
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 factors are given by the columns of the low
22 rank factorization. Geometrically, traditional PCA can be interpreted as providing a point estimate of
23 a low dimensional hyper-plane that is closest to a cloud of data points. For binary or integer valued
24 matrices, such as graph adjacency matrices arising in network science, Matrix Factorization (NMF)
25 [7] or Exponential Family PCA (EPCA) [4] are often used to find such low dimensional latent factors.

26 Probabilistic PCA (PPCA) [13] and Bayesian Exponential Family PCA (BXPCA) [10], posit proba-
27 bilistic generative models that are equivalent to PCA in the limit of decreasing noise [12, chapt. 12.2].
28 This probabilistic approach is attractive because it allows a straightforward way, via Bayesian in-
29 ference, to quantify the uncertainty in our estimates (to prevent overfitting) and conduct hypothesis
30 testing, as is often used in the sciences. Furthermore, probabilistic models are amenable to expansion
31 and can serve as modules within larger probabilistic graphical models. This is important in a real-
32 world setting where true generative models are often complex and where they do not necessarily
33 follow the simple generative process set forth by PPCA. We illustrate both of these properties in our
34 examples Section (ref).

In practice, conducting full Bayesian inference on PPCA and related models involves one or more unknown orthonormal matrices. This is difficult because orthonormal matrices form a rather particular subset out of all possible realizable matrices yielding a probability measure on a manifold within the full space with 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 $V_{n,p}$ (cite). When first considering this problem, it may seem that this geometry may rule out a straight-forward 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 are sure to lead to posteriors that assign mass to invalid regions of parameter space.

To address these issues, there has been some recent work on posterior inference of orthonormal matrices based on modified versions of Hamiltonian Monte Carlo (HMC) with sampling for constrained parameters (cite). These methods use modified numerical integrators to ensure that exploration only occurs in valid areas of parameter space where the necessary constraints are satisfied (cite). Because these methods use modified integrators for constrained parameters, they require approaches in practice for keeping track of the type of each variable and for which type of integrator to use on each variable. This adds an extra layer of algorithmic complexity, especially for large complex probabilistic graphical models. This further makes it difficult to implement these methods in a scalable fashion and within widely used probabilistic programming languages and packages, such as Stan [3]. Furthermore, we have found while these methods represent some progress on this challenging problems the numerical methods are sometimes not rigorous or robust and at times can suffer from errors and instabilities that jeopardize the integrity of samples. In addition, we have found these methods can at times induce unnecessary multi-modality that are an artifact of the chosen representation and not intrinsic to the model under study. This arises in part from the choice of representations directly in terms of unconstrained matrices in these methods (cite). As a consequence, the matrix in the PPCA model has equal likelihood when making an arbitrary change of sign of one of its columns. Given this equivalent relation this can result in significant multi-modality that must be handled with some care in the final analysis.

In the case of positively constrained parameters, posterior distributions are in practice rather routinely inferred using both MCMC and VI. This can be accomplished by transforming the constrained variables to an unconstrained space using the one-to-one mapping $f : \text{supp}(z_{\text{constr}}) \rightarrow \mathbb{R}^D$. The $\text{supp}(z_{\text{constr}})$ is the support of the constrained random variable z_{constr} . To our knowledge, no such transformation has been proposed in the past to map orthonormal matrices to a comperable unconstrained space.

In this paper, we introduce just such a transform based on the Givens Reductions of a matrix. This expresses the matrix in terms of a sequence of funamental rotations through given angles. We call our approach the Givens Transform (GT) PCA or (GT-PCA) for short. The transform allows for a straight-forward implementation of PPCA and related models that involve orthonormal matrices. This allows for use in stand-alone HMC implementations or in the context of a probabilistic programming languages like Stan (cite). We first discuss the details of our approach and then illustrate how this can be done using Stan. A particular advantage of using Stan is the ability to use the built-in NUTS and VI implementations. When we consider more complicated models in our examples Section (ref), we will show how the Givens Transform also provides through this alternative representation of orthonormal matrices useful ways to work with and interpret our models. For instance, by limiting the range of the Givens angle representation we can avoid naturally the issues involved with other representations that have unneccessary multi-modal posteriors. Furthermore, our alternative representation 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 comperable to the Langevin prior (cite).

We give a brief overview of probabilistic dimensionality reduction in Section 2. We discuss how orthonormal matrices arise naturally for performing inference and introduce a precise description of the Stiefel manifold in Section 2.2. We also briefly review prior work on Bayesian dimensionality

reduction using orthonormal matrices in Section 2.2. We introduce the Givens Transform (GT) and use for reductions to obtain our representations in Section 2.4. We also discuss practical consideration for GT in computing and using the representation in Section 2.4. We then show how our methods can be used in practice for Bayesian inference on a few probabilistic graphical models in Section (ref). This includes illustration of the basic aspects of the technique on XYZ and presenting a more advanced real-world application to patient data for factors affecting coagulopathy in Section (ref). We conclude discussing how our GT-PPCA methods can be utilized on other problems and by discussing our implementation in Stan that can be utilized by others interested in our approach.

2 Probabilistic Dimensionality Reduction

PJA: Discuss general context here... linear dimensionality reductions... motivations briefly etc...

PJA: Discuss specific approach... PPCA...

2.1 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 using a much 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 see and estimate of the subspace. Formally, in PPCA this consists of using the following generative process

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

The W is an $n \times p$ orthonormal matrix and Λ is a $p \times p$ diagonal matrix with positive elements. For simplicity in our presentation of PPCA, we have assumed here that the data has only zero mean but the more general case can also readily be considered (cite). We also remark that generative model could also be expanded to the case of non-Gaussian data by replacing equation 1 with an exponential family member whose natural parameters are given by $\text{Expon}(W\Lambda\mathbf{z}^i)$ where $\text{Expon}(\cdot)$ is an appropriate link function [10].

2.2 Importance of the Orthonormality Condition

We mention that the orthonormal constraint on the matrix W plays an important role in obtaining robust methods for in making inferences in probabilistic PCA. If one were to relax the orthonormal constraint the likelihood function would assign identical probability to a whole equivalence class of matrices $W \sim V$ where the span is the same linear subspace $\text{span}\{W\} = \text{span}\{V\}$. While this might not seem theoretically too problematic, in practice this presents a number of major challenges. This first is that the matrices in a given equivalence class are not all equally well-conditioned numerically in defining the linear subspace and round-off errors and truncation errors become problematic in practical calculations. Secondly, these issues with the representation further manifest in the log-likelihood objective function where regions arise of particularly large curvature. This causes significant numerical issues for variational inference (VI) in nonlinear optimization methods and in monte-carlo (MC) approaches with samplers having slow mixing times Murphy [12], Mohamed et al. [10], Holbrook et al. [6, chapt. 12.1.3]. These issues render making inferences in PPCA with such non-orthonormal matrices impractical resulting from a statistics stand-point in effective unidentifiability.

2.3 Inference using the Stiefel Manifold

To develop robust approaches for PPCA we restrict inferences to be over matrices that are constrained to be orthonormal Murphy [12, chapt. 12.1.3]. In the space of $n \times p$ orthonormal matrices defines the Stiefel Manifold [11]

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

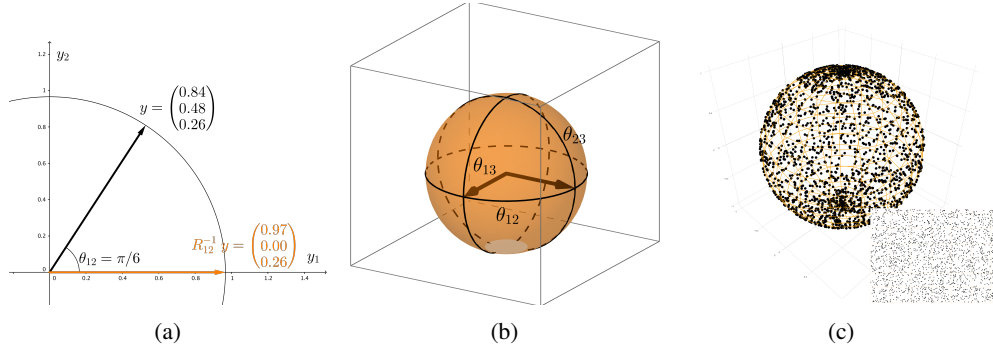


Figure 1: Visualizing the Givens Transform. (a) How the Givens Reduction “zeros out” a column vector. (b) A geometric view of the Stiefel manifold, two-frame in three dimensions. (c) Sampling without a proper measure adjustment. **PJA: Good start on figures. Please remember to use vector format .PDF final output. Can use Inkscape or Adobe Illustrator to process PDF and SVG files from plotters like Python Matplotlib or Matlab or Stan, etc... Please be sure to save both .SVGZ (source file) and .PDF files in Git repo. This allows for easy edits later to figures as needed. PJA: Fonts are a little too small on Fig. a. Color in Fig. b would be better to use light blue or something more neutral. Use Red for labels... remember color balance. Point sizes a little bigger in in-set in Fig. C. Save both sphere dots and square separately and use Inkscape / Illustrator to combine.**

137 The Stiefel Manifold has an intrinsic dimension of $\dim V^{n,p} = np - p(p+1)/2$. This arises from
 138 the constraints on the columns of the matrix that impose orthonormality. This dimensionality can
 139 be seen by observing, that the first column of $Y \in V_{n,p}$ must have norm one and hence has one
 140 constraint placed on it. The second column must also have norm one and also must be orthogonal to
 141 the first column hence with two constraints placed on it. Continuing to the third column through the
 142 n^{th} one arrives at the conclusion that each point of the Stiefel Manifold has only $np - p(p+1)/2$
 143 degrees of freedom. We shall discuss how to develop coordinate-charts for representing points within
 144 the Stiefel Manifold using Givens transformations in Section (ref).

145 2.4 Givens Transform (GT) approach to PPCA (GT-PPCA)

146 **PJA: I would give a streamlined discussion of the GT-PPCA approach giving the basics of the GT**
 147 **and how the representation is used.**

148 **PJA: I will move much of the material below to an appendix for now and move pieces or re-writes**
 149 **up into this section as needed to make clear. Can always refer reader to the Appendix for the more**
 150 **technical parts. A streamlined discussion GT should be given enough to develop the PPCA, but not**
 151 **too much technicalities on differential geometry and all the rest that an expert could piece together. It**
 152 **is an art at what level to cover things and I can certainly help with that.**

153 3 Examples

154 3.1 Synthetic Data

155 Show how point estimation of subspace and dimensionality can be misleading. Show multi-modality
 156 of Symone’s method.

157 3.2 FMRI

158 Illustrate hypothesis testing.

159 3.3 School Network

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

162 3.4 Coagulopathy using hierarchical subspace models

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

164 4 Discussion

165 Acknowledgments

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

168 References

- 169 [1] Marcus Brubaker, Mathieu Salzmann, and Raquel Urtasun. A family of mcmc methods on
170 implicitly defined manifolds. In *Artificial Intelligence and Statistics*, pages 161–172, 2012.
- 171 [2] Simon Byrne and Mark Girolami. Geodesic monte carlo on embedded manifolds. *Scandinavian*
172 *Journal of Statistics*, 40(4):825–845, 2013.
- 173 [3] Bob Carpenter, Andrew Gelman, Matt Hoffman, Daniel Lee, Ben Goodrich, Michael Betan-
174 court, Michael A Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. Stan: A probabilistic
175 programming language. *Journal of Statistical Software*, 20, 2016.
- 176 [4] Michael Collins, Sanjoy Dasgupta, and Robert E Schapire. A generalization of principal
177 component analysis to the exponential family. In *Nips*, volume 13, page 23, 2001.
- 178 [5] Alan Edelman. 18.325: Finite random matrix theory volumes and integration. 2005.
- 179 [6] Andrew Holbrook, Alexander Vandenberg-Rodes, and Babak Shahbaba. Bayesian inference on
180 matrix manifolds for linear dimensionality reduction. *arXiv preprint arXiv:1606.04478*, 2016.
- 181 [7] Daniel D Lee and H Sebastian Seung. Algorithms for non-negative matrix factorization. In
182 *Advances in neural information processing systems*, pages 556–562, 2001.
- 183 [8] Benedict Leimkuhler and Sebastian Reich. *Simulating hamiltonian dynamics*, volume 14.
184 Cambridge University Press, 2004.
- 185 [9] Carl D Meyer. *Matrix analysis and applied linear algebra*, volume 2. Siam, 2000.
- 186 [10] Shakir Mohamed, Zoubin Ghahramani, and Katherine A Heller. Bayesian exponential family
187 pca. In *Advances in neural information processing systems*, pages 1089–1096, 2009.
- 188 [11] Robb J Muirhead. *Aspects of multivariate statistical theory*, volume 197. John Wiley & Sons,
189 2009.
- 190 [12] Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.
- 191 [13] Michael E Tipping and Christopher M Bishop. Probabilistic principal component analysis.
192 *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):611–622,
193 1999.

194 Appendix

195 4.1 Givens Reductions of Matrices

196 **PJA:** Give a basic description of the background on Givens reductions of a matrix that we can refer to
197 with additional details than we need in the main text.

198 4.2 Differential Geometry of the Steifel Manifold

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

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

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

$$(R_{1n}^{-1} \cdots R_{13}^{-1} R_{12}^{-1})Y = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & * & \cdots & * \\ 0 & * & \cdots & * \\ \vdots & \vdots & \vdots & \vdots \\ 0 & * & \cdots & * \\ 0 & * & \cdots & * \\ \vdots & \vdots & \vdots & \vdots \\ 0 & * & \cdots & * \end{pmatrix}. \quad (4)$$

$$(R_{2n}^{-1} \cdots R_{23}^{-1})(R_{1n}^{-1} \cdots R_{12}^{-1})Y = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & * \\ 0 & 0 & \cdots & * \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & * \\ 0 & 0 & \cdots & * \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & * \end{pmatrix}. \quad (5)$$

203 Continuing in this fashion yields

$$(R_{pn}^{-1} \cdots R_{p,p+1}^{-1}) \cdots (R_{2n}^{-1} \cdots R_{23}^{-1})(R_{1n}^{-1} \cdots R_{12}^{-1})Y = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 \end{pmatrix} = I_{np}. \quad (6)$$

204 5 Misc LaTeX

205 5.1 Related works using HMC

206 **PJA: I'd move this material in some form to the introduction and give a brief overview of how out**
 207 **methods are different. Then maybe mention briefly during discussions prior work when our approach**
 208 **differs and explain why (as you do below, but in a sentence or two in the particular context).**

209 Before discussing details of our approach in more detail, we briefly discuss related prior works.
 210 Recently, two approaches to posterior inference of orthonormal matrices and other constrained
 211 parameters have been proposed. These methods work by modifying the leap-frog integrator used to
 212 simulate Hamiltonian dynamics in HMC to ensure that any trajectory in parameter space remains on
 213 the Stiefel manifold. Brubaker et al. [1] uses the SHAKE integrator [8] to ensure trajectories through
 214 parameter space continuously satisfy the required constraints. The integrator works by repeatedly
 215 taking a step forward that may be off the manifold using ordinary leap frog, then projecting back
 216 down to the nearest point on the manifold. This projection is done via Newton iterations, which may
 217 not converge in practice, possibly jeopardizing the ergodicity of a Markov chain.

218 Byrne and Girolami [2] took a different approach, exploiting the fact that closed form solutions are
 219 known for the geodesic equations over the Stiefel manifold in the embedded coordinates, W . While
 220 this method is completely explicit, requiring no Newton iterations, in practice we found that for larger
 221 step sizes, the integrator steps off the Stiefel manifold, due to the numerical imprecision of the matrix
 222 exponential function.

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

230 5.2 Givens Reductions

231 **PJA: I've temporarily moved this to the misc text section, to be added back into the main exposition**
 232 **in a more streamlined way.** We provide a brief exposition on Givens Reductions, which motivate the
 233 Givens Transform, then describe the Givens Transform along with relevant practical considerations.

234 Define R_{ij} to be the $n \times n$ rotation matrix that performs a counter-clockwise rotation in the (i, j) -
 235 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
 236 counter-clockwise rotation of angle θ_{ij} in the (x, y) , (x, z) and (y, z) planes respectively. Rotation
 237 matrices have the following key properties:

- 238 1. They preserve length and angles between vectors, i.e. for two vectors $u, v \in \mathbb{R}^n$, $R_{ij}u, R_{ij}v$
 239 are the same length as u and v respectively, and if u and v are orthogonal then so are $R_{ij}u$
 240 and $R_{ij}v$.

241 2. They are invertible and their inverse is their transpose $R_{ij}^{-1} = R_{ij}^T$. Their inverse corresponds
 242 to a clockwise rotation in the (i, j) -plane.

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

247 Similarly, we can apply consecutive rotations $R_{13}^{-1}, R_{14}^{-1}, \dots, R_{1n}^{-1}$ to this result so that all entries
 248 of the first column besides the first element are zero. The first column of the resulting matrix,
 249 $R_{1n}^{-1} \dots R_{13}^{-1} R_{12}^{-1} Y$, will be the length one vector $(1 \ 0 \ \dots \ 0)^T$ which lies entirely along the first axis.
 250 If one takes the perspective that these rotations are applied to the columns of Y , then with the two
 251 properties of rotation matrices we mentioned earlier, it is evident that columns 2 through n must have
 252 zero in their first element because they will be orthogonal to the first column, $(1 \ 0 \ \dots \ 0)^T$.

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

$$(R_{pn}^{-1} \dots R_{p,p+2}^{-1} R_{p,p+1}^{-1}) \dots (R_{2n}^{-1} \dots R_{24}^{-1} R_{23}^{-1})(R_{1n}^{-1} \dots R_{13}^{-1} R_{12}^{-1})Y = I_{n,p}, \quad (7)$$

256 where $I_{n,p}$ consists of the first p columns of the $n \times n$ identity matrix. This process of applying
 257 consecutive rotation matrices to a matrix is known in numerical analysis as the Givens reduction
 258 [9], and is applied more generally to square matrices for matrix-vector solves. In total we will have
 259 applied $(n-1) + (n-2) + \dots + (n-p) = np - p(p+1)/2$ rotations matrices. We note that
 260 because rotation matrices are very sparse in high dimensions, multiplication by a rotation matrix is
 261 computationally much less intensive than matrix multiplication otherwise would be.

262 5.3 Givens Transform

263 Because rotation matrices are invertible, equation 7 implies that we can rewrite the $n \times p$ orthonormal
 264 matrix Y as the product of counter-clockwise rotation matrices and $I_{n,p}$:

$$Y = (R_{12} \dots R_{1n}) \dots (R_{23} \dots R_{2n})(R_{p+1,n} \dots R_{pn})I_{n,p}. \quad (8)$$

265 Recall that each of the $np - p(p+1)/2$ rotation matrices have an associated angle
 266 $(\theta_{12} \dots \theta_{1n}) \dots (\theta_{23} \dots \theta_{2n})(\theta_{p+1,n} \dots \theta_{pn})$, that we collectively refer to as Θ . In this way we
 267 have reparameterized all $n \times p$ orthonormal matrices¹, a constrained space, in terms of unconstrained
 268 angles², using a transform $\Theta : V_{n,p} \rightarrow \mathbb{R}^{np-p(p+1)/2}$. We refer to 8 as the Givens representation or
 269 Givens transform.

270 5.3.1 A geometric perspective

271 The Givens transform can be visualized in the simple case when $n = 3$ and $p = 2$. This is the case
 272 where data is three-dimensional and we seek a two-dimensional subspace to represent the data. In
 273 this case, we can imagine a two dimensional bases, visualized as two perpendicular vectors that rotate
 274 rigidly together along three angles of rotations. This is referred to as a 2-frame in differential geometry
 275 [11] and is depicted in Figure 1b. Recall θ_{12} controls the amount of rotation in the $(1, 2)$ -plane, θ_{13}
 276 in the $(1, 3)$ -plane, and θ_{23} controls a rotation of the second basis vector about the first. All rotations
 277 will preserve the length and orthogonality of the basis vectors. Given a flat, pancake-like cloud of
 278 points in 3D, the two dimensional bases will move according to the appropriate angles to be aligned
 279 with the cloud of points, if one were conducting a (Maximum A-Posteriori) MAP estimation over the
 280 Stiefel manifold.

¹other than a set of measure zero we explain in the next subsection

²the 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

5.3.2 Practical considerations of topology and angles

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 $S \subset V_{n,p}$, that 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 $\theta_{23} \in (-\pi/2, \pi/2)$. Luckily this set is of measure zero (under the proper measure for the Stiefel manifold, see section 5.3.4), and thus, with probability one, the orthonormal matrix that describes the true subspace our data lie in will not be in that set.

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

Lastly, we note that if the true bases lies near a pole, i.e. θ_{ij} is close to $-\pi/2$ or $\pi/2$, then posteriors will tend to be multi-modal as the region in parameter space close to the boundaries will be close to 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 chart so that $\theta_{ij} \in (0, \pi)$, creating a uni-modal posterior in the new coordinate system, and alleviating numerical issues. In Stan this is straight-forward, as one simply has to change the lower and upper bound of the angle parameter.

5.3.3 Jacobian under a change of variables

The Givens transform 8 allows us to represent orthonormal matrices as angles $Y(\Theta)$. This in turn allows us to write probability densities $p_Y(Y)$ in terms of angles, so that we can conduct inference in an unconstrained space. It is well known in probability theory that the transform of a random variable is in general not the density of the transform, i.e. $p_\Theta(\Theta) \neq p_Y(Y(\Theta))$ [12, 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:

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

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.

5.3.4 Differential forms

We offer an intuitive high-level overview of differential forms. For a thorough account we recommend Muirhead [11], Edelman [5], 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

330 $p_{\text{Euc}}(x, y, z)$ on the sphere, the “natural” way to measure the probability of a spherical random
 331 variable falling within some area on the surface of that sphere is by first covering that area with tiny
 332 rectangles tangent to the sphere, taking the average density of each rectangle, multiplying that density
 333 by the area of that rectangle, and summing over the resulting products. The probability of the random
 334 variable falling within that area is defined to be the limit of that result as the size of the rectangles
 335 go to zero. Area forms, or two-forms can be thought of as these infinitesimal rectangles. They are
 336 written as the wedge product of two differentials, e.g. $dx \wedge dy$, and they are objects that are inserted
 337 into integrals to obtain measures. Differential forms form (no pun intended) a vector space useful
 338 for measuring areas and high-dimensional volumes. For example, the area forms on a sphere can be
 339 written as a linear combination of the standard-basis two forms

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

340 One can then integrate this over a sub-area of the sphere to obtain the measure of that sub-area.
 341 The beauty of differential forms is that a different coordinate system such as polar coordinates,
 342 simply correspond to using a different basis for representing forms, which again are vectors. In fact
 343 writing a form in a different bases simply involves taking partial derivatives, e.g. $dx = (\partial x / \partial \theta) d\theta +$
 344 $(\partial x / \partial \varphi) d\varphi$. If we substitute the forms in the angle bases in to 10 and simplify using the well-defined
 345 anti-symmetric and distributive properties of wedge products and differential forms (see [11]), we
 346 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
 347 then be integrated using a double integral in polar coordinates. The absolute value of this area-form
 348 evaluated at some point specified in angle coordinates, (θ_0, φ_0) , intuitively measures how much the
 349 area of tiny square in angle space gets shrunk or stretched when the area is mapped to an area on the
 350 sphere.

351 Analogously, we can measure volumes on the Stiefel manifold. For $n \times p$ orthonormal matrices, there
 352 are only $np - p(p + 1)/2$ free parameters and so the proper form to measure sets of orthonormal
 353 matrices is in fact a $np - p(p + 1)/2$ -form. For an orthonormal, $n \times p$ matrix, Y , we can find an
 354 orthonormal $n \times n$ matrix G such that $G^T Y = I_{n,p}$. In fact G just comes from the product of the
 355 appropriate rotation matrices that comes from the Givens Reduction 7. Muirhead [11] shows that
 356 the correct form comes from wedging the elements of the $n \times p$ matrix $G^T dY$ that lie below the
 357 diagonal i.e.

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

358 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
 359 coordinates we simply obtain dY in angle coordinates. dY_i can be obtained in terms of the angle
 360 coordinates by the following relationship, $dY_i = J_{Y_i}(\Theta) d\Theta$, where J_{Y_i} is the Jacobian of Y_i with
 361 respect to the angle coordinates. Once we obtain the form 11 in terms of the angle coordinates, the
 362 result is a wedge product of $np - p(p + 1)/2$ vectors that are $np - p(p + 1)/2$ dimensional, which
 363 reduces to the determinant of these vectors aligned side by side as a $np - p(p + 1)/2 \times np - p(p + 1)/2$
 364 matrix. This determinant is analogous to and serves the same purpose as Jacobian adjustment that
 365 comes from transforming random variables. We can insert it in to the log-probability of a model to
 366 avoid the sort of unintended sampling behavior depicted in Figure 1c. We incorporate the form 11 in
 367 to the log-probability of all of our Stan examples.