Random Orthogonal Matrices with Applications in Statistics

by

Michael Jauch

Department of Statistical Science
Duke University

Date: ______________________
Approved:

____
Peter D. Hoff, Supervisor

____
David B. Dunson, Co-supervisor

____
Sayan Mukherjee

____
Galen Reeves

Dissertation submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in the Department of Statistical Science
in the Graduate School of Duke University
2019

Abstract

Random Orthogonal Matrices with Applications in Statistics

by

Michael Jauch

Department of Statistical Science
Duke University

Date: ______________________
Approved:

____
Peter D. Hoff, Supervisor

____
David B. Dunson, Co-supervisor

____
Sayan Mukherjee

____
Galen Reeves

An abstract of a dissertation submitted in partial fulfillment of the requirements for
the degree of Doctor of Philosophy in the Department of Statistical Science
in the Graduate School of Duke University
2019

Abstract

This dissertation focuses on random orthogonal matrices with applications in statistics. While Bayesian inference for statistical models with orthogonal matrix parameters is a recurring theme, several of the results on random orthogonal matrices may be of interest to those in the broader probability and random matrix theory communities. In Chapter 2, we parametrize the Stiefel and Grassmann manifolds, represented as subsets of orthogonal matrices, in terms of Euclidean parameters using the Cayley transform and then derive Jacobian terms for change of variables formulas. This allows for Markov chain Monte Carlo simulation from probability distributions defined on the Stiefel and Grassmann manifolds. We also establish an asymptotic independent normal approximation for the distribution of the Euclidean parameters corresponding to the uniform distribution on the Stiefel manifold. In Chapter 3, we present *polar expansion*, a general approach to Monte Carlo simulation from probability distributions on the Stiefel manifold. When combined with modern Markov chain Monte Carlo software, polar expansion allows for routine and flexible posterior inference in models with orthogonal matrix parameters. Chapter 4 addresses prior distributions for structured orthogonal matrices. We introduce an approach to constructing prior distributions for structured orthogonal matrices which leads to tractable posterior simulation via polar expansion. We state two main results which support our approach and offer a new perspective on approximating the entries of random orthogonal matrices.

To my family

Contents

List of Tables

List of Figures

1

Introduction

This dissertation focuses on random orthogonal matrices with applications in statistics. While Bayesian inference for statistical models with orthogonal matrix parameters is a recurring theme, several of the results on random orthogonal matrices may be of interest to those in the broader probability and random matrix theory communities.

In Chapter 2, 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 describe and illustrate through examples a Markov chain Monte Carlo approach to simulating from distributions on the Stiefel and Grassmann manifolds. Finally, we establish an asymptotic independent normal approximation for the distribution of the Euclidean parameters corresponding to the uniform distribution on the Stiefel manifold. This result contributes to the growing literature on normal approximations to the entries of random orthogonal matrices or transformations thereof.

Chapter 3 introduces *polar expansion*, a general approach to Monte Carlo simulation from probability distributions on the Stiefel manifold. To bypass many of the well-established challenges of simulating from the distribution of a random orthogonal matrix $\boldsymbol{Q}$, we construct a distribution for an unconstrained random matrix $\boldsymbol{X}$ such that $\boldsymbol{Q}_X$, the orthogonal component of the polar decomposition of $\boldsymbol{X}$, is equal in distribution to $\boldsymbol{Q}$. The distribution of $\boldsymbol{X}$ is amenable to MCMC simulation using standard methods, and an approximation to the distribution of $\boldsymbol{Q}$ can be recovered from a Markov chain on the unconstrained space. When combined with modern MCMC software, polar expansion allows for routine and flexible posterior inference in models with orthogonal matrix parameters. We find that polar expansion with adaptive Hamiltonian Monte Carlo is an order of magnitude more efficient than competing MCMC approaches in a benchmark protein interaction network application. We also propose a new approach to Bayesian functional principal components analysis which we illustrate in a meteorological time series application.

Chapter 4 addresses prior distributions for structured orthogonal matrices. Structural assumptions regarding unknown parameters play a critical role in statistical theory and practice. In Bayesian statistics, these structural assumptions are commonly reflected in the prior distribution. While there is a substantial literature on prior distributions for real-valued vectors, matrices, or arrays whose entries are sparse or dependent, there has been little work on analogous priors for orthogonal matrices. We introduce an approach to constructing prior distributions for structured orthogonal matrices which leads to tractable posterior simulation via polar expansion. In particular, we consider prior distributions which are the $\boldsymbol{Q}_X$-margin of the distribution of a real random matrix $\boldsymbol{X}$. We state two main results showing that features of the distribution of $\boldsymbol{X}$ are inherited by its $\boldsymbol{Q}_X$-margin. Thus, if we want the prior distribution for an orthogonal matrix parameter to reflect structural assumptions such as sparsity or row dependence or to satisfy an invariance property, we can build these features into the distribution of $\boldsymbol{X}$. Beyond its relevance to statistical modeling, the second of our main results offers a new perspective on approximating the entries of random orthogonal matrices, a topic of significant interest in random matrix theory.

Chapter 2 can be read independently of Chapters 3 and 4. Chapter 4 assumes the reader is familiar with the content of Chapter 3.

2

Random orthogonal matrices and the Cayley transform

2.1 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 $d_{\mathcal{G}} = (p-k)k$. Typically, points in the Grassmann manifold are 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 $\mathcal{O}(k)$. In Section 2.3.2, we elaborate and expand upon the contributions of Shepard et al. (2015) 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 chapter, 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 $\boldsymbol{Q} \sim P_{\mathcal{V}(k,p)}$, then $\boldsymbol{U}\boldsymbol{Q}\boldsymbol{V} \overset{\text{dist.}}{=} \boldsymbol{Q}$ for all $\boldsymbol{U} \in \mathcal{O}(p)$ and $\boldsymbol{V} \in \mathcal{O}(k)$. 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 push-forward 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, as we discuss in Section 2.6, and in statistics, where they appear in foundational work on multivariate theory (James, 1954).

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 lying on compact Riemannian manifolds in the field of directional statistics (Mardia and Jupp, 2009). Also, statistical models having the Stiefel or Grassmann manifold as their parameter space are increasingly common (Hoff, 2007, 2009b; Cook et al., 2010). 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 (Hoff, 2009b; Byrne and Girolami, 2013; Rao et al., 2016).

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 Forrester

(2017) identify as the earliest substantial mathematical contribution to modern random matrix theory, Hurwitz (1897) 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. Anderson et al. (1987) 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 (León et al., 2006; Cron and West, 2016).

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. (2015) 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 chapter, 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. (2015) 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 describe and illustrate through examples an approach to MCMC simulation from distributions on these sets. Finally, we establish an asymptotic independent normal approximation for the distribution of the Euclidean parameters which corresponds to the uniform distribution on the Stiefel manifold. 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 chapter and to simulate from distributions on $\mathcal{V}(k,p)$ and $\mathcal{V}^+(k,p)$ is available at `https://github.com/michaeljauch/cayley`.

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

In this section, we introduce tools for defining and manipulating probability distributions on a 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. (2013) 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 $\mathbb{R}^m$. As in Diaconis et al. (2013) and Byrne and Girolami (2013), we take the Hausdorff measure as our reference measure. 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

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

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

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

and $\alpha_m = \Gamma(\frac{1}{2})^m / \Gamma(\frac{m}{2} + 1)$, the volume of the unit ball in $\mathbb{R}^n$. 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 $H^m(\mathcal{M} \backslash \mathcal{I}) = 0$;

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. Define the m-dimensional Jacobian of f at $\boldsymbol{\phi}$ as $J_m f(\boldsymbol{\phi}) = \left| Df(\boldsymbol{\phi})^T Df(\boldsymbol{\phi}) \right|^{1/2}$. 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 Diaconis et al. (2013). The change of variables formula is given in the following theorem, which is essentially a restatement of the main result of Traynor (1993):

Theorem 1. *For all Borel subsets $A \subset \mathcal{D}$*

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

and hence

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

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 $\boldsymbol{y} \overset{\text{dist.}}{=} f(\boldsymbol{\phi})$ when the distribution of $\boldsymbol{\phi} \in \mathcal{D}$ has a density proportional to $g[f(\boldsymbol{\phi})]J_m f(\boldsymbol{\phi})$.

2.3 The Cayley parametrizations

The Cayley transform, as introduced in Cayley (1846), is a map from skew-symmetric matrices to special orthogonal matrices. Given $\boldsymbol{X}$ in $\text{Skew}(p) = \{\boldsymbol{X} \in \mathbb{R}^{p\times p} | \boldsymbol{X} = -\boldsymbol{X}^T\}$, the (original) Cayley transform of $\boldsymbol{X}$ is the special orthogonal matrix

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

We work instead with a modified version of the Cayley transform described in Sheppard et al. (2015). In this version, the Cayley transform of $\boldsymbol{X}$ is the $p \times k$ orthogonal matrix

$$C(\boldsymbol{X}) = (\boldsymbol{I}_p + \boldsymbol{X})(\boldsymbol{I}_p - \boldsymbol{X})^{-1}\boldsymbol{I}_{p\times k}$$

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

In this section, we parametrize (in the sense of Section 2.2) 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 2.4. 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 $\mathcal{V}^+(k,p)$.

2.3.1 Cayley parametrization of the Stiefel manifold

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

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

Let $\boldsymbol{X} \in \mathcal{D}_{\mathcal{V}}$ and set $\boldsymbol{Q} = C(\boldsymbol{X})$. Partition $\boldsymbol{Q} = [\boldsymbol{Q}_1^T \quad \boldsymbol{Q}_2^T]^T$ so that $\boldsymbol{Q}_1$ is square. We can write the blocks of $\boldsymbol{Q}$ in terms of the blocks of $\boldsymbol{X}$ as

$$\boldsymbol{Q}_1 = (\boldsymbol{I}_k - \boldsymbol{A}^T\boldsymbol{A} + \boldsymbol{B})(\boldsymbol{I}_k + \boldsymbol{A}^T\boldsymbol{A} - \boldsymbol{B})^{-1}$$

$$\boldsymbol{Q}_2 = 2\boldsymbol{A}(\boldsymbol{I}_k + \boldsymbol{A}^T\boldsymbol{A} - \boldsymbol{B})^{-1}.$$

The matrix $\boldsymbol{I}_k + \boldsymbol{A}^T\boldsymbol{A} - \boldsymbol{B}$ 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 $\boldsymbol{X} \in \mathcal{D}_{\mathcal{V}}$. We can recover the matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ from $\boldsymbol{Q}$:

$$\boldsymbol{F} = (\boldsymbol{I}_k - \boldsymbol{Q}_1)(\boldsymbol{I}_k + \boldsymbol{Q}_1)^{-1} \tag{2.1}$$

$$\boldsymbol{B} = \frac{1}{2}(\boldsymbol{F}^T - \boldsymbol{F}) \tag{2.2}$$

$$\boldsymbol{A} = \frac{1}{2}\boldsymbol{Q}_2(\boldsymbol{I}_k + \boldsymbol{F}). \tag{2.3}$$

We are now in a position to verify the first two conditions of Section 2.2. The image of $\mathcal{D}_V$ under C is the set

$$\mathcal{I}_{\mathcal{V}} = \{\boldsymbol{Q} = [\boldsymbol{Q}_1^T \quad \boldsymbol{Q}_2^T]^T \in \mathcal{V}(k,p) \mid \boldsymbol{I}_k + \boldsymbol{Q}_1 \text{ is nonsingular}\}$$

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 2.1 - 2.3. 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 $\boldsymbol{X}$. Let $\boldsymbol{b}$ be the $k(k-1)/2$-dimensional vector of independent entries of $\boldsymbol{B}$ obtained by eliminating diagonal and supradiagonal elements from the vectorization $\operatorname{vec}\boldsymbol{B}$. The vector $\boldsymbol{\varphi} = (\boldsymbol{b}^T, \operatorname{vec}\boldsymbol{A}^T)^T$ then contains each of the independent entries of $\boldsymbol{X}$. Let $\boldsymbol{X}_{\boldsymbol{\varphi}} \in \mathcal{D}_{\mathcal{V}}$ 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 $\boldsymbol{\varphi}$:

$$C(\boldsymbol{\varphi}) = (\boldsymbol{I}_p + \boldsymbol{X}_{\boldsymbol{\varphi}})(\boldsymbol{I}_p - \boldsymbol{X}_{\boldsymbol{\varphi}})^{-1}\boldsymbol{I}_{p\times k}.$$

As a function of $\boldsymbol{\varphi}$, the Cayley transform is a bijection between the set $\mathcal{D}_{\mathcal{V}}^{\boldsymbol{\varphi}} = \{\boldsymbol{\varphi} \in \mathbb{R}^{d_{\mathcal{V}}} : \boldsymbol{X}_{\boldsymbol{\varphi}} \in \mathcal{D}_{\mathcal{V}}\} = \mathbb{R}^{d_{\mathcal{V}}}$ and $\mathcal{I}_{\mathcal{V}}$. The inverse Cayley transform is defined in the obvious way as the map $C^{-1} : \boldsymbol{Q} \mapsto (\boldsymbol{b}^T, \operatorname{vec}\boldsymbol{A}^T)^T$ where $\boldsymbol{B}$ and $\boldsymbol{A}$ are computed from $\boldsymbol{Q}$ according to equations 2.1-2.3 and $\boldsymbol{b}$ contains the independent entries of $\boldsymbol{B}$ as before.

The next lemma provides an explicit linear map $\boldsymbol{\Gamma}_{\mathcal{V}}$ from $\boldsymbol{\varphi}$ to $\operatorname{vec}\boldsymbol{X}_{\boldsymbol{\varphi}}$, 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 $\widetilde{\boldsymbol{D}}_k\boldsymbol{b} = \operatorname{vec}\boldsymbol{B}$, both of which are discussed in Magnus (1988) and defined explicitly in Appendix A.2 . Set $\boldsymbol{\Theta}_1 = [\boldsymbol{I}_k \quad \boldsymbol{0}_{k\times p-k}]$ and $\boldsymbol{\Theta}_2 = [\boldsymbol{0}_{p-k\times k} \quad \boldsymbol{I}_{p-k}]$.

Lemma 2. *The equation* $\operatorname{vec} X_{\boldsymbol{\varphi}} = \boldsymbol{\Gamma}_{\mathcal{V}}\, \boldsymbol{\varphi}$ *is satisfied by the matrix*

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

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

Proposition 3. *The Cayley transform is continuously differentiable on* $\mathcal{D}_{\mathcal{V}}^{\boldsymbol{\varphi}} = \mathbb{R}^{d_{\mathcal{V}}}$ *with derivative matrix*

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

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 $\boldsymbol{X}_{\varphi}$, while the matrix $\boldsymbol{\Gamma}_{\mathcal{V}}$ arises from differentiating $\boldsymbol{X}_{\varphi}$ with respect to $\boldsymbol{\varphi}$.

2.3.2 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 Chikuse (2003). We propose instead to represent $\mathcal{G}(k,p)$ by the subset of orthogonal matrices

$$\mathcal{V}^{+}(k,p) = \left\{\boldsymbol{Q} = [\boldsymbol{Q}_1^T\, \boldsymbol{Q}_2^T]^T \in \mathcal{V}(k,p) \,\middle|\, \boldsymbol{Q}_1 > \boldsymbol{0}\right\}.$$

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 4. *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

$$\mathcal{D}_{\mathcal{G}} = \left\{ \boldsymbol{X} = \begin{bmatrix} \boldsymbol{0} & -\boldsymbol{A}^T \\ \boldsymbol{A} & \boldsymbol{0}_{p-k} \end{bmatrix} \,\middle|\, \begin{array}{l} \boldsymbol{A} \in \mathbb{R}^{p-k \times k}, \\ \mathrm{eval}_i(\boldsymbol{A}^T\boldsymbol{A}) \in [0,1) \text{ for } 1 \leqslant i \leqslant k \end{array} \right\} \subset \mathrm{Skew}(p)$$

where the notation $\mathrm{eval}_i(\boldsymbol{A}^T\boldsymbol{A})$ indicates the ith eigenvalue of the matrix $\boldsymbol{A}^T\boldsymbol{A}$. Let $\boldsymbol{X} \in \mathcal{D}_{\mathcal{G}}$ and set $\boldsymbol{Q} = C(\boldsymbol{X})$. Again, partition $\boldsymbol{Q} = [\boldsymbol{Q}_1^T \quad \boldsymbol{Q}_2^T]^T$ so that $\boldsymbol{Q}_1$ is square. We can write the blocks of $\boldsymbol{Q}$ in terms of $\boldsymbol{A}$ as

$$\boldsymbol{Q}_1 = (\boldsymbol{I}_k - \boldsymbol{A}^T\boldsymbol{A})(\boldsymbol{I}_k + \boldsymbol{A}^T\boldsymbol{A})^{-1}$$

$$\boldsymbol{Q}_2 = 2\boldsymbol{A}(\boldsymbol{I}_k + \boldsymbol{A}^T\boldsymbol{A})^{-1}.$$

We can recover the matrix $\boldsymbol{A}$ from $\boldsymbol{Q}$:

$$\boldsymbol{F} = (\boldsymbol{I}_k - \boldsymbol{Q}_1)(\boldsymbol{I}_k + \boldsymbol{Q}_1)^{-1} \tag{2.4}$$

$$\boldsymbol{A} = \frac{1}{2}\boldsymbol{Q}_2(\boldsymbol{I}_k + \boldsymbol{F}). \tag{2.5}$$

We turn to verifying the conditions of Section 2.2. The first two conditions are satisfied as a consequence of the following proposition:

Proposition 5. *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 $\boldsymbol{\psi} = \operatorname{vec} \boldsymbol{A}$ containing each of the independent entries of $\boldsymbol{X}$. Let $\boldsymbol{X}_{\boldsymbol{\psi}} \in \mathcal{D}_{\mathcal{G}}$ be the matrix having $\boldsymbol{\psi}$ as its corresponding vector of independent entries. The Cayley transform can now be thought of as a function of $\boldsymbol{\psi}$:

$$C(\boldsymbol{\psi}) = (\boldsymbol{I}_p + \boldsymbol{X}_{\boldsymbol{\psi}})(\boldsymbol{I}_p - \boldsymbol{X}_{\boldsymbol{\psi}})^{-1}\boldsymbol{I}_{p\times k}.$$

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

Lemma 6. *The equation $\operatorname{vec} X_{\boldsymbol{\psi}} = \boldsymbol{\Gamma}_{\mathcal{G}}\boldsymbol{\psi}$ is satisfied by the matrix*

$$\boldsymbol{\Gamma}_{\mathcal{G}} = (\boldsymbol{I}_{p^2} - \boldsymbol{K}_{p,p})(\boldsymbol{\Theta}_1^T \otimes \boldsymbol{\Theta}_2^T).$$

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

Proposition 7. *The Cayley transform is continuously differentiable on $\mathcal{D}_{\mathcal{G}}^{\boldsymbol{\psi}}$ with derivative matrix*

$$DC(\boldsymbol{\psi}) = 2\left[\boldsymbol{I}_{p\times k}^T(\boldsymbol{I}_p - \boldsymbol{X}_{\boldsymbol{\psi}})^{-T} \otimes (\boldsymbol{I}_p - \boldsymbol{X}_{\boldsymbol{\psi}})^{-1}\right]\boldsymbol{\Gamma}_{\mathcal{G}}.$$

2.4 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 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

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

where

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

and $\boldsymbol{\Gamma}_{\mathcal{V}}$ is defined as in Section 2.3.1. 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 8. *(Change of Variables Formulas) For all Borel sets* $A \subset \mathcal{D}_{\mathcal{V}}^{\boldsymbol{\varphi}}$

$$\int_A J_{d_{\mathcal{V}}}C(\boldsymbol{\varphi}) L^{d_{\mathcal{V}}}(d\boldsymbol{\varphi}) = H^{d_{\mathcal{V}}}[C(A)] \tag{2.6}$$

and hence

$$\int_A g\left[C(\boldsymbol{\varphi})\right] J_{d_{\mathcal{V}}}C(\boldsymbol{\varphi}) L^{d_{\mathcal{V}}}(d\boldsymbol{\varphi}) = \int_{C(A)} g(\boldsymbol{Q}) H^{d_{\mathcal{V}}}(d\boldsymbol{Q}). \tag{2.7}$$

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 1, Theorem 8 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 $\boldsymbol{Q} \overset{\text{dist.}}{=} C(\boldsymbol{\varphi})$ when the distribution of $\boldsymbol{\varphi} \in \mathcal{D}_{\mathcal{V}}^{\boldsymbol{\varphi}} = \mathbb{R}^{d_{\mathcal{V}}}$ has a density proportional to $g\left[C(\boldsymbol{\varphi})\right] J_{d_{\mathcal{V}}} C(\boldsymbol{\varphi})$. In particular, let $g \propto 1$ so that $\boldsymbol{Q} \sim P_{\mathcal{V}(k,p)}$. Then $\boldsymbol{Q} \overset{\text{dist.}}{=} C(\boldsymbol{\varphi})$ 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 $\mathcal{V}^{+}(k, p.)$

2.5 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 orthogonal 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 2.3. The recent work of Pourzanjani et al. (2017) 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 (2009b), 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 (2013). 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 $\boldsymbol{Q}$ 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 $t = 0, ..., T$:

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 $u \sim \text{Unif}(0,1)$. If $u \leqslant r$, 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) (Neal, 2011), with implementations in software such as Carpenter et al. (2017) and Salvatier et al. (2016), is a natural choice. For settings in which evaluation and automatic differentiation of the Jacobian term are not prohibitively expensive, the Cayley transform approach offers a relatively straightforward path to MCMC simulation on the sets $\mathcal{V}(k,p)$ and $\mathcal{V}^+(k,p)$.

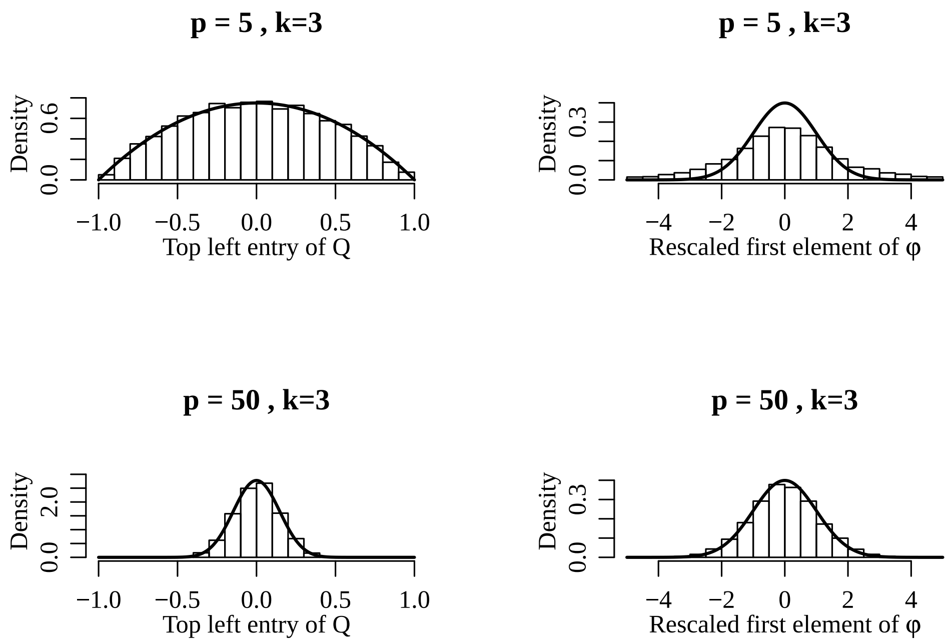

FIGURE 2.1: The top left panel is a histogram of the top left entry of simulated values of $\boldsymbol{Q} \sim P_{\mathcal{V}(3,5)}$ plotted against the exact density. The bottom left panel is the analogous plot for $\boldsymbol{Q} \sim P_{\mathcal{V}(3,50)}$. The top right panel is a histogram of the first entry, rescaled by $\sqrt{p/2}$, of $\boldsymbol{\varphi} = C^{-1}(\boldsymbol{Q})$ 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 $k = 3$.

2.5.1 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 (Carpenter et al., 2017) to simulate $\boldsymbol{Q} \sim P_{\mathcal{V}(k,p)}$. Of 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 2.1 provides plots based on $10,000$ simulated values of $\boldsymbol{Q} \sim P_{\mathcal{V}(k,p)}$. 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 $\boldsymbol{Q}$ plotted against the exact density, given in Proposition 7.3 of Eaton (1989). 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 $\boldsymbol{\varphi} = C^{-1}(Q)$. These are plotted against a standard normal density. Theorem 9 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.

2.5.2 Example: Bayesian inference for the spiked covariance model

Suppose the rows of an $n \times p$ data matrix $\boldsymbol{Y}$ are independent samples from a mean zero multivariate normal population with covariance matrix $\boldsymbol{\Sigma}$. The spiked covariance model, considered by Johnstone (2001) and others, assumes the covariance matrix $\boldsymbol{\Sigma}$ can be decomposed as $\boldsymbol{\Sigma} = \sigma^2 \left(\boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^T + \boldsymbol{I}_p\right)$ with $\boldsymbol{Q} \in \mathcal{V}(k, p)$ and $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, ..., \lambda_k)$ where $\lambda_1 > ... > \lambda_k > 0$. 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 (Hoff, 2009b), having density

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

We compare two MCMC approaches to simulating from the matrix Bingham distribution: the Gibbs sampling method of Hoff (2009b) and the Cayley transform

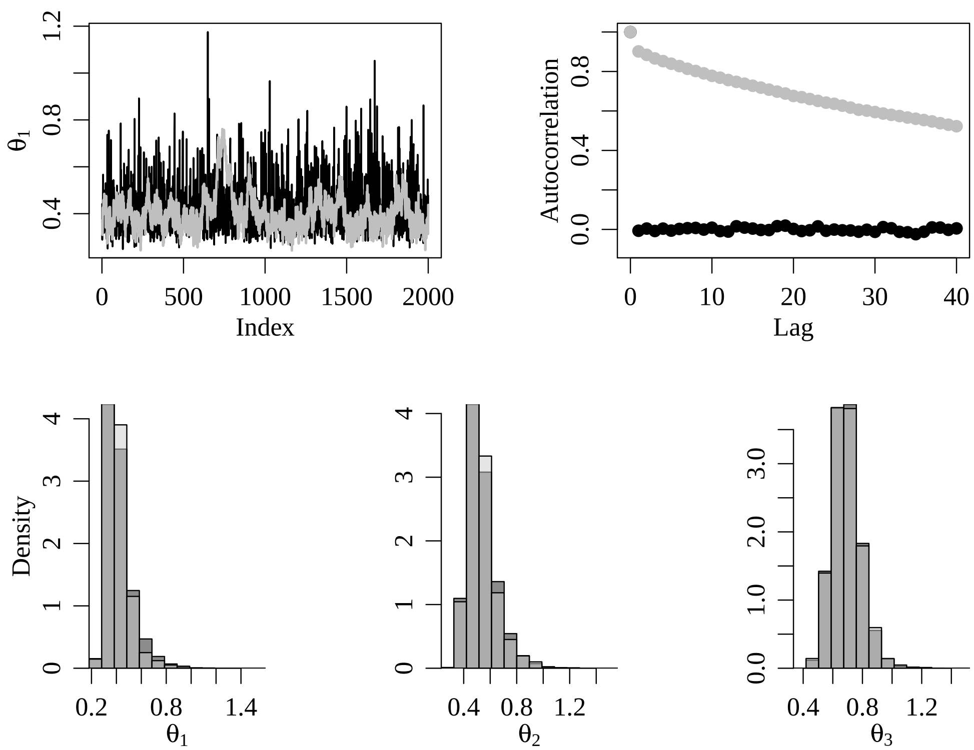

FIGURE 2.2: The plot in the top left overlays trace plots of the first principal angle calculated from a portion of each of the chains, while the plot in the top right shows the correlation between lagged values of the first principal angle. The black lines and dots correspond to our MCMC approach, while the gray lines and dots correspond to that of Hoff (2009b). The plots in the bottom half compare histogram approximations of the marginal posterior distributions of the principal angles.

approach (again, with HMC as implemented in Stan). The dimensions are chosen as $n = 100$, $p = 50$, and $k = 3$. We set $\sigma^2 = 1$, $\boldsymbol{\Lambda} = \text{diag}(5, 3, 1.5)$, and choose a true value of $\boldsymbol{Q}$ uniformly from $\mathcal{V}(3, 50)$. We then generate a data matrix according to the model $\text{vec}\,\boldsymbol{Y} \sim \mathcal{N}\left[\mathbf{0}, \sigma^2\left(\boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^T + \boldsymbol{I}_p\right) \otimes \boldsymbol{I}_{100}\right]$. 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 $j = 1, ..., 3$, 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 2.2 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 (2009b). 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 (2009b). Together, the plots in the top half of the figure indicate that, compared to the approach of Hoff (2009b), the Cayley transform approach produces a Markov chain with less autocorrelation, reducing Monte Carlo error in the resulting posterior inferences.

2.6 An asymptotic independent normal approximation

In Section 2.4, 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 term $J_{d_{\mathcal{V}}} C(\boldsymbol{\varphi})$. 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 9 of this section provides a mathematical explanation for this empirical phenomenon.

In order to understand Theorem 9 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 (Mehler, 1866), Maxwell (Maxwell, 1875, 1878), and Borel (Borel, 1906). A theorem of Borel shows that $\Pr(\sqrt{p}q_p \leqslant x) \to \Phi(x)$ 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 Diaconis and Freedman (1987), while an overview is given in D'Aristotile et al. (2003) and Jiang (2006).

Much of this literature is devoted to normal approximations to the joint distribution of entries of random orthogonal matrices. Letting $\boldsymbol{q}_p$ be the first column of $\boldsymbol{Q}_p$ for each p (again, any other column would also work), Stam (1982) proved that the total variation distance between the distribution of the first m_p coordinates of $\sqrt{p}\boldsymbol{q}_p$ 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 (1987) strengthened this result, showing that it holds for $m_p = o(p)$. Diaconis et al. (1992) prove that the total variation distance between the distribution of the top left $m_p \times n_p$ block of $\sqrt{p}\boldsymbol{Q}_p$ 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 $n_p \leqslant k_p$ 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 (2006) solved this problem, finding the largest orders to be $o(p^{1/2})$. Recent work has further explored this topic (Stewart, 2018; Jiang and Ma, 2017).

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. (2003) and Meckes (2008) study the convergence of linear combinations of the entries of the matrices in the sequence $\{\boldsymbol{Q}_p\}$ to normality as $p \to \infty$. Diaconis and Shahshahani (1994); Stein (1995); Johansson (1997), and Rains (1997) addresses the normality of traces of powers of random orthogonal and unitary matrices. In the latter category, Chatterjee and Meckes (2008) and Jiang and Ma (2017) consider probability metrics other than total variation distance. Jiang (2006) also considers a notion of approximation other than total variation distance, and Theorem 3 in that work is particularly important in understanding our Theorem 9.

Theorem 3 of Jiang (2006) 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 (2006) shows that one can construct a sequence of pairs of random matrices $\{\boldsymbol{Z}_p, \boldsymbol{Q}_p\}$ with each pair defined on the same probability space such that

(i) The entries of the $p \times k_p$ matrix $\boldsymbol{Z}_p$ are independent standard normals;

(ii) The matrix $\boldsymbol{Q}_p$ is uniform on $\mathcal{V}(k_p, p)$;

(iii) The quantity $\epsilon_p = \|\sqrt{p}\boldsymbol{Q}_p - \boldsymbol{Z}_p\|_{\max} \to 0$ in probability as p grows 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 $\boldsymbol{Q}_p$ be the result of the Gram-Schmidt orthogonalization procedure applied to $\boldsymbol{Z}_p$.

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 $\sqrt{p}\boldsymbol{Q}_p$ on a computer with finite precision. One could simulate a matrix $\boldsymbol{Z}_p$ of independent standard normals, obtain $\boldsymbol{Q}_p$ using Gram-Schmidt, then multiply by $\sqrt{p}$ to arrive at $\sqrt{p}\boldsymbol{Q}_p$. However, for a fixed machine precision and p sufficiently large (again, both in an absolute sense and relative to k_p), the matrix $\sqrt{p}\boldsymbol{Q}_p$ would be indistinguishable from $\boldsymbol{Z}_p$ with high probability.

Our Theorem 9 establishes that the distribution of $\boldsymbol{\varphi}_p = C_p^{-1}(\boldsymbol{Q}_p)$, which we know to have a density proportional to $J_{d_\mathcal{V}} C_p(\boldsymbol{\varphi})$, 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

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

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

Theorem 9. *One can construct a sequence of pairs of random vectors $\{\boldsymbol{z}_p, \boldsymbol{\varphi}_p\}$ such that*

(i) The entries of the vector $\boldsymbol{z}_p$ are independent standard normals;

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

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

The construction of the coupling is more elaborate than in Theorem 3 of Jiang (2006). 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 $\operatorname{vec} \widetilde{B}_p(\boldsymbol{M})$ while $\widetilde{A}_p(\boldsymbol{M}) = \boldsymbol{M}_2$. The approximate inverse Cayley transform is then

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

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

$$\begin{aligned} \boldsymbol{z}_p &= \boldsymbol{\Pi}_p \, \widetilde{C}_p^{-1}(p^{-1/2} \boldsymbol{Z}_p) \\ \boldsymbol{\varphi}_p &= C_p^{-1}(\boldsymbol{Q}_p). \end{aligned}$$

The details of the proof of Theorem 9 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 10. *(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* $k_p = o\left(\frac{p^{1/4}}{\sqrt{\log p}}\right)$, *and (ii) the quantity* $\|\boldsymbol{\Pi}_p \widetilde{\boldsymbol{\varphi}}_p - \boldsymbol{z}_p\|_\infty \to 0$ *in probability as* $p \to \infty$ *provided* $k_p = o\left(\frac{p}{\log p}\right)$.

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

3

Monte Carlo simulation on the Stiefel manifold via polar expansion

3.1 Introduction

Probability distributions on the Stiefel manifold, the set of orthogonal matrices $\mathcal{V}(k,p) = \left\{\boldsymbol{Q} \in \mathbb{R}^{p\times k} \,|\, \boldsymbol{Q}^\top\boldsymbol{Q} = \boldsymbol{I}_k\right\}$ with $p \geqslant k$, play a number of roles throughout statistics. The uniform distribution on the Stiefel manifold appears in foundational work on multivariate theory (James, 1954), while non-uniform distributions on $\mathcal{V}(k,p)$ arise in modern statistical applications. Distributions on the Stiefel manifold model directions, axes, planes, and rotations in the field of directional statistics (Mardia and Jupp, 2009). They also represent prior or posterior distributions in Bayesian analyses of models with orthogonal matrix parameters. In this chapter, we are primarily motivated by applications in Bayesian statistics, but the discussion is relevant more broadly.

Statistical models for multivariate data are often naturally parametrized by a set of orthogonal matrices. Parametrization in terms of orthogonal matrices is common in low-rank matrix or tensor estimation, dimension reduction, and covariance modeling. For example, we might model an $n \times p$ data matrix as $\boldsymbol{Y} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top + \sigma\boldsymbol{E}$, where $\boldsymbol{U} \in \mathcal{V}(k, n)$, $\boldsymbol{V} \in \mathcal{V}(k, p)$, $\boldsymbol{D}$ is a $k \times k$ diagonal matrix with positive entries on the diagonal, $\boldsymbol{E}$ is a matrix of errors, and $\sigma > 0$. This model and variants are important in matrix denoising problems (Donoho and Gavish, 2014) and model-based principal component analysis (PCA) (Hoff, 2009b).

Bayesian analyses of models with orthogonal matrix parameters are increasingly common but raise computational challenges. In modern Bayesian statistics, analytic calculation of posterior expectations or exact Monte Carlo simulation from the posterior is typically infeasible. Instead, one constructs a Markov chain whose stationary distribution is the posterior using Markov chain Monte Carlo (MCMC) methods. For models with orthogonal matrix parameters, this Markov chain must lie on the Stiefel manifold. However, the constraints which define the manifold complicate MCMC simulation to the extent that Bayesian analyses of models with orthogonal matrix parameters are often prohibitively difficult.

A number of authors have addressed simulation from distributions on the Stiefel manifold, but there remains a need for more routine and flexible methodology for posterior simulation in models with orthogonal matrix parameters. In the directional statistics literature, which focuses on exact Monte Carlo simulation in a low dimensional setting, rejection sampling is common. See, for example, Kent et al. (2013). These rejection sampling approaches are not well suited for routine and flexible posterior simulation, as they must be tailored to particular distributions, and acceptance rates can decrease rapidly with increasing dimension or concentration of the target distribution. Hoff (2009b) proposes a Gibbs sampler for the Bingham-von Mises-Fisher family of distributions on the Stiefel manifold and applies it to posterior simulation for the network eigenmodel discussed in Section 3.5.1. The Gibbs sampler of Hoff (2009b) can be a practical option for posterior simulation but is applicable only when the conditional posterior distributions belong to the designated family.

As we will see in Section 3.5.1, Gibbs sampling can also produce Markov chains with high autocorrelation. Furthermore, simulation from conditional distributions is performed via rejection sampling, and acceptance rates can be vanishingly small, as described in Brubaker et al. (2012). Byrne and Girolami (2013) introduce geodesic Monte Carlo (GMC), an elegant and well-motivated algorithm extending Hamiltonian Monte Carlo (HMC) (Neal, 2011) to distributions defined on the Stiefel manifold and other manifolds embedded in Euclidean spaces. However, without methodology for adaptive tuning parameter selection or a robust software implementation, GMC does not yet offer routine and flexible posterior simulation. Jauch et al. (2018) and Pourzanjani et al. (2017) reparametrize the Stiefel manifold in terms of unconstrained Euclidean parameters, derive the Jacobian term required to map the target distribution from the Stiefel manifold to Euclidean space, then leverage MCMC software to simulate from the transformed distribution. The core idea of recasting a constrained simulation problem as an easier unconstrained problem is compelling, but the cost of computing the Jacobian term (in Jauch et al. (2018)) and the pathologies introduced in mapping between topologically distinct spaces are drawbacks of these reparametrization approaches.

In this chapter, we present *polar expansion*, a general approach to Monte Carlo simulation from probability distributions on the Stiefel manifold. To bypass many of the well-established challenges of simulating from the distribution of a random orthogonal matrix $\boldsymbol{Q} \in \mathcal{V}(k,p)$, we construct a distribution for an unconstrained random matrix $\boldsymbol{X} \in \mathbb{R}^{p\times k}$ such that $\boldsymbol{Q}_X$, the orthogonal component of the polar decomposition, is equal in distribution to $\boldsymbol{Q}$. The distribution of $\boldsymbol{X}$ is amenable to Markov chain Monte Carlo simulation using standard methods, and an approximation to the distribution of $\boldsymbol{Q}$ can be recovered from a Markov chain on the unconstrained space. When combined with modern MCMC software, polar expansion allows for routine and flexible posterior inference in models with orthogonal matrix parameters. Polar expansion can be seen as a generalization of the method for simulating from the unit sphere $\mathcal{V}(1, p)$ built into Stan at the time of writing (Stan Development Team, 2019).

We provide an outline of what follows. In Section 3.2, we present polar expansion in detail. In Section 3.3, we build intuition through simple examples in which exact Monte Carlo simulation is possible. That discussion serves as a prelude for Section 3.4, which addresses polar expansion and MCMC simulation in more complex settings, including posterior simulation for models with orthogonal matrix parameters. In Section 3.5, we illustrate the practical importance of polar expansion in applications. We find that polar expansion with adaptive HMC is an order of magnitude more efficient than competing MCMC approaches in a benchmark protein interaction network application. We also propose a new approach to Bayesian functional principal components analysis which we illustrate in a meteorological time series application. We conclude with a brief discussion in Section 3.6. Code to reproduce the figures and analyses in this chapter is available at `https://github.com/michaeljauch/polar`.

3.2 Polar expansion via change of variables

The polar decomposition is the unique representation of a full rank matrix $\boldsymbol{X} \in \mathbb{R}^{p \times k}$ as the product $\boldsymbol{X} = \boldsymbol{Q}_X \boldsymbol{S}_X^{1/2}$ where $\boldsymbol{Q}_X \in \mathcal{V}(k, p)$, $\boldsymbol{S}_X$ is a $k \times k$ symmetric positive definite (SPD) matrix, and $\boldsymbol{S}_X^{1/2}$ is the symmetric square root of $\boldsymbol{S}_X$. As the name suggests, the polar decomposition is analogous to the polar form $z = e^{i\varphi} r$ of a nonzero complex number, with $\boldsymbol{Q}_X$ being the analog of $e^{i\varphi}$ and $\boldsymbol{S}_X^{1/2}$ being the analog of r. The components of the polar decomposition can be computed from $\boldsymbol{X}$ as $\boldsymbol{Q}_X = \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1/2}$ and $\boldsymbol{S}_X = \boldsymbol{X}^\top \boldsymbol{X}$. In terms of the singular value decomposition $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top$, we have $\boldsymbol{Q}_X = \boldsymbol{U}\boldsymbol{V}^\top$ and $\boldsymbol{S}_X^{1/2} = \boldsymbol{V}\boldsymbol{D}\boldsymbol{V}^\top$. Additionally, the orthogonal component $\boldsymbol{Q}_X$ has an intuitive geometric interpretation as the closest matrix in

$\mathcal{V}(k,p)$ to $\boldsymbol{X}$ in the Frobenius norm, i.e. $\boldsymbol{Q}_X = \text{argmin}_{\boldsymbol{Q}\in\mathcal{V}(k,p)}\|\boldsymbol{X}-\boldsymbol{Q}\|_F$.

Given a density f_Q defined with respect to the uniform measure on $\mathcal{V}(k,p)$, we would like to simulate a random orthogonal matrix $\boldsymbol{Q}$ whose distribution has density f_Q. Our strategy, motivated by the relative ease of unconstrained simulation, is to simulate a random matrix $\boldsymbol{X}$ from a distribution whose $\boldsymbol{Q}_X$-margin has density f_Q. The distributions on $\boldsymbol{X}$ that have the desired marginal distribution for $\boldsymbol{Q}_X$ can be identified via a change of variables. The mapping from a real, full rank matrix $\boldsymbol{X}$ to the components $(\boldsymbol{Q}_X, \boldsymbol{S}_X)$ of its polar decomposition is one-to-one, so the density of the distribution of $\boldsymbol{X}$ can be derived from the density of the joint distribution of $\boldsymbol{Q}_X$ and $\boldsymbol{S}_X$ as

$$f_X(\boldsymbol{X}) = f_{S_X|Q_X}(\boldsymbol{S}_X \mid \boldsymbol{Q}_X)\, f_{Q_X}(\boldsymbol{Q}_X) \times J(\boldsymbol{Q}_X, \boldsymbol{S}_X; \boldsymbol{X}).$$

The Jacobian of the transformation from $\boldsymbol{X}$ to $(\boldsymbol{Q}_X, \boldsymbol{S}_X)$ is provided in Chikuse (2003):

$$J(\boldsymbol{Q}_X, \boldsymbol{S}_X; \boldsymbol{X}) = \frac{\Gamma_k(\frac{p}{2})}{\pi^{\frac{pk}{2}}}\, |\boldsymbol{S}_X|^{-\frac{p-k-1}{2}}.$$

If $\boldsymbol{Q}_X$ is to have marginal density f_Q, we must have

$$f_X(\boldsymbol{X}) = f_{S_X|Q_X}(\boldsymbol{S}_X|\boldsymbol{Q}_X)\, f_Q(\boldsymbol{Q}_X) \times J(\boldsymbol{Q}_X, \boldsymbol{S}_X; \boldsymbol{X}).$$

Putting these observations together, we arrive at the following proposition:

Proposition 11. *The $\boldsymbol{Q}_X$-margin of an absolutely continuous random matrix $\boldsymbol{X}$ has density f_Q if and only if*

$$f_X(\boldsymbol{X}) = f_{S_X|Q_X}(\boldsymbol{S}_X|\boldsymbol{Q}_X)\, f_Q(\boldsymbol{Q}_X) \times J(\boldsymbol{Q}_X, \boldsymbol{S}_X; \boldsymbol{X}). \tag{3.1}$$

There is not a unique distribution for $\boldsymbol{X}$ which has the desired $\boldsymbol{Q}_X$-margin. From Proposition 11, we see there is one such distribution for each choice of conditional density $f_{S_X|Q_X}$. For some simulation problems, there is an obvious choice for the distribution of $\boldsymbol{X}$ having the desired $\boldsymbol{Q}_X-$margin, and the conditional density $f_{S_X|Q_X}$

is an afterthought. For others, there is no obvious choice. In that case, we construct a distribution for $\boldsymbol{X}$ by choosing a conditional density $f_{S_X|Q_X}$ and plugging it into Equation (3.1).

The term "parameter expansion" applies to methods which expand the parameter space of a statistical model by introducing redundant working parameters for computational purposes. The working parameters render the expanded parametrization non-identifiable, but the original parameters of interest can still be recovered. Parameter expansion has been successfully applied in the context of the expectation maximization algorithm (Liu et al., 1998) and MCMC simulation (Liu and Wu, 1999; Van Dyk and Meng, 2001). As the name suggests, polar expansion fits this pattern. When applied to posterior simulation in a model with a parameter $\boldsymbol{Q} \in \mathcal{V}(k, p)$, polar expansion replaces the orthogonal matrix $\boldsymbol{Q}$ having $pk - k(k-1)/2$ free parameters with an unconstrained matrix $\boldsymbol{X} \in \mathbb{R}^{p \times k}$ having pk free parameters. The expanded model is non-identifiable, but the original parameter of interest $\boldsymbol{Q}$ can be recovered via the polar decomposition of $\boldsymbol{X}$.

3.3 Polar expansion and exact Monte Carlo

There are some simple, well-known distributions for a random orthogonal matrix $\boldsymbol{Q}$ which are the $\boldsymbol{Q}_X$-margin of a standard distribution for $\boldsymbol{X}$. If exact Monte Carlo simulation of $\boldsymbol{X}$ is possible, then the same is true of $\boldsymbol{Q}$. To simulate a random orthogonal matrix $\boldsymbol{Q}$ with the desired distribution, we simply simulate $\boldsymbol{X}$ and then set $\boldsymbol{Q} = \boldsymbol{Q}_X$. We go through the following examples, in order of increasing generality, to build intuition about polar expansion and familiarity with the required calculations. We will draw on these foundations in Section 3.4, which addresses polar expansion in more complex settings.

Uniform distribution on the sphere Suppose we want to simulate a random vector $\boldsymbol{Q}$ which is uniformly distributed on the unit sphere $\mathcal{V}(1,p)$. A well-known approach described, for example, in Marsaglia (1972) is to simulate $\boldsymbol{X} \sim N(\mathbf{0}, \boldsymbol{I}_p)$ and then set $\boldsymbol{Q} = \boldsymbol{Q}_X = \boldsymbol{X}/\sqrt{\boldsymbol{X}^\top \boldsymbol{X}}$. The random variable $S_X > 0$ is independent of $\boldsymbol{Q}_X$ and χ^2_p distributed.

Uniform distribution on the Stiefel manifold Now suppose we want to simulate a random orthogonal matrix $\boldsymbol{Q}$ which is uniformly distributed on the Stiefel manifold $\mathcal{V}(k,p)$. We can do so by simulating a random matrix $\boldsymbol{X} \in \mathbb{R}^{p\times k}$ with independent standard normal entries and then setting $\boldsymbol{Q} = \boldsymbol{Q}_X$. This construction of a uniform orthogonal matrix is also well-known (Eaton, 1989). The random SPD matrix $\boldsymbol{S}_X$ is independent of $\boldsymbol{Q}_X$ and Wishart $\mathrm{W}_p(\boldsymbol{I}_k)$ distributed.

Matrix angular central Gaussian The random orthogonal matrix $\boldsymbol{Q}$ is said to have a matrix angular central Gaussian MACG($\boldsymbol{\Sigma}$) distribution if $\boldsymbol{Q} \stackrel{d}{=} \boldsymbol{Q}_X$ where $\boldsymbol{X} \sim N_{p,k}(\mathbf{0}, \boldsymbol{\Sigma}, \boldsymbol{I})$ (Chikuse, 2003). The notation $N_{p,k}(\mathbf{0}, \boldsymbol{\Sigma}, \boldsymbol{I})$ indicates a centered matrix normal distribution with $\boldsymbol{\Sigma}$ as its row covariance matrix and the identity as its column covariance matrix (Srivastava and Khatri, 1979; Dawid, 1981). The MACG($\boldsymbol{\Sigma}$) distribution has density $f_{\boldsymbol{Q}}(\boldsymbol{Q}) = |\boldsymbol{\Sigma}|^{-k/2}|\boldsymbol{Q}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{Q}|^{-p/2}$ and is uniform on the Stiefel manifold when $\boldsymbol{\Sigma} = \boldsymbol{I}$. Clearly, we can simulate $\boldsymbol{Q} \sim$ MACG($\boldsymbol{\Sigma}$) by first simulating $\boldsymbol{X} \sim N_{p,k}(\mathbf{0}, \boldsymbol{\Sigma}, \boldsymbol{I})$ and then setting $\boldsymbol{Q} = \boldsymbol{Q}_X$. The random SPD matrix $\boldsymbol{S}_X$ is independent of $\boldsymbol{Q}_X$ with

$$f_{S_X|Q_X}(\boldsymbol{S}_X \mid \boldsymbol{Q}_X) = \frac{{}_0F_0^{(p)}\left(-\frac{1}{2}\boldsymbol{\Sigma}^{-1}, \boldsymbol{S}_X\right)}{2^{pk/2}\Gamma_k(\frac{p}{2})\,|\boldsymbol{\Sigma}|^{k/2}}\,|\boldsymbol{S}_X|^{(p-k-1)/2}\,. \tag{3.2}$$

See Chikuse (2003) for a discussion of the hypergeometric function ${}_0F_0^{(p)}$ of matrix argument.

As we indicated before, the examples are listed in order of increasing generality. In each case, the distribution of $\boldsymbol{Q}$ is MACG. More generally, any distribution which is the $\boldsymbol{Q}_X$-margin of a standard distribution for $\boldsymbol{X}$ lends itself to exact Monte Carlo simulation via polar expansion.

3.4 Polar expansion and MCMC

In many simulation problems of interest, the target distribution of the random orthogonal matrix $\boldsymbol{Q}$ is not the $\boldsymbol{Q}_X$-margin of a standard distribution for $\boldsymbol{X}$. While exact Monte Carlo simulation from these distributions is out of reach, we can still apply polar expansion to construct a distribution for $\boldsymbol{X}$ which has the desired $\boldsymbol{Q}_X$-margin and is amenable to MCMC simulation. We first consider the scenario in which the distribution of $\boldsymbol{Q}$ is a posterior arising from an MACG prior. Guided by the examples of the previous section, we propose a simple way to construct a distribution for $\boldsymbol{X}$ with the desired $\boldsymbol{Q}_X$-margin. We then consider the very general scenario in which the distribution of $\boldsymbol{Q}$ is specified by a density f_Q which is known up to a multiplicative constant. In this general scenario, we construct a distribution for $\boldsymbol{X}$ which has the desired $\boldsymbol{Q}_X$-margin by choosing a conditional density $f_{S_X|Q_X}$ and plugging it into Equation (3.1). Finally, we motivate our recommendation of HMC for MCMC simulation from the distribution of $\boldsymbol{X}$.

3.4.1 Posterior simulation with an MACG prior

We consider the case in which the distribution of $\boldsymbol{Q}$ is a posterior arising from an MACG($\boldsymbol{\Sigma}$) prior. The MACG($\boldsymbol{\Sigma}$) distribution is uniform when $\boldsymbol{\Sigma} = \boldsymbol{I}$ but can incorporate prior structure such as row dependence when $\boldsymbol{\Sigma} \neq \boldsymbol{I}$. We take advantage of this flexibility in the functional PCA application of Section 3.5.2.

Suppose we have data $\boldsymbol{y}$ whose distribution given the unknown parameter $\boldsymbol{Q} \in \mathcal{V}(k,p)$ has density $p(\boldsymbol{y} \mid \boldsymbol{Q})$. The MACG prior density is $p(\boldsymbol{Q}) = |\boldsymbol{\Sigma}|^{-k/2}|\boldsymbol{Q}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{Q}|^{-p/2}$

and the posterior density satisfies $p(\boldsymbol{Q} \mid \boldsymbol{y}) \propto p(\boldsymbol{y} \mid \boldsymbol{Q})\, p(\boldsymbol{Q})$. To approximate the posterior distribution of $\boldsymbol{Q}$, we propose constructing a Markov chain $\{\boldsymbol{X}_t\}_{t=1}^T$ whose stationary distribution has density

$$
\begin{aligned}
f_X(\boldsymbol{X}) &= p(\boldsymbol{X} \mid \boldsymbol{y}) \\
&\propto p(\boldsymbol{y} \mid \boldsymbol{Q}_X)\, N_{p,k}(\boldsymbol{X} \mid \boldsymbol{0}, \boldsymbol{\Sigma}, \boldsymbol{I})
\end{aligned} \tag{3.3}
$$

and taking $\{\boldsymbol{Q}_{X_t}\}_{t=1}^T$ as our approximation. The $\boldsymbol{Q}_X$-margin of the distribution for $\boldsymbol{X}$ specified by the density (3.3) is the posterior distribution of $\boldsymbol{Q}$. This can be verified formally via a change of variables from $\boldsymbol{X}$ to the components of its polar decomposition. The distribution of $\boldsymbol{X}$ is nonstandard, but knowing its density allows us to apply standard MCMC methods.

An analogous approach to posterior simulation is available whenever the prior distribution for $\boldsymbol{Q}$ is the $\boldsymbol{Q}_X$-margin of a standard distribution for $\boldsymbol{X}$. One can simply replace the matrix normal density in Equation (3.3) with the alternative density for $\boldsymbol{X}$. We emphasize the MACG distribution because of its utility as a prior distribution and because, as far as we are aware, it is the only distribution in the literature which is the $\boldsymbol{Q}_X$-margin of a standard distribution for $\boldsymbol{X}$.

3.4.2 General simulation problems

There are important settings in which the distribution of an orthogonal matrix $\boldsymbol{Q}$ is neither the $\boldsymbol{Q}_X$-margin of a standard distribution for $\boldsymbol{X}$ nor a posterior arising from such a prior. In particular, the distribution of $\boldsymbol{Q}$ might belong to the Bingham-von Mises-Fisher family (Hoff, 2009b) or be a posterior distribution arising from a prior which is not the $\boldsymbol{Q}_X$-margin of a standard distribution for $\boldsymbol{X}$. With these examples in mind, we consider simulating from a distribution for $\boldsymbol{Q}$ specified by a density f_Q which is known up to a multiplicative constant.

In this general scenario, unlike the previous examples, there is no obvious choice for the distribution of $\boldsymbol{X}$ which has the desired $\boldsymbol{Q}_X$-margin. Instead, we construct a distribution for $\boldsymbol{X}$ by choosing a conditional density $f_{S_X|Q_X}$ and plugging it into Equation (3.1). We propose to let $f_{S_X|Q_X} = \mathrm{W}_p(\boldsymbol{S}_X; \boldsymbol{I}_k)$. That is, the conditional density $f_{S_X|Q_X}$ is a Wishart density with p degrees of freedom and $\boldsymbol{I}_k$ as its scale matrix. With this choice, the density of the distribution of $\boldsymbol{X}$ simplifies to

$$f_X(\boldsymbol{X}) = (2\pi)^{-pk/2}\,\mathrm{etr}\left(-\boldsymbol{X}^\top \boldsymbol{X}/2\right) f_{\boldsymbol{Q}}(\boldsymbol{Q}_X). \tag{3.4}$$

When $f_{\boldsymbol{Q}}(\boldsymbol{Q}) \propto 1$ and the distribution of $\boldsymbol{Q}$ is uniform, the entries of $\boldsymbol{X}$ are independent standard normal random variables. This appealing correspondence between the uniform distribution on $\mathcal{V}(k,p)$ and the distribution of pk independent standard normals is one motivation for our choice of conditional density $f_{S_X|Q_X}$. Furthermore, when applied to the problem of simulating from the unit sphere $\mathcal{V}(1,p)$, our proposed approach is equivalent to the method for simulating from $\mathcal{V}(1,p)$ built into Stan at the time of writing (Stan Development Team, 2019).

3.4.3 Hamiltonian Monte Carlo

To simulate from the distribution of $\boldsymbol{X}$, we recommend Hamiltonian Monte Carlo (Neal, 2011). Hamiltonian Monte Carlo (originally Hybrid Monte Carlo (Duane et al., 1987)) is a class of MCMC methods which simulates Hamiltonian dynamics in order to propose long distance moves in the state space while maintaining high acceptance rates. Markov chains produced by HMC typically converge more quickly to their stationary distribution and exhibit less autocorrelation than those produced by random walk Metropolis or Gibbs sampling algorithms. Through their automatic differentiation and adaptive tuning functionality, software implementations such as Stan (Carpenter et al., 2017) greatly simplify applications of HMC. They also provide a powerful set of diagnostics which alert the user to potential problems that may lead to poor Monte Carlo estimates.

3.5 Applications

3.5.1 Network eigenmodel for protein interaction data

We compare polar expansion to competing MCMC approaches in a benchmark protein interaction network application. Using polar expansion with adaptive HMC as implemented in Stan, GMC without parallel tempering, and the Gibbs sampler of Hoff (2009b), we simulate from the posterior distribution of the network eigenmodel of Hoff (2009b) applied to the protein interaction data first appearing in Butland et al. (2005). Compared to GMC, its strongest competitor, polar expansion with adaptive HMC is an order of magnitude more efficient in terms of effective sample size per iteration and comparable in terms of iterations per second.

The application which we use as a benchmark was first introduced in Hoff (2009b). The interactions of $p = 270$ proteins of *Escherichia coli* are recorded in the binary, symmetric $p \times p$ matrix $\boldsymbol{Y} = (y_{i,j})$. If protein i and protein j interact, then $y_{i,j} = 1$. Otherwise, $y_{i,j} = 0$. The edge probabilities are assumed to have a low-rank structure with

$$P(y_{i,j} = 1) = \Phi\left[c + \left(\boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^\top\right)_{i,j}\right] \tag{3.5}$$

where Φ is the cumulative distribution function of a standard normal random variable and $(c, \boldsymbol{Q}, \boldsymbol{\Lambda})$ are unknown parameters. The parameter $\boldsymbol{Q}$ is a $p \times 3$ orthogonal matrix, $\boldsymbol{\Lambda} = \text{diag}(\lambda_1, \lambda_2, \lambda_3)$ is a 3×3 diagonal matrix, and c is a real number. Following Hoff (2009b), $\boldsymbol{Q}$ is *a priori* uniform on $\mathcal{V}(3, p)$, the diagonal elements of $\boldsymbol{\Lambda}$ have independent $N(0, p)$ prior distributions, and $c \sim N(0, 10^2)$.

Hoff (2009b) proposes a Gibbs sampler for posterior simulation. As discussed in Albert and Chib (1993), the probit link function admits a simple data augmentation scheme which often leads to standard conditional posterior distributions. After taking advantage of this data augmentation scheme, the conditional posterior distribution of the orthogonal matrix parameter $\boldsymbol{Q}$ is matrix Bingham-von Mises-Fisher.

Hoff (2009b) provides a column-wise strategy for simulating from this conditional posterior distribution.

An approximation to the posterior distribution of the parameters $(c, \boldsymbol{Q}, \boldsymbol{\Lambda})$ can also be obtained using polar expansion with adaptive HMC as implemented in Stan. To carry out posterior simulation with Stan's adaptive HMC algorithm, we must provide the log posterior density, modulo an additive constant. Applying polar expansion, the log posterior density is

$$\begin{aligned}\log p(c, \boldsymbol{X}, \boldsymbol{\Lambda} \mid \boldsymbol{Y}) = &\sum_{i>j} y_{i,j} \Phi \left[c + \left(\boldsymbol{Q}_X \boldsymbol{\Lambda} \boldsymbol{Q}_X^\top \right)_{i,j} \right] \\ &+ \sum_{i>j} (1 - y_{i,j}) \left\{ 1 - \Phi \left[c + \left(\boldsymbol{Q}_X \boldsymbol{\Lambda} \boldsymbol{Q}_X^\top \right)_{i,j} \right] \right\} \\ &- \frac{c^2}{2 \times 10^2} - \frac{\boldsymbol{X}^\top \boldsymbol{X}}{2} - \sum_{j=1}^{3} \frac{\lambda_j^2}{2p} + C \end{aligned} \tag{3.6}$$

where C is a constant which does not depend upon the parameters. Given a Markov chain $\{c_t, \boldsymbol{X}_t, \boldsymbol{\Lambda}_t\}_{t=1}^T$ whose stationary distribution has density (3.6), we approximate the posterior distribution of $(c, \boldsymbol{Q}, \boldsymbol{\Lambda})$ by $\{c_t, \boldsymbol{Q}_{X_t}, \boldsymbol{\Lambda}_t\}_{t=1}^T$.

Figure 3.1 provides traceplots for the diagonal elements of $\boldsymbol{\Lambda} = \text{diag}(\lambda_1, \lambda_2, \lambda_3)$ based on polar expansion with adaptive HMC, GMC without parallel tempering, and the Gibbs sampler of Hoff (2009b). Stan's diagnostics did not give any indication of problems which would lead to poor Monte Carlo estimates. For GMC, we used the tuning parameters given in Byrne and Girolami (2013). Even visually, we can tell that the Markov chain produced by polar expansion with adaptive HMC exhibits less autocorrelation than those produced via GMC or the Gibbs sampler. This is confirmed by the calculations in Table 3.1 which show that the effective sample size per iteration of our approach is an order of magnitude greater than that of the competing methods. Effective sample size per second is the truly relevant quantity to compare, but variability in code quality and random initializations make such

Parameter	Polar Exp.	GMC	Gibbs
λ_1	0.835	0.031	0.030
λ_2	0.886	0.038	0.030
λ_3	0.683	0.033	0.036

Table 3.1: Effective sample sizes per iteration for the diagonal elements of $\boldsymbol{\Lambda}$ calculated using the `R` package `mcmcse` (Flegal et al., 2017). The calculations are based on 5000 post warm up Markov chain iterations.

comparisons challenging. We remark only that simulating 5000 post warm up Markov chain iterations with our approach took a similar amount of time to the equivalent task with GMC and far less time compared to the Gibbs sampler.

Because one can simultaneously permute the columns of $\boldsymbol{Q}$ and $\boldsymbol{D}$ and change their signs without changing the value of the posterior density, the posterior distribution of the network eigenmodel has multiple symmetric modes. None of the MCMC methods we compare are capable of switching between these symmetric modes. However, this lack of switching does not affect inferences about identifiable parameters. Byrne and Girolami (2013) combine GMC with parallel tempering and show that the resulting Markov chains do switch between symmetric modes. They also describe how, without parallel tempering, Markov chains produced by GMC can become stuck in a local mode with negligible posterior mass. Markov chains produced by HMC applied to the distribution with the log posterior density (3.6) are likewise vulnerable to becoming stuck in this mode. However, all the Markov chains in Figure 3.1 have converged to the same mode as in Byrne and Girolami (2013) and Hoff (2009b).

3.5.2 Principal components analysis of functional data

We propose a new approach to Bayesian functional principal components analysis which we illustrate in a meteorological time series application. Principal component analysis linearly transforms a set of high-dimensional, correlated variables into a lower-dimensional set of uncorrelated "principal component scores," accounting for as much variation in the original data as possible. PCA has become an essential

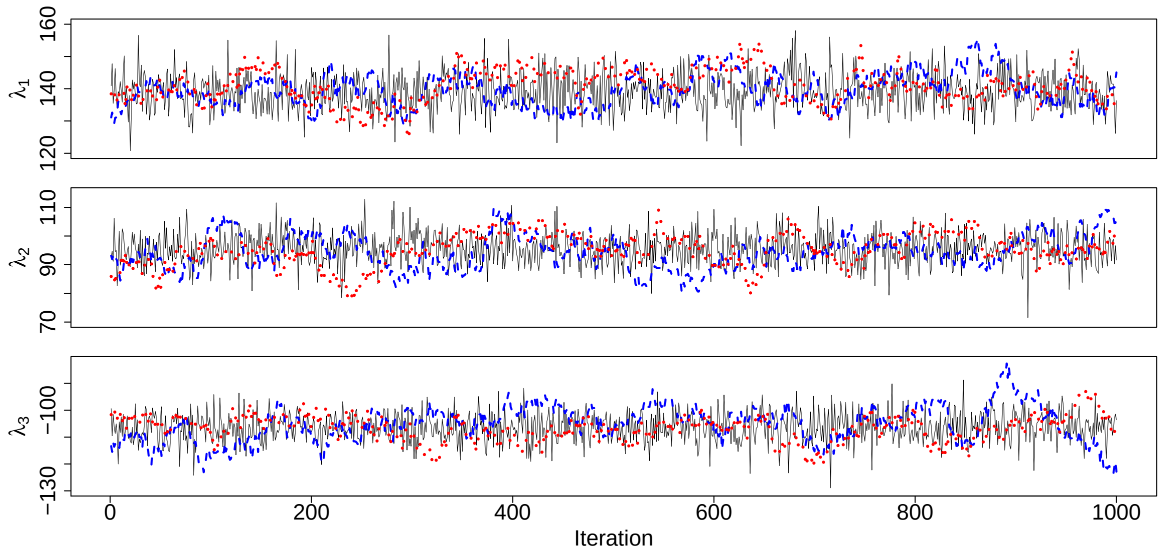

FIGURE 3.1: Traceplots for the diagonal elements of $\mathbf{\Lambda} = \text{diag}(\lambda_1, \lambda_2, \lambda_3)$ based on the three MCMC methods. The solid black lines correspond to polar expansion with adaptive HMC, the dashed blue lines correspond GMC without parallel tempering, and the dotted red lines correspond to the Gibbs sampler of Hoff (2009b).

tool for exploratory data analysis and dimension reduction, and has inspired a vast literature of related methodology. When applied to data arising from an underlying curve or surface, however, classical PCA fails to take the functional structure into account and, as a result, can be excessively noisy. Ramsay and Silverman's influential book (Ramsay and Silverman, 1997) describes how to adapt PCA to functional data from a penalized optimization perspective. As an alternative, our Bayesian approach to principal components analysis of functional data has a number of potential advantages: functional structure can be incorporated through the prior distribution, smoothing parameters can be estimated rather than chosen via cross-validation, and parameter uncertainty is reflected in posterior distribution. Additionally, our method can easily accommodate certain types of missing data and can be flexibly modified or extended.

We consider the Canadian weather data previously analyzed in Ramsay and Silverman (1997) and Suarez and Ghosal (2017). The Canadian weather data set, available in the `R` (Team, 2019) package `FDA` (Ramsay et al., 2018), includes average daily temperatures for 35 weather stations throughout Canada. The raw data matrix $\boldsymbol{Y}_{\text{raw}}$ has $n = 35$ rows and $p = 365$ columns with entry (i, j) recording the average temperature in city i on day j. The columns of $\boldsymbol{Y}_{\text{raw}}$ are plotted in the top left panel of Figure 3.2. Immediately, we see the functional nature of the data, large differences in the average yearly temperature across cities, and a roughly sinusoidal pattern of seasonal variation. Large differences in average yearly temperature are to be expected, given that the data set includes weather stations from Victoria, British Columbia to Inuvik, Northwest Territories. The roughly sinusoidal pattern of seasonal variation is also unsurprising. The aim of our functional principal component analysis is to identify subtler modes of variation present in the data, while taking into account its functional nature.

We subtract row and column means from $\boldsymbol{Y}_{\text{raw}}$ and model the resulting matrix as $\boldsymbol{Y} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top + \sigma\boldsymbol{E}\boldsymbol{\Omega}(\varphi)^{1/2}$. The unknown parameters are the orthogonal matrices $\boldsymbol{U} \in \mathcal{V}(k, n)$ and $\boldsymbol{V} \in \mathcal{V}(k, p)$, the diagonal matrix $\boldsymbol{D} = \text{diag}(d_1, ..., d_k)$ with $d_1, ..., d_k > 0$, the scale parameter $\sigma > 0$, and the correlation parameter φ. The entries of the matrix $\boldsymbol{E}$ are independent standard normal random variables, and $\boldsymbol{\Omega}(\varphi)$ is the correlation matrix of an AR(1) process with parameter φ. The low rank matrix $\boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top$ is intended to capture long term, seasonal variation in temperature. The rows of $\boldsymbol{U}\boldsymbol{D}$ contain the principal component scores for each weather station, while the columns of $\boldsymbol{V}$ form the corresponding basis of principal component curves. In this analysis, we set $k = 3$. The matrix $\sigma\boldsymbol{E}\boldsymbol{\Omega}(\varphi)^{1/2}$ is intended to capture short term, day to day variation in temperature. Conditional on σ and φ, each column of $\sigma\boldsymbol{E}\boldsymbol{\Omega}(\varphi)^{1/2}$ is an independent AR(1) process.

We assign the parameter $\boldsymbol{V}$ a hierarchical prior chosen to reflect the functional nature of the temperature data. Because we intend $\boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top$ to capture long term,

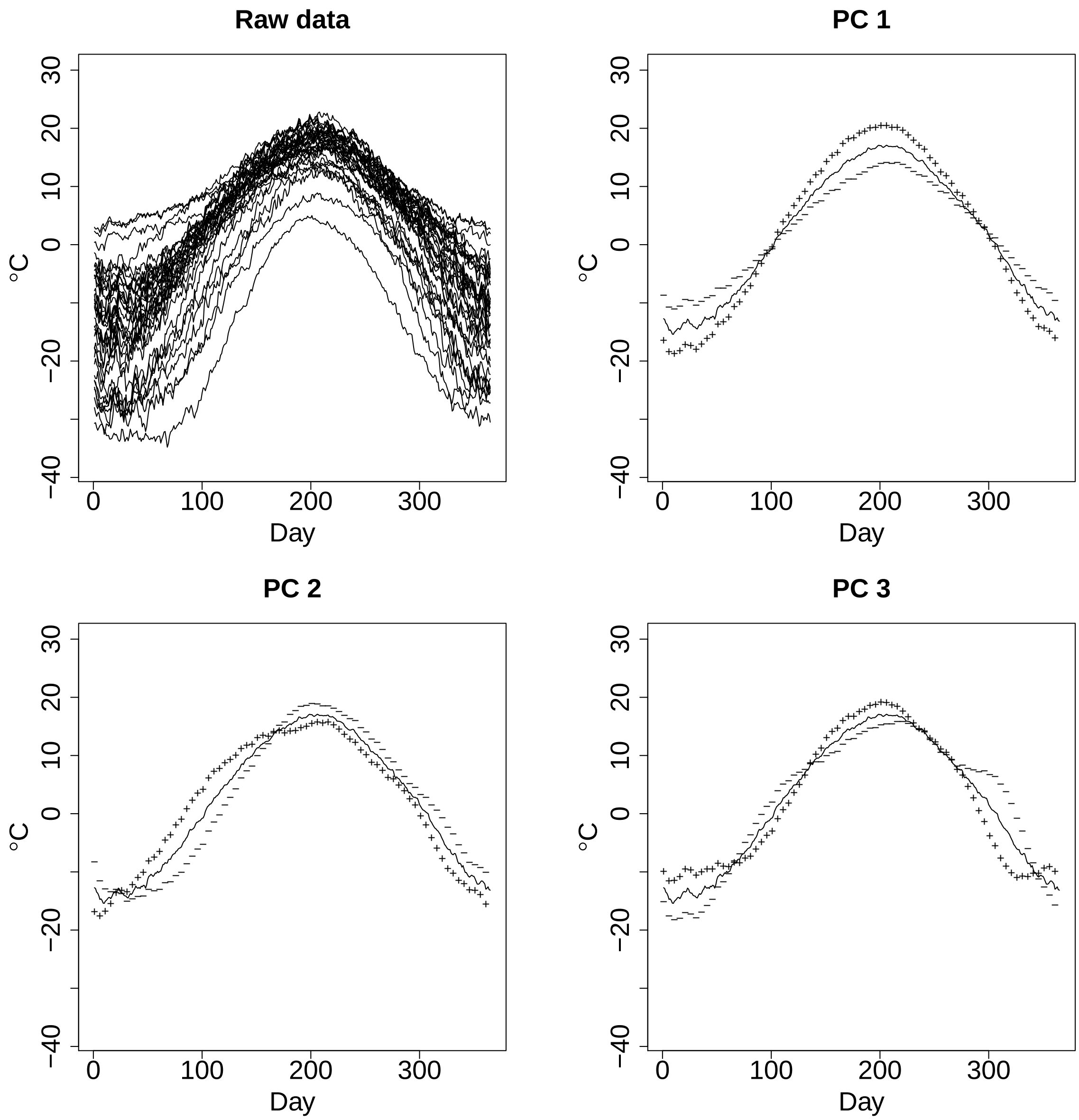

FIGURE 3.2: The columns of $\boldsymbol{Y}_{\text{raw}}$ are plotted in the top left panel. The other three panels plot the column means of $\boldsymbol{Y}_{\text{raw}}$ plus and minus a suitable multiple of the principal component curves.

seasonal variation in temperature, we want the principal component curves in each column of $\boldsymbol{V}$ to look like the discretization of a smooth function. This functional structure can be represented by a MACG($\boldsymbol{K}$) prior when $\boldsymbol{K} = (k_{i,j})$ is constructed using, for instance, the squared exponential correlation function (Rasmussen and Williams, 2006) with $k_{i,j} = \exp\left[-(i-j)^2/(2\rho^2)\right]$. The length-scale hyperparameter ρ controls the "wiggliness" of the principal component curves. We assign ρ an inverse gamma prior, yielding the following hierarchical prior for $\boldsymbol{V}$:

$$\boldsymbol{V} \mid \rho \sim \mathrm{MACG}(\boldsymbol{K})$$
$$1/\rho \sim \mathrm{Ga}(\alpha, \beta).$$

When $\boldsymbol{V}$ is MACG($\boldsymbol{K}$) and $p \gg k$, each column of $\boldsymbol{V}$ behaves like a centered Gaussian process (GP) with a squared exponential correlation function and length-scale ρ. Such a GP is infinitely differentiable, and the expected number of zero crossings in an interval of length T is $T/(2\pi\rho)$ (Rasmussen and Williams, 2006; Adler, 1981). Motivated by the latter observation, we choose the inverse gamma hyperparameters α and β so that ρ has a prior mean of $365/(4\pi)$ and prior standard deviation of five. If ρ were fixed at its prior mean, the expected number of zero crossings of the principal component curves would be approximately two. The proposed inverse gamma prior also puts very little prior mass on small values of ρ. Together, these attributes reflect our intention that $\boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top$ capture long term, seasonal variation in temperature.

We now specify priors for the remaining parameters. The rows of $\boldsymbol{U}$ correspond to locations throughout Canada. We could try to incorporate this spatial structure through the prior distribution, but in this analysis we simply assign $\boldsymbol{U}$ a uniform prior. To the correlation parameter φ, we assign the arc-sine prior discussed in Fosdick and Raftery (2012). The priors for σ^2 are inverse gamma and truncated normal:

$$1/\sigma^2 \sim \mathrm{Ga}\left(\frac{\nu}{2}, \frac{\nu}{2}s^2\right)$$

$$p(d_1, ..., d_k) \propto \mathbb{1}\{d_1, ..., d_k > 0\} \prod_{i=1}^{k} N(d_i\,;\,0, \tau^2).$$

We use an empirical Bayes strategy to select the hyperparameters ν, s^2, and τ. Let $\widehat{\boldsymbol{Y}}$ be the best rank-k approximation to $\boldsymbol{Y}$ in the Frobenius norm (Eckart and Young, 1936), and let $\widehat{\sigma^2}$ be the sample variance of the entries of the residual matrix $\boldsymbol{Y} - \widehat{\boldsymbol{Y}}$. The prior variance of σ^2 is decreasing as a function of the hyperparameter ν, which has an interpretation as a prior sample size in a normal model (Hoff, 2009a). We let $\nu = 1$ and then set $s^2 = 3\,\widehat{\sigma^2}$ so that the prior mode for σ^2 is $\widehat{\sigma^2}$. We choose τ^2 so that the prior expectation of $\sum_{i=1}^{k} d_i^2$ is equal to $\mathrm{Tr}(\widehat{\boldsymbol{Y}}^\top \widehat{\boldsymbol{Y}})$.

We simulate from the posterior of the proposed functional principal components model using polar expansion with adaptive HMC. Again, Stan's diagnostics did not give any indication of problems which would lead to poor Monte Carlo estimates. As our point estimate of $\boldsymbol{V}$, we take the first $k = 3$ right singular vectors of the posterior mean of $\boldsymbol{UDV}^\top$. Figure 3.3 compares our point estimate to the results of classical PCA. The black lines are our estimated principal component curves, while the gray lines are the corresponding values based on classical PCA. Compared to the results of classical PCA, the principal component curves produced by our method are smoother and less noisy.

Figure 3.2 aids in interpreting the principal component curves. The top left panel is a plot of the raw temperature data. The other three panels plot the column means of $\boldsymbol{Y}_{\text{raw}}$ plus and minus a suitable multiple of the principal component curves. The multiple is chosen subjectively for the sake of interpretability. This approach to visualizing principal components analyses of functional data is described in Ramsay and Silverman (1997). We see that the first principal component relates to the difference between summer and winter temperatures, with a higher principal component score corresponding to a larger difference. The second principal component relates to a time shift effect. The third principal component is hardest to interpret, but a higher value appears to indicate a later spring and an earlier end to Autumn.

The left hand side of Figure 3.4 compares a histogram estimate of the marginal posterior density of ρ with its prior density. The marginal posterior distribution of ρ is more concentrated than the prior and has a higher mean, indicating that the proposed method can learn a suitable value for ρ without resorting to cross-validation. The right hand side shows simulated posterior values of the third principal component curve (in gray) and the point estimate (in black). The simulated posterior values, not just the point estimate, are smooth, and their variation about the point estimate reflects the parameter uncertainty remaining.

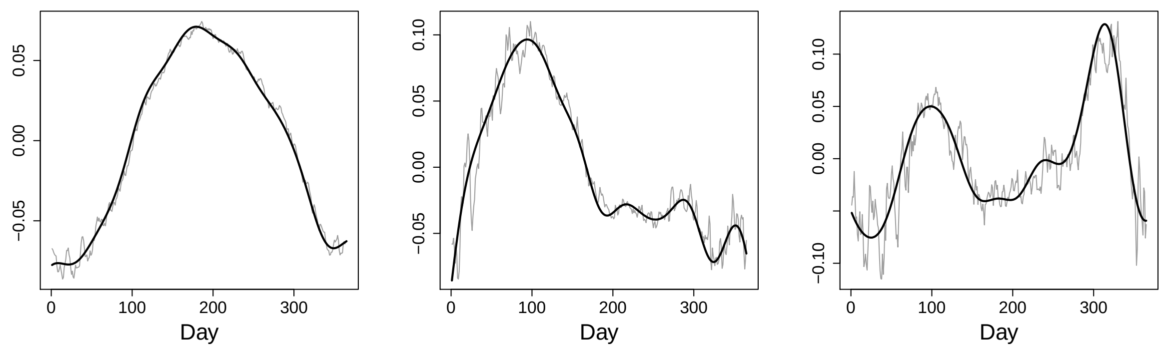

FIGURE 3.3: A comparison of our point estimate of $\boldsymbol{V}$ to the results of classical PCA. The black lines are our estimated principal component curves, while the gray lines are the corresponding values based on classical PCA.

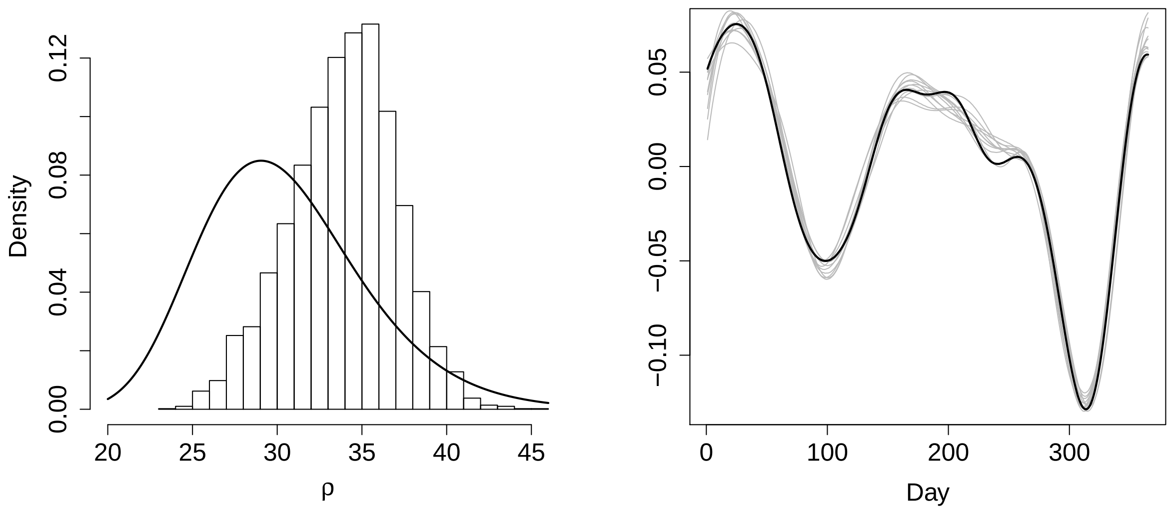

FIGURE 3.4: The left hand side compares a histogram estimate of the marginal posterior density of ρ with its prior density. The right hand side shows simulated posterior values of the third principal component curve (in gray) and its point estimate (in black).

3.6 Discussion

Together with modern MCMC software, polar expansion allows for routine and flexible simulation from probability distributions on the Stiefel manifold, including posterior distributions arising from statistical models with orthogonal matrix parameters. The key idea is to transform the constrained simulation problem into an easier unconstrained problem using the polar decomposition and its Jacobian. We described how to apply polar expansion in simulation problems and then considered two applications. In the first, we found that polar expansion with adaptive HMC is an order of magnitude more efficient than competing MCMC approaches in a benchmark protein interaction network application. In the second, we proposed a new approach to Bayesian functional principal components analysis which we illustrated in a meteorological time series application.

We briefly describe a few directions for future work. Proposition 11 tells us we have a great deal of flexibility in our choice of conditional density $f_{S_X|Q_X}$. The choices in Section 3.4 are motivated by the simplicity of the resulting distribution for $\boldsymbol{X}$. While these choices work well in a wide range of simulation problems, it would be interesting to explore more systematically how the choice of conditional density impacts subsequent MCMC simulation. Thus far, we have made a case for polar expansion based on its practical performance. Recent work on the convergence of HMC may provide tools to analyze polar expansion from a theoretical perspective (Durmus et al., 2017; Livingstone et al., 2018; Bou-Rabee and Sanz-Serna, 2017). Sections 3.4 and 3.5 demonstrate that prior distributions which are the $\boldsymbol{Q}_X$-margin of a standard distribution for $\boldsymbol{X}$ are tractable and useful. An interesting direction is to study the relationship between the distribution of $\boldsymbol{X}$ and its $\boldsymbol{Q}_X$-margin.

4

Priors for structured orthogonal matrices

4.1 Introduction

Structural assumptions regarding unknown parameters play a critical role in statistical theory and practice. In high-dimensional linear regression, where the number of observations is much smaller than the number of covariates, the regression coefficients are often assumed to be sparse. When data are spatially or temporally indexed, it is common to share information across related inferences by assuming that parameters associated with nearby observations will be similar. When structural assumptions such as sparsity and dependence are warranted, they allow us to make meaningful inferences which would otherwise be impossible.

In Bayesian statistics, structural assumptions are commonly reflected in the prior distribution. While there is a substantial literature on prior distributions for real-valued vectors, matrices, or arrays whose entries are sparse or dependent, there has been little work on similar priors for orthogonal matrices. In part, this is because defining probability distributions on the Stiefel manifold which reflect such prior information and lead to tractable posterior inference is challenging.

Nevertheless, there are many applications where these priors would be useful. We saw one example already. In the functional principal components analysis of the Canadian weather data in Section 3.5.2, the prior distribution for the matrix $\boldsymbol{V}$ featured row dependence in order to reflect the functional nature of the temperature data. The resulting point estimates for the column curves were substantially less noisy than those of classical PCA. As we will discuss, priors for sparse orthogonal matrices should prove similarly useful in sparse PCA and inference for pairwise interactions in regression.

In this chapter, we introduce an approach to constructing prior distributions for structured orthogonal matrices which leads to tractable posterior simulation. Let $\boldsymbol{Z} = (z_{i,j})$ be a $p \times k$ matrix with i.i.d. real entries having mean zero and unit variance, let $\boldsymbol{\Omega}$ be a $p \times p$ correlation matrix, and set $\boldsymbol{X} = \boldsymbol{\Omega}^{1/2}\boldsymbol{Z}$. The proposed prior distribution is the $\boldsymbol{Q}_X$-margin of the distribution of $\boldsymbol{X}$. We saw in Section 3.4 that prior distributions which are the $\boldsymbol{Q}_X$-margin of a standard distribution for $\boldsymbol{X}$ lead to tractable posterior inference via polar expansion. As we will demonstrate, features of the distribution of $\boldsymbol{X}$ are inherited by its $\boldsymbol{Q}_X$-margin. If we want the prior distribution for an orthogonal matrix parameter to reflect structural assumptions such as sparsity or row dependence or to satisfy an invariance property, we build these features into the distribution of $\boldsymbol{X}$.

We consider two examples. In the first, the entries of $\boldsymbol{Z}$ are i.i.d. normal-gamma random variables (Griffin and Brown, 2010) and $\boldsymbol{\Omega} = \boldsymbol{I}_p$. The normal-gamma distribution is commonly used as a prior in regression problems when the coefficients are assumed to be sparse. As $\lambda \to 0$, the normal-gamma distribution assigns more mass to a neighborhood of zero and more mass in the tails. We choose $\lambda = 1/5$ and then set $\gamma = \sqrt{\frac{1}{2\lambda}}$ to ensure unit variance. Figure 4.1 compares the columns of a single realization $\boldsymbol{X}_0$ of $\boldsymbol{X}$ to those of $\sqrt{p}\,\boldsymbol{Q}_{X_0}$ when $p = 100$ and $k = 3$. The entries of $\boldsymbol{X}_0$ appear as red dots while the entries of $\sqrt{p}\,\boldsymbol{Q}_{X_0}$ appear as blue circles. In most cases, the blue circles lie directly on top of the red dots, lending support to our claim that features of the distribution of $\boldsymbol{X}$ are inherited by its $\boldsymbol{Q}_X$-margin.

In the second example, which relates to the functional principal components application of Section 3.5.2, the entries of $\boldsymbol{Z}$ are i.i.d. standard normal random variables and $\boldsymbol{\Omega} = (\omega_{i,j})$ is a correlation matrix constructed with the squared exponential correlation function (Rasmussen and Williams, 2006). More precisely, $\omega_{i,j} = \exp\left[-(i-j)^2/(2\rho^2)\right]$. The columns of $\boldsymbol{X}$ are independent multivariate normal random vectors and $\boldsymbol{Q}_X \sim MACG(\boldsymbol{\Omega})$. As with our prior for $\boldsymbol{V}$ in the functional principal components application, each column of $\boldsymbol{X}$ or $\boldsymbol{Q}_X$ looks like the discretization of a smooth function, and the length-scale hyperparameter ρ controls the "wigginess" of these curves. Figure 4.2 compares the columns of a single realization $\boldsymbol{X}_0$ of $\boldsymbol{X}$ to those of $\sqrt{p}\,\boldsymbol{Q}_{X_0}$ when $p = 100, k = 3$, and $\rho = 5$. The blue circles overlap the red dots, and we see again that features of the distribution of $\boldsymbol{X}$ are inherited by its $\boldsymbol{Q}_X$-margin.

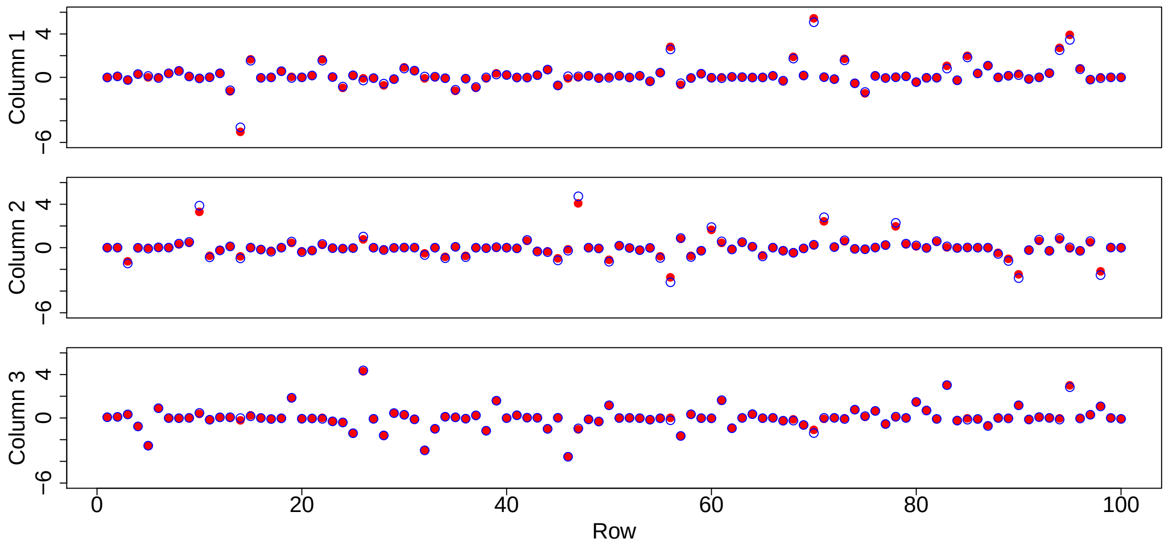

FIGURE 4.1: A comparison of the columns of a single realization $\boldsymbol{X}_0$ of $\boldsymbol{X}$ to those of $\sqrt{p}\,\boldsymbol{Q}_{X_0}$ in the normal-gamma example with $p = 100$ and $k = 3$. The entries of $\boldsymbol{X}_0$ appear as red dots while the entries of $\sqrt{p}\,\boldsymbol{Q}_{X_0}$ appear as blue circles.

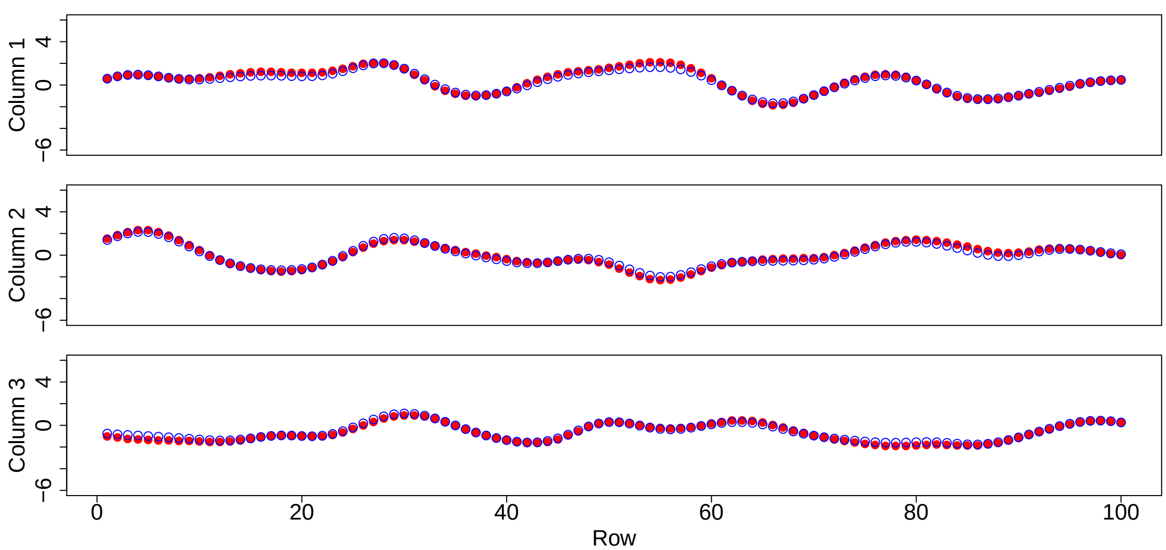

FIGURE 4.2: A comparison of the columns of a single realization $\boldsymbol{X}_0$ of $\boldsymbol{X}$ to those of $\sqrt{p}\,\boldsymbol{Q}_{X_0}$ in the squared exponential correlation example with $p = 100$ and $k = 3$. The entries of $\boldsymbol{X}_0$ appear as red dots while the entries of $\sqrt{p}\,\boldsymbol{Q}_{X_0}$ appear as blue circles.

As theoretical support for our claim that features of the distribution of $\boldsymbol{X}$ are inherited by its $\boldsymbol{Q}_X$-margin, we present two main results. Theorem 12 states that $\boldsymbol{Q}_X$ is invariant to left and right multiplication by given subsets of orthogonal matrices whenever $\boldsymbol{X}$ is. Theorem 13 states that the distribution of finitely many elements of $\sqrt{p}\boldsymbol{Q}_X$ converges weakly to the distribution of the corresponding elements of $\boldsymbol{X}$ in the asymptotic setting in which $p, k \to \infty$ and $k/p \to 0$. Beyond its relevance to statistical modeling, Theorem 13 offers a new perspective on approximating the entries of random orthogonal matrices, a topic of significant interest in random matrix theory.

We outline the remainder of the chapter. In Section 4.2, we present the main results and discuss their relevance to statistical modeling and the wider probability literature. In Section 4.3, we give sketches of applications in which an orthogonal matrix parameter might be assumed to have structure such as sparsity or row dependence.

4.2 Main results

4.2.1 Invariance theorem

The first of our main results addresses invariance of the distribution of $\boldsymbol{X}$ and its $\boldsymbol{Q}_X$-margin to left and right multiplication by subsets of orthogonal matrices. Let $\mathcal{L} \subseteq \mathcal{O}(p)$ and $\mathcal{R} \subseteq \mathcal{O}(k)$ be subsets of the orthogonal groups in dimensions p and k, respectively. A $p \times k$ random matrix $\boldsymbol{Y}$ is invariant to left multiplication by elements of $\mathcal{L}$ if $\boldsymbol{L}\boldsymbol{Y} \overset{d}{=} \boldsymbol{Y}$ for all $\boldsymbol{L} \in \mathcal{L}$. The random matrix $\boldsymbol{Y}$ is invariant to right multiplication by elements of $\mathcal{R}$ if $\boldsymbol{Y}\boldsymbol{R} \overset{d}{=} \boldsymbol{Y}$ for all $\boldsymbol{R} \in \mathcal{R}$. Finally, the random matrix $\boldsymbol{Y}$ is invariant to left multiplication by elements of $\mathcal{L}$ and right multiplication by elements of $\mathcal{R}$ if $\boldsymbol{L}\boldsymbol{Y}\boldsymbol{R} \overset{d}{=} \boldsymbol{Y}$ for all $\boldsymbol{L} \in \mathcal{L}$ and $\boldsymbol{R} \in \mathcal{R}$.

Theorem 12. *If the random matrix $\boldsymbol{X}$ is invariant to left multiplication by elements* *of $\mathcal{L}$ and right multiplication by elements of $\mathcal{R}$, then $\boldsymbol{Q}_X$ is as well.*

To see the relevance of Theorem 12 to prior construction, we consider a few examples. Suppose our prior distribution for an orthogonal matrix parameter is the $\boldsymbol{Q}_X$-margin of the distribution of $\boldsymbol{X}$. The uniform distribution on $\mathcal{V}(k,p)$ is the unique distribution which is invariant to left multiplication by elements of $\mathcal{O}(p)$ and right multiplication by elements of $\mathcal{O}(k)$. Theorem 12 tells us our prior distribution is uniform whenever $\boldsymbol{X}$ is invariant to left multiplication by elements of $\mathcal{O}(p)$ and right multiplication by elements of $\mathcal{O}(k)$. When we expect an orthogonal matrix parameter to have structure such as sparsity or row dependence, a uniform prior is inappropriate. However, less stringent invariance properties may still be desirable. For example, if row indices provide no substantive information, $\boldsymbol{Q}_X$ should be invariant to left multiplication by permutation matrices. Furthermore, if there is no prior information regarding the signs of its entries, $\boldsymbol{Q}_X$ should be invariant to left multiplication by signed permutation matrices. Theorem 12 tells us how to achieve these goals.

4.2.2 Limit theorem

The second of our main results provides an explanation of our observation that the distributions of $\sqrt{p}\boldsymbol{Q}_X$ and $\boldsymbol{X}$ are similar when p is large and $p \gg k$. For each $p \in \mathbb{N}$, let $\boldsymbol{Z}_p = (z_{i,j})$ be a $p \times k_p$ matrix with i.i.d. real entries satisfying $\mathbb{E}[z_{i,j}] = 0$, $\mathbb{V}[z_{i,j}] = 1$, and having finite fourth moments. The subscript on k_p indicates that the number of columns may grow with the number of rows. Let $\boldsymbol{\Omega}_p$ be a $p \times p$ correlation matrix, set $\boldsymbol{X}_p = \boldsymbol{\Omega}_p^{1/2}\boldsymbol{Z}_p$, and let the random orthogonal matrix $\boldsymbol{Q}_{X_p}$ be the orthogonal component of the polar decomposition of $\boldsymbol{X}_p$.

Theorem 13. *When $\boldsymbol{\Omega}_p = \boldsymbol{I}_p$, the distribution of finitely many entries of $\sqrt{p}\boldsymbol{Q}_{X_p}$ converges weakly to the distribution of the corresponding entries of $\boldsymbol{X}_p$ as $p, k \to \infty$ with $k_p/p \to 0$.*

In the proof, we show that a Wasserstein distance between the distributions goes to zero as $p, k \to \infty$ with $k_p/p \to 0$. The random matrices $\boldsymbol{X}_p$ and $\boldsymbol{Q}_{X_p}$ provide a natural coupling of the two distributions from which we derive an upper bound for the Wasserstein distance. The upper bound is $\mathcal{O}\left((k/p)^{1/4}\right)$. Our proof does not obviously extend to the case $\boldsymbol{\Omega}_p \neq \boldsymbol{I}_p$, but we expect that some version of Theorem 13 still holds.

There is significant interest in normal approximations to the entries of random orthogonal matrices in the random matrix theory community. See Section 2.6 for an overview. We highlight an especially close connection between Theorem 13 and this literature. A result of Stam (1982) shows that the distribution of m entries of a uniformly distributed unit vector converges to the distribution of m independent standard normal random variables. Watson (1983) extends this result to a uniformly distributed orthogonal matrix with a fixed number of columns. Theorem 13 offers a new perspective on this topic. It shows that the results of Stam (1982) and Watson (1983), in which the entries of a uniformly distributed orthogonal matrix are approximated by independent standard normals, describe a special case of a more general phenomenon. When the distribution of a random orthogonal matrix is the $\boldsymbol{Q}_X$-margin of a real random matrix $\boldsymbol{X}$, the distribution of its entries can be approximated by the distribution of the corresponding entries of $\boldsymbol{X}$ under weak assumptions.

4.3 Sketches of applications

We sketch applications in which an orthogonal matrix parameter might be assumed to have structure such as sparsity or row dependence. We have already seen one example involving row dependence. In the functional principal components analysis of the Canadian weather data in Section 3.5.2, the prior distribution for the matrix $\boldsymbol{V}$ of principal component curves featured row dependence in order to reflect the functional nature of the temperature data. The resulting point estimates for the column curves were significantly less noisy than those of classical PCA. Priors for sparse orthogonal matrices should prove similarly useful.

Sparse PCA As described in Chapter 3, principal component analysis linearly transforms a set of high-dimensional, correlated variables into a lower-dimensional set of uncorrelated principal component scores, accounting for as much variation in the original data as possible. Suppose $\boldsymbol{Y}$ is an $n \times p$ containing n observations of p variables. A model-based principal components analysis might assume $\boldsymbol{Y} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top + \sigma\boldsymbol{E}$, where $\boldsymbol{U} \in \mathcal{V}(k, n)$, $\boldsymbol{V} \in \mathcal{V}(k, p)$, $\boldsymbol{D}$ is a $k \times k$ diagonal matrix with positive entries on the diagonal, $\boldsymbol{E}$ is a matrix of errors, and $\sigma > 0$. The rows of $\boldsymbol{U}\boldsymbol{D}$ contain the principal component scores for each observation, while the columns of $\boldsymbol{V}$ form the corresponding basis. When $\boldsymbol{V}$ is a dense matrix, principal component scores are linear combinations of every original variable and thus challenging to interpret. This limitation, along with the challenge of estimating the basis vectors in high-dimensional problems (Johnstone and Lu, 2009), has motivated sparse PCA methodology. The main approaches involve optimization with sparsity-inducing penalties (Witten et al., 2009) or Bayesian analyses with appropriate priors (Gao and Zhou, 2015; Yoshida and West, 2010; Cron and West, 2016). Typically, these variations on PCA relax some features of the original formulation, such as the orthogonality of the principal axes. Polar expansion and the prior distributions described in this chapter offer an alternative. Suppose the prior distribution for $\boldsymbol{V}$ is the $\boldsymbol{Q}_X$-margin of a random matrix $\boldsymbol{X}$ whose entries are i.i.d. normal-gamma random variables with unit variance and a small value for λ. When p is large and $p \gg k$, as is typical in PCA, the prior distribution for $\sqrt{p}\boldsymbol{V}$ resembles the distribution of $\boldsymbol{X}$, with many entries close to zero and a small number of large entries.

Inference for pairwise interactions in regression Suppose y_i is a real-valued response variable and $\boldsymbol{x}_i$ is a $p \times 1$ vector of covariates. For example, the response might be a health outcome and the covariates might be chemical exposures, as in the environmental health application of Ferrari (2019). In many applications, pairwise interactions are of interest. For example, exposure to a particular pair of chemicals may be especially harmful, beyond what would be predicted by a linear regression model including only main effects. The quadratic regression model

$$
\begin{aligned}
&y_i = \mu + \boldsymbol{x}_i^\top \boldsymbol{\omega} + \boldsymbol{x}_i^\top \boldsymbol{\Omega} \boldsymbol{x}_i + \epsilon_i \\
&\epsilon_i \sim N(0, \sigma^2) \\
&\boldsymbol{\Omega} = \boldsymbol{\Omega}^\top
\end{aligned}
$$

includes all pairwise interactions, but has $p + \binom{p}{2}$ more parameters than the main effects model. Estimating such a large number of parameters is unrealistic in most applications, unless we make additional structural assumptions on $\boldsymbol{\Omega}$. As a first step, we might assume that $\boldsymbol{Q}$ has rank $k \ll p$. In that case, $\boldsymbol{Q} = \boldsymbol{V}\boldsymbol{D}\boldsymbol{V}^\top$ where $\boldsymbol{V} \in \mathcal{V}(k, p)$ and $\boldsymbol{D} = \text{diag}(d_1, ..., d_k)$. If we also expect that only a small number of interaction terms are nonzero, we can assign $\boldsymbol{V}$ the prior discussed in the sparse PCA sketch. The induced prior for $\boldsymbol{\Omega}$ reflects our structural assumptions that $\boldsymbol{\Omega}$ is sparse and low rank.

5

Conclusion

This dissertation has focused on random orthogonal matrices with applications in statistics. Bayesian inference for statistical models with orthogonal matrix parameters has been a recurring theme, but several of the results on random orthogonal matrices will be of interest to those in the broader probability and random matrix theory communities. We conclude with a brief summary of the contributions along with directions for future work.

Chapter 2 deals with random orthogonal matrices and the Cayley transform. We parametrize the Stiefel and Grassmann manifolds, represented as subsets of orthogonal matrices, in terms of Euclidean parameters using the Cayley transform and then derived Jacobian terms for change of variables formulas. This allows for Markov chain Monte Carlo simulation from probability distributions defined on the Stiefel and Grassmann manifolds. We also establish an asymptotic independent normal approximation for the distribution of the Euclidean parameters corresponding to the uniform distribution on the Stiefel manifold.

This chapter leaves open a few directions for future work. The Jacobian terms related to the Cayley transform are opaque and expensive to compute. Finding a way to simplify these terms could provide insight the distribution of the Euclidean parameters corresponding to the uniform distribution on the Stiefel manifold and make MCMC simulation more practical. Theorem 9 holds when the number of columns k_p is $o\left(\frac{p^{1/4}}{\sqrt{\log p}}\right)$. It is an open question whether this is as fast as k_p can grow or whether Theorem 9 can be strengthened. Theorem 9 shows that the distribution of the Euclidean parameters can be approximated by independent normals "in probability." It would be interesting to know whether the approximation holds in a stronger sense.

In Chapter 3, we present *polar expansion*, a general approach to Monte Carlo simulation from probability distributions on the Stiefel manifold. When combined with modern MCMC software, polar expansion allows for routine and flexible posterior inference in models with orthogonal matrix parameters.

Future work could consider the choice of conditional density $f_{S_X|Q_X}$ and theoretical properties of polar expansion. Proposition 11 tells us we have a great deal of flexibility in our choice of conditional density $f_{S_X|Q_X}$. The choices in Section 3.4 are motivated by the simplicity of the resulting distribution for $\boldsymbol{X}$. While these choices work well in a wide range of simulation problems, it would be interesting to explore more systematically how the choice of conditional density impacts subsequent MCMC simulation. Chapter 3 makes a case for polar expansion based on its practical performance. Recent work on the convergence of HMC may provide tools to analyze polar expansion from a theoretical perspective (Durmus et al., 2017; Livingstone et al., 2018; Bou-Rabee and Sanz-Serna, 2017).

Chapter 4 addresses prior distributions for structured orthogonal matrices. We introduce an approach to constructing prior distributions for structured orthogonal matrices which leads to tractable posterior simulation via polar expansion. We state two main results which support our approach and offer a new perspective on approximating the entries of random orthogonal matrices.

There is significant room to build upon the ideas in Chapter 4. In its current form, Theorem 13 requires that $\boldsymbol{\Omega} = \boldsymbol{I}_{k_p}$. We expect that some version of Theorem 13 holds when $\boldsymbol{\Omega} \neq \boldsymbol{I}_{k_p}$, and we intend to pursue this direction in future work. We also intend to go beyond the sketches provided in Chapter 4 and apply the proposed priors in real applications.

Appendix A

Appendix to Chapter 2

A.1 Proofs

A.1.1 *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

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

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.1.2 *Proof of Proposition 3*

One can compute, following Magnus and Neudecker (1988), 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

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

By Lemma 2, we have that $\operatorname{vec} X_{\boldsymbol{\varphi}} = \boldsymbol{\Gamma}_{\mathcal{V}}\boldsymbol{\varphi}$. Thus,

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

Using the first identification table of Magnus and Neudecker (1988), we identify the derivative matrix

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

A.1.3 Proof of Proposition 4

We first show that l is injective on $\mathcal{V}^+(k,p)$. Let $\boldsymbol{Q} = \left[\boldsymbol{Q}_1^T\, \boldsymbol{Q}_2^T\right]^T$, $\boldsymbol{Q}' = \left[\boldsymbol{Q}_1'^T\, \boldsymbol{Q}_2'^T\right]^T \in \mathcal{V}^+(k,p)$ and suppose that $l(\boldsymbol{Q}) = l\left(\boldsymbol{Q}'\right)$, i.e. the columns of $\boldsymbol{Q}$ and $\boldsymbol{Q}'$ span the same subspace. There must exist $\boldsymbol{R} \in \mathcal{O}(k)$ such that $\boldsymbol{Q} = \boldsymbol{Q}'\boldsymbol{R}$ so that $\boldsymbol{Q}_1 = \boldsymbol{Q}_1'\boldsymbol{R}$. Because the matrix $\boldsymbol{Q}_1$ 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 Eaton (1983)). We conclude that $\boldsymbol{R} = \boldsymbol{I}_k$ and $\boldsymbol{Q}_1 = \boldsymbol{Q}_1'$. 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

$$\mathcal{V}^{\mathrm{N}}(k,p) = \left\{\boldsymbol{Q} = \begin{bmatrix}\boldsymbol{Q}_1^T & \boldsymbol{Q}_2^T\end{bmatrix}^T \in \mathcal{V}(k,p) : \boldsymbol{Q}_1 \text{ is nonsingular}\right\}.$$

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

Proof. The direction $l\left[\mathcal{V}^+(k,p)\right] \subseteq l\left[\mathcal{V}^{\mathrm{N}}(k,p)\right]$ is immediate. Now, let $S \in l\left[\mathcal{V}^{\mathrm{N}}(k,p)\right]$. There must exist $\boldsymbol{Q} = \left[\boldsymbol{Q}_1^T\, \boldsymbol{Q}_2^T\right]^T \in \mathcal{V}^{\mathrm{N}}(k,p)$ having S as its column space. Let $\boldsymbol{Q}_1 = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T$ be the singular value decomposition of $\boldsymbol{Q}_1$ and set $\boldsymbol{Q}' = \boldsymbol{Q}\boldsymbol{V}\boldsymbol{U}^T$. Then $l\left(\boldsymbol{Q}'\right) = S$ because $\boldsymbol{V}\boldsymbol{U}^T \in \mathcal{O}(k)$ and $\boldsymbol{Q}' \in \mathcal{V}^+(k,p)$ because its square top block $\boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^T$ is symmetric positive definite. Thus $S \in l\left[\mathcal{V}^+(k,p)\right]$ and we conclude that $l\left[\mathcal{V}^{\mathrm{N}}(k,p)\right] \subseteq l\left[\mathcal{V}^+(k,p)\right]$. $\square$

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 $P_{\mathcal{G}(k,p)}(A) = P_{\mathcal{V}(k,p)}\left[l^{-1}(A)\right]$. Thus,

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

A.1.4 Proof of Proposition 5

We begin with $\boldsymbol{Q} = [\boldsymbol{Q}_1^T\, \boldsymbol{Q}_2^T] \in \mathcal{V}^{+}(k,p)$ and we want to verify that the matrix $\boldsymbol{A}$ obtained by equations 2.4-2.5 satisfies $\mathrm{eval}_i(\boldsymbol{A}^T\boldsymbol{A}) \in [0,1)$ for $1 \leqslant i \leqslant k$. Let $\boldsymbol{Q}_1 = \boldsymbol{V}\mathrm{diag}(\lambda_1, ..., \lambda_k)\boldsymbol{V}^T$ be the eigendecomposition of $\boldsymbol{Q}_1$. That each eigenvalue is positive follows from the condition $\boldsymbol{Q}_1 > \boldsymbol{0}$. We know that each eigenvalue is less than or equal to one because

$$\begin{aligned}
\boldsymbol{0}_k &\preceq \boldsymbol{Q}_2^T\boldsymbol{Q}_2 \\
&= \boldsymbol{I}_k - \boldsymbol{Q}_1^T\boldsymbol{Q}_1 \\
&= \boldsymbol{V}\mathrm{diag}\left(1-\lambda_1^2, ..., 1-\lambda_k^2\right)\boldsymbol{V}^T.
\end{aligned}$$

Therefore, $\lambda_i \in (0,1]$ for each i. As in equations 2.4-2.5, set

$$\begin{aligned}
\boldsymbol{F} &= (\boldsymbol{I}_k - \boldsymbol{Q}_1)(\boldsymbol{I}_k + \boldsymbol{Q}_1)^{-1} \\
\boldsymbol{A} &= \frac{1}{2}\boldsymbol{Q}_2(\boldsymbol{I}_k + \boldsymbol{F}).
\end{aligned}$$

We can write $\boldsymbol{A}^T\boldsymbol{A}$ as

$$\begin{aligned}
\boldsymbol{A}^T\boldsymbol{A} &= \frac{1}{4}(\boldsymbol{I}_k + \boldsymbol{F})^T\boldsymbol{Q}_2^T\boldsymbol{Q}_2(\boldsymbol{I}_k + \boldsymbol{F}) \\
&= \frac{1}{4}(\boldsymbol{I}_k + \boldsymbol{F})^T\left(\boldsymbol{I}_k - \boldsymbol{Q}_1^T\boldsymbol{Q}_1\right)(\boldsymbol{I}_k + \boldsymbol{F})
\end{aligned}$$

from which it follows that

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

for each i. Since the eigenvalues of $\boldsymbol{Q}_1$ lie in the interval $(0,1]$, the eigenvalues $\boldsymbol{A}^T\boldsymbol{A}$ lie in the interval $[0,1)$.

Now we start with $\boldsymbol{A}$ such that $\delta_i = \text{eval}_i(\boldsymbol{A}^T\boldsymbol{A}) \in [0,1)$ for each i and we want to check that $\boldsymbol{Q}_1 = (\boldsymbol{I}_k - \boldsymbol{A}^T\boldsymbol{A})(\boldsymbol{I}_k + \boldsymbol{A}^T\boldsymbol{A})^{-1} \succ \boldsymbol{0}$. Let $\boldsymbol{A}^T\boldsymbol{A} = \boldsymbol{W}\text{diag}(\delta_1, ..., \delta_k)\boldsymbol{W}^T$ be the eigendecomposition of $\boldsymbol{A}^T\boldsymbol{A}$. Then

$$\boldsymbol{Q}_1 = \boldsymbol{W}\text{diag}\left(\frac{1-\delta_1}{1+\delta_1}, ..., \frac{1-\delta_k}{1+\delta_k}\right)\boldsymbol{W}^T \succ \boldsymbol{0}.$$

A.1.5 Proof of Proposition 7

The proof is nearly identical to that of Proposition 3. We simply replace $\boldsymbol{\varphi}$ with $\boldsymbol{\psi}$ and $\boldsymbol{\Gamma}_{\mathcal{V}}$ with $\boldsymbol{\Gamma}_{\mathcal{G}}$.

A.1.6 Proof of Proposition 10 part (i)

Denote the square top block of $\boldsymbol{Q}_p$ by $\boldsymbol{Q}_{p,1}$ and the bottom block by $\boldsymbol{Q}_{p,2}$. Define matrices

$$\begin{aligned}
\boldsymbol{F}_p &= (\boldsymbol{I}_{k_p} - \boldsymbol{Q}_{p,1})(\boldsymbol{I}_{k_p} + \boldsymbol{Q}_{p,1})^{-1} \\
\boldsymbol{B}_p &= \frac{1}{2}(\boldsymbol{F}_p^T - \boldsymbol{F}_p) \\
\boldsymbol{A}_p &= \frac{1}{2}\boldsymbol{Q}_{p,2}(\boldsymbol{I}_{k_p} + \boldsymbol{F}_p)
\end{aligned}$$

as in equations 2.1-2.3. Let $\boldsymbol{b}_p$ be the vector of independent entries of $\boldsymbol{B}_p$ obtained by eliminating diagonal and supradiagonal elements from $\text{vec}\,\boldsymbol{B}_p$. When the Frobenius norm $\|\boldsymbol{Q}_{p,1}\|_F$ is less than one, the matrices admit series representations:

$$\begin{aligned}
\boldsymbol{F}_p &= \boldsymbol{I}_{k_p} - 2\boldsymbol{Q}_{p,1} + 2\boldsymbol{Q}_{p,1}^2 - 2\boldsymbol{Q}_{p,1}^3 + ... \\
\boldsymbol{B}_p &= (\boldsymbol{Q}_{p,1} - \boldsymbol{Q}_{p,1}^T) - (\boldsymbol{Q}_{p,1}^2 - \boldsymbol{Q}_{p,1}^{2T}) + (\boldsymbol{Q}_{p,1}^3 - \boldsymbol{Q}_{p,1}^{3T}) - ... \\
\boldsymbol{A}_p &= \boldsymbol{Q}_{p,2} - \boldsymbol{Q}_{p,2}\boldsymbol{Q}_{p,1} + \boldsymbol{Q}_{p,2}\boldsymbol{Q}_{p,1}^2 - \boldsymbol{Q}_2\boldsymbol{Q}_{p,1}^3 + ...
\end{aligned}$$

It follows that

$$\widetilde{\boldsymbol{B}}_p(\boldsymbol{Q}_p) - \boldsymbol{B}_p = \sum_{i=2}^{\infty}(-1)^i\left(\boldsymbol{Q}_{p,1}^i - \boldsymbol{Q}_{p,1}^{iT}\right)$$

$$\widetilde{\boldsymbol{A}}_p(\boldsymbol{Q}_p) - \boldsymbol{A}_p = \sum_{i=1}^{\infty}(-1)^{i-1}\boldsymbol{Q}_{p,2}\boldsymbol{Q}_{p,1}^i$$

when $\|\boldsymbol{Q}_{p,1}\|_F < 1$. See Hubbard and Hubbard (2009) for a discussion of matrix geometric series.

Inequalities relating the Frobenius and max norms will prove useful. Let $\boldsymbol{W}_1, \boldsymbol{W}_2$ be $d_1 \times d_2$ and $d_2 \times d_3$ dimensional matrices, respectively. Then

$$\|\boldsymbol{W}_1\|_F \leqslant \sqrt{d_1 d_2}\|\boldsymbol{W}_1\|_{\max}$$

$$\|\boldsymbol{W}_1\boldsymbol{W}_2\|_{\max} \leqslant d_2\|\boldsymbol{W}_1\|_{\max}\|\boldsymbol{W}_2\|_{\max}.$$

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

We will also need the following somewhat technical lemma:

Lemma 15. *The following quantities, which will appear in later inequalities, go to zero in probability as p grows provided that* $k_p = o\left(\frac{p^{1/4}}{\sqrt{\log p}}\right)$ *:*

(i) $\|p^{1/2}\boldsymbol{Q}_p - \boldsymbol{Z}_p\|_{\max}$ $\quad$ (ii) $k_p\|\boldsymbol{Q}_{p,1}\|_{\max}$

(iii) $k_p^2 p^{1/2}\|\boldsymbol{Q}_{p,1}\|_{\max}^2$ $\quad$ (iv) $k_p p^{1/2}\|\boldsymbol{Q}_p\|_{\max}^2$.

Proof. Suppose $k_p = o\left(\frac{p^{1/4}}{\sqrt{\log p}}\right)$. This implies that $k_p = o\left(\frac{p}{\log p}\right)$ and quantity (i) goes to zero in probability by Theorem 3 of Jiang (2006). The quantities (ii)-(iv) are nonnegative and each is smaller than either $k_p^2\, p^{1/2}\|\boldsymbol{Q}_p\|_{\max}^2$ or its square root. If we can show that either this quantity or its square root goes to zero in probability, we are done. We have the following upper bound for the square root:

$$\begin{aligned} k_p\, p^{1/4}\,\|\boldsymbol{Q}_p\|_{\max} &= k_p\, p^{1/4}\left\|\boldsymbol{Q}_p - p^{-1/2}\boldsymbol{Z}_p + p^{-1/2}\boldsymbol{Z}_p\right\|_{\max} \\ &\leqslant k_p\, p^{1/4}\left(\left\|\boldsymbol{Q}_p - p^{-1/2}\boldsymbol{Z}_p\right\|_{\max} + \left\|p^{-1/2}\boldsymbol{Z}_p\right\|_{\max}\right) \end{aligned}$$

$$= k_p \, p^{-1/4} \left\| p^{1/2} \boldsymbol{Q}_p - \boldsymbol{Z}_p \right\|_{\max} + k_p \, p^{-1/4} \left\| \boldsymbol{Z}_p \right\|_{\max}.$$

We know that the first summand goes to zero in probability by the condition on k_p and the fact that quantity (i) goes to zero in probability. Using a well-known inequality involving the maximum of independent standard normal random variables (see Problem 5.1 in van Handel (2016)), we obtain the following bound on the expected value of the second summand:

$$\begin{aligned} k_p \, p^{-1/4} \mathbb{E}\left[\| \boldsymbol{Z}_p \|_{\max} \right] &\leqslant k_p \, p^{-1/4} \sqrt{2 \log p k_p} \\ &\leqslant k_p \, p^{-1/4} \sqrt{2 \log p^2} \\ &= 2 \frac{k_p}{\left(\frac{p^{1/4}}{\sqrt{\log p}} \right)}. \end{aligned}$$

The condition on k_p implies that the expectation of the second summand goes to zero. We conclude that the second summand goes to zero in mean and thus in probability. We have shown that an upper bound for the square root of $k_p^2 \, p^{1/2} \| \boldsymbol{Q}_p \|_{\max}^2$ goes to zero in probability, which is sufficient to prove the lemma. □

Now set $a_p = \| \boldsymbol{\Pi}_p \widetilde{\boldsymbol{\varphi}}_p - \boldsymbol{\Pi}_p \boldsymbol{\varphi}_p \|_\infty$, assume that $k_p = o\left(\frac{p^{1/4}}{\sqrt{\log p}} \right)$, 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

$$\begin{aligned} \Pr\{a_p > \epsilon\} = & \Pr\left\{ a_p > \epsilon \mid k_p \| \boldsymbol{Q}_{p,1} \|_{\max} < 1 \right\} \Pr\left\{ k_p \| \boldsymbol{Q}_{p,1} \|_{\max} < 1 \right\} + \\ & \Pr\left\{ a_p > \epsilon \mid k_p \| \boldsymbol{Q}_{p,1} \|_{\max} \geqslant 1 \right\} \Pr\left\{ k_p \| \boldsymbol{Q}_{p,1} \|_{\max} \geqslant 1 \right\} \\ \leqslant & \Pr\left\{ a_p > \epsilon \mid k_p \| \boldsymbol{Q}_{p,1} \|_{\max} < 1 \right\} + \Pr\left\{ k_p \| \boldsymbol{Q}_{p,1} \|_{\max} \geqslant 1 \right\}. \end{aligned}$$

Lemma 15 implies that $\Pr\{k_p \| \boldsymbol{Q}_{p,1} \|_{\max} \geqslant 1\}$ goes to zero. It follows that $\Pr\{a_p > \epsilon\}$ goes to zero if $\Pr\left\{ a_p > \epsilon \mid k_p \| \boldsymbol{Q}_{p,1} \|_{\max} < 1 \right\}$ does. Therefore, we only need to verify this condition.

For each p, we have

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

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

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

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

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

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

and

$$\sqrt{p}\, \left\| \widetilde{\boldsymbol{A}}_p(\boldsymbol{Q}_p) - \boldsymbol{A}_p \right\|_{\max} = \sqrt{p} \left\| \sum_{i=1}^{\infty} (-1)^{i-1} \boldsymbol{Q}_{p,2} \boldsymbol{Q}_{p,1}^i \right\|_{\max}$$

$$
\begin{aligned}
&\leqslant \sqrt{p} \sum_{n=1}^{\infty} \left\| \boldsymbol{Q}_{p,2} \boldsymbol{Q}_{p,1}^{n} \right\|_{\max} \\
&\leqslant \sqrt{p} \sum_{n=1}^{\infty} k_p \left\| \boldsymbol{Q}_{p,2} \right\|_{\max} \left\| \boldsymbol{Q}_{p,1}^{n} \right\|_{\max} \\
&\leqslant \sqrt{p} \left\| \boldsymbol{Q}_{p,2} \right\|_{\max} \sum_{n=1}^{\infty} \left(k_p \left\| \boldsymbol{Q}_{p,1} \right\|_{\max} \right)^{n} \\
&= \sqrt{p} \left\| \boldsymbol{Q}_{p,2} \right\|_{\max} \left(\frac{1}{1 - k_p \left\| \boldsymbol{Q}_{p,1} \right\|_{\max}} - 1 \right) \\
&= \sqrt{p} \left\| \boldsymbol{Q}_{p,2} \right\|_{\max} \left(\frac{k_p \left\| \boldsymbol{Q}_{p,1} \right\|_{\max}}{1 - k_p \left\| \boldsymbol{Q}_{p,1} \right\|_{\max}} \right) \\
&\leqslant \frac{k_p \sqrt{p} \max \left\{ \left\| \boldsymbol{Q}_{p,1} \right\|_{\max}, \left\| \boldsymbol{Q}_{p,2} \right\|_{\max} \right\}^{2}}{1 - k_p \left\| \boldsymbol{Q}_{p,1} \right\|_{\max}} \\
&= \frac{k_p \sqrt{p} \left\| \boldsymbol{Q}_{p} \right\|_{\max}^{2}}{1 - k_p \left\| \boldsymbol{Q}_{p,1} \right\|_{\max}}.
\end{aligned}
$$

Thus, the upper bound

$$
\begin{aligned}
a_p &\leqslant \sqrt{p/2} \left\| \widetilde{\boldsymbol{B}}_p(\boldsymbol{Q}_p) - \boldsymbol{B}_p \right\|_{\max} + \sqrt{p} \left\| \widetilde{\boldsymbol{A}}_p(\boldsymbol{Q}_p) - \boldsymbol{A}_p \right\|_{\max} \\
&\leqslant \sqrt{2} \frac{k_p^2 \sqrt{p} \left\| \boldsymbol{Q}_{p,1} \right\|_{\max}^{2}}{1 - k_p \left\| \boldsymbol{Q}_{p,1} \right\|_{\max}} + \frac{k_p \sqrt{p} \left\| \boldsymbol{Q}_{p} \right\|_{\max}^{2}}{1 - k_p \left\| \boldsymbol{Q}_{p,1} \right\|_{\max}} := u_p
\end{aligned}
$$

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

$$
\begin{aligned}
\Pr\left\{ a_p > \epsilon \mid k_p \|\boldsymbol{Q}_{p,1}\|_{\max} < 1 \right\} &\leqslant \Pr\left\{ u_p > \epsilon \mid k_p \|\boldsymbol{Q}_{p,1}\|_{\max} < 1 \right\} \\
&= \frac{\Pr\left\{ u_p > \epsilon \cap k_p \|\boldsymbol{Q}_{p,1}\|_{\max} < 1 \right\}}{\Pr\left\{ k_p \|\boldsymbol{Q}_{p,1}\|_{\max} < 1 \right\}} \\
&\leqslant \frac{\Pr\left\{ u_p > \epsilon \right\}}{\Pr\left\{ k_p \|\boldsymbol{Q}_{p,1}\|_{\max} < 1 \right\}}.
\end{aligned}
$$

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 15. 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.1.7 Proof of Proposition 10 part (ii)

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

$$\begin{aligned}
\|\boldsymbol{\Pi}_p \widetilde{\boldsymbol{\varphi}}_p - \boldsymbol{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}\operatorname{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}\operatorname{vec}\widetilde{A}_p(p^{-1/2}\boldsymbol{Z}_p) \end{bmatrix} \right\|_\infty \\
&= \max\left\{ \sqrt{p/2}\left\|\widetilde{B}_p(\boldsymbol{Q}_p) - \widetilde{B}_p(p^{-1/2}\boldsymbol{Z}_p)\right\|_{\max}, \right. \\
&\qquad\qquad \left. \sqrt{p}\left\|\widetilde{A}_p(\boldsymbol{Q}_p) - \widetilde{A}_p(p^{-1/2}\boldsymbol{Z}_p)\right\|_{\max} \right\} \\
&\leqslant \sqrt{p/2}\left\|\widetilde{B}_p(\boldsymbol{Q}_p) - \widetilde{B}_p(p^{-1/2}\boldsymbol{Z}_p)\right\|_{\max} + \\
&\quad \sqrt{p}\left\|\widetilde{A}_p(\boldsymbol{Q}_p) - \widetilde{A}_p(p^{-1/2}\boldsymbol{Z}_p)\right\|_{\max} \\
&= \sqrt{p/2}\left\|\boldsymbol{Q}_{p,1} - p^{-1/2}\boldsymbol{Z}_{p,1} + p^{-1/2}\boldsymbol{Z}_{p,1}^T - \boldsymbol{Q}_{p,1}^T\right\|_{\max} + \\
&\quad \sqrt{p}\left\|\boldsymbol{Q}_{p,2} - p^{-1/2}\boldsymbol{Z}_{p,2}\right\|_{\max} \\
&\leqslant \sqrt{p/2}\left(\left\|\boldsymbol{Q}_{p,1} - p^{-1/2}\boldsymbol{Z}_{p,1}\right\| + \left\|p^{-1/2}\boldsymbol{Z}_{p,1}^T - \boldsymbol{Q}_{p,1}^T\right\|_{\max}\right) + \\
&\quad \sqrt{p}\left\|\boldsymbol{Q}_{p,2} - p^{-1/2}\boldsymbol{Z}_{p,2}\right\|_{\max} \\
&= 2\sqrt{p/2}\left\|\boldsymbol{Q}_{p,1} - p^{-1/2}\boldsymbol{Z}_{p,1}\right\|_{\max} + \sqrt{p}\left\|\boldsymbol{Q}_{p,2} - p^{-1/2}\boldsymbol{Z}_{p,2}\right\|_{\max} \\
&= \sqrt{2}\left\|\sqrt{p}\boldsymbol{Q}_{p,1} - \boldsymbol{Z}_{p,1}\right\|_{\max} + \left\|\sqrt{p}\boldsymbol{Q}_{p,2} - \boldsymbol{Z}_{p,2}\right\|_{\max} \\
&\leqslant \sqrt{2}\left\|\sqrt{p}\boldsymbol{Q}_p - \boldsymbol{Z}_p\right\|_{\max} + \left\|\sqrt{p}\boldsymbol{Q}_p - \boldsymbol{Z}_p\right\|_{\max} \\
&= (\sqrt{2}+1)\left\|\sqrt{p}\boldsymbol{Q}_p - \boldsymbol{Z}_p\right\|_{\max}.
\end{aligned}$$

Theorem 3 of Jiang (2006) implies that this upper bound goes to zero in probability as p grows. Therefore, the quantity $\|\boldsymbol{\Pi}_p \widetilde{\boldsymbol{\varphi}}_p - \boldsymbol{z}_p\|_\infty$ does as well.

A.2 Special matrices

In constructing the linear transformations given in Lemmas 2 and 6, 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 Magnus and Neudecker (1979), while the matrix $\widetilde{\boldsymbol{D}}_n$ was introduced in Neudecker (1983). Our presentation follows that of Magnus (1988).

A.2.1 The commutation matrix $\boldsymbol{K}_{m,n}$

Let $\boldsymbol{A}$ 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 Magnus (1988) states that $\boldsymbol{K}_{p,m}(\boldsymbol{A} \otimes \boldsymbol{B}) = (\boldsymbol{B} \otimes \boldsymbol{A})\boldsymbol{K}_{q,n}$. Section 3.3 of Magnus (1988) 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 Magnus (1988) tells us that

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

A.2.2 The matrix $\widetilde{\boldsymbol{D}}_n$

The matrix $\widetilde{\boldsymbol{D}}_n$ proves useful when differentiating expressions involving skew-symmetric matrices. Let $\boldsymbol{A}$ 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 $\operatorname{vec} \boldsymbol{A}$. Definition 6.1 of Magnus (1988) defines $\widetilde{\boldsymbol{D}}_n$ as the unique $n^2 \times n(n-1)/2$ matrix with the property $\widetilde{\boldsymbol{D}}_n \tilde{v}(A) = \operatorname{vec} \boldsymbol{A}$ for every skew-symmetric matrix $\boldsymbol{A}$. Theorem 6.1 of Magnus (1988) gives an explicit expression for $\widetilde{\boldsymbol{D}}_n$. Let $\boldsymbol{E}_{i,j}$ be the $n \times n$ matrix with 1 in the i,jth position and zeros everywhere else and set $\tilde{\boldsymbol{T}}_{i,j} = \boldsymbol{E}_{i,j} - \boldsymbol{E}_{j,i}$. 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

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

A.3 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

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

where

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

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

$$\boldsymbol{C}_{11} = (\boldsymbol{I}_k - \boldsymbol{B} + \boldsymbol{A}^T \boldsymbol{A})^{-1} \qquad\qquad \boldsymbol{C}_{12} = -\boldsymbol{C}_{11} \boldsymbol{A}^T$$

$$\boldsymbol{C}_{21} = \boldsymbol{A}\boldsymbol{C}_{11} \qquad\qquad \boldsymbol{C}_{22} = \boldsymbol{I}_{p-k} - \boldsymbol{A}\boldsymbol{C}_{11}\boldsymbol{A}^T.$$

The blocks of the matrices

$$\boldsymbol{G}_{\mathcal{V}} = \begin{bmatrix} \boldsymbol{G}_{11} & \boldsymbol{G}_{12} \\ \boldsymbol{G}_{21} & \boldsymbol{G}_{22} \end{bmatrix} \quad \boldsymbol{H}_{\mathcal{V}} = \begin{bmatrix} \boldsymbol{H}_{11} & \boldsymbol{H}_{12} \\ \boldsymbol{H}_{21} & \boldsymbol{H}_{22} \end{bmatrix}$$

can be written in terms of the blocks of $\boldsymbol{C}_{\mathcal{V}}$ as

$$\boldsymbol{H}_{11} = \boldsymbol{C}_{11}^T\boldsymbol{C}_{11} + \boldsymbol{C}_{21}^T\boldsymbol{C}_{21} \qquad \boldsymbol{H}_{12} = \boldsymbol{C}_{11}^T\boldsymbol{C}_{12} + \boldsymbol{C}_{21}^T\boldsymbol{C}_{22}$$

$$\boldsymbol{H}_{21} = \boldsymbol{C}_{12}^T\boldsymbol{C}_{11} + \boldsymbol{C}_{22}^T\boldsymbol{C}_{21} \qquad \boldsymbol{H}_{22} = \boldsymbol{C}_{12}^T\boldsymbol{C}_{12} + \boldsymbol{C}_{22}^T\boldsymbol{C}_{22}$$

and

$$\boldsymbol{G}_{11} = \boldsymbol{C}_{11}\boldsymbol{C}_{11}^T \qquad \boldsymbol{G}_{12} = \boldsymbol{C}_{11}\boldsymbol{C}_{21}^T$$

$$\boldsymbol{G}_{21} = \boldsymbol{C}_{21}\boldsymbol{C}_{11}^T \qquad \boldsymbol{G}_{22} = \boldsymbol{C}_{21}\boldsymbol{C}_{21}^T.$$

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

$$\begin{aligned} DC(\boldsymbol{\varphi})^T DC(\boldsymbol{\varphi}) &= 2^2 \boldsymbol{\Gamma}_{\mathcal{V}}^T \left(\boldsymbol{G}_{\mathcal{V}} \otimes \boldsymbol{H}_{\mathcal{V}}\right) \boldsymbol{\Gamma}_{\mathcal{V}} \\ &= 2^2 \begin{bmatrix} \boldsymbol{\Omega}_{11} & \boldsymbol{\Omega}_{12} \\ \boldsymbol{\Omega}_{21} & \boldsymbol{\Omega}_{22} \end{bmatrix} \end{aligned}$$

where

$$\begin{aligned} \boldsymbol{\Omega}_{11} &= \widetilde{\boldsymbol{D}}_k^T(\boldsymbol{G}_{11} \otimes \boldsymbol{H}_{11})\widetilde{\boldsymbol{D}}_k \\ \boldsymbol{\Omega}_{12} &= \widetilde{\boldsymbol{D}}_k^T(\boldsymbol{G}_{11} \otimes \boldsymbol{H}_{12}) - \widetilde{\boldsymbol{D}}_k^T(\boldsymbol{G}_{12} \otimes \boldsymbol{H}_{11})\boldsymbol{K}_{p-k,k} \\ \boldsymbol{\Omega}_{21} &= \boldsymbol{\Omega}_{12}^T \\ \boldsymbol{\Omega}_{22} &= (\boldsymbol{G}_{11} \otimes \boldsymbol{H}_{22} + \boldsymbol{H}_{11} \otimes \boldsymbol{G}_{22}) - (\boldsymbol{G}_{12} \otimes \boldsymbol{H}_{21} + \boldsymbol{H}_{12} \otimes \boldsymbol{G}_{21})\boldsymbol{K}_{p-k,k}. \end{aligned}$$

Then

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

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

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

Appendix B

Appendix to Chapter 4

B.1 Proofs

B.1.1 Proof of Proposition 12

Suppose that $\boldsymbol{X}$ is invariant to left multiplication by elements of $\mathcal{L}$ and right multiplication by elements of $\boldsymbol{R}$, and let $\boldsymbol{L} \in \mathcal{L}$ and $\boldsymbol{R} \in \mathcal{R}$ be given. We have that $\boldsymbol{X} \overset{d}{=} \boldsymbol{LXR}$. Let $\boldsymbol{VDV}^\top$ be the eigendecomposition of $\boldsymbol{X}^\top\boldsymbol{X}$. Then

$$\begin{aligned}
\boldsymbol{Q}_X &\overset{d}{=} \boldsymbol{LXR}(\boldsymbol{R}^\top\boldsymbol{X}^\top\boldsymbol{L}^\top\boldsymbol{LXR})^{-1/2} \\
&= \boldsymbol{LXR}(\boldsymbol{R}^\top\boldsymbol{X}^\top\boldsymbol{XR})^{-1/2} \\
&= \boldsymbol{LXR}(\boldsymbol{R}^\top\boldsymbol{VDV}^\top\boldsymbol{R})^{-1/2} \\
&= \boldsymbol{LXRR}^\top\boldsymbol{VD}^{-1/2}\boldsymbol{V}^\top\boldsymbol{R} \\
&= \boldsymbol{LXVD}^{-1/2}\boldsymbol{V}^\top\boldsymbol{R} \\
&= \boldsymbol{LX}(\boldsymbol{X}^\top\boldsymbol{X})^{-1/2}\boldsymbol{R} \\
&= \boldsymbol{LQ}_X\boldsymbol{R}.
\end{aligned}$$

B.1.2 Proof of Proposition 13

To prove the result, we will show that a Wasserstein distance between the distribution of finitely many entries of $\sqrt{p}\boldsymbol{Q}_{X_p}$ and the distribution of the corresponding entries of $\boldsymbol{X}_p$ goes to zero as $p, k \to \infty$ with $k/p \to 0$. This is sufficient because, by Theorem 6.9 of Villani (2009), the Wasserstein distance metrizes weak convergence. From Villani (2009):

Definition 1. *Let $(\mathcal{X}, d)$ be a Polish metric space and let $\ell \in [1, \infty)$. For any two probability measures μ, ν on $\mathcal{X}$, the Wasserstein distance of order ℓ between μ and ν is defined by the formula*

$$W_\ell(\mu, \nu) = \inf_{X,Y} \left\{ \left[\mathbb{E} d(X,Y)^\ell \right]^{\frac{1}{\ell}} \mid law(X) = \mu,\ law(Y) = \nu \right\}.$$

Let $W_2\left[\text{law}\left(\sqrt{p}\,\boldsymbol{Q}_{X_p}\right), \text{law}\left(\boldsymbol{X}_p\right)\right]^2$ be the squared Wasserstein distance of order two between $\text{law}\left(\sqrt{p}\,\boldsymbol{Q}_{X_p}\right)$ and $\text{law}\left(\boldsymbol{X}_p\right)$ with the metric induced by the Frobenius norm. Then $W_2\left[\text{law}\left(\sqrt{p}\,\boldsymbol{Q}_{X_p}\right), \text{law}\left(\boldsymbol{X}_p\right)\right]^2$ is bounded above by $\mathbb{E}\left\|\sqrt{p}\,\boldsymbol{Q}_{X_p} - \boldsymbol{X}_p\right\|_F^2$. The squared Frobenius norm can be written as

$$\begin{aligned}
\left\|\sqrt{p}\,\boldsymbol{Q}_{X_p} - \boldsymbol{X}_p\right\|_F^2 &= \text{Tr}\left[\left(\sqrt{p}\,\boldsymbol{Q}_{X_p} - \boldsymbol{X}_p\right)^\top \left(\sqrt{p}\,\boldsymbol{Q}_{X_p} - \boldsymbol{X}_p\right)\right] \\
&= \text{Tr}\left[p\,\boldsymbol{Q}_{X_p}^\top \boldsymbol{Q}_{X_p}\right] - 2\,\text{Tr}\left[\sqrt{p}\,\boldsymbol{Q}_{X_p}^\top \boldsymbol{X}_p\right] + \text{Tr}\left[\boldsymbol{X}_p^\top \boldsymbol{X}_p\right].
\end{aligned}$$

We evaluate the three terms and their expectations separately and then combine the results. In terms of $\boldsymbol{Z}_p$ and $\boldsymbol{\Omega}_p$, the matrix $\boldsymbol{Q}_{X_p}$ is given as

$$\begin{aligned}
\boldsymbol{Q}_{X_p} &= \boldsymbol{X}_p(\boldsymbol{X}_p^\top \boldsymbol{X}_p)^{-1/2} \\
&= \boldsymbol{\Omega}_p^{1/2} \boldsymbol{Z}_p (\boldsymbol{Z}_p^\top \boldsymbol{\Omega}_p \boldsymbol{Z}_p)^{-1/2}.
\end{aligned}$$

Thus, the first term is

$$\text{Tr}\left[p\,\boldsymbol{Q}_{X_p}^\top \boldsymbol{Q}_{X_p}\right] = p\,\text{Tr}\left[(\boldsymbol{Z}_p^\top \boldsymbol{\Omega}_p \boldsymbol{Z}_p)^{-1/2} \boldsymbol{Z}_p^\top \boldsymbol{\Omega}_p^{1/2} \boldsymbol{\Omega}_p^{1/2} \boldsymbol{Z}_p (\boldsymbol{Z}_p^\top \boldsymbol{\Omega}_p \boldsymbol{Z}_p)^{-1/2}\right]$$

$$= p\,\mathrm{Tr}\left[(\boldsymbol{Z}_p^\top \boldsymbol{\Omega}_p \boldsymbol{Z}_p)^{-1} \boldsymbol{Z}_p^\top \boldsymbol{\Omega}_p \boldsymbol{Z}_p\right]$$

$$= p\,\mathrm{Tr}\left[\boldsymbol{I}_{\boldsymbol{k_p}}\right]$$

$$= pk_p.$$

Then, of course, we have $\mathbb{E}\,\mathrm{Tr}\left[p\,\boldsymbol{Q}_{X_p}^\top \boldsymbol{Q}_{X_p}\right] = pk_p$. The third term is $\mathrm{Tr}\left[\boldsymbol{X}_p^\top \boldsymbol{X}_p\right] = \mathrm{Tr}\left[\boldsymbol{Z}_p^\top \boldsymbol{\Omega}_p \boldsymbol{Z}_p\right]$. Let $\boldsymbol{z}_j$ denote the jth column of $\boldsymbol{Z}_p$. Then

$$\begin{aligned}
\mathbb{E}\,\mathrm{Tr}\left[\boldsymbol{Z}_p^\top \boldsymbol{\Omega}_p \boldsymbol{Z}_p\right] &= \mathbb{E}\left[\sum_{j=1}^{k_p} \boldsymbol{z}_j^\top \boldsymbol{\Omega}_p \boldsymbol{z}_j\right] \\
&= k_p \mathbb{E}\left[\sum_{j=1}^{k_p} \boldsymbol{z}_1^\top \boldsymbol{\Omega}_p \boldsymbol{z}_1\right] \\
&= k_p\,\mathrm{Tr}\,\boldsymbol{\Omega}_p \\
&= pk_p.
\end{aligned}$$

The second term is

$$\begin{aligned}
-2\,\mathrm{Tr}\left[\sqrt{p}\,\boldsymbol{Q}_{X_p}^\top \boldsymbol{X}_p\right] &= -2\,\mathrm{Tr}\left[\sqrt{p}\left(\boldsymbol{Z}_p^\top \boldsymbol{\Omega}_p \boldsymbol{Z}_p\right)^{-1/2} \boldsymbol{Z}_p^\top \boldsymbol{\Omega}_p^{1/2} \boldsymbol{\Omega}_p^{1/2} \boldsymbol{Z}_p\right] \\
&= -2\,\mathrm{Tr}\left[\left(p^{-1} \boldsymbol{Z}_p^\top \boldsymbol{\Omega}_p \boldsymbol{Z}_p\right)^{-1/2} \boldsymbol{Z}_p^\top \boldsymbol{\Omega}_p \boldsymbol{Z}_p\right] \\
&= -2p\,\mathrm{Tr}\left[\left(p^{-1} \boldsymbol{Z}_p^\top \boldsymbol{\Omega}_p \boldsymbol{Z}_p\right)^{-1/2} p^{-1} \boldsymbol{Z}_p^\top \boldsymbol{\Omega}_p \boldsymbol{Z}_p\right] \\
&= -2p\,\mathrm{Tr}\left[\left(p^{-1} \boldsymbol{Z}_p^\top \boldsymbol{\Omega}_p \boldsymbol{Z}_p\right)^{1/2}\right] \\
&= -2p\,\mathrm{Tr}\left[\boldsymbol{S}_p^{1/2}\right]
\end{aligned}$$

where $\boldsymbol{S}_p = p^{-1} \boldsymbol{Z}_p^\top \boldsymbol{\Omega}_p \boldsymbol{Z}_p$. Let $\lambda_1 \geqslant \ldots \geqslant \lambda_{k_p} \geqslant 0$ be the eigenvalues of $\boldsymbol{S}_p$. Putting this all together, we have that

$$\begin{aligned}
\mathbb{E}\left\|\sqrt{p}\,\boldsymbol{Q}_{X_p} - \boldsymbol{X}_p\right\|_F^2 &= 2pk_p - 2p\mathbb{E}\,\mathrm{Tr}\left[\boldsymbol{S}_p^{1/2}\right] \\
&= 2p\mathbb{E}\,\mathrm{Tr}\left[\boldsymbol{I}_{k_p} - \boldsymbol{S}_p^{1/2}\right] \\
&= 2p\mathbb{E}\left[\sum_{j=1}^{k_p}(1 - \lambda_j^{1/2})\right]
\end{aligned}$$

$$\leqslant 2p\mathbb{E}\left[\sum_{j=1}^{k_p}|1-\lambda_j^{1/2}|\right]$$

$$\leqslant 2p\mathbb{E}\left[\sum_{j=1}^{k_p}|1-\lambda_j|\right]$$

$$\leqslant 2pk_p\mathbb{E}\left[\max_{1\leqslant j\leqslant k_p}|1-\lambda_j|\right]$$

$$= 2pk_p\mathbb{E}\left\|\boldsymbol{S}_p-\boldsymbol{I}_{k_p}\right\|$$

where $\left\|\boldsymbol{S}_p-\boldsymbol{I}_{k_p}\right\|$ denotes the spectral norm of $\boldsymbol{S}_p-\boldsymbol{I}_{k_p}$. From Chen and Pan (2012), we know $\mathbb{E}\left\|\boldsymbol{S}_p-\boldsymbol{I}_{k_p}\right\|=\mathcal{O}(\sqrt{k/p})$.

For each p, let $\boldsymbol{q}_p=\left(q_1^{(p)},...,q_m^{(p)}\right)^\top$ be a vector of m entries of $\boldsymbol{Q}_{X_p}$ corresponding to a fixed set of indices. Let $\boldsymbol{x}_p=\left(x_1^{(p)},...,x_m^{(p)}\right)^\top$ be the vector of corresponding entries of $\boldsymbol{X}_p$. Let $W_2\left[\text{law}\left(\sqrt{p}\,\boldsymbol{q}_p\right),\text{law}\left(\boldsymbol{x}_p\right)\right]^2$ be the squared Wasserstein distance of order two between law $\left(\sqrt{p}\,\boldsymbol{q}_p\right)$ and law $(\boldsymbol{x}_p)$ with the metric induced by the Euclidean norm. When $\boldsymbol{\Omega}_p=\boldsymbol{I}_p$, we have

$$\begin{aligned}W_2\left[\text{law}\left(\sqrt{p}\,\boldsymbol{q}_p\right),\text{law}\left(\boldsymbol{x}_p\right)\right]^2 &\leqslant \mathbb{E}\left\|\sqrt{p}\,\boldsymbol{q}_p-\boldsymbol{x}_p\right\|^2\\ &= \frac{m}{pk_p}\mathbb{E}\left\|\sqrt{p}\,\boldsymbol{Q}_{X_p}-\boldsymbol{X}_p\right\|_F^2.\end{aligned}$$

Thus, $W_2\left[\text{law}\left(\sqrt{p}\,\boldsymbol{q}_p\right),\text{law}\left(\boldsymbol{x}_p\right)\right]^2=\mathcal{O}\left(\sqrt{k_[/p}\right)$ and $W_2\left[\text{law}\left(\sqrt{p}\,\boldsymbol{q}_p\right),\text{law}\left(\boldsymbol{x}_p\right)\right]=\mathcal{O}\left((k_p/p)^{1/4}\right)$. The Wasserstein distance between law $\left(\sqrt{p}\,\boldsymbol{q}_p\right)$ and law $(\boldsymbol{x}_p)$ goes to zero, and we can conclude that law $\left(\sqrt{p}\,\boldsymbol{q}_p\right)$ converges weakly to law $(\boldsymbol{x}_p)$.

Bibliography

Adler, R. (1981), *The Geometry of Random Fields*, Wiley, Chichester.

Albert, J. H. and Chib, S. (1993), "Bayesian Analysis of Binary and Polychotomous Response Data," *Journal of the American Statistical Association*, 88, 669–679.

Anderson, T. W., Olkin, I., and Underhill, L. G. (1987), "Generation of Random Orthogonal Matrices," *SIAM Journal on Scientific and Statistical Computing*, 8, 625–629.

Borel, É. (1906), "Sur les principes de la théorie cinétique des gaz," *Annales scientifiques de l'École Normale Supérieure*, 23, 9–32.

Bou-Rabee, N. and Sanz-Serna, J. M. (2017), "Randomized Hamiltonian Monte Carlo," *Annals of Applied Probability*, 27, 2159–2194.

Brubaker, M. A., Salzmann, M., and Urtasun, R. (2012), "A family of MCMC methods on implicitly defined manifolds," in *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, eds. N. D. Lawrence and M. Girolami, vol. 22, pp. 161–172, Proceedings of Machine Learning Research.

Butland, G., Peregrín-Alvarez, J. M., Li, J., Yang, W., Yang, X., Canadien, V., Starostine, A., Richards, D., Beattie, B., Krogan, N., Davey, M., Parkinson, J., Greenblatt, J., and Emili, A. (2005), "Interaction network containing conserved and essential protein complexes in Escherichia coli," *Nature*, pp. 531–537.

Byrne, S. and Girolami, M. (2013), "Geodesic Monte Carlo on embedded manifolds," *Scandinavian Journal of Statistics*, 40, 825–845.

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," *Journal of Statistical Software*, 76, 1–32.

Cayley, A. (1846), "Sur quelques propriétés des déterminants gauches," *Journal für die reine und angewandte Mathematik*, 32, 119–123.

Chatterjee, S. and Meckes, E. (2008), "Multivariate normal approximation using exchangeable pairs," *Latin American Journal of Probability and Mathematical Statistics*, 4, 257–283.

Chen, B. and Pan, G. (2012), "Convergence of the largest eigenvalue of normalized sample covariance matrices when p and n both tend to infinity with their ratio converging to zero," *Bernoulli*, 18, 1405–1420.

Chikuse, Y. (2003), *Statistics on Special Manifolds*, Springer New York.

Cook, R. D., Li, B., and Chiaromonte, F. (2010), "Envelope models for parsimonious and efficient multivariate linear regression," *Statistica Sinica*, 20, 927–960.

Cron, A. and West, M. (2016), "Models of Random Sparse Eigenmatrices and Bayesian Analysis of Multivariate Structure," in *Statistical Analysis for High-Dimensional Data*, pp. 125–153, Springer International Publishing.

D'Aristotile, A., Diaconis, P., and Newman, C. M. (2003), "Brownian motion and the classical groups," *Lecture Notes-Monograph Series*, 41, 97–116.

Dawid, A. P. (1981), "Some matrix-variate distribution theory: Notational considerations and a Bayesian application," *Biometrika*, 68, 265–274.

Diaconis, P. and Forrester, P. J. (2017), "Hurwitz and the origins of random matrix theory in mathematics," *Random Matrices: Theory and Applications*, 6.

Diaconis, P. and Freedman, D. (1987), "A dozen de Finetti-style results in search of a theory," *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 23, 397–423.

Diaconis, P. and Shahshahani, M. (1994), "On the Eigenvalues of Random Matrices," *Journal of Applied Probability*, 31, 49–62.

Diaconis, P., Eaton, M., and Lauritzen, S. (1992), "Finite De Finetti Theorems in Linear Models and Multivariate Analysis," *Scandinavian Journal of Statistics*, 19, 289–315.

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*, pp. 102–125, Institute of Mathematical Statistics.

Donoho, D. and Gavish, M. (2014), "Minimax risk of matrix denoising by singular value thresholding," *The Annals of Statistics*, 42, 2413–2440.

Duane, S., Kennedy, A., Pendleton, B. J., and Roweth, D. (1987), "Hybrid Monte Carlo," *Physics Letters B*, 195, 216–222.

Durmus, A., Moulines, E., and Saksman, E. (2017), "On the convergence of Hamiltonian Monte Carlo," *arXiv:1705.00166*.

Eaton, M. L. (1983), *Multivariate Statistics: a Vector Space Approach*, John Wiley & Sons.

Eaton, M. L. (1989), "Group Invariance Applications in Statistics," in *Regional Conference Series in Probability and Statistics*, vol. 1, pp. i–133, Institute of Mathematical Statistics.

Eckart, C. and Young, G. (1936), "The approximation of one matrix by another of lower rank," *Pyschometrika*, 1, 211–218.

Ferrari, F. (2019), "Bayesian Factor Analysis for Inference on Interactions," *arXiv:1904.11603*.

Flegal, J. M., Hughes, J., Vats, D., and Dai, N. (2017), "mcmcse: Monte Carlo Standard Errors for MCMC," .

Fosdick, B. K. and Raftery, A. E. (2012), "Estimating the Correlation in Bivariate Normal Data With Known Variances and Small Sample Sizes," *The American Statistician*, 66, 34–41.

Gao, C. and Zhou, H. H. (2015), "Rate-optimal posterior contraction for sparse PCA," *Ann. Statist.*, 43, 785–818.

Griffin, J. E. and Brown, P. J. (2010), "Inference with normal-gamma prior distributions in regression problems," *Bayesian Analysis*, 5, 171–188.

Hoff, P. (2009a), *A First Course in Bayesian Statistical Methods*, Springer.

Hoff, P. D. (2007), "Model Averaging and Dimension Selection for the Singular Value Decomposition," *Journal of the American Statistical Association*, 102, 674–685.

Hoff, P. D. (2009b), "Simulation of the matrix Bingham-von Mises-Fisher Distribution, With Applications to Multivariate and Relational Data," *Journal of Computational and Graphical Statistics*, 18, 438–456.

Hubbard, J. H. and Hubbard, B. B. (2009), *Vector Calculus, Linear Algebra, and Differential Forms: A Unified Approach*, Matrix Editions.

Hurwitz, A. (1897), "Über die Erzeugung der invarianten durch integration," *Nachrichten von der Gesellschaft der Wissenschaften zu Göttingen, Mathematisch-Physikalische Klasse*, 1897, 71–90.

James, A. T. (1954), "Normal Multivariate Analysis and the Orthogonal Group," *Annals of Mathematical Statistics*, 25, 40–75.

Jauch, M., Hoff, P. D., and Dunson, D. B. (2018), "Random orthogonal matrices and the Cayley transform," *arXiv:1810.02881*.

Jiang, T. (2006), "How many entries of a typical orthogonal matrix can be approximated by independent normals?" *Annals of Probability*, 34, 1497–1529.

Jiang, T. and Ma, Y. (2017), "Distances between random orthogonal matrices and independent normals," *arXiv:1704.05205v1*.

Johansson, K. (1997), "On Random Matrices from the Compact Classical Groups," *Annals of Mathematics*, 145, 519–545.

Johnstone, I. M. (2001), "On the distribution of the largest eigenvalue in principal components analysis," *The Annals of Statistics*, 29, 295–327.

Johnstone, I. M. and Lu, A. Y. (2009), "On Consistency and Sparsity for Principal Components Analysis in High DimensionsNo Title," *Journal of the American Statistical Association*, 104, 682–693.

Kent, J. T., Ganeiber, A. M., and Mardia, K. V. (2013), "A new method to simulate the Bingham and related distributions in directional data analysis with applications," *arXiv:1310.8110v1*.

León, C. A., Massé, J.-C., and Rivest, L.-P. (2006), "A Statistical Model for Random Rotations," *Journal of Multivariate Analysis*, 97, 412–430.

Liu, C., Rubin, D. B., and Wu, Y. N. (1998), "Parameter expansion to accelerate EM: The PX-EM algorithm," *Biometrika*, 85, 755–770.

Liu, J. S. and Wu, Y. N. (1999), "Parameter Expansion for Data Augmentation," *Journal of the American Statistical Association*, 94, 1264–1274.

Livingstone, S., Betancourt, M., Byrne, S., and Girolami, M. (2018), "On the geometric ergodicity of Hamiltonian Monte Carlo," *arXiv:1601.08057v4*.

Magnus, J. R. (1988), *Linear Structures*, Oxford University Press.

Magnus, J. R. and Neudecker, H. (1979), "The Commutation Matrix: Some Properties and Applications," *The Annals of Statistics*, 7, 381–394.

Magnus, J. R. and Neudecker, H. (1988), *Matrix differential calculus with applications in statistics and econometrics*, Wiley Series in Probability and Mathematical Statistics.

Mardia, K. V. and Jupp, P. E. (2009), *Directional Statistics*, Wiley Series in Probability and Statistics.

Marsaglia, G. (1972), "Choosing a Point from the Surface of a Sphere," *The Annals of Mathematical Statistics*, 43, 645–646.

Maxwell, J. C. (1875), *Theory of Heat*, Longmans, London, 4th edn.

Maxwell, J. C. (1878), "On Boltzmann's theorem on the average distribution of energy in a system of material points," *Transactions of the Cambridge Philosophical Society*, 12, 547–575.

Meckes, E. (2008), "Linear functions on the classical matrix groups," *Transactions of the American Mathematical Society*, 360, 5355–5366.

Mehler, F. G. (1866), "Ueber die Entwicklung einer Function von beliebig vielen. Variablen nach Laplaschen Functionen höherer Ordnung," *Crelle's Journal*, 66, 161–176.

Neal, R. M. (2011), "MCMC Using Hamiltonian Dynamics," in *Handbook of Markov Chain Monte Carlo*, pp. 113–162, CRC Press.

Neudecker, H. (1983), "On Jacobians of transformations with skew-symmetric, strictly (lower) triangular or diagonal matrix arguments," *Linear and Multilinear Algebra*, 14, 271–295.

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," *arXiv:1710.09443v2*.

Rains, E. M. (1997), "High powers of random elements of compact Lie groups," *Probability Theory and Related Fields*, 107, 219–241.

Ramsay, J. and Silverman, B. (1997), *Functional Data Analysis*, Springer-Verlag, New York.

Ramsay, J., Wickham, H., Graves, S., and Hooker, G. (2018), "fda: Functional Data Analysis," .

Rao, V., Lin, L., and Dunson, D. B. (2016), "Data augmentation for models based on rejection sampling," *Biometrika*, 103, 319–335.

Rasmussen, C. E. and Williams, C. K. I. (2006), *Gaussian processes for Machine learning*, MIT Press.

Salvatier, J., Wiecki, T. V., and Fonnesbeck, C. (2016), "Probabilistic programming in Python using PyMC3." *PeerJ Computer Science*, 2, e55.

Shepard, R., Brozell, S. R., and Gidofalvi, G. (2015), "The Representation and Parametrization of Orthogonal Matrices," *The Journal of Physical Chemistry A*, 119, 7924–7939.

Srivastava, M. S. and Khatri, C. G. (1979), *An Introduction to Multivariate Statistics*, North-Holland/New York.

Stam, A. J. (1982), "Limit theorems for uniform distributions on spheres in high-dimensional Euclidean spaces," *Journal of Applied Probability*, 19, 221–228.

Stan Development Team (2019), *Stan Modeling Language Users Guide and Reference Manual, Version 2.19.0*.

Stein, C. (1995), "The accuracy of the normal approximation to the distribution of the traces of powers of random orthogonal matrices," Tech. Rep. No. 470, Stanford University Department of Statistics.

Stewart, K. (2018), "Total variation approximation of random orthogonal matrices by Gaussian matrices," *arXiv:1704.06641v2*.

Suarez, A. J. and Ghosal, S. (2017), "Bayesian Estimation of Principal Components for Functional Data," *Bayesian Analysis*, 12, 311–333.

Team, R. C. (2019), "R: A Language and Environment for Statistical Computing," .

Traynor, T. (1993), "Change of variable for Hausdorff measure," *Rendiconti dell'Istituto di Matematica dell'Università di Trieste*, 1, 327–347.

Van Dyk, D. A. and Meng, X.-L. (2001), "The Art of Data Augmentation," *Journal of Computational and Graphical Statistics*, 10, 1–50.

van Handel, R. (2016), "Probability in high dimension," Tech. rep., Princeton University.

Villani, C. (2009), *Optimal transport: old and new*, Springer-Verlag.

Watson, G. S. (1983), "Limit theorems on high dimensional spheres and Stiefel manifolds," in *Studies in Econometrics, Time Series, and Multivariate Statistics*, pp. 559–570, Academic Press.

Witten, D. M., Tibshirani, R., and Hastie, T. (2009), "A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis," *Biostatistics*, 10, 515–534.

Yoshida, R. and West, M. (2010), "Bayesian learning in sparse graphical factor models via annealed entropy," *Journal of Machine Learning Research*, 11, 1771–1798.