A Deflation Method for Probabilistic PCA

Rajiv Khanna UT Austin

Joydeep Ghosh UT Austin

Russel Poldrack Stanford University

Oluwasanmi Koyejo Stanford University

Abstract

Modern treatments of principal component analysis often focus on the estimation of a single factor under various structural assumptions or priors e.g. sparsity and smoothness, then the procedure is extended to multiple factors by sequential estimation interleaved with deflation. While prior work has highlighted the importance of proper deflation for ensuring the quality of the estimated factors, 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. Using tools recently developed for constrained probabilistic estimation via information projection, we propose a deflation method for probabilistic principal component analysis. The factors 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. Experimental evaluation on neuroimaging data show that deflation leads to improved interpretability and can improve variance explained by each factor.

1 Introduction

Principal Component Analysis (PCA) is a well known technique for data exploration and dimensionality reduction [3]. The goal of PCA is to represent a centered data matrix as a linear combination of few basis vectors. In the classical deterministic setting, factors 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 other domain structure, or are designed to incorporate useful statistical properties such as noise tolerance in high dimensions [4, 16, 1, 5, 11] .

Most modern treatments of principal component analysis and its extensions focus on the estimation of a single factor, leaving multi-factor extensions to sequential estimation interleaved with deflation. Informally, the purpose of deflation is to minimize the influence of previously computed factors on subsequent factors, most often by assuming that subsequent factors are mutually orthogonal. Mackey [10] investigated the effect of deflation choices on the quality of inferences from sparse PCA, showing that careless deflation did not preserve orthogonality and could lead to pathological results such as estimating the same factor multiple times without explaining additional variance. Solving for factors one at a time is more than a mere convenience, as sequential estimation may be necessary to enable scalability for modern "big data" problems. Further, selecting the appropriate number of factors (the rank) via sequential estimation avoids the significant computational overhead of re-estimating all the factors each time the rank is changed, which is required without proper deflation. Several authors have explored probabilistic variants of principal components analysis [2, 12, 6]. Despite the rich prior literature on PPCA, research has primarily focused on batch inferences and do not incorporate notions of proper deflation. Further, proposed techniques for deflation have only focused non-probabilistic principal component analyses, and are not trivially extended to probabilistic analyses. In this manuscript, we seek to bridge this gap in the literature by highlighting issues that may occur with improper deflation. As a remedy, we propose a deflation method for probabilistic treatments of principal components analysis.

Our contributions are as follows: (1) we propose a novel deflation technique for probabilistic PCA via information projection of the means of subsequent factors to orthogonal subspaces based on recent techniques for probabilistic estimation subject to constraints via information projection by Koyejo et al. [9]; (2) we explore an application of the proposed deflation approach to *sparse* probabilistic PCA by extending a recent technique for sparse submodular probabilistic PCA [6]; (3) 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 (this discussion is presented in the supplement).

Experimental evaluation on neuroimaging data shows that deflation leads to improved interpretability and can improve variance explained by each factor.

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 †. The i^{th} row of a matrix $\mathbf{M}$ is indexed as $\mathbf{M}_{i,\cdot}$, while j^{th} column is $\mathbf{M}_{\cdot,j}$. Sets are represented by sans serif fonts e.g. S. For a vector $\mathbf{u} \in \mathbb{R}^d$, and a set S of indices with $|\mathsf{S}| = k, k \le d$, $\mathbf{u}_{\mathsf{S}} \in \mathbb{R}^k$ denotes subvector of $\mathbf{u}$ supported on S.

2 Information Projection onto Subspaces

In this section, we illustrate the use of a recent technique of information projection to restrict a distribution to a subspace in the Euclidean space. These results are applied for deflation in probabilistic PCA by restricting the support of subsequent factors to be orthogonal to the subspace spanned by means of previously extracted factors. For completeness, we present the relevant background in the supplement.

Let $\mathcal{M}$ be the target subspace onto which we aim to restrict a probability density p. The following proposition is a special case of a result from Koyejo et al. [9].

Proposition 1. *Define* characteristic function $\phi_{\mathcal{M}} : \mathbb{R}^d \to \mathbb{R}$ *as* $\phi_{\mathcal{M}}(x) = 0$ *if* $x \in \mathcal{M}$*, and* $\phi_{\mathcal{M}}(x) > 0$ *if* $x \notin \mathcal{M}$*. The restriction of the density* p *to a subspace* $\mathcal{M}$ *can be obtained as:*

$$\arg\min \mathrm{KL}(q\|p) \text{ s.t. } \mathbb{E}_q[\phi_{\mathcal{M}}(x)] = 0. \tag{1}$$

2.1 Information Projection of Gaussians onto Subspaces

The special case where p is a Gaussian is of particular to our development of deflation for (sparse) PPCA. Let $\mathcal{N}(\boldsymbol{\mu}, \mathbf{S})$ represent a multivariate Gaussian distribution with mean $\mu \in \mathbb{R}^d$ 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 $\mathrm{P}_{\mathcal{M}}$. The characteristic function of the set $\mathcal{M}$ is given by $\phi_{\mathcal{M}}(x) := x^\dagger \mathrm{P}_{\mathcal{M}_\perp} x$. It is clear that $x \in \mathcal{M} \implies \phi_{\mathcal{M}}(x) = 0$ and $x \notin \mathcal{M} \implies \phi_{\mathcal{M}}(x) > 0$.

When p is Gaussian, it is known that the information projection onto $\mathrm{P}_{\mathcal{M}}$ is also a Gaussian distribution [8, 7]. We emphasize that this is not an assumption, but rather a property of information projection. Thus, the search for the information 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 (1) can be expanded as $\mathbb{E}_q[\phi_{\mathcal{M}}(x)] = 0 \implies \mathrm{tr}\big(\mathrm{P}_{\mathcal{M}_\perp}\mathbf{a}\mathbf{a}^\dagger + \mathrm{P}_{\mathcal{M}_\perp}\mathbf{B}\big) = 0 \implies \mathrm{tr}\big(\mathrm{P}_{\mathcal{M}_\perp}\mathbf{a}\mathbf{a}^\dagger\big) = 0$ and $\mathrm{tr}(\mathrm{P}_{\mathcal{M}_\perp}\mathbf{B}) = 0$ (since all projection matrices are positive semidefinite). Expanding Equation 1 using definition of KL divergence between two multivariate Gaussian distributions results in the decoupled optimization problems:

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

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

As outlined in [8, 7], 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 $q^* \sim \mathcal{N}(\mathbf{a}^*, \mathbf{B}^*)$.

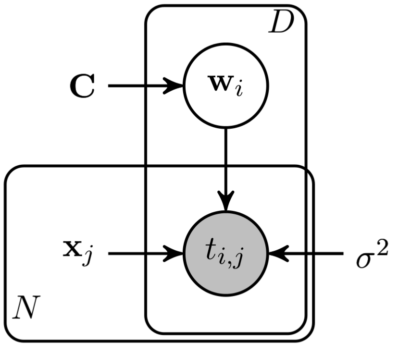
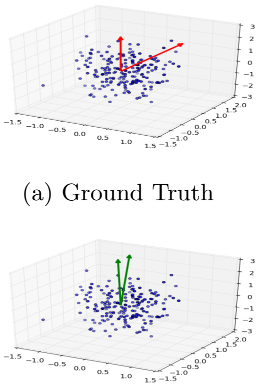


Figure 1: Plate model for Probabilistic PCA. The matrix $\mathbf{C}$ is the prior design matrix.



(a) Ground Truth

(b) Estimated model

Figure 2: Simulated data example showing incorrect estimates using the naïve deflation.

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 zero mean. 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, $\mathbf{C}$, $\mathbf{I}$). The observation model is $\mathbf{T} = \mathbf{XW} + \boldsymbol{\epsilon}$, 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 factors and sample the loadings 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 [13] sequentially using the naïve deflation based on the estimated expected factor. As shown in Figure 2, the procedure estimated degenerate factors 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 (joint estimation of all factors) or with the proposed deflation that enforces orthogonality, where the correct factors were recovered. Fitting a full model however is less scalable, and loses on the opportunity of run time model selection. The interpretation of individual factors 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 [15].

3.1 Orthogonal Deflation

Probabilistic PCA is typically solved by an EM algorithm. We modifiy the inference by restricting the E-step. 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. Let $\mathcal{M}^i$ be subspace spanned by means of first i factors i.e. the subspace spanned by $\bigcup_{j=1}^{i} \mathbb{E}[\mathbf{W}_{i.,}]$. We restrict the support of factor $(i+1)$ to be $\mathcal{M}^i_\perp$.

Let $\mathbf{Z}_i = \mathbf{T} - \Sigma_{j<i}\mathbf{X}_{.,j}\mathbb{E}[\mathbf{W}_{j,.}]$. The variational E-step update for the i^{th} factor is given by $q_i(\mathbf{W}_{i,.}) \sim \mathcal{N}(\mathbf{m}_i, \mathbf{\Sigma}_i)$ where:

$$\mathbf{\Sigma}_i^{-1} = \mathrm{P}_{\mathcal{M}_\perp^{(i-1)}} \left(\frac{1}{\sigma^2} (\mathbf{X}^\dagger_{.,i}\mathbf{X}_{.,i})\mathbf{I} + \mathbf{C}^{-1} \right) \mathrm{P}_{\mathcal{M}_\perp^{(i-1)}}, \;\; \mathbf{m}_i = \frac{1}{\sigma^2} \mathbf{\Sigma}_i \mathrm{P}_{\mathcal{M}_\perp^{(i-1)}} \mathbf{Z}_i^\dagger \mathbf{X}_{.,i}. \qquad (2)$$

The M-step is also straightforward to derive and is presented in the supplement. We term the resulting procedure of sequential estimation and deflation Orthogonal Probabilistic PCA (oPPCA).

3.2 Deflation for Sparse Probabilistic PCA (soPPCA)

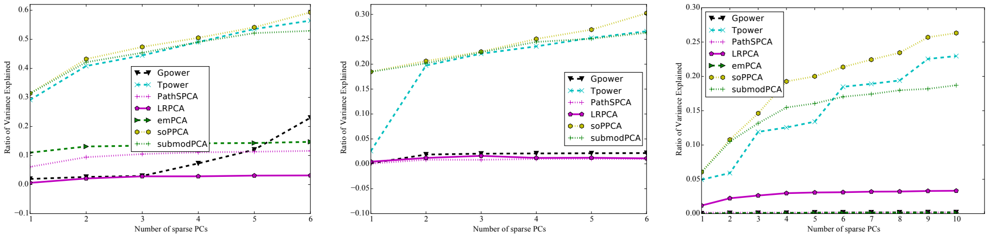
The proposed deflation may also be extended to sparse Probabilistic PCA, where the support of factors is to be restricted to a few dimensions. We focus on the approach proposed by Khanna et al. [6] as it directly utilizes information projection to impose sparsity on the factors, and is a special case of our framework for restriction to subspaces. Thus, for factor 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. Khanna et al. [6] show that a greedy search for the best k_i dimensions is efficient by exploiting supermodularity of the cost function to be minimized. Let $\mathcal{S}_i^*$ be the support set selected for factor i in this fashion. Combining with the orthogonal deflation approach developed in 3.1, the resulting variational E-step is solved as $\bar{q}_i \sim \mathcal{N}(\mathbf{c}_i, \mathbf{D}_i)$:

$$(\mathbf{D})^{-1} = \mathrm{P}_{\mathcal{S}_i^*} \mathbf{\Sigma}_i^{-1} \mathrm{P}_{\mathcal{S}_i^*}, \;\; \mathbf{c}_i = \mathbf{D} \mathrm{P}_{\mathcal{S}_i^*} \Sigma_i^{-1} \mathbf{m}_i \qquad (3)$$

We term the overall procedure sparse orthogonal probabilistic PCA (soPPCA). More details are in the supplement.

4 Experiments

In this section we present empirical results to illustrate the utility of orthogonality in probabilistic PCA models in practice. The details of the preprocessing the data are presented in the supplement. For the three fMRI datasets, we compare the ratio of variance explained by first 6 sparse components to the total variance in the dataset. For $d = 100, 1000, 10000$, each sparse component has sparsity $k = 10, 10, 60$, respectively. To illustrate the predictive power 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 [5] (Gpower), PCA via Low rank [11] (LRPCA), Truncated Power Method [14] (Tpower), Full Regularized Path Sparse PCA [1] (PathSPCA), emPCA [12], submodPCA [6]. We plot the ratio of explained variance along with all the above mentioned methods. Figure 3 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.



(a) d=100

(b) d=1000

(c) d=10000

Figure 3: Performance on fMRI Data Out of Sample

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

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

[4] 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.

[5] 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.

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

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

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

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

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

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

[12] 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.

[13] 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.

[14] 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.

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

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

Supplement: A Deflation method for Orthogonal Probabilistic PCA

1 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 factors $\mathbf{W} \in \mathbb{R}^{r \times d}$ and loadings $\mathbf{X} \in \mathbb{R}^{n \times r}$.

In the classical deterministic setting, factors 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:

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

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 (1) at the solution. To obtain the next factor, the covariance matrix is *deflated* to remove the variance component explained, then (1) is re-solved with the deflated covariance. Using Hotelling's deflation [2], the subsequent covariance matrix at step $i+1$ is computed from the i^{th} covariance matrix as:

$$\mathbf{\Sigma}_{i+1} = \mathbf{\Sigma}_i - \mathbf{w}_i \mathbf{w}_i^\dagger \mathbf{\Sigma}_i \mathbf{w}_i \mathbf{w}_i^\dagger, \tag{2}$$

where $\mathbf{\Sigma}_0 = \mathbf{\Sigma}$ and $\mathbf{w}_0$ is the first factor. Alternatives to Hotelling's deflation include Schur complement deflation and orthogonal projection (see [6] for more details).

While the covariance approach is perhaps the most popular, an alternative and equivalent approach is to estimate both the factors and the loadings $\mathbf{s} \in \mathbb{R}^d$ to minimize the reconstruction error Pearson [9] as:

$$\min_{\mathbf{x},||\mathbf{w}||_2=1} ||\mathbf{T} - \mathbf{x}\mathbf{w}^\dagger||_F^2. \tag{3}$$

The optimal $\mathbf{w}$ is given by the left singular vector of the data matrix which is associated with the largest singular value and $\mathbf{x}$ is the corresponding right singular vector multiplied by the singular value. The associated *variance explained* can be computed using the same equation as the covariance approach (1) at the solution. The reconstruction error view suggests the naïve deflation for the subsequent factors by replacing the data matrix with the residual in (2.1), where the residual is given by:

$$\mathbf{T}_{i+1} = \mathbf{T}_i - \mathbf{x}_i \mathbf{w}_i^\dagger, \tag{4}$$

where $\mathbf{T}_0 = \mathbf{T}$ and $\mathbf{x}_0$ is the *first* loading 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 factorization can then be obtained by maximizing the log likelihood, typically by EM algorithm to solve for all r factors at the same time.

Related Work on PPCA: Probabilistic PCA was first proposed by Tipping and Bishop [12] based on an extension of the well established factor models in statistics. Tipping and Bishop [12] showed that the result was equivalent to standard PCA under certain choices of hyperparameters, and generalized PPCA to incorporate priors on the loadings. Šmídl and Quinn [11] extended this work to a full Bayesian treatment which included priors on both factors and loadings, and considered the use of a appropriate priors on the factors to enforce orthogonality. Beyond standard PCA, several authors have proposed additional priors to encourage sparsity or non-negativity on the factors[1, 10]. Perhaps the work closest to ours is Khanna et al. [3] who applied the information projection approach for sparse probabilistic PCA. However, like other prior work, Khanna et al. [3] focused on single factor estimation and did not consider deflation. It is clear that the reconstruction error view of classic PCA is equivalent to modeling using the Gaussian likelihood.

Transpose of Tipping and Bishop [12]: We note an important difference between our modeling approach from that of Tipping and Bishop [12]. Tipping and Bishop [12] assumed that the factors are fixed parameters, and the loadings are random variables. For problems of interest, we are more interested in incorporating structural priors on the factors, so we assume that the factors are random variables while the loading are fixed parameters. In practice, both approaches are trivially equivalent by replacing the data matrix by its transpose.

1.1 Constrained Probabilistic Inference via Information Projection

In the interest of a self contained discussion, this section outlines relevant background constrained probabilistic inference via information projection, which will be useful for constructing our proposed deflation technique. We begin with a definition Kullback-Leibler divergence and the the information projection.

Let X represent the sample space of interest. Let $\mathcal{P}$ represent the set of bounded densities supported on X.

Information Projection: Let $p \in \mathcal{P}$ and $q \in \mathcal{P}$, then the the Kullback-Leibler divergence [5] between p and q is defined as

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

Given a set $\mathcal{Q} \subseteq \mathcal{P}$, the *information projection* of $p \in \mathcal{Q}$ 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. Following Koyejo et al. [4], Khanna et al. [3], 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 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: $p_\mathsf{S}(x) = 0$ if $x \notin \mathsf{S}$, $p_\mathsf{S}(x) = \frac{p(x)}{\int_{s\in\mathsf{S}} p(s)ds}$ if $x \in \mathsf{S}$.

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, and provides an alternative way to estimate the restriction

Lemma 1 (Koyejo et al. [4]). *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 Iterated Projections

Theorem 2 (Iterated Information Projection [4]). *Let $\pi : [n] \mapsto [n]$ be a permutation function and $\{\mathsf{C}_{\pi(i)} \mid \mathsf{C}_{\pi(i)} \subset \mathsf{X}\}$ represent a sequence of sets with non empty intersection $\mathsf{B} = \bigcap \mathsf{C}_i \neq \emptyset$. Given a base density p, let $q_0 = p$, and define the sequence of information projections:*

$$q_i = \operatorname*{arg\,min}_{q\in\mathcal{F}_{\mathsf{C}_{\pi(i)}}} \mathrm{KL}(q\|q_{i-1}),$$

then $q_* = q_N$ is independent of π. Further:

$$q_* = \min_{q\in\mathcal{F}_\mathsf{B}} \mathrm{KL}(q\|p).$$

2.1 Inference for Probabilistic PCA via Variational EM

Probabilistic PCA is typically solved by an EM algorithm. The EM obviates construction of the full covariance matrix, and instead enables working with the data matrix while returning both the loadings and factors. Expectation Maximization can be described using the free energy interpretation given by Neal and Hinton [8]. Maximizing the negative log-likelihood can be shown to be equivalent to maximizing a free energy function $\mathscr{F}$ (see Equation 5). 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 6), and the M-step can be interpreted to be the search over the parameter space, keeping the latent variables distribution q fixed (Equation 7). The cost function for the EM is given by [8]:

$$\mathscr{F}(q(\mathbf{W}), \Theta) = -\mathrm{KL}(q(\mathbf{W}) \| p(\mathbf{W}|\mathbf{T};\Theta)) + \log p(\mathbf{T};\Theta). \tag{5}$$

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

$$\text{E-step:} \qquad \max_{q} \mathscr{F}(q(\mathbf{W}), \Theta), \tag{6}$$

$$\text{M-step:} \qquad \max_{\Theta} \mathscr{F}(q(\mathbf{W}), \Theta). \tag{7}$$

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 Equation 6 returns the posterior $p(\mathbf{W}|\mathbf{T};\Theta)$. Variational methods perform the search for best q over a constrained set [13] using constrained KL minimization. Let D be the set of distributions over $\mathbf{W}$ that fully factorize over individual rows 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. More commonly, 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 factors $q(\mathbf{W}_{\setminus i})$ when estimating the distribution of the new factor $q(\mathbf{w}_i)$. Following the E-step 6, the effect on the model is straightforward to compute as (up to additive and multiplicative constants):

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

With factors $j < i$ fixed, it is clear that this is equivalent to the naïve deflation of (4) using the estimated posterior mean.

3 Reduction of oPPCA to PCA

The naïve 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 the deflation equation, we compute:

$$\mathbf{Z}_i = \mathbf{T}(\mathbf{I} - \Sigma_{j<i}\alpha_j\mathbf{m}_j\mathbf{m}_j^{\dagger}) \tag{8}$$

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

Proposition 3. *If $\mathbf{C} = \mathbf{I}$ the means of factors estimated by oPPCA correspond to the factors estimated by deterministic PCA.*

Proof Sketch. 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 see that solving for the first factor $\mathbf{m}_1$ is equivalent to performing power iterations on $\mathbf{T}^{\dagger}\mathbf{T}$. For the subsequent factors, solving for $\mathbf{m}_i$ is equivalent to performing power iterations on the deflated matrix $(\mathbf{I} - \Sigma_{j<i}\alpha_j m_j m_j^{\dagger})\mathbf{T}^{\dagger}\mathbf{T}(\mathbf{I} - \Sigma_{j<i}\alpha_j m_j m_j^{\dagger})$. □

3.1 Deflation for Sparse Probabilistic PCA (soPPCA)

The proposed deflation using the framework may also be extended to sparse Probabilistic PCA, where the support of factors is to be restricted to a few dimensions. We focus on the approach proposed by Khanna et al. [3] as it directly utilizes information projection to impose sparsity on the factors, and is a special case of our framework for restriction to subspaces. Thus, for factor 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. Khanna et al. [3] show that a greedy search for the best k_i dimensions is efficient by exploiting supermodularity of the cost function to be 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} factor. The variational E-step for sparse factor $\mathbf{W}_{i,.}$ is given by:

$$\min_{\substack{\text{Supp}(\bar{q}_i(\mathbf{W}_{i,.})) \in (\mathrm{P}_{M_{\perp}^{(i-1)}} \cap \mathcal{S}) \\ \mathcal{S} \in \mathcal{S}_{k_i}}} \mathrm{KL}\big(\bar{q}_i(\mathbf{W}_{i,.}) \| \bar{p}_i(\mathbf{W}_{i,.} | \mathbf{Z}_i; \mathbf{X}_{.,i}, \sigma^2)\big). \tag{9}$$

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 supplement and [4] for details). Following the optimization for the support dimensions $\mathcal{S}_i^*$ (which is solved greedily owing to the supermodularity), and combining with the proposed orthogonal deflation, the resulting variational E-step is solved as $\bar{q}_i \sim \mathcal{N}(\mathbf{c}_i, \mathbf{D}_i)$:

$$(\mathbf{D})^{-1} = \mathrm{P}_{\mathcal{S}_i^*} \mathbf{\Sigma}_i^{-1} \mathrm{P}_{\mathcal{S}_i^*}, \ \mathbf{c}_i = \mathbf{D} \mathrm{P}_{\mathcal{S}_i^*} \Sigma_i^{-1} \mathbf{m}_i \tag{10}$$

We term the overall procedure sparse orthogonal probabilistic PCA (soPPCA).

4 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 [14] for details and recovery guarantees). The following proposition shows an equivalence between a single factor from soPPCA and the truncated power method.

Proposition 4 (Reduction to the truncated power method)**.** *If* $\mathbf{C} = \mathbf{I}$*, the normalized mean of the factor* $\mathbf{m}_1$ *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 reduces to (by combining E-step and M-step, and ignoring scaling constants since they vanish when normalizing):

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

□

Orthogonal projection deflation of the covariance matrix involves a Gram-Schmidt procedure to build orthogonal set of factors from possibly non-orthogonal ones obtained after projection deflation [6]. The following corollary shows an equivalence between soPPCA and the truncated power method with orthogonal projection covariance deflation.

Corollary 5. *If* $\mathbf{C} = \mathbf{I}$*, the means of the factors estimated by soPPCA recover the sparse eigenvectors obtained by the truncated power method with orthogonal projection deflation.*

Proof Sketch. Follows from Proposition 4, the projection deflation formula 8 and the fact that projection deflation with truncated power method is equivalent to orthogonal projection deflation. □

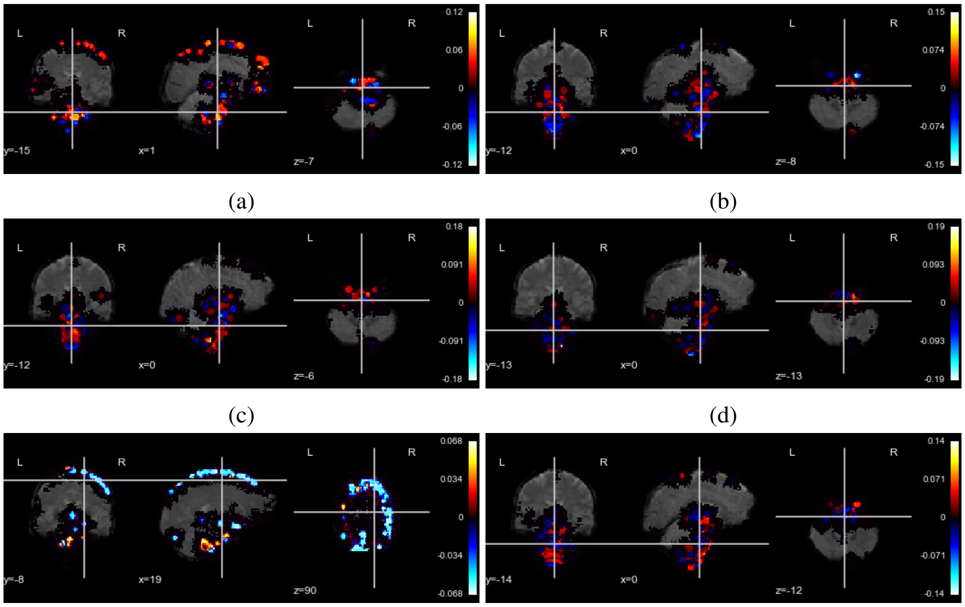
Table 1: Ratio of variance explained for first 10 factors

	PPCA+deflation by subtract.	oPPCA
	0.036	0.085

5 Extra experiments

5.1 Simulated data

To validate the oPPCA model, we generate simulated data as follows. We fix n=1000, d=10000. We fix the low rank r=100, and generate low rank factors and loadings from normal gaussian. We take the outer product of these factors and loadings to get an n X d matrix of rank r. We run PPCA with standard subtraction deflation, and compare with oPPCA for 10 factors. The results are in Table 1. As expected, using oPPCA which restricts the support of subsequent factors to be orthogonal to already selected factors gives much better variance explained.



(a) (b)

(c) (d)

(e) (f)

Figure 1: Brain plots of few factors extracted from fMRI data. The top-6 extracted factors are consistent with the motion artifacts similar to those obtained by ICA during fMRI processing.

6 fMRI data details

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. We cluster the original set of dimensions to fewer dimensions using the spatially constrained Ward hierarchical clustering approach of [7], 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 $\mathbf{W}$. $\mathbf{C}$ is obtained as covariance corresponding to MRF with neighboring voxels connected with unit weight.

References

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

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

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

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

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

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

[7] 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.

[8] 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.

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

[10] 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.

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

[12] 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.

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

[14] 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.