Bayesian Inference over the Stiefel Manifold via the Givens Representation

김성민

서울대학교
통계학과, 베이즈통계 연구실

2021. 08. 18

CONTENTS

CONTENTS

Stiefel Manifold

- We denote it as $V_{p,n}$ which are known as p-frames.
- A p-frame is an orthogonal set of p n-dimensional unit-length vector, where $p \leq n$
- p-frames naturally correspond to $p \times n$ orthogonal matrices.

$$V_{p,n} = \{Y \in \mathbb{R}^{n \times p} Y^T Y = I\}$$

- A simple case is $V_{1,3}$, which consists of a single vector, u_1, on the unit sphere.

Motivation

- Statistical models parameterized in terms of orthogonal matrices are ubiquitous.
- Posterior inference in Bayesian models with orthogonal matrix parameters remains a challenge.
  - Inference under constraint. (HMC-based methods rely on specialized HMC update rules)
  - Transforming the parameters to an unconstraint space. (pose challenges related to the change in measure, topology, and parameterization)

CONTENTS

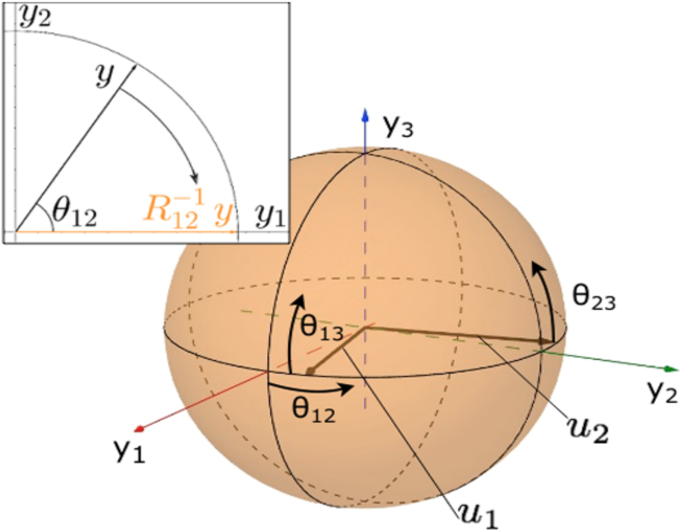
Givens Rotations and Reductions

- Given any $n \times p$ matrix, A, the Givens reduction algorithm is a numerical algorithm for finding the QR-factorization of A (an $n \times p$ orthogonal matrix Q,$p \times p$ upper-triangular matrix R, such that A=QR)
- The algorithm works by successively applying a series of Givens rotation matrices so as to "zero-out" the elements $\{A_{ij} : i > j\}$ of A

Givens Rotations and Reductions

- The rotation matrix $R_{ij}(\theta_{ij})$ has the effect of rotating the vector counter-clockwise in the (i,j)-plane.
- $R_{ij}(\theta_{ij})$, form of an identity matrix except for the (i,i) and (j,j) positions which are replaced by $\cos(\theta_{ij})$, and the (i,j) and (j,i) positions which are replaced by $-\sin(\theta_{ij})$ and $\sin(\theta_{ij})$ repectively.

Givens Rotation



Givens Rotation

$$\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1p} \\ a_{21} & a_{22} & \cdots & a_{2p} \\ a_{31} & a_{32} & \cdots & a_{3p} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{np} \end{pmatrix} \Rightarrow \begin{pmatrix} * & * & \cdots & * \\ 0 & * & \cdots & * \\ a_{31} & a_{32} & \cdots & a_{3p} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{np} \end{pmatrix} \Rightarrow \cdots \Rightarrow \begin{pmatrix} * & * & \cdots & * \\ 0 & * & \cdots & * \\ 0 & * & \cdots & * \\ \vdots & \vdots & \ddots & \vdots \\ 0 & * & \cdots & * \end{pmatrix} \Rightarrow \cdots \Rightarrow \begin{pmatrix} * & * & \cdots & * \\ 0 & * & \cdots & * \\ 0 & 0 & \cdots & * \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{pmatrix}$$

Figure 2: The Givens reduction eliminates lower diagonal elements of an $n \times p$ matrix one column at a time. Because each rotation, $R_{ij}(\theta_{ij})$, only affects rows i and j, previously zeroed out elements do not change.

$$\begin{aligned} R_* &= R_{pn}^{-1}(\theta_{pn}) \cdots R_{p,p+1}^{-1}(\theta_{p,p+1}) \cdots R_{1n}^{-1}(\theta_{1n}) \cdots R_{12}^{-1}(\theta_{12}) A \\ &\Rightarrow Q_* R_* = A \end{aligned}$$

Givens representation

- When applied to an $n \times p$ orthogonal matrix Y, the Givens reduction yields

$$R_{pn}^{-1}(\theta_{pn}) \cdots R_{p,p+1}^{-1}(\theta_{p,p+1}) \cdots R_{1n}^{-1}(\theta_{1n}) \cdots R_{12}^{-1}(\theta_{12}) Y = I_{n,p}$$
$$Y = R_{12}(\theta_{12}) \cdots R_{1n}(\theta_{1n}) \cdots R_{p,p+1}(\theta_{p,p+1}) R_{pn}(\theta_{pn}) I_{n,p}$$

- $\Theta = (\theta_{12} \cdots \theta_{1n} \cdots \theta_{23} \cdots \theta_{2n} \cdots \theta_{p,p+1} \cdots \theta_{pn})$ effectively parameterizing the Stiefel manifold.

CONTENTS

Transformation of Measure

- $Y = R_{12}(\theta_{12}) \cdots R_{1n}(\theta_{1n}) \cdots R_{p,p+1}(\theta_{p,p+1}) R_{pn}(\theta_{pn}) I_{n,p}$
- $p_\Theta(\theta) = p_Y(y(\theta)) |J_{Y(\Theta)}(\theta)|$
- Unfortunately, the Givens representation, $Y(\Theta)$, is a map from a space of dimension $d = np - p(p+1)/2$ to a space of dimension np.
- Its Jacobian is non-square and the determinant of the Jacobian is undefined.

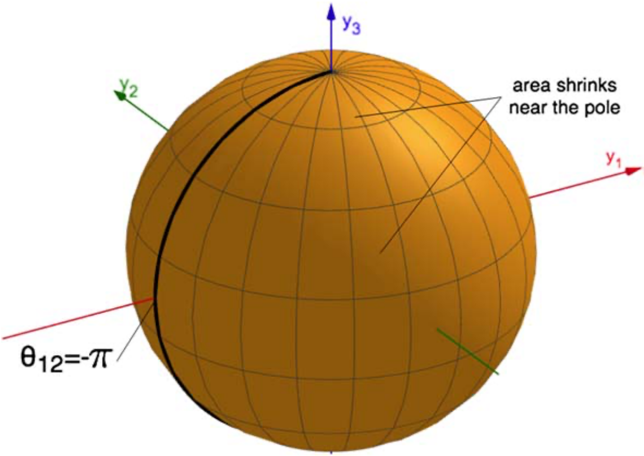
Transformation of Measure

- Muirhead (2009)
- Edelman (2005)
- James (1954)
- $J_{Y(\Theta)}(\theta) = \Pi_{i=1}^{p} \Pi_{j=i+1}^{n} \cos^{j-i-1} \theta_{ij}$

Issue(1)

- Let $\theta_{12}, \theta_{23}, \cdots \theta_{p,p+1}$ range from $-\pi$ to π (longitudinal coordinates)
- Let remaining coordinates range from $-\pi/2$ to $\pi/2$ (latitudinal coordinates)
- $\theta_{12} = \pi$ are disconnected

Issue(1)



- To address this issue, auxiliary parameters to the Givens representation is introduced.

Auxiliary Parameters for Addressing Connectedness

- Introduce for each angle parameter, θ_{ij}, an independent auxiliary parameter, r_{ij}.
- $x_{ij} = r_{ij} \cos\theta_{ij}$ and $y_{ij} = r_{ij} \sin\theta_{ij}$

$$p_{x,y}(x, y) = p_{\theta,r}\left(arctan(y/x), \sqrt{x^2 + y^2}\right) \frac{1}{r}$$

- In practice, we set $p_r(r)$ to a $N(1, 0.1^2)$

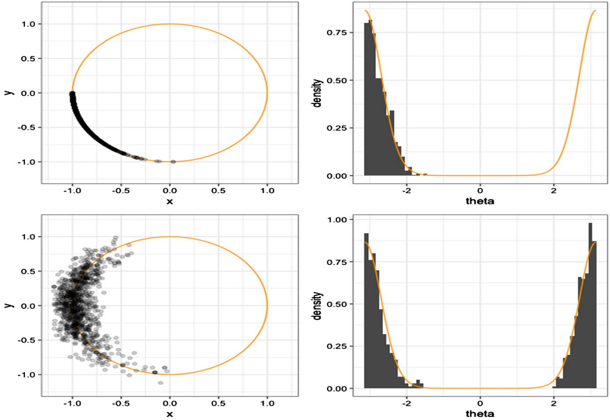
Von Mises distribution

- The Von Mises pdf for the angle x is given by:

$$f(x|\mu,\kappa) = \frac{\exp\{\kappa\cos(x-\mu)\}}{2\pi I_0(\kappa)}$$

- μ : measure of location
- κ : measure of concentration
  - If κ is zero, the distribution is uniform.
  - If κ is large, the distribution becomes very concentrated about the angle μ

Auxiliary Parameters for Addressing Connectedness



- Upper : 1000 samples from Von Mises distribution ($\mu = -\pi$, $\kappa = 5$) sampled over the space $\theta \in [-\pi, \pi]$
- Lower : 1000 samples from the equivalent distribution sampled over the (x,y)-space.

Issue(2)

- $J_{Y(\Theta)}(\theta) = \Pi_{i=1}^{p} \Pi_{j=i+1}^{n} \cos^{j-i-1} \theta_{ij}$
- $J_{Y(\Theta)}(\theta)$ approaches zero near the "poles" (where the latitudinal coordinates equal $-\pi/2$ or $\pi/2$)
- In practice, this prevents algorithms such as HMC from obtaining samples in a small region near the "poles".

Transformation of Densities Near the Poles

- Limiting all latitudinal angles to the region $[-\pi/2+\epsilon, \pi/2-\epsilon]$ where ϵ is a small value.

Each of these individual probabilities is proportional to $\cos^{j-i-1}\theta_{ij}$, which for small ϵ can be bounded by ϵ^{j-i-1} over the interval $[\pi/2-\epsilon, \pi/2]$. Thus the probability of falling within the ϵ-region is bounded by a constant times the following quantity:

$$\sum_{i=1}^{p}\sum_{j=i+2}^{n} 2\int_{\pi/2-\epsilon}^{\pi/2} \epsilon^{j-i-1} d\theta_{ij} = \sum_{i=1}^{p}\sum_{j=i+2}^{n} 2\epsilon^{j-i} = \sum_{i=1}^{p}\mathcal{O}(\epsilon^2) = \mathcal{O}(p\epsilon^2). \tag{4.7}$$

Transformation of Densities Near the Poles

p	n	$\epsilon = 0.1$	$\epsilon = 0.05$	$\epsilon = 0.025$	$\epsilon = 0.0125$	$\epsilon = 1e-5$
1	10	490	114	22	4	0
1	20	499	118	25	4	0
1	50	570	148	32	6	0
3	10	1,612	381	79	15	0
3	20	1,665	398	78	19	0
3	50	1,712	416	100	24	0
10	10	4,260	1,071	258	59	0
10	20	5,342	1,336	357	91	0
10	50	5,266	1,368	334	90	0

Table 1: The number of uniform samples out of 100,000 that fell within the ϵ region for various values of n, p, and ϵ. Samples are taken uniformly from the Stiefel manifold using the QR factorization method. As the theoretical bound suggests, the number of samples falling in this region increases modestly for fixed p and increasing n. It increases linearly with p, and it decreases quadratically with ϵ. In particular, whenever ϵ is halved, the number of samples falling within the region decreases by about a fourth. We also note that for $\epsilon = 1e-5$, the value we used for most of our experiments, the number of samples falling within the ϵ region was zero for all settings.

Transformation of Densities Near the Poles

	Givens					Wood
κ	$\epsilon = 0.1$	$\epsilon = 0.05$	$\epsilon = 0.025$	$\epsilon = 0.0125$	$\epsilon = 1e-5$	
1	1.2027	1.2042	1.2008	1.1995	1.1986	1.2012
10	0.4181	0.4065	0.4031	0.4012	0.4019	0.4015
100	0.1657	0.1377	0.1290	0.1258	0.1261	0.1255
1,000	0.1092	0.0657	0.0483	0.0422	0.0396	0.0398

Table 3: The empirical expectation of the principal angle, $\arccos(\mu^T Y)$, sampled under the von Mises Fisher distribution with $\mu = (0, 0, 1)$ and $\kappa = 1, 10, 100$ and 1000 using the Givens representation in Stan with various sizes of the ϵ area and using the method of Wood (1994). As ϵ decreases, the empirical expectation computed using the Givens representation becomes much closer to those taken via the method of Wood (1994). For small κ the expectations do not differ much even for large ϵ because much less mass concentrates near the ϵ regions.

CONTENTS

Probabilistic Principal Component Analysis(PPCA)

- PPCA posits the following generative process for how a sequence of high-dimensional data vectors $x_i \in \mathbb{R}^n$, $i = 1, , N$ arise from some low-dimensional latent representations $z_i \in \mathbb{R}^p$ $(p < n)$.

$$z_i \sim N_p(0, I)$$
$$x_i | z_i, W, \Lambda, \sigma^2 \sim N_n(W\Lambda z_i, \sigma^2 I).$$

- To ensure identifiability, W is constrained to be an orthogonal $n \times p$ matrix while Λ is a diagonal matrix with positive, ordered elements.

$$x_i | W, \Lambda, \sigma^2 \sim N_n(0, C)$$
$$C = W\Lambda^2 W^T + \sigma^2 I$$

Probabilistic Principal Component Analysis(PPCA)

$$
\begin{aligned}
p(x_1, \cdots, x_N | W, \Lambda, \sigma^2) &= -\frac{N}{2} \log|C| - \frac{1}{2} \sum x_i^T c^{-1} x_i \\
&= -\frac{N}{2} \log|C| - \frac{N}{2} tr(c^{-1} \hat{\Sigma})
\end{aligned}
$$

- Traditional PCA corresponds to the closed-form maximum likelihood estimator for W in the limit as $\sigma^2 \to 0$
- Sampling the posterior of a model both provides a measure of uncertainty for parameter estimates and is possible even for more elaborate models.

Probabilistic Principal Component Analysis(PPCA)

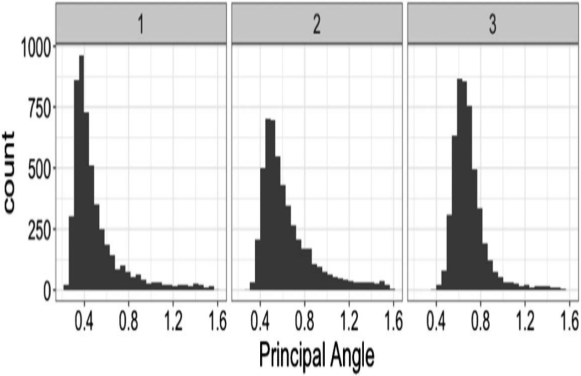
- Setting
  - $n = 50, \; p = 3$
  - $sigma^2$: uniform prior
  - W ; uniform prior over the Stiefel manifold
  - $\Lambda^2 = diag(5, 3, 1.5), \; \sigma^2 = 1$
- In the Gives representation

$$p(\Theta, \Lambda, \sigma^2 | x_1, \cdots, x_N) \propto p(x_1, , x_N | W(\Theta), \Lambda, \sigma^2) | J_{Y(\Theta)}(\theta) |$$

Histogram of principal angle

- W_j : columns of posterior draws of W
- E_j : columns of the first three eigenvectors of

$$\phi = \arccos\left(E_j^T W_j\right),\ j = 1, 2, 3$$



Marginal posterior distributions of the elements of W

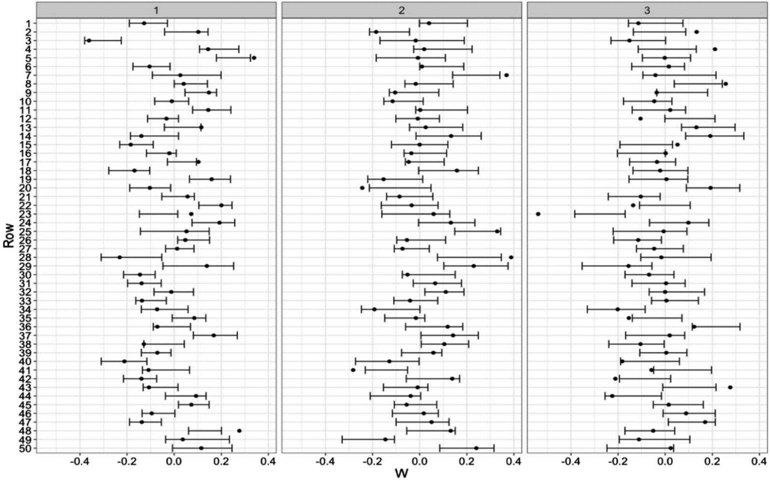


Figure 10: True values of W used in the simulation along with 90% credible intervals computed using draws of the posterior. Each facet corresponds to one of the three columns of W.

References I

Edelman, A. (2005). 18.325: Finite random matrix theory: Jacobians of matrix transforms (with wedge products).

James, A. T. (1954). Normal multivariate analysis and the orthogonal group, *The Annals of Mathematical Statistics* **25**(1): 40–75.

Muirhead, R. J. (2009). *Aspects of multivariate statistical theory*, Vol. 197, John Wiley & Sons.

Pourzanjani, A. A., Jiang, R. M., Mitchell, B., Atzberger, P. J. and Petzold, L. R. (2021). Bayesian inference over the stiefel manifold via the givens representation, *Bayesian Analysis* **16**(2): 639–666.