General Bayesian Inference over the Stiefel Manifold via the Givens Transform

Arya A Pourzanjani **Richard M Jiang** **Paul J Atzberger** **Linda R Petzold**
University of California Santa Barbara

Abstract

We introduce the Givens Transform, a novel transform between the space of orthonormal matrices and $\mathbb{R}^D$. The Givens Transform allows for the application of any general Bayesian inference algorithm to probabilistic models containing constrained unit-vectors or orthonormal matrix parameters. This includes a variety of matrix factorizations and dimensionality reduction models such as Probabilistic PCA (PPCA), Exponential Family PPCA (BXPCA), and Canonical Correlation Analysis (CCA). While previous Bayesian approaches to these models relied on separate sampling update rules for constrained and unconstrained parameters, the Givens Transform enables the treatment of unit-vectors and orthonormal matrices agnostically as unconstrained parameters. Thus any Bayesian inference algorithm can be used on these models without modification. This opens the door to not just sampling algorithms, but Variational Inference (VI) as well. We illustrate with several examples and supplied code, how the Givens Transform allows end-users to easily build complex models in their favorite Bayesian modeling framework such as Stan, Edward, or PyMC3, a task that was previously intractable due to technical constraints.

1 Introduction

The Bayesian modeling paradigm involves setting up a probabilistic model describing how data was generated, assigning prior distributions over unknown model parameters, and then calculating a posterior distribution over these parameters [4; 17]. In practice, this posterior distribution is intractable to compute exactly for all but the simplest models, and one must resort to approximate posterior inference algorithms. Fortunately, much work has been done on this problem, resulting in state-of-the-art algorithms such as Hamiltonian Monte Carlo (HMC) [18], the No-U-Turn Sampler (NUTS) [9], Automatic Differentiation VI (ADVI) [11] and Black Box VI[20]. These algorithms are applicable to a wide class of models, and are readily available in popular Probabilistic Programming languages such as Stan, Edward, and PyMC3[3; 25; 22].

One class of models these algorithms do not generally apply to are models with parameters constrained to be unit-vectors or orthonormal matrices. This most notably precludes models arising in several domains such as materials science [19], biology [7], and robotics [13], and also models arising in probabilistic dimensionality reduction [2; 10] such as PPCA, BXPCA [15], mixture of PPCA [5], CCA [17, Chapt. 12.5], and the examples we showcase in our empirical studies section. For simple constrained parameters, researchers and practitioners have typically bypassed this hurdle by transforming constrained model parameters to an unconstrained space and conducting inference in the resultant space. For example, if a model contains some parameter $\sigma > 0$ that is constrained to be positive, one can simply take the log of this parameter and conduct inference over $\log \sigma$, which is unconstrained. This procedure is done routinely in Stan [3] and is the basis for ADVI [11].

Unfortunately, for more complex constraints, such transformations have not been mathematically derived. In response, various sampling algorithms have been devised for obtaining distributions over constrained parameters such as unit-vectors and orthonormal matrices [8; 1; 2; 10]. These algorithms use different update rules on constrained and unconstrained parameters, making them difficult to implement in standard software packages, and precluding practical use on larger, more complex models. In particular, no VI

Preliminary work. Under review by AISTATS 2018. Do not distribute.

methods have been proposed for models with unit-vector and orthonormal matrix parameters, making inference on larger models with these parameters particularly problematic. Being able to use a transform would be ideal as it would more cleanly modularize models from inference algorithms.

To this end, we introduce the Givens Transform, a novel transform between the space of orthonormal matrices and the unconstrained space, which, to the best of our knowledge, is the first transform that allows for the application of any general inference algorithm to models containing unit-vectors and orthonormal matrix parameters. The transform is easy to implement and does not require any specialized inference algorithms or modifications to existing algorithms or software. This allows users to rapidly build and prototype complex probabilistic models with orthonormal matrix parameters in any common software framework such as Stan, Edward, or PyMC3 without having to worrying about messy implementation details. Users can then subsequently conduct fully Bayesian inference using any state-of-the-art inference algorithm available in these packages, including variational inference, which was previously not possible. We stress that allowing users to use models with orthonormal matrix parameters in common modeling packages opens up use of a wide class of new models, and frees them up to focus on modeling rather than implementation and debugging of custom modeling code and inference algorithms. Furthermore, by treating parameters agnostically as unconstrained, the Givens Transform allows inference algorithm designers to focus on more general algorithms rather than separate model specific ones.

In addition, the Givens Transform, which represents orthonormal matrices in terms of a sequence of fundamental rotations through given angles, yields geometric insights into novel and useful ways to work with and interpret models with orthonormal matrix parameters. This helps in addressing a number of previously unresolved issues. Specifically, the elegant geometric representation lets us see how, by limiting the range of the parameters in the Givens Transform, we can naturally avoid issues of unidentifiability that arise when working with orthonormal matrices. The Givens Transform also enables new and creative ways to generate and use prior distributions on orthonormal matrices, and thus subspaces, a task that had previously been rather complicated due to the difficulty of evaluating densities of orthonormal matrix distributions for even small problem sizes [8]. As we shall discuss in more detail, our method allows for a natural way to specify prior distributions over orthonormal matrices comparable to the Matrix Langevin prior [16].

In Section 2 we discuss previous methods for conducting inference over unit-vector and orthonormal matrix parameters. We briefly explain how transformations are typically used in Bayesian inference in Section 3. In Section 4 we discuss the geometry of the Stiefel Manifold, the space of orthonormal matrices, setting the stage for the Givens Transform which we discuss in Section 5. In Section 6 we present several examples from our own applied work where we utilize the Givens Transform in Stan to implement several complex models containing unit-vector and orthonormal matrix parameters. We finish with a brief discussion in Section 7.

2 Related Work

A few sampling-based methods have been developed to obtain posteriors over orthonormal matrix parameters. Brubaker et al. [1] proposes the use of the SHAKE integrator [12] to simulate Hamiltonian dynamics and generate proposals. For constrained parameters, 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 using Newton Iterations. Byrne and Girolami [2] as well as Holbrook et al. [10] exploit the fact that closed form solutions are known for the geodesic equations in the space of orthonormal matrices in the embedded coordinates, W. They utilize these equations to update constrained parameters in a different manner than for unconstrained parameters in their derived Embedded Manifold HMC (EMHMC) algorithm.

As these methods all use modified integrators for constrained parameters, they require additional bookkeeping of the support and the integrator of each model parameter, unlike the Givens Transform which treats these parameters as unconstrained parameters. This makes them incompatible with the current widely available Probabilistic Programming languages such as Stan and Edward, which typically do not expose the underlying inference algorithm to the user. Furthermore, these algorithms are unable to take advantage of improved samplers such as NUTS and optimization based approximate methods such as ADVI, limiting their scalability to large models.

3 Bayesian Inference of Constrained Parameters Using Transformations

Given a constrained random variable Z, and a (possibly unnormalized) density $p_Z(z)$, one can use a smooth one-to-one mapping, $T : \text{support}(Z) \rightarrow \mathbb{R}^D$ to obtain a new density $p_U(u) = p_Z(T^{-1}(u))|J_{T^{-1}}(u)|$ in terms of an unconstrained random variable, U. Here $|J_{T^{-1}}(u)|$ is the determinant of the Jacobian of T^{-1}

and is a standard term that accounts for how a unit volume changes under the transformation [4; 17; 11]. Under the transformed density $p_U(u)$, one can obtain posterior samples $u_1, \cdots, u_N$ or a variational distribution, $q_\gamma(u)$ over the unconstrained parameter that corresponds to the original constrained parameter of interest. Samples can then be freely mapped back to the original constrained space using the inverse transform T^{-1} to obtain posteriors in the original constrained space $z_1 = T^{-1}(u_1), \cdots, z_N = T^{-1}(u_N)$. To this end, we derive just such a transform for orthonormal matrix parameters by first appealing to the geometry of the space in which they reside.

4 Geometry of the Stiefel Manifold

To make better sense of the space of orthonormal matrices, we can analyze its geometric properties. Just as three-dimensional unit vectors form a sphere in $\mathbb{R}^3$, which is a sub-manifold of $\mathbb{R}^3$, the set of $n \times p$ orthonormal matrices form a sub-manifold in the space of general $n \times p$ matrices, known as the Stiefel Manifold and denoted $V_{n,p}$[16]. $V_{n,p}$ is formally defined as

$$V_{n,p} := \{Y \in \mathbb{R}^{n \times p} : YY^T = I\}. \tag{1}$$

The elements of $V_{n,p}$ are referred to as p-frames, a collection of p orthonormal vectors that lie in n-dimensional space, and can collectively be represented by an $n \times p$ orthonormal matrix, Y. Given a p-frame, Y, we can get any other p-frame by simultaneously and rigidly rotating the columns of Y about any combination of axes an arbitrary number of times.

Even though $n \times p$ orthonormal matrices are typically represented by np elements, the intrinsic dimension of the Stiefel Manifold, $V_{n,p}$, is actually $np - p(p+1)/2$. 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 has two constraints placed on it. Continuing to the third column through the n^{th}. one arrives at the conclusion that each point of the Stiefel Manifold has only $np-(1+2+\cdots+p) = np-p(p+1)/2$ degrees of freedom. The reduced dimensionality motivates the Givens Transform, which can be thought of as an $np - p(p+1)/2$-dimensional set of coordinates Φ, that represent elements of the Stiefel manifold.

As a concrete example, the specific case where $n = 3$ and $p = 1$ corresponds to aforementioned unit vector in $\mathbb{R}^3$ whose position can be represented in terms of two angles of rotation: one representing rotation in the xy-plane, θ_{12} (latitude), and one representing rotation in the xz-plane θ_{13} (latitude). This is the standard spherical coordinates system with the radius free parameter set to exactly 1. Extending this to $n = 3$ and $p = 2$, we can imagine adding a second unit vector with position defined to be orthonormal to the first unit vector. However, now to define any other element of $V_{3,2}$ from any other, we must take care to keep the two orthonormal. This means we are constrained to rotate the second unit vector in reference to the first. Thus, this whole system is represented by three angles: two angles to represent the position of the first vector, and a third angle, θ_{23} that controls how much the second basis vector is rotated about the first (Figure 1).

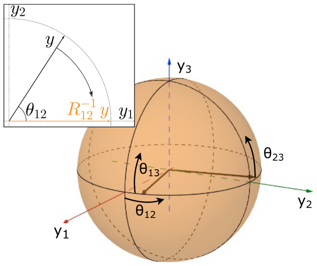

Figure 1: To obtain different elements of the Stiefel Manifold we rigidly rotate p-frames. This motivates the connection to Givens Reductions which work by rotating in some plane (inset).

5 The Givens Transform

More generally, one can represent any $n \times p$ orthonormal matrix by a $np - p(p+1)/2$-dimensional vector $\Theta := (\theta_{12} \cdots \theta_{1n}) \cdots (\theta_{23} \cdots \theta_{2n})(\theta_{p+1,n} \cdots \theta_{pn})$ by successively applying clock-wise rotation matrices with these angles to the matrix $I_{n,p}$, which we define to be first p columns of the $n \times n$ identity matrix. That is we can represent any orthonormal matrix Y, as a function of these angles:

$$Y(\Theta) = (R_{12}^{\theta_{12}} \cdots R_{1n}^{\theta_{1n}}) \cdots (R_{23}^{\theta_{23}} \cdots R_{2n}^{\theta_{2n}})(R_{p+1,n}^{\theta_{p+1,n}} \cdots R_{pn}^{\theta_{pn}}) I_{n,p}. \tag{2}$$

The elements of Θ will be constrained to lie in either the interval $[-\pi, \pi)$ or $[-\pi/2, \pi/2)$, however we can use a simple logistic transform [4] applied element-wise to obtain the unconstrained vector Φ, which can represent any orthonormal matrix by $Y(\Theta(\Phi))$. This establishes the Givens Transform, a continuous one-to-one map between the space of orthonormal matrices and unconstrained space, $Y : \mathbb{R}^{np-p(p+1)/2} \rightarrow V_{n,p}$. The transform enables use of orthonormal matrices in any probabilistic programming language by declaring Φ as unconstrained parameters and transforming to Y, which can then be used in any likelihood. Algorithm 1 shows how this transformation is formally computed.

We first note that the formula in Equation 5 will always return an orthonormal matrix, that is continuous in Θ. This is evident as $I_{n,p}$ is orthonormal and application of rotation matrices does not affect magnitude or orthonormality.

We establish that the Givens Transform is a valid mapping in three steps using the remainder of this section. Namely we establish surjectivity, by showing that any $n \times p$ orthonormal matrix can be represented by some combination of angles Θ, and hence Φ. We then show injectivity by showing that no two different combinations of Θ, and hence Φ lead to the same angles. Lastly, we show we can identify and explicitly compute a form akin to the Jacobian adjustment term described in Section 3 for Bayesian inference.

5.1 Surjectivity

To establish surjectivity of the Givens Transform, we rely on the Givens Reduction algorithm from numerical analysis, an algorithm traditionally used for obtaining a QR factorization of an $n \times n$ matrix A[14]. The algorithm works by applying a series of counterclockwise rotation 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 ultimately resulting in the upper-triangular matrix

$$\begin{aligned} R \quad = \quad & ((R_{pn}^{\theta_{pn}})^{-1} \cdots (R_{p+1,n}^{\theta_{p+1,n}})^{-1}) \\ & \cdots ((R_{2n}^{\theta_{2n}})^{-1} \cdots (R_{23}^{\theta_{23}})^{-1})((R_{1n}^{\theta_{1n}})^{-1} \cdots (R_{12}^{\theta_{12}})^{-1})A \end{aligned} \tag{3}$$

where R_{ij}^{θ} denotes the matrix representing a counter-clock-wise rotation of angle θ in the i-j plane and its inverse denotes the corresponding clock-wise rotation. Note that each subsequent rotation has the property of not "undoing" the the rotations before it. One can similarly apply this algorithm to an orthonormal matrix:

Theorem 5.1. *Let Y be any arbitrary $n \times p$ matrix with orthonormal columns. Applying the Givens Reduction algorithm applied to Y, i.e. replacing A with Y in Equation 5.1 results in the matrix $I_{n,p}$.*

Proof. By construction, the Givens Reduction will make all lower diagonal elements of R zero. Thus the first column of R will be non-zero only in the first element. In fact this first element will be one since rotations do not affect magnitude and the column started with length one by virtue of Y being orthonormal.

Now assuming that the first J columns of R match the first J columns of $I_{n,p}$ after the Givens Reduction algorithm has "zeroed out" the first J columns, we show that the algorithm will go on to make the first $J+1$ columns match as well. In fact the first J columns will be left unchanged by this transformation since rotation will only take place on elements after the Jth row, and these columns will have zeros in those rows. Furthermore, by construction, the $J+1$ column will have zeros past the $J+1$ element, but since simultaneous rotations of the columns of R retain the orthogonality between the $J+1$ column and the preceding columns, the $J+1$ column will also have zeros above the $J+1$ entry as well, otherwise it would not be orthogonal to the preceding columns. Lastly, because rotations do not change the length of a vector, the $J+1$ column must be of length one, and hence its only non-zero entry, the $J+1$ will be one. □

Because rotations are invertible one can take all the rotations in Equation 5.1 to the left hand side obtaining the representation in Equation 5. Since Theorem 5.1 only stipulated Y to be any arbitrary $n \times p$ matrix with orthonormal columns, surjectivity is established.

For thoroughness we note that topologically, $V_{n,p}$ is locally equivalent to Euclidean space, but not globally equivalent, meaning it is impossible to find a one-to-one map between the Stiefel manifold and Euclidean space. Technically speaking, the Givens transform can map angles to all of $V_{n,p}$ except for a subset $S \subset V_{n,p}$ of measure zero [1]. In the $n = 3, p = 1$ case (the sphere), this corresponds to a sliver where $\theta_{12} = \pi$ and $\theta_{13} \in (-\pi/2, \pi/2)$. As a consequence, with probability one, the orthonormal matrix that describes the true subspace our data lies in will not be in that set [21].

5.2 Injectivity

To establish injectivity it is enough to see that different angles inserted in to Equation 5 cannot possible lead to different orthonormal matrices, as differing angles would produce unique invertible rotation matrices

5.3 Transformation of Measure Under the Givens Transform

To fully specify our method so that Bayesian inference can be conducted, we need to define the Jacobian

[1] S is of measure zero under the Lebesgue measure or any continuous density over the Stiefel Manifold

```
Input: A set of $np - p(p+1)/2$ angles to rotate
       about, $\theta$
Result: An orthonormal $n \times p$ matrix $Y$ representing
        the coordinates from the angles.
$Z = I_n$; idx = 0
for $i$ in 1:p do
    for $j$ in i+1:n do
        //create Rotation matrix defined by theta
        $T = I_n$;
        $T[i,i] = \cos(\theta_{idx})$; $T[i,j] = \sin(\theta_{idx})$;
        $T[j,i] = -\sin(\theta_{idx})$; $T[j,j] = \cos(\theta_{idx})$;
        //apply rotation matrix
        $Z = ZT$;
        $idx = idx + 1$;
    end
end
$Y = Z[:, :p]$;
```

Algorithm 1: Psuedo-code for performing a Givens Transform on a set of $np - p(p+1)/2$ angles.

adjustment term to account for the change in unit volume under the transformation, as described in Section 3. Failure to include such a term can result in measures that are incorrectly concentrated at different locations in parameter space (Figure 2).

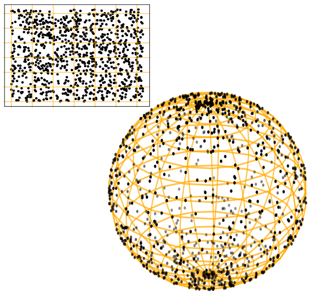

Figure 2: Even if we sample uniformly in angle coordinates (inset), without a proper measure adjustment accounting for how volumes are warped under the Givens Transform, samples will not be uniform when transformed to the sphere. In this case, samples are sparse near the equator and congregate near the poles. Intuitively, areas that are near the poles are shrunk far more than areas near the equator, so when mapped back onto the sphere, points will congregate closer to the poles of the sphere than the equator.

In the case of the Givens Transform, as the Jacobian is not square, one can not simply use the determinant of the Jacobian as the volume correction factor as that factor is undefined. An $n \times p$ orthonormal matrix is np-dimensional and the Givens transform, $\Phi(Y)$, maps this set to an $np - p(p+1)/2$-dimensional set of angles, Φ. To obtain the correct morphing factor, we appeal to the calculus of differential forms, which roughly-speaking, measures how a transform warps an infinitesimal volume from one space to another. For accessibility, we provide psuedo-code in the supplementary materials, as well as actual Stan code.

For $n \times p$ orthonormal matrices, there are $np - p(p+1)/2$ free parameters and so the proper form to measure sets of orthonormal matrices is a $np - p(p+1)/2$-form. For an orthonormal, $n \times p$ matrix, Y, we can find an orthonormal $n \times n$ matrix G such that $G^T Y = I_{n,p}$. In fact G just comes from the product of the appropriate rotation matrices that arises in the Givens Reduction, Q. Muirhead [16] shows that the correct form for measuring volumes on the Stiefel manifold comes from wedging the elements of the $n \times p$ matrix $G^T dY$ that lie below the diagonal i.e.

$$\bigwedge_{i=1}^{p} \bigwedge_{j=i+1}^{n} G_j^T \, dY_i, \tag{4}$$

where G_j is the jth column of G and Y_i is the ith column of Y. To obtain the form in angle coordinates, we obtain dY_i in terms of the angle 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 (4) 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 into the log-probability of a model to avoid the sort of unintended sampling behavior depicted in Figure 2. We incorporate the form (4) in to the log-probability of all of our Stan examples.

6 Empirical Studies

To demonstrate the use of the Givens Transform, we construct several models with orthonormal matrices or unit vectors as parameters and perform fully Bayesian inference on them in Stan.

6.1 Avoiding Unidentifiability in Neural Network Models Using Unit-Vectors

A single layer neural network with Rectified Linear Unit(ReLU) nonlinearities and H hidden nodes maps inputs $x \in \mathbb{R}^D$ to outputs $y \in \mathbb{R}$ via the relationship $y = \max\{0, W_1 x + b_1\} W_2$ where $W_1 \in \mathbb{R}^{H \times D}$ and $b_1, W_2 \in \mathbb{R}^H$. It is well known that several equivalent values of W_1, b_1, and W_2 result in the same exact function. For example, the ith row of W_1 and the ith entry of b_1 can be scaled by α as long as the ith entry of W_2 is scaled by $1/\alpha$ [6, p.277].

In Bayesian analysis, this leads to thin, ridge-like posteriors with high curvatures that are difficult to sample and approximate using VI, often times leading to variational posteriors that underestimate model uncertainty (Figure 3). By replacing the columns of W_1 with unit-vectors using the Givens Transform, and sampling over the space of unconstrained angles, we can obtain much more well behaved posteriors.

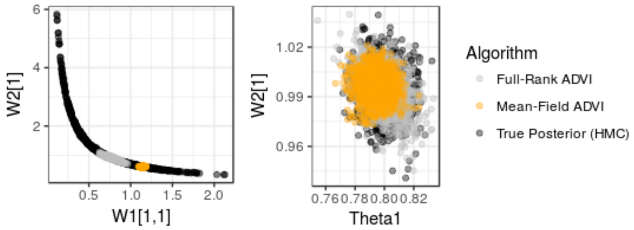

Figure 3: (Left) Posterior HMC samples of $W_1[1,1]$ and $W_2[1]$ reveal how the scaling unidentifiability of ReLu networks manifests as ridge-like posteriors that are difficult to sample and approximate using VI, as oppose to the more well-behaved posteriors in the Givens coordinates (Right).

6.2 Probabilistic PCA

Factor Analysis (FA) and Probabilistic PCA (PPCA) [24] posit a probabilistic generative model where high-dimensional data is determined by a linear function of some low-dimensional latent state [17, Chapt. 12]. Geometrically, for a three-dimensional set of points forming a flat, pancake-like cloud, PCA can be thought of as finding the best 2-frame that aligns with this cloud (Figure 4). Formally, PPCA posits the following generative process for how a sequence of high-dimensional data vectors $\mathbf{x}_i \in \mathbb{R}^n$, $i = 1, \cdots, N$ arise from some low dimensional latent representations $\mathbf{z}_i \in \mathbb{R}^p$ $(p < n)$, via a linear transformation, or matrix $W \in \mathbb{R}^{n \times p}$:

$$\begin{aligned} p(\mathbf{z}_i) &\sim \mathcal{N}_p(0, I) \\ p(\mathbf{x}_i|\mathbf{z}_i, W, \sigma^2) &\sim \mathcal{N}_n(W\mathbf{z}_i, \sigma^2 I). \end{aligned} \tag{5}$$

A closed-form maximum likelihood estimator for W is known for this model in the limit as $\sigma^2 \to 0$, but as we shall see, for more complicated models/likelihoods, closed-form maximum-likelihood estimators are almost never known. This has often been dealt with by using Expectation Maximization (EM) in these models to obtain a point estimate [17, Chapt. 12.2.5]. In Bayesian inference we are typically interested in the entire distribution over possible solutions, i.e. a posterior distribution over unknown parameters to quantify uncertainty.

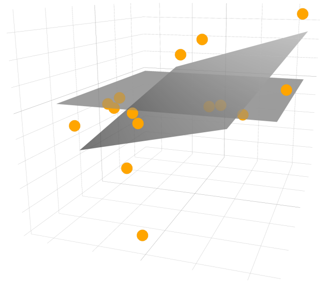

Figure 4: PCA finds the orthonormal matrix in the Stiefel Manifold that best describes the subspace that the data lie in. In the Figure, the point estimate misleads us from the true subspace, which in this case is the flat plane.

One can show that the W parameter in PPCA is unidentifiable [17, chapt. 12.1.3], as it can be rotated to achieve an identical likelihood; thus the model must be changed to make the posterior distribution interpretable. Furthermore, this rotational unidentifiability manifests in the log-likelihood function as regions in parameter space where large curvature arise, causing numerical problems in HMC, as pointed out by Holbrook et al. [10]. To this end, those in the Bayesian dimensionality reduction community have used a modified form of the model (5), whereby the matrix W is replaced by a new term $W\Lambda$ where W is an $n \times p$ orthonormal matrix and Λ is a $p \times p$ diagonal matrix with positive elements [2; 10].

6.2.1 Test on Synthetic Data

We used the Givens Transform to fit this modified PPCA using Stan's NUTS inference algorithm, which otherwise would be unusable on this model with an orthonormal matrix parameter. We generated a synthetic, three-dimensional dataset that lies on a two-dimensional plane with $N = 15$ observations according to the modified version of (5) (data shown in Figure 4). We choose $\mathrm{diag}(\Lambda) = \mathrm{diag}(1, 1)$, $\sigma^2 = 1$, and W to be $I_{3,2}$, which in the Givens representation corresponds to $\theta_{12} = \theta_{13} = \theta_{23} = 0$ i.e. the horizontal plane, which contrasts with the slanted plane that we obtain from a classical PCA maximum likelihood estimate (Figure 4). In this case the advantage of the full posterior estimate the Bayesian framework affords is clear. Posterior samples of θ_{13}, which if we recall from Figure 1 is the Givens Transform angle that controls the upwards tilt of the plane, reveal a wide posterior which cautions us against the spurious maximum likelihood estimate of $\hat{\theta}_{13} = -0.15$ (Figure 5).

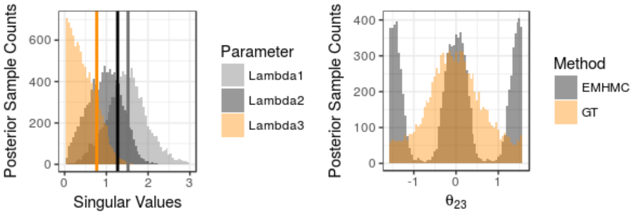

Figure 5: Inference for three-dimensional synthetic data. (Left) Posterior draws of the Λ parameter are more informative in dimensionality selection than plane point estimates. (Right) By limiting the angles of rotation in the Givens Transform we can avoid unidentifiability in our problem and eliminate multimodal posteriors that show up in other methods such as EMHMC.

We also note that the representation of the Givens Transform can in certain models allow us to avoid issues of identifiability that are present in the sampling algorithms of [1; 2]. While most identifiability issues are alleviated by using the modified PPCA with an orthonormal matrix, the PPCA likelihood is still equivalent for an orthonormal matrix W and any permutation of the columns of W being negative [17; 10, Chapt. 12.1.3]. Taking a geometric view of the Stiefel Manifold (Figure 1), this means that a mirroring of the p-frame would yield an identical value in the likelihood of even the modified PPCA. As such, even the methods of Brubaker et al. [1] and Byrne and Girolami [2] will lead to multi-modal posteriors that can be avoided in a straightforward manner by simply limiting the angles in the Givens Transform from a range of $(-\pi, \pi)$ to a range of $(-\pi/2, \pi/2)$, a change that is much more evidently afforded when working in the angle coordinates (Figure 5 [right]).

As a practical matter, if the true basis lies near a pole, i.e. θ_{ij} is close to $-\pi/2$ or $\pi/2$, then posteriors might still tend to be multi-modal as the region in parameter space close to the boundaries will be nearly equally valid, while the region near zero will not be valid and thus contain little probability mass. In these cases, one can simply change the coordinate bounds (chart) so that $\theta_{ij} \in (0, \pi)$ will have a unimodal posterior in the new coordinate system, alleviating possible exploration issues in HMC. In Stan this is straightforward, as one simply has to change the lower and upper bound of the angle parameter.

6.3 Hierarchical subspace models for grouped multi-view medical data

We modeled grouped multi-view hospital data for injured patients using a hierarchical CCA model [17, Chapt. 15.2]. CCA can model two types (or views) of data as being a function of two respective latent low dimensional states, but also a common latent state that captures the common information contained in both views (Figure 6 [left]). In our case we compared blood protein measurements and clot strength measurements for injured patients belonging to one of four groups, depending on the type of injury. While the four types of injuries were different enough so that we could not use a single CCA model to capture the characteristics of all models at once, the four groups were not so different as to warrant separate CCA models for each. To share information between the CCA models, we placed a hierarchical prior over the angles of the Givens Transform representing the distinct orthonormal matrix parameter for each group.

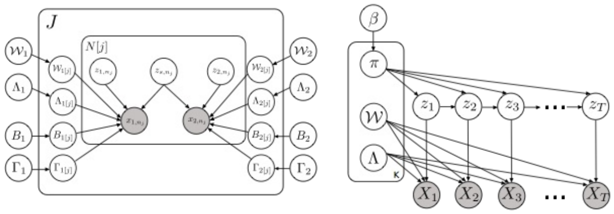

Figure 6: Probabilistic graphical models for Hierarchical CCA Model (left) and Network HMM (right).

While distributions on the Stiefel Manifold such as the Matrix Langevin distribution [16] exist, these distributions are difficult to use in practice, as computing their density requires evaluating an expensive matrix sum [8]. By appealing to the Givens Transform and placing a hierarchical prior over the angles of the different orthonormal matrices, we were able to build a hierarchical model over subspaces, a previously intractable task. The hierachical prior “shrinks” the posterior median of the orthonormal matrices towards a common mean in addition to reducing the variance of these estimates (Figure 7). This is particularly helpful for groups with only a smaller number of observations such as the SW group, which contains only 16 patients, in comparison with the GSW group of 86 patients. Comparing the angle between the first principal components for the SW and GSW groups illustrates how using a hierarchical prior shrinks estimates of subspaces together towards a common hierarchical subspace (Figure 8).

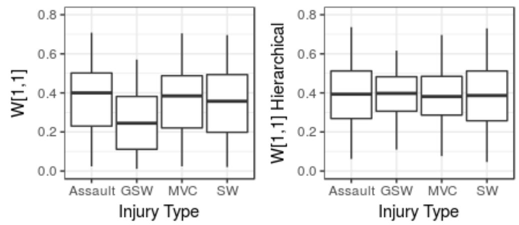

Figure 7: (Left) When estimated separately, estimates of the matrix parameter W have high uncertainty. (Right) Placing a hierarchical prior over these matrices with GT-PPCA shrinks these parameters to a common hierarchical mean and results in smaller posterior intervals.

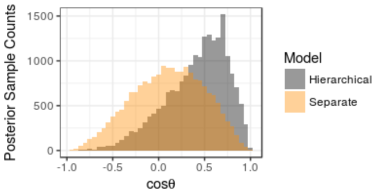

Figure 8: Geometrically the respective first principal components of two different groups are shrunk closer together in a hierarchical model.

6.4 Social Networks

We built an HMM subspace model for count data to model the hidden time-dependent structure of a social network of school children. RF sensors were used to track the interactions between school children in 12 different classes (two classes for grades 1-6) for an entire school day so as to better understand how disease spreads throughout a network [23]. We collated the number of interactions between each pair of classes into 11-minute contiguous time windows, giving us 177 symmetric matrices of counts representing the network interactions between different classrooms throughout the day (Figure 9 [lower row]). We modeled the elements of these count matrices as each coming from a Poisson distribution with rate defined by a symmetric matrix $R = \exp(W\Lambda W^T)$, where the orthonormal matrix W captures the low-dimensional structure of the network. To model the time varying structure of the network, we posited that the network was always in one of three latent states, that evolve according to a Markov Chain (Figure 6 [right]). The three states each have their own associated orthonormal matrix W_i that captures the low-dimensional latent network structure for that state.

The posterior modes capture the latent structure of the rate matrices of the three hidden states (Figure 9 top row). Interaction rates are visibly low when students are in class, high during lunch, and nonexistent when out of school out of class. Posteriors of the orthonormal components of the rate matrices are shown in (Figure 10 [left]). We also generated samples from the posterior distribution over states from posterior samples of the Markov Chain, enabling us to provide a posterior over which of the hidden states the network is in at a given time (Figure 10 [right]), a common inference task in disease networks as well as fMRI networks.

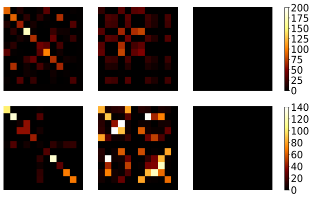

Figure 9: Posterior modes of rate matrices for the three states (top) capture the pattern found in example count matrices belonging to each of these three states (bottom).

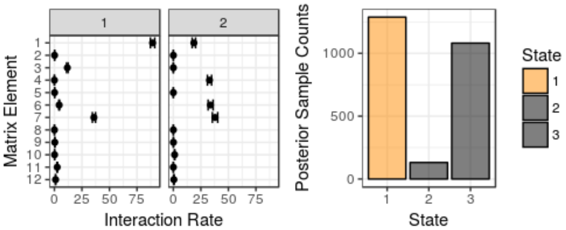

Figure 10: (Left) Posterior intervals from GT-PPCA with NUTS capture uncertainty in the orthonormal matrix estimates for the first two columns of the rate matrix for the first hidden state. (Right) Posterior draws can tell us the posterior probability that the network was in a certain state given the data.

7 Discussion

We introduce the Givens Transform, a parameter transformation that represents orthonormal matrices in an alternative coordinate system, enabling the construction and inference of complex probabilistic models with unit-vector and orthonormal matrix parameters in a probabilistic framework like Stan. We show using real-world examples how one can use Stan and the Givens Transform to obtain uncertainties over such parameter all while avoiding multi-modal posteriors and analyzing the parameters using an easy to understand angle representation.

Acknowledgements

Acknowledgements temporarily anonymized for review.

References

[1] M. Brubaker, M. Salzmann, and R. Urtasun. A family of mcmc methods on implicitly defined manifolds. In *Artificial Intelligence and Statistics*, pages 161–172, 2012.

[2] S. Byrne and M. Girolami. Geodesic monte carlo on embedded manifolds. *Scandinavian Journal of Statistics*, 40(4):825–845, 2013.

[3] B. Carpenter, A. Gelman, M. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. A. Brubaker, J. Guo, P. Li, and A. Riddell. Stan: A probabilistic programming language. *Journal of Statistical Software*, 20, 2016.

[4] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian data analysis*, volume 2. Chapman & Hall/CRC Boca Raton, FL, USA, 2014.

[5] Z. Ghahramani, G. E. Hinton, et al. The em algorithm for mixtures of factor analyzers. Technical report, Technical Report CRG-TR-96-1, University of Toronto, 1996.

[6] I. Goodfellow, Y. Bengio, and A. Courville. *Deep learning*. MIT press, 2016.

[7] T. Hamelryck, J. T. Kent, and A. Krogh. Sampling realistic protein conformations using local structural bias. *PLoS Computational Biology*, 2 (9):e131, 2006.

[8] P. D. Hoff. Simulation of the matrix bingham–von mises–fisher distribution, with applications to multivariate and relational data. *Journal of Computational and Graphical Statistics*, 18(2): 438–456, 2009.

[9] M. D. Hoffman and A. Gelman. The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. *Journal of Machine Learning Research*, 15(1):1593–1623, 2014.

[10] A. Holbrook, A. Vandenberg-Rodes, and B. Shahbaba. Bayesian inference on matrix manifolds for linear dimensionality reduction. *arXiv preprint arXiv:1606.04478*, 2016.

[11] A. Kucukelbir, R. Ranganath, A. Gelman, and D. Blei. Fully automatic variational inference of differentiable probability models. In *NIPS Workshop on Probabilistic Programming*, 2014.

[12] B. Leimkuhler and S. Reich. *Simulating hamiltonian dynamics*, volume 14. Cambridge University Press, 2004.

[13] F. Lu and E. Milios. Robot pose estimation in unknown environments by matching 2d range scans. *Journal of Intelligent and Robotic systems*, 18(3): 249–275, 1997.

[14] C. D. Meyer. *Matrix analysis and applied linear algebra*, volume 2. Siam, 2000.

[15] S. Mohamed, Z. Ghahramani, and K. A. Heller. Bayesian exponential family pca. In *Advances in neural information processing systems*, pages 1089–1096, 2009.

[16] R. J. Muirhead. *Aspects of multivariate statistical theory*, volume 197. John Wiley & Sons, 2009.

[17] K. P. Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

[18] R. M. Neal et al. Mcmc using hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, 2 (11), 2011.

[19] S.-H. Oh, L. Staveley-Smith, K. Spekkens, P. Kamphuis, and B. S. Koribalski. 2d bayesian automated tilted-ring fitting of disk galaxies in large hi galaxy surveys: 2dbat. *Monthly Notices of the Royal Astronomical Society*, 2017.

[20] R. Ranganath, S. Gerrish, and D. Blei. Black box variational inference. In *Artificial Intelligence and Statistics*, pages 814–822, 2014.

[21] S. I. Resnick. *A probability path*. Springer Science & Business Media, 2013.

[22] J. Salvatier, T. V. Wiecki, and C. Fonnesbeck. Probabilistic programming in python using pymc3. *PeerJ Computer Science*, 2:e55, 2016.

[23] J. Stehlé, N. Voirin, A. Barrat, C. Cattuto, L. Isella, J.-F. Pinton, M. Quaggiotto, W. Van den Broeck, C. Régis, B. Lina, et al. High-resolution measurements of face-to-face contact patterns in a primary school. *PloS one*, 6(8): e23176, 2011.

[24] M. E. Tipping and C. M. Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):611–622, 1999.

[25] D. Tran, A. Kucukelbir, A. B. Dieng, M. Rudolph, D. Liang, and D. M. Blei. Edward: A library for probabilistic modeling, inference, and criticism. *arXiv preprint arXiv:1610.09787*, 2016.