Random orthogonal matrices and the Cayley transform

MICHAEL JAUCH[1], PETER D. HOFF[2,*] and DAVID B. DUNSON[2,†]

[1]*Center for Applied Mathematics, Cornell University, Ithaca, NY 14850, USA.*
E-mail: michael.jauch@cornell.edu
[2]*Department of Statistical Science, Duke University, Durham, NC 27708, USA.*
*E-mail: *peter.hoff@duke.edu; †dunson@duke.edu*

Random orthogonal matrices play an important role in probability and statistics, arising in multivariate analysis, directional statistics, and models of physical systems, among other areas. Calculations involving random orthogonal matrices are complicated by their constrained support. Accordingly, we parametrize the Stiefel and Grassmann manifolds, represented as subsets of orthogonal matrices, in terms of Euclidean parameters using the Cayley transform. We derive the necessary Jacobian terms for change of variables formulas. Given a density defined on the Stiefel or Grassmann manifold, these allow us to specify the corresponding density for the Euclidean parameters, and vice versa. As an application, we present a Markov chain Monte Carlo approach to simulating from distributions on the Stiefel and Grassmann manifolds. Finally, we establish that the Euclidean parameters corresponding to a uniform orthogonal matrix can be approximated asymptotically by independent normals. This result contributes to the growing literature on normal approximations to the entries of random orthogonal matrices or transformations thereof.

Keywords: Gaussian approximation; Grassmann manifold; Jacobian; Markov chain Monte Carlo; Stiefel manifold

1. Introduction

Random orthogonal matrices play an important role in probability and statistics. They arise, for example, in multivariate analysis, directional statistics, and models of physical systems. The set of $p \times k$ orthogonal matrices $\mathcal{V}(k, p) = \{ \boldsymbol{Q} \in \mathbb{R}^{p \times k} \mid \boldsymbol{Q}^T \boldsymbol{Q} = \boldsymbol{I}_k \}$, known as the Stiefel manifold, is a $d_{\mathcal{V}} = pk - k(k+1)/2$ dimensional submanifold of $\mathbb{R}^{pk}$. There are two notable special cases: $\mathcal{V}(1, p)$ is equivalent to the unit hypersphere, while $\mathcal{V}(p, p)$ is equivalent to the orthogonal group $\mathcal{O}(p)$. Closely related to the Stiefel manifold is the Grassmann manifold $\mathcal{G}(k, p)$, the set of k-dimensional linear subspaces of $\mathbb{R}^p$. The Grassmann manifold has dimension $d_{\mathcal{G}} = (p - k)k$. Points in the Grassmann manifold are often thought of as equivalence classes of $\mathcal{V}(k, p)$, where two orthogonal matrices belong to the same class if they share the same column space or, equivalently, if one matrix can be obtained from the other through right multiplication by an element of $\mathcal{O}(k)$ [22]. In Section 3.2, we elaborate and expand upon the contributions of Shepard et al. [50] to provide another representation of the Grassmann manifold $\mathcal{G}(k, p)$ as the subset $\mathcal{V}^+(k, p)$ of $p \times k$ orthogonal matrices having a symmetric positive definite (SPD) top block. In this article, we focus on orthogonal matrices having fewer columns than rows.

Both the Stiefel and Grassmann manifolds can be equipped with a uniform probability measure, also know as an invariant or Haar measure. The uniform distribution $P_{\mathcal{V}(k,p)}$ on $\mathcal{V}(k, p)$

is characterized by its invariance to left and right multiplication by orthogonal matrices: If $Q \sim P_{\mathcal{V}(k,p)}$, then $UQV \stackrel{d}{=} Q$ for all $U \in \mathcal{O}(p)$ and $V \in \mathcal{O}(k)$. Letting $l : \mathcal{V}(k,p) \to \mathcal{G}(k,p)$ be the function taking an orthogonal matrix to its column space, the uniform distribution $P_{\mathcal{G}(k,p)}$ on $\mathcal{G}(k,p)$ is the pushforward measure of $P_{\mathcal{V}(k,p)}$ under l. In other words, the measure of $A \subseteq \mathcal{G}(k,p)$ is $P_{\mathcal{G}(k,p)}[A] = P_{\mathcal{V}(k,p)}[l^{-1}(A)]$. The uniform distributions on these manifolds have a long history in probability theory, as we discuss in Section 6, and in statistics, where they appear in foundational work on multivariate analysis [28].

Non-uniform distributions on the Stiefel and Grassmann manifolds play an important role in modern statistical applications. They are used to model directions, axes, planes, rotations, and other data in the field of directional statistics [39]. Also, statistical models having the Stiefel or Grassmann manifold as their parameter space are increasingly common [10,24,25]. In particular, Bayesian analyses of multivariate data often involve prior and posterior distributions on $\mathcal{V}(k,p)$ or $\mathcal{G}(k,p)$. Bayesian inference typically requires simulating from these posterior distributions, motivating the development of new Markov chain Monte Carlo (MCMC) methodology [5,25,29,48].

The challenge of performing calculations with random orthogonal matrices has motivated researchers to parametrize sets of square orthogonal matrices in terms of Euclidean parameters. We provide a few examples. In what Diaconis and Forester [14] identify as the earliest substantial mathematical contribution to modern random matrix theory, Hurwitz [27] parametrizes the special orthogonal and unitary groups using Euler angles and computes the volumes of their invariant measures. An implication of these computations is that the Euler angles of a uniformly distributed matrix follow independent beta distributions. Over a century later, connections to analytic number theory [34] have motivated extensions to more general measures and groups [3,4]. Anderson et al. [1] discuss the potential of various parametrizations in the simulation of uniformly distributed square orthogonal matrices. Other authors have made use of parametrizations of square orthogonal or rotation matrices in statistical applications [12,35].

In contrast, the topic of parametrizing random orthogonal matrices having fewer columns than rows has received little attention. The recent work of Shepard et al. [50] extends four existing approaches to parametrizing square orthogonal matrices to the case when $k < p$ and to the scenario in which only the column space of the orthogonal matrix is of interest. Naturally, this latter scenario is closely related to the Grassmann manifold. The tools needed to use these parametrizations in a probabilistic setting are still largely missing.

In this article, we lay foundations for application of the Cayley parametrization of the Stiefel and Grassmann manifolds in a probabilistic setting. There are three main contributions. First, we elaborate and expand upon the work of Shepard et al. [50] to show that the Grassmann manifold $\mathcal{G}(k,p)$ can be represented by the subset $\mathcal{V}^+(k,p)$ of orthogonal matrices having an SPD top block and that this subset can be parametrized in terms of Euclidean elements using the Cayley transform. Next, we derive the necessary Jacobian terms for change of variables formulas. Given a density defined on $\mathcal{V}(k,p)$ or $\mathcal{V}^+(k,p)$, these allow us to specify the corresponding density for the Euclidean parameters, and vice versa. As an application, we present a Markov chain Monte Carlo approach to simulating from distributions on the Stiefel and Grassmann manifolds. Finally, we establish that the Euclidean parameters corresponding to a uniform orthogonal matrix can be approximated asymptotically by independent normals. This result contributes to

the growing literature on normal approximations to the entries of random orthogonal matrices or transformations thereof.

Code to replicate the figures in this article and to simulate from distributions on $\mathcal{V}(k, p)$ is available at https://github.com/michaeljauch/cayley.

2. Probability distributions on submanifolds of $\mathbb{R}^n$

In this section, we introduce tools for defining and manipulating probability distributions on an m-dimensional submanifold $\mathcal{M}$ of $\mathbb{R}^n$. In particular, we discuss the reference measure with respect to which we define densities on $\mathcal{M}$, we make precise what it means for us to parametrize $\mathcal{M}$ in terms of Euclidean parameters, and we state a change of variables formula applicable in this setting. Our formulation of these ideas follows that of Diaconis et al. [16] somewhat closely. This general discussion will form the basis for our handling of the specific cases in which $\mathcal{M}$ is $\mathcal{V}(k, p)$ or $\mathcal{V}^+(k, p)$.

In order to specify probability distributions on $\mathcal{M}$ in terms of density functions, we need a reference measure on that space, analogous to Lebesgue measure L^m on $\mathbb{R}^m$. We take the Hausdorff measure as our reference measure. A referee and Diaconis et al. [17] point out that the Hausdorff measure is a more natural fit for rough sets (see, e.g., the Cantor set or fractional Brownian motion) or manifolds with kinks (see, e.g., the barbell of Grayson [23]) than for smooth Riemannian manifolds such as $\mathcal{V}(k, p)$ or $\mathcal{V}^+(k, p)$. Still, the Hausdorff measure is intuitive, coincides with alternative measures on submanifolds of $\mathbb{R}^n$, allows for a simple statement of the required results, and appears in related work of Diaconis et al. [16] and Byrne and Girolami [5].

Heuristically, the m-dimensional Hausdorff measure of $A \subset \mathbb{R}^n$ is the m-dimensional area of A. More formally, the m-dimensional Hausdorff measure $H^m(A)$ of A is defined

$$H^m(A) = \lim_{\delta \to 0} \inf_{\substack{A \subset \bigcup_i S_i \\ \mathrm{diam}(S_i) < \delta}} \sum_i \alpha_m \left(\frac{\mathrm{diam}(S_i)}{2} \right)^m,$$

where the infimum is taken over countable coverings $\{S_i\}_{i \in \mathbb{N}}$ of A with

$$\mathrm{diam}(S_i) = \sup \big\{ |x - y| : x, y \in S_i \big\}$$

and $\alpha_m = \Gamma(\frac{1}{2})^m / \Gamma(\frac{m}{2} + 1)$, the volume of the unit ball in $\mathbb{R}^m$. The $d_\mathcal{V}$-dimensional Hausdorff measure on $\mathcal{V}(k, p)$ coincides with $P_{\mathcal{V}(k, p)}$ up to a multiplicative constant.

Let $g : \mathcal{M} \to \mathbb{R}$ be proportional to a density with respect to the m-dimensional Hausdorff measure on $\mathcal{M}$. Furthermore, suppose $\mathcal{M}$ can be parametrized by a function f from an open domain $\mathcal{D} \in \mathbb{R}^m$ to $\mathcal{I} = f(\mathcal{D}) \subseteq \mathcal{M}$ satisfying the following conditions:

1. Almost all of $\mathcal{M}$ is contained in the image $\mathcal{I}$ of $\mathcal{D}$ under f so that $H^m(\mathcal{M} \setminus \mathcal{I}) = 0$.
2. The function f is injective on $\mathcal{D}$.
3. The function f is continuously differentiable on $\mathcal{D}$ with the derivative matrix $Df(\boldsymbol{\phi})$ at $\boldsymbol{\phi} \in \mathcal{D}$.

In this setting, we obtain a simple change of variables formula. This formula will be applied to $\mathcal{V}(k, p)$ and $\mathcal{V}^+(k, p)$ in the derivation of Theorem 4.1. Define the m-dimensional Jacobian of f at $\boldsymbol{\phi}$ as $J_m f(\boldsymbol{\phi}) = |Df(\boldsymbol{\phi})^T Df(\boldsymbol{\phi})|^{1/2}$. Like the familiar Jacobian determinant, this term acts as a scaling factor in a change of variables formula. However, it is defined even when the derivative matrix is not square. For more details, see the discussion in [16]. The change of variables formula is given in the following theorem, which is essentially a restatement of the main result of Traynor [54]:

Theorem 2.1. *For all Borel subsets $A \subset \mathcal{D}$*

$$\int_A J_m f(\boldsymbol{\phi}) L^m(d\boldsymbol{\phi}) = H^m\big[f(A)\big]$$

and hence

$$\int_A g\big[f(\boldsymbol{\phi})\big] J_m f(\boldsymbol{\phi}) L^m(d\boldsymbol{\phi}) = \int_{f(A)} g(\boldsymbol{y}) H^m(d\boldsymbol{y}).$$

Naturally, the change of variables formula has an interpretation in terms of random variables. Let $\boldsymbol{y}$ be a random element of $\mathcal{M}$ whose distribution has a density proportional to g. Then $\boldsymbol{y} \overset{d}{=} f(\boldsymbol{\phi})$ when the distribution of $\boldsymbol{\phi} \in \mathcal{D}$ has a density proportional to $g[f(\boldsymbol{\phi})] J_m f(\boldsymbol{\phi})$.

3. The Cayley parametrizations

The Cayley transform, as introduced in [7], is a map from skew-symmetric matrices to special orthogonal matrices. Given X in $\text{Skew}(p) = \{X \in \mathbb{R}^{p \times p} | X = -X^T\}$, the (original) Cayley transform of X is the special orthogonal matrix

$$C_{\text{orig.}}(X) = (I_p + X)(I_p - X)^{-1}.$$

We work instead with a modified version of the Cayley transform described in [50]. In this version, the Cayley transform of X is the $p \times k$ orthogonal matrix

$$C(X) = (I_p + X)(I_p - X)^{-1} I_{p \times k}$$

where $I_{p \times k}$ denotes the $p \times k$ matrix having the identity matrix as its top block and the remaining entries zero. The matrix $I_p - X$ is invertible for any $X \in \text{Skew}(p)$, so the Cayley transform is defined everywhere.

In this section, we parametrize (in the sense of Section 2) the sets $\mathcal{V}(k, p)$ and $\mathcal{V}^+(k, p)$ using the Cayley transform C. We are able to parametrize these distinct sets by restricting the domain of C to distinct subsets of $\text{Skew}(p)$. The third condition of the previous section requires that C be continuously differentiable on its domain. We verify this condition by computing the derivative matrices of C, clearing a path for the statement of change of variables formulas in Section 4. We also state and discuss an important proposition which justifies our claim that the Grassmann manifold $\mathcal{G}(k, p)$ can be represented by the set $\mathcal{V}^+(k, p)$.

3.1. Cayley parametrization of the Stiefel manifold

To parametrize the Stiefel manifold $\mathcal{V}(k, p)$, the domain of C is restricted to the subset

$$\mathcal{D}_{\mathcal{V}} = \left\{ X = \begin{bmatrix} B & -A^T \\ A & 0_{p-k} \end{bmatrix} \middle| \begin{array}{l} B \in \text{Skew}(k), \\ A \in \mathbb{R}^{p-k \times k} \end{array} \right\} \subset \text{Skew}(p).$$

Let $X \in \mathcal{D}_{\mathcal{V}}$ and set $Q = C(X)$. Partition $Q = [Q_1^T \ Q_2^T]^T$ so that Q_1 is square. We can write the blocks of Q in terms of the blocks of X as

$$Q_1 = (I_k - A^T A + B)(I_k + A^T A - B)^{-1},$$

$$Q_2 = 2A(I_k + A^T A - B)^{-1}.$$

The matrix $I_k + A^T A - B$ is the sum of a symmetric positive definite matrix and a skew-symmetric matrix and therefore nonsingular (see the appendix for a proof). This observation guarantees (again) that the Cayley transform is defined for all $X \in \mathcal{D}_{\mathcal{V}}$. We can recover the matrices A and B from Q:

$$F = (I_k - Q_1)(I_k + Q_1)^{-1}, \tag{3.1}$$

$$B = \frac{1}{2}(F^T - F), \tag{3.2}$$

$$A = \frac{1}{2}Q_2(I_k + F). \tag{3.3}$$

We are now in a position to verify the first two conditions of Section 2. The image of $\mathcal{D}_V$ under C is the set

$$\mathcal{I}_{\mathcal{V}} = \left\{ Q = [Q_1^T \quad Q_2^T]^T \in \mathcal{V}(k, p) \mid I_k + Q_1 \text{ is nonsingular} \right\},$$

which has measure one with respect to $P_{\mathcal{V}(k,p)}$. The injectivity of C on $\mathcal{D}_{\mathcal{V}}$ follows from the existence of the inverse mapping described in equations (3.1)–(3.3). All that remains to be verified is the third condition, that C is continuously differentiable on its domain.

As the first step in computing the derivative matrix of C, we define a $d_{\mathcal{V}}$-dimensional vector $\boldsymbol{\varphi}$ containing each of the independent entries of X. Let b be the $k(k-1)/2$-dimensional vector of independent entries of B obtained by eliminating diagonal and supradiagonal elements from the vectorization vec B. The vector $\boldsymbol{\varphi} = (b^T, \text{vec } A^T)^T$ then contains each of the independent entries of X. Let $X_{\boldsymbol{\varphi}} \in \mathcal{D}_{\mathcal{V}}$ be the matrix having $\boldsymbol{\varphi}$ as its corresponding vector of independent entries.

The Cayley transform can now be thought of as a function of $\boldsymbol{\varphi}$:

$$C(\boldsymbol{\varphi}) = (I_p + X_{\boldsymbol{\varphi}})(I_p - X_{\boldsymbol{\varphi}})^{-1} I_{p \times k}.$$

As a function of $\boldsymbol{\varphi}$, the Cayley transform is a bijection between the set $\mathcal{D}_{\mathcal{V}}^{\boldsymbol{\varphi}} = \{\boldsymbol{\varphi} \in \mathbb{R}^{d_{\mathcal{V}}} : X_{\boldsymbol{\varphi}} \in \mathcal{D}_{\mathcal{V}}\} = \mathbb{R}^{d_{\mathcal{V}}}$ and $\mathcal{I}_{\mathcal{V}}$. The inverse Cayley transform is defined in the obvious way as the

map $C^{-1} : \boldsymbol{Q} \mapsto (\boldsymbol{b}^T, \text{vec} \, \boldsymbol{A}^T)^T$ where $\boldsymbol{B}$ and $\boldsymbol{A}$ are computed from $\boldsymbol{Q}$ according to equations (3.1)–(3.3) and $\boldsymbol{b}$ contains the independent entries of $\boldsymbol{B}$ as before.

The next lemma provides an explicit linear map $\boldsymbol{\Gamma}_\mathcal{V}$ sending $\boldsymbol{\varphi}$ to vec $X_{\boldsymbol{\varphi}}$, greatly simplifying our calculation of the derivative matrix $DC(\boldsymbol{\varphi})$. The entries of $\boldsymbol{\Gamma}_\mathcal{V}$ belong to the set $\{-1, 0, 1\}$. The construction of $\boldsymbol{\Gamma}_\mathcal{V}$ involves the commutation matrix $\boldsymbol{K}_{p,p}$ and the matrix $\widetilde{\boldsymbol{D}}_k$ satisfying $\widetilde{\boldsymbol{D}}_k \boldsymbol{b} = \text{vec} \, \boldsymbol{B}$, both of which are discussed in [36] and defined explicitly in Appendix B. Set $\boldsymbol{\Theta}_1 = [\boldsymbol{I}_k \; \boldsymbol{0}_{k \times p-k}]$ and $\boldsymbol{\Theta}_2 = [\boldsymbol{0}_{p-k \times k} \; \boldsymbol{I}_{p-k}]$.

Lemma 3.1. *The equation* vec $X_{\boldsymbol{\varphi}} = \boldsymbol{\Gamma}_\mathcal{V} \boldsymbol{\varphi}$ *is satisfied by the matrix*

$$\boldsymbol{\Gamma}_\mathcal{V} = \big[(\boldsymbol{\Theta}_1^T \otimes \boldsymbol{\Theta}_1^T) \widetilde{\boldsymbol{D}}_k \quad (\boldsymbol{I}_{p^2} - \boldsymbol{K}_{p,p})(\boldsymbol{\Theta}_1^T \otimes \boldsymbol{\Theta}_2^T) \big].$$

With these pieces in place, we can now identify the derivative matrix $C(\boldsymbol{\varphi})$.

Proposition 3.1. *The Cayley transform is continuously differentiable on* $\mathcal{D}_\mathcal{V}^{\boldsymbol{\varphi}} = \mathbb{R}^{d_\mathcal{V}}$ *with derivative matrix*

$$DC(\boldsymbol{\varphi}) = 2 \big[\boldsymbol{I}_{p \times k}^T (\boldsymbol{I}_p - X_{\boldsymbol{\varphi}})^{-T} \otimes (\boldsymbol{I}_p - X_{\boldsymbol{\varphi}})^{-1} \big] \boldsymbol{\Gamma}_\mathcal{V}.$$

The form of the derivative matrix reflects the composite structure of the Cayley transform as a function of $\boldsymbol{\varphi}$. The Kronecker product term arises from differentiating C with respect to $X_{\boldsymbol{\varphi}}$, while the matrix $\boldsymbol{\Gamma}_\mathcal{V}$ arises from differentiating $X_{\boldsymbol{\varphi}}$ with respect to $\boldsymbol{\varphi}$.

3.2. The Cayley parametrization of the Grassmann manifold

While the definition of the Grassmann manifold as a collection of subspaces is fundamental, we often need a more concrete representation. One simple idea is to represent a subspace $S \in \mathcal{G}(k, p)$ by $\boldsymbol{Q} \in \mathcal{V}(k, p)$ having S as its column space. However, the choice of $\boldsymbol{Q}$ is far from unique. Alternatively, the subspace S can be represented uniquely by the orthogonal projection matrix onto S, as in [9]. We propose instead to represent $\mathcal{G}(k, p)$ by the subset of orthogonal matrices

$$\mathcal{V}^+(k, p) = \big\{ \boldsymbol{Q} = \big[\boldsymbol{Q}_1^T \quad \boldsymbol{Q}_2^T \big]^T \in \mathcal{V}(k, p) \mid \boldsymbol{Q}_1 \succ \boldsymbol{0} \big\}.$$

As the next proposition makes precise, almost every element of $\mathcal{G}(k, p)$ can be represented uniquely by an element of $\mathcal{V}^+(k, p)$. Recall that we defined $l : \mathcal{V}(k, p) \to \mathcal{G}(k, p)$ as the map which sends each element of $\mathcal{V}(k, p)$ to its column space.

Proposition 3.2. *The map l is injective on $\mathcal{V}^+(k, p)$ and the image of $\mathcal{V}^+(k, p)$ under l has measure one with respect to the uniform probability measure $P_{\mathcal{G}(k,p)}$ on $\mathcal{G}(k, p)$.*

In turn, the set $\mathcal{V}^+(k, p)$ is amenable to parametrization by the Cayley transform. In this case, the domain of the Cayley transform C is restricted to the subset

$$\mathcal{D}_\mathcal{G} = \left\{ X = \begin{bmatrix} \boldsymbol{0} & -\boldsymbol{A}^T \\ \boldsymbol{A} & \boldsymbol{0}_{p-k} \end{bmatrix} \,\middle|\, \begin{matrix} \boldsymbol{A} \in \mathbb{R}^{p-k \times k}, \\ \boldsymbol{0} \preceq \boldsymbol{A}^\top \boldsymbol{A} \prec \boldsymbol{I}_k \end{matrix} \right\} \subset \text{Skew}(p).$$

Let $X \in \mathcal{D}_\mathcal{G}$ and set $Q = C(X)$. Again, partition $Q = [Q_1^T \ Q_2^T]^T$ so that Q_1 is square. We can write the blocks of Q in terms of A as

$$Q_1 = (I_k - A^T A)(I_k + A^T A)^{-1},$$

$$Q_2 = 2A(I_k + A^T A)^{-1}.$$

We can recover the matrix A from Q:

$$F = (I_k - Q_1)(I_k + Q_1)^{-1}, \tag{3.4}$$

$$A = \frac{1}{2}Q_2(I_k + F). \tag{3.5}$$

We turn to verifying the conditions of Section 2. The first two conditions are satisfied as a consequence of the following proposition:

Proposition 3.3. *The Cayley transform $C : \mathcal{D}_\mathcal{G} \to \mathcal{G}(k, p)$ is one-to-one.*

Only the third condition, that C is continuously differentiable on $\mathcal{D}_G$, remains.

The process of computing the derivative matrix is the same as before. We define a $d_\mathcal{G}$-dimensional vector $\psi = \text{vec } A$ containing each of the independent entries of X. Let $X_\psi \in \mathcal{D}_\mathcal{G}$ be the matrix having ψ as its corresponding vector of independent entries. The Cayley transform can now be thought of as a function of ψ:

$$C(\psi) = (I_p + X_\psi)(I_p - X_\psi)^{-1} I_{p \times k}.$$

As a function of ψ, the Cayley transform is a bijection between the set $\mathcal{D}_\mathcal{G}^\psi = \{\psi \in \mathbb{R}^{d_\mathcal{G}} : X_\psi \in \mathcal{D}_\mathcal{G}\}$ and $\mathcal{V}^+(k, p)$. The next lemma provides an explicit linear map sending ψ to $\text{vec } X_\psi$ in the form of a $\{-1, 0, 1\}$ matrix $\Gamma_\mathcal{G}$:

Lemma 3.2. *The equation $\text{vec } X_\psi = \Gamma_\mathcal{G}\psi$ is satisfied by the matrix*

$$\Gamma_\mathcal{G} = (I_{p^2} - K_{p,p})(\Theta_1^T \otimes \Theta_2^T).$$

In the next proposition, we identify the derivative matrix $DC(\phi)$.

Proposition 3.4. *The Cayley transform is continuously differentiable on $\mathcal{D}_\mathcal{G}^\psi$ with derivative matrix*

$$DC(\psi) = 2[I_{p \times k}^T (I_p - X_\psi)^{-T} \otimes (I_p - X_\psi)^{-1}]\Gamma_\mathcal{G}.$$

4. Change of variables formulas

We now state change of variables formulas for the Cayley parametrizations of $\mathcal{V}(k, p)$ and $\mathcal{V}^+(k, p)$. Given the results of the previous section, the formulas follow directly from Theorem 2.1. The $d_\mathcal{V}$-dimensional Jacobian $J_{d_\mathcal{V}} C(\boldsymbol{\varphi})$ of the Cayley transform C at $\boldsymbol{\varphi}$ is equal to

$$J_{d_\mathcal{V}} C(\boldsymbol{\varphi}) = \left| DC(\boldsymbol{\varphi})^T DC(\boldsymbol{\varphi}) \right|^{1/2}$$
$$= \left| 2^2 \boldsymbol{\Gamma}_\mathcal{V}^T (\boldsymbol{G}_\mathcal{V} \otimes \boldsymbol{H}_\mathcal{V}) \boldsymbol{\Gamma}_\mathcal{V} \right|^{1/2},$$

where

$$\boldsymbol{G}_\mathcal{V} = (\boldsymbol{I}_p - \boldsymbol{X}_{\boldsymbol{\varphi}})^{-1} \boldsymbol{I}_{p \times k} \boldsymbol{I}_{p \times k}^T (\boldsymbol{I}_p - \boldsymbol{X}_{\boldsymbol{\varphi}})^{-T},$$
$$\boldsymbol{H}_\mathcal{V} = (\boldsymbol{I}_p - \boldsymbol{X}_{\boldsymbol{\varphi}})^{-T} (\boldsymbol{I}_p - \boldsymbol{X}_{\boldsymbol{\varphi}})^{-1}$$

and $\boldsymbol{\Gamma}_\mathcal{V}$ is defined as in Section 3.1. The equations above hold for the $d_\mathcal{G}$-dimensional Jacobian $J_{d_\mathcal{G}} C(\boldsymbol{\phi})$ if we replace the symbols $\mathcal{V}$ and $\boldsymbol{\varphi}$ with the symbols $\mathcal{G}$ and $\boldsymbol{\psi}$, respectively. In the supplementary material, we describe how to compute the Jacobian terms, taking advantage of their block structure. Let $g : \mathcal{V}(k, p) \to \mathbb{R}$ be proportional to a density with respect to the $d_\mathcal{V}$-dimensional Hausdorff measure on $\mathcal{V}(k, p)$.

Theorem 4.1 (Change of variables formulas). *For all Borel sets $A \subset \mathcal{D}_\mathcal{V}^{\varphi}$*

$$\int_A J_{d_\mathcal{V}} C(\boldsymbol{\varphi}) L^{d_\mathcal{V}}(d\boldsymbol{\varphi}) = H^{d_\mathcal{V}}\left[C(A) \right] \tag{4.1}$$

and hence

$$\int_A g\left[C(\boldsymbol{\varphi}) \right] J_{d_\mathcal{V}} C(\boldsymbol{\varphi}) L^{d_\mathcal{V}}(d\boldsymbol{\varphi}) = \int_{C(A)} g(\boldsymbol{Q}) H^{d_\mathcal{V}}(d\boldsymbol{Q}). \tag{4.2}$$

If instead we have $g : \mathcal{V}^+(k, p) \to \mathbb{R}$ proportional to a density with respect to the $d_\mathcal{G}$-dimensional Hausdorff measure on $\mathcal{V}^+(k, p)$, the statement is true when we replace the symbols $\mathcal{V}$ and $\boldsymbol{\varphi}$ with the symbols $\mathcal{G}$ and $\boldsymbol{\psi}$, respectively.

Similarly to Theorem 2.1, Theorem 4.1 has an interpretation in terms of random variables. Let the distribution of $\boldsymbol{Q} \in \mathcal{V}(k, p)$ have a density proportional to g. Then $\boldsymbol{Q} \overset{d}{=} C(\boldsymbol{\varphi})$ when the distribution of $\boldsymbol{\varphi} \in \mathcal{D}_\mathcal{V}^{\varphi} = \mathbb{R}^{d_\mathcal{V}}$ has a density proportional to $g[C(\boldsymbol{\varphi})] J_{d_\mathcal{V}} C(\boldsymbol{\varphi})$. In particular, let $g \propto 1$ so that $\boldsymbol{Q} \sim P_{\mathcal{V}(k,p)}$. Then $\boldsymbol{Q} \overset{d}{=} C(\boldsymbol{\varphi})$ when the distribution of $\boldsymbol{\varphi}$ has a density proportional to $J_{d_\mathcal{V}} C(\boldsymbol{\varphi})$. Analogous statements hold when $\boldsymbol{Q}$ is a random element of $\mathcal{V}^+(k, p)$.

5. Simulating from $\mathcal{V}(k, p)$ and $\mathcal{V}^+(k, p)$

Practical applications often require simulating a random orthogonal matrix $\boldsymbol{Q}$ whose distribution has a prescribed density g. For instance, Bayesian analyses of statistical models with an orthog-

onal matrix parameter yield posterior densities on the Stiefel manifold, and inference typically requires simulating from these densities. In many cases, generating independent samples is too challenging and MCMC methods are the most sensible option.

In this section, we present an MCMC approach to simulating from a distribution having a density on the set $\mathcal{V}(k, p)$ or $\mathcal{V}^+(k, p)$ which takes advantage of the Cayley parametrizations described in Section 3. The recent work of Pourzanjani et al. [46] explores a similar idea based on a Givens rotation parametrization of the Stiefel manifold. When it is not too computationally expensive, our approach may have certain advantages over existing methods. Unlike Hoff [25], it can be applied regardless of whether conditional distributions belong to a particular parametric family, and it is arguably simpler to implement than the approach of Byrne and Girolami [5]. In statistical applications where interest lies in a subspace rather than a particular orthogonal basis, our representation of the Grassmann manifold $\mathcal{G}(k, p)$ by the set $\mathcal{V}^+(k, p)$ may suggest an appealing parametrization, with the MCMC approach of this section offering the machinery for Bayesian inference.

We illustrate the basic idea with the Stiefel manifold. (Simulating from $\mathcal{V}^+(k, p)$ involves analogous steps.) In order to simulate Q whose distribution has density g on the Stiefel manifold $\mathcal{V}(k, p)$, we construct a Markov chain whose stationary distribution has density $g[C(\boldsymbol{\varphi})]J_{d_\mathcal{V}}C(\boldsymbol{\varphi})$ on the set $\mathcal{D}_\mathcal{V}^{\boldsymbol{\varphi}} = \mathbb{R}^{d_\mathcal{V}}$. Then we simply transform the realized Markov chain back to $\mathcal{V}(k, p)$ using the Cayley transform. By doing so, we avoid the difficulty of choosing and simulating from an efficient proposal distribution defined on the Stiefel manifold.

To make things more concrete, we describe the approach with the Metropolis–Hastings algorithm as our MCMC method. We start with an initial value $\boldsymbol{\varphi}_0 \in \mathcal{D}_\mathcal{V}^{\boldsymbol{\varphi}}$ for our chain and a density $q(\boldsymbol{\varphi}'|\boldsymbol{\varphi})$ for the proposal $\boldsymbol{\varphi}'$ given the previous value $\boldsymbol{\varphi}$. The Metropolis–Hastings algorithm for simulating from the distribution having density $g[C(\boldsymbol{\varphi})]J_{d_\mathcal{V}}C(\boldsymbol{\varphi})$ on $\mathcal{D}_\mathcal{V}^{\boldsymbol{\varphi}}$ proceeds as follows. For $t = 0, \ldots, T$:

1. Generate $\boldsymbol{\varphi}'$ from $q(\boldsymbol{\varphi}'|\boldsymbol{\varphi}_t)$.
2. Compute the acceptance ratio

$$r = \frac{g[C(\boldsymbol{\varphi}')]J_{d_\mathcal{V}}C(\boldsymbol{\varphi}')}{g[C(\boldsymbol{\varphi}_t)]J_{d_\mathcal{V}}C(\boldsymbol{\varphi}_t)} \frac{q(\boldsymbol{\varphi}_t|\boldsymbol{\varphi}')}{q(\boldsymbol{\varphi}'|\boldsymbol{\varphi}_t)}.$$

3. Sample $u \sim \mathrm{Unif}(0, 1)$. If $u \leq r$, set $\boldsymbol{\varphi}_{t+1} = \boldsymbol{\varphi}'$. Otherwise, set $\boldsymbol{\varphi}_{t+1} = \boldsymbol{\varphi}_t$.

For a broad class of g and q, the orthogonal matrices $\{C(\boldsymbol{\varphi}_t)\}_{t=0}^T$ approximate the distribution having density g when T is large enough.

In place of this simple Metropolis–Hastings algorithm, we can substitute other MCMC methods. Hamiltonian Monte Carlo (HMC) [44], with implementations in software such as [6] and [49], is a natural choice. For settings in which evaluation of the Jacobian term is not prohibitively expensive, the Cayley transform approach offers a relatively straightforward path to MCMC simulation on the sets $\mathcal{V}(k, p)$ and $\mathcal{V}^+(k, p)$.

5.1. Example: The uniform distribution on $\mathcal{V}(k, p)$

Using the MCMC approach described above, we simulate from the uniform distribution on $\mathcal{V}(k, p)$. Specifically, we use HMC as implemented in Stan [6] to simulate $Q \sim P_{\mathcal{V}(k, p)}$. Of
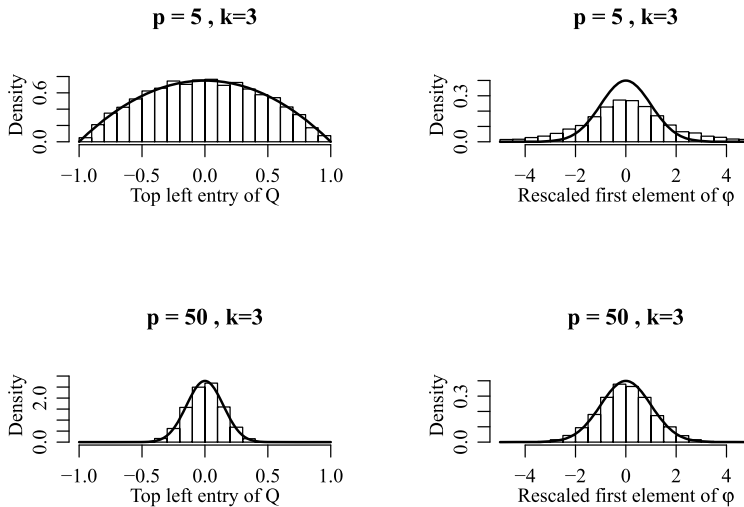
Figure 1. The top left panel is a histogram of the top left entry of simulated values of $Q \sim P_{\mathcal{V}(3,5)}$ plotted against the exact density. The bottom left panel is the analogous plot for $Q \sim P_{\mathcal{V}(3,50)}$. The top right panel is a histogram of the first entry, rescaled by $\sqrt{p/2}$, of $\varphi = C^{-1}(Q)$ when $p = 5$ and $k = 3$. The histogram is plotted against a standard normal density. The bottom right panel is the analogous plot for the case when $p = 50$ and $k = 3$.

course, there exist many algorithms for the simulation of independent, uniformly distributed orthogonal matrices. The uniform distribution only serves as a starting point for illustrating the proposed approach.

Figure 1 provides plots based on 10,000 simulated values of $Q \sim P_{\mathcal{V}(k,p)}$. The top row of the figure deals with the case $p = 5$ and $k = 3$, while the bottom row deals with the case $p = 50$ and $k = 3$. The histograms on the left show the top left entry of the simulated values of Q plotted against the exact density, given in Proposition 7.3 of [21]. As we expect, there is close agreement between the histogram density estimate and the exact density. The histograms on the right show the first entry, rescaled by $\sqrt{p/2}$, of the simulated values of the vector $\varphi = C^{-1}(Q)$. These are plotted against a standard normal density. Theorem 6.1 tells us that the histogram density estimate and the standard normal density should agree when p is large (both in an absolute sense and relative to k), which is what we observe in the plot on the bottom right. When k is similar in magnitude to p, the standard normal density is a poor approximation, as we see in the plot on the top right.

5.2. Example: Bayesian inference for the spiked covariance model

Suppose the rows of an $n \times p$ data matrix Y are independent samples from a mean zero multivariate normal population with covariance matrix Σ. The spiked covariance model, considered by Johnstone [33] and others, assumes the covariance matrix Σ can be decomposed as

$\boldsymbol{\Sigma} = \sigma^2(\boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^T + \boldsymbol{I}_p)$ with $\boldsymbol{Q} \in \mathcal{V}(k, p)$ and $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_k)$ where $\lambda_1 > \cdots > \lambda_k > 0$. Under this model, the covariance matrix is partitioned into the low rank "signal" component $\sigma^2\boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^T$ and the isotropic "noise" component $\sigma^2\boldsymbol{I}_p$.

Given priors for the unknown parameters, a conventional Bayesian analysis will approximate the posterior distribution by a Markov chain having the posterior as its stationary distribution. Inference for the trivially constrained parameters σ^2 and $\boldsymbol{\Lambda}$ is easily handled by standard MCMC approaches, so we treat these parameters as fixed and focus on inference for the orthogonal matrix parameter $\boldsymbol{Q}$. With a uniform prior for $\boldsymbol{Q}$, the posterior distribution is matrix Bingham [25], having density

$$p(\boldsymbol{Q} \mid \boldsymbol{Y}, \sigma^2, \boldsymbol{\Lambda}) \propto \mathrm{etr}\{[(\boldsymbol{\Lambda}^{-1} + \boldsymbol{I}_k)^{-1}/(2\sigma^2)]\boldsymbol{Q}^T[\boldsymbol{Y}^T\boldsymbol{Y}]\boldsymbol{Q}\}.$$

We compare two MCMC approaches to simulating from the matrix Bingham distribution: the Gibbs sampling method of Hoff [25] and the Cayley transform approach (again, with HMC as implemented in Stan). The dimensions are chosen as $n = 100$, $p = 50$, and $k = 3$. We set $\sigma^2 = 1$, $\boldsymbol{\Lambda} = \mathrm{diag}(5, 3, 1.5)$, and choose a true value of $\boldsymbol{Q}$ uniformly from $\mathcal{V}(3, 50)$. We then generate a data matrix according to the model $\mathrm{vec}\,\boldsymbol{Y} \sim \mathcal{N}[\boldsymbol{0}, \sigma^2(\boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^T + \boldsymbol{I}_p) \otimes \boldsymbol{I}_{100}]$. We run each Markov chain for 12,000 steps and discard the first 2000 steps as burn-in. In order to summarize the high-dimensional posterior simulations in terms of lower dimensional quantities, we compute the principal angles between the columns of the simulated $\boldsymbol{Q}$ matrices and the corresponding columns of the posterior mode $\boldsymbol{V}$, computed from the eigendecomposition $\boldsymbol{A} = \boldsymbol{V}\boldsymbol{D}\boldsymbol{V}^T$. For $j = 1, \ldots, 3$, the principal angles are

$$\theta_j = \cos^{-1}\left(\frac{|\boldsymbol{q}_j^T\boldsymbol{v}_j|}{\|\boldsymbol{q}_j\|\|\boldsymbol{v}_j\|}\right),$$

where $\boldsymbol{q}_j$ and $\boldsymbol{v}_j$ are the jth columns of $\boldsymbol{Q}$ and $\boldsymbol{V}$, respectively.

Figure 2 displays the principal angles calculated from the two Markov chains. The plots in the bottom half of the figure compare histogram approximations of the marginal posterior distributions of the principal angles. There is considerable overlap, suggesting that the two chains have found their way to equivalent modes of their matrix Bingham stationary distribution. The plot in the top left of the figure overlays trace plots of the first principal angle calculated from a portion of each of the chains. The black line corresponds to our MCMC approach, while the gray line corresponds to that of Hoff [25]. The plot in the top right shows the correlation between lagged values of the first principal angle. Again, the black dots correspond to our MCMC approach, while the gray dots correspond to that of Hoff [25]. Together, the plots in the top half of the figure indicate that, compared to the approach of Hoff [25], the Cayley transform approach produces a Markov chain with less autocorrelation, reducing Monte Carlo error in the resulting posterior inferences.

6. An asymptotic independent normal approximation

In Section 4, we derived the density for the distribution of $\boldsymbol{\varphi} = C^{-1}(\boldsymbol{Q})$ when $\boldsymbol{Q}$ is distributed uniformly on the Stiefel manifold. However, the expression involves the rather opaque Jacobian
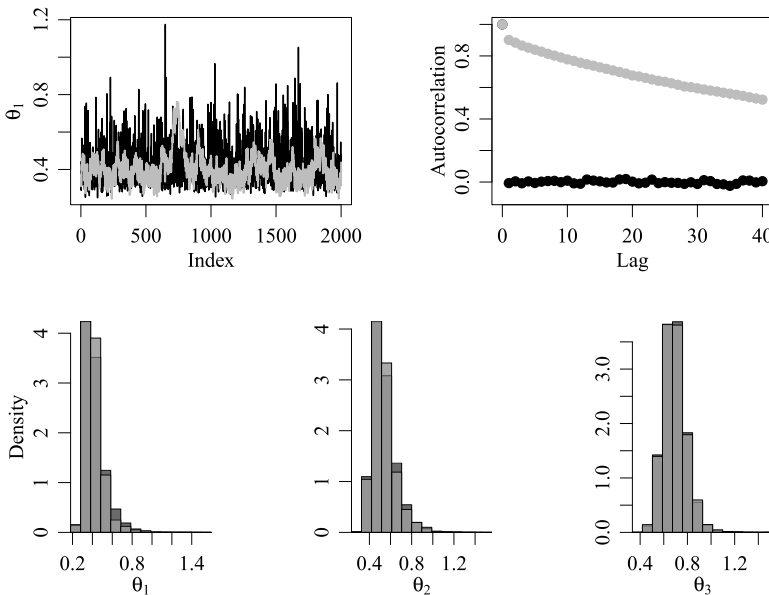
Figure 2. The plot in the top left overlays trace plots of the first principal angle calculated from a portion of each of the chains, while the plot in the top right shows the correlation between lagged values of the first principal angle. The black lines and dots correspond to our MCMC approach, while the gray lines and dots correspond to that of [25]. The plots in the bottom half compare histogram approximations of the marginal posterior distributions of the principal angles.

term $J_{d_V} C(\boldsymbol{\varphi})$. Instead of analyzing the density function, we can gain insight into the distribution of $\boldsymbol{\varphi}$ by other means. The critical observation, evident in simulations, is the following: If $\boldsymbol{Q} \sim P_{\mathcal{V}(k,p)}$ with p large and $p \gg k$ then, in some sense, the elements of $\boldsymbol{\varphi} = C^{-1}(\boldsymbol{Q})$ are approximately independent and normally distributed. Theorem 6.1 of this section provides a mathematical explanation for this empirical phenomenon.

In order to understand Theorem 6.1 and its broader context, it is helpful to review the literature relating to normal approximations to the entries of random orthogonal matrices or transformations thereof. For the sake of consistency and clarity, the notation and formulation of the relevant results have been modified slightly. Let $\{\boldsymbol{Q}_p\}$ be a sequence of random orthogonal matrices with each element $\boldsymbol{Q}_p$ uniform on $\mathcal{V}(k_p, p)$. The notation k_p indicates that the number of columns may grow with the number of rows. For each p, let q_p be the top left entry of $\boldsymbol{Q}_p$ (any other entry would also work). It has long been observed that q_p is approximately normal when p is large. The earliest work in this direction relates to the equivalence of ensembles in statistical mechanics and is due to Mehler [43], Maxwell [40,41], and Borel [2]. A theorem of Borel shows that $\Pr(\sqrt{p} q_p \leq x) \to \Phi(x)$ as p grows, where Φ is the cumulative distribution function of a standard normal random variable. Since then, a large and growing literature on this sort of normal approximation has emerged. A detailed history is given in [15], while an overview is given in [13] and [30].

Much of this literature is devoted to normal approximations to the joint distribution of entries of random orthogonal matrices. Letting q_p be the first column of Q_p for each p (again, any other column would also work), Stam [51] proved that the total variation distance between the distribution of the first m_p coordinates of $\sqrt{p}q_p$ and the distribution of m_p independent standard normal random variables converges to zero as p gets large so long as $m_p = o(\sqrt{p})$. Diaconis and Freedman [15] strengthened this result, showing that it holds for $m_p = o(p)$. Diaconis et al. [19] prove that the total variation distance between the distribution of the top left $m_p \times n_p$ block of $\sqrt{p}Q_p$ and the distribution of $m_p n_p$ independent standard normals goes to zero as $p \to \infty$ if $m_p = o(p^\gamma)$ and $n_p = o(p^\gamma)$ for $\gamma = 1/3$. (Clearly, we must have $n_p \le k_p$ for this result to make sense.) Their work drew attention to the problem of determining the largest orders of m_p and n_p such that the total variation distance goes to zero. Jiang [30] solved this problem, finding the largest orders to be $o(p^{1/2})$. Recent work has further explored this topic [31,53].

Many authors have also considered transformations of random orthogonal matrices or notions of approximation not based on total variation distance. In the former category, D'Aristotile et al. [13] and Meckes [42] study the convergence of linear combinations of the entries of matrices in the sequence $\{Q_p\}$ to normality as $p \to \infty$. Diaconis and Shahshahani [18], Stein [52], Johansson [32], and Rains [47] address the normality of traces of powers of random orthogonal and unitary matrices. In the latter category, Chatterjee and Meckes [8] and Jiang and Ma [31] consider probability metrics other than total variation distance. Jiang [30] also considers a notion of approximation other than total variation distance, and Theorem 3 in that work is particularly important in understanding our Theorem 6.1.

Theorem 3 of [30] tells us that the distribution $P_{\mathcal{V}(k_p,p)}$ can be approximated by the distribution of pk_p independent normals provided that p is sufficiently large (both in an absolute sense and relative to k_p). The form of approximation in the theorem is weaker and likely less familiar than one based on total variation distance. Define the max norm $\|\cdot\|_{\max}$ of a matrix as the maximum of the absolute values of its entries. Jiang [30] shows that one can construct a sequence of pairs of random matrices $\{Z_p, Q_p\}$ with each pair defined on the same probability space such that

    (i) The entries of the $p \times k_p$ matrix Z_p are independent standard normals.
    (ii) The matrix Q_p is uniform on $\mathcal{V}(k_p, p)$.
    (iii) The quantity $\epsilon_p = \|\sqrt{p}Q_p - Z_p\|_{\max} \to 0$ in probability as $p \to \infty$ provided that $k_p = o(p/\log p)$, and this is the largest order of k_p such that the result holds.

The coupling is constructed by letting Q_p be the result of the Gram–Schmidt orthogonalization procedure applied to Z_p.

To better understand this type of approximation, which we refer to as 'approximation in probability,' consider the problem of simulating from the distribution of the random matrix $\sqrt{p}Q_p$ on a computer with finite precision. One could simulate a matrix Z_p of independent standard normals, obtain Q_p using Gram–Schmidt, then multiply by $\sqrt{p}$ to arrive at $\sqrt{p}Q_p$. However, for a fixed machine precision and p sufficiently large (again, both in an absolute sense and relative to k_p), the matrix $\sqrt{p}Q_p$ would be indistinguishable from Z_p with high probability.

Our Theorem 6.1 establishes that the distribution of $\varphi_p = C_p^{-1}(Q_p)$, which we know to have a density proportional to $J_{d_\mathcal{V}}C_p(\varphi)$, can be approximated in probability by independent normals. (Since we now have a sequence of matrices of different dimensions, we denote the Cayley

transform parametrizing $\mathcal{V}(k_p, p)$ by C_p.) For each p, define the diagonal scale matrix

$$\boldsymbol{\Pi}_p = \begin{bmatrix} \sqrt{p/2}\boldsymbol{I}_{k_p(k_p-1)/2} & \mathbf{0} \\ \mathbf{0} & \sqrt{p}\boldsymbol{I}_{(p-k_p)k_p} \end{bmatrix},$$

and recall that the infinity norm $\|\cdot\|_\infty$ of a vector is equal to the maximum of the absolute values of its entries.

Theorem 6.1. *One can construct a sequence of pairs of random vectors* $\{z_p, \boldsymbol{\varphi}_p\}$ *such that*

(i) *The entries of the vector* z_p *are independent standard normals.*

(ii) *The vector* $\boldsymbol{\varphi}_p \stackrel{d}{=} C_p^{-1}(\boldsymbol{Q}_p)$ *where* $\boldsymbol{Q}_p \sim P_{\mathcal{V}(k_p, p)}$.

(iii) *The quantity* $\epsilon_p := \|\boldsymbol{\Pi}_p \boldsymbol{\varphi}_p - z_p\|_\infty \to 0$ *in probability as* $p \to \infty$ *provided* $k_p = o(\frac{p^{1/4}}{\sqrt{\log p}})$.

The construction of the coupling is more elaborate than in Theorem 3 of [30]. We first introduce a function $\widetilde{C}_p^{-1}$ which approximates the inverse Cayley transform. Given a matrix $\boldsymbol{M} \in \mathbb{R}^{p \times k_p}$ having a square top block $\boldsymbol{M}_1$ and a bottom block $\boldsymbol{M}_2$, the vector $\widetilde{b}_p(\boldsymbol{M})$ contains the independent entries of $\widetilde{B}_p(\boldsymbol{M}) = \boldsymbol{M}_1 - \boldsymbol{M}_1^T$ obtained by eliminating diagonal and supradiagonal entries from vec $\widetilde{B}_p(\boldsymbol{M})$ while $\widetilde{A}_p(\boldsymbol{M}) = \boldsymbol{M}_2$. The approximate inverse Cayley transform is then

$$\widetilde{C}_p^{-1}(\boldsymbol{M}) = \begin{bmatrix} \widetilde{b}_p(\boldsymbol{M}) \\ \text{vec } \widetilde{A}_p(\boldsymbol{M}) \end{bmatrix}.$$

Now let $\boldsymbol{Z}_p$ be a $p \times k_p$ matrix of independent standard normals and let $\boldsymbol{Q}_p$ be the result of applying the Gram–Schmidt orthogonalization procedure to $\boldsymbol{Z}_p$. It follows that $\boldsymbol{Q}_p \sim P_{\mathcal{V}(k_p, p)}$. Finally, set

$$z_p = \boldsymbol{\Pi}_p \widetilde{C}_p^{-1}(p^{-1/2}\boldsymbol{Z}_p),$$

$$\boldsymbol{\varphi}_p = C_p^{-1}(\boldsymbol{Q}_p).$$

The details of the proof of Theorem 6.1 appear in the appendix, but we provide a sketch here. Part (i) is straightforward to verify and (ii) is immediate. Part (iii) requires more work. The proof of (iii) involves verifying the following proposition, which involves a third random vector $\widetilde{\boldsymbol{\varphi}}_p = \widetilde{C}_p^{-1}(\boldsymbol{Q}_p)$:

Proposition 6.1. (i) *The quantity* $\|\boldsymbol{\Pi}_p \widetilde{\boldsymbol{\varphi}}_p - \boldsymbol{\Pi}_p \boldsymbol{\varphi}_p\|_\infty \to 0$ *in probability as* $p \to \infty$ *provided* $k_p = o(\frac{p^{1/4}}{\sqrt{\log p}})$, *and* (ii) *the quantity* $\|\boldsymbol{\Pi}_p \widetilde{\boldsymbol{\varphi}}_p - z_p\|_\infty \to 0$ *in probability as* $p \to \infty$ *provided* $k_p = o(\frac{p}{\log p})$.

Part (iii) then follows from the proposition by the triangle inequality.

Whether $k_p = o(\frac{p^{1/4}}{\sqrt{\log p}})$ is the largest order for k_p such that the approximation in Theorem 6.1 holds is an open question. As we saw in the literature review, approximation results like Theorem 6.1 are often strengthened in subsequent work using different techniques. We believe, at least, that we have found the largest order for k_p that our proof strategy allows. For a more detailed discussion of the condition on k_p, see the proof of Lemma A.2 in Section A.6.

Appendix A: Proofs

A.1. The sum of a symmetric positive definite matrix and skew-symmetric matrix is nonsingular

Let $\boldsymbol{\Sigma}$ and $\boldsymbol{S}$ be symmetric positive definite and skew-symmetric, respectively, of equal dimension. Their sum can be written

$$\boldsymbol{\Sigma} + \boldsymbol{S} = \boldsymbol{\Sigma}^{1/2}\big(\boldsymbol{I} + \boldsymbol{\Sigma}^{-1/2}\boldsymbol{S}\boldsymbol{\Sigma}^{-1/2}\big)\boldsymbol{\Sigma}^{1/2}.$$

The matrix $\boldsymbol{\Sigma}^{-1/2}\boldsymbol{S}\boldsymbol{\Sigma}^{-1/2}$ is skew-symmetric, which implies that $\boldsymbol{I} + \boldsymbol{\Sigma}^{-1/2}\boldsymbol{S}\boldsymbol{\Sigma}^{-1/2}$ is nonsingular. Because it can be written as the product of nonsingular matrices, the sum $\boldsymbol{\Sigma} + \boldsymbol{S}$ is nonsingular.

A.2. Proof of Proposition 3.1

One can compute, following [38], that

$$d\boldsymbol{Q} = 2(\boldsymbol{I}_p - \boldsymbol{X}_{\boldsymbol{\varphi}})^{-1}\, d\boldsymbol{X}_{\boldsymbol{\varphi}}(\boldsymbol{I}_p - \boldsymbol{X}_{\boldsymbol{\varphi}})^{-1}\boldsymbol{I}_{p\times k},$$

so that

$$d\operatorname{vec}\boldsymbol{Q} = 2\big[\boldsymbol{I}_{p\times k}^T(\boldsymbol{I}_p - \boldsymbol{X}_{\boldsymbol{\varphi}})^{-T} \otimes \big(\boldsymbol{I}_p - \boldsymbol{X}_{\boldsymbol{\varphi}}^{-1}\big)\big]\, d\operatorname{vec}\boldsymbol{X}_{\boldsymbol{\varphi}}.$$

By Lemma 3.1, we have that $\operatorname{vec}\boldsymbol{X}_{\boldsymbol{\varphi}} = \boldsymbol{\Gamma}_{\mathcal{V}}\boldsymbol{\varphi}$. Thus,

$$d\operatorname{vec}\boldsymbol{Q} = 2\big[\boldsymbol{I}_{p\times k}^T(\boldsymbol{I}_p - \boldsymbol{X}_{\boldsymbol{\varphi}})^{-T} \otimes (\boldsymbol{I}_p - \boldsymbol{X}_{\boldsymbol{\varphi}})^{-1}\big]\boldsymbol{\Gamma}_{\mathcal{V}}\, d\boldsymbol{\varphi}.$$

Using the first identification table of [38], we identify the derivative matrix

$$D\operatorname{Cay}_{\mathcal{V}}(\boldsymbol{\varphi}) = 2\big[\boldsymbol{I}_{p\times k}^T(\boldsymbol{I}_p - \boldsymbol{X}_{\boldsymbol{\varphi}})^{-T} \otimes (\boldsymbol{I}_p - \boldsymbol{X}_{\boldsymbol{\varphi}})^{-1}\big]\boldsymbol{\Gamma}_{\mathcal{V}}.$$

A.3. Proof of Proposition 3.2

We first show that l is injective on $\mathcal{V}^+(k, p)$. Let $\boldsymbol{Q} = [\boldsymbol{Q}_1^T\ \boldsymbol{Q}_2^T]^T$, $\boldsymbol{Q}' = [\boldsymbol{Q}_1'^T\ \boldsymbol{Q}_2'^T]^T \in \mathcal{V}^+(k, p)$ and suppose that $l(\boldsymbol{Q}) = l(\boldsymbol{Q}')$, i.e. the columns of $\boldsymbol{Q}$ and $\boldsymbol{Q}'$ span the same subspace. There

must exist $R \in \mathcal{O}(k)$ such that $Q = Q'R$ so that $Q_1 = Q_1'R$. Because the matrix Q_1 is nonsingular, its left polar decomposition into the product of a symmetric positive definite matrix and an orthogonal matrix is unique (see, for example, Proposition 5.5 of [20]). We conclude that $R = I_k$ and $Q_1 = Q_1'$. Thus, l is injective on $\mathcal{V}^+(k, p)$.

Next, we prove that the image of $\mathcal{V}^+(k, p)$ under l has measure one with respect to the uniform probability measure $P_{\mathcal{G}(k,p)}$ on $\mathcal{G}(k, p)$. Define $\mathcal{V}^{\mathrm{N}}(k, p)$ as the set

$$\mathcal{V}^{\mathrm{N}}(k, p) = \left\{ Q = \begin{bmatrix} Q_1^T & Q_2^T \end{bmatrix}^T \in \mathcal{V}(k, p) : Q_1 \text{ is nonsingular} \right\}.$$

The set $\mathcal{V}^{\mathrm{N}}(k, p)$ has measure one with respect to $P_{\mathcal{V}(k,p)}$ and the following lemma holds:

Lemma A.1. *The images of $\mathcal{V}^+(k, p)$ and $\mathcal{V}^{\mathrm{N}}(k, p)$ under l are equal.*

Proof. The direction $l[\mathcal{V}^+(k, p)] \subseteq l[\mathcal{V}^{\mathrm{N}}(k, p)]$ is immediate. Now, let $S \in l[\mathcal{V}^{\mathrm{N}}(k, p)]$. There must exist $Q = [Q_1^T \; Q_2^T]^T \in \mathcal{V}^{\mathrm{N}}(k, p)$ having S as its column space. Let $Q_1 = UDV^T$ be the singular value decomposition of Q_1 and set $Q' = QVU^T$. Then $l(Q') = S$ because $VU^T \in \mathcal{O}(k)$ and $Q' \in \mathcal{V}^+(k, p)$ because its square top block UDU^T is symmetric positive definite. Thus $S \in l[\mathcal{V}^+(k, p)]$ and we conclude that $l[\mathcal{V}^{\mathrm{N}}(k, p)] \subseteq l[\mathcal{V}^+(k, p)]$. $\qquad\square$

Recall that the measure $P_{\mathcal{G}(k,p)}$ is the pushforward of $P_{\mathcal{V}(k,p)}$ by l, i.e. for a subset $A \subset \mathcal{G}(k, p)$ we have $P_{\mathcal{G}(k,p)}(A) = P_{\mathcal{V}(k,p)}[l^{-1}(A)]$. Thus,

$$\begin{aligned}
P_{\mathcal{G}(k,p)}\left\{ l[\mathcal{V}^+(k, p)] \right\} &= P_{\mathcal{G}(k,p)}\left\{ l[\mathcal{V}^{\mathrm{N}}(k, p)] \right\} \\
&= P_{\mathcal{V}(k,p)}\left\{ \mathcal{V}^{\mathrm{N}}(k, p) \right\} \\
&= 1.
\end{aligned}$$

A.4. Proof of Proposition 3.3

We begin with $Q = [Q_1^T \; Q_2^T] \in \mathcal{V}^+(k, p)$ and we want to verify that the matrix A obtained by equations (3.4)–(3.5) satisfies $0 \preceq A^\top A \prec I_k$. Let $Q_1 = V \operatorname{diag}(\lambda_1, \ldots, \lambda_k) V^T$ be the eigendecomposition of Q_1. That each eigenvalue of Q_1 is positive follows from the condition $Q_1 \succ 0$. We know that each eigenvalue is less than or equal to one because

$$\begin{aligned}
0_k &\preceq Q_2^T Q_2 \\
&= I_k - Q_1^T Q_1 \\
&= V \operatorname{diag}\left(1 - \lambda_1^2, \ldots, 1 - \lambda_k^2\right) V^T.
\end{aligned}$$

Therefore, $\lambda_i \in (0, 1]$ for each i. As in equations (3.4)–(3.5), set

$$F = (I_k - Q_1)(I_k + Q_1)^{-1},$$

$$A = \frac{1}{2} Q_2(I_k + F).$$

We can write $A^T A$ as

$$A^T A = \frac{1}{4}(I_k + F)^T Q_2^T Q_2 (I_k + F)$$

$$= \frac{1}{4}(I_k + F)^T (I_k - Q_1^T Q_1)(I_k + F)$$

from which it follows that

$$\text{eval}_i (A^T A) = \frac{1}{4}(1 - \lambda_i^2)\left(1 + \frac{1 - \lambda_i}{1 + \lambda_i}\right)^2$$

for each i. The notation $\text{eval}_i (A^T A)$ indicates the ith eigenvalue of $A^T A$. Since the eigenvalues of Q_1 lie in the interval $(0, 1]$, the eigenvalues of $A^T A$ lie in the interval $[0, 1)$.

Now we start with A such that $\delta_i = \text{eval}_i (A^T A) \in [0, 1)$ for each i and we want to check that $Q_1 = (I_k - A^T A)(I_k + A^T A)^{-1} \succ 0$. Let $A^T A = W \, \text{diag}(\delta_1, \ldots, \delta_k) W^T$ be the eigendecomposition of $A^T A$. Then

$$Q_1 = W \, \text{diag}\left(\frac{1 - \delta_1}{1 + \delta_1}, \ldots, \frac{1 - \delta_k}{1 + \delta_k}\right) W^T \succ 0.$$

A.5. Proof of Proposition 3.4

The proof is nearly identical to that of Proposition 3.1. We simply replace φ with ψ and $\Gamma_\mathcal{V}$ with $\Gamma_\mathcal{G}$.

A.6. Proof of Proposition 6.1 part (i)

Denote the square top block of Q_p by $Q_{p,1}$ and the bottom block by $Q_{p,2}$. Define matrices

$$F_p = (I_{k_p} - Q_{p,1})(I_{k_p} + Q_{p,1})^{-1},$$

$$B_p = \frac{1}{2}(F_p^T - F_p),$$

$$A_p = \frac{1}{2} Q_{p,2}(I_{k_p} + F_p)$$

as in equations (3.1)–(3.3). Let b_p be the vector of independent entries of B_p obtained by eliminating diagonal and supradiagonal elements from vec B_p. When the Frobenius norm $\| Q_{p,1} \|_F$ is less than one, the matrices admit series representations:

$$F_p = I_{k_p} - 2 Q_{p,1} + 2 Q_{p,1}^2 - 2 Q_{p,1}^3 + \cdots,$$

$$B_p = (Q_{p,1} - Q_{p,1}^T) - (Q_{p,1}^2 - Q_{p,1}^{2T}) + (Q_{p,1}^3 - Q_{p,1}^{3T}) - \cdots,$$

$$A_p = Q_{p,2} - Q_{p,2} Q_{p,1} + Q_{p,2} Q_{p,1}^2 - Q_2 Q_{p,1}^3 + \cdots.$$

It follows that

$$\widetilde{B}_p(Q_p) - B_p = \sum_{i=2}^{\infty} (-1)^i \left(Q_{p,1}^i - Q_{p,1}^{iT} \right),$$

$$\widetilde{A}_p(Q_p) - A_p = \sum_{i=1}^{\infty} (-1)^{i-1} Q_{p,2} Q_{p,1}^i,$$

when $\|Q_{p,1}\|_F < 1$. See [26] for a discussion of matrix geometric series.

Inequalities relating the Frobenius and max norms will prove useful. Let W_1, W_2 be $d_1 \times d_2$ and $d_2 \times d_3$ dimensional matrices, respectively. Then

$$\|W_1\|_F \le \sqrt{d_1 d_2} \|W_1\|_{\max},$$

$$\|W_1 W_2\|_{\max} \le d_2 \|W_1\|_{\max} \|W_2\|_{\max}.$$

The first inequality implies that the condition $\|Q_{p,1}\|_F < 1$ under which our series representations converge is satisfied when $k_p \|Q_{p,1}\|_{\max} < 1$. The second inequality implies that $\|Q_{p,1}^i\|_{\max} \le k_p^{i-1} \|Q_{p,1}\|_{\max}^i$ for each natural number i.

We will also need the following somewhat technical lemma.

Lemma A.2. *The following quantities, which will appear in later inequalities, go to zero in probability as p grows provided that $k_p = o(\frac{p^{1/4}}{\sqrt{\log p}})$:*

(i) $\left\| p^{1/2} Q_p - Z_p \right\|_{\max}$,

(ii) $k_p \| Q_{p,1} \|_{\max}$,

(iii) $k_p^2 p^{1/2} \| Q_{p,1} \|_{\max}^2$,

(iv) $k_p p^{1/2} \| Q_p \|_{\max}^2$.

Proof. Suppose $k_p = o(\frac{p^{1/4}}{\sqrt{\log p}})$. This implies that $k_p = o(\frac{p}{\log p})$ and quantity (i) goes to zero in probability as p grows by Theorem 3 of [30]. The quantities (ii)–(iv) are nonnegative and bounded above by either $k_p^2 p^{1/2} \| Q_p \|_{\max}^2$ or its square root. It is sufficient to bound the square root above and show that the upper bound goes to zero in probability. The square root satisfies

$$k_p p^{1/4} \| Q_p \|_{\max} = k_p p^{1/4} \left\| Q_p - p^{-1/2} Z_p + p^{-1/2} Z_p \right\|_{\max}$$

$$\le k_p p^{1/4} \left(\left\| Q_p - p^{-1/2} Z_p \right\|_{\max} + \left\| p^{-1/2} Z_p \right\|_{\max} \right)$$

$$= k_p p^{-1/4} \left\| p^{1/2} Q_p - Z_p \right\|_{\max} + k_p p^{-1/4} \| Z_p \|_{\max}.$$

The first summand goes to zero in probability by the condition on k_p and part (i) of the lemma. The expected value of the second summand can be bounded above using a well-known inequality

involving the maximum of independent standard normal random variables (see Problem 5.1 in [55]):

$$k_p p^{-1/4} \mathbb{E}\|\boldsymbol{Z}_p\|_{\max} \le k_p p^{-1/4}\sqrt{2\log pk_p}$$

$$\le k_p p^{-1/4}\sqrt{2\log p^2}$$

$$= 2\frac{k_p}{\left(\frac{p^{1/4}}{\sqrt{\log p}}\right)}. \tag{A.1}$$

The condition on k_p implies that the second summand goes to zero in mean and thus in probability. The upper bound for the square root of $k_p^2 p^{1/2}\|\boldsymbol{Q}_p\|_{\max}^2$ goes to zero in probability as p grows and thus quantities (ii)–(iv) do as well.

One may wonder whether we can relax the condition $k_p = o(\frac{p^{1/4}}{\sqrt{\log p}})$ by treating quantities (ii)–(iv) separately or by being more careful in inequality (A.1). It turns out we cannot. For the proof to hold, we would still need that $k_p\sqrt{\log k_p} = o(p^{1/4})$ or, equivalently, that

$$k_p = o\big[e^{W_0(2\sqrt{p})/2}\big],$$

where W_0 is the principal branch of Lambert's W function [11]. This is not an improved order for k_p, because

$$\lim_{p\to\infty} \frac{e^{W_0(2\sqrt{p})/2}}{\left(\frac{p^{1/4}}{\sqrt{\log p}}\right)} = 2. \qquad \square$$

Now set $a_p = \|\boldsymbol{\Pi}_p\widetilde{\boldsymbol{\varphi}}_p - \boldsymbol{\Pi}_p\boldsymbol{\varphi}_p\|_\infty$, assume that $k_p = o(\frac{p^{1/4}}{\sqrt{\log p}})$, and let $\epsilon > 0$ be given. We want to show that $\Pr\{a_p > \epsilon\} \to 0$ as p gets large. We express this probability as

$$\Pr\{a_p > \epsilon\} = \Pr\{a_p > \epsilon \mid k_p\|\boldsymbol{Q}_{p,1}\|_{\max} < 1\}\Pr\{k_p\|\boldsymbol{Q}_{p,1}\|_{\max} < 1\}$$

$$+ \Pr\{a_p > \epsilon \mid k_p\|\boldsymbol{Q}_{p,1}\|_{\max} \ge 1\}\Pr\{k_p\|\boldsymbol{Q}_{p,1}\|_{\max} \ge 1\}$$

$$\le \Pr\{a_p > \epsilon \mid k_p\|\boldsymbol{Q}_{p,1}\|_{\max} < 1\} + \Pr\{k_p\|\boldsymbol{Q}_{p,1}\|_{\max} \ge 1\}.$$

Lemma A.2 implies that $\Pr\{k_p\|\boldsymbol{Q}_{p,1}\|_{\max} \ge 1\}$ goes to zero. It follows that $\Pr\{a_p > \epsilon\}$ goes to zero if $\Pr\{a_p > \epsilon \mid k_p\|\boldsymbol{Q}_{p,1}\|_{\max} < 1\}$ does. Therefore, we only need to verify this condition.

For each p, we have

$$a_p = \|\boldsymbol{\Pi}_p\widetilde{\boldsymbol{\varphi}}_p - \boldsymbol{\Pi}_p\boldsymbol{\varphi}_p\|_\infty$$

$$= \left\|\begin{bmatrix} \sqrt{p/2}\,\widetilde{\boldsymbol{b}}_p(\boldsymbol{Q}_p) \\ \sqrt{p}\,\mathrm{vec}\,\widetilde{\boldsymbol{A}}_p(\boldsymbol{Q}_p) \end{bmatrix} - \begin{bmatrix} \sqrt{p/2}\,\boldsymbol{b}_p \\ \sqrt{p}\,\mathrm{vec}\,\boldsymbol{A}_p \end{bmatrix}\right\|_\infty$$

$$= \max\{\sqrt{p/2}\|\widetilde{\boldsymbol{B}}_p(\boldsymbol{Q}_p) - \boldsymbol{B}_p\|_{\max}, \sqrt{p}\|\widetilde{\boldsymbol{A}}_p(\boldsymbol{Q}_p) - \boldsymbol{A}_p\|_{\max}\}$$

$$\le \sqrt{p/2}\|\widetilde{\boldsymbol{B}}_p(\boldsymbol{Q}_p) - \boldsymbol{B}_p\|_{\max} + \sqrt{p}\|\widetilde{\boldsymbol{A}}_p(\boldsymbol{Q}_p) - \boldsymbol{A}_p\|_{\max}.$$

When $k_p \| \boldsymbol{Q}_{p,1} \|_{\max} < 1$, we have

$$
\begin{aligned}
\sqrt{p/2} \big\| \widetilde{\boldsymbol{B}}_p(\boldsymbol{Q}_p) - \boldsymbol{B}_p \big\|_{\max} &= \sqrt{p/2} \left\| \sum_{i=2}^{\infty} (-1)^i \big(\boldsymbol{Q}_{p,1}^i - \boldsymbol{Q}_{p,1}^{iT} \big) \right\|_{\max} \\
&\leq \sqrt{p/2} \sum_{i=2}^{\infty} \big\| \boldsymbol{Q}_{p,1}^i - \boldsymbol{Q}_{p,1}^{iT} \big\|_{\max} \\
&\leq \sqrt{p/2} \sum_{i=2}^{\infty} 2 \big\| \boldsymbol{Q}_{p,1}^i \big\|_{\max} \\
&= \sqrt{2p} \sum_{i=2}^{\infty} \big\| \boldsymbol{Q}_{p,1}^i \big\|_{\max} \\
&\leq \sqrt{2p} \sum_{i=2}^{\infty} k_p^{i-1} \| \boldsymbol{Q}_{p,1} \|_{\max}^i \\
&= \sqrt{2p} \sum_{i=2}^{\infty} \big(k_p^{\frac{i-1}{i}} \| \boldsymbol{Q}_{p,1} \|_{\max} \big)^i \\
&\leq \sqrt{2p} \sum_{i=2}^{\infty} \big(k_p \| \boldsymbol{Q}_{p,1} \|_{\max} \big)^i \\
&= \sqrt{2p} \left(\frac{1}{1 - k_p \| \boldsymbol{Q}_{p,1} \|_{\max}} - k_p \| \boldsymbol{Q}_{p,1} \|_{\max} - 1 \right) \\
&= \sqrt{2} \frac{k_p^2 \sqrt{p} \| \boldsymbol{Q}_{p,1} \|_{\max}^2}{1 - k_p \| \boldsymbol{Q}_{p,1} \|_{\max}}
\end{aligned}
$$

and

$$
\begin{aligned}
\sqrt{p} \big\| \widetilde{\boldsymbol{A}}_p(\boldsymbol{Q}_p) - \boldsymbol{A}_p \big\|_{\max} &= \sqrt{p} \left\| \sum_{i=1}^{\infty} (-1)^{i-1} \boldsymbol{Q}_{p,2} \boldsymbol{Q}_{p,1}^i \right\|_{\max} \\
&\leq \sqrt{p} \sum_{n=1}^{\infty} \big\| \boldsymbol{Q}_{p,2} \boldsymbol{Q}_{p,1}^n \big\|_{\max} \\
&\leq \sqrt{p} \sum_{n=1}^{\infty} k_p \| \boldsymbol{Q}_{p,2} \|_{\max} \big\| \boldsymbol{Q}_{p,1}^n \big\|_{\max} \\
&\leq \sqrt{p} \| \boldsymbol{Q}_{p,2} \|_{\max} \sum_{n=1}^{\infty} \big(k_p \| \boldsymbol{Q}_{p,1} \|_{\max} \big)^n
\end{aligned}
$$

$$= \sqrt{p} \| \boldsymbol{Q}_{p,2} \|_{\max} \left(\frac{1}{1 - k_p \| \boldsymbol{Q}_{p,1} \|_{\max}} - 1 \right)$$

$$= \sqrt{p} \| \boldsymbol{Q}_{p,2} \|_{\max} \left(\frac{k_p \| \boldsymbol{Q}_{p,1} \|_{\max}}{1 - k_p \| \boldsymbol{Q}_{p,1} \|_{\max}} \right)$$

$$\leq \frac{k_p \sqrt{p} \max\{ \| \boldsymbol{Q}_{p,1} \|_{\max}, \| \boldsymbol{Q}_{p,2} \|_{\max} \}^2}{1 - k_p \| \boldsymbol{Q}_{p,1} \|_{\max}}$$

$$= \frac{k_p \sqrt{p} \| \boldsymbol{Q}_p \|_{\max}^2}{1 - k_p \| \boldsymbol{Q}_{p,1} \|_{\max}}.$$

Thus, the upper bound

$$a_p \leq \sqrt{p/2} \| \widetilde{\boldsymbol{B}}_p(\boldsymbol{Q}_p) - \boldsymbol{B}_p \|_{\max} + \sqrt{p} \| \widetilde{\boldsymbol{A}}_p(\boldsymbol{Q}_p) - \boldsymbol{A}_p \|_{\max}$$

$$\leq \sqrt{2} \frac{k_p^2 \sqrt{p} \| \boldsymbol{Q}_{p,1} \|_{\max}^2}{1 - k_p \| \boldsymbol{Q}_{p,1} \|_{\max}} + \frac{k_p \sqrt{p} \| \boldsymbol{Q}_p \|_{\max}^2}{1 - k_p \| \boldsymbol{Q}_{p,1} \|_{\max}} := u_p$$

is valid when $k_p \| \boldsymbol{Q}_{p,1} \|_{\max} < 1$. Then

$$\Pr\{ a_p > \epsilon \mid k_p \| \boldsymbol{Q}_{p,1} \|_{\max} < 1 \} \leq \Pr\{ u_p > \epsilon \mid k_p \| \boldsymbol{Q}_{p,1} \|_{\max} < 1 \}$$

$$= \frac{\Pr\{ u_p > \epsilon \cap k_p \| \boldsymbol{Q}_{p,1} \|_{\max} < 1 \}}{\Pr\{ k_p \| \boldsymbol{Q}_{p,1} \|_{\max} < 1 \}}$$

$$\leq \frac{\Pr\{ u_p > \epsilon \}}{\Pr\{ k_p \| \boldsymbol{Q}_{p,1} \|_{\max} < 1 \}}.$$

Because $\Pr\{ k_p \| \boldsymbol{Q}_{p,1} \|_{\max} < 1 \} \to 1$, we only need to show $\Pr\{ u_p > \epsilon \} \to 0$ as p grows. Since ϵ is arbitrary, this is equivalent to showing that u_p goes to zero in probability, which follows from the continuous mapping theorem and Lemma A.2. We have shown that $\Pr\{ a_p > \epsilon \mid k_p \| \boldsymbol{Q}_{p,1} \|_{\max} < 1 \}$ goes to zero as p gets large, which is sufficient to prove part (i) of the proposition.

A.7. Proof of Proposition 6.1 part (ii)

Assume that $k_p = o(\frac{p}{\log p})$. For each p, we have:

$$\| \boldsymbol{\Pi}_p \widetilde{\boldsymbol{\varphi}}_p - z_p \|_\infty = \| \boldsymbol{\Pi}_p \widetilde{C}_p^{-1}(\boldsymbol{Q}_p) - \boldsymbol{\Pi}_p \widetilde{C}_p^{-1}(p^{-1/2} \boldsymbol{Z}_p) \|_\infty$$

$$= \left\| \begin{bmatrix} \sqrt{p/2} \widetilde{b}_p(\boldsymbol{Q}_p) \\ \sqrt{p} \, \mathrm{vec} \, \widetilde{A}_p(\boldsymbol{Q}_p) \end{bmatrix} - \begin{bmatrix} \sqrt{p/2} \widetilde{b}_p(p^{-1/2} \boldsymbol{Z}_p) \\ \sqrt{p} \, \mathrm{vec} \, \widetilde{A}_p(p^{-1/2} \boldsymbol{Z}_p) \end{bmatrix} \right\|_\infty$$

$$= \max\{ \sqrt{p/2} \| \widetilde{B}_p(\boldsymbol{Q}_p) - \widetilde{B}_p(p^{-1/2} \boldsymbol{Z}_p) \|_{\max},$$

$$\sqrt{p}\big\|\widetilde{A}_p(\boldsymbol{Q}_p) - \widetilde{A}_p\big(p^{-1/2}\boldsymbol{Z}_p\big)\big\|_{\max}\big\}$$

$$\leq \sqrt{p/2}\big\|\widetilde{B}_p(\boldsymbol{Q}_p) - \widetilde{B}_p\big(p^{-1/2}\boldsymbol{Z}_p\big)\big\|_{\max}$$

$$+ \sqrt{p}\big\|\widetilde{A}_p(\boldsymbol{Q}_p) - \widetilde{A}_p\big(p^{-1/2}\boldsymbol{Z}_p\big)\big\|_{\max}$$

$$= \sqrt{p/2}\big\|\boldsymbol{Q}_{p,1} - p^{-1/2}\boldsymbol{Z}_{p,1} + p^{-1/2}\boldsymbol{Z}_{p,1}^T - \boldsymbol{Q}_{p,1}^T\big\|_{\max}$$

$$+ \sqrt{p}\big\|\boldsymbol{Q}_{p,2} - p^{-1/2}\boldsymbol{Z}_{p,2}\big\|_{\max}$$

$$\leq \sqrt{p/2}\big(\big\|\boldsymbol{Q}_{p,1} - p^{-1/2}\boldsymbol{Z}_{p,1}\big\| + \big\|p^{-1/2}\boldsymbol{Z}_{p,1}^T - \boldsymbol{Q}_{p,1}^T\big\|_{\max}\big)$$

$$+ \sqrt{p}\big\|\boldsymbol{Q}_{p,2} - p^{-1/2}\boldsymbol{Z}_{p,2}\big\|_{\max}$$

$$= 2\sqrt{p/2}\big\|\boldsymbol{Q}_{p,1} - p^{-1/2}\boldsymbol{Z}_{p,1}\big\|_{\max} + \sqrt{p}\big\|\boldsymbol{Q}_{p,2} - p^{-1/2}\boldsymbol{Z}_{p,2}\big\|_{\max}$$

$$= \sqrt{2}\|\sqrt{p}\,\boldsymbol{Q}_{p,1} - \boldsymbol{Z}_{p,1}\|_{\max} + \|\sqrt{p}\,\boldsymbol{Q}_{p,2} - \boldsymbol{Z}_{p,2}\|_{\max}$$

$$\leq \sqrt{2}\|\sqrt{p}\,\boldsymbol{Q}_p - \boldsymbol{Z}_p\|_{\max} + \|\sqrt{p}\,\boldsymbol{Q}_p - \boldsymbol{Z}_p\|_{\max}$$

$$= (\sqrt{2} + 1)\|\sqrt{p}\,\boldsymbol{Q}_p - \boldsymbol{Z}_p\|_{\max}.$$

Theorem 3 of [30] implies that this upper bound goes to zero in probability as p grows. Therefore, the quantity $\|\boldsymbol{\Pi}_p\widetilde{\boldsymbol{\varphi}}_p - z_p\|_\infty$ does as well.

Appendix B: Special matrices

In constructing the linear transformations given in Lemmas 3.1 and 3.2, we rely upon two special matrices: the commutation matrix $\boldsymbol{K}_{m,n}$ and the matrix $\widetilde{\boldsymbol{D}}_n$. An early reference related to the matrix $\boldsymbol{K}_{m,n}$ is [37], while the matrix $\widetilde{\boldsymbol{D}}_n$ was introduced in [45]. Our presentation follows that of [36].

B.1. The commutation matrix $\boldsymbol{K}_{m,n}$

Let A be an $m \times n$ matrix and B be a $p \times q$ matrix. The commutation matrix $\boldsymbol{K}_{m,n}$ is the unique $mn \times mn$ permutation matrix with the property that $\boldsymbol{K}_{m,n} \operatorname{vec} A = \operatorname{vec} A^T$. The critical property of the commutation matrix is that it allows us to exchange the order of the matrices in a Kronecker product. Theorem 3.1 of [36] states that $\boldsymbol{K}_{p,m}(A \otimes B) = (B \otimes A)\boldsymbol{K}_{q,n}$. Section 3.3 of [36] gives an explicit expression for the commutation matrix $\boldsymbol{K}_{m,n}$. Let $\boldsymbol{H}_{i,j}$ be the $m \times n$ matrix having a 1 in the i, jth position and zeros everywhere else. Then Theorem 3.2 of [36] tells us that

$$\boldsymbol{K}_{m,n} = \sum_{i=1}^m \sum_{j=1}^n \big(\boldsymbol{H}_{i,j} \otimes \boldsymbol{H}_{i,j}^T\big).$$

B.2. The matrix $\widetilde{D}_n$

The matrix $\widetilde{D}_n$ proves useful when differentiating expressions involving skew-symmetric matrices. Let A be an $n \times n$ matrix and let $\tilde{v}(A)$ be the $n(n-1)/2$-dimensional vector obtained by eliminating diagonal and supradiagonal elements from the vectorization vec A. Definition 6.1 of [36] defines $\widetilde{D}_n$ as the unique $n^2 \times n(n-1)/2$ matrix with the property $\widetilde{D}_n \tilde{v}(A) = \text{vec } A$ for every skew-symmetric matrix A. Theorem 6.1 of [36] gives an explicit expression for $\widetilde{D}_n$. Let $E_{i,j}$ be the $n \times n$ matrix with 1 in the i,jth position and zeros everywhere else and set $\tilde{T}_{i,j} = E_{i,j} - E_{j,i}$. Also, let $\tilde{u}_{i,j}$ be the $n(n-1)/2$-dimensional vector having 1 in its $(j-1)n + i - j(j+1)/2$ place and zeros everywhere else. Then Theorem 6.1 tells us that

$$\widetilde{D}_n = \sum_{i>j}(\text{vec } \tilde{T}_{i,j})\tilde{u}_{i,j}^T.$$

Appendix C: Evaluating the Jacobian terms

Taking a naive approach to evaluating the Jacobian term $J_{d_\mathcal{V}}C(\boldsymbol{\varphi})$ becomes prohibitively expensive for even small dimensions. Recall that

$$J_{d_\mathcal{V}}C(\boldsymbol{\varphi}) = \left|DC(\boldsymbol{\varphi})^T DC(\boldsymbol{\varphi})\right|^{1/2}$$
$$= \left|2^2\boldsymbol{\Gamma}_\mathcal{V}^T(\boldsymbol{G}_\mathcal{V} \otimes \boldsymbol{H}_\mathcal{V})\boldsymbol{\Gamma}_\mathcal{V}\right|^{1/2},$$

where

$$\boldsymbol{G}_\mathcal{V} = (\boldsymbol{I}_p - \boldsymbol{X}_{\boldsymbol{\varphi}})^{-1}\boldsymbol{I}_{p \times k}\boldsymbol{I}_{p \times k}^T(\boldsymbol{I}_p - \boldsymbol{X}_{\boldsymbol{\varphi}})^{-T},$$
$$\boldsymbol{H}_\mathcal{V} = (\boldsymbol{I}_p - \boldsymbol{X}_{\boldsymbol{\varphi}})^{-T}(\boldsymbol{I}_p - \boldsymbol{X}_{\boldsymbol{\varphi}})^{-1}.$$

The Kronecker product $\boldsymbol{G}_\mathcal{V} \otimes \boldsymbol{H}_\mathcal{V}$ has dimension $p^2 \times p^2$. Evaluating this Kronecker product and computing its matrix product with $\boldsymbol{\Gamma}_\mathcal{V}$ is extremely costly for large p. In this section, we describe a more efficient approach which takes advantage of the block stucture of the matrices involved. (The Jacobian term $J_{d_\mathcal{G}}C(\boldsymbol{\psi})$ can be evaluated analogously.)

Let $\boldsymbol{C}_\mathcal{V} = (\boldsymbol{I} - \boldsymbol{X}_{\boldsymbol{\varphi}})^{-1}$. Then

$$\boldsymbol{C}_\mathcal{V} = \begin{bmatrix} \boldsymbol{C}_{11} & \boldsymbol{C}_{12} \\ \boldsymbol{C}_{21} & \boldsymbol{C}_{22} \end{bmatrix},$$

where

$$\boldsymbol{C}_{11} = \left(\boldsymbol{I}_k - \boldsymbol{B} + \boldsymbol{A}^T\boldsymbol{A}\right)^{-1}, \qquad \boldsymbol{C}_{12} = -\boldsymbol{C}_{11}\boldsymbol{A}^T,$$
$$\boldsymbol{C}_{21} = \boldsymbol{A}\boldsymbol{C}_{11}, \qquad \boldsymbol{C}_{22} = \boldsymbol{I}_{p-k} - \boldsymbol{A}\boldsymbol{C}_{11}\boldsymbol{A}^T.$$

The blocks of the matrices

$$G_{\mathcal{V}} = \begin{bmatrix} G_{11} & G_{12} \\ G_{21} & G_{22} \end{bmatrix}, \qquad H_{\mathcal{V}} = \begin{bmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{bmatrix}$$

can be written in terms of the blocks of $C_{\mathcal{V}}$ as

$$H_{11} = C_{11}^T C_{11} + C_{21}^T C_{21}, \qquad H_{12} = C_{11}^T C_{12} + C_{21}^T C_{22},$$

$$H_{21} = C_{12}^T C_{11} + C_{22}^T C_{21}, \qquad H_{22} = C_{12}^T C_{12} + C_{22}^T C_{22}$$

and

$$G_{11} = C_{11} C_{11}^T, \qquad G_{12} = C_{11} C_{21}^T,$$

$$G_{21} = C_{21} C_{11}^T, \qquad G_{22} = C_{21} C_{21}^T.$$

We can express the matrix $DC(\boldsymbol{\varphi})^T DC(\boldsymbol{\varphi})$ in blocks as

$$DC(\boldsymbol{\varphi})^T DC(\boldsymbol{\varphi}) = 2^2 \boldsymbol{\Gamma}_{\mathcal{V}}^T (G_{\mathcal{V}} \otimes H_{\mathcal{V}}) \boldsymbol{\Gamma}_{\mathcal{V}}$$

$$= 2^2 \begin{bmatrix} \boldsymbol{\Omega}_{11} & \boldsymbol{\Omega}_{12} \\ \boldsymbol{\Omega}_{21} & \boldsymbol{\Omega}_{22} \end{bmatrix},$$

where

$$\boldsymbol{\Omega}_{11} = \widetilde{D}_k^T (G_{11} \otimes H_{11}) \widetilde{D}_k,$$

$$\boldsymbol{\Omega}_{12} = \widetilde{D}_k^T (G_{11} \otimes H_{12}) - \widetilde{D}_k^T (G_{12} \otimes H_{11}) K_{p-k,k},$$

$$\boldsymbol{\Omega}_{21} = \boldsymbol{\Omega}_{12}^T,$$

$$\boldsymbol{\Omega}_{22} = (G_{11} \otimes H_{22} + H_{11} \otimes G_{22}) - (G_{12} \otimes H_{21} + H_{12} \otimes G_{21}) K_{p-k,k}.$$

Then

$$J_{d_{\mathcal{V}}} C(\boldsymbol{\varphi}) = \left| 2^2 \begin{bmatrix} \boldsymbol{\Omega}_{11} & \boldsymbol{\Omega}_{12} \\ \boldsymbol{\Omega}_{21} & \boldsymbol{\Omega}_{22} \end{bmatrix} \right|^{1/2}$$

$$= 2^{d_{\mathcal{V}}} |\boldsymbol{\Omega}_{22}| |\boldsymbol{\Omega}_{11} - \boldsymbol{\Omega}_{12} \boldsymbol{\Omega}_{22}^{-1} \boldsymbol{\Omega}_{21}|.$$

The Kronecker products in the formulas for $\boldsymbol{\Omega}_{11}$, $\boldsymbol{\Omega}_{12}$, $\boldsymbol{\Omega}_{21}$, and $\boldsymbol{\Omega}_{22}$ are much smaller than $G_{\mathcal{V}} \otimes H_{\mathcal{V}}$, which leads to a significant computational savings compared to the naive approach.

Acknowledgements

References

[1] Anderson, T.W., Olkin, I. and Underhill, L.G. (1987). Generation of random orthogonal matrices. *SIAM J. Sci. Statist. Comput.* **8** 625–629. MR0892309 https://doi.org/10.1137/0908055

[2] Borel, É. (1906). Sur les principes de la théorie cinétique des gaz. *Ann. Sci. Éc. Norm. Supér.* (3) **23** 9–32. MR1509063

[3] Bourgade, P., Nikeghbali, A. and Rouault, A. (2007). Hua–Pickrell measures on general compact groups. Available at arXiv:0712.0848v1.

[4] Bourgade, P., Nikeghbali, A. and Rouault, A. (2011). Ewens measures on compact groups and hypergeometric kernels. In *Séminaire de Probabilités XLIII. Lecture Notes in Math.* **2006** 351–377. Berlin: Springer. MR2790381 https://doi.org/10.1007/978-3-642-15217-7_15

[5] Byrne, S. and Girolami, M. (2013). Geodesic Monte Carlo on embedded manifolds. *Scand. J. Stat.* **40** 825–845. MR3145120 https://doi.org/10.1111/sjos.12036

[6] Carpenter, B., Gelman, A., Hoffman, M.D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P. and Riddell, A. (2017). Stan: A probabilistic programming language. *J. Stat. Softw.* **76** 1–32.

[7] Cayley, A. (1846). Sur quelques propriétés des déterminants gauches. *J. Reine Angew. Math.* **32** 119–123. MR1578518 https://doi.org/10.1515/crll.1846.32.119

[8] Chatterjee, S. and Meckes, E. (2008). Multivariate normal approximation using exchangeable pairs. *ALEA Lat. Am. J. Probab. Math. Stat.* **4** 257–283. MR2453473

[9] Chikuse, Y. (2003). *Statistics on Special Manifolds. Lecture Notes in Statist.* **174**. New York: Springer. MR1960435 https://doi.org/10.1007/978-0-387-21540-2

[10] Cook, R.D., Li, B. and Chiaromonte, F. (2010). Envelope models for parsimonious and efficient multivariate linear regression. *Statist. Sinica* **20** 927–960. MR2729839

[11] Corless, R.M., Gonnet, G.H., Hare, D.E.G., Jeffrey, D.J. and Knuth, D.E. (1996). On the Lambert *W* function. *Adv. Comput. Math.* **5** 329–359. MR1414285 https://doi.org/10.1007/BF02124750

[12] Cron, A. and West, M. (2016). Models of random sparse eigenmatrices and Bayesian analysis of multivariate structure. In *Statistical Analysis for High-Dimensional Data. Abel Symp.* **11** 125–153. Cham: Springer. MR3616267

[13] D'Aristotile, A., Diaconis, P. and Newman, C.M. (2003). Brownian motion and the classical groups. In *Probability, Statistics and Their Applications: Papers in Honor of Rabi Bhattacharya. Institute of Mathematical Statistics Lecture Notes—Monograph Series* **41** 97–116. Beachwood, OH: IMS. MR1999417 https://doi.org/10.1214/lnms/1215091660

[14] Diaconis, P. and Forrester, P.J. (2017). Hurwitz and the origins of random matrix theory in mathematics. *Random Matrices Theory Appl.* **6** Art. ID 1730001. MR3612265 https://doi.org/10.1142/S2010326317300017

[15] Diaconis, P. and Freedman, D. (1987). A dozen de Finetti-style results in search of a theory. *Ann. Inst. Henri Poincaré Probab. Stat.* **23** 397–423. MR0898502

[16] Diaconis, P., Holmes, S. and Shahshahani, M. (2013). Sampling from a manifold. In *Advances in Modern Statistical Theory and Applications: A Festschrift in Honor of Morris L. Eaton. Inst. Math. Stat.* (*IMS*) *Collect.* **10** 102–125. Beachwood, OH: IMS. MR3586941

[17] Diaconis, P., Seiler, C. and Holmes, S. (2014). Connections and extensions: A discussion of the paper by Girolami and Byrne. *Scand. J. Stat.* **41** 3–7. MR3181123 https://doi.org/10.1111/sjos.12070

[18] Diaconis, P. and Shahshahani, M. (1994). On the eigenvalues of random matrices. *J. Appl. Probab.* **31** 49–62. MR1274717 https://doi.org/10.2307/3214948

[19] Diaconis, P.W., Eaton, M.L. and Lauritzen, S.L. (1992). Finite de Finetti theorems in linear models and multivariate analysis. *Scand. J. Stat.* **19** 289–315. MR1211786

[20] Eaton, M.L. (1983). *Multivariate Statistics*: *A Vector Space Approach*. *Wiley Series in Probability and Mathematical Statistics*: *Probability and Mathematical Statistics*. New York: Wiley. MR0716321

[21] Eaton, M.L. (1989). *Group Invariance Applications in Statistics*. *NSF–CBMS Regional Conference Series in Probability and Statistics* **1**. Hayward, CA: IMS; Alexandria, VA: Amer. Statist. Assoc. MR1089423

[22] Edelman, A., Arias, T.A. and Smith, S.T. (1999). The geometry of algorithms with orthogonality constraints. *SIAM J. Matrix Anal. Appl.* **20** 303–353. MR1646856 https://doi.org/10.1137/S0895479895290954

[23] Grayson, M.A. (1989). A short note on the evolution of a surface by its mean curvature. *Duke Math. J.* **58** 555–558. MR1016434 https://doi.org/10.1215/S0012-7094-89-05825-0

[24] Hoff, P.D. (2007). Model averaging and dimension selection for the singular value decomposition. *J. Amer. Statist. Assoc.* **102** 674–685. MR2325118 https://doi.org/10.1198/016214506000001310

[25] Hoff, P.D. (2009). Simulation of the matrix Bingham–von Mises–Fisher distribution, with applications to multivariate and relational data. *J. Comput. Graph. Statist.* **18** 438–456. MR2749840 https://doi.org/10.1198/jcgs.2009.07177

[26] Hubbard, J.H. and Hubbard, B.B. (2009). *Vector Calculus*, *Linear Algebra*, *and Differential Forms*: *A Unified Approach*, 5th ed. Ithaca, NY: Matrix Editions.

[27] Hurwitz, A. (1897). Über die Erzeugung der invarianten durch integration. *Nachr. Akad. Wiss. Göttingen Math.-Phys. Kl. II* **1897** 71–90.

[28] James, A.T. (1954). Normal multivariate analysis and the orthogonal group. *Ann. Math. Stat.* **25** 40–75. MR0060779 https://doi.org/10.1214/aoms/1177728846

[29] Jauch, M., Hoff, P.D. and Dunson, D.B. (2019). Monte Carlo simulation on the Stiefel manifold via polar expansion. Available at arXiv:1906.07684.

[30] Jiang, T. (2006). How many entries of a typical orthogonal matrix can be approximated by independent normals? *Ann. Probab.* **34** 1497–1529. MR2257653 https://doi.org/10.1214/009117906000000205

[31] Jiang, T. and Ma, Y. (2019). Distances between random orthogonal matrices and independent normals. *Trans. Amer. Math. Soc.* **372** 1509–1553. MR3976569 https://doi.org/10.1090/tran/7470

[32] Johansson, K. (1997). On random matrices from the compact classical groups. *Ann. of Math.* (2) **145** 519–545. MR1454702 https://doi.org/10.2307/2951843

[33] Johnstone, I.M. (2001). On the distribution of the largest eigenvalue in principal components analysis. *Ann. Statist.* **29** 295–327. MR1863961 https://doi.org/10.1214/aos/1009210544

[34] Keating, J.P. and Snaith, N.C. (2000). Random matrix theory and $\zeta(1/2 + it)$. *Comm. Math. Phys.* **214** 57–89. MR1794265 https://doi.org/10.1007/s002200000261

[35] León, C.A., Massé, J.-C. and Rivest, L.-P. (2006). A statistical model for random rotations. *J. Multivariate Anal.* **97** 412–430. MR2234030 https://doi.org/10.1016/j.jmva.2005.03.009

[36] Magnus, J.R. (1988). *Linear Structures*. *Griffin's Statistical Monographs & Courses* **42**. London: Charles Griffin & Co., Ltd.; New York: The Clarendon Press. MR0947343

[37] Magnus, J.R. and Neudecker, H. (1979). The commutation matrix: Some properties and applications. *Ann. Statist.* **7** 381–394. MR0520247

[38] Magnus, J.R. and Neudecker, H. (1988). *Matrix Differential Calculus with Applications in Statistics and Econometrics*. *Wiley Series in Probability and Statistics*. Chichester: Wiley.

[39] Mardia, K.V. and Jupp, P.E. (2009). *Directional Statistics*. *Wiley Series in Probability and Statistics*. Chichester: Wiley.

[40] Maxwell, J.C. (1875). *Theory of Heat*, 4th ed. London: Longmans.

[41] Maxwell, J.C. (1878). On Boltzmann's theorem on the average distribution of energy in a system of material points. *Trans. Camb. Philos. Soc.* **12** 547–575.

[42] Meckes, E. (2008). Linear functions on the classical matrix groups. *Trans. Amer. Math. Soc.* **360** 5355–5366. MR2415077 https://doi.org/10.1090/S0002-9947-08-04444-9

[43] Mehler, F.G. (1866). Ueber die Entwicklung einer Function von beliebig vielen Variablen nach Laplaceschen Functionen höherer Ordnung. *J. Reine Angew. Math.* **66** 161–176. MR1579340 https://doi.org/10.1515/crll.1866.66.161

[44] Neal, R.M. (2011). MCMC using Hamiltonian dynamics. In *Handbook of Markov Chain Monte Carlo. Chapman & Hall/CRC Handb. Mod. Stat. Methods* 113–162. Boca Raton, FL: CRC Press. MR2858447

[45] Neudecker, H. (1983). On Jacobians of transformations with skew-symmetric, strictly (lower) triangular or diagonal matrix arguments. *Linear Multilinear Algebra* **14** 271–295. MR0718955 https://doi.org/10.1080/03081088308817563

[46] Pourzanjani, A.A., Jiang, R.M., Mitchell, B., Atzberger, P.J. and Petzold, L.R. (2017). General Bayesian inference over the Stiefel manifold via the givens transform. Available at arXiv:1710.09443v2.

[47] Rains, E.M. (1997). High powers of random elements of compact Lie groups. *Probab. Theory Related Fields* **107** 219–241. MR1431220 https://doi.org/10.1007/s004400050084

[48] Rao, V., Lin, L. and Dunson, D.B. (2016). Data augmentation for models based on rejection sampling. *Biometrika* **103** 319–335. MR3509889 https://doi.org/10.1093/biomet/asw005

[49] Salvatier, J., Wiecki, T.V. and Fonnesbeck, C. (2016). Probabilistic programming in Python using PyMC3. *PeerJ Comput. Sci.* **2** Art. ID e55.

[50] Shepard, R., Brozell, S.R. and Gidofalvi, G. (2015). The representation and parametrization of orthogonal matrices. *J. Phys. Chem. A* **119** 7924–7939. https://doi.org/10.1021/acs.jpca.5b02015

[51] Stam, A.J. (1982). Limit theorems for uniform distributions on spheres in high-dimensional Euclidean spaces. *J. Appl. Probab.* **19** 221–228. MR0644435 https://doi.org/10.2307/3213932

[52] Stein, C. (1995). The accuracy of the normal approximation to the distribution of the traces of powers of random orthogonal matrices. Technical Report 470 Dept. Statistics, Stanford Univ.

[53] Stewart, K. (2019). Total variation approximation of random orthogonal matrices by Gaussian matrices. *J. Theoret. Probab.* https://doi.org/10.1007/s10959-019-00900-5

[54] Traynor, T. (1994). Change of variable for Hausdorff measure (from the beginning). *Rend. Istit. Mat. Univ. Trieste* **26** 327–347. MR1408955

[55] van Handel, R. (2016). Probability in high dimension. Technical report, Princeton Univ.