we have cases where the dimensionality exceeds the number of datapoints. We incorporate smoothness via spatial correlation matrix $\mathbf{C}$ on the prior on $\mathbf{W}$.

For the three datasets, we compare the ratio of variance explained by first 6 sparse components to the total variance in the dataset. For $d = 100$, each sparse component has sparsity 10, $d = 1000$ and $d = 10000$ have sparsity of 10 amd 20 respectively in each of their principal components. To illustrate the generalization ability obtained by the use of proper priors, we split the data 50-50 training and testing. We find the k sparse principal components on the training data, and use the extracted components to estimate the variance explained on the out of sample test data. We compare against: Generalized Power Method [7] (Gpower), PCA via Low rank [16] (LRPCA), Truncated Power Method [23] (Tpower), Full Regularized Path Sparse PCA [1] (PathSPCA), emPCA [18], submodPCA [8]. We plot the ratio of explained variance along with all the above mentioned methods. Figure 4 shows the plots for all the three datasets. soPPCA performs better than all the other methods on the three datasets. Of special note is the gain in performance over submodPCA which uses naïve deflation as opposed to the orthogonal deflation proposed in this paper. The gain in performance is more apparent with increase in dimensionality. For $d = 10000$ and 10 principal components, our deflation technique achieves more than 20% performance improvement with respect to the state-of-the-art TPower, more than 50% improvement over the state-of-the-art probabilistic PCA method using naïve deflation, and more than 400% improvement vs emPCA. We note that our improvement is even greater than the corresponding improvement obtained by [13] for deterministic deflation techniques. Qualitative analysis is presented in Figure 3. The brain maps presented reflect overlapping but distinct task activation networks, all involving activation of visual cortices. The networks vary in the lateralization of prefrontal engagement, with (a) showing largely right-lateralized ventral prefrontal, (b) showing largely bilateral dorsolateral prefrontal, and (c) showing bilateral premotor and left-lateralized ventrolateral activation

Timing experiments To illustrate the importance of deflation for scalability, we select the one of the datasets used above (with $d = 1000$), and show the difference in time taken to infer top-6 components for non-sparse Probabilistic PCA (Figure 1) using traditional joint estimation of all 6 components ([21]) vs our approach ofsequential estimation with orthogonal deflation outlined in Section 3.2. The results are presented in Figure 5. On the x-axis is the number of top-k components extracted, while on the y-axis is the amount of time taken in seconds. We plot total time taken to estimate top-k components using both the methods. The plot clearly shows the importance of deflation for scalability – the time taken for joint estimation grows exponentially as opposed to growing linearly when using deflation. We further state that these results were obtained on the smallest dataset of the three datasets used in this manuscript for illustrative purposes, and difference in the time taken on bigger datasets are even more profound. We hope that this analysis and our presented method can aid in being a useful tool for scaling up probabilistic methods to fit the needs for modern day data analysis.

6 Conclusion and Future Work

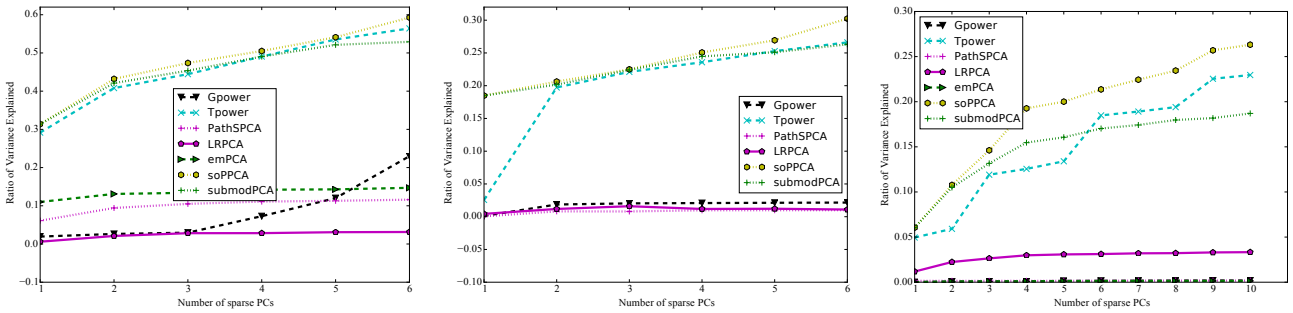
We proposed a general purpose technique that can be applied to incorporate deflation into probabilistic PCA and its sparse



(a)

(b) (c)

Figure 3: Brain plots of top-3 components extracted from the fMRI data. They reflect overlapping but distinct task activation networks, all involving activation of visual cortices. The networks vary in the lateralization of prefrontal engagement, with (a) showing largely right-lateralized ventral prefrontal, (b) showing largely bilateral dorsolateral prefrontal, and (c) showing bilateral premotor and left-lateralized ventrolateral activation



(a) d=100 (b) d=1000 (c) d=10000

Figure 4: Quantitative Performance on fMRI Data. For $d = 10000$ and 10 principal components, our deflation technique achieves more than 20% performance improvement with respect to the state-of-the-art TPower, more than 50% improvement over the state-of-the-art probabilistic PCA method using naïve deflation, and more than 400% improvement vs emPCA.

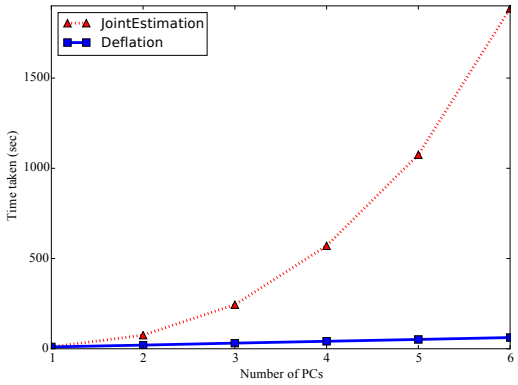


Figure 5: Time taken for estimating top-6 components using traditional joint estimation vs estimation one component at a time using deflation

variants. We note that the application of sparse PCA is for illustrating the flexibility and power of our proposed method, as it can be readily applied to *any* structured probabilistic PCA model with a minor modification. The proposed approach enables large-scale applications of probabilistic PCA where one is interested in the top few (and potentially structured) components given high-dimensional data. Our approach enables such components to be efficiently computed in a serial fashion, interleaved with deflation. We showed that the resulting components regain the interpretability and decomposability of optimization based techniques, while retaining the rich prior structure of probabilistic techniques. Experimental results demonstrate the utility of the proposed deflation approach. We showed that using proper deflation improves the variance explained over several existing methods which use naïve deflation. Further we presented an experiment for illustrating the importance of sequential estimation for scalability. We also showed the equivalence between various probabilistic decomposition techniques and their deterministic counterparts. We are also interested in developing analogues of the proposed techniques for non-linear decomposition.

References

[1] Alexandre d'Aspremont, Francis R. Bach, and Laurent El Ghaoui. Full regularization path for sparse principal component analysis. In *ICML*, pages 177–184, 2007.

[2] Gene H Golub and Charles F Van Loan. *Matrix computations*, volume 3. JHU Press, 2012.

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

[4] Harold Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417, 1933.

[5] Ian Jolliffe. *Principal component analysis*. Wiley Online Library, 2002.

[6] 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.

[7] 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.

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

[9] Oluwasanmi Koyejo. Constrained relative entropy minimization with applications to multitask learning. *PhD Thesis*, 2013.

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

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

[12] Solomon Kullback. Information theory and statistics, 1959.

[13] Lester W Mackey. Deflation methods for sparse pca. In *Advances in neural information processing systems*, pages 1017–1024, 2009.

[14] 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.

[15] 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.

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

[17] K. Pearson. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2(6):559–572, 1901.

[18] 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.

[19] Václav Šmídl and Anthony Quinn. On bayesian principal component analysis. *Computational statistics & data analysis*, 51(9):4101–4123, 2007.

[20] 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.

[21] 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.

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

[23] 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.

[24] Dejiao Zhang and Laura Balzano. Global convergence of a grassmannian gradient descent algorithm for subspace estimation. *arXiv:1506.07405*, 2015.

[25] Fuzhen Zhang. *The Schur complement and its applications*, volume 4. Springer Science & Business Media, 2006.

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