Duke
UNIVERSITY

# Bayesian analysis with orthogonal matrix parameters

by

Michael Jauch

June 27, 2018

# Contents

# Introduction

# The parameter space

Our parameter space will be the set of $p \times k$ orthogonal matrices

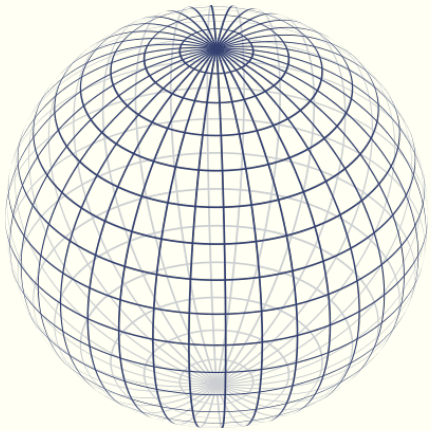
$$\mathcal{V}_{k,p} = \{\boldsymbol{Q} \in \mathbb{R}^{p \times k} : \boldsymbol{Q}^T \boldsymbol{Q} = \boldsymbol{I}_k\}$$

often referred to as the Stiefel manifold.

# The parameter space

The Stiefel manifold has two notable special cases:

- $\mathcal{V}_{p,p}$ is equivalent to the orthogonal group $\mathcal{O}_p$.
- $\mathcal{V}_{1,p}$ is equivalent to the the unit sphere $\mathcal{S}^{p-1}$ in $\mathbb{R}^p$.



$\mathcal{V}_{1,3}$ is the unit sphere in $\mathbb{R}^3$

# The parameter space

In a Bayesian setting, we are concerned with prior and posterior distributions on our parameter space.

When $\mathcal{V}_{k,p}$ is our parameter space, prior and posterior distributions will typically be defined in terms of densities with respect to the uniform probability measure $\mathcal{U}$ on $\mathcal{V}_{k,p}$.

This is the unique probability measure with the property that if $\boldsymbol{Q} \in \mathcal{V}_{k,p}$ is distributed according to $\mathcal{U}$, then $\boldsymbol{UQV} \stackrel{d}{=} \boldsymbol{Q}$ for all $\boldsymbol{U} \in \mathcal{O}_p$ and $\boldsymbol{V} \in \mathcal{O}_k$.

The uniform measure $\mathcal{U}$ on $\mathcal{V}_{k,p}$ will play a role analagous to that of Lebesgue measure on Euclidean space.

# When does this come up?

Models for multivariate data are often naturally parametrized in terms of orthogonal matrix parameters. Two common settings:

- Matrix/tensor decompositions
- Dimension reduction

There are many examples in the literature: Hoff [2007, 2009a,b, 2016], Franks and Hoff [2016], Cron and West [2016], Yoshida and West [2010], Johnstone [2001], Yang and Berger [1994], Cunningham and Ghahramani [2015], Cook et al. [2010], Khare et al. [2017], etc.

# Examples: Spiked covariance model

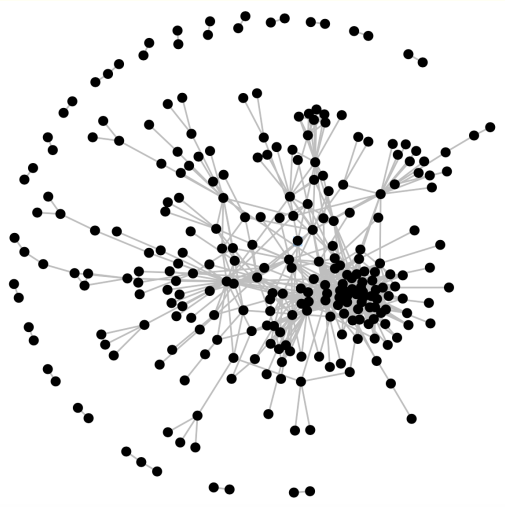
As a simple matrix decomposition example, consider the following 'spiked covariance' model for data $\boldsymbol{y}_1, ..., \boldsymbol{y}_N \in \mathbb{R}^p$ :

$$\begin{aligned}
\boldsymbol{y}_i &\overset{iid}{\sim} \mathrm{N}(\boldsymbol{0}, \boldsymbol{\Sigma}) \\
\boldsymbol{\Sigma} &= \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^T + \sigma^2 \boldsymbol{I}_p \\
\boldsymbol{Q} &\in \mathcal{V}_{k,p} \\
\boldsymbol{\Lambda} &= \mathrm{diag}(\lambda_1, ..., \lambda_k) \\
\lambda_1 &> ... > \lambda_k > 0.
\end{aligned}$$

The covariance matrix is modeled as the sum of a low-rank component, represented by its eigendecomposition, and a scaled identity matrix.

# Examples: Network eigenmodel

Hoff [2009a] describes a model for the pairwise interactions of 270 proteins of *E. Coli*. For each pair of proteins $i$ and $j$, we observe a binary variable $y_{ij}$ with $y_{ij} = 1$ if the pair interacts and $y_{ij} = 0$ otherwise.



# Examples: Network eigenmodel

Let $\boldsymbol{Q} \in \mathcal{V}_{3,270}$, $\boldsymbol{\Lambda} = \text{diag}(\lambda_1, \lambda_2, \lambda_3)$, and $\Phi$ be the CDF of a standard normal. The model specification is as follows:

$$\begin{aligned} \Pr(y_{ij} = 1) &= \Phi[c + (\boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^T)_{ij}] \\ \boldsymbol{Q} &\sim \mathcal{U} \\ \lambda_l &\overset{iid}{\sim} \mathcal{N}(0, 270) \\ c &\sim \mathcal{N}(0, 10^2). \end{aligned}$$

We'll return to this model later.

# Examples: Covariance estimation

In Yang and Berger [1994], the authors discuss a default prior for a (full rank) covariance matrix $\boldsymbol{\Sigma} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^T$ in a multivariate normal model.

They argue for their prior over the Jeffreys prior, in part, by inspecting the implied (improper) prior on the eigenvalues of the covariance matrix:

$$\pi_R(\boldsymbol{Q}, \boldsymbol{\Lambda}) \propto \frac{1}{|\boldsymbol{\Lambda}|}$$

vs.

$$\pi_J(\boldsymbol{Q}, \boldsymbol{\Lambda}) \propto \frac{\prod_{i<j}(\lambda_i - \lambda_j)}{|\boldsymbol{\Lambda}|^{(p+1)/2}}.$$

# Motivation

The motivation for parametrizing a model in terms of orthogonal matrices is often

- that the resulting parameters correspond to something fundamental in the problem of interest
- that the resulting parametrization is identifiable
- to make our prior assumptions regarding the identifiable quantities more transparent
- to open another avenue for constructing prior distributions

# Challenges

Bayesian analysis with orthogonal matrix parameters presents two major challenges:

- posterior sampling on the constrained parameter space
- incorporation of prior information, such as sparsity

This talk will address these two challenges.

# Challenges: Sampling from $\mathcal{V}_{k,p}$

How do people do MC or MCMC sampling from the Stiefel manifold?

**Rejection sampling** is common in the directional statistics literature. **Cons**: These methods are tailored for specific distributions and perform badly in even moderate dimensions and other important regimes.

**Gibbs samplers** are available in some cases, e.g. Hoff [2009a]. **Cons:** These methods work for problems of moderate dimension but are only applicable when the full conditional distributions belong to certain parametric families. They typically rely upon rejection samplers and inherit some of their problems.

# Challenges: Sampling from $\mathcal{V}_{k,p}$

**Gradient based approaches** such as Byrne and Girolami [2013] don't require conjugacy and scale to reasonably large problems.
**Cons:** These algorithms are somewhat tricky to implement and involve tuning parameters which are critical to the perfomance of the sampler.

# Challenges: Sampling from $\mathcal{V}_{k,p}$

We would like to come up with a method to simulate from a smooth density on the Stiefel manifold which

- is simple to implement in standard statistical software (e.g. Stan)
- does not require conjugacy
- scales to handle problems of realistic size

# Challenges: Prior information

It can be challenging to incorporate prior information regarding a parameter $\boldsymbol{Q} \in \mathcal{V}_{k,p}$.

In the second part of this talk, we propose a method applicable when $\boldsymbol{Q}$ is tall and skinny of constructing prior distributions having element-wise marginal distributions approximately matching (up to rescaling) a desired distribution.

We are particularly interested in how our proposed method can incorporate the prior information that $\boldsymbol{Q}$ is 'nearly sparse.'

# Polar expansion

# Overview

Suppose we want to simulate the random orthogonal matrix $\boldsymbol{Q} \in \mathcal{V}_{k,p}$ having density $g_{\boldsymbol{Q}}$ (e.g. $\boldsymbol{Q}$ is a parameter of interest and $g_{\boldsymbol{Q}}$ is its posterior density).

Using the Jacobian associated with the polar decomposition, we can write down a density $g_{\boldsymbol{X}}$ for a random matrix $\boldsymbol{X} \in \mathbb{R}^{p \times k}$ such $\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1/2} \stackrel{d}{=} \boldsymbol{Q}$.

Since $\boldsymbol{X}$ is unconstrained, we can use familiar MCMC approaches to sample from $g_{\boldsymbol{X}}$ and then transform the resulting samples to simulate from $g_{\boldsymbol{Q}}$.

# The polar decomposition

We can uniquely decompose a full rank matrix $\boldsymbol{X} \in \mathbb{R}^{p \times k}$ as the product of an orthogonal matrix and a symmetric positive definite matrix $\boldsymbol{X} = \boldsymbol{Q}\boldsymbol{S}^{1/2}$ where

$$\boldsymbol{Q} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1/2}$$
$$\boldsymbol{S} = \boldsymbol{X}^T\boldsymbol{X}.$$

For the sake of building intuition, we mention that matrix $\boldsymbol{Q}$ satisfies

$$\boldsymbol{Q} = \underset{\boldsymbol{Q}' \in \mathcal{V}_{k,p}}{\arg\min} \|\boldsymbol{Q}' - \boldsymbol{X}\|_F.$$

# The trick

Given $\boldsymbol{Q} \in \mathcal{V}_{k,p}$ having density $g_{\boldsymbol{Q}}$, we can introduce an auxiliary random $k \times k$ s.p.d. matrix $\boldsymbol{S}$ and denote the joint density by $g_{\boldsymbol{Q},\boldsymbol{S}}$.

The density $g_{\boldsymbol{X}}$ of $\boldsymbol{X} = \boldsymbol{Q}\boldsymbol{S}^{1/2}$ is

$$g_{\boldsymbol{X}}(\boldsymbol{X}) = \frac{\Gamma_k(p/2)}{\pi^{pk/2}} |\boldsymbol{S}_{\boldsymbol{X}}|^{-(p-k-1)/2} g_{\boldsymbol{Q},\boldsymbol{S}}(\boldsymbol{Q}_{\boldsymbol{X}}, \boldsymbol{S}_{\boldsymbol{X}})$$

where

$$\boldsymbol{Q}_{\boldsymbol{X}} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1/2}$$
$$\boldsymbol{S}_{\boldsymbol{X}} = \boldsymbol{X}^T\boldsymbol{X}$$

and $\Gamma_k$ is the multivariate gamma function [Chikuse, 2003].

But how do we choose the distribution of $\boldsymbol{S}$?

# The Wishart case

We focus on the case in which $\boldsymbol{S} \perp \boldsymbol{Q}$ and $\boldsymbol{S} \sim \mathrm{W}_p(\boldsymbol{I})$.

Then

$$g_{\boldsymbol{X}}(\boldsymbol{X}) = (2\pi)^{-pk/2}\, g_{\boldsymbol{Q}}\left(\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1/2}\right) \operatorname{etr}\left(-\frac{1}{2}\boldsymbol{X}^T\boldsymbol{X}\right).$$

In particular, when $g_{\boldsymbol{Q}}$ is uniform then $\operatorname{vec}\boldsymbol{X} \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{I}_k \otimes \boldsymbol{I}_p)$.

If $g_{\boldsymbol{Q}}$ is smooth then

$$g^{\epsilon}_{\boldsymbol{X}}(\boldsymbol{X}) = (2\pi)^{-pk/2}\, g_{\boldsymbol{Q}}\left(\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X} + \epsilon\boldsymbol{I})^{-1/2}\right) \operatorname{etr}\left(-\frac{1}{2}\boldsymbol{X}^T\boldsymbol{X}\right)$$

is smooth on all of $\mathbb{R}^{p\times k}$ and $g^{\epsilon}_{\boldsymbol{X}} \to g_{\boldsymbol{X}}$ pointwise as $\epsilon \to 0$.

# Application to sampling

To simulate from $g_{\boldsymbol{Q}}$, one can simulate

$$\boldsymbol{X}^{(1)}, \ldots, \boldsymbol{X}^{(T)}$$

from $g_{\boldsymbol{X}}$ and then compute

$$\boldsymbol{Q}_{\boldsymbol{X}^{(1)}}, \ldots, \boldsymbol{Q}_{\boldsymbol{X}^{(T)}}.$$

Since $g_{\boldsymbol{X}}$ is defined on $\mathbb{R}^{p \times k}$ rather than the Stiefel manifold, we can use familiar MCMC approaches.

A general and relatively hassle-free approach is to use Hamiltonian Monte Carlo [Neal, 2010] as implemented in software such as Stan [Carpenter et al., 2017].

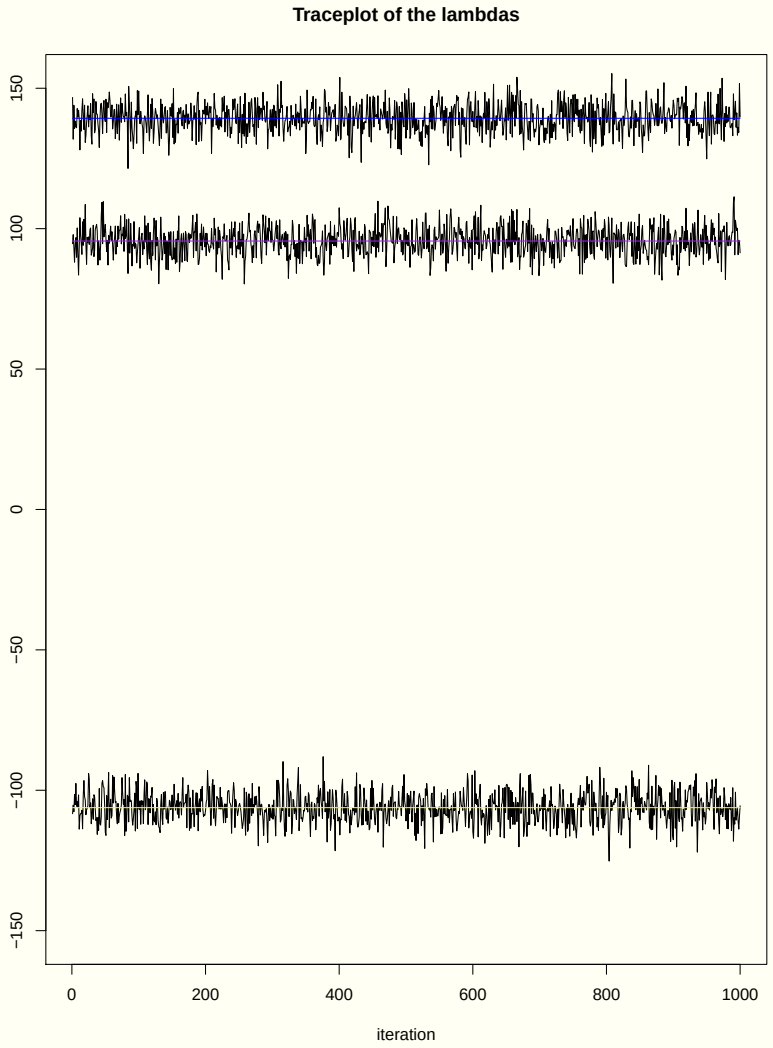
# The network eigenmodel revisited

Hoff [2009a] and Byrne and Girolami [2013] introduce methodology for sampling from the Stiefel manifold and apply it to the problem of posterior sampling for the network eigenmodel described in the introduction.

We have done posterior sampling for the same model using polar expansion with $\boldsymbol{S} \perp \boldsymbol{Q}$ and $\boldsymbol{S} \sim \mathrm{W}_p(\boldsymbol{I})$, which has several advantages over the existing approaches:

- It can be implemented quite simply in Stan.
- We are free to modify the model without worrying about losing the required conjugacy.
- The samples show less auto-correlation than with the other approaches.

# The network eigenmodel revisited



# Goals related to polar expansion

The recent work of Durmus et al. [2017] provides sufficient conditions for a Markov chain arising from HMC to be geometrically ergodic.

To further justify our general approach and our choice $\boldsymbol{S} \perp \boldsymbol{Q}$ and $\boldsymbol{S} \sim \mathrm{W}_p(\boldsymbol{I})$, we'd like to use these results establish the geometric ergodicity of HMC targeting $g^{\epsilon}_{\boldsymbol{X}}$ for a broad class of densities $g_{\boldsymbol{Q}}$.

# Approx. marginal matching prior distributions

# Overview

What if we want to want to incorportate prior knowledge regarding the marginal distributions of the entries of $\boldsymbol{Q} \in \mathcal{V}_{k,p}$?

For example, we might want a prior distribution for $\boldsymbol{Q}$ such that each entry $q_{ij}$ is distributed according to some conventional 'sparsity favoring' prior.

This topic hasn't been addressed in the literature.

We present a practical approach to constructing prior distributions for $\boldsymbol{Q} \in \mathcal{V}_{k,p}$ in the case that $\boldsymbol{Q}$ is tall and skinny for which the elementwise marginals approximately match (up to rescaling) a desired distribution.

# The motivating observation

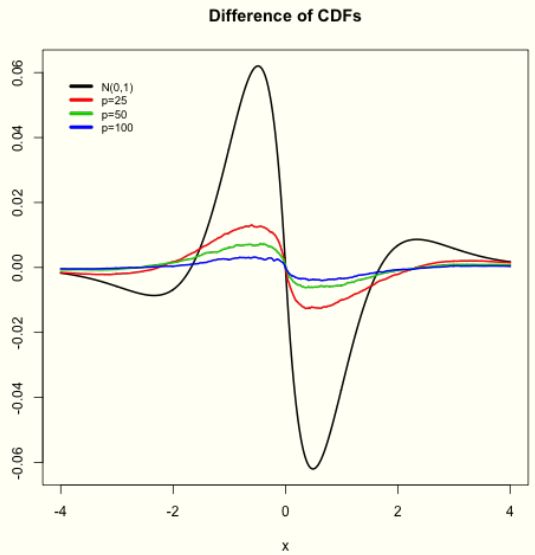
Suppose we have a tall and skinny random matrix $\boldsymbol{X} = (x_{ij}) \in \mathbb{R}^{p \times k}$ such that $x_{ij} \overset{iid}{\sim} \mu$ with $\mathbb{E}[x_{ij}] = 0$ and $\mathbb{V}[x_{ij}] = 1$.

If we set $\boldsymbol{Q} = (q_{ij}) = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{1/2}$, it appears that $\sqrt{p}\, q_{ij} \sim \mu$ is a very good approximation.

# The motivating observation

For example, suppose we fix $k = 3$, set $\mu = \text{Laplace}(0, \sqrt{1/2})$, and consider the marginal distribution of $\sqrt{p}\, q_{11}$ for $p \in \{25, 50, 100\}$.

We plot the difference between the empirical C.D.F. of $\sqrt{p}\, q_{11}$ and the C.D.F. of the Laplace distribution for each value of $p$ :



# An explanation

**Proposition**

*Let* $k \in \mathbb{N}$ *be fixed. For each* $p \in \mathbb{N}$, *let* $\boldsymbol{X}^{(p)} = \left(x_{ij}^{(p)}\right) \in \mathbb{R}^{p \times k}$ *be a random matrix such that* $x_{ij}^{(p)} \overset{iid}{\sim} \mu$ *with* $\mathbb{E}\left[x_{ij}^{(p)}\right] = 0$ *and* $\mathbb{V}\left[x_{ij}^{(p)}\right] = 1$. *Define* $\boldsymbol{Q}^{(p)} = \left(q_{ij}^{(p)}\right) \in \mathcal{V}_{k,p}$ *as* $\boldsymbol{Q}^{(p)} = \boldsymbol{X}^{(p)} \left(\boldsymbol{X}^{(p)T} \boldsymbol{X}^{(p)}\right)^{-1/2}$. *For any* $i \in \mathbb{N}$ *and* $j \in \{1, ..., k\}$, *the sequence* $\left\{\sqrt{p} q_{ij}^{(p)}\right\}_{p>i}$ *converges to* $\mu$ *in distribution.*

# Posterior sampling with an AMM

Suppose we observe data $\boldsymbol{y}$ generated according to $p(\boldsymbol{y}|\boldsymbol{Q})$. Let $p(\boldsymbol{X})$ be the density of $\boldsymbol{X}$ and $p(\boldsymbol{Q})$ be the induced AMM prior on $\boldsymbol{Q}$.

We sample $\boldsymbol{X}^{(1)}, ..., \boldsymbol{X}^{(T)}$ from the density proportional to

$$p(\boldsymbol{y}|\boldsymbol{Q}_{\boldsymbol{X}})p(\boldsymbol{X})$$

where again $\boldsymbol{Q}_{\boldsymbol{X}} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1/2}$.

Then $\boldsymbol{Q}_{\boldsymbol{X}^{(1)}}, ..., \boldsymbol{Q}_{\boldsymbol{X}^{(T)}}$ are samples from the density

$$p(\boldsymbol{Q}|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{Q})p(\boldsymbol{Q}).$$

# Applications

In many matrix decomposition and dimension reduction applications it would be useful to incorporate prior information that $\boldsymbol{Q}$ is 'nearly sparse.'

In this case, we have many possible $\mu$ to choose from in the literature on continuous shrinkage priors.

There are a few papers which involve priors with exact zeros for orthogonal matrices [Cron and West, 2016, Yoshida and West, 2010, Gao and Zhou, 2015], but their approaches are quite different from ours.

# Conclusion

# Conclusion

We've presented methodology to address two challenges related to Bayesian analysis with orthogonal matrix parameters:

- We introduced a parameter expansion scheme based on the polar decomposition which allows for relatively easy posterior inference for a wide class of models.
- We introduced a new class of prior distributions which allow one to incorporate prior information regarding the marginal distributions of the entries of a tall and skinny orthogonal matrix parameter.

Thanks!

# References

S. Byrne and M. Girolami. Geodesic monte carlo on embedded manifolds. *Scandinavian Journal of Statistics*, 40(4):825–845, 2013.

B. Carpenter, A. Gelman, M. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. Brubaker, J. Guo, P. Li, and A. Riddell. Stan: A probabilistic programming language. *Journal of Statistical Software, Articles*, 76(1):1–32, 2017.

Y. Chikuse. *Statistics on Special Manifolds*. Springer, February 2003.

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

A. J. Cron and M. West. Models of random sparse eigenmatrices matrices and bayesian analysis of multivariate structure. In *Statistical Analysis for High Dimensional Data*, pages 123–154. Springer International Publishing, Switzerland, 2016.

J. P. Cunningham and Z. Ghahramani. Linear dimensionality reduction: Survey, insights, and generalizations. *Journal of Machine Learning Research*, 16:2859–2900, 2015.

A. Durmus, E. Moulines, and E. Saksman. On the convergence of hamiltonian monte carlo, 2017.

A. Franks and P. Hoff. Shared subspace models for multi-group covariance estimation. 2016.

C. Gao and H. H. Zhou. Rate-optimal posterior contraction for sparse pca. *Ann. Statist.*, 43(2):785–818, 04 2015.

P. D. Hoff. Model averaging and dimension selection for the singular value decomposition. *Journal of the American Statistical Association*, 102(478):674–685, 2007.

P. D. Hoff. Simulation of the matrix Bingham-von Mises-Fisher distribution, with applications to multivariate and relational data. *Journal of Computational and Graphical Statistics*, 18(2): 438–456, 2009a.

P. D. Hoff. A hierarchical eigenmodel for pooled covariance estimation. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 71(5):971–992, 2009b.

P. D. Hoff. Equivariant and scale-free tucker decomposition models. *Bayesian Anal.*, 11(3):627–648, 09 2016.

I. M. Johnstone. On the distribution of the largest eigenvalue in principal components analysis. *Ann. Statist.*, 29(2):295–327, 04 2001.

K. Khare, S. Pal, and Z. Su. A bayesian approach for envelope models. *Ann. Statist.*, 45(1):196–222, 02 2017.

R. M. Neal. MCMC using Hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, 54:113–162, 2010.

R. Yang and J. O. Berger. Estimation of a covariance matrix using the reference prior. *Ann. Statist.*, 22(3):1195–1211, 09 1994.

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