<table>
<thead>
<tr>
<th><strong>Statistica Sinica Preprint No:</strong> SS-2016-0017</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Title</strong></td>
</tr>
<tr>
<td><strong>Manuscript ID</strong></td>
</tr>
<tr>
<td><strong>URL</strong></td>
</tr>
<tr>
<td><strong>DOI</strong></td>
</tr>
</tbody>
</table>
| **Complete List of Authors** | Lizhen Lin  
Vinayak Rao  
David Dunson |
| **Corresponding Author** | Lizhen Lin |
| **E-mail** | lizhen.lin@austin.utexas.edu |
Bayesian nonparametric inference on the Stiefel manifold

Lizhen Lin, Vinayak Rao and David Dunson

Department of Statistics and Data Sciences, The University of Texas at Austin, USA.
Department of Statistics, Purdue University, USA.
Department of Statistical Science, Duke University, USA.

Abstract: The Stiefel manifold $V_{p,d}$ is the space of all $d \times p$ orthonormal matrices, with the $d-1$ hypersphere and the space of all orthogonal matrices constituting special cases. In modeling data lying on the Stiefel manifold, parametric distributions such as the matrix Langevin distribution are often used; however, model misspecification is a concern and it is desirable to have nonparametric alternatives. Current nonparametric methods are mainly Fréchet-mean based. We take a fully generative nonparametric approach, which relies on mixing parametric kernels such as the matrix Langevin. The proposed kernel mixtures can approximate a large class of distributions on the Stiefel manifold, and we develop theory showing posterior consistency. While there exists work developing general posterior consistency results, extending these results to this particular manifold requires substantial new theory. Posterior inference is illustrated on a dataset of near-Earth objects.

Key words and phrases: Bayesian nonparametric, kernel mixture, matrix Langevin, orthonormal matrices, posterior consistency, Stiefel manifold, von Mises Fisher

1. Introduction

Statistical analysis of matrices with orthonormal columns has diverse applications including principal components analysis, estimation of rotation matrices, as well as in analyzing orbit data of the orientation of comets and asteroids. Central to probabilistic models involving such matrices are probability distributions on the Stiefel manifold, the space of all $d \times p$ orthonormal matrices. Popular examples of parametric distributions are the matrix von Mises-Fisher distribution (Khatri and Mardia (1977); Hornik and Grün (2013)) (also known as the matrix Langevin (Chikuse (1993, 2003a, 2006))), and its generalization, the Bingham-von Mises-Fisher distribution (Hoff (2009)). Maximum likelihood estimation is often used in estimating the parameters, while recently Rao et al. (2016) proposed a sampling algorithm allowing Bayesian inference for such distributions.
Current parametric models are overly simple for most applications, and non-parametric inference has been mostly limited to estimation of Fréchet means (Bhattacharya and Bhattacharya (2012); Bhattacharya and Lin (2016)). Chikuse (1998) proposes a frequentist density estimator using kernel density estimation on the Stiefel manifold and studies its asymptotic behavior. Model-based non-parametric Bayesian inference has several advantages, including providing a fully generative model for prediction and characterization of uncertainty, while allowing adaptation to the complexity of the data. We propose a class of nonparametric models based on mixing parametric kernels on the Stiefel manifold. Such models have appealing properties including large support, posterior consistency, and straightforward computation adapting the sampler of Rao et al. (2016). Depending on the application, our models can be used to characterize the data directly, or to describe latent components of a hierarchical model.

Section 2 provides some details on the geometry of the Stiefel manifold. Section 3 introduces the matrix Langevin distribution, the nonparametric model, and the posterior consistency theory. Section 4 illustrates the model through application to an object orbits data set. All proofs are included in the appendix and our code is available at https://github.com/varao/stiefel.

2. Geometry of the Stiefel manifold

The Stiefel manifold $V_{p,d}$ is the space of all $p$-frames in $\mathbb{R}^d$, with a $p$-frame consisting of $p$ ordered orthonormal vectors in $\mathbb{R}^d$. Writing $M(d, p)$ for the space of all $d \times p$ real matrices, and letting $I_p$ represent the $p \times p$ identity matrix, the Stiefel manifold can be represented as

$$V_{p,d} = \{X \in M(d, p) : X^T X = I_p\}. \quad (2.1)$$

The Stiefel manifold $V_{p,d}$ has the $d - 1$ hypersphere $S^{d-1}$ as a special case when $p = 1$. When $p = d$, this is the space of all the orthogonal matrices $O(d)$. $V_{p,d}$ is a Riemannian manifold of dimension $dp - p - p(p - 1)/2 = p(2d - p - 1)/2$. It can be embedded into the Euclidean space $M(d, p)$ of dimension $dp$ with the inclusion map as a natural embedding, and is thus a submanifold of $\mathbb{R}^{dp}$.

Let $G \in V_{p,d}$, and $G_1$ be a matrix of size $d \times (d - p)$ such that $[G : G_1]$, the augmented matrix obtained by concatenating the rows of $G$ and $G_1$, is in $O(d)$, the group of $d$ by $d$ orthogonal matrices. The volume form on the manifold is $\lambda(dG) = \Lambda_{i=1}^p \Lambda_{j=i+1}^d g_j^T d g_i$ where $g_1, \ldots, g_p$ are the columns of $G$, $g_{p+1}, \ldots, g_d$...
Bayesian inference on the Stiefel manifold

are the columns of $G_1$, and $\wedge$ represents the wedge product (Muirhead (2005)). If $p = d$, that is when $G \in O(d)$, one can represent $\lambda(dG) = \wedge_{i<j} g_{ij}^T dg_i$. Note that $\lambda(dG)$ is invariant under the left action of the orthogonal group $O(d)$ and the right action of the orthogonal group $O(p)$, and forms the Haar measure on the Stiefel manifold. For more details on the Riemannian structure of the Stiefel manifold, we refer to Edelman et al. (1998).

3. Bayesian nonparametric model

Let $X$ be a random variable on $V_{p,d}$. A popular parametric distribution of $X$ is the matrix Langevin distribution that has, with respect to the invariant Haar volume measure on $V_{p,d}$, the density

$$P_{\text{ML}}(X|F) = \text{etr}(F^T X) / Z(F),$$

(3.1)

where etr stands for the exponential trace function. The parameter $F$ is a $d \times p$ matrix, and the normalization constant $Z(F) = \text{e}_1(F_{1/2}^T F_{1/2})$ is the hypergeometric function with matrix arguments (Herz (1955)), evaluated at $\frac{1}{4} F^T F$ (Chikuse (2003b)). Write the singular value decomposition (SVD) of $F$ as $F = GH^T$, with $G$ and $H$, $d \times p$ and $p \times p$ orthonormal matrices, and $\kappa$ a diagonal matrix with positive elements. One can think of $G$ and $H$ as orientations, with $\kappa$ controlling the concentration in the directions determined by these orientations. Large values of $\kappa$ imply concentration along the associated directions, while setting $\kappa$ to zero recovers the uniform distribution on the Stiefel manifold. Khatri and Mardia (1977) show that $\text{e}_1(F_{1/2}^T F_{1/2}) = \text{e}_1(F_{1/2}^T F_{1/2})$, so that the normalization constant depends only on $\kappa$, and we write it as $Z(\kappa)$. The mode of the distribution is given by $GH^T$ and, from the characteristic function of $X$, one can show $E(X) = FU$, where the $(i,j)$th element of the matrix $U$ is given by

$$U_{ij} = 2 \frac{\partial \log \text{e}_1(F_{1/2}^T F_{1/2})}{\partial (F^T F)_{ij}}.$$

Consider $n$ observations $X_1, \ldots, X_n$ drawn i.i.d. from $P_{\text{ML}}(X|F)$. A simple approach to characterizing these observations is via a maximum likelihood estimate of the parameter $F$. This is complicated by the dependence of the normalization constant $Z(F)$ on $F$, and Chikuse (2003b) describes an approach based on an asymptotic approximation to $Z(F)$. The intractable normalizing constant makes Bayesian estimation of $F$ even more challenging, since quantifying
the effects of such approximations is very difficult. Rao et al. (2016) propose an exact sampling scheme based on a data augmentation technique to solve this intractability problem.

In many situations, assuming the observations come from a particular parametric family such as matrix Langevin is restrictive, and raises concerns about model misspecification. Nonparametric alternatives, on the other hand, have much wider applicability, and we consider these in the following.

Denote by $\mathcal{M}$ the space of all densities on $V_{p,d}$ with respect to the Haar measure $\lambda$. Let $g(X,G,\kappa)$ be a parametric kernel on the Stiefel manifold with a ‘location parameter’ $G$ and a vector of concentration parameters $\kappa = (\kappa_1, \ldots, \kappa_p)$. One can place a prior $\Pi$ on $\mathcal{M}$ by modelling the random density $f$ as

$$f(X) = \int g(X,G,\kappa)P(d\kappa d G), \quad (3.2)$$

with the mixing measure $P$ a random probability measure. A popular prior over $P$ is the Dirichlet process (Ferguson (1973)), parametrized by a base probability measure $P_0$ on the product space $\mathbb{R}_+^p \times V_{p,d}$, and a concentration parameter $\alpha > 0$. We denote by $\Pi_1$ the DP prior on the space of mixing measures, and assume $P_0$ has full support on $\mathbb{R}_+^p \times V_{p,d}$.

The model in (3.2) is a ‘location-scale’ mixture model, and corresponds to an infinite mixture model where each component has its own location and scale. One can also define the following ‘location’ mixture model given by

$$f(X) = \int g(X,G,\kappa)P(dG)\mu(d\kappa), \quad (3.3)$$

where $P$ is given a nonparametric prior like the DP and $\mu(d\kappa)$ is a parametric measure (like a product of Gamma or Weibull distributions). In this model, all components are constrained to have the same scale parameters $\kappa$. This model is analogous to a mixture of Gaussians with all components constrained to have the same covariance. We show later that with an appropriate prior over $\kappa$, this constrained model is still asymptotically consistent. However, in practical settings, care must be taken to ensure that this assumption is appropriate, if not, the model can infer an inappropriately large number of mixture components.

When $\Pi_1$ corresponds to a DP prior, one can precisely quantify the mean of
the induced density \( \Pi \). For model (3.2), the prior mean is given by

\[
E(f(X)) = \int g(X, G, \kappa)E(P(d\kappa dG)) = \int g(X, G, \kappa)P_0(d\kappa dG),
\]

(3.4)

while for model (3.3), it is

\[
E(f(X)) = \int g(X, G, \kappa)\mu(d\kappa)P_0(dG).
\]

(3.5)

The parameter \( \alpha \) governs the number of components in the mixing density and roughly controls the concentration of the prior around the mean density, and one can place a hyperprior on \( \alpha \) as well.

In the following, we set \( g(X, G, \kappa) \) to be the matrix Langevin distribution with parameter \( F = GK \). Thus,

\[
g(X, G, \kappa) = \operatorname{etr}(\kappa G^T X)/Z(\kappa) = C(\kappa)\operatorname{etr}(\kappa G^T X),
\]

(3.6)

with \( C(\kappa) = 1/Z(\kappa) = 1/0F_1(\frac{1}{2},\frac{1}{4}\kappa^T \kappa) \). We have restricted ourselves to the special case where the matrix Langevin parameter \( F \) has orthogonal columns (or equivalently, where \( H = I_p \)). While it is easy to apply our ideas to the general case, we demonstrate below that even with this restricted kernel, our nonparametric model has such properties as large support and consistency.

### 3.1. Posterior consistency

With our choice of parametric kernel, a DP prior on \( \Pi_1 \) induces an infinite mixture of matrix Langevin distributions on \( \mathcal{M} \). Call this distribution \( \Pi \); we will show that this has large support on \( \mathcal{M} \), and that the resulting posterior distribution concentrates around any true data generating density in \( \mathcal{M} \). Our modelling framework and theory builds on Bhattacharya and Dunson (2010, 2012), who developed consistency theorems for density estimation on compact Riemannian manifolds, and considered DP mixtures of kernels appropriate to the manifold under consideration. However, they only considered simple manifolds, and showing that our proposed models have large support and consistency properties requires substantial new theory.

We first introduce some notions of distance and neighborhoods on \( \mathcal{M} \). A weak neighborhood of \( f_0 \) with radius \( \epsilon \) is defined as

\[
W_\epsilon(f_0) = \left\{ f : \left| \int z\lambda(dX) - z f_0 \lambda(dX) \right| \leq \epsilon, \text{for all } z \in C_b(V_p,d) \right\},
\]

(3.7)
where $C_b(V_{p,d})$ is the space of all continuous and bounded functions on $V_{p,d}$. The Hellinger distance $d_H(f, f_0)$ is defined as

$$d_H(f, f_0) = \left( \frac{1}{2} \int (\sqrt{f(X)} - \sqrt{f_0(X)})^2 \lambda(dX) \right)^{1/2}.$$  

We let $U_\epsilon(f_0)$ denote an $\epsilon$-Hellinger neighborhood around $f_0$ with respect to $d_H$. The Kullback-Leibler (KL) divergence between $f_0$ and $f$ is defined to be

$$d_{KL}(f_0, f) = \int f_0(X) \log \frac{f_0(X)}{f(X)} \lambda(dX),$$

with $K_\epsilon(f_0)$ denoting an $\epsilon$-KL neighborhood of $f_0$.

Let $X_1, \ldots, X_n$ be $n$ observations drawn i.i.d. from some true density $f_0$ on $V_{p,d}$. Under our model, the posterior probability $\Pi_n$ of some neighborhood $W_\epsilon(f_0)$ is given by

$$\Pi_n(W_\epsilon(f_0)|X_1, \ldots, X_n) = \frac{\int_{W_\epsilon(f_0)} \prod_{i=1}^n f(X_i) \pi(df)}{\int_{M} \prod_{i=1}^n f(X_i) \Pi(df)}.$$ (3.9)

The posterior is weakly consistent if for all $\epsilon > 0$,

$$\Pi_n(W_\epsilon(f_0)|X_1, \ldots, X_n) \rightarrow 1 \ a.s. \ Pf_0^\infty \text{ as } n \rightarrow \infty,$$ (3.10)

where $Pf_0^\infty$ represents the true probability measure for $(X_1, X_2, \ldots)$.

We assume the true density $f_0$ is continuous with $F_0$ as its probability distribution. The following theorem on the weak consistency of the posterior under the mixture prior is for both models (3.2) and (3.3), the proof of which is included in the Appendix.

**Theorem 1.** The posterior $\Pi_n$ in the DP-mixture of matrix Langevin distributions is weakly consistent.

We now consider the consistency property of the posterior $\Pi_n$ with respect to the Hellinger neighborhood $U_\epsilon(f_0)$; this is referred to as strong consistency.

**Theorem 2.** Let $\pi_\kappa$ be the prior on $\kappa$, and let $\Pi$ be the prior on $\mathcal{M}$ induced by $\Pi_1$ and $\pi_\kappa$ via the mixture model (3.3). Let $\Pi_1 \sim DP_{\alpha P_0}$ with $P_0$ a base measure having full support on $V_{p,d}$. Assume $\pi_\kappa(\kappa : \phi(\kappa) \leq n^a) \leq \exp(-n\beta)$ eventually for some $a < 1/(p + 2)dp$ and $\beta > 0$ with $\phi(\kappa) = \sqrt{\sum_{i=1}^p (\kappa_i + 1)^2}$. Then the posterior $\Pi_n$ is consistent with respect to the Hellinger distance $d_H$. 

Bayesian inference on the Stiefel manifold

**Remark 1.** For prior $\pi_\kappa$ on the concentration parameter $\kappa$ to satisfy the condition $\pi_\kappa(\kappa : \phi(\kappa) \leq n^a) < \exp(-n\beta)$, for some $a < 1/(dp(p+2))$ and $\beta > 0$, requires fast decay of the tails for $\pi_\kappa$. One can check that an independent Weibull prior for $\kappa_i$, $i = 1, \ldots, p$ with $\kappa_i \sim \kappa_i^{(1/a) - 1} \exp(-b\kappa_i^{(1/a)})$ satisfies the tail condition.

Another choice is to allow $\pi_\kappa$ to be sample size dependent as suggested by Bhattacharya and Dunson (2012). In this case, one can choose independent Gamma priors for $\kappa_i$ with $\kappa_i \sim \kappa_i c \exp(-b n \kappa_i)$ where $c > 0$ and $n^{1-a}/b \to 0$ with $0 < a < 1/(dp(p+2))$.

### 4. Inference for the nonparametric model

A common approach to posterior inference for the Dirichlet process is Markov chain Monte Carlo (MCMC) based on the Chinese restaurant process (CRP) representation of the DP (Neal (2000)). The Chinese restaurant process describes the distribution over partitions of observations that results from integrating out the random probability measure $\Pi_1$, and a CRP-based Gibbs sampler updates this partition by reassigning each observation to a cluster conditioned on the rest. The probability of an observation $X_i$ joining a cluster with parameters $(G, \kappa)$ is proportional to the likelihood $g(X_i, G, \kappa)$ times the number of observations already in that cluster (for an empty cluster, the latter is the concentration parameter $\alpha$). Our case is complicated by the intractable likelihood $g(\cdot)$; this also makes updating the cluster parameters not straightforward. One possibility is to use an asymptotic approximation to the normalization constant $Z(\kappa)$ (Hoff (2009)). We instead use a recently proposed data augmentation scheme by Rao et al. (2016) to construct a Markov chain with the exact stationary distribution.

This approach builds on a rejection sampling scheme by Hoff (2009) that produces samples from a matrix Langevin distribution by accepting or rejecting proposals from a simpler, tractable distribution on the Stiefel manifold. Under this scheme, every sample from the matrix Langevin distribution is preceded by a sequence of rejected proposals from the proposal distribution. Updating the parameters of this proposal distribution (which are the same parameters as the matrix Langevin distribution) is easy, however this requires imputing the rejected proposals that precede each observation. Rao et al. (2016) show how to carry out this step, and thus run MCMC on the augmented (and now tractable) space.
We refer the reader to that paper for more details about this auxiliary variable Gibbs sampler. Below we detail the steps of the algorithm. We write $\theta = (\kappa, G)$, and $q_\theta(X)$ for the proposal distribution of Hoff (2009).

**Algorithm 1**: MCMC sampler for DP mixture of Matrix Langevin distributions

**Input**: A partition of observations and a set of cluster parameters

**Output**: A new partition and a new set of cluster parameters

1. **Update cluster assignments**: For each observation $x$:
   - Let $\theta^*$ be the parameter of its current cluster.
   - Sample from $q_{\theta^*}$ till acceptance (Hoff (2009)), calling the rejected proposals $Y$.
   - Treat the vector $(x, Y)$ as the actual observation, with the likelihood corresponding to drawing its components independently from the tractable $q_\theta$. Under this likelihood, assign $(x, Y)$ to a new cluster according to the usual Chinese restaurant process (Neal (2000)). Then discard $Y$.

2. **Update cluster parameters**: For each cluster $c$ (with parameters $\theta^* = (\kappa^*, G^*)$):
   - Write $X_c$ for all observations at this cluster, and $|X_c|$ for the cardinality.
   - Sample independently from $q^*_c$ until $|X_c|$ samples are accepted.
   - Write $Y_c$ for all rejected proposals.
   - Update $\kappa^*$ as if $(X_c, Y_c)$ were observations at this cluster with likelihood $q_\theta$.
   - Discard $Y_c$ and update $G^*$.

We apply our nonparametric model to a dataset of near-Earth astronomical objects (comets and asteroids). Inferences were based on 5,000 samples from the MCMC sampler, after a burn-in period of 1,000 samples.

### 4.1. Near Earth Objects dataset

The Near Earth Objects dataset was collected by the Near Earth Object Program of the National Aeronautics and Space Administration and consists of 162 measurements of Near-Earth Comets (NECs). Each data point characterizes the orientation of a two-dimensional elliptical orbit in three-dimensional space.

Downloaded from http://neo.jpl.nasa.gov/cgi-bin/neo_elem
Bayesian inference on the Stiefel manifold

Figure 1: The Near Earth Objects dataset (left), and the adjacency matrix inferred by the DP mixture model (right)

Figure 2: Posterior over the number of clusters for the Near Earth Objects dataset (left), and location and scale parameters of an MCMC sample with three clusters (right). The circles associated with each cluster correspond to 75% predictive probability regions for the associated component.

and thus lies on the Stiefel manifold $V_{3,2}$. Analysis of such data is important towards better understanding the origin and evolution of the NEOs population (Morbidelli et al. (2002)). The left subplot in Figure 1 shows these data, with each 2-frame represented as two orthonormal unit vectors. The first column (representing the latitude of perihelion) is the set of cyan lines arranged as two horizontal cones. The magenta lines (arranged as two vertical cones) form the second column, the longitude of perihelion.

We model this dataset as a DP mixture of matrix Langevin distributions. We set the DP concentration parameter $\alpha$ to 1 and, for the DP base measure, placed...
Figure 3: Posterior over the number of clusters for $\kappa$ restricted to $[1, \infty)$ (left) and $[3, \infty)$ (middle). The rightmost plot shows a traceplot of the number of clusters over MCMC iterations.

independent probability measures on the matrices $G$ and $\kappa$. For the former, we used a uniform prior (as in Section 3); however we found that an uninformative prior on $\kappa$ resulted in high posterior probability for a single diffuse cluster with no interesting structure. To discourage this, we sought to penalize small values of $\kappa_i$. One way to do this is to use a Gamma prior with a large shape parameter. Another is to use a hard constraint to bound the $\kappa_i$’s away from small values. We took the latter approach, placing independent exponential priors restricted to $[5, \infty)$ on the diagonal elements of $\kappa$. Our choice was motivated by the fact that for the one-dimensional von Mises distribution on the unit circle, $\kappa = 5$ gives a distribution of angles with standard deviation approximately equal to one.

The right plot in Figure 1 shows the adjacency matrix summarizing the posterior distribution over clusterings. An off-diagonal element $(i, j)$ gives the number of times observations $i$ and $j$ were assigned to the same cluster under the posterior. We see a highly coupled set of observations (from around observation 20 to 80 keeping the ordering of the downloaded dataset). This cluster corresponds to a tightly grouped set of observations, visible as a pair of bold lines in the left plot of Figure 1.

To investigate the underlying structure more carefully, we plot in Figure 2 the posterior distribution over the number of clusters. The figure shows this number is peaked at 4, extending up to 9. However, in most instances, most clusters have a small number of observations, with the posterior dominated by 2 or 3 large clusters. A typical two-cluster realization is fairly intuitive, with each cluster
corresponding to one of the two pairs of cones at right angles, and these clusters were identified quite consistently across all posterior samples. Occasionally, one or both of these might be further split into two smaller clusters, resulting in 3 or 4 clusters. A different example of a three cluster structure is shown in the right subfigure (this instance corresponded to the last MCMC sample of our chain that had three large clusters). In addition to the two aforementioned clusters, this assigns the bunched group of observations mentioned earlier (see the bunched cyan lines in figure 1) to their own cluster. In figure 3, we show the number of clusters when the diagonal elements of $\kappa$ are limited to $[1, \infty)$ (left) and $[3, \infty)$ (middle). In the former case, the posterior is dominated by a single large cluster, while the latter places more posterior mass on 2 to 3 clusters, the ideal solution. We also repeated the analysis of this section with a more general kernel that was not limited to having $H$ equal to the identity matrix. The results we obtained were largely the same, the only difference being a slight but insignificant decrease in the number of clusters observed under the posterior. This is to be expected. Finally, in the rightmost subplot, we plot a trace of the number of clusters under the posterior, demonstrating that our sampler mixes well.

Parametric analysis of this dataset typically requires identifying this cluster and treating it as a single observation (Sei et al. (2013)); by contrast, our non-parametric approach handles this much more naturally. Further, our Bayesian approach allows incorporating such an analysis as part of a larger hierarchical model. Figure 4 show the log predictive-probabilities of observations given this...
dataset, with the left subplot giving the distribution of the first component, and the right, the second. The peak of this distribution (the red spot to the right for the first plot, and the spot to the bottom left for the second), corresponds to the bunched set of observations mentioned earlier.

**Remark 2.** Chikuse (1998) proposed a kernel density estimator on the Stiefel manifold. The estimator is slightly technical and requires proper estimation of the smoothing parameter. Our model is fully generative and allows fully MCMC based inference for estimation, prediction, and uncertainty quantification. In addition, fitting our model via MCMC returns a clustering of the data, which is useful in many applications.

5. Proofs

**Proof of Theorem 4** In order to show weak consistency of the posterior distribution, it suffices to show that the prior distribution assigns positive mass to Kullback-Leibler neighborhoods of the true density $f_0$. This is a well-known result from Schwartz (1965). For density estimation on a manifold, Bhattacharya and Dunson (2012) derive some sufficient conditions for the KL support condition to hold on a general Riemannian manifold. We strive to check these conditions under our model.

By slightly abuse of notation, let $f$ be any continuous function on $V_{p,d}$ in this proof. We start by checking the KL condition. Bhattacharya and Dunson (2012) derive the following sufficient conditions for the KL support condition to hold on a general Riemannian manifold.

1. The kernel $g(X,G,\kappa)$ is continuous in all of its arguments.

2. The set $\{F_0\} \times D^p_\epsilon$ intersects the support of $\Pi_1 \times \pi_\kappa$ with $D^p_\epsilon$ as the interior of $D_\epsilon$, a compact neighborhood of some $\{\kappa_1, \ldots, \kappa_p\}$ in $\mathbb{R}^p$.

3. For any continuous function $f$ on $M$, there exists a compact neighborhood $D_\epsilon$ of $\{\kappa_1, \ldots, \kappa_p\}$, such that

$$\sup_{X \in V_{p,d}, \kappa \in D_\epsilon} \left\| f(X) - \int g(X,G,\kappa)f(G)\lambda(dG) \right\| \leq \epsilon.$$
Bayesian inference on the Stiefel manifold

For (1) one can write

$$g(X, G, \kappa) = C(\kappa) \exp \left( \sum_{i=1}^{p} \kappa_i G_{[i]} X_{[i]} \right).$$

Here $g$ is continuous with respect to $\kappa$ since the hypergeometric function $C(\kappa)$ is continuous and $\exp(\cdot)$ is clearly continuous with respect to $\kappa$ as the exponential term can be viewed as a linear combination of $\kappa_i$’s.

Now rewrite the density as

$$g(X, G, \kappa) = C(\kappa) \exp \left( \sum_{i=1}^{p} \kappa_i^2 - \rho(F, X)^2 \right),$$

where $\rho$ is the Frobenius distance between two matrices $F$ and $X$. Therefore $\exp(\cdot)$ is a continuous density of $X$ with respect to the Frobenius distance. $V_{p,d}$ can be embedded onto the Euclidean space $M(d, p)$ via the inclusion map, so one can equip $V_{p,d}$ with a metric space structure via the extrinsic distance $\rho$ in the Euclidean space. From the symmetry between $G$ and $X$, $g$ is also continuous with respect to $G$.

To prove (2), note that DP has weak support on all the measures whose support is contained by the base measure $P_0$ (See Theorem 3.2.4 in [Ghosh and Ramamoorthi (2003), pp. 104]. As $P_0$ and $\pi_\kappa$ have full support, (2) follows immediately.

Let $I(X) = f(X) - \int g(X, G, \kappa) f(G)dG$. For the last condition, we must show that there exists some compact subset in $\mathbb{R}^p$ with non-empty interior, $D_\epsilon$, such that

$$\sup_{X \in V_{p,d}, \kappa \in D_\epsilon} \|I(X)\| \leq \epsilon. \quad (5.1)$$

From symmetry of $g$ with respect to $G$ and $X$, one can write

$$I(X) = C(\kappa) \int (f(X) - f(G)) \exp(\cdot) \lambda(dG).$$

Let $\hat{G} = Q(d)^T G$, where $Q(d)$ is an orthogonal matrix with first $p$ columns as $X$. Then $G = Q(d) \hat{G}$. As the volume form is invariant under the group action
of the orthogonal matrices $O(d)$ on the left, one has $\lambda(dG) = \lambda(d\widehat{G})$. First,
\[
\rho^2(X, Q(d)\widehat{G}) = \text{Trace}\left( \left( X - Q(d)\widehat{G} \right) \left( X - Q(d)\widehat{G} \right)^T \right)
\]
\[
= 2p - 2 \text{Trace} \left( Q(d)^T X \widehat{G}^T \right)
\]
\[
= 2 \sum_{i=1}^p (1 - \widehat{g}_{ii}) ,
\]
with $\widehat{g}_{ii}$ being the diagonal elements of $\widehat{G}$. Let $(1 - \widehat{g}_{ii}) = \frac{1}{\kappa_i} s_{ii}$ for $i = 1, \ldots, p$, with $s_{ii} \in [0, 2\kappa_i]$. Then $\rho^2(X, Q(d)\widehat{G}) = 2 \sum_{i=1}^p \frac{1}{\kappa_i} s_{ii}$ for any given $\kappa_i$. Since the Stiefel manifold is compact and $f$ is continuous with respect to $\rho$, $f$ is uniformly continuously on $V_{p,d}$ with respect to the distance. Therefore, when $\kappa_i \to \infty$ for all $i = 1, \ldots, p$, one has for $s = \{s_{11}, \ldots, s_{pp}\}$,
\[
\sup_{X \in V_{p,d}} \left| \left( f(X) - f(Q(d)\widehat{G}) \right) \right| \to 0. \tag{5.2}
\]
Let $\hat{F}$ be the matrix whose $k$th column is $\kappa_k Q(d)\widehat{G}_{[k]}$. One has
\[
\sup_{X \in V_{p,d}} |I(X)| \leq \sup_{X \in V_{p,d}} C(\kappa) \int \left| \left( f(X) - f(Q(d)\widehat{G}) \right) \right| \text{etr}(\hat{F}^T X) \lambda(d\widehat{G})
\]
\[
\leq C(\kappa) \int \left\{ \sup_{X \in V_{p,d}} \left| \left( f(X) - f(Q(d)\widehat{G}) \right) \right| \right\} \exp \left( \sum_{i=1}^p \kappa_i \widehat{g}_{ii} \right) \lambda(d\widehat{G})
\]
\[
= C(\kappa) \exp \left( \sum_{i=1}^p \kappa_i \right) \int \left\{ \sup_{X \in V_{p,d}} \left| \left( f(X) - f(Q(d)\widehat{G}) \right) \right| \right\} \exp \left( - \sum_{i=1}^p s_{ii} \right) \lambda(d\widehat{G}). \tag{5.3}
\]
Let $\pi_1$ be the transformation given by $\pi_1(\widehat{g}_{ij}) = \widehat{g}_{ij}$ when $i \neq j$ and $\pi_1(\widehat{g}_{ii}) = s_{ii} = \kappa_i(1 - \widehat{g}_{ii})$. Let $\lambda(d\widehat{G}_s)$ be new volume measure after change of variables with respect to $\pi_1$. Let $J_1$ be the Jacobian of the map $\pi_1$. Rewrite $\lambda(d\widehat{G}) = \varphi(\widehat{G})d\widehat{g}_{11} \wedge d\widehat{g}_{12} \cdots \wedge d\widehat{g}_{dp}$ where $\varphi(\widehat{G})$ is some function of $\widehat{G}$. Similarly, let $\lambda(d\widehat{G}_s) = \overline{\varphi}(\widehat{G}_s)ds_{11} \wedge \cdots \wedge ds_{dp}$. One has
\[
\lambda(d\widehat{G}_s) = \overline{\varphi}(\widehat{G}_s) \prod_{i=1}^p \kappa_i d\widehat{g}_{11} \wedge d\widehat{g}_{12} \cdots \wedge d\widehat{g}_{dp}
\]
\[
= \overline{\varphi}(\widehat{G}_s) \prod_{i=1}^p \frac{1}{\varphi(\widehat{G})} \lambda(d\widehat{G}). \tag{5.4}
\]
Bayesian inference on the Stiefel manifold

The last term of (5.3) is

\[
C(\kappa) \exp \left( \sum_{i=1}^{p} \frac{\kappa_i}{\kappa_i} \right) \prod_{i=1}^{p} \frac{1}{\kappa_i} \int \left\{ \sup_{X \in V_{p,d}} | f(X) - f(Q(d)\hat{G}) | \right\} \times \\
\exp \left( \sum_{i=1}^{p} -s_{ii} \right) \frac{\varphi(\hat{G})}{\varphi(G_s)} \lambda(dG_s),
\]

(5.5)

with appropriate change of the range of integration. It is not hard to see that

\[
\int \exp \left( - \sum_{i=1}^{p} s_{ii} \right) \frac{\varphi(\hat{G})}{\varphi(G_s)} \lambda(dG_s) < \infty.
\]

(5.6)

We now proceed to show that even as \( \kappa_i \to \infty \),

\[
C(\kappa) \exp \left( \sum_{i=1}^{p} \frac{\kappa_i}{\kappa_i} \right) \prod_{i=1}^{p} \frac{1}{\kappa_i} < \infty.
\]

(5.7)

One has

\[
C(\kappa) \exp \left( \sum_{i=1}^{p} \frac{\kappa_i}{\kappa_i} \right) \prod_{i=1}^{p} \frac{1}{\kappa_i} = \prod_{i=1}^{p} \frac{1}{\kappa_i} 0_F^1 \left( \frac{1}{2}, \frac{1}{4} \text{Diag} \{ \kappa_1^2, \ldots, \kappa_p^2 \} \right) \prod_{i=1}^{p} \exp(\kappa_i).
\]

Write (see Butler and Wood (2003))

\[
0_F^1 \left( \frac{1}{2}, \frac{1}{4} \text{Diag} \{ \kappa_1^2, \ldots, \kappa_p^2 \} \right) = \int_{O_p} \text{etr} \left( \text{Diag} \{ \kappa_1, \ldots, \kappa_p \} T \right) dT
\]

(5.8)

with \( T \in O_p \) the group of all the \( p \) by \( p \) orthogonal matrices with \( dT \) given by \( \wedge_{i<j} t_{ij}^T dt_i \). When \( \kappa_i \geq 1 \) for \( i = 1, \ldots, p \), one looks at

\[
\int_{O_p} \text{etr} \left( \text{Diag} \{ \kappa_1, \ldots, \kappa_p \} T \right) \prod_{i=1}^{p} \exp(\kappa_i) dT = \int_{O_p} \exp \left( - \left( \sum_{i=1}^{p} \kappa_i(1 - t_{ii}) \right) \right) dT,
\]

where \( t_{ii} \) are the diagonal elements of \( T \). For \( \pi_2 \) the map such that \( \pi_2(t_{ij}) = t_{ij} \) for \( i \neq j \), and \( u_{ii} = \pi_2(t_{ii}) = \kappa_i(1 - t_{ii}) \), one has \( u_{ii} \in [0, 2\kappa_i] \). Let \( d\hat{T} \) be the volume form after change of variable. By the argument given in (5.4), we have

\[
\int_{O_p} \exp \left( - \left( \sum_{i=1}^{p} \kappa_i(1 - t_{ii}) \right) \right) dT = \prod_{i=1}^{p} 1/\kappa_i \int_{O_p} \exp \left( - \left( \sum_{i=1}^{p} u_{ii} \right) \right) \frac{1}{\det(J_2)} d\hat{T},
\]
where \( \det(J_2) \) corresponds to determinants of the Jacobian of maps \( \pi_2 \), which is essentially the same map as \( \pi_1 \) but with domain \( T \in O_p \). Note that 
\[
\int \exp \left( - \left( \sum_{i=1}^{p} u_i \right) \right) \frac{1}{\det(J_2)} d\hat{T}
\]
is bounded away from zero and infinity as \( \kappa_i \to \infty \). Therefore, we can conclude
\[
C(\kappa) \exp \left( \sum_{i=1}^{p} \kappa_i \right) \prod_{i=1}^{p} \frac{1}{\kappa_i} < \infty. \tag{5.9}
\]
Combining (5.2) and (5.9), by the Dominated Convergence Theorem, one has
\[
\sup_{X \in V_{p,d}} |I(X)| \to 0
\]
as \( \kappa_i \to \infty \) for all \( i = 1, \ldots, p \). Thus for all \( \epsilon > 0 \), there exists \( M_i \) large enough such that, when \( \kappa_i > M_i \), \( \sup_{X \in V_{p,d}} |I(X)| \leq \epsilon \). One can take \( D_\epsilon \) to be an \( \epsilon \)-neighborhood of \( \{\kappa_1, \ldots, \kappa_p\} \) with \( \kappa_i > \max\{M_i, i = 1, \ldots, p\} \).

**Proof of Theorem 3.** In order to establish strong consistency, it is not sufficient for the prior \( \Pi \) to assign positive mass to any Kullback-Leibler neighborhood of \( f_0 \); we need to construct high mass sieves with metric entropy \( N(\epsilon, \mathcal{F}) \) bounded by certain order, where \( N(\epsilon, \mathcal{F}) \) is defined as the logarithm of the minimum number of balls with Hellinger radius \( \epsilon \) to cover the space \( \mathcal{F} \). We refer to Barron et al. (1996) for some general strong consistency theorems. We first proceed to verify two conditions on the kernel \( g(X, G, \kappa) \).

(a) There exists positive constants \( k_0, a_1, \text{ and } A_1 \) such that for all \( k > k_0 \), \( G_1, G_2 \in V_{p,d} \), one has
\[
\sup_{X \in V_{p,d}, \kappa \in \phi^{-1}[0,k]} |g(X, G_1, \kappa) - g(X, G_2, \kappa)| \leq A_1 k^{a_1} \rho(G_1, G_2), \tag{5.10}
\]
where \( \phi : \mathbb{R}^p \to [0, \infty) \) is some continuous function of \( \kappa \).

(b) There exists positive constants \( a_2 \) and \( A_2 \) such that for all \( \kappa, \tilde{\kappa} \in \phi^{-1}[0,k], \ k \geq k_0 \),
\[
\sup_{X, G \in V_{p,d}} |g(X, G, \kappa) - g(X, G, \tilde{\kappa})| \leq A_2 k^{a_2} \rho_2(\kappa, \tilde{\kappa}), \tag{5.11}
\]
where \( \rho_2 \) is the Euclidean distance \( \| : \|_2 \) on \( \mathbb{R}^p \).
Bayesian inference on the Stiefel manifold

Let $G_1, G_2 \in V_{p,d}$ and $F_1$ and $F_2$ be such that their $i$th columns are given by $\kappa_i G_{1,i}$ and $\kappa_i G_{2,i}$, respectively. For $s, t \in [0, c]$ and $c > 0$, one has

$$\left| \exp \left( -\frac{s^2}{2} \right) - \exp \left( -\frac{t^2}{2} \right) \right| \leq |\eta \exp \left( -\frac{\eta^2}{2} \right) (s-t)| \leq c |s-t|,$$

where $\eta$ is some point between $s$ and $t$. Let $k_{\text{max}} = \max \{\kappa_1, \ldots, \kappa_p\}$. A little calculation shows that $\rho(F, X) \leq \sqrt{\sum_{i=1}^p (\kappa_i + 1)^2}$, so that

$$\sup_{X \in V_{p,d}, \kappa \in \phi^{-1}[0,k]} |g(X, G_1, \kappa) - g(X, G_2, \kappa)|$$

$$= \sup_{X \in V_{p,d}, \kappa \in \phi^{-1}[0,k]} \left| C(\kappa) \exp \left( \frac{p}{2} \right) \exp \left( \frac{\sum_{i=1}^p \kappa_i}{2} \right) \left( \exp \left( -\rho^2(F_1, X) \right) - \exp \left( -\rho^2(F_2, X) \right) \right) \right|$$

$$\leq \sup_{X \in V_{p,d}, \kappa \in \phi^{-1}[0,k]} \left| C(\kappa) \exp \left( \frac{p}{2} \right) \rho(F_1, F_2) \left( \sum_{i=1}^p (\kappa_i + 1)^2 \right) \right|$$

$$\leq 2 \exp \left( \frac{p}{2} \right) \sup_{X \in V_{p,d}, \kappa \in \phi^{-1}[0,k]} \left| C(\kappa) \exp \left( \frac{\sum_{i=1}^p \kappa_i}{2} \right) \sqrt{\sum_{i=1}^p \kappa_i^2 \rho(G_1, G_2)} \right|$$

$$\leq 2 \exp \left( \frac{p}{2} \right) \sup_{X \in V_{p,d}, \kappa \in \phi^{-1}[0,k]} \left| C \prod_{i=1}^p \kappa_i \sqrt{\sum_{i=1}^p \kappa_i^2 \rho(G_1, G_2)} \right|$$

where $C$ is some constant according to (5.9). Let $\phi(\kappa) = \sqrt{\sum_{i=1}^p (\kappa_i + 1)^2}$. If $\phi(\kappa) \leq k$, then $\sqrt{\sum_{i=1}^p \kappa_i^2} \leq \phi(\kappa) \leq k$ and $\kappa_i \leq k$ for each $i$. Thus $\prod_{i=1}^p \kappa_i \leq k^p$. Therefore,

$$\sup_{X \in V_{p,d}, \kappa \in \phi^{-1}[0,k]} |g(X, G_1, \kappa) - g(X, G_2, \kappa)| \leq C_1 k^p + 2 \rho(G_1, G_2),$$

with $C_1$ some constant. With $a_1 = p + 2$, Condition (a) holds.

Let $\kappa, \tilde{\kappa} \in \mathbb{R}^p$ be two vectors of the concentration parameters. By the Mean Value Theorem, one has, for some $t \in (0, 1)$,

$$g(X, G, \kappa) - g(X, G, \tilde{\kappa}) = (\nabla g(X, G, (1-t)\kappa + t\tilde{\kappa})) \cdot (\kappa - \tilde{\kappa}),$$

where $\nabla g(X, G, (1-t)\kappa + t\tilde{\kappa})$ is the gradient of $g(X, G, \kappa)$ with respect to $\kappa$ evaluated at $(1-t)\kappa + t\tilde{\kappa}$ and $\cdot$ denotes the inner product. By the Cauchy-Schwarz
inequality, one has

\[ |g(X, G, \kappa) - g(X, G, \tilde{\kappa})| \leq \| \bigtriangledown g(X, G, (1 - t)\kappa + t\tilde{\kappa}) \|_2 \| \kappa - \tilde{\kappa} \|_2. \]

Note that for \( i = 1, \ldots, p, \)

\[ \frac{\partial g}{\partial \kappa_i} = \exp \left( - \sum_{i=1}^{p} \kappa_i (1 - G_{[i]}^T X_{[i]}) \right) \left( C(\kappa) G_{[i]}^T X_{[i]} \exp \left( \sum_{i=1}^{p} \kappa_i \right) + \frac{\partial C(\kappa)}{\partial \kappa_i} \exp \left( \sum_{i=1}^{p} \kappa_i \right) \right) \]

\[ - C^2(\kappa) \frac{\partial_0 F_1 \left( \frac{1}{2} d, \frac{1}{4} \text{Diag} \{ \kappa_1^2, \ldots, \kappa_p^2 \} \right)}{\partial \kappa_i} \exp \left( \sum_{i=1}^{p} \kappa_i \right). \]

By applying the general Leibniz rule for differentiation under an integral sign, one has

\[ \frac{\partial_0 F_1 \left( \frac{1}{2} d, \frac{1}{4} \text{Diag} \{ \kappa_1^2, \ldots, \kappa_p^2 \} \right)}{\partial \kappa_i} \exp \left( \sum_{i=1}^{p} \kappa_i \right) \leq \frac{1}{C(\kappa)}. \]

Then one has

\[ \left| \frac{\partial g(X, G, \kappa)}{\partial \kappa_i} \right| \leq C(\kappa) \exp \left( \sum_{i=1}^{p} \kappa_i \right) + C^2(\kappa) \frac{\partial_0 F_1 \left( \frac{1}{2} d, \frac{1}{4} \text{Diag} \{ \kappa_1^2, \ldots, \kappa_p^2 \} \right)}{\partial \kappa_i} \exp \left( \sum_{i=1}^{p} \kappa_i \right) \]

\[ \leq 2C(\kappa) \exp \left( \sum_{i=1}^{p} \kappa_i \right) \leq C_2 \prod_{i=1}^{p} \kappa_i, \]

for some constant \( C_2 \) by \([5,9]\). Therefore, \( \| \bigtriangledown g(X, G, (1 - t)\kappa + t\tilde{\kappa}) \|_2 \leq C_2 q^p, \)

and one has \( |g(X, G, \kappa) - g(X, G, \tilde{\kappa})| \leq C_2 q^p \| \kappa - \tilde{\kappa} \|_2. \) Letting \( a_2 = p, \) Condition (b) is verified.

We proceed to verify two entropy conditions:

(c) For any \( k \geq k_0, \) the subset \( \phi^{-1}[0, k] \) is compact and its \( \epsilon \)-covering number is bounded by \( (k\epsilon^{-1})^{b_2} \) for some constant \( b_2 \) independent of \( \kappa \) and \( \epsilon. \)
(d) The $\epsilon$ covering number of the manifold $V_{p,d}$ is bounded by $A_3\epsilon^{-a_3}$ for any $\epsilon > 0$.

It is easy to verify Condition (c) as $\phi^{-1}([0, k]) = \{\kappa, \sum_{i=1}^{p}(\kappa_i + 1)^2 \leq k^2\}$, which is a subset of a shifted Euclidean ball in $\mathbb{R}^p$ with radius $k$. With a direct argument using packing numbers [Pollard, 1990, Section 4], one can obtain a bound for the entropy of $\phi^{-1}[0, k]$ given by $\frac{3k_p}{e_p}$. Thus Condition (c) holds with $b_2 = p$.

Let $N(\epsilon)$ be the entropy of $V_{p,d}$ and $N_E(\epsilon)$ be the entropy of $V_{p,d}$ viewed as a subset of $\mathbb{R}^{pd}$ (points covering $V_{p,d}$ do not necessarily lie on $V_{p,d}$ for the latter case). One can show that $N(2\epsilon) \leq N_E(\epsilon)$. Here $V_{p,d} \subset [-1, 1]^{pd}$, which is a subset of a Euclidean ball of radius $\sqrt{dp}$ centered at zero, the $\epsilon$ number of which is bounded by $\left(\frac{3\sqrt{dp}}{\epsilon}\right)^{dp}$. Therefore, Condition (d) holds with $a_3 = dp$. Then by Corollary 1 in Bhattacharya and Dunson (2012), strong consistency follows.
Acknowledgements The authors sincerely thank an associate editor and the referees for their valuable input, which has led to improvements of the paper. This work was supported by grant R01ES017240 from the National Institute of Environmental Health Sciences (NIEHS) of the National Institute of Health (NIH) and a National Science Foundation (NSF) grant IIS1546331.

Bibliography


Department of Statistics and Data Sciences, The University of Texas at Austin, Austin, Texas, 78712, USA.
E-mail: lizhen.lin@austin.utexas.edu

Department of Statistics, Purdue University, West Lafayette, Indiana, 47907, USA.
E-mail: varao@purdue.edu

Department of Statistical Science, Duke University, Durham, North Carolina, 27707, USA.
E-mail: dunson@duke.edu