

---

# Givens Transform Approach for Efficient Probabilistic Principle Component Analysis for Bayesian Dimensionality Reduction (GT-PPCA)

---

Anonymous Author(s)

Affiliation

Address

email

## Abstract

1 We develop scalable and flexible Probabilistic Principal Component Analysis  
2 (PPCA) methods for determining posterior distributions of spanning frames based  
3 on a Givens Representation of the PCA which we term (GT-PPCA). This addresses  
4 significant challenges that arise with latent variable in a traditional formulation of  
5 PPCA. For sampling posterior distributions we develop Hamiltonian Monte-Carlo  
6 Methods (HMC) for sampling on the Stiefel Manifold the PCA orthogonal frame  
7 sets. We demonstrate our approach on several challenging example problems  
8 including tests problems XYZ and problems arising in our recent work on under-  
9 standing medical patient data associated with coagulopathy (factors influencing  
10 blood clotting). We show our methods provides ways to identify when data sets  
11 contain a mixture of low dimensional structures that would not be resolved with  
12 traditional PCA approaches. We further show how our approach can be used to  
13 develop heirarchical models in terms of low dimensional structures learned from  
14 the data sets or to develop prior distributions useful in generalizing low dimensional  
15 structures to new settings. To facilitate use of our GT-PPCA method we provide a  
16 package with the widely-used Stan statistics package.

## 17 1 Introduction

18 Principal Component Analysis (PCA) is a widely used dimensionality-reduction tool for exploratory  
19 analysis and modeling in both the natural and social sciences. By factorizing an empirical covariance  
20 matrix into a product of low rank matrices, PCA effectively finds a low dimensional subspace that  
21 describes the dataset in terms of the latent factors. These latent factors are given by the columns of  
22 the low rank factorization. Geometrically, traditional PCA can be interpreted as providing a point  
23 estimate of a low dimensional hyper-plane that is closest to a cloud of data points. PCA and other  
24 variants are also routinely applied to binary or integer valued matrices, such as item response tables  
25 in recommendation systems or graph adjacency matrices arising in network science [7, 4].

26 Probabilistic PCA (PPCA) [13] posits a probabilistic generative model that is equivalent to PCA in the  
27 limit of decreasing noise [12, chapt. 12.2]. This probabilistic approach is attractive because it enables  
28 a straightforward methodology, via Bayesian inference, to quantify the uncertainty in our estimates  
29 (to prevent overfitting) and conduct hypothesis testing. For example, given high dimensional medical  
30 data for patients with two different type of injuries, it would be desirable to find posterior distributions  
31 of low-dimensional subspaces that describe the data, then find the probability given the data that these  
32 subspaces are different for the two groups of patients. While uncertainty quantification of the latent  
33 factors in PPCA has been explored in the literature [? 1, 2], there are currently no out-of-the-box  
34 solutions available to researchers.

In addition to enabling uncertainty quantification and hypothesis testing, probabilistic models are amenable to expansion and can serve as modules within larger probabilistic graphical models. This is important in real-world settings where we seek to utilize any known prior information in our inference or when true generative models do not necessarily follow the simple generative process set forth by PPCA. If we believe our latent factors to be sparse, we can add Laplace or Cauchy priors to our PPCA model yielding a probabilistic sparse PCA[? ]. Following the medical example, we may believe subspaces for different groups of patients come from some common prior distribution of subspaces, in which case we can build a hierarchical model to do transfer learning. Similarly, we can expand the PPCA graphical model to conduct non-linear dimensionality reduction via Mixtures of Factor Analyzers [? ]. To handle binary or discrete data we can expand PPCA using a link function as in Bayesian Exponential Family PCA (BXPCA) [10].

While expanded PPCA models have shown promise on a variety of problems, they have not been fully explored because their implementation remains elusive, and most inference schemes such as Expectation Maximization (EM) only provide point estimates. The availability of PPCA in a simple framework for building probabilistic graphical models like Stan [3] would allow rapid building and prototyping of such models in a fully Bayesian way that provides uncertainty around any point estimates.

**Bayesian inference of orthonormal matrices** Many of the difficulties in conducting full Bayesian inference on PPCA and related models stem from having to infer one or more unknown orthonormal matrix parameters. This is difficult because orthonormal matrices form a rather particular subset (or submanifold) of all possible realizable matrices, analogous to three-dimensional unit vectors lying on the sphere (a submanifold of  $\mathbb{R}^3$ ). This yields a probability distribution on a sub-manifold within the full space of matrices that has a non-trivial geometry. More specifically, the prior and posterior distributions of an orthonormal matrix  $W$  must have support over the set of  $n \times p$  orthonormal matrices. This is known as the Stiefel manifold and denoted  $V_{n,p}$  [11].

When first considering this problem, it may seem that this geometry may rule out a straightforward use of two of the most prominent techniques for posterior inference posterior inference, Markov Chain Monte Carlo (MCMC) and Variational Inference (VI). Intuitively, for MCMC, we have no way of guaranteeing that a chain over  $W$  will explore only valid regions of parameter space that satisfy the orthonormality constraints. Similarly for Variational Inference, positing a common variational posterior distribution such as a Gaussian over the elements of  $W$  is sure to lead to posteriors that assign mass to invalid regions of parameter space.

**Transformed random variables** Posterior distributions for constrained parameters in probabilistic graphical models are routinely inferred by transforming such parameters to an unconstrained space and seeking posterior distributions over the transformed parameter [3? ]. This requires a smooth one-to-one transformation  $f : \text{supp}(z_{\text{constr}}) \rightarrow \mathbb{R}^D$ , where  $\text{supp}(z_{\text{constr}})$  is the support of the constrained random variable  $z_{\text{constr}}$ . To our knowledge, no such transformation has been proposed to map orthonormal matrices to a comparable unconstrained space.

In this paper we draw on techniques from Differential Geometry and Numerical Analysis to introduce a novel and geometrically elegant way to represent orthonormal matrices. In our approach we express orthonormal matrices in terms of a sequence of fundamental rotations through given angles. This gives insight into the geometry of the Stiefel manifold, and results in a transform we call the Givens Transform, that maps orthonormal matrices to an unconstrained space. We apply the Givens Transform to inference of PPCA-based models, and collectively refer to this as GT-PPCA.

**GT-PPCA** GT-PPCA is straightforward to implement in stand-alone inference schemes, but is particularly useful in the context of probabilistic programming framework like Stan [3], where we can use it for uncertainty quantification and hypothesis testing of PPCA models, as well as extending PPCA to more complex probabilistic graphical models, two previously intractable tasks. We provide Stan code for our example models allowing for use and expansion by researchers and scientists out-of-the-box. We demonstrate how inference of these models in Stan yields good empirical performance on large probabilistic graphical models that were previously intractable to implement especially if fully-Bayesian posterior analysis is desired. Specifically we present a hierarchical subspace model for grouped, multi-view medical data and a PPCA Hidden Markov Model (HMM) for disease-network data.

In addition to opening the door for straightforward implementation of large PPCA models, GT-PPCA yields insight in to novel and useful ways to work with and interpret our models. For instance, the elegant geometric representation lets us see how by limiting the range of the parameters in GT-PPCA, we can naturally avoid issues of unidentifiability and multi-modal posteriors that arise in other methods. GT-PPCA also allows us new and creative ways to generate and use prior distributions on orthonormal matrices. In the setting of using the matrix directly this task has previously been rather complicated and rather intractable for even small problem sizes. This is linked to the difficulty of evaluating densities of orthonormal matrix distributions in other representations. As we shall discuss in more detail, our GT representation provides a rather natural way to specify prior distributions comparable to the Matrix Langevin prior [11].

**Related work** While previous authors have developed methods for posterior sampling of distributions orthonormal matrices, these methods can at times suffer from numerical issues, they are difficult to implement on large probabilistic graphical models, and they can not be used in any general inference scheme such as VI or Maximum A-Posteriori (MAP) estimation like GT-PPCA can. Brubaker et al. [1] and Byrne and Girolami [2] used separate approaches to modify the Leap-Frog integrator typically used in Hamiltonian Monte Carlo (HMC), so that Hamiltonian exploration, and thus MCMC samples of posteriors, satisfied any necessary constraints at all times. Specifically, Brubaker et al. [1] uses the SHAKE integrator [8] to simulate Hamiltonian dynamics and generate proposals. The integrator works by repeatedly taking a step forward that may be off the manifold using ordinary leap frog, then projecting back down to the nearest point on the manifold. This projection is done via Newton iterations, which may converge to the wrong local minimum in practice or perhaps not converge at all, possibly jeopardizing the ergodicity of a Markov Chain, and the integrity of samples [? ]. Byrne and Girolami [2] took a different approach, exploiting the fact that closed form solutions are known for the geodesic equations over the Stiefel manifold in the embedded coordinates,  $W$ . While this method is completely explicit, requiring no Newton iterations, in practice we found that for larger step sizes, the integrator steps off the Stiefel manifold, due to the numerical imprecision of the matrix exponential function. Because these methods use modified integrators for constrained parameters, in practice they require keeping track of the support of each variable and which type of integrator to use on each variable. This adds an extra layer of implementation complexity, especially for large complex probabilistic graphical models, that makes it difficult to implement these methods within a probabilistic programming language such as Stan. This precludes the rapid prototyping and building of models as well as the flexibility to use different inference schemes that Stan provides. Lastly, we remark that for inference on orthonormal matrices, these methods can lead to multi-modal posteriors, that can be avoided in a straight-forward way using the Givens transform.

**Paper outline** We give a brief overview of probabilistic dimensionality reduction in Section 2. We discuss the geometry of the Stiefel Manifold in Section 3, before finally introducing the Givens Transform (GT) in Section 4. Finally, we present various empirical studies where we used GT-PPCA in Stan for practical uncertainty quantification and hypothesis testing, as well as for building complex probabilistic graphical models.

## 2 Probabilistic Principle Component Analysis (PPCA)

In probabilistic principle component analysis (PPCA) one starts by considering a collection of data points in a typically high-dimensional vector space and seeks to find a posterior distribution over a reduced representations of the data in the form of a lower dimensional subspace. The central postulate is that for a data vector  $\mathbf{x} \in \mathbb{R}^n$  there exists an unknown low-dimensional latent representation  $\mathbf{z} \in \mathbb{R}^p$  where  $p < n$ , (ideally with  $p \ll n$ ). The two representations are related to each other by a single unknown linear transformation  $\mathbf{x} \rightarrow \mathbf{z}$ . Mathematically, we consider a finite collection of sampled data vectors  $\mathbf{x}_i \in \mathbb{R}^n$ ,  $i = 1, \dots, N$  and try to estimate this subspace. Formally, PPCA consists of the following generative process

$$\begin{aligned} p(\mathbf{z}_i) &\sim \mathcal{N}_p(\mathbf{0}, I) \\ p(\mathbf{x}_i | \mathbf{z}_i, W, \Lambda, \sigma^2) &\sim \mathcal{N}_n(W\Lambda\mathbf{z}_i, \sigma^2 I). \end{aligned} \quad (1)$$

137 The  $W$  is an  $n \times p$  orthonormal matrix and  $\Lambda$  is a  $p \times p$  diagonal matrix with positive elements. For  
 138 simplicity in our presentation of PPCA, we have assumed here that the data has only zero mean but  
 139 the more general case can also readily be considered [12, chapt. 12.1].

140 **Quantifying uncertainty** Inference for PPCA is typically conducted by obtaining a point estimate  
 141 for  $W$  via a closed-form estimator, or for expanded PPCA models via (EM) [12, chapt. 12.2], neither  
 142 of which provide a notion of uncertainty for our point estimates. Without information regarding  
 143 the uncertainty of our estimates these point estimates could be far from the true value of  $W$  and  
 144 thus mislead our conclusions, especially for larger models and/or when there is relatively little data  
 145 available. Furthermore, point estimates do not allow for hypothesis testing e.g. statistically testing  
 146 whether two different groups of observations lie in the same subspace. We show with examples how  
 147 GT-PPCA in Stan makes it easy to achieve these tasks as we show in section 5.

148 **Expanding models** As alluded to previously, PPCA generative model can be flexibly expanded  
 149 in several ways as modelers see fit. To build a probabilistic sparse PCA, one can place a Laplace  
 150 or Cauchy prior over the elements of  $W$ . If we have meaningfully grouped data, such as data from  
 151 hospital patients with different types of injury, it might be desirable to designate a separate  $W$   
 152 parameter (subspace) for each group, then place prior over these subspaces to garner the benefits of  
 153 hierarchical modeling [?, chapt. 5]. Mohamed et al. [10] showed that we can model non-Gaussian  
 154 data,  $\mathbf{x}_i$ , by replacing equation 1 with an exponential family member whose natural parameters are  
 155 given by  $\text{Expon}(W\Lambda\mathbf{z}_i)$  where  $\text{Expon}(\cdot)$  is an appropriate link function. Again, in the context of a  
 156 probabilistic programming language such as Stan, these extensions to the base PPCA model become  
 157 trivial to implement as we illustrate with examples in section 5.

158 **Importance of the Orthonormality Condition** The orthonormal constraint on the matrix  $W$   
 159 plays an important role in obtaining robust methods for making inferences in probabilistic PCA  
 160 because it alleviates identifiability and numerical issues. If one were to relax the orthonormality  
 161 constraint the likelihood function would assign identical probability to a whole equivalence class  
 162 of matrices  $W \sim V$  where the span is the same linear subspace  $\text{span}\{W\} = \text{span}\{V\}$  Murphy [12,  
 163 chapt. 12.1.3]. Besides resulting in an unidentifiable model, in practice this presents a number of  
 164 major challenges. This first is that the matrices in a given equivalence class are not all equally well-  
 165 conditioned numerically and round-off errors and truncation errors become problematic in practical  
 166 calculations. Secondly, these issues with the representation further manifest in the log-likelihood  
 167 objective function where regions arise of particularly large curvature as pointed out by [6]. This  
 168 causes significant numerical issues for variational inference (VI) in nonlinear optimization methods  
 169 and in Monte-Carlo (MC) approaches with samplers having slow mixing times [6]. We note that,  
 170 while most identifiability issues and numerical issues are alleviated by constraining inference to  
 171 orthonormal matrices, the PPCA likelihood is equivalent for an orthonormal matrix  $W$  and any  
 172 permutation of the columns of  $W$  being negative as pointed out by both Murphy [12, chapt. 12.1.3]  
 173 and Holbrook et al. [6]. As such, even the methods of Brubaker et al. [1] and Byrne and Girolami [2]  
 174 will lead to multi-modal posteriors, that can be avoided in a straight-forward way by appealing to  
 175 insights revealed by the Givens Transform, as we explain in Section 4.

### 176 3 Geometry of the Stiefel Manifold

177 The set of  $n \times p$  orthonormal matrices  $V_{n,p}$ , form a sub-manifold in the space of general  $n \times p$   
 178 matrices known as the Stiefel Manifold [11] and formally defined as

$$V_{n,p} := \{Y \in \mathbb{R}^{n \times p} : YY^T = I\}. \quad (2)$$

179 Intuitively, the elements of  $V_{n,p}$  can be thought of not as orthonormal matrices, but as  $p$ -frames which  
 180 are comprised of  $p$  orthonormal vectors that lie in  $n$ -dimensional space. To move about the Stiefel  
 181 manifold, one can rigidly rotate the vectors in the  $p$ -frame about any combination of axes an arbitrary  
 182 number of times. In the case where  $n = 3$  and  $p = 2$ , this is almost identical to sphere (Figure 1a),  
 183 but with an extra angle,  $\theta_{23}$  that controls how much the second basis vector is rotated about the first.  
 184 For a three-dimensional set of points forming a flat, pancake-like cloud, PPCA can be thought of as  
 185 finding the best 2-frame that aligns with this cloud.

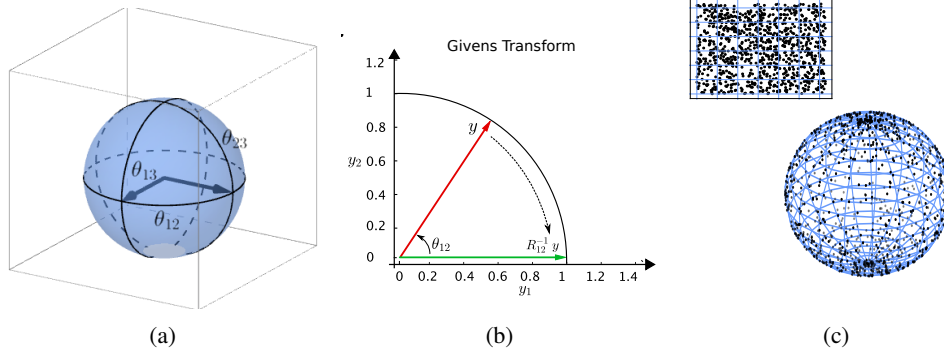


Figure 1: Visualizing the Givens Transform. (a) How the Givens Reduction “zeros out” a column vector. (b) A geometric view of the Stiefel manifold, two-frame in three dimensions. (c) Sampling without a proper measure adjustment. **PJA: I’ve performed edits of the middle figure to illustrate Givens Transform. The left and right figure still need work.**

While  $n \times p$  orthonormal matrices are represented by  $np$  elements, the Stiefel Manifold  $V_{n,p}$ , has an intrinsic dimension of  $np - p(p+1)/2$ . This arises from the constraints on the columns of the matrix that impose orthonormality. This dimensionality can be seen by observing, that the first column of  $Y \in V_{n,p}$  must have norm one and hence has one constraint placed on it. The second column must also have norm one and also must be orthogonal to the first column hence with two constraints placed on it. Continuing to the third column through the  $n^{\text{th}}$  one arrives at the conclusion that each point of the Stiefel Manifold has only  $np - (1 + 2 + \dots + p) = np - p(p+1)/2$  degrees of freedom. The Givens transform can be thought of as an  $np - p(p+1)/2$ -dimensional set of coordinates  $\Theta$ , that represent elements of the Stiefel manifold.

#### 4 Givens Transform (GT) approach to PPCA (GT-PPCA)

For several types of constrained parameters, posterior distributions are in practice rather routinely inferred using both MCMC and VI by transforming the constrained variables to an unconstrained space using a one-to-one mapping  $T : \text{supp}(z_{\text{constr}}) \rightarrow \mathbb{R}^D$ , where  $\text{supp}(z_{\text{constr}})$  is the support of the constrained random variable  $z_{\text{constr}}$ . One can obtain a posterior over the unconstrained parameter that corresponds to the original constrained parameter of interest, then map inferences back to the original constrained space. This procedure requires computing the Jacobian,  $J_{T^{-1}}$  of the transformation, to obtain  $f_Y(y) = f_{z_{\text{constr}}}(T^{-1}(y))J_{T^{-1}}(y)$  where  $Y$  is an unconstrained random variable with probability density function (PDF)  $f_Y$  and  $f_{z_{\text{constr}}}$  is the probability density of  $z_{\text{constr}}$ , which for PPCA comes from equation 1. The extra Jacobian term accounts for how the a unit volume under the transformation changes [? ]. Without this extra Jacobian factor, inference between the two spaces is incomparable. For example, uniformly sampling in spherical coordinates (unconstrained space) does not correspond to uniformly sampling on the sphere (constrained space), unless we include an appropriate term accounting for how volumes are warped under the transformation (see Figure 1c). Conducting inference in a transformed space is most notably used in ADVI and Stan’s HMC routines [3? ]. In sub-section 4.1 we briefly discuss Givens Reductions, motivating the Givens Transform. In sub-section 4.2 we discuss geometric aspects of the Givens transform such as avoiding multi-modality and including a term that measures how volume is changed under the transform that is analogous to the Jacobian described above.

##### 4.1 Givens Reductions and the Givens Transform

The Givens Reduction is a numerical analysis technique for reducing a square matrix  $A$  to upper-triangular form [9]. The technique works by applying a series of rotation of matrices to  $A$  such that elements below the diagonal are “zeroed out” starting with the second element of the first column, and moving down the first column before zeroing out the appropriate elements of the subsequent columns. For example if  $A$  is a  $3 \times 3$  matrix and its first column is  $(0.84, 0.48, 0.26)^T$ , the Givens Reduction would apply a rotation in the  $(1, 2)$ -plane,  $R_{12}^{-1}$  so as to annihilate the second element

221 of this column (Figure 1b). For an  $n \times p$  orthonormal matrix  $Y$ , applying the Givens Reduction  
 222 requires multiplication by  $(n-1) + (n-2) + \dots + (n-p) = np - p(p+1)/2$  rotations matrices  
 223 each with their own respective angles and results in the matrix  $I_{n,p}$ , whose columns are the first  $p$   
 224 standard basis vectors.

$$(R_{pn}^{-1} \dots R_{p,p+2}^{-1} R_{p,p+1}^{-1}) \dots (R_{2n}^{-1} \dots R_{24}^{-1} R_{23}^{-1})(R_{1n}^{-1} \dots R_{13}^{-1} R_{12}^{-1})Y = I_{n,p}, \quad (3)$$

225 From the perspective of  $p$ -frames, the Givens Reductions maps all  $p$ -frames to the canonical  $p$ -frame  
 226  $I_{n,p}$ . Geometrically, this is because if  $Y$  is already orthonormal then applying rotation matrices  
 227 will rigidly rotate all columns of  $Y$  at once preserving their orthogonality, and leaving their length  
 228 unchanged. Because rotations are invertible we can rewrite 3 as

$$Y = (R_{12} \dots R_{1n}) \dots (R_{23} \dots R_{2n})(R_{p+1,n} \dots R_{pn})I_{n,p}. \quad (4)$$

229 which we refer to as the Givens Representation of an orthonormal matrix  $Y$ . Since each of the  $np -$   
 230  $p(p+1)/2$  rotation matrices have an associated angle  $(\theta_{12} \dots \theta_{1n}) \dots (\theta_{23} \dots \theta_{2n})(\theta_{p+1,n} \dots \theta_{pn})$ ,  
 231 that we collectively refer to as  $\Theta$ , we can use these angles to represent any  $n \times p$  orthonormal matrix.  
 232 In this way we have reparameterized all  $n \times p$  orthonormal matrices<sup>1</sup>, a constrained space, in terms  
 233 of unconstrained angles<sup>2</sup>, using a transform  $\Theta : V_{n,p} \rightarrow \mathbb{R}^{np-p(p+1)/2}$ . We refer to 4 as the Givens  
 234 representation or Givens Transform.

## 235 4.2 Geometry of the Givens Transform

236 limiting angles and area form.

237 **PJA: Discuss the basic differential geometry of the Steifel Manifold and how we handle these various**  
 238 **issues. Degenerate regions and metric factors (Jacobians). Discuss how we use "multi-coordinate**  
 239 **charts" when necessary to avoid being close to regions with bad metric factors and degeneracy, etc...**

## 240 5 Empirical Studies

### 241 5.1 Synthetic Data

242 We generated a synthetic, three-dimensional dataset that lies on a two-dimensional plane with  $N = 15$   
 243 observations according to the generative process of PPCA 1. We chose  $\text{diag}(\Lambda) = \text{diag}(1, 1)$ ,  $\sigma^2 = 1$ ,  
 244 and  $W$  to be  $I_{3,2}$  which in the Givens representation corresponds to  $\theta_{12} = \theta_{13} = \theta_{23} = 0$ . This  
 245 example illustrates how simply running GT-PPCA in Stan can alleviate overfitting issues in the  
 246 common use-case where one seeks to carry out dimensionality reduction in a low-observation regime.  
 247 A standard classical PCA analysis yields the singular values  $\text{diag}(\hat{\Lambda}) = (1.52, 1.27, 0.77)$ , possibly  
 248 suggesting that our data lie close to some two-dimensional plane, since the third singular value has a  
 249 larger drop off from the first two than the second has from the first. Figure 2a illustrates geometrically  
 250 a point estimate of the subspace found by PCA. This corresponds to the subspace spanned by the  
 251 PCA point estimates of the latent factor loadings. Because of relatively low signal to noise ratio  
 252 and modest sample size, the point estimate is drastically affected by only a few observations and  
 253 is characteristically different from the flat plane, which we know to be the truth in this case. The  
 254 PCA point estimate  $\theta_{13}$ , which if we recall from Figure 1a is the Givens Transform angle that  
 255 controls the upwards tilt of the plane, is  $\hat{\theta}_{13} = -0.15$ . Meanwhile, posterior HMC samples from  
 256 GT-PPCA in Stan yields a median value of -0.24 and a 95% posterior interval of  $(-1, 0.78)$ . This  
 257 lets us know that there is high uncertainty around our point estimate given the data, and suggests  
 258 that any conclusions drawn from point estimates may be overfit to the data, thus protecting us from  
 259 concluding false-positive results and suggesting to the experimenter that more data is needed to make  
 260 a conclusive statement. Alternatively, we can incorporate any prior knowledge we have about the  
 261 problem, such as knowledge about the structure of  $W$  or knowledge about the  $W$  of a closely related

<sup>1</sup>other than a set of measure zero, that is thus negligible

<sup>2</sup>the angles are themselves constrained to lie in certain intervals e.g.  $[0, \pi)$  but these sorts of constraints are routine to deal with using a one-to-one diffeomorphism between intervals and the real line e.g. the sigmoid transform

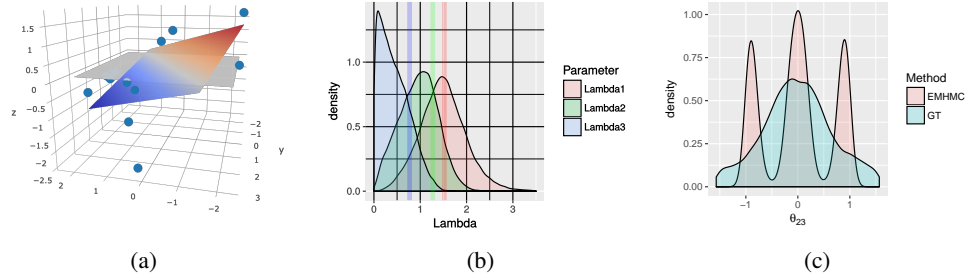


Figure 2: Inferences for three-dimensional synthetic data. (a) Three-dimensional points, true subspace (grey), and classical PCA point estimate of subspace (colored). (b) Estimated densities from posterior draws of  $\Lambda$  parameters A.K.A the singular values, and point estimates from classical PCA show as colored bars. (c) Avoidance of multi-modal behavior in GT-PPCA versus EMHMC. **PJA: This is a good start, but need bigger font size on the fonts in all these plots, since doing 3-panel. Also use black grid-lines, not white... very hard to read. Fine tune in Inkscape and also save a .svgz as well as .pdf copy for future edits.**

Figure 3: Posterior intervals for elements of  $W$  in sparse PCA.

group of samples, in the form of a prior distribution of our angles in our Stan model as we do in the following subsections.

The fully Bayesian approach provided by GT-PPCA in Stan also allows us to examine posterior draws of  $\Lambda$  to make probabilistic statements about the inherent dimensionality in our data. Figure 2b shows estimated densities from posterior draws of  $\Lambda$ . The posterior of  $\Lambda_3$  for example places considerable mass close to zero (58% of samples were less than 0.5), providing strong evidence that our data is inherently two, not three, dimensional. This is as oppose to classical PCA where we heuristically assess dimensionality based solely on the magnitude of our point estimates.

Lastly, Figure 2c compares posterior samples of our synthetic data from Embedded Manifold HMC (EMHMC) and GT-PPCA in Stan. As explained in section, EMHMC explores the entire Stiefel manifold which includes multiple equivalent modes, where as with GT-PPCA we can eliminate this multi-modal behavior by simply constraining the Givens Transform angle parameters. This is useful both for interpretation and in higher dimensional problems where the number of modes grows exponentially and HMC can not visit all of them.

## 5.2 Sparse PCA

## 5.3 Coagulopathy using hierarchical subspace models

Include figure for CCA probabilistic graphical model.

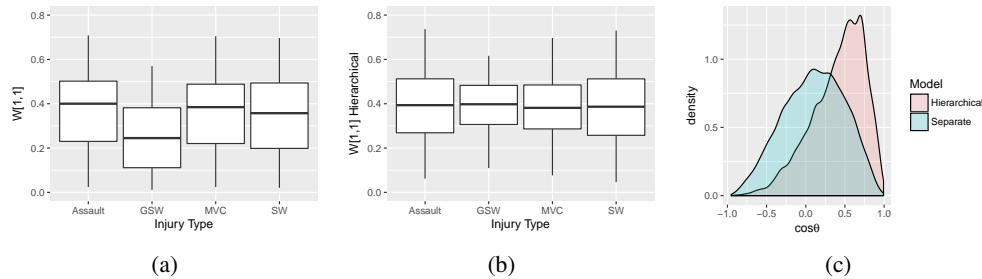


Figure 4: Inferences for Hierarchical CCA model.

## 5.4 School Network

Show  $W$  for each of the three states then posterior probabilities of what state you're in. Show probabilistic graphical model. Point how this can be used for disease networks and also recognizing states in fMRI data.

## 6 Discussion

### Acknowledgments

Use unnumbered third level headings for the acknowledgments. All acknowledgments go at the end of the paper. Do not include acknowledgments in the anonymized submission, only in the final paper.

### References

- [1] Marcus Brubaker, Mathieu Salzmann, and Raquel Urtasun. A family of mcmc methods on implicitly defined manifolds. In *Artificial Intelligence and Statistics*, pages 161–172, 2012.
- [2] Simon Byrne and Mark Girolami. Geodesic monte carlo on embedded manifolds. *Scandinavian Journal of Statistics*, 40(4):825–845, 2013.
- [3] Bob Carpenter, Andrew Gelman, Matt Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Michael A Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. Stan: A probabilistic programming language. *Journal of Statistical Software*, 20, 2016.
- [4] Michael Collins, Sanjoy Dasgupta, and Robert E Schapire. A generalization of principal component analysis to the exponential family. In *Nips*, volume 13, page 23, 2001.
- [5] Alan Edelman. 18.325: Finite random matrix theory volumes and integration. 2005.
- [6] Andrew Holbrook, Alexander Vandenberg-Rodes, and Babak Shahbaba. Bayesian inference on matrix manifolds for linear dimensionality reduction. *arXiv preprint arXiv:1606.04478*, 2016.
- [7] Daniel D Lee and H Sebastian Seung. Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems*, pages 556–562, 2001.
- [8] Benedict Leimkuhler and Sebastian Reich. *Simulating hamiltonian dynamics*, volume 14. Cambridge University Press, 2004.
- [9] Carl D Meyer. *Matrix analysis and applied linear algebra*, volume 2. Siam, 2000.
- [10] Shakir Mohamed, Zoubin Ghahramani, and Katherine A Heller. Bayesian exponential family pca. In *Advances in neural information processing systems*, pages 1089–1096, 2009.
- [11] Robb J Muirhead. *Aspects of multivariate statistical theory*, volume 197. John Wiley & Sons, 2009.
- [12] Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.
- [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.



$$R_{ij} := \begin{pmatrix} 1 & & & & & & & & & \\ & \ddots & & & & & & & & \\ & & 1 & & & & & & & \\ & & & \cos \theta_{ij} & & & & & & \\ & & & & 1 & & & & & \\ & & & & & \ddots & & & & \\ & & & & & & 1 & & & \\ & & \sin \theta_{ij} & & & & & \cos \theta_{ij} & & \\ & & & & & & & & 1 & \\ & & & & & & & & & \ddots \\ & & & & & & & & & & 1 \end{pmatrix} \begin{matrix} i \\ j \end{matrix}$$

$$R_{12}^{-1}Y = R_{12}^{-1} \begin{pmatrix} y_{11} & y_{12} & \cdots & y_{1p} \\ y_{21} & y_{22} & \cdots & y_{2p} \\ y_{31} & y_{32} & \cdots & y_{3p} \\ \vdots & \vdots & \vdots & \vdots \\ y_{p1} & y_{p2} & \cdots & y_{pp} \\ y_{p+1,1} & y_{p+1,2} & \cdots & y_{p+1,p} \\ \vdots & \vdots & \vdots & \vdots \\ y_{n1} & y_{n2} & \cdots & y_{np} \end{pmatrix} = \begin{pmatrix} * & * & \cdots & * \\ 0 & * & \cdots & * \\ y_{31} & y_{32} & \cdots & y_{3p} \\ \vdots & \vdots & \vdots & \vdots \\ y_{p1} & y_{p2} & \cdots & y_{pp} \\ y_{p+1,1} & y_{p+1,2} & \cdots & y_{p+1,p} \\ \vdots & \vdots & \vdots & \vdots \\ y_{n1} & y_{n2} & \cdots & y_{np} \end{pmatrix}. \quad (5)$$

$$(R_{1n}^{-1} \cdots R_{13}^{-1} R_{12}^{-1})Y = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & * & \cdots & * \\ 0 & * & \cdots & * \\ \vdots & \vdots & \vdots & \vdots \\ 0 & * & \cdots & * \\ 0 & * & \cdots & * \\ \vdots & \vdots & \vdots & \vdots \\ 0 & * & \cdots & * \end{pmatrix}. \quad (6)$$

$$(R_{2n}^{-1} \cdots R_{23}^{-1})(R_{1n}^{-1} \cdots R_{12}^{-1})Y = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & * \\ 0 & 0 & \cdots & * \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & * \\ 0 & 0 & \cdots & * \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & * \end{pmatrix}. \quad (7)$$

315 Continuing in this fashion yields

$$(R_{pn}^{-1} \cdots R_{p,p+1}^{-1}) \cdots (R_{2n}^{-1} \cdots R_{23}^{-1})(R_{1n}^{-1} \cdots R_{12}^{-1})Y = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 \end{pmatrix} = I_{np}. \quad (8)$$

## 316 7 Misc LaTeX

### 317 7.1 Givens Reductions

#### 318 7.1.1 Practical considerations of topology and angles

319 Topologically,  $V_{n,p}$  is locally equivalent to Euclidean space, but not globally equivalent, meaning it is  
 320 impossible to find a one-to-one map between the Stiefel manifold and Euclidean space. Technically  
 321 speaking, the Givens transform can map angles to all of  $V_{n,p}$  except for a subset  $S \subset V_{n,p}$ , that  
 322 in the  $n = 3, p = 2$  case corresponds to a sliver when  $\theta_{12} \in (-\pi, \pi)$ ,  $\theta_{13} \in (-\pi/2, \pi/2)$ , and  
 323  $\theta_{23} \in (-\pi/2, \pi/2)$ . Luckily this set is of measure zero (under the proper measure for the Stiefel  
 324 manifold, see section 7.1.3), and thus, with probability one, the orthonormal matrix that describes the  
 325 true subspace our data lie in will not be in that set.

326 In practice, we actually limit the angle  $\theta_{12}$  to an interval of length  $\pi$  rather than an integral of length  
 327  $2\pi$ , that traverses the entire Stiefel manifold. Examining the angles of the Givens transform makes it  
 328 evident that in the latter case, two equivalent bases that are the negation of each other can be reached,  
 329 resulting in a multi-modal posterior that makes sampling and VI more difficult and harder to interpret.  
 330 To avoid this multi-modality using the modified integrator methods would require a mechanism to  
 331 avoid boundaries, which are not as intuitively defined in the default embedded coordinates as in the  
 332 Givens transform.

333 Lastly, we note that if the true bases lies near a pole, i.e.  $\theta_{ij}$  is close to  $-\pi/2$  or  $\pi/2$ , then posteriors  
 334 will tend to be multi-modal as the region in parameter space close to the boundaries will be close to  
 335 equally valid, while the region near zero, will not be valid and thus contain little probability mass. In  
 336 these cases, one can simply change the chart so that  $\theta_{ij} \in (0, \pi)$ , creating a uni-modal posterior in  
 337 the new coordinate system, and alleviating numerical issues. In Stan this is straight-forward, as one  
 338 simply has to change the lower and upper bound of the angle parameter.

#### 339 7.1.2 Jacobian under a change of variables

340 The Givens transform 4 allows us to represent orthonormal matrices as angles  $Y(\Theta)$ . This in turn  
 341 allows us to write probability densities  $p_Y(Y)$  in terms of angles, so that we can conduct inference in  
 342 an unconstrained space. It is well known in probability theory that the transform of a random variable  
 343 is in general not the density of the transform, i.e.  $p_\Theta(\Theta) \neq p_Y(Y(\Theta))$  [12, chap. 2.6]. To be more  
 344 precise, densities are measured against volumes and integrated to get actual probabilities (otherwise  
 345 known as probability mass). Under a transformation, densities are unaffected, but volumes (or rather  
 346 the way in which volumes are measured) may change. This is important in the context of posterior  
 347 inference, as not including the Jacobian adjustment would result in different priors than we intend.

348 Figure 1c depicts how samples that are uniform in the angle space are not uniform on the sphere.  
 349 Samples congregate at the poles of the sphere because a patch of area in the angle space that is near  
 350 the top corresponds to a very tiny patch of the sphere near the pole. In practice this would lead to  
 351 posteriors that bias towards the poles, when what we really intend is a prior that is uniform on the  
 352 Stiefel manifold.

353 For a  $K$ -dimensional random vector  $Y$  and a transformation  $f : \text{supp}(Y) \rightarrow \mathbb{R}^K$  the proper way to  
 354 measure probability under a transform is by multiplying by the determinant of the Jacobian of the the  
 355 inverse transform:

$$\int p_{\Theta}(\Theta) d\Theta = \int p_Y(Y(\Theta)) |\det J_{f^{-1}}(\Theta)| d\Theta. \quad (9)$$

356 In our case this poses a problem however, because an  $n \times p$  orthonormal matrix is  $np$ -dimensional,  
 357 but the Givens transform,  $\Theta(Y)$ , maps this set to a  $np - p(p+1)/2$ -dimensional set. In this more  
 358 general scenario, one can not simply take the determinant of the Jacobian as the volume morphing  
 359 factor, because the Jacobian is not even square and hence the determinant is undefined. To obtain the  
 360 correct factor one must appeal to the calculus of differential forms.

### 361 7.1.3 Differential forms

362 We offer an intuitive high-level overview of differential forms. For a thorough account we recommend  
 363 Muirhead [11], Edelman [5], or any standard text in differential geometry. The simplest non-trivial  
 364 example of differential-forms between spaces of different sizes arises when trying to measure  
 365 probability using spherical coordinates. Spherical coordinates give us a map from  $\mathbb{R}^2$  to the sphere,  
 366  $(\theta, \varphi) \mapsto (x(\theta, \varphi), y(\theta, \varphi), z(\theta, \varphi))$ . Although a sphere lies in 3-D space, if we have a density  
 367  $p_{\text{Euc}}(x, y, z)$  on the sphere, the “natural” way to measure the probability of a spherical random  
 368 variable falling within some area on the surface of that sphere is by first covering that area with tiny  
 369 rectangles tangent to the sphere, taking the average density of each rectangle, multiplying that density  
 370 by the area of that rectangle, and summing over the resulting products. The probability of the random  
 371 variable falling within that area is defined to be the limit of that result as the size of the rectangles  
 372 go to zero. Area forms, or two-forms can be thought of as these infinitesimal rectangles. They are  
 373 written as the wedge product of two differentials, e.g.  $dx \wedge dy$ , and they are objects that are inserted  
 374 into integrals to obtain measures. Differential forms form (no pun intended) a vector space useful  
 375 for measuring areas and high-dimensional volumes. For example, the area forms on a sphere can be  
 376 written as a linear combination of the standard-basis two forms

$$a(x, y, z) dx \wedge dy + b(x, y, z) dx \wedge dz + c(x, y, z) dy \wedge dz. \quad (10)$$

377 One can then integrate this over a sub-area of the sphere to obtain the measure of that sub-area.  
 378 The beauty of differential forms is that a different coordinate system such as polar coordinates,  
 379 simply correspond to using a different basis for representing forms, which again are vectors. In fact  
 380 writing a form in a different bases simply involves taking partial derivatives, e.g.  $dx = (\partial x / \partial \theta) d\theta +$   
 381  $(\partial x / \partial \varphi) d\varphi$ . If we substitute the forms in the angle bases in to 10 and simplify using the well-defined  
 382 anti-symmetric and distributive properties of wedge products and differential forms (see [11]), we  
 383 obtain a proper area-form 10 in spherical coordinates,  $S(x(\theta, \varphi), y(\theta, \varphi), z(\theta, \varphi)) d\theta \wedge d\varphi$ , that can  
 384 then be integrated using a double integral in polar coordinates. The absolute value of this area-form  
 385 evaluated at some point specified in angle coordinates,  $(\theta_0, \varphi_0)$ , intuitively measures how much the  
 386 area of tiny square in angle space gets shrunk or stretched when the area is mapped to an area on the  
 387 sphere.

388 Analogously, we can measure volumes on the Stiefel manifold. For  $n \times p$  orthonormal matrices, there  
 389 are only  $np - p(p+1)/2$  free parameters and so the proper form to measure sets of orthonormal  
 390 matrices is in fact a  $np - p(p+1)/2$ -form. For an orthonormal,  $n \times p$  matrix,  $Y$ , we can find an  
 391 orthonormal  $n \times n$  matrix  $G$  such that  $G^T Y = I_{n,p}$ . In fact  $G$  just comes from the product of the  
 392 appropriate rotation matrices that comes from the Givens Reduction 3. Muirhead [11] shows that  
 393 the correct form comes from wedging the elements of the  $n \times p$  matrix  $G^T dY$  that lie below the  
 394 diagonal i.e.

$$\bigwedge_{i=1}^p \bigwedge_{j=i+1}^n G_j^T dY_i \quad (11)$$

395 where  $G_j$  is the  $j$ th column of  $G$  and  $Y_i$  is the  $i$ th column of  $Y$ . To obtain the form in angle  
 396 coordinates we simply obtain  $dY$  in angle coordinates.  $dY_i$  can be obtained in terms of the angle

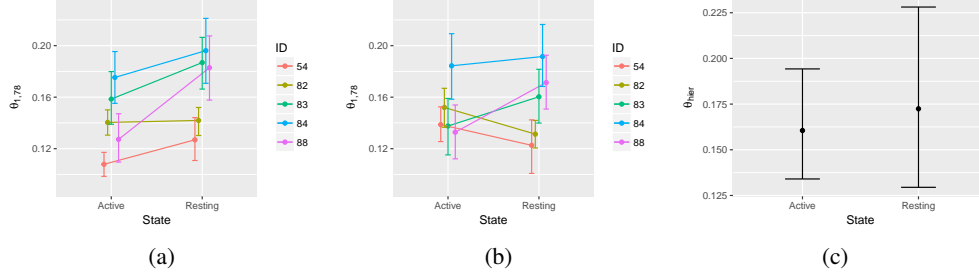


Figure 5: Inferring a hierarchical subspace model for fMRI data. (a) Individual posterior estimates of Givens Transform angle  $\theta_{1,78}$ . (b) Hierarchical estimates. (c) Posterior mean of hierarchical mean parameter.

coordinates by the following relationship,  $dY_i = J_{Y_i}(\Theta) d\Theta$ , where  $J_{Y_i}$  is the Jacobian of  $Y_i$  with respect to the angle coordinates. Once we obtain the form 11 in terms of the angle coordinates, the result is a wedge product of  $np - p(p+1)/2$  vectors that are  $np - p(p+1)/2$  dimensional, which reduces to the determinant of these vectors aligned side by side as a  $np - p(p+1)/2 \times np - p(p+1)/2$  matrix. This determinant is analogous to and serves the same purpose as Jacobian adjustment that comes from transforming random variables. We can insert it in to the log-probability of a model to avoid the sort of unintended sampling behavior depicted in Figure 1c. We incorporate the form 11 in to the log-probability of all of our Stan examples.

## 7.2 A hierarchical PPCA model for fMRI data

To illustrate the usefulness of GT-PPCA in a real-life setting, we applied it to a brain fMRI dataset and fit a hierarchical model using VI in Stan. We obtained 90-dimensional fMRI scans of five human subjects under an “active” condition where subjects were asked to perform a task, and a “resting” condition where subjects were not given any tasks or instructions. Each of the 90 dimensions corresponds to an activity level of a distinct area of the brain. From the data scientists wanted to discern whether fMRI activities were markedly different under the active and resting tasks.

It is well known that brain fMRI data is highly correlated so that our high dimensional data actually lies on some low dimensional subspace. In preliminary analysis we saw that the first principal component was highly tied to brain region 78 when subjects were in the resting state. We sought to test whether under the two states the first principal component had differing correlations to this brain region. Examining the posterior distribution of the GT parameter  $\theta_{1,78}$  we can intuitively understand how correlated with the 78th brain region the first component is. Drawing from our intuition from Figure ??, when this parameter is close to  $\pi/2$  or  $-\pi/2$  then the first component is highly correlated to the 78th brain region. Figure 5b show posterior estimates of this angle parameter,  $\theta_{1,78}$ , for our five subjects under the two states. While it seems as though the resting state is individually higher for this parameter, it is not clear whether this relationship is statistically significant.

A-priori, we know our subjects will have similar principal components when in the resting states and different, but also similar to each other, principal components in the active state. To reflect this prior knowledge we place a truncated normal prior over  $\theta_{1,78}^i$ ,  $i = 1, \dots, 5$ , and estimate the hyper-parameters  $\mu_0, \sigma_0$  of this prior for the active and resting state respectively. Not including the hierarchical prior is a special case of this where the hierarchical prior is completely flat, i.e.  $\sigma_0 = \infty$ , so that if there is a shared information between subjects our model can pick it up, and if there is not our estimates will be unaffected. Figure 5b shows posterior estimates for  $\theta_{1,78}^i$ , after we used a hierarchical prior on each respective state. The hierarchical prior has the regularization effect of shrinking together estimates in the two respective states, as information about one individual in the resting state tells us information about other individuals in the resting state.

Finally, to test our original hypothesis we compared the posterior distribution of  $\mu_0^{\text{resting}}$  and  $\mu_0^{\text{active}}$  in Figure 5c. While the posterior median of  $\mu_0^{\text{resting}}$  is lower, confidence intervals indicate that we would need more data to be able to conclude this with confidence, a conclusion that would be difficult to discern in a non-probabilistic framework.