

---

# Givens Transform Approach for Efficient PPCA in Bayesian Dimensionality Reduction

---

Anonymous Author(s)

Affiliation

Address

email

## Abstract

The abstract paragraph should be indented  $\frac{1}{2}$  inch (3 picas) on both the left- and right-hand margins. Use 10 point type, with a vertical spacing (leading) of 11 points. The word **Abstract** must be centered, bold, and in point size 12. Two line spaces precede the abstract. The abstract must be limited to one paragraph.

## 1 Introduction

**PJA: I would streamline this, since most readers will likely be experts on PCA. Instead focus on what is interesting about our use of PCA.** Principal Component Analysis (PCA) is a widely used dimensionality-reduction tool for exploratory analysis and modeling in both the natural and social sciences. By factorizing an empirical covariance matrix into a product of low rank matrices, PCA effectively finds a lower dimensional description of a dataset in terms of the latent factors that are given by the columns of the low rank factorization. Geometrically, PCA can be interpreted as providing a point estimate of a low dimensional hyper-plane that is closest to a cloud of data points. For binary or integer valued matrices, such as graph adjacency matrices arising in network science, Matrix Factorization (NMF) [?] or Exponential Family PCA (EPCA) [?] are often used for finding low dimensional latent factors.

**PJA: Good give based of PPCA** Probabilistic PCA (PPCA) [?] and Bayesian Exponential Family PCA (BXPCA) [?], posit probabilistic generative models that are equivalent to PCA in the limit of decreasing noise [?, chapt. 12.2]. This probabilistic approach is attractive because it allows a straightforward way, via Bayesian inference, to quantify the uncertainty in our estimates (to prevent overfitting) and conduct hypothesis testing, as is often done in the sciences. Furthermore, probabilistic models are amenable to expansion and serving as modules in larger probabilistic graphical models, which is important in a real world setting where true generative models are often complex and do not necessarily follow the simple generative process set forth by PPCA. We illustrate both of these properties in our examples section.

In practice, conducting full Bayesian inference on PPCA and other models involving one or more unknown orthonormal matrices is difficult because of the unusual support of such matrices. Namely, prior and posterior distributions of an orthonormal matrix,  $W$ , must have support over the set of  $n \times p$  orthonormal matrices, known as the Stiefel manifold and denoted  $V_{n,p}$ . At first glance, this rules out posterior inference via Markov Chain Monte Carlo (MCMC) and Variational Inference (VI), the two most prominent techniques for posterior inference. 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  $W$  are sure to lead to posteriors that assign mass to invalid regions of parameter space.

Despite this, there have been recent approaches to posterior inference of orthonormal matrices based on modified versions of Hamiltonian Monte Carlo (HMC) sampling for constrained parameters. These methods use modified numerical integrators to ensure that exploration only occurs in valid areas of parameter space where the necessary constraints are satisfied. Because these methods use modified integrators for constrained parameters, in practice they require keeping track of the type of each variable and which integrator to use on each variable. This adds an extra layer of programming complexity, especially for large, complex probabilistic graphical models, and makes difficult the implementation of these methods in the framework of a probabilistic programming language such as Stan [?]. Furthermore, in practice we found these methods to at times suffer from numerical errors that jeopardize the integrity of samples. Lastly, these methods can at times suffer from issues of unnecessary multi-modality. This occurs because the likelihood of the PPCA model is equivalent for an orthonormal matrix and arbitrary sign changes of its columns.

For inference of constrained parameters other than orthonormal matrices, such as positively constrained parameters, posterior distributions are routinely inferred in practice using both MCMC and VI. This is usually accomplished by transforming the constrained variables to an unconstrained space via a one-to-one mapping  $f : \text{supp}(z_{\text{constr}}) \rightarrow \mathbb{R}^D$ , where  $\text{supp}(z_{\text{constr}})$  is the support of some constrained random variable  $z_{\text{constr}}$ . To our knowledge, no such transformation that maps orthonormal matrices to an unconstrained space exists in the literature.

We introduce just such a transform based on Givens Reductions, that we call the Givens transform. The transform allows for straightforward implementation of PPCA and other models involving orthonormal matrices in any stand-alone HMC implementation or in the context of a probabilistic programming language like Stan. We illustrate this with our own provided Stan code which allows us to use Stan’s built in NUTS and VI implementations, as well as build more complicated models we describe in our examples section. Our transform provides an alternative representation of orthonormal matrices in terms of angles that we find useful for model interpretation. By limiting the range of these angles we can also avoid issues involving unnecessarily multi-modal posteriors in a straightforward way. Furthermore, the alternative representation allows for new and creative use of prior distributions on orthonormal matrices, a task which was previously intractable for even small problem sizes due to the difficulty of evaluating densities of orthonormal matrix distributions.

In section 2 we give an overview of probabilistic dimensionality reduction, then in section 3 we describe why orthonormal matrices should be used in inference, formally introduce the Stiefel manifold, and examine works related to Bayesian dimensionality reduction using orthonormal matrices. In section 4 we introduce Givens Reductions, which motivate the Givens Transform, as well the Givens Transform itself and practical considerations. We finish by showcasing examples of Bayesian inference on various probabilistic graphical models using the Givens Transform in Stan.

## 2 Probabilistic dimensionality reduction

PPCA posits that for a set of high-dimensional data vectors,  $x_i \in \mathbb{R}^n$ ,  $i = 1, \dots, N$ , there exists an unknown low-dimensional latent representation,  $z_i \in \mathbb{R}^p$ , of each vector, and that the two representations are related to each other by a single, unknown linear transformation. Formally PPCA consists of the following generative process:

$$\begin{aligned} p(z_i) &\sim \mathcal{N}_p(0, I) \\ p(x_i|z_i, W, \Lambda, \sigma^2) &\sim \mathcal{N}_n(W\Lambda z_i, \sigma^2 I) \end{aligned} \quad (1)$$

where  $W$  is an  $n \times p$  orthonormal matrix and  $\Lambda$  is a  $p \times p$  diagonal matrix with positive elements<sup>1</sup>. This model can be expanded to non-Gaussian data by replacing equation 1 with an exponential family member whose natural parameters are given by  $\text{Expon}(W\Lambda z_i)$ , where  $\text{Expon}(\cdot)$  is an appropriate link function [?].

---

<sup>1</sup>we assume the zero mean case here for simplicity, but the more general case easily follows

## 79 3 Bayesian inference of orthonormal matrices

### 80 3.1 Unidentifiability

81 If we assume momentarily that the matrix  $W$  need not be orthonormal, then as pointed out in ? ,  
82 chapt. 12.1.3], the likelihood will assign identical probability density values to different values of  
83  $W$ . In other words, without assuming orthonormality, the PPCA model is unidentifiable. While one  
84 can proceed with inference using HMC as in [? ], as pointed out by ? ] the unidentifiability in the  
85 model leads to high curvature areas of the log-posterior which itself leads to numerical issues for any  
86 sampler as well as slow mixing times, making inference with non-orthonormal matrices impractical.

### 87 3.2 Inference on the Stiefel manifold

88 To alleviate issues of unidentifiability, ? , chapt. 12.1.3] suggests forcing  $W$  to be orthonormal.  
89 However, as noted earlier, conducting full Bayesian inference on the space of orthonormal matrices is  
90 difficult because of the complicated constraints. We denote the set of  $n \times p$  orthonormal matrices  
91  $V_{n,p}$ . Formally we define

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

92  $V_{n,p}$  is known as the Stiefel manifold in differential geometry [? ]. While elements of the Stiefel  
93 manifold can be described as  $n \times p$  matrices with  $np$  elements, due to the constraint of the columns  
94 of the matrix being orthonormal, there are actually only  $np - p(p+1)/2$  degrees of freedom on the  
95 Stiefel manifold. One can see this by observing, that the first column of  $Y \in V_{n,p}$  must be of length  
96 one and hence has one constraint placed on it. The second column must also be of length one, but  
97 must also be orthogonal to the first column, and hence has two constraints placed on it, etc.

### 98 3.3 Related works using HMC

99 Recently, two approaches to posterior inference of orthonormal matrices and other constrained  
100 parameters have been proposed. These methods work by modifying the leap-frog integrator used to  
101 simulate Hamiltonian dynamics in HMC to ensure that any trajectory in parameter space remains on  
102 the Stiefel manifold. ? ] uses the SHAKE integrator [? ] to ensure trajectories through parameter  
103 space continuously satisfy the required constraints. The integrator works by repeatedly taking a step  
104 forward that may be off the manifold using ordinary leap frog, then projecting back down to the  
105 nearest point on the manifold. This projection is done via Newton iterations, which may not converge  
106 in practice, possibly jeopardizing the ergodicity of a Markov chain.

107 ? ] took a different approach, exploiting the fact that closed form solutions are known for the geodesic  
108 equations over the Stiefel manifold in the embedded coordinates,  $W$ . While this method is completely  
109 explicit, requiring no Newton iterations, in practice we found that for larger step sizes, the integrator  
110 steps off the Stiefel manifold, due to the numerical imprecision of the matrix exponential function.

111 Both methods rely on applying different integrators to constrained parameters. This adds imple-  
112 mentation difficulty in practice, as one must keep track of the type of each variable (unconstrained  
113 or constrained, and type of constraint), and apply the appropriate integrator. Additionally, we note  
114 that even using orthonormal matrices, the PPCA likelihood is equivalent for a matrix  $W$  and any  
115 permutation of the columns of  $W$  being negative. As such, these methods lead to multi-modal  
116 posteriors, that can be avoided in a straight-forward way, as we show in the next section, using the  
117 Givens transform.

## 118 4 Givens Transform

119 We provide a brief exposition on Givens Reductions, which motivate the Givens Transform, then  
120 describe the Givens Transform along with relevant practical considerations.

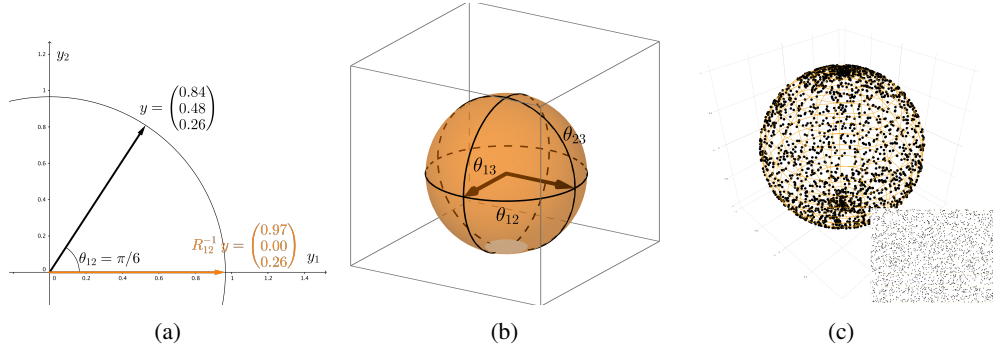


Figure 1: Visualizing the Givens Transform. (a) How the Givens Reduction “zeros out” a column vector. (b) A geometric view of the Stiefel manifold, two-frame in three dimensions. (c) Sampling without a proper measure adjustment.

#### 121 4.1 Givens Reductions

122 Define  $R_{ij}$  to be the  $n \times n$  rotation matrix that performs a counter-clockwise rotation in the  $(i, j)$ -  
 123 plane of  $\mathbb{R}^n$ , where  $j > i$ . In  $\mathbb{R}^3$ , there are three such matrices,  $R_{12}$ ,  $R_{13}$ , and  $R_{23}$ . They perform  
 124 counter-clockwise rotation of angle  $\theta_{ij}$  in the  $(x, y)$ ,  $(x, z)$  and  $(y, z)$  planes respectively. Rotation  
 125 matrices have the following key properties:

- 126 1. They preserve length and angles between vectors, i.e. for two vectors  $u, v \in \mathbb{R}^n$ ,  $R_{ij}u, R_{ij}v$   
 127 are the same length as  $u$  and  $v$  respectively, and if  $u$  and  $v$  are orthogonal then so are  $R_{ij}u$   
 128 and  $R_{ij}v$ .
- 129 2. They are invertible and their inverse is their transpose  $R_{ij}^{-1} = R_{ij}^T$ . Their inverse corresponds  
 130 to a clockwise rotation in the  $(i, j)$ -plane.

131 Now we consider an  $n \times p$  matrix  $Y$ , with orthonormal columns. In general, the first column is a  
 132 vector in  $\mathbb{R}^n$  with a non-zero second element. However, we can apply an invertible clockwise rotation  
 133 in the  $(1, 2)$ -plane,  $R_{12}^{-1}$ , to “zero out” the second element of the first column. Figure 1a depicts this  
 134 visually for a 3D vector projected on to the  $(1, 2)$ -plane.

135 Similarly, we can apply consecutive rotations  $R_{13}^{-1}, R_{14}^{-1}, \dots, R_{1n}^{-1}$  to this result so that all entries  
 136 of the first column besides the first element are zero. The first column of the resulting matrix,  
 137  $R_{1n}^{-1} \dots R_{13}^{-1} R_{12}^{-1} Y$ , will be the length one vector  $(1 \ 0 \ \dots \ 0)^T$  which lies entirely along the first axis.  
 138 If one takes the perspective that these rotations are applied to the columns of  $Y$ , then with the two  
 139 properties of rotation matrices we mentioned earlier, it is evident that columns 2 through  $n$  must have  
 140 zero in their first element because they will be orthogonal to the first column,  $(1 \ 0 \ \dots \ 0)^T$ .

141 To “zero out” the elements of the second column, one can similarly apply rotations  $R_{23}^{-1}, \dots, R_{2n}^{-1}$ .  
 142 These rotations will leave the first column and first row unaffected, as the first column no lies entirely  
 143 on the 1st axis, which these rotations do not involve. Continuing in this fashion yields

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

144 where  $I_{n,p}$  consists of the first  $p$  columns of the  $n \times n$  identity matrix. This process of applying  
 145 consecutive rotation matrices to a matrix is known in numerical analysis as the Givens reduction [?]  
 146 ], and is applied more generally to square matrices for matrix-vector solves. In total we will have  
 147 applied  $(n-1) + (n-2) + \dots + (n-p) = np - p(p+1)/2$  rotations matrices. We note that  
 148 because rotation matrices are very sparse in high dimensions, multiplication by a rotation matrix is  
 149 computationally much less intensive than matrix multiplication otherwise would be.

## 150 4.2 Givens Transform

151 Because rotation matrices are invertible, equation 3 implies that we can rewrite the  $n \times p$  orthonormal  
152 matrix  $Y$  as the product of counter-clockwise rotation matrices and  $I_{n,p}$ :

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

153 Recall that each of the  $np - p(p+1)/2$  rotation matrices have an associated angle  
154  $(\theta_{12} \cdots \theta_{1n}) \cdots (\theta_{23} \cdots \theta_{2n}) (\theta_{p+1,n} \cdots \theta_{pn})$ , that we collectively refer to as  $\Theta$ . In this way we  
155 have reparameterized all  $n \times p$  orthonormal matrices<sup>2</sup>, a constrained space, in terms of unconstrained  
156 angles<sup>3</sup>, using a transform  $\Theta : V_{n,p} \rightarrow \mathbb{R}^{np-p(p+1)/2}$ . We refer to 4 as the Givens representation or  
157 Givens transform.

### 158 4.2.1 A geometric perspective

159 The Givens transform can be visualized in the simple case when  $n = 3$  and  $p = 2$ . This is the case  
160 where data is three-dimensional and we seek a two-dimensional subspace to represent the data. In  
161 this case, we can imagine a two dimensional bases, visualized as two perpendicular vectors that rotate  
162 rigidly together along three angles of rotations. This is referred to as a 2-frame in differential geometry  
163 [?] and is depicted in Figure 1b. Recall  $\theta_{12}$  controls the amount of rotation in the  $(1, 2)$ -plane,  $\theta_{13}$   
164 in the  $(1, 3)$ -plane, and  $\theta_{23}$  controls a rotation of the second basis vector about the first. All rotations  
165 will preserve the length and orthogonality of the basis vectors. Given a flat, pancake-like cloud of  
166 points in 3D, the two dimensional bases will move according to the appropriate angles to be aligned  
167 with the cloud of points, if one were conducting a (Maximum A-Posteriori) MAP estimation over the  
168 Stiefel manifold.

### 169 4.2.2 Practical considerations of topology and angles

170 Topologically,  $V_{n,p}$  is locally equivalent to Euclidean space, but not globally equivalent, meaning it is  
171 impossible to find a one-to-one map between the Stiefel manifold and Euclidean space. Technically  
172 speaking, the Givens transform can map angles to all of  $V_{n,p}$  except for a subset  $S \subset V_{n,p}$ , that  
173 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  
174  $\theta_{23} \in (-\pi/2, \pi/2)$ . Luckily this set is of measure zero (under the proper measure for the Stiefel  
175 manifold, see section 4.2.4), and thus, with probability one, the orthonormal matrix that describes the  
176 true subspace our data lie in will not be in that set.

177 In practice, we actually limit the angle  $\theta_{12}$  to an interval of length  $\pi$  rather than an interval of length  
178  $2\pi$ , that traverses the entire Stiefel manifold. Examining the angles of the Givens transform makes it  
179 evident that in the latter case, two equivalent bases that are the negation of each other can be reached,  
180 resulting in a multi-modal posterior that makes sampling and VI more difficult and harder to interpret.  
181 To avoid this multi-modality using the modified integrator methods would require a mechanism to  
182 avoid boundaries, which are not as intuitively defined in the default embedded coordinates as in the  
183 Givens transform.

184 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  
185 will tend to be multi-modal as the region in parameter space close to the boundaries will be close to  
186 equally valid, while the region near zero, will not be valid and thus contain little probability mass. In  
187 these cases, one can simply change the chart so that  $\theta_{ij} \in (0, \pi)$ , creating a uni-modal posterior in  
188 the new coordinate system, and alleviating numerical issues. In Stan this is straight-forward, as one  
189 simply has to change the lower and upper bound of the angle parameter.

### 190 4.2.3 Jacobian under a change of variables

191 The Givens transform 4 allows us to represent orthonormal matrices as angles  $Y(\Theta)$ . This in turn  
192 allows us to write probability densities  $p_Y(Y)$  in terms of angles, so that we can conduct inference in  
193 an unconstrained space. It is well known in probability theory that the transform of a random variable

<sup>2</sup>other than a set of measure zero we explain in the next subsection

<sup>3</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

is in general not the density of the transform, i.e.  $p_\Theta(\Theta) \neq p_Y(Y(\Theta))$  [?, chapt. 2.6]. To be more precise, densities are measured against volumes and integrated to get actual probabilities (otherwise known as probability mass). Under a transformation, densities are unaffected, but volumes (or rather the way in which volumes are measured) may change. This is important in the context of posterior inference, as not including the Jacobian adjustment would result in different priors than we intend.

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

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

$$\int p_\Theta(\Theta) d\Theta = \int p_Y(Y(\Theta)) |\det J_{f^{-1}}(\Theta)| d\Theta. \quad (5)$$

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

#### 4.2.4 Differential forms

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

$$a(x, y, z) dx \wedge dy + b(x, y, z) dx \wedge dz + c(x, y, z) dy \wedge dz. \quad (6)$$

One can then integrate this over a sub-area of the sphere to obtain the measure of that sub-area. The beauty of differential forms is that a different coordinate system such as polar coordinates, simply correspond to using a different basis for representing forms, which again are vectors. In fact writing a form in a different bases simply involves taking partial derivatives, e.g.  $dx = (\partial x / \partial \theta) d\theta + (\partial x / \partial \varphi) d\varphi$ . If we substitute the forms in the angle bases in to 6 and simplify using the well-defined anti-symmetric and distributive properties of wedge products and differential forms (see [?]), we obtain a proper area-form 6 in spherical coordinates,  $S(x(\theta, \varphi), y(\theta, \varphi), z(\theta, \varphi)) d\theta \wedge d\varphi$ , that can then be integrated using a double integral in polar coordinates. The absolute value of this area-form evaluated at some point specified in angle coordinates,  $(\theta_0, \varphi_0)$ , intuitively measures how much the area of tiny square in angle space gets shrunk or stretched when the area is mapped to an area on the sphere.

Analogously, we can measure volumes on the Stiefel manifold. For  $n \times p$  orthonormal matrices, there are only  $np - p(p+1)/2$  free parameters and so the proper form to measure sets of orthonormal matrices is in fact a  $np - p(p+1)/2$ -form. For an orthonormal,  $n \times p$  matrix,  $Y$ , we can find an

242 orthonormal  $n \times n$  matrix  $G$  such that  $G^T Y = I_{n,p}$ . In fact  $G$  just comes from the product of the  
 243 appropriate rotation matrices that comes from the Givens Reduction 3. ? ] shows that the correct  
 244 form comes from wedging the elements of the  $n \times p$  matrix  $G^T dY$  that lie below the diagonal i.e.

$$\bigwedge_{i=1}^p \bigwedge_{j=i+1}^n G_j^T dY_i \quad (7)$$

245 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  
 246 coordinates we simply obtain  $dY$  in angle coordinates.  $dY_i$  can be obtained in terms of the angle  
 247 coordinates by the following relationship,  $dY_i = J_{Y_i}(\Theta) d\Theta$ , where  $J_{Y_i}$  is the Jacobian of  $Y_i$  with  
 248 respect to the angle coordinates. Once we obtain the form 7 in terms of the angle coordinates, the  
 249 result is a wedge product of  $np - p(p+1)/2$  vectors that are  $np - p(p+1)/2$  dimensional, which  
 250 reduces to the determinant of these vectors aligned side by side as a  $np - p(p+1)/2 \times np - p(p+1)/2$   
 251 matrix. This determinant is analogous to and serves the same purpose as Jacobian adjustment that  
 252 comes from transforming random variables. We can insert it in to the log-probability of a model to  
 253 avoid the sort of unintended sampling behavior depicted in Figure 1c. We incorporate the form 7 in  
 254 to the log-probability of all of our Stan examples.

## 255 5 Examples

### 256 5.1 Synthetic Data

257 Show how point estimation of subspace and dimensionality can be misleading. Show multi-modality  
 258 of Symone's method.

### 259 5.2 FMRI

260 Illustrate hypothesis testing.

### 261 5.3 School Network

262 Mainly here to show we can do exponential families easy too. Possibly make this a change point  
 263 detection?

### 264 5.4 Coagulopathy using hierarchical subspace models

265 Don't forget to mention sparse PCA, via Cauchy priors.

## 266 6 Discussion

### 267 Acknowledgments

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

## 270 References

- 271 [ ] Marcus Brubaker, Mathieu Salzmann, and Raquel Urtasun. A family of mcmc methods on  
 272 implicitly defined manifolds. In *Artificial Intelligence and Statistics*, pages 161–172, 2012.
- 273 [ ] Simon Byrne and Mark Girolami. Geodesic monte carlo on embedded manifolds. *Scandinavian*  
 274 *Journal of Statistics*, 40(4):825–845, 2013.
- 275 [ ] Bob Carpenter, Andrew Gelman, Matt Hoffman, Daniel Lee, Ben Goodrich, Michael Betan-  
 276 court, Michael A Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. Stan: A probabilistic  
 277 programming language. *Journal of Statistical Software*, 20, 2016.

- 278 [] Michael Collins, Sanjoy Dasgupta, and Robert E Schapire. A generalization of principal  
279 component analysis to the exponential family. In *Nips*, volume 13, page 23, 2001.
- 280 [] Alan Edelman. 18.325: Finite random matrix theory volumes and integration. 2005.
- 281 [] Andrew Holbrook, Alexander Vandenberg-Rodes, and Babak Shahbaba. Bayesian inference on  
282 matrix manifolds for linear dimensionality reduction. *arXiv preprint arXiv:1606.04478*, 2016.
- 283 [] Daniel D Lee and H Sebastian Seung. Algorithms for non-negative matrix factorization. In  
284 *Advances in neural information processing systems*, pages 556–562, 2001.
- 285 [] Benedict Leimkuhler and Sebastian Reich. *Simulating hamiltonian dynamics*, volume 14.  
286 Cambridge University Press, 2004.
- 287 [] Carl D Meyer. *Matrix analysis and applied linear algebra*, volume 2. Siam, 2000.
- 288 [] Shakir Mohamed, Zoubin Ghahramani, and Katherine A Heller. Bayesian exponential family  
289 pca. In *Advances in neural information processing systems*, pages 1089–1096, 2009.
- 290 [] Robb J Muirhead. *Aspects of multivariate statistical theory*, volume 197. John Wiley & Sons,  
291 2009.
- 292 [] Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.
- 293 [] Michael E Tipping and Christopher M Bishop. Probabilistic principal component analysis.  
294 *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):611–622,  
295 1999.

## 296 Appendix

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

$$R_{12}^{-1}Y = R_{12}^{-1} \begin{pmatrix} y_{11} & y_{12} & \cdots & y_{1p} \\ y_{21} & y_{22} & \cdots & y_{2p} \\ y_{31} & y_{32} & \cdots & y_{3p} \\ \vdots & \vdots & \vdots & \vdots \\ y_{p1} & y_{p2} & \cdots & y_{pp} \\ y_{p+1,1} & y_{p+1,2} & \cdots & y_{p+1,p} \\ \vdots & \vdots & \vdots & \vdots \\ y_{n1} & y_{n2} & \cdots & y_{np} \end{pmatrix} = \begin{pmatrix} * & * & \cdots & * \\ 0 & * & \cdots & * \\ y_{31} & y_{32} & \cdots & y_{3p} \\ \vdots & \vdots & \vdots & \vdots \\ y_{p1} & y_{p2} & \cdots & y_{pp} \\ y_{p+1,1} & y_{p+1,2} & \cdots & y_{p+1,p} \\ \vdots & \vdots & \vdots & \vdots \\ y_{n1} & y_{n2} & \cdots & y_{np} \end{pmatrix}. \quad (8)$$



$$(R_{1n}^{-1} \cdots R_{13}^{-1} R_{12}^{-1})Y = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & * & \cdots & * \\ 0 & * & \cdots & * \\ \vdots & \vdots & \vdots & \vdots \\ 0 & * & \cdots & * \\ 0 & * & \cdots & * \\ \vdots & \vdots & \vdots & \vdots \\ 0 & * & \cdots & * \end{pmatrix}. \quad (9)$$

$$(R_{2n}^{-1} \cdots R_{23}^{-1})(R_{1n}^{-1} \cdots R_{12}^{-1})Y = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & * \\ 0 & 0 & \cdots & * \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & * \\ 0 & 0 & \cdots & * \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & * \end{pmatrix}. \quad (10)$$

297 Continuing in this fashion yields

$$(R_{pn}^{-1} \cdots R_{p,p+1}^{-1}) \cdots (R_{2n}^{-1} \cdots R_{23}^{-1})(R_{1n}^{-1} \cdots R_{12}^{-1})Y = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 \end{pmatrix} = I_{np}. \quad (11)$$