University of California
Santa Barbara

Bayesian Inference on the Stiefel Manifold: Models, Applications and Algorithms

A dissertation submitted in partial satisfaction
of the requirements for the degree

Doctor of Philosophy
in
Statistics and Applied Probability

by

Fanqi Meng

Committee in charge:

Professor Alexander Franks, Chair
Professor Alexander Shkolnik
Professor Sang-Yun Oh

September 2021

The Dissertation of Fanqi Meng is approved.

Professor Alexander Shkolnik

Professor Sang-Yun Oh

Professor Alexander Franks, Committee Chair

September 2021

Bayesian Inference on the Stiefel Manifold: Models, Applications and Algorithms

by

Fanqi Meng

To my academic advisers, family, and friends who
supported me along the journey.

Acknowledgements

I would like to express my deepest appreciation to my academic committee chair, Professor Alexander Franks, who introduced me into the field of Bayesian Statistics on the Stiefel manifold and covariance estimation. I am extremely thankful for his diligence and devotion in academics, encouragements and accommodation on my mental health, as well as financial support during the summer of 2019. This dissertation would not have been possible without the support and nurturing of him. Under his supervision, I have made considerable improvements in critical thinking, model building, software development, as well as academic writing. Through the participation of group laboratory meetings, collaborative discussions, seminars and presentations, I gained abundant experiences in academic research and systematic problem solving. I always feel lucky and honored to be his first Ph.D. student.

I would also like to extend my deepest gratitude to Prof. Alexander Shkolnik, who had introduced me to the financial world and supported me by generously offering multiple valuable resources and discussion sessions. His encouragement and support had made my research smooth, pleasant, and full of productivity. Many thanks to my committee member Prof. Sang-Yun Oh, from whom I learned tremendously about the research process, problem solving skills, and technical knowledge in optimization and graphical models.

I am grateful to Prof. Carter for recruiting me to the department and offering me the fellowship during the first year, and to Prof. Jammalamadaka for his excellent instructions on statistical theory. Meanwhile, I am indebted to Prof. Fouque, Prof. Feldman, and Prof. Ludkovski, Prof. Wang, Prof. Meiring, and Prof. Hsu, for helping me build a strong academic foundation for in-depth studies in advanced statistics.

Furthermore, special thanks to my friends and colleagues for their support and friendship. I am particularly thankful to Ke Wang, Jiajing Zheng, Yi Zheng, Zhipu Zhou, Yuanbo Wang, Joshua Bang, Megan Elcheikhali, Javier Zapata, Andrea Angiuli, Ming Yin, Wenhan Xiong, Tsung-Han Yeh, Yao Xuan, Ruimeng Hu, Yichen Feng.

Last but not least, I want to express my gratitude to my parents and my uncle, who have supported me for the past ten years since undergraduate. Finally, I am extremely grateful to two elderly ladies, my grandmother Shuqin Xue, and my childhood mentor Xiulan Yu. Without their guidance and nurturing since childhood, I would never have imagined obtaining the highest degree in statistics in my life.

Curriculum Vitæ
Fanqi Meng

Education

2021	Ph.D. in Statistics and Applied Probability, University of California, Santa Barbara.
2016	B.S. in Mathematics, The Chinese University of Hong Kong.

Professional Experiences

2016 - 2021	Teaching Assistant, University of California, Santa Barbara.
2020	Quantitative Strategist Internship, BofA Securities (BAML), New York City, NY.
2014	Data Scientist, Cluster Technology Limited, Hong Kong.

Publications

1. *StanStiefel*: Rstan package for Bayesian Statistics on the Stiefel manifold, in preparation.

2. *BayDynCov*: R package for Bayesian dynamic covariance estimation, in preparation.

3. Bayesian Time Series Modeling for Dynamic Covariance Estimation, in preparation.

4. Bayesian Autoregressive Covariance Modeling for Financial Markets and its Implications, in preparation.

Awards

2016	Regents Fellowship, awarded by University of California, Santa Barbara.
2011	Full Mainland Admission Scholarship, awarded by The Chinese University of Hong Kong.

Abstract

Bayesian Inference on the Stiefel Manifold: Models, Applications and Algorithms

by

Fanqi Meng

In finance, it is crucial to use recent data to model the relationship between the companies since the market environment is evolving constantly. In particular, estimating time-varying covariance matrices has been an important topic for both portfolio optimization and risk management. Market measures such as betas for companies, beta dispersion, and market volatility are also closely related to the eigenvectors and eigenvalues of the covariance matrices. The current approaches for dynamic covariance estimation are focused on vector autoregressive processes and have shared parameters for the eigenvalues and eigenvectors. This inevitably introduces dependencies and fails to reveal the relationships between the model parameters. We contribute to the field of time-varying covariance estimation by proposing a Bayesian autoregressive model on the Stiefel manifold for high dimensional data. Our model considers the eigenvalues and eigenvectors separately, and provides a reliable solution to the relationships between the eigenvalues and eigenvectors. To our knowledge, this is the first attempt for an autoregressive time series model on the Stiefel manifold, and it can be extended to a class of models that are widely applicable to datasets in finance, biology, climate changes, etc.

Our Bayesian model involves sampling and inference on the Stiefel manifold, which has been a challenging task. We contribute to Bayesian modeling on the Stiefel manifold by writing a new package using the Stan program. In our *StanStiefel* package, we extend the sampling method in [Jauch et al., 2020], and propose novel parameter inference methods for popular distributions. Our package takes much less time to generate comparable amount of effective samples than the *rstiefel* package, especially in high dimensions.

Contents

Chapter 1

Introduction

Many modern statistical applications involve time-varying covariances. In finance, practitioners need to keep up with the changing financial environment and make constant updates for the portfolio weights and risk measures. In biology, people are interested in how the relationships amongst the metabolites evolve as an individual ages, so as to improve our understanding about age-related diseases. Current time series covariance models either focus too much on vector autoregressive processes, or have shared parameters for modeling the eigenvalues and eigenvectors. This hinders the discovery of true relationships between the model parameters, and makes it an inferior choice for some applications.

We propose the first Bayesian autoregressive model for dynamic covariance estimation on the Stiefel manifold. Our model successfully addresses the high dimensional data issue by assuming a spiked covariance model at each time point, and utilizing information across all time points. It allows separate modeling of the eigenvalues and eigenvectors, which is crucial for some applications where the relationship between eigenvalues and eigenvectors is of great importance. We apply our new Bayesian dynamic covariance model on the S&P500 historical returns dataset. Our model can validate the dynamic nature of market beta, and reveal the relationship between the beta dispersion and market volatility.

The novel model is in fact a general framework, and can be easily extended to different variations. We can generate new models by either changing the autoregressive processes on the eigenvectors, or the eigenvalues. In addition, the model requires sampling on the Stiefel manifold, which is a challenging task in its own right. My work reinforces the foundation of Bayesian statistics on the Stiefel manifold. I summarized the state-of-the-art sampling algorithms and developed the inference algorithms for various popular distributions. These algorithms are wrapped up in the handy *StanStiefel* package. This package outperforms the *rstiefel* package in high dimensions, and serves as a powerful toolbox for more complicated models with orthogonal matrix parameters.

Dissertation Organization

The dissertation consists of three major components. We start with the theoretical part, which involves an exploration of the statistical foundations on the Stiefel manifold. The building blocks of Bayesian statistics are then discussed, including the essential distributions, as well as ways to generate samples from them. In the second component, we propose a Bayesian autoregressive model for dynamic covariance estimation. To our best knowledge, this model is the first attempt for Bayesian autoregressive time series models on the Stiefel manifold, and it can be easily tailored to model datasets in different domains. Lastly, we consider a sophisticated problem motivated by the financial context, and provide our solution via the novel Bayesian dynamic covariance estimation approach.

Chapter 2

Statistics on the Stiefel Manifold and the *StanStiefel* Package

2.1 Introduction

The Stiefel manifold $\mathcal{V}_{p,r}$ is the Riemannian submanifold referring to the collections of unit-length vectors in high-dimensional spaces. It consists of all $p \times r$ orthonormal matrices in $\mathbb{R}^{p\times r}$. The elements of the Stiefel manifold $\mathcal{V}_{p,r}$ are known as r-frames, which is an orthogonal set of r p-dimensional unit-length vectors. The orthogonal constraints between the r vectors can be succinctly expressed as

$$\mathcal{V}_{p,r} := \{Y \in \mathbb{R}^{p\times r} : Y^T Y = I_r\}.$$

The column vectors all have Euclidean norm 1 and they are orthogonal with each other. These constraints reduce the degree of freedom of $\mathcal{V}_{p,r}$ from pr to $pr - \frac{1}{2}r(r+1)$. One special case is the unit hypersphere $\mathcal{V}_{p,1}$ in $\mathbb{R}^p$, where each element corresponds to a particular direction pointed by the unit vector. Another extreme is the orthogonal group, where elements represent rotation matrices. In general, $\mathcal{V}_{p,r}$ can be considered an orientation extending the notion of a direction in $\mathbb{R}^p$.

The Stiefel manifold is drawing more and more attentions as many statistical models can be parameterized in terms of orthogonal matrices. In Chrétien and Guedj (2020), the latent variable matrix is modeled with the orthogonal group. Tan et al. (2019) proposes a stabilized alternating direction method of multipliers (ADMM) solution to solve the sparse PCA problem directly over the Stiefel manifold. This avoids deflation technique and convex relaxations, which usually suffer from approximation errors. Moreover, Yang and Bauwens (2018) develops multivariate state-space models where the latent states follow a conditional matrix Langevin distribution over the Stiefel manifold. Outside of the statistical realm, the Stiefel manifold has been traditionally and consistently emphasized in computer vision and robotics, such as in Turaga et al. (2008) and Lui (2012).

In recent years, there has been a surge in utilizing orthogonality on the parameter matrices for estimating neural networks. This brings up multi-faceted benefits. Bansal et al. (2018) shows that orthogonality improves accuracy and boosts convergence rate for convolutional neural network models, and Cogswell et al. (2015) shows that orthonormality helps with tackling the overfitting problem. In addition, it is shown in Arjovsky et al. (2016) that orthogonal parameters can reduce the ingrained vanishing and exploding gradient problems for recurrent neural network models. The benefits obtained from imposing the orthogonal constraints highlight its power and mark its potentials in the machine learning community.

Besides the above-mentioned developments, Bayesian models over orthogonal parameters have also gained their significance. In particular, Pal et al. (2020) establishes a unified Bayesian framework for inference on the Stiefel manifold with the matrix Langevin distribution, and Lin et al. (2017) takes a generative non-parametric approach which takes advantages of kernel mixtures that can approximate a large class of distributions on the Stiefel manifold. For applications, a stream of thoughts applied the Bayesian technique on covariance estimations with orthogonal matrices parameters, as shown in Hoff (2009a) and Franks and Hoff (2019).

In this chapter we take an overview of the substantial probability distributions on the Stiefel manifold with discussions on their properties. Then we move on investigating the various procedures for sampling from these distributions via different parametrizations. Last but not least, we propose inference algorithms that complete the Bayesian framework. The sampling and inference algorithms are well-packed in the handy *StanStiefel* package.

2.2 Probability Distributions over the Stiefel Manifold

2.2.1 Uniform Distribution

By definition, the uniform distribution assumes identical density values at all elements in $\mathcal{V}_{p,r}$. For $X \in \mathcal{V}_{p,r}$, both left rotations and right rotations give other elements in the same space.

Namely, $QX \in \mathcal{V}_{p,r}$ and $XH \in \mathcal{V}_{p,r}$ for $Q \in \mathcal{O}_{p,p}$ and $H \in \mathcal{O}_{r,r}$. The uniform density ought to reflect this rotational invariance property. Denote the uniform density by f, then it should follow

$$f(X) = f(QX) = f(XH), \quad \forall Q \in \mathcal{O}_{p,p},\ H \in \mathcal{O}_{r,r}, \tag{2.1}$$

which are called the left-invariance and right-invariance.

To be exact, the invariant measure on the Stiefel manifold $\mathcal{V}_{p,r}$ is denoted by $[(dX^T)X]$. Here $X_0(p \times p) = (X \quad X_1)$, $X_0^T X_0 = I_p$, and X_1 is the complement of X that makes X_0 an orthogonal matrix. It has been shown in Gupta and Nagar (2018) that

$$Vol(\mathcal{V}_{p,r}) = \int_{\mathcal{V}_{p,r}} [(dX^T)X] = \frac{2^r \pi^{\frac{pr}{2}}}{\Gamma_r(\frac{p}{2})}. \tag{2.2}$$

Therefore,

$$\frac{1}{Vol(\mathcal{V}_{p,r})}[(dX^T)X] \tag{2.3}$$

defines the probability element of the invariant distribution of random matrix X on $\mathcal{V}_{p,r}$. There are a few more related theorems on the uniform distribution, which can be referred to on page 281 of Gupta and Nagar (2018).

2.2.2 Matrix Von Mises-Fisher Distribution

The matrix von Mises-Fisher distribution is also well-known as the matrix Langevin distribution. The distribution is obtained by imposing the orthogonality constraints on a multivariate normal distribution. Denoted by $L(p, r; F)$, the explicit expression of the density function for random variable X is

$$\frac{1}{a(F)} \operatorname{etr}\left(F^T X\right), \tag{2.4}$$

where F is a $p \times r$ matrix and $a(F)$ the normalizing constant.

In particular, we can write out the singular value decomposition of F as $F = \Gamma\Lambda\Theta^T$, where $\Gamma \in \mathcal{V}_{p,r}, \Theta \in \mathcal{O}(r)$, and $\Lambda = \operatorname{diag}(\{\lambda_1, \dots, \lambda_r\})$, $\lambda_1 \geq \cdots \geq \lambda_r \geq 0$. The λ_i's control the concentrations in the directions pointed by the orientations Γ and Θ. The mode is unique when the $\lambda' s$ are distinct, and is given by $X_0 = \Gamma\Theta^T$. Because

$$\max_X \operatorname{tr}(F^T X) = \operatorname{tr} F^T X_0 = \operatorname{tr} \Lambda = \sum_{i=1}^{n} \lambda_i. \tag{2.5}$$

According to Gupta and Nagar (2018), the normalizing constant $a(F)$ can be evaluated by

$$\begin{aligned} a(F) &= \int_{X \in O(p,r)} \operatorname{etr}\left(F^T X\right) dX \\ &= {}_0F_1\left(\frac{1}{2}p; \frac{1}{4}F^T F\right) \\ &= {}_0F_1\left(\frac{1}{2}p; \frac{1}{4}\Lambda^2\right). \end{aligned} \tag{2.6}$$

Notice that the normalizing constant depends on F only through the singular values Λ, not the principal components Γ or Θ.

The Langevin distribution plays an important role in directional statistics. Its vector version is primarily used in modeling high-dimensional vector dynamics. An interesting probabilistic property is that the first exit point from the $p-1$ dimensional sphere of the drifted Wiener process on $\mathbb{R}^p$ starting from the origin follows a von Mises-Fisher distribution, see Gatto (2013) for more details.

2.2.3 Matrix Bingham Distribution

The Bingham distribution is the analogue on the sphere of the isotropic bivariate normal distribution in the plane. A random matrix X on $\mathcal{V}_{p,r}$ is said to have the matrix Bingham distribution if its density function has the form:

$$\frac{1}{b(G)} \operatorname{etr}\left(X^T G X\right), \tag{2.7}$$

where G is a p by p symmetric matrix. We can write out the spectral decomposition of G as $G = VAV^T$, where $V \in \mathcal{O}(r)$, and $A = \operatorname{diag}\left(\{a_1, \ldots, a_p\}\right)$. The normalizing constant $b(G)$ can be explicitly calculated as

$$\begin{aligned} b(G) &= \int_{X \in O(p,r)} \operatorname{etr}\left(X^T G X\right) dX \\ &= {}_1F_1\left(\frac{1}{2}r; \frac{1}{2}p; G\right). \end{aligned} \tag{2.8}$$

There are a few characteristics marking the specialness of this distribution.

Right-rotational Invariance

The matrix Bingham distribution is invariant under right-orthogonal transformation. Let $X_1 = XH$, for $H \in \mathcal{O}_r$, then

$$\operatorname{tr}(X_1^T G X_1) = \operatorname{tr}(H^T X^T G X H) = \operatorname{tr}(H H^T X^T G X) = \operatorname{tr}(X^T G X), \tag{2.9}$$

since $HH^T = I_r$.

Identifiability

There is identifiability issue with the Bingham distribution since for $A' = A + aI$, we have

$$\begin{aligned} \operatorname{tr}(X^T V A' V^T X) &= \operatorname{tr}(X^T V(A + aI)V^T X) \\ &= \operatorname{tr}(X^T V A V^T X) + \operatorname{tr}(X^T V aI V^T X) \\ &= \operatorname{tr}(X^T G X) + bp \end{aligned}$$

However, bp is a constant and it will be subsumed in the normalizing constant. Therefore the kernels for both distributions are the same, suggesting that the distribution depends on A only through its differences. To ensure its identifiability, we choose the version where $a_1 \geq a_2 \geq ... \geq a_{p-1} \geq a_p = 0$. In case G is low-rank, we can remove the constraint on $a_p = 0$, as it is implied already by the low-rank property.

Antipodal Symmetry

A density $f(x)$ for vector x is considered antipodally symmetric if $f(x) = f(-x)$ for all values $x \in \mathbb{R}^p$. Bingham (1974) discusses in detail this property for the vector Bingham distribution on the sphere. For matrix distributions, by analogy, the definition implies that flipping the signs of any columns of X does not change the density value. As for the matrix Bingham distribution, for any diagonal matrix S with diagonal elements equal to ±1, let $X' = XS$,

$$\begin{aligned} \operatorname{tr}(X'^T G X') &= \operatorname{tr}(S^T X^T G X S) = \operatorname{tr}(X^T G X S S^T) \\ &= \operatorname{tr}(X^T G X). \end{aligned}$$

The last step holds since S is a diagonal matrix with elements ±1. Hence the antipodal symmetry is confirmed, which makes it a good candidate for modelling axes, instead of directions. This property is highly respected in the field of directional statistics, see Prentice (1982) for more details.

Multi-modality

The matrix Bingham distribution is an analogy of the centered normal distribution on the Stiefel manifold, where the mode and concentration are controlled by V and A. Meanwhile, since the signs of the columns can be flipped arbitrarily, there are 2^p modals even when the diagonal elements of A are all distinct.

2.2.4 Generalized Matrix Bingham Distribution

The generalized Bingham distribution introduces an extra parameter H in addition to the usual Bingham parameter G. It has a density of the format

$$p(X|G,H) = \frac{1}{c(H,G)} \operatorname{etr}(HX^TGX), \tag{2.10}$$

when $X \in \mathcal{V}_{p,r}$, G is a $p\times p$ symmetric matrix and H is an $r\times r$ symmetric matrix. Let $G = VAV^T$ and $H = WBW^T$, $V \in \mathcal{O}_p$, $W \in \mathcal{O}_r$ and A and B are diagonal matrices of non-negative values. the density can then be rewritten as

$$p(X|A,B,V,W) \propto \operatorname{etr}(WBW^TX^TVAV^TX) = \operatorname{etr}(B(W^TX^TV)A(V^TXW)). \tag{2.11}$$

Notice that W^TX^TV is the transpose of V^TXW. For any fixed V and W, there is a one-to-one relationship between V^TXW and X on the same space, therefore the density is equivalent to

$$p(X|A,B,V,W) \propto \operatorname{etr}(BX^TAX). \tag{2.12}$$

Identifiability

The density depends on A and B only through the differences between their respective diagonal elements. This can be seen by considering

$$\begin{aligned}\operatorname{tr}\{(B+dI)X^T(A+cI)X\} &= \operatorname{tr}(BX^TAX) + d\operatorname{tr}(X^TAX) + c\operatorname{tr}(BX^TX) + cd\operatorname{tr}(X^TX)\\ &= \operatorname{tr}(BX^TAX) + d\operatorname{tr}(X^TAX) + c\operatorname{tr}(B) + cdr.\end{aligned}$$

$c\operatorname{tr}(B)$ and cdr are constants not associated with X, so they will be subsumed in the normalizing constant. We can see the density depends on $A + cI$ only through the elements of A. Therefore, only the differences amongst the diagonal elements of A matter. A special case is when $p = r$, then $d\operatorname{tr}(X^TAX) = d\operatorname{tr}(AXX^T) = d\operatorname{tr}(A)$, in which case the density depends on both A and B only through the differences amongst the diagonal elements. Furthermore, for any constant $c > 0$, we have

$$\operatorname{etr}((cB)X^T(c^{-1}A)X) = \operatorname{etr}(BX^TAX). \tag{2.13}$$

Hence scaling A and B together while keeping the product won't affect the density value. In light of the above two properties, we choose the version of parameterization when $p > r$:

$$\begin{aligned}&\operatorname{diag}(A) = (a_1, a_2, ..., a_p),\ a_1 \geq a_2 \geq ... \geq a_p = 0,\\ &\operatorname{diag}(B) = (b_1, b_2, ..., b_r),\ b_1 \geq b_2 \geq ... \geq b_r > 0,\\ &\quad a_1 = b_1.\end{aligned}$$

And when $p = r$:

$$\begin{aligned} \operatorname{diag}(A) &= (a_1, a_2, ..., a_p),\ a_1 \geq a_2 \geq ... \geq a_p = 0, \\ \operatorname{diag}(B) &= (b_1, b_2, ..., b_r),\ b_1 \geq b_2 \geq ... \geq b_p = 0, \\ a_1 &= b_1. \end{aligned}$$

Condition for Antipodal Symmetry

The antipodal symmetry is guaranteed for Bingham distribution, but not for the generalized Bingham. Hoff (2009a) provided the necessary and sufficient conditions for a generalized Bingham distribution to have the antipodal symmetry as follows:

Proposition 1. If G and H both have more than one distinct eigenvalue, then a necessary and sufficient condition for the generalized Bingham density 2.10 to be antipodally symmetric in the columns of U is that H be a diagonal matrix.

Parameter Interpretations

Restricting ourselves to the antipodally symmetric generalized Bingham distributions, we can write them as

$$p_B(X|A, B, V) = c(A, B)\,\operatorname{etr}(BX^TVAV^TX), \tag{2.14}$$

based on the above necessary and sufficient conditions. Here A and B are diagonal matrices with $a_1 \geq a_2 \geq ... \geq a_p = 0, b_1 \geq b_2 \geq ... \geq b_r > 0$ and $V \in \mathcal{O}_p$. The diagonal elements of A and B play important roles in controlling the variability of X around V, as well as similarities between eigenvectors. To interpret the parameters, it is easier to write the density in terms of the vector inner products.

$$\operatorname{tr}(BX^TVAV^TX) = \sum_{i=1}^{p}\sum_{j=1}^{r} a_i b_j (v_i^T x_j)^2. \tag{2.15}$$

When a_1 and b_1 are large, the density would be large when $v_1^T x_1$ is close to 1, suggesting that the samples x_1 would center around v_1. Meanwhile, the orthogonality constraint prevents $v_1^T x_2$ and $v_2^T x_1$ from being large since $x_1^T x_2 = 0$ and $v_1^T v_2 = 0$. This sets v_2 and x_2 free so they can be closer when $a_2 b_2$ is large. Other similarities between the vectors can be deduced in the same manner. For more details on the implications on eigenvectors, one can refer to Hoff (2009a).

2.2.5 Matrix Bingham–von Mises–Fisher Distribution

The matrix Bingham-von Mises-Fisher distribution is a flexible probability distribution involving both the linear and quadratic terms. The distribution is the general normal distribution on the Stiefel manifold and the density can be written as

$$p_{BMF}(X|I_p, \Phi, M) = \{K(I_p, \Phi, M)\}^{-1} \operatorname{etr}((X - M)^T \Phi (X - M)). \tag{2.16}$$

The normalizing constant $K(I_p, \Phi, M)$ is complicated and is given in page 286 of Gupta and Nagar (2018). After expanding the parentheses and simplification, the density can be rewritten concisely as

$$p_{BMF}(X|A, B, C) \propto \operatorname{etr}(C^T X + BX^T AX), \tag{2.17}$$

where A and B can be assumed to be symmetric and diagonal matrices, respectively. The matrix von Mises-Fisher Distribution is a special case where either B or A has all 0 diagonal elements, and the generalized matrix Bingham distribution is a special case where C is the zero matrix.

2.3 Literature Review

Generating samples from target distributions are the fundamental building blocks for Bayesian statistics. There has been a plethora of sampling techniques developed for various domains and objective distributions. Some basic techniques are the rejection sampling and the importance sampling, which are only efficient for a special class of problems. In addition, Laplace approximation and variational Bayes methods are designed in the spirit of replacing the target posterior distribution with computationally feasible alternatives. Another prominent stream of thought, which is well-known as Markov chain Monte Carlo (MCMC), is based on constructing a Markov chain with the target distribution as the stationary distribution. The Metropolis-Hastings algorithm and Gibbs sampling both fall into this category. Recently, a sub-class of MCMC methods gains popularity with their abilities to propose long distance moves in the state space and high acceptance rates. Being known as Hamiltonian Monte Carlo (Neal et al. (2011)), the method simulates Hamiltonian dynamics in an augmented parameter space, and the trajectories that are projected back to the original space are retained as samples.

In the unconstrained scalar or vector $\mathbb{R}^p$ domains, the famous Stan program (Gelman et al. (2015)) can be utilized to explore thoroughly most distributions with high efficiency. However, when it comes to sampling on the Stiefel manifold, the unit-length constraints and the orthogonality constraints create extra difficulties in addition to the multivariate sampling problem. In this section, we review some excellent papers on addressing this issue from various perspectives.

2.3.1 Gibbs Sampling

Hoff (2009b) discusses the Gibbs sampling algorithm for sampling from the matrix Bingham-von Mises-Fisher distribution. It starts with some discussions on rejection sampling based on a uniform envelope, which is a feasible method with increasing rejection rate as the dimensions grow. Gibbs sampling schemes corresponding to different distributions were proposed with tweaks to accommodate respective properties. We summarize the most clever techniques in relaxing the tough constraints, which is of great help for related sampling problems.

Removing the Orthogonality Constraint

For $X \in \mathcal{V}_{p,r}$, the columns of X are orthogonal to each other. In order to sample X using MCMC techniques, it would be beneficial to consider sampling the columns of X iteratively, instead of regarding X as a whole variable. Bearing that in mind, we rewrite X as $X = \{X_{[,1]}, X_{[,-1]}\}$. Notice that there are two constraints involved with $X_{[,1]}$, the unit length and the orthogonality. Since $X_{[,-1]}^T X_{[,1]} = 0_{r-1}$, we know $X_{[,1]}$ is in the null space of $X_{[,-1]}$, which we will denote by $N_{p,r-1}$. Since N stands for an orthonormal basis we have $N^T N = I$. There exists $z \in \mathcal{S}_{r-1}$ such that $X_{[,1]} = Nz$, and $z = N^T X_{[,1]}$. The newly-introduced variable, z, now only has to reside on the unit sphere. We then move on considering the posterior distribution of z instead of $X_{[,1]}$.

Removing the Unit Length Constraint

As for the unit length constraint, consider $y \in \mathcal{S}_p$, $\sum_{i=1}^p y_i^2 = 1$. Suppose we are dealing with a simple Bingham distribution with $\Lambda = \text{diag}\{\lambda_1, \lambda_2, ..., \lambda_p\}$, then

$$\begin{aligned} p(y|\Lambda) &= c(\Lambda)\exp(y^T \Lambda y) \\ &\propto \exp(\sum_{i=1}^p \lambda_i y_i^2) \end{aligned}$$

Due to the constraint, we will write $y_p^2 = 1 - \sum_{i=1}^{p-1} y_i^2$. Then the density with respect to Lebesgue measure on $y_1, y_2, ..., y_{p-1}$ becomes

$$p(y|\Lambda) \propto \exp(\sum_{i=1}^p \lambda_i y_i^2)|y_p|^{-1}, \quad y_p^2 = 1 - \sum_{i=1}^{p-1} y_i^2. \tag{2.18}$$

Now instead of sampling $y_1, y_2, ..., y_{p-1}$, we let $\theta = y_1^2$ and $q = \{y_1^2/(1-y_1^2), ..., y_p^2/(1-y_1^2)\}$, so that $\{y_1, y_2, ..., y_p^2\} = \{\theta, (1-\theta)q_{-1}\}$. The reason is to relax the constraints and encourage the mixing of Markov chains. In addition, it is apparent that the signs of $y_i's$ do not affect the density. Hence we would assign the positive and negative signs to $y_i's$ randomly with equal probability.

Removing the Orthogonality Constraints for p by p Orthogonal Matrices

For the technique of removing the orthogonality constraints, there is a problem when the variable is a p by p orthogonal matrix. In that case, conditioning on $X_{[,-1]}$ will result in a one-dimensional vector subspace containing only $X_{[,1]}$ and its reverse direction. Therefore, the samples would only be changing signs of the columns of the starting point. In order to effectively explore the parameter space, we sample two columns at a time and apply a similar procedure. To be specific, suppose the goal is to sample from

$$p(X) \propto \text{etr}(C^T X + BX^T AX), \tag{2.19}$$

where X is a $p \times p$ orthogonal matrix. In this case sampling column 1 and 2 is done in the following way. Let $X = \{X_{[,(1,2)]}, X_{[,-(1,2)]}\}$, and N be the null space of $X_{[,-(1,2)]}$. Then there exists a 2×2 orthogonal matrix Z such that $X_{[,(1,2)]} = NZ$. The density of Z given $X_{[,-(1,2)]}$ is

$$p(Z|X_{[,-(1,2)]}) \propto \text{etr}(\tilde{C}^T Z + \tilde{B} Z^T \tilde{A} Z), \tag{2.20}$$

where $\tilde{C} = N^T C_{[,(1,2)]}$, $\tilde{B} = \text{diag}(\{b_{1,1}, b_{2,2}\})$ and $\tilde{A} = N^T AN$. Notice that Z is a 2×2 orthogonal matrix, we can parametrize it as

$$Z = \begin{pmatrix} \cos(\phi) & s\sin(\phi) \\ \sin(\phi) & -s\cos(\phi) \end{pmatrix} \tag{2.21}$$

for some $\phi \in (0, 2\pi)$ and $s = \pm 1$. The joint density of (ϕ, s) is $p(Z(\phi, s))$, and now it becomes sampling ϕ and s. See Hoff (2009b) for more details.

Gibbs sampling algorithm is one of the most intuitive and natural algorithms for sampling from multivariate distributions. This is the first effective attempt to achieve great sampling performance on this challenging manifold. The main spirit is to simplify the problem by relaxing the constraints and focusing on elementary cases like vectors. However, we see from the implementations that there are many techniques involved and also special cases being treated separately. A more unified and holistic method is still highly desirable. Meanwhile, the efficiency is limited by the essence of Metropolis Hastings algorithms, which is a lot inferior than the modern Hamiltonian Monte Carlo algorithms.

2.3.2 Givens Representation

An orthogonal matrix can be reparametrized based on Givens representations (Shepard et al. (2015)). The idea is to decompose an orthogonal matrix into a sequential product of rotational matrices, which can be represented using the angles with respect to corresponding axes. Pourzanjani et al. (2021) adopts this notion and proposes an innovative sampling algorithm by sampling on the unconstrained angle space. Meanwhile, it also addresses the topological issues that occur during the transformation, providing insights into how topology plays a role in the shift of probability spaces.

Givens Representation

The Givens rotations can be used to zero out individual isolated entries in a matrix. To start with, we consider a 2×2 orthogonal matrix

$$\begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix}$$

applied to a vector $y = [y_1, y_2]^T$. If $y_2 \neq 0$, the rotation matrix that can zero out the second entry of y can be computed via

$$y_1 \sin(\theta) + y_2 \cos(\theta) = 0. \tag{2.22}$$

Hence $\theta = \operatorname{arccot}(-\frac{y_1}{y_2})$. The observation is that we can zero out an entry by applying a rotation matrix on essentially a two dimensional plane. For higher dimensional matrices, in order to zero out the (i, j)th entry, we would consider the plane rotation matrix

$$G_{i,j,\theta} = \begin{bmatrix} I & 0 & 0 & 0 & 0 \\ 0 & \cos(\theta) & 0 & -\sin(\theta) & 0 \\ 0 & 0 & I & 0 & 0 \\ 0 & \sin(\theta) & 0 & \cos(\theta) & 0 \\ 0 & 0 & 0 & 0 & I \end{bmatrix},$$

where the ith row contains $\cos(\theta)$ and $-\sin(\theta)$, and the jth row contains $\sin(\theta)$ and $\cos(\theta)$.

Now we fully understand the idea to zero out a single entry. This step can be applied iteratively to perform the QR-factorization. For any $n \times p$ ($n \geq p$) matrix A, the QR-factorization finds an orthogonal matrix Q and a right-triangular matrix R, such that $A = QR$. Since all the entries below R's diagonal are null, we can create R by eliminating the entries one at a time. Following the notation in Pourzanjani et al. (2021),

$$R = R_{pn}^{-1}(\theta_{pn}) \cdots R_{p,p+1}^{-1}(\theta_{p,p+1}) \cdots R_{1n}^{-1}(\theta_{1n}) \cdots R_{12}^{-1}(\theta_{12})A, \tag{2.23}$$

where $R_{ij}^{-1}(\theta_{ij})$ is the rotation matrix to zero out the ij-th entry in A. Therefore,

$$A = R_{12}(-\theta_{12}) \cdots R_{1n}(-\theta_{1n}) \cdots R_{p,p+1}(-\theta_{p,p+1}) \cdots R_{pn}(-\theta_{pn})R. \tag{2.24}$$

Let $Q = R_{12}(-\theta_{12}) \cdots R_{1n}(-\theta_{1n}) \cdots R_{p,p+1}(-\theta_{p,p+1}) \cdots R_{pn}(-\theta_{pn})$. Now Q is a again an orthogonal matrix since it is a product of orthogonal matrices.

In case the original matrix A has orthogonal and unit-length columns, we actually get a stronger result that R is the first p columns of the $n \times n$ identity matrix. The result can be argued inductively. For the first column, only the first entry is non-zero. Therefore it is constrained to be $a_{11} = \pm 1$. For the second column, suppose the first two elements are (a_{12}, a_{22}). We would have

$$\begin{aligned} a_{12}^2 + a_{22}^2 &= 1 \\ a_{11}a_{12} &= 0. \end{aligned}$$

Again, $a_{12} = 0$ and $a_{22} = \pm 1$. Inductively one can get $a_{pp} = \pm 1$ for all p, while all the other entries are 0. See section 3.3 of Pourzanjani et al. (2021) for more detailed arguments.

Change-of-measure Adjustment

Givens transformation achieves one-to-one mapping between an element in the Stiefel manifold and an element in $[0, 2\pi]^{np - \frac{p(p+1)}{2}}$. As is often the case, we need to account for the distortion of probability measure whenever there is a transformation and adjust accordingly. However, in our context the Givens representation is map from $np - \frac{p(p+1)}{2}$ dimensional space to a space of dimension np. The counterpart of a Jabobian term is non-square and we need to resort to differential forms to find its value. To be exact, let

$$G = R_{12}(\theta_{12}) \cdots R_{1n}(\theta_{1n}) \cdots R_{p,p+1}(\theta_{p,p+1}) \cdots R_{pn}(\theta_{pn}). \tag{2.25}$$

Muirhead (2009) shows that the signed surface element measure is given by the absolute value of the differential form:

$$\bigwedge_{i=1}^{p} \bigwedge_{j=i+1}^{n} G_j^T dA_i, \tag{2.26}$$

where G_j is the jth column of G and dA_i are differential 1-forms representing infinitesimal directions on along the Stiefel manifold. Pourzanjani et al. (2021) provides a succinct formula for this measure adjustment term as:

$$J_{A(\Theta)}(\Theta) = \prod_{i=1}^{p} \prod_{j=i+1}^{n} \cos^{j-i-1} \theta_{ij}, \tag{2.27}$$

and the details can be referred to in the appendix of the paper.

Sampling via Givens representation is a very innovative approach since it cleverly takes advantage of the orthonormal property. The thought of zeroing out individual entries via rotation on an extracted two dimensional space decomposes Q into a product of a sequence of rotation matrices, whose inverses can be easily determined by substituting $-\theta_{ij}$. However, this approach requires the computation of a complicated change-of-measure term, which involves the knowledge of differential forms. In addition, the topology of the Stiefel manifold and the Given representation differ slightly, and that leads to some topological issues, for which the paper introduces some ad-hoc techniques and empirical proof.

2.3.3 Cayley's Transformation

Cayley (1846) introduced the classical result that for any $n \times n$ rotation matrix R, if R does not admit -1 as an eigenvalue, then there is a unique skew-symmetric matrix S, such that

$$R = (I - S)(I + S)^{-1}. \tag{2.28}$$

Here R is called the Cayley transform of S. Beautiful as it is, the result is limited to square orthogonal matrices without -1 as its eigenvalues. Fortunately, Shepard et al. (2015) extends the result to general $m \times n$ orthogonal matrices Q. To be exact, the matrix X is taken as

$$X = \begin{bmatrix} B & -A^T \\ A & 0 \end{bmatrix} \tag{2.29}$$

with $B \in \mathbb{R}^{n \times n}$, $B^T = -B$ and $A \in \mathbb{R}^{(m-n) \times n}$. Q can be represented as

$$\begin{aligned} Q &= (I_m + X)(I_m - X)^{-1} I_{m \times n} \\ &= \begin{bmatrix} I_n - F \\ 2A \end{bmatrix} (I_n + F)^{-1}, \quad F = A^T A - B, \end{aligned}$$

and $I_{m\times n}$ is defined as the matrix having the identity matrix as its top block and the remaining entries zero. Reversely, given an orthogonal $m \times n$ matrix Q, we can find X by computing

$$\begin{aligned} F &= (I_n - Q_1)(I_n + Q_1)^{-1} \\ B &= \frac{1}{2}(F^T - F) \\ A &= \frac{1}{2}Q_2(I_n + F), \end{aligned}$$

where $Q = [Q_1^T, Q_2^T]^T$. Under this definition, the Cayley transform is well-defined for any skew-symmetric matrix X since $I_m - X$ is always invertible.

Based on this improved Cayley transformation for general orthogonal matrices, Jauch et al. (2020b) developed a novel sampling scheme. It first proves the Cayley transform is a continuously differentiable map, then computes the derivative matrix according to

$$\frac{\partial Q}{\partial X_{jk}} = 2(I - X)^{-1}\frac{\partial X}{\partial X_{jk}}(I - X)^{-1}. \tag{2.30}$$

The challenging question is again to compute the change-of-variable adjustment term. Ben-Israel (1999) resolves this problem by introducing the concept of matrix volume. Essentially, for a derivative matrix A, the volume of A is defined as the product of the singular values of A, hence if A is of full column rank,

$$\operatorname{vol} A = \sqrt{\det(A^T A)}. \tag{2.31}$$

It is easier to understand this concept in terms of singular value decomposition. Let $A = UDV^T$,

$$\begin{aligned} \det(A^T A) &= \det(VD^T U^T UDV^T) \\ &= \det(D^T(U^T U)D(V^T V)) \\ &= \det(D^2) \\ &= \prod_{i=1}^{n} d_i^2. \end{aligned}$$

So $\operatorname{vol} A = |\prod_{i=1}^{n} d_i|$. The geometric meaning of singular values characterizes how each dimension is scaled in the transformed space. Therefore, the absolute value of the product represents the volume of the transformed cube from the unit cube, which is exactly the adjustment term for the change of probability measures.

After obtaining the change-of-measure term, we are able to derive the density on the unconstrained space, where the Markov chain will be built on. After all the samples are generated, we simply compute their Cayley's transform to obtain desired samples on the Stiefel manifold.

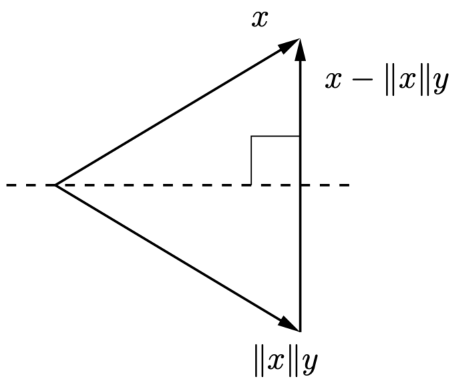
This work is similar to Pourzanjani et al. (2021), which was based on a Givens rotation parametrization of the Stiefel manifold. They both involve reparametrizations of the Stiefel manifold and computations of the change-of-measure adjustment terms. Markov chains will be constructed on the transformed (unconstrained) spaces via Metropolis-Hastings algorithm or more modern Hamiltonian Monte Carlo (HMC) algorithm. Desired samples from the original Stiefel manifold will be generated via corresponding inverse transformations.

2.3.4 Householder Transformation

Nirwan and Bertschinger (2019) tries to resolve the identifiability of the Bayesian principal component analysis models, which are notoriously known for their rotational symmetry. The corresponding posterior distributions possess continuous subspaces of equal density, making it difficult to infer and interpret the parameters. The Householder Transformation is able to eliminate the rotational symmetry in the posterior and bridge the gaps between Bayesian models on the Stiefel manifold and modern sampling softwares, such as Stan.

Householder Transformation

Householder transformations are orthogonal transformations that can introduce zeros into the lower triangle of a matrix, in the same spirits of Gaussian elimination algorithm and the Givens rotations. It is primarily used as a stable way to implement the QR-decomposition. Geometrically, the Householder transformation of a vector x with respect to a unit normal vector y is the reflectional symmetry of x in the direction pointed by y, as shown below.



The line that evenly separates the angle formed by x and y is the axis of symmetry. The unit vector y provides the direction of the transformed result, while the norm of x characterizes its magnitude.

Algebraically, the Householder operator can be expressed as $H = I - 2vv^T$, where $u = x - ||x||y$ and $v = u/||u||$. Notice that

$$\begin{aligned}\left(I - 2vv^T\right) x &= x - 2\frac{(x + \|x\|y)\left(x^T x + \|x\|x^T y\right)}{\|x\|^2 + 2x^T y\|x\| + \|x\|^2\|y\|^2} \\ &= x - (x - \|x\|y) \\ &= \|x\|y.\end{aligned}$$

For a $p \times r$ matrix Z, if all the elements of Z are i.i.d. Gaussian with zero mean and unit variance, the orthogonal matrix Q satisfying $Z = QR$ would be Haar distributed. To implement the QR-decomposition, we follow the procedure indicated by the following theorem, which was summarized in Mezzadri (2006).

Theorem 1 *Let $v_p, v_{p-1}, ..., v_1$ be uniformly distributed on the unit sphere $\mathbb{S}^{p-1}, \mathbb{S}^{p-2}, ..., \mathbb{S}^0$ respectively, where $\mathbb{S}^{p-1}$ is the unit sphere in $\mathbb{R}^p$. Furthermore, let $H_n(v_n)$ be the nth Householder transformation defined as*

$$H_n = \begin{pmatrix} I & 0 \\ 0 & \tilde{H}_n \end{pmatrix},$$

where

$$\tilde{\boldsymbol{H}}_{\boldsymbol{n}}(\boldsymbol{v}_n) = -\operatorname{sgn}(\boldsymbol{v}_{n1})\left(\boldsymbol{I} - 2\boldsymbol{u}_n\boldsymbol{u}_n^T\right) \in \mathbb{R}^{n\times n}$$

and

$$\boldsymbol{u}_n = \frac{\boldsymbol{v}_n + \operatorname{sgn}(\boldsymbol{v}_{n1})\|\boldsymbol{v}_n\|\,\boldsymbol{e}_1}{\|\boldsymbol{v}_n + \operatorname{sgn}(\boldsymbol{v}_{n1})\|\,\boldsymbol{v}_n\,\|\boldsymbol{e}_1\|}.$$

Finally,

$$Q = H_p(v_p)H_{p-1}(v_{p-1})\cdots H_1(v_1)$$

is a random orthogonal matrix with distribution given by the Haar measure on $\mathcal{O}(p)$, and a draw from the Stiefel manifold $\mathcal{V}_{p,r}$ is formulated by multiplying the first r matrices.

The procedure based on Householder transformation avoids the expensive computation of the Jacobian determinant term, which is constantly required for other transformation-based methods. In addition, the Householder parameters v are unconstrained and we won't encounter the dilemma of hitting the boundary of the space, or topological issues arising from mappings between different spaces. Despite the above advantages, it also suffers the inherent combinatorial symmetry, which is akin to the label switching problem in Gaussian mixture models. To resolve that, post-processing needs to be conducted, such as making the first entry of each column of the eigenvector to be positive, or always taking the direction which has acute angles to benchmark vectors.

This method circumvents rejection sampling and variational approximations. When combined with Stan, it extends the realm of Bayesian models on the Stiefel manifold since it does not require conditional conjugacy for the prior distributions. The prior can be flexibly added to the *target* variable in the Stan program. Nirwan and Bertschinger (2019) integrates this idea with the prominent Gaussian process latent variable model (GPLVM). This move opens the door to incorporating orthogonal matrix parameters to modern machine learning models, such as deep neural networks.

2.3.5 Monte Carlo Simulation via Polar Expansion

This ingenious method was proposed in Jauch et al. (2020a), which can be seen as a generalization of the method for simulating from the unit sphere $\mathcal{V}_{p,1}$. It avoids most of the well-established challenges in simulating from the Stiefel manifold. The idea is in the same spirit as Pourzanjani et al. (2021) and Jauch et al. (2020a). To sample random orthogonal matrix $Q \in \mathcal{V}_{p,k}$ $(p \geq k)$, we will construct a Markov chain on the unconstrained random matrix space X, such that Q_X, the orthogonal component of the polar decomposition, is equal in distribution to Q.

To start, we first briefly review how to get samples on $\mathcal{V}_{p,1}$. Based on Muller (1959) and Marsaglia et al. (1972), we can implement the algorithm in the following two steps:

1. Generate N independent standard normal random samples $x_1, x_2, ..., x_N$.

2. Locate a point y on the unit N-sphere by

$$y_i = \frac{x_i}{\sqrt{\sum_{i=1}^N x_i^2}}$$

The point y then will be uniformly distributed on the unit N-sphere. As an intuitive non-rigorous proof, we can consider the joint density of $(x_1, x_2, ..., x_n)$:

$$\begin{aligned} p(x_1, x_2, ..., x_n) &= \prod_{i=1}^{N} p(x_i) \\ &= \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}} e^{-\frac{x_i^2}{2}} \\ &= \left(\frac{1}{\sqrt{2\pi}}\right)^N e^{-\frac{||y||^2}{2}}. \end{aligned}$$

The joint density depends on y only through its magnitude, rather than the directions. Hence at each fixed radius, the density is uniformly distributed on its spherical surface area.

Now for a matrix $X \in \mathbb{R}^{p\times k}$, its singular value decomposition is denoted as $X = UDV^T$, let

$$\begin{aligned} Q_X &= X(X^TX)^{-1/2} = UV^T, \\ S_X &= X^TX = VD^TU^TUDV^T = VD^TDV^T, \\ S_X^{1/2} &= VDV^T. \end{aligned}$$

Then $X = Q_XS_X^{1/2}$ and $S_X = S_X^{1/2}S_X^{1/2}$, where Q_X is an orthogonal matrix while $S_X^{1/2}$ is a symmetric positive definite matrix. Analogous to the polar expansion $z = re^{i\phi}$ for complex numbers, $S_X^{1/2}$ is the counterpart for r while Q_X is comparable to $e^{i\phi}$.

To sample $Q_{p\times k}$ from the Stiefel manifold $\mathcal{V}_{k,p}$, we aim to sample $X_1, X_2, ..., X_N$ from an appropriate density f_X on the unconstrained real matrix $X \in \mathbb{R}^{p\times k}$. For each X_i, we compute its polar expansion $Q_i = X_i(X_i^TX_i)^{-1/2}$, the distribution of Q would match our target distribution and thus $Q_1, Q_2, ..., Q_N$ are samples from the target distribution f_Q. To achieve these, we need to find how f_Q and f_X relate to each other.

The advantage of introducing Q_X and S_x together is that now the mapping from a real, full rank matrix X to the components (Q_X, S_X) of its polar decomposition is one-to-one, and the density f_X can be derived as

$$f_X(X) = f_{S_X|Q_X}(S_X|Q_X)f_{Q_X}(Q_X) \times J(Q_X, S_X; X). \tag{2.32}$$

In contrast with Cayley's transformation and Givens representation, where it is expensive to compute the Jacobian, $J(Q_X, S_X; X)$ is a standard result shown in Chikuse (2012).

$$J\left(Q_X, S_X; X\right) = \frac{\Gamma_k\left(\frac{p}{2}\right)}{\pi^{\frac{pk}{2}}} \left|S_X\right|^{-\frac{p-k-1}{2}}. \tag{2.33}$$

This convenience makes this approach much more attractive than other competitors.

As indicated above, $f_{Q_X}(Q_X)$ would be our target distribution f_Q. Therefore, once the conditional distribution of $f_{S_X|Q_X}$ is determined, we would have a corresponding density on X. It is easily seen that there are various densities $f_X(X)$ that have the margin distribution matching our desired distribution.

As a default choice, Jauch et al. (2020a) recommended $f_{S_X|Q_X}$ to be the density of the Wishart distribution $W_k(p, I_k)$ and it is independent of Q_X. With this choice, the density of the distribution of X simplifies to

$$f_X(X) = \left(\frac{1}{\sqrt{2\pi}}\right)^{pk} \operatorname{etr}(-X^TX/2)f_Q(Q_X). \tag{2.34}$$

In particular, if we consider the problem of sampling uniformly from the Stiefel manifold, $f_Q(Q_X) \propto 1$, then the density of X will be

$$f_X(X) = \left(\frac{1}{\sqrt{2\pi}}\right)^{pk} \operatorname{etr}(-X^T X/2). \tag{2.35}$$

This density shows that all the entries of X are independent standard normal random variables. Notice that this is equivalent to the situation of sampling from the unit sphere. This correspondence motivates the author to select the Wishart distribution as the default choice.

This is by far the most elegant and flexible method for sampling on the Stiefel manifold. It bypasses the obstacles encountered in Pourzanjani et al. (2021), Jauch et al. (2020a) and Nirwan and Bertschinger (2019), including but not limited to topological inconsistency, the expensive computation of super complicated change-of-measure adjustment terms, and sophisticated construction of transformations. It is also flexible for a wide range of distributions, in contrast to the confinements to specific distributions in Hoff (2009b), which involves lots of ad-hoc tweaks to satisfy the various properties on the Stiefel manifold.

2.4 Sampling Algorithms

We consider sampling algorithms for popular distributions on the Stiefel manifold. Fortunately, they can all be derived from 2.34. The corresponding distributions are displayed in the following table.

Distribution	$f_Q(Q_X)$
Uniform Distribution	1
von Mises-Fisher Distribution	$\operatorname{etr}(F^T Q_X)$
Bingham Distribution	$\operatorname{etr}(Q_X^T G Q_X)$
Matrix Bingham-von Mises-Fisher Distribution	$\operatorname{etr}(C^T Q_X + B Q_X^T A Q_X)$
Generalized Bingham Distribution	$\operatorname{etr}(B Q_X^T A Q_X)$

For example, if we want to draw n samples from the Bingham distribution, we substitute $f_Q(Q_X) = \operatorname{etr}(Q_X^T G Q_X)$ in 2.34. Essentially we would be sampling X in

$$f_X(X) = \left(\frac{1}{\sqrt{2\pi}}\right)^{pk} \operatorname{etr}(-X^T X/2) \operatorname{etr}(Q_X^T G Q_X). \tag{2.36}$$

Once we get samples $X_1, X_2, \cdots, X_n$, the orthogonal matrices samples can be obtained by $Q_i =$

$X_i(X_i^T X_i)^{-1/2}$, $i = 1, 2, \cdots, n$. The algorithm can be implemented in Stan directly, and other distributions can be sampled in the same manner.

2.5 Inference Algorithms

2.5.1 Inference for the von Mises-Fisher Distribution

Problem Setup

To infer the parameter F of a von Mises-Fisher distribution, we collect samples $U_1, U_2, ..., U_n$ from the target distribution 2.4. Then the likelihood for F is

$$\begin{aligned} L(F|U_1, U_2, ..., U_n) &= p(U_1, U_2, ..., U_n|F) \\ &= \prod_{i=1}^{n} \frac{1}{a(F)} \operatorname{etr}\left(F^T U_i\right) \\ &= a(F)^{-n} \operatorname{etr}\left(F^T \sum_{i=1}^{n} U_i\right). \end{aligned} \tag{2.37}$$

By 2.6, $a(F) = {}_0F_1\left(\frac{1}{2}p; \frac{1}{4}\Lambda^2\right)$, which is a hypergeometric function of one matrix argument, and Λ is the diagonal matrix of the singular values of F. One should notice that the value of the hypergeometric function depends on F only through its singular values. Therefore, we can write

$$a(\Lambda) = a(F) = {}_0F_1\left(\frac{1}{2}p; \frac{1}{4}\Lambda^2\right).$$

In addition, Butler et al. (2002) provides the approximation result using Laplace approximation. If $X = \operatorname{diag}\{x_1, ..., x_m\}$,

$${}_0\hat{F}_1(n/2; XX^T/4) = R_{0,1}^{-1/2} \prod_{i=1}^{m} \{(1 - \hat{y}_i^2)^{n/2} e^{x_i \hat{y}_i}\}, \tag{2.38}$$

where

$$R_{0,1} = \prod_{i=1}^{m} \prod_{j=i}^{m} (1 - \hat{y}_i^2 \hat{y}_j^2)$$

and $\hat{y}_i = \hat{y}(x_i)$ is given by

$$\hat{y}(x) = u/(\sqrt{(u^2 + 1)} + 1),$$

where $u = 2x/n$.

Inference

Let us get the singular value decomposition of F, $F = \Gamma\Lambda\Theta^T$, both Γ and Θ are orthogonal matrices, and Λ is a diagonal matrix. Inferencing F can be achieved by finding the joint posterior distribution of Γ, Λ and Θ. The posterior distribution is

$$p(\Gamma, \Lambda, \Theta | U_1, ..., U_n) \propto p(\Gamma, \Lambda, \Theta) p(U_1, U_2, ..., U_n | F) \tag{2.39}$$

In most cases, we will assume a uniform prior $p(\Gamma, \Lambda, \Theta) \propto 1$, or one can assume independence and add prior distributions for Γ, Λ, Θ, respectively. By 2.37 the posterior distribution becomes

$$p(\Gamma, \Lambda, \Theta | U_1, ..., U_n) \propto a(\Lambda)^{-n} \operatorname{etr}\left(\Theta\Lambda\Gamma^T S\right), \tag{2.40}$$

where $S = \sum_{i=1}^{n} U_i$.

Direct inferences in this joint matrix parameter space is challenging but fortunately we can utilize the MCMC technique, which converts the inference problem into a few sampling problems. The conditional distributions can be easily obtained,

$$p(\Gamma | \Theta, \Lambda, S) \propto \operatorname{etr}((S\Theta\Lambda)^T \Gamma) \tag{2.41}$$

$$p(\Theta | \Gamma, \Lambda, S) \propto \operatorname{etr}((S^T \Gamma \Lambda^T)^T \Theta) \tag{2.42}$$

$$p(\Lambda | \Theta, \Gamma, S) \propto a(\Lambda)^{-n} \operatorname{etr}(\Gamma^T S \Theta \Lambda) \tag{2.43}$$

The first two are von-Mises Fisher distributions on Γ and Θ, whose samples can be generated by our sampling algorithm. The conditional distribution of Λ can be considered as a multivariate distribution of r scalars, which can be efficiently explored by the Stan software.

Algorithm

Algorithm 1: Bayesian Algorithm for von-Mises Fisher Parameter

```
Result: Bayesian samples of $F, \Gamma, \Lambda, \Theta$.
Initialization: initialize $\Gamma, \Lambda, \Theta$ ;
for $i$ in 1 : Iterations do
    Sample $\Gamma$ from $p(\Gamma|\Theta, \Lambda, S)$;
    Sample $\Theta$ from $p(\Theta|\Gamma, \Lambda, S)$;
    Sample $\Lambda$ from $p(\Lambda|\Theta, \Gamma, S)$;
    Compute $F$ by $F = \Gamma\Lambda\Theta^T$.
end
```

2.5.2 Inference for the Bingham Distribution

Problem Setup

To infer the parameter G of a matrix Bingham distribution, we collect samples $U_1, U_2, ..., U_n$ from the target distribution 2.7. Then the likelihood for G is

$$\begin{aligned} L(G|U_1, U_2, ..., U_n) &= p(U_1, U_2, ..., U_n|G) \\ &= \prod_{i=1}^{n} \frac{1}{b(G)} \operatorname{etr}\left(U_i^T G U_i\right) \\ &= b(G)^{-n} \operatorname{etr}\left(G \sum_{i=1}^{n} U_i U_i^T\right). \end{aligned} \tag{2.44}$$

By 2.8, $b(G) = {}_1F_1\left(\frac{1}{2}r; \frac{1}{2}p; G\right)$, which is a hypergeometric function of one matrix argument. Since G is a positive definite matrix, there exist V and Λ such that $G = V\Lambda V^T$. Similar to the normalizing constant of von-Mises Fisher distribution, $b(G) = b(\Lambda) = {}_1F_1\left(\frac{1}{2}r; \frac{1}{2}p; \Lambda\right)$, where Λ is the diagonal matrix containing the eigenvalues of G. Butler et al. (2002) provides the approximation result based on Laplace approximation. Let $X = \operatorname{diag}\{x_1, x_2, ..., x_p\}$, the raw Laplace approximation is given by

$${}_1\tilde{F}_1(a; b; X) = 2^{p/2}\pi^{p(p+1)/4} B_p(a, b-a)^{-1} J_{1,1}^{-1/2} \prod_{i=1}^{p} \left\{\hat{y}_i^a \left(1-\hat{y}_i\right)^{b-a} e^{x_i \hat{y}_i}\right\},$$

where

$$J_{1,1} = \prod_{i=1}^{p} \prod_{j=i}^{p} \{a(1-\hat{y}_i)(1-\hat{y}_j) + (b-a)\hat{y}_i\hat{y}_j\}.$$

The calibrated approximation ${}_1\hat{F}_1(a;b;X)$ is given by

$$\begin{aligned} {}_1\hat{F}_1(a;b;X) &= \frac{{}_1\tilde{F}_1(a;b;X)}{{}_1\tilde{F}_1\left(a,b;0_p\right)} \\ &= b^{bp-p(p+1)/4} R_{1,1}^{-1/2} \prod_{i=1}^{p} \left\{ \left(\frac{\hat{y}_i}{a}\right)^a \left(\frac{1-\hat{y}_i}{b-a}\right)^{b-a} e^{x_i \hat{y}_i} \right\} \\ R_{1,1} &= \prod_{i=1}^{p}\prod_{j=i}^{p} \left\{ \frac{\hat{y}_i\hat{y}_j}{a} + \frac{(1-\hat{y}_i)(1-\hat{y}_j)}{b-a} \right\}, \end{aligned}$$

where

$$\hat{y}(x) = \frac{2a}{b-x+\sqrt{(x-b)^2+4ax}}.$$

Inference

For a Bingham distribution, G is a positive definite matrix, and we represent $G = V\Lambda V^T$ as before, where V is an orthogonal matrix and Λ is a diagonal matrix consists of all the eigenvalues. It suffices to obtain estimates of V and Λ separately to get estimates of G. The posterior distribution is

$$p(V,\Lambda|U_1,...,U_n) \propto p(V,\Lambda)p(U_1,U_2,...,U_n|G). \tag{2.45}$$

We might choose independent priors on V and Λ or just use the uniform prior. Suppose the uniform prior is adopted and let $S = \sum_{i=1}^n U_i U_i^T$, the posterior distribution is now

$$p(V,\Lambda|U_1,...,U_n) \propto b(\Lambda)^{-n}\,\mathrm{etr}\left(V\Lambda V^T S\right). \tag{2.46}$$

It is difficult to sample V and Λ simultaneously since the Bingham distribution is highly irregular with multiple modes. Nevertheless, multiple modes may corresponds to the same parameter G as one can always flip the signs of the columns of V. Here we apply Gibbs sampling to decompose into conditional posterior distributions for V and Λ separately.

$$\begin{aligned} p(V|\Lambda,U_1,...,U_n) &\propto \mathrm{etr}\left(\Lambda V^T S V\right), \\ p(\Lambda|V,U_1,...,U_n) &\propto b(\Lambda)^{-n}\,\mathrm{etr}\left(\Lambda V^T S V\right). \end{aligned}$$

The first conditional distribution is a generalized Bingham distribution, whereas the second one can be simplified to a sophisticated multivariate likelihood, which can be explored by Stan.

Algorithm

Algorithm 2: Bayesian Algorithm for Bingham Parameters

```
Result: Bayesian samples of $V, \Lambda, G$.
Initialization: initialize $V, \Lambda, G$ ;
for i in 1 : Iterations do
    Sample $V$ from $p(V|\Lambda, S)$;
    Sample $\Theta$ from $p(\Lambda|V, S)$;
    Compute $G$ by $G = V\Lambda V^T$.
end
```

Special Case

There is a special case for Bingham distribution with full rank. The above sampling method works for random variables of dimension $p \times r$ where $r < p$. When $p = r$, the Bingham distribution degenerates to the uniform distribution on the space. Because of full rank, we have $U^TU = UU^T = I_p$. Therefore,

$$\begin{aligned} p(U|G) &= \frac{1}{b(G)} \operatorname{etr}(U^TGU) \\ &= \frac{1}{b(G)} \operatorname{etr}(GUU^T) \\ &= \frac{1}{b(G)} \operatorname{etr}(G) \\ &= \frac{1}{b(\Lambda)} \operatorname{etr}(V\Lambda V^T) \\ &= \frac{1}{b(\Lambda)} \operatorname{etr}(\Lambda V^T V) \\ &= \frac{1}{b(\Lambda)} \operatorname{etr}(\Lambda) \end{aligned}$$

We can see this holds for all values U on the space, hence this is actually the uniform distribution. If we want to have a full rank distribution of the Bingham type, it has to be the generalized Bingham distribution, which is to be discussed in the next section.

2.5.3 Inference for the Generalized Bingham Distribution

Problem Setup

To infer the parameters A, B and V for a generalized matrix Bingham distribution, we collect samples $U_1, U_2, ..., U_n$ from the target distribution

$$U_i \sim \frac{1}{c(A,B)} \operatorname{etr}(BU_i^T(VAV^T)U_i) \tag{2.47}$$

Since we have two groups of parameters, A, B and V, hence we consider Gibbs sampling. The conditional likelihood for A, B is

$$\begin{aligned} L(A,B|V,U_1,U_2,\cdots,U_n) &= p(U_1,U_2,\cdots,U_n|A,B,V) \\ &= \prod_{i=1}^{n} \frac{1}{c(A,B)} \operatorname{etr}\left(BU_i^T VAV^T U_i\right) \\ &= c(A,B)^{-n} \operatorname{etr}\left(\sum_{i=1}^{n} BU_i^T VAV^T U_i\right). \\ &= c(A,B)^{-n} \exp\left(\sum_{i=1}^{r}\sum_{j=1}^{r} a_i b_j (\sum_{k=1}^{n} X_{ij}^{(k)^2})\right), \end{aligned} \tag{2.48}$$

where $X^{(k)} = V^T U_k$. And the conditional distribution for V is

$$\begin{aligned} L(V|A,B,U_1,U_2,\cdots,U_n) &= p(U_1,U_2,\cdots,U_n|A,B,V) \\ &= \prod_{i=1}^{n} \frac{1}{c(A,B)} \operatorname{etr}\left(BU_i^T VAV^T U_i\right) \\ &\propto \operatorname{etr}\left(AV^T(\sum_{i=1}^{n} U_i BU_i^T)V\right). \end{aligned} \tag{2.49}$$

This is a generalized Bingham distribution, from which we can already sample. Therefore we should focus on devising a method for sampling A, B from 2.48.

Inference

In order to estimate A and B, we need to find an adequate numerical approximation of $c(A, B)$. According to corollary 2.1 in Constantine and Muirhead (1976):

If R_1 and S are $k \times k$ and $m \times m$ diagonal matrices respectively, $k \leq m$, with unequal elements ordered in descending order, then

$$\int_{V(k,m)} \exp(\mathrm{tr}(1/2)n\ R_1 H_1^T S H_1)(dH_1)$$
$$\sim 2^k \exp\left((1/2)n\sum_{i=1}^{k} r_i s_i\right)\prod_{i<j}^{k}(\frac{2\pi}{nc_{ij}})^{1/2}\prod_{i=1}^{k}\prod_{j=k+1}^{m}(\frac{2\pi}{nd_{ij}})^{1/2},$$

where $c_{ij} = (r_i - r_j)(s_i - s_j)$ and $d_{ij} = r_i(s_i - s_j)$ for $i = 1, 2, \cdots, k$ and $j = k+1, \cdots, m$. $V(k, m)$ is the Stiefel manifold consisting of all $m \times k$ matrices H_1 with orthonormal columns.

In the above corollary, we take $n = 2, m = p, k = r, R_1 = B$, the first r diagonal elements of S to be A, and the last $p - r$ elements to 0. We obtain a good approximation of $c(A, B)$ as

$$2^r \pi^{\frac{2pr-r(r+1)}{4}} \exp\left(\sum_{i=1}^{r} a_i b_i\right)\prod_{i<j}^{r}(a_i - a_j)^{-1/2}(b_i - b_j)^{-1/2}\prod_{i=1}^{r}(a_i b_i)^{\frac{r-p}{2}}. \tag{2.50}$$

Notice that as parameters of a generalized Bingham distribution, A and B are non-identifiable under some transformations. As mentioned in Hoff (2009a), the likelihood $p(A, B|U_i's)$ behaves the same as that with $p(kA, \frac{1}{k}B|U_i's)$ for $k > 0$. Meanwhile, $p(A + cI, B + dI|U_i's)$ gets a density proportional to that with A and B, and that suggests only the differences amongst the diagonal elements matter. Taking these properties into consideration, we reparametrize A and B as:

$$\mathrm{diag}(A) = (a_1, \ldots, a_r) = \sqrt{w}\,(\alpha_1, \ldots, \alpha_r) \tag{2.51}$$
$$\mathrm{diag}(B) = (b_1, \ldots, b_r) = \sqrt{w}\,(\beta_1, \ldots, \beta_r)\,, \tag{2.52}$$

where $w > 0, 1 = \alpha_1 > \alpha_2 > \cdots > \alpha_{r-1} > \alpha_r > 0$ and $1 = \beta_1 > \beta_2 > \cdots > \beta_{r-1} > \beta_r > 0$. The final expression using w, $\alpha's$ and $\beta's$ can be coded into a Stan program. In order to represent the positive real numbers in (0, 1], we create ordered positive numbers α_i' and represent $\alpha_i = \frac{\alpha_i'}{\alpha_i'+1}$. Similar reparameterization was done for $\beta_i's$.

The inference for V is achieved by sampling from a standard generalized Bingham distribution.

Algorithm

The sampling algorithm is summarized as follows:

2.5.4 Inference for the Matrix Bingham–Von Mises–Fisher Distribution

This is a very challenging problem up till the moment the dissertation was written. In particular, there is no sufficient closed-form approximation for the normalizing constant, which is a compli-

Algorithm 3: Bayesian Algorithm for Generalized Bingham Parameters

```
Result: Bayesian samples of $A, B, V$.
Initialization: initialize $A, B, V$ ;
for i in 1 : Iterations do
    Sample $A, B$ from $p(A, B|V, U_1, U_2, \cdots, U_n)$;
    Compute Sample $V$ from $p(V|A, B, U_1, U_2, \cdots, U_n)$;
    Save samples $A, B, V$.
end
```

cated function of two matrix arguments F and A. The current best achievements in estimating the normalizing constant of the Fisher-Bingham distributions are Kume et al. (2013), Kume and Sei (2018) and Chen and Tanaka (2020). They adopt the maximum likelihood perspective and try to estimate the numerical values using gradient descent approaches.

However, the lack of closed-form approximations in terms of the parameter values makes it extremely hard to utilize the Stan program to inference the parameters. As more accurate closed-form approximations are exploited, it is promising to use the same procedure as above to tackle this challenging task.

2.6 The *StanStiefel* Package

We wrap up all the functions for sampling and inference the various distributions on the Stiefel manifold. We create a new package based on the Stan program, and name it *StanStiefel*. Here we provide a pictorial description of the structure of the package in Figure 2.1. This package contains fundamental Bayesian algorithms for sampling from popular distributions and inferring parameters for the distributions.

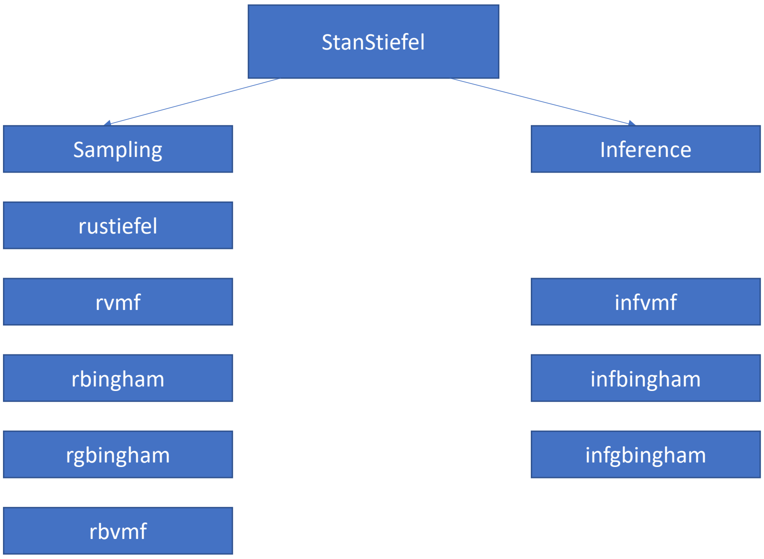


Figure 2.1: *StanStiefel* Package

2.7 Comparisons with *rstiefel* Package

The current default package for Bayesian sampling on the Stiefel manifold is *rstiefel*, which is developed and maintained based on Hoff (2009b). In this section we conduct an experiment to compare the performance of our new *StanStiefel* package with it. Under different setups, the performance may vary. Since both the matrix Langevin distribution and the generalized Bingham distribution are special cases of the matrix Bingham-von Mises-Fisher distribution, we conduct the experiment under the most general distribution. In particular, we consider four cases where C, and A, B can be either large or small, and we vary the dimensions for $p \in \{5, 10, 50, 100, 200, 500\}$, and fix the number of factors to $r = 5$. We keep track of the time elapsed for 200 iterations with the first 100 as burn-in samples. And for the obtained samples, we compute counterparts for effective sample sizes (ESS), which is designed for computing the equivalent number of independent samples by adjusting for the autocorrelations. Moreover, we are interested in the effective sample sizes per second (ESPS). For vector samples, we compute two alternative measures for effective sample sizes.

Entry-wise Average ESPS

This is a straight-forward analogue of the scalar samples. We compute the effective sample sizes for all the entries. For a sequence of samples of N $p \times r$ matrices. We first compute the pr effective sample sizes for each element in the matrix samples, then compute the average of the pr effective sample sizes. Finally divide the time spent to get the ESPS.

Vector Angles Average ESPS

Alternatively, we consider the angles between the columns of samples to a fixed anchor. We choose the eigenvectors of A as our anchor V, since the sample are supposed to be close to that when B and the eigenvalues of A are large. For a matrix sample X, we compute the r angles between the columns of V and X, the average effective sample sizes of the angles, and divide the time spent to get the ESPS.

When C is small, C is a uniformly chosen orthogonal matrix on the Stiefel manifold, and when C is large, it refers to a uniformly chosen orthogonal matrix multiplied by ten. When A and B are small and large, they are set to

$$A = B = \operatorname{diag}(\{10, 8, 6, 4, 2\})$$
$$A = B = \sqrt{10} \times \operatorname{diag}(\{10, 8, 6, 4, 2\}),$$

respectively. The results are collected in Table 2.1.

From the results, we see when p is equal to r, both methods require more time since the problem becomes sampling a full rank orthogonal matrix. This is a special case, and both methods take more time than other cases with similar dimensions. When $p > r$, the sampling time starts to increase drastically as the dimension increases. In particular, when both C and A, B are large, $p = 5$ takes more than one hour to run.

For each case, as p increases, the time for *rstiefel* package increases significantly, whereas that for the *StanStiefel* package grows at a much slower rate. The chains for both methods mixed well and in most of the experiments, *StanStiefel* package has a higher effective sample size under the above two measures. Combining the effects from both aspects, *StanStiefel* package achieves a much higher ESPS at all cases where $p \geq 100$.

		rstiefel			*StanStiefel*		
	p	Time(s)	Entrywise ESPS	Angle ESPS	Time(s)	Entrywise ESPS	Angle ESPS
Small C, small A and B	5	1.99	**42.74**	**47.34**	14.56	6.86	6.86
	10	0.39	**223.46**	**233.79**	0.92	102.21	100.46
	50	0.96	**100.58**	**104.17**	1.02	94.19	98.04
	100	4.31	22.11	20.87	1.68	**57.45**	**59.52**
	200	23.88	4.06	3.94	3.18	**30.78**	**29.11**
	500	300.25	0.32	0.33	9.69	**10.07**	**10.32**
Big C, small A and B	5	49.48	0.5	0.47	14.4	**6.7**	**6.94**
	10	0.43	87.49	80.32	0.68	**144.37**	**147.06**
	50	0.85	**103.42**	90.59	1.06	92.27	**94.34**
	100	4.36	22.22	22.94	1.62	**59.5**	**57.25**
	200	24.36	4	3.9	3.4	**28.67**	**24.8**
	500	308.26	0.31	0.32	9.92	**9.68**	**10.08**
Small C, big A and B	5	141.98	0.67	0.7	14.09	**6.97**	**7.1**
	10	0.43	**204.37**	**185.07**	1.37	71.3	72.99
	50	1.08	**90.58**	**91.49**	1.91	49.79	48.47
	100	4.57	21.45	21.88	3.04	**31.41**	**32.89**
	200	23.81	4.03	3.98	4.53	**21.35**	**22.08**
	500	300.63	0.32	0.29	13	**7.51**	**7.69**
Big C, big A and B	5	¿ 3600	N/A	N/A	14.06	**7.03**	**7.11**
	10	0.56	**82.75**	43.45	1.61	58.5	**51.25**
	50	1.31	**57.18**	**53.31**	2.01	48.2	44.07
	100	4.49	20.76	18.57	2.91	**33.74**	**34.36**
	200	23.63	4.02	4.23	4.47	**21.55**	**19.01**
	500	299	0.32	0.3	13.18	**7.35**	**7.59**

Table 2.1: Running Time and ESPS Comparisons of *rstiefel* and *StanStiefel*

2.8 Example

2.8.1 Bayesian Principal Component Analysis

In this example, we demonstrate a Bayesian principal component analysis model on simulated data. In the simulation, p is set to 50 and r is chosen as 4, the four principal eigenvalues are spaced out as $\{50, 30, 10, 5\}$, and the idiosyncratic variance is 1. We choose the spiked covariance model,

$$\Sigma = U\Lambda U^T + \sigma^2 I_p,$$
$$\Lambda = \text{diag}(\{50, 30, 10, 5\}),$$

and generate 100 observations from this covariance matrix. The goal is to apply our Bayesian principal component analysis model on the observations data to retrieve the parameters U, Λ and σ^2. The MCMC algorithm was run for 400 iterations with the first half as burn-in samples, and the samples was saved every 2 iterations after the burn-in period.

In Figure 2.2, 2.3, 2.4 and 2.5 we show the posterior distribution of the first 4 eigenvalues, respectively. The red vertical lines indicate the true eigenvalues, and the blue vertical lines represent the empirical estimates. Our bayesian PCA model is able to produce samples concentrating on the maximum likelihood estimates from observations, with appropriate posterior uncertainties.

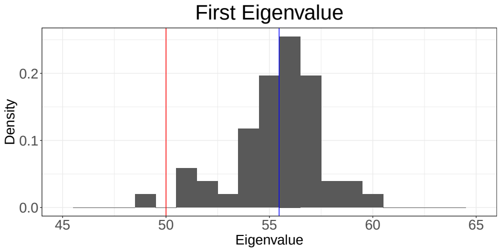


Figure 2.2: First Eigenvalue

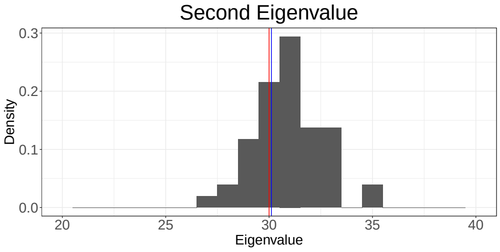


Figure 2.3: Second Eigenvalue

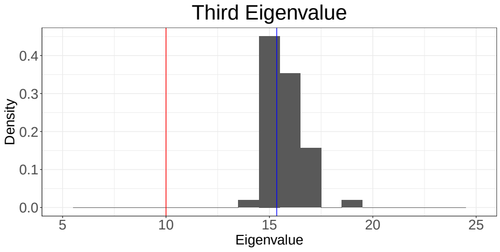


Figure 2.4: Third Eigenvalue

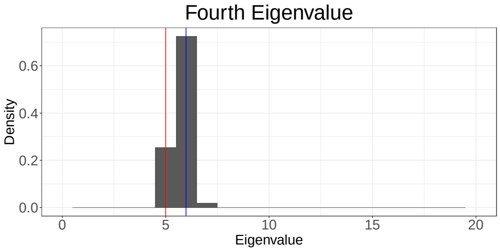


Figure 2.5: Fourth Eigenvalue

In Figure 2.6, 2.7, 2.8 and 2.9 we plot posterior summaries of the first four eigenvectors. In our model, the likelihood is invariant to the direction of the axes, so we summarize the results by the absolute values of the inner product between posterior samples and a particular fixed target vector. We compare the posterior samples to both the empirical eigenvectors and the true eigenvector, and show the corresponding histograms on the same plot. Intuitively, when the Bayesian eigenvectors are close to the target, the absolute inner product ought to be close to 1. In all four figures, we see the red histograms all have high mass near 1, indicating our Bayesian estimates concentrate at the empirical values. In 2.6 and 2.7, the blue histograms also have high mass near 1, meaning our estimated first and second eigenvectors are closer to the truth as well. In contrast, since the fourth eigenvalue is relatively smaller than the rest, and there is more posterior uncertainty because the eigenvalue is smaller.

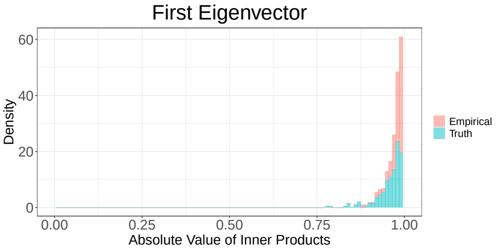


Figure 2.6: First Eigenvector

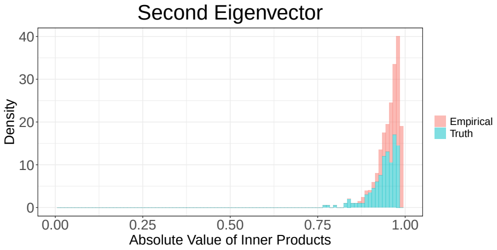


Figure 2.7: Second Eigenvector

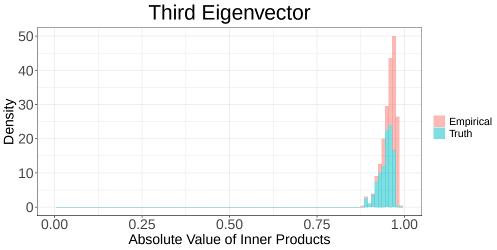


Figure 2.8: Third Eigenvector

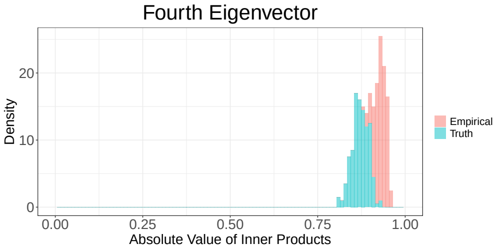


Figure 2.9: Fourth Eigenvector

Next we consider the comparison between the Bayesian samples and the true eigenvectors. The conditional distribution for the eigenvectors is proportional to

$$\text{etr}(\Omega U^T \frac{S}{2\sigma^2} U), \tag{2.53}$$

where $S = YY^T$, $\omega_i = \frac{\lambda_i}{\lambda_i + \sigma^2}$, and $\Omega = \text{diag}(\{\omega_1, \omega_2, \cdots \omega_r\})$. Hence the log-likelihood is

$$\text{tr}(\Omega U^T \frac{S}{2\sigma^2} U). \tag{2.54}$$

It is well-known that the mode of the Bayesian estimates would be the empirical eigenvectors, as shown in Hoff (2009a). Once we observe the generated data, S is fixed, in the following we consider the conditional distribution of the eigenvectors, and compare our estimated samples with the true eigenvectors. To achieve that, we use the true values of Ω and σ^2 to compute the log-likelihood. The histogram for the log-likelihood of the samples is shown in Figure 2.10, in which the red line represents the log-likelihood computed with the truth. It is clear that the log-likelihood computed with the truth near the mode of the Bayesian samples, indicating that our Bayesian samples resemble estimates for the true eigenvectors as well.

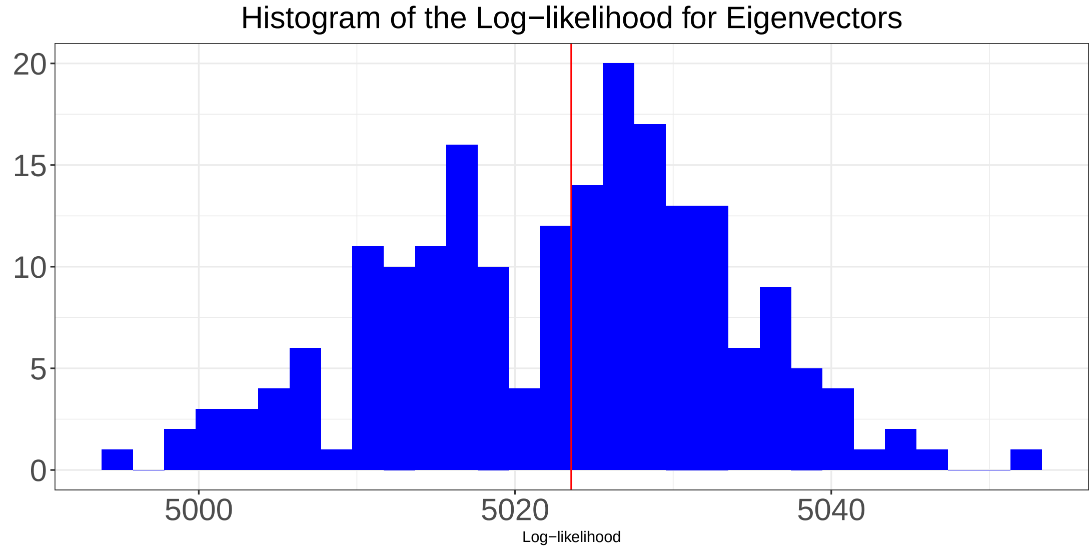


Figure 2.10: Histogram for Log-likelihood

In total, our Bayesian principal component analysis works well based on the sampling algorithm on the Stiefel manifold. The model would provide less accurate Bayesian samples when the eigenvalues are less distinguishable. We demonstrate the effectiveness of the sampling algorithm for the generalized Bingham distribution, algorithms for other distributions can also be illustrated under a similar simulation setup.

Chapter 3

Time Series Modeling on Stiefel Manifold and Applications on Covariance Estimations

3.1 Introduction

Most time series models are focused on time-varying or conditional means, assuming homoskedastic covariances over time. However, many modern statistical applications involve heterogeneous covariances. In finance, practitioners need to keep up with the changing financial environment and make constant updates for the portfolio weights and risk measures. Hence they need information about dynamic covariance matrices, as indicated in Harris et al. (2017) and Engle et al. (2019). In biology, people are interested in how the relationships amongst the metabolites evolve as an individual ages, so as to improve our understanding about age-related diseases, such as Hwangbo et al. (2021) and Franks and Hoff (2019). Heterogeneous covariance matrices are also critical for gaussian processes, quadratic discriminant analysis, and other predictive techniques. Historically, some prominent models have been proposed for modeling heterogeneous covariances. As the pioneer in this subfield, Flury (1987) developed the common principal components model, where the covariance matrices share the same eigenvectors. Based on Flurry's work, Boik (2002) relaxed the assumption to eigenvectors being shared among some or all the groups. In lieu of the strict assumption where eigenvectors are either shared or completely different, a school of Bayesian statisticians have developed hierarchical models to accommodate both similarities and differences. Specifically, Hoff (2009a) established a hierarchical model, where the eigenvector matrices are samples from the same matrix Bingham distribution. The shared parameters of the distribution are designed for the similarities, whereas the randomness characterizes the distinction. With the mind of explicitly expressing the shared information about the subspace spanned by group-level eigenvectors, Franks and Hoff (2019) proposed a shared subspace model to characterize the resemblance at the subspace level.

As a special case for heterogeneous covariances, time-varying covariance estimation has gained its significance in recent years. Engle (2002) applied the idea of generalized autoregressive conditional heteroskedastic (GARCH) to the factored components of conditional correlation matrices. Engle and Kroner (1995) further proposed multivariate extension of GARCH called BEKK, where the dynamic covariances follow an autoregressive moving average model (ARMA) process. Wu et al. (2013) improved the computational efficiency and provided uncertainty quantification by solving a variation of BEKK using Bayesian approaches. Analogous to linear regression, Hoff and Niu (2012) proposed a covariance regression model, where the covariance is posited to be the sum of a fixed positive definite matrix and a quadratic form involving the explanatory variables. Moreover, Franks and Hoff (2019) successfully extended the model to high-dimensional settings. This advance was quite beneficial for applied tasks since there are numerous cases in which the problem dimension exceeds the size of available data. In biological and medical applications, not only the data are difficult and expensive to collect, but practitioners also face privacy and moral issues. Meanwhile, there are hundreds, if not thousands of chemicals influencing the human metabolism, as well as other biological reactions. In finance, data are vastly noisy, and the relationship between the stock returns is constantly changing due to social, political and psychological factors. In order to obtain satisfactory analysis, one would better use more recent and relevant data, which is usually scarce. For low to mid-frequency investors, the goal is usually to discover profitable patterns amongst enormously available products with the help of several years' daily data, which again manifests a high-dimensional problem.

Fortunately, high-dimensional data generally can be approximated by a smaller number of factors, enabling insights to be drawn from limited samples. In biological applications, Heimberg et al. (2016) showed that the effective dimensionality is thought to scale with the number of gene regulatory modules, not the number of genes themselves. Analogously, Fama and French (1992) discovered a linear relationship between mean excess returns and exposures to three factors, the market factor, a size-based factor and a book-to-market-based factor. In 2015, they appended the profitability and investment factors to deliver better performance, see Fama and French (2015). Based on the above evidences, it is believed that with appropriate modeling techniques and experiences, people are still able to exploit the small samples and extract useful insights.

In this paper, we develop a modern Bayesian model to sufficiently estimate a temporal sequence of covariance matrices. In particular, we would like our model to be well-suited to high-dimensional data, such as biology and finance. To address the dimensionality issue, at each time point, we adopt the spiked principal components model (spiked PCA), which is studied scrupulously in Johnstone (2001). Mathematically, we assume a low dimensional structure of covariance matrices

$$\Sigma_t = U_t \Lambda_t U_t^T + \sigma_t^2 I, \tag{3.1}$$

where U_t is a $p \times r$ matrix, and Λ_t is an $r \times r$ diagonal matrix, and $p \gg r$. The leading r factors that dominate the covariance matrix corresponds to the first r eigenvalues, and σ_t^2 models the idiosyncratic variances for errors. This special structure preserves the importance of dominant factors, while simultaneously models the idiosyncratic variances σ_t^2. Meanwhile, our model is motivated by the benefits of hierarchical modeling in Hoff (2009a), where the shrinkage modeling on the eigenvectors was accomplished over the Stiefel manifold towards a grand pooled target. To tailor the idea to the time series context, we extend the shrinkage to a fixed target to an autoregressive process among the neighboring eigenvectors. This extension reduces the high variance for high-dimensional data and captures the evolution in the eigenvectors. Moreover, to further reduce the variability resulting from the scarcity of the data, we also propose a shrinkage model over the eigenvalues to exert proper regularizations. The final full Bayesian model can be efficiently inferred with Markov chain Monte Carlo algorithms via R and Stan.

Our model contributes to the community of Bayesian modeling in high-dimensional dynamic covariance matrices as a general framework encompassing various specific models. Compared with existing literature, which focus mostly on vector autoregressive processes, our approach separately models the dynamics of eigenvectors and eigenvalues. Our model allows unrelated priors and separate parameters that prevents the introduction of correlation through prior knowledge and model structure. The model is particular useful for problems where people want to discover the relationship between eigenvectors and eigenvalues. For instance, in finance the eigenvectors can be interpreted as a scaled version of the market beta, which characterizes the risk that an asset is exposed to with respect to a well-diversed market portfolio. The eigenvalues describe the variances of the latent factor that explain the cross-sectional correlations. In particular, the first eigenvalue, which corresponds to the market factor, serves as a proxy for market volatility. The discussion about the relationship between beta and volatility has been active in the financial literature, and our model offers a distinct perspective in this topic. The details can be found in the next chapter of this dissertation. Moreover, we are the first to propose a Bayesian autoregressive model on the Stiefel manifold, which opens the door to a new set of time series models on non-Euclidean space. Similar models were demonstrated in Chikuse (2006) and Yang and Bauwens (2018). In Chikuse (2006), both the observation and the latent variables locate on the Stiefel manifold, whereas in Yang and Bauwens (2018), only the latent variables stay on the Stiefel manifold. Our work provides a Bayesian version for the dynamics. Insofar, our work is the pioneer in weaving ideas from Stiefel manifold sampling, Bayesian autoregressive modeling and time series modeling.

In Section 2 we expound the model definition in greater detail. Separate explanations will be provided on how the eigenvector matrices and eigenvalues are manipulated with shrinkage models. Section 3 is devoted for thoroughly dissecting the Markov chain Monte Carlo algorithm for inferencing all the parameters and hyper-parameters. The comprehensive full posterior distribution is provided and full conditional distributions for different Gibbs sampling procedures will be extracted. In Section 4 we run simulations to demonstrate the effectiveness and power of the model in mitigating high-dimensional variabilities and improving the accuracy. Finally, the model is applied to a real temperature dataset on the mean surface temperature change of the domains from Food and Agriculture Organization of the United Nations (FAO). We aim to discover the latent factors that influence the covariance of change of temperatures for different areas, and investigate the dynamics of these factors over time.

3.2 Model Definition

In this paper we aim to address a high-dimensional dynamical covariance estimation problem. There are T time points, and n_t observations were collected at time point t. We would like to estimate T corresponding covariance matrices simultaneously. Specifically, the n_t observations of dimension p are assumed to be independent and they are identically distributed. We model the marginal distributions of the observations as multivariate normal distributions $N(0_p, \Sigma_t)$. The mean-adjusted observations are denoted as $y_t^{(1)}, y_t^{(2)}, ..., y_t^{(n_t)}$, and are stacked column-wise to create the data matrix Y_t at time t. As a result, Y_t is a $p \times n_t$ matrix with $p \gg n_t$, and $S_t = Y_t Y_t^T$ follows a possibly degenerate Wishart(Σ_t, n_t) distribution with density

$$p(S_t|\Sigma_t, n_t) \propto l(\Sigma_t : S_t) = |\Sigma_t|^{-n_t/2} \operatorname{etr}(-\Sigma_t^{-1} S_t/2), \tag{3.2}$$

where etr is the exponentiated trace.

At each time point, $p \gg n_t$, hence the empirical covariance matrix fails to be a full rank matrix. Thus, it is not an ideal candidate for estimating the true covariance matrix. Fortunately, according to Udell and Townsend (2019), high-dimensional data often manifest a low rank structure and can be explained by a few significant factors. We thus posit a spiked covariance model, which involves a low rank component representing the dominant factors, and a diagonal component that models the idiosyncratic variances for the errors. The diagonal component bridges the gap between the low rank structure and a full rank covariance matrix. Johnstone (2001), Paul (2007) and Baik and Silverstein (2006) studies the theory of the asymptotics of the spiked covariance model. Under this model, the covariance matrix at time point t can be represented as

$$Y_t = U_t X_t + \epsilon_t, \tag{3.3}$$

$$X_t \sim N(0, \Lambda_t), \tag{3.4}$$

$$\Lambda_t = \text{diag}(\{\lambda_t^{(1)}, \lambda_t^{(2)}, \cdots, \lambda_t^{(r)}\}). \tag{3.5}$$

$$\epsilon_t \sim N(0, \sigma_t^2 I_p), \tag{3.6}$$

$$\Sigma_t = U_t \Lambda_t U_t^T + \sigma_t^2 I_p, \tag{3.7}$$

U_t is a p by r orthogonal matrix denoting the leading r eigenvectors, and $U_t^T U_t = I_r$. Notice that σ_t represents the common variance of the idiosyncratic factors, and $\{\lambda_t^{(1)} + \sigma_t^2, \lambda_t^{(2)} + \sigma_t^2, \cdots, \lambda_t^{(r)} + \sigma_t^2\}$ represent the variances of the latent factors at time t. Meanwhile, Σ_t is fully determined by three groups of parameters: $\{U_t\}, \{\sigma_t^2\}, \{\lambda_t^{(1)}, \lambda_t^{(2)}, \cdots, \lambda_t^{(r)}\}$.

In fact, for the problem we are studying, we can assume that the eigenvectors and eigenvalues evolve smoothly, and the high variations in the empirical estimates due mainly to the sampling variabilities under the high-dimensional setup. In our model, we strive to exploit the similarities amongst time points and combine information across the whole timeframe, and we propose autoregressive time series models on eigenvectors and eigenvalues to achieve those ideas.

3.2.1 Time Series Modeling for Eigenvector Parameters

At each time point t, U_t is by definition an element on the Stiefel manifold $\mathcal{V}_{p,r}$, which contains all $p \times r$ semi-orthogonal matrices in $\mathbb{R}^p$, such that $U_t^T U_t = I_r$. Columns of U_t reflects the principal axes, and the directions are not identifiable since multiplying by -1 on any column does not change the model in any sense. Since we are mainly interested in the axes rather than the directions, sign-agnostic distributions will be adopted. To translate the time series idea into probability, we propose an autoregressive model on the sequence of eigenvector matrices with the generalized Bingham distribution. Mathematically, the conditional distribution can be expressed as

$$U_t|U_{t-1}, A, B \sim c(A, B)\,\mathrm{etr}(BU_t^T U_{t-1} A U_{t-1}^T U_t), \tag{3.8}$$

$$A = \mathrm{diag}(\{a_1, a_2, \cdots, a_r\}), \quad a_1 \geq a_2 \geq \cdots \geq a_r > 0 \tag{3.9}$$

$$B = \mathrm{diag}(\{b_1, b_2, \cdots, b_r\}), \quad b_1 \geq b_2 \geq \cdots \geq b_r > 0, \tag{3.10}$$

where U_{t-1} denotes the eigenvectors at time $t-1$, and $c(A, B)$ is the inverse of the normalizing constant for the generalized Bingham distribution. Notice that the parameters A and B are diagonal matrices shared across different time points, and they facilitate the alignment of the columns of $U_t's$.

The idea is analogous to Hoff (2009a), where the population of eigenvectors are modeled as samples from a common distribution. Our model serves as a similar counterpart for time series modeling. We briefly discuss the interpretations of A and B in the following, and leave more interested readers to check out section 2 in Hoff (2009a). In general, consider two matrices U and V in the Stiefel manifold $\mathcal{V}_{p,r}$, and

$$U \sim c(A, B)\,\mathrm{etr}(BU^T V A V^T U), \tag{3.11}$$

then

$$\mathrm{tr}(BU^T V A V^T U) = \sum_{i=1}^{r}\sum_{j=1}^{r} a_i b_j (v_i^T u_j)^2 = \sum_{j=1}^{r} b_j u_j^T \left(VAV^T\right) u_j. \tag{3.12}$$

Based on the principle of maximum likelihood, when a_i and b_j are large, u_j will be close to v_i. For instance, if a_1 and b_1 are larger than the rest, we would expect $a_1 b_1$ to be large, hence u_1 revolves around v_1. One should also keep in mind the built-in orthogonality amongst $u_i's$ and $v_i's$. Therefore, when u_1 is close to v_1, other columns of U must be nearly perpendicular to v_1. Alternatively, $a_i = a_{i+1}$ implies $v_i^T u_j$ behaves the same in distribution as $v_{i+1}^T u_j$, and $b_j = b_{j+1}$ implies that u_j follows the same distribution as u_{j+1}.

3.2.2 Modeling Principal Eigenvalues

Since there are only relatively few observations at each time point compared with the number of features, the empirical eigenvalues tend to overestimate the truth, making them poor estimates of the true eigenvalues. Moreover, evidence shows that the first several empirical eigenvalues can be volatile across time, which contradicts our assumption that the eigenvalues evolve smoothly. To introduce smoothness into the model, we propose time series models to allow information sharing across all time points. Possible choices involving traditional statistics, such as linear regression models, autoregressive model $AR(p)$, Gaussian processes models and non-parametric regression models. Notice that there is no perfect solution in the choice of models. Simpler models are often preferred for at least two reasons. First of all, the main purpose of the shrinkage model is to bring the estimates closer, so as to avoid variabilities created by the high-dimensional issue. Flexible models can easily overfit the data and concentrate on the empirical values, thus mitigates the shrinkage effect. An overly flexible model, which might have a reduced bias, will inevitably add high variance into the context, thus contaminate the estimated results.

In light of the above concerns on the selection of the shrinkage model, we hereby discuss two convenient well-studied options: the constrained simple linear regression and the constrained first-order autoregression. One should keep in mind that there might be better-suited alternatives based on the problem structure and prior knowledge. It is encouraged to try a few different alternatives before deciding on the one with the best interpretations for the underlying dataset. In the following, we demonstrate the above model. Due to the positiveness of eigenvalues, the parameters are constrained to the positive real line, which can be handled in Stan by setting the lower bound to 0.

Constrained Simple Linear Regression

We propose a constrained simple linear regression using the time index as the explanatory variable:

$$\begin{aligned}&\lambda_t^{(j)} > 0,\\ &\lambda_t^{(j)} = \alpha^{(j)} + \beta^{(j)} t + \epsilon_t^{(j)}, \quad \epsilon_t^{(j)} \sim N(0, \tau^{(j)^2}).\end{aligned}$$

where $j \in \{1, 2, ..., r\}$, representing the index of the principal eigenvalues, and $t \in \{1, 2, ..., T\}$ stands for the time index. This model is appropriate when the researchers postulate that the variance of the factor evolves in a linear fashion, such as the variance of the concentration of chemicals in a person's body. In addition, we might put a prior on $\tau^{(j)^2}$ to achieve smoother estimates.

Constrained First-order Autoregression

The constrained first-order autoregression is proposed as

$$\begin{aligned} &\lambda_t^{(j)} > 0, \\ &\lambda_t^{(j)} = c^{(j)} + \varphi^{(j)} \lambda_{t-1}^{(j)} + \epsilon_t^{(j)}, \quad \epsilon_t^{(j)} \sim N(0, {\tau^{(j)}}^2). \end{aligned}$$

where $j \in \{1, 2, ..., r\}$, representing the index of the principal eigenvalues, and $t \in \{2, ..., T\}$ stands for the time index. This model is used to capture the situation where the variance of the factor is believed to be dominated mostly by the one-step-ahead predecessor, such as the variance of the factors that explain the financial returns, see Fama and French (1992) and Fama and French (2015). In addition, we might alter the distribution of the error term ϵ_t to accommodate the black swan events in the market.

An alternative approach is to model the eigenvalues in their logarithmic scales to avoid the positive constraint. Here we show an example, and other models can be constructed similarly.

First-order Autoregression on Logarithmic Scale

The first-order autoregression on logarithmic scale is proposed as

$$\log(\lambda_t^{(j)}) = c^{(j)} + \varphi^{(j)} \log(\lambda_{t-1}^{(j)}) + \epsilon_t^{(j)}, \quad \epsilon_t^{(j)} \sim N(0, {\tau^{(j)}}^2).$$

The advantage is getting refrained from the constraint. However, biases on the logarithmic of the eigenvalues would induce higher biases when they are transformed back to the original scale via the exponential function.

3.2.3 Modeling the Idiosyncratic Variances

To satisfy the full-rank assumption of the covariance matrix, we have the diagonal part $\sigma_t^2 I$ in the model. For time point t, the common trailing eigenvalue is denoted by σ_t^2. We choose not to utilize information from other time points, as opposed to estimating the principal eigenvalues. On one hand, there is sufficient information at time t to provide a good estimate of σ_t^2. Evidences show that the median of the trailing empirical eigenvalues serves as a satisfactory candidate. On the other hand, we reserve this parameter for adjusting for the uniqueness for time point t.

3.2.4 Model Summary

From the temporal perspective, the model can be summarized pictorially in Figure 3.1. For the eigenvalues, here we take the general model for illustration purposes. Model summaries for specific eigenvalue processes, such as the first-order autoregressive models or linear regressions, can be constructed similarly.

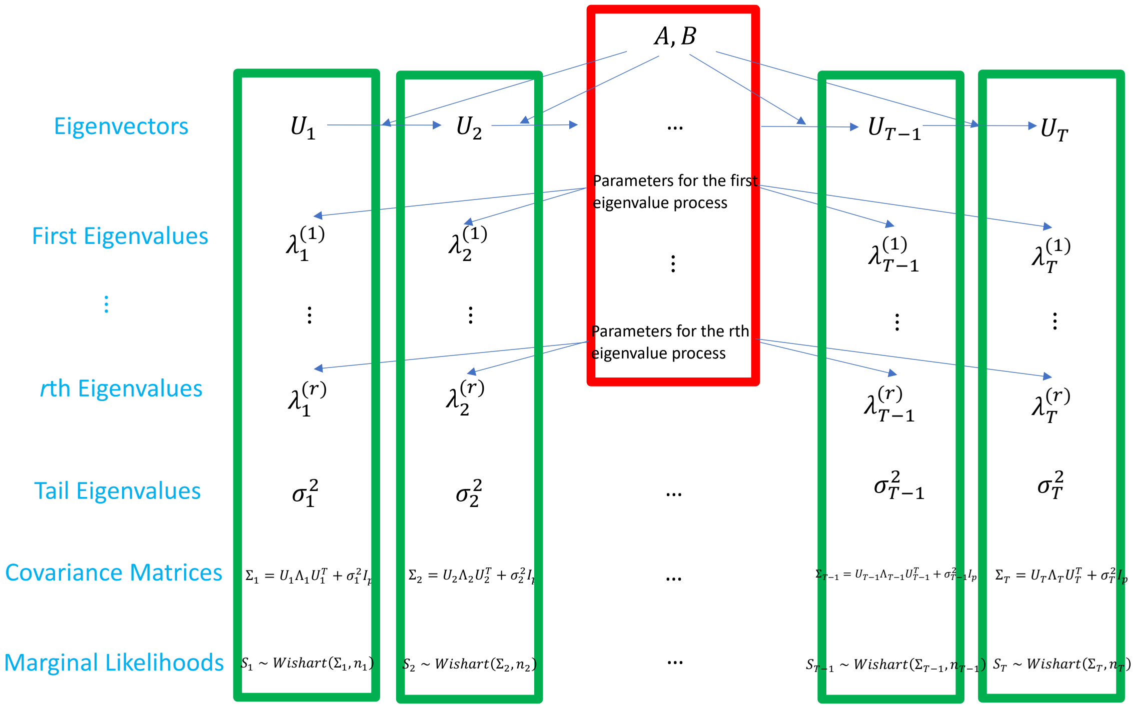

A, B
Eigenvectors
First Eigenvalues
⋮
rth Eigenvalues
Tail Eigenvalues
Covariance Matrices
Marginal Likelihoods
U_1
U_2
...
U_{T-1}
U_T
Parameters for the first eigenvalue process
Parameters for the rth eigenvalue process
$\lambda_1^{(1)}$
$\lambda_2^{(1)}$
$\lambda_{T-1}^{(1)}$
$\lambda_T^{(1)}$
$\lambda_1^{(r)}$
$\lambda_2^{(r)}$
$\lambda_{T-1}^{(r)}$
$\lambda_T^{(r)}$
σ_1^2
σ_2^2
σ_{T-1}^2
σ_T^2
$\Sigma_1 = U_1\Lambda_1U_1^T + \sigma_1^2I_p$
$\Sigma_2 = U_2\Lambda_2U_2^T + \sigma_2^2I_p$
$\Sigma_{T-1} = U_{T-1}\Lambda_{T-1}U_{T-1}^T + \sigma_{T-1}^2I_p$
$\Sigma_T = U_T\Lambda_TU_T^T + \sigma_T^2I_p$
$S_1 \sim Wishart(\Sigma_1, n_1)$
$S_2 \sim Wishart(\Sigma_2, n_2)$
$S_{T-1} \sim Wishart(\Sigma_{T-1}, n_{T-1})$
$S_T \sim Wishart(\Sigma_T, n_T)$

Figure 3.1: Model Summary

3.3 Model Inference

This section is devoted to the technical details about inferring the parameters. There are many parameters and they can be grouped up as across-group and group-specific parameters. We first derive the full posterior distribution and move on to the conditional distributions for the Gibbs Sampling. The full Bayesian hierarchical model can be decomposed into four components depending on their functionalities, as shown in 3.2. The observations contribute to the likelihood, which are products of normal distributions for all the observations at different time points. The model parameters can be divided into two main groups, across-group and group-specific. Across-group parameters involve the hyperparameters for the generalized Bingham distribution, as well as the β's and τ's for different eigenvalue sequences, assuming the first-order autoregressive processes is adopted. The group-specific parameters at time point t contain the eigenvector U_t, r principal eigenvalues $\lambda_t^{(1)}, \lambda_t^{(2)}, \cdots, \lambda_t^{(r)}$, and the trailing eigenvalue σ_t^2. Standard priors or conjugate priors are adopted to facilitate the inference process. Figure 3.2 shows the organization of the model parameters.

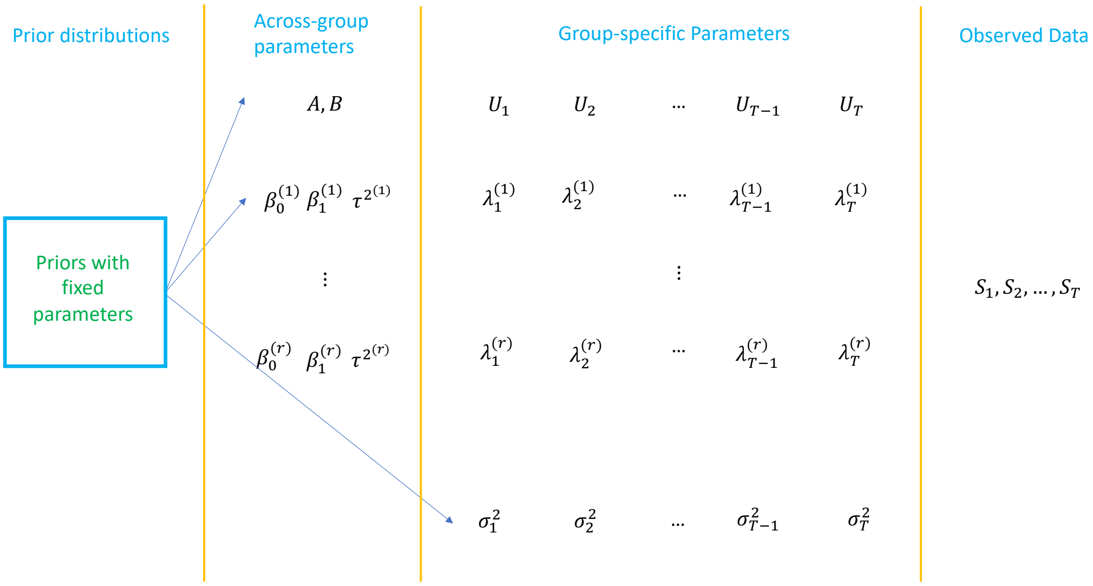

Prior distributions
Across-group parameters
Group-specific Parameters
Observed Data
Priors with fixed parameters
A, B
$\beta_0^{(1)}$ $\beta_1^{(1)}$ $\tau^{2(1)}$
⋮
$\beta_0^{(r)}$ $\beta_1^{(r)}$ $\tau^{2(r)}$
U_1 U_2 … U_{T-1} U_T
$\lambda_1^{(1)}$ $\lambda_2^{(1)}$ … $\lambda_{T-1}^{(1)}$ $\lambda_T^{(1)}$
⋮
$\lambda_1^{(r)}$ $\lambda_2^{(r)}$ … $\lambda_{T-1}^{(r)}$ $\lambda_T^{(r)}$
σ_1^2 σ_2^2 … σ_{T-1}^2 σ_T^2
$S_1, S_2, \dots, S_T$

Figure 3.2: All Model Parameters

3.3.1 Derivation of Full Posterior Distribution

The centered marginal likelihoods are normal, $y_t^{(k)} \sim N(0, \Sigma_t)$ for $k = 1, 2, \cdots, n_t$. The full likelihood for all the observations across all time points is

$$p(S_1, ..., S_T | \Sigma_1, ..., \Sigma_T, n_1, ..., n_T) \propto \prod_{t=1}^{N} \prod_{k=1}^{n_t} f(y_t^{(k)} | \Sigma_t), \tag{3.13}$$

where $f(y_t^{(k)}|\Sigma_t)$ represents the centered multivariate normal likelihood with covariance matrix Σ_t. Furthermore, under the spiked covariance model assumption, the determinant and the inverse of Σ_t can be expressed as:

$$|\Sigma_t| = \det(\Sigma_t) = (\sigma_t^2)^p \prod_{j=1}^{r} \frac{\lambda_t^{(j)} + \sigma_t^2}{\sigma_t^2}, \tag{3.14}$$

and

$$\Sigma_t^{-1} = \frac{1}{\sigma_t^2}\left(I_p - U_t \Omega_t U_t^T\right), \tag{3.15}$$

where Ω_t is a diagonal matrix and $w_t^{(j)} = \frac{\lambda_t^{(j)}}{\lambda_t^{(j)}+\sigma_t^2}$, $j \in \{1, 2, ..., r\}$.

Therefore, the full likelihood can be written explicitly in terms of the parameters as

$$p(S_1, ..., S_T|\Sigma_1, ..., \Sigma_T, n_1, ..., n_T) \propto \prod_{t=1}^{T} (\sigma_t^2)^{-\frac{n_t p}{2}} \operatorname{etr}\left(\frac{1}{2\sigma_t^2}(U_t \Omega_t U_t^T - I)S_t\right) \prod_{j=1}^{r} \left(\frac{\sigma_t^2}{\lambda_t^{(j)} + \sigma_t^2}\right)^{\frac{n_t}{2}} \tag{3.16}$$

The priors, on the other hand, are induced by the autoregressive processes. The prior distribution for the eigenvectors is a sequence of generalized Bingham distributions characterizing the evolution:

$$\prod_{t=2}^{N} c(A, B) \operatorname{etr}(B U_t^T U_{t-1} A U_{t-1}^T U_t). \tag{3.17}$$

The priors for the eigenvalues depend on which eigenvalue model is adopted. For simpler models such as linear regressions or first-order autoregressive models, the priors can be easily written as

$$\prod_{j=1}^{r} \prod_{t=2}^{T} \frac{1}{\sqrt{2\pi\tau^{2(j)}}} \exp\left(-\frac{(\lambda_t^{(j)} - \beta_0^{(j)} - \beta_1^{(j)} t)^2}{2\tau^{2(j)}}\right). \tag{3.18}$$

or

$$\prod_{j=1}^{r} \prod_{t=2}^{T} \frac{1}{\sqrt{2\pi\tau^{2(j)}}} \exp\left(-\frac{(\lambda_t^{(j)} - \beta_0^{(j)} - \beta_1^{(j)} \lambda_{t-1}^{(j)})^2}{2\tau^{2(j)}}\right). \tag{3.19}$$

As for the trailing eigenvalues, the family of inverse gamma distributions serves as good conjugate priors. We eventually decide on the uninformative prior on the positive real line to avoid biases when we are not equipped with enough domain knowledge.

Finally, the fixed priors on the across-group parameter are chosen at the discretion of the modeler. Non-informative priors on corresponding domains are selected provided no possession of prior knowledge.

The full posterior distribution is formulated via multiplying the full likelihood 3.16, priors on the eigenvectors 3.17, priors on leading eigenvalues (3.18 or 3.19 or other models) and priors on idiosyncratic variances, which is assumed to be non-informative, as well as all the fixed priors on the across-group parameters. The final result is too complicated to be classified as any standard distribution and it would be extremely slow and troublesome to attempt working in the whole parameter space. Therefore, we are going to apply the Gibbs sampling technique.

3.3.2 Markov Chain Monte Carlo Algorithm

Inference A and B

Since A and B are diagonal matrices, essentially we are inferring $2r$ parameters in total. They can be done separately, subject to the order constraints. The full conditional distribution is

$$p(A, B|U_t's) \sim \prod_{t=2}^{T} c(A, B)\,\mathrm{etr}(BU_t^T U_{t-1} A U_{t-1}^T U_t). \tag{3.20}$$

In order to estimate A and B, we need to find an adequate numerical approximation of $c(A, B)$. According to corollary 2.1 in Constantine and Muirhead (1976):

If R_1 and S are $k \times k$ and $m \times m$ diagonal matrices respectively, $k \leq m$, with unequal elements ordered in descending order, then

$$\int_{V(k,m)} \exp(\mathrm{tr}(1/2)n\ R_1 H_1^T S H_1)(dH_1)$$
$$\sim 2^k \exp\left((1/2)n\sum_{i=1}^{k} r_i s_i\right)\prod_{i<j}^{k}(\frac{2\pi}{nc_{ij}})^{1/2}\prod_{i=1}^{k}\prod_{j=k+1}^{m}(\frac{2\pi}{nd_{ij}})^{1/2},$$

where $c_{ij} = (r_i - r_j)(s_i - s_j)$ and $d_{ij} = r_i(s_i - s_j)$ for $i = 1, 2, \cdots, k$ and $j = k+1, \cdots, m$. $V(k, m)$ is the Stiefel manifold consisting of all $m \times k$ matrices H_1 with orthonormal columns.

In the above corollary, we take $n = 2, m = p, k = r, R_1 = A$, the first r diagonal elements of S to be B, and the last $p - r$ elements to 0. We obtain a good approximation of $c(A, B)$ as

$$2^{-r}\pi^{\frac{1}{2}(\frac{r(r+1)}{2}-pr)} \exp\left(-\sum_{i=1}^{r} a_i b_i\right)\prod_{i<j}^{r}(a_i - a_j)^{1/2}(b_i - b_j)^{1/2}\prod_{i=1}^{r}(a_i b_i)^{\frac{p-r}{2}}. \tag{3.21}$$

Notice that as parameters of a generalized Bingham distribution, A and B are non-identifiable under some transformations. As mentioned in Hoff (2009a), the likelihood $p(A, B|U_i's)$ behaves the same as that with $p(kA, \frac{1}{k}B|U_i's)$ for $k > 0$. Meanwhile, $p(A + cI, B + dI|U_i's)$ gets a density proportional to that with A and B, and that suggests only the differences amongst the diagonal elements matter. Taking these properties into consideration, we reparametrize A and B as:

$$\mathrm{diag}(A) = (a_1, \ldots, a_r) = \sqrt{w}\,(\alpha_1, \ldots, \alpha_r) \tag{3.22}$$
$$\mathrm{diag}(B) = (b_1, \ldots, b_r) = \sqrt{w}\,(\beta_1, \ldots, \beta_r)\,, \tag{3.23}$$

where $w > 0, 1 = \alpha_1 > \alpha_2 > \cdots > \alpha_{r-1} > \alpha_r > 0$ and $1 = \beta_1 > \beta_2 > \cdots > \beta_{r-1} > \beta_r > 0$. The final expression using w, $\alpha's$ and $\beta's$ can be coded into a Stan program, which can explore the parameter space well.

Inference Parameters for the Eigenvalue Processes

Again, the full conditional distribution of the parameters underlying the eigenvalues model is not fixed. And one has lots of flexibility to choose the desired shrinkage model. In case of a simple linear regression model, the full conditional distribution is given by

$$p(\beta_0^{(j)}, \beta_1^{(j)}, \tau^{2^{(j)}} | \Lambda_t's) \propto \prod_{t=2}^{T} \frac{1}{\sqrt{2\pi\tau^{2^{(j)}}}} \exp\left(-\frac{(\lambda_t^{(j)} - \beta_0^{(j)} - \beta_1^{(j)} t)^2}{2\tau^{2^{(j)}}} \right), \tag{3.24}$$

and that for a first-order autoregressive model would be

$$p(\beta_0^{(j)}, \beta_1^{(j)}, \tau^{2^{(j)}} | \Lambda_t's) \propto \prod_{t=2}^{T} \frac{1}{\sqrt{2\pi\tau^{2^{(j)}}}} \exp\left(-\frac{(\lambda_t^{(j)} - \beta_0^{(j)} - \beta_1^{(j)} \lambda_{t-1}^{(j)})^2}{2\tau^{2^{(j)}}} \right). \tag{3.25}$$

To achieve shrinkage effects, we ought to put another shrinkage prior on $\tau^{2^{(j)}}$, the exact value of the prior distribution depends on the scale of the problem as well. Meanwhile, if we postulate that the eigenvalues should follow a stationary process we might also put a prior on $\beta_1^{(j)}$ and restrict it to the range of $(0, 1)$. Besides, there is no ordering issue as long as we keep the correspondence between the factors (columns of $U's$) and their variances.

Inference U_t

The full conditional distribution of U_t varies, based on the location of U_t and the number of neighbors it has. There are three cases, the first time point, the last time point, and any time point in between.

1. U_1.

   It only has one neighbor, U_2. The full conditional distribution is

$$p(U_1 | A, B, U_2, U_3, \cdots, U_T) \propto \text{etr}(BU_2^T U_1 A U_1^T U_2)\, \text{etr}\left(\frac{1}{2\sigma_1^2} U_1 \Omega_1 U_1^T S_1 \right) \tag{3.26}$$

2. U_t, $t \in \{2, \cdots, T-1\}$.

   There are two neighbors: U_{t-1} and U_{t+1}.

$$p(U_t | A, B, U_1, \cdots, U_{t-1}, U_{t+1}, \cdots, U_T) \propto$$
$$\text{etr}(BU_t^T U_{t-1} A U_{t-1}^T U_t)\, \text{etr}(BU_{t+1}^T U_t A U_t^T U_{t+1})\, \text{etr}\left(\frac{1}{2\sigma_t^2} U_t \Omega_t U_t^T S_t \right) \tag{3.27}$$

3. U_T.

It has only one neighbor, the second last time point U_{T-1}.

$$p(U_T|A, B, U_1, U_2, \cdots, U_{T-1}) \propto \operatorname{etr}(BU_T^T U_{T-1} A U_{T-1}^T U_T) \operatorname{etr}\left(\frac{1}{2\sigma_T^2} U_T \Omega_T U_T^T S_T\right) \quad (3.28)$$

After algebraic manipulations, they can be unified in the general format $\operatorname{etr}(AU^T BU + CU^T DU + EU^T FU)$, where A, C, E are diagonal matrices and B, D, F are $p \times p$ matrices.

The sampling on the Stiefel manifold is challenging and there are attempts from various aspects. Some basic techniques are the rejection sampling and the importance sampling, which are only efficient for a special class of problems. In addition, Laplace approximation and variational Bayes methods are designed in the spirit of replacing the target posterior distribution with a computationally feasible alternative. Another prominent stream of thought, which is well-known as Markov chain Monte Carlo (MCMC), is based on constructing a Markov chain with the target distribution as the stationary distribution. The Metropolis-Hastings algorithm and Gibbs sampling both fall into this category. Recently, a sub-class of MCMC methods gains popularity with their ability to propose long distance moves in the state space and high acceptance rates. Being known as Hamiltonian Monte Carlo (Neal et al. (2011)), the method simulates Hamiltonian dynamics in an augmented parameter space and the projected trajectories are retained as samples.

Hoff (2009b) discusses the Gibbs sampling algorithm for sampling from the matrix Bingham-von Mises-Fisher distribution. Reparameterization was adopted to remove the built-in constraints of the Stiefel manifold. In most cases, we can use Gibbs sampling to sample the column vectors iteratively, and special treatments need to be applied for the full rank case. Pourzanjani et al. (2021) utilizes the idea of Givens representation to develop a nice algorithm in Stan. However, it takes great efforts to theoretically compute the change-of-measure term and it pays to adjust to the topological difference between the transformed parameter space and the original space. Moreover, Jauch et al. (2020b) developed a novel sampling scheme on the basis of the Cayley transformation, and Nirwan and Bertschinger (2019) works on the Householder transformation. In this paper we are going to adopt the latest, and probably the best algorithm devised by Jauch et al. (2020a), which reparametrizes the Stiefel manifold by unconstrained matrices of the same dimension. For a matrix $X \in \mathbb{R}^{p\times k}$, its singular value decomposition is denoted as

$X = UDV^T$, let

$$\begin{aligned} Q_X &= X(X^TX)^{-1/2} = UV^T, \\ S_X &= X^TX = VD^TU^TUDV^T = VD^TDV^T, \\ S_X^{1/2} &= VDV^T. \end{aligned}$$

Then $X = Q_X S_X^{1/2}$ and $S_X = S_X^{1/2} S_X^{1/2}$, where Q_X is an orthogonal matrix while $S_X^{1/2}$ is a symmetric positive definite matrix. Analogous to the polar expansion $z = re^{i\phi}$ for complex numbers, $S_X^{1/2}$ is the counterpart for r while Q_X is comparable to $e^{i\phi}$.

The advantage of introducing Q_X and S_x together is that now the mapping from a real, full rank matrix X to the components (Q_X, S_X) of its polar decomposition is one-to-one, and the density f_X can be derived as

$$f_X(X) = f_{S_X|Q_X}(S_X|Q_X) f_{Q_X}(Q_X) \times J(Q_X, S_X; X). \tag{3.29}$$

In contrast with Cayley's transformation and Givens representation, where it is expensive to compute the Jacobian, $J(Q_X, S_X; X)$ is a standard result shown in Chikuse (2012).

$$J\left(Q_X, S_X; X\right) = \frac{\Gamma_k\left(\frac{p}{2}\right)}{\pi^{\frac{pk}{2}}} \left|S_X\right|^{-\frac{p-k-1}{2}}. \tag{3.30}$$

This convenience makes this approach much more attractive than other competitors.

As indicated above, $f_{Q_X}(Q_X)$ would be our target distribution f_Q. Therefore, once the conditional distribution of $f_{S_X|Q_X}$ is determined, we would have a corresponding density on X. It is easily seen that there are various densities $f_X(X)$ that have the margin distribution matching our desired distribution.

As a default choice, Jauch et al. (2020a) recommended $f_{S_X|Q_X}$ to be the density of the Wishart distribution $W_k(p, I_k)$ and it is independent of Q_X. With this choice, the density of the distribution of X simplifies to

$$f_X(X) = \left(\frac{1}{\sqrt{2\pi}}\right)^{pk} \operatorname{etr}(-X^TX/2) f_Q(Q_X). \tag{3.31}$$

In particular, if we consider the problem of sampling uniformly from the Stiefel manifold, $f_Q(Q_X) \propto 1$, then the density of X will be

$$f_X(X) = \left(\frac{1}{\sqrt{2\pi}}\right)^{pk} \operatorname{etr}(-X^TX/2). \tag{3.32}$$

This density shows that all the entries of X are independent standard normal random variables. Notice that this is equivalent to the situation of sampling from the unit sphere. This correspondence motivates the author to select the Wishart distribution as the default choice.

Inference $\lambda_t^{(j)}$

Besides the specific model used for the eigenvalue processes, the posterior distribution also depends on the location and the number of neighbors. Here we demonstrate using the first-order autoregressive model, where j represents the index of the eigenvalues, and $j \in \{1, 2, \cdots, r\}$.

1. $\lambda_1^{(j)}$. It only has one neighbor, $\lambda_2^{(j)}$. Hence the full conditional distribution is:

$$p(\lambda_1^{(j)} | \beta_0^{(j)}, \beta_1^{(j)}, \tau^{2^{(j)}}, \lambda_2^{(j)}, \cdots, \lambda_T^{(j)}) \propto \exp\left(-\frac{(\lambda_2^{(j)} - \beta_0^{(j)} - \beta_1^{(j)}\lambda_1^{(j)})^2}{2\tau^{2^{(j)}}}\right) \operatorname{etr}\left(\frac{1}{2\sigma_1^2} U_1 \Omega_1 U_1^T S_1\right) \left(\frac{\sigma_1^2}{\lambda_1^{(j)} + \sigma_1^2}\right)^{\frac{n_1}{2}} \tag{3.33}$$

2. $\lambda_t^{(j)}$, $t \in \{2, \cdots, T-1\}$. Every element here has two neighbors.

$$p(\lambda_t^{(j)} | \beta_0^{(j)}, \beta_1^{(j)}, \tau^{2^{(j)}}, \lambda_1^{(j)}, \cdots, \lambda_{t-1}^{(j)}, \lambda_{t+1}^{(j)} \cdots \lambda_T^{(j)}) \propto \exp\left(-\frac{(\lambda_t^{(j)} - \beta_0^{(j)} - \beta_1^{(j)}\lambda_{t-1}^{(j)})^2}{2\tau^{2^{(j)}}} - \frac{(\lambda_{t+1}^{(j)} - \beta_0^{(j)} - \beta_1^{(j)}\lambda_t^{(j)})^2}{2\tau^{2^{(j)}}}\right) \operatorname{etr}\left(\frac{1}{2\sigma_t^2} U_t \Omega_t U_t^T S_t\right) \left(\frac{\sigma_t^2}{\lambda_t^{(j)} + \sigma_i^2}\right)^{\frac{n_t}{2}} \tag{3.34}$$

3. $\lambda_T^{(j)}$. Its only neighbor is the second last time point.

$$p(\lambda_T^{(j)} | \beta_0^{(j)}, \beta_1^{(j)}, \tau^{2^{(j)}}, \lambda_1^{(j)}, \cdots, \lambda_{T-1}^{(j)}) \propto \exp\left(-\frac{(\lambda_T^{(j)} - \beta_0^{(j)} - \beta_1^{(j)}\lambda_{T-1}^{(j)})^2}{2\tau^{2^{(j)}}}\right) \operatorname{etr}\left(\frac{1}{2\sigma_T^2} U_T \Omega_T U_T^T S_T\right) \left(\frac{\sigma_T^2}{\lambda_N^{(j)} + \sigma_T^2}\right)^{\frac{n_T}{2}} \tag{3.35}$$

These are complicated univariate distributions, from which we can use Stan programs to sample. Notice that we need to constrain the $\lambda_t^{(j)}$'s to the positive real line since they are parameters for the eigenvalues.

Inference σ_t^2

Without prior knowledge, we will put a non-informative uniform prior on $[0, \infty)$.

$$p(\sigma_t^2 | U_t, \lambda_t^{(j)}, S_t) \propto (\sigma_t^2)^{-\frac{n_t p}{2}} \operatorname{etr}\left(\frac{1}{2\sigma_t^2}(U_t \Omega_t U_t^T - I) S_t\right) \prod_{j=1}^{r} \left(\frac{\sigma_t^2}{\lambda_t^{(j)} + \sigma_t^2}\right)^{\frac{n_t}{2}} \tag{3.36}$$

This is again a complicated univariate distribution, and we resort to Stan programs.

3.3.3 Summary of Markov Chain Monte Carlo Algorithm

Initialization

The initial values of the parameters are assigned by the empirical values provided by the data. In particular, at time point t, U_t takes the first r empirical eigenvectors and $\{\lambda_t^{(1)}, \lambda_t^{(2)}, ..., \lambda_t^{(r)}\}$ take the leading r empirical eigenvalues, whereas σ_t^2 is initialized as the median of the $p - r$ tail eigenvalues. The initial values of the across-group hyper-parameters are assigned by running the corresponding Gibbs sampling steps once with the initialized group-specific parameters.

Sampling Algorithm

Algorithm 4: MCMC Algorithm for Dynamic Covariance Estimation

```
Result: Samples of $U_t$'s, $\Lambda_t$'s, $\sigma_t^2$'s, $\Sigma_t$'s.
Initialization: initialize $U_t$'s, $\Lambda_t$'s, $\sigma_t^2$'s using empirical values;
for i in 1 : (Burn-in + Iterations) do
    Update the across-group parameters:
    1. Sample $A, B$ with 3.21;
    2. Update general parameters for the eigenvalue processes;
    Update the group-specific parameters:
    3. Update $U_t$ with 3.26, 3.27 and 3.28;
    4. Update $\{\lambda_t^{(1)}, \lambda_t^{(2)}, ..., \lambda_t^{(r)}\}$
    5. Update $\sigma_t^2$ with 3.36;
    Save the samples for every 5 iterations;
end
```

3.4 Simulation Results

For the simulation results, we aim to show that our method is able to recover the correct parameters when the data are indeed generated from the specified model, which includes the parameters for the dynamic process on the Stiefel manifold, the eigenvectors, and the eigenvalues. The model should work nicely when the eigenvalues are well-separated, which suggests that the eigenvectors be clearly identified.

In the following simulation study, we consider a three-factor dynamic model in the 100 dimensional space, namely $p = 100$, $r = 3$. We assume that there are 30 time points for the temporal evolution process and there are twenty observations at each time point. The true distribution that governs the first-order Markov dynamics on the Stiefel manifold is a generalized Bingham distribution with $A = B = \text{diag}(\{50, 20, 10\})$. For the eigenvalues, the truth comes from stationary distributions of three first-order auto-regressive processes

$$\lambda_t^{(j)} = c^{(j)} + \varphi^{(j)} \lambda_{t-1}^{(j)} + \epsilon_t^{(j)},\ \epsilon_t^{(j)} \sim N(0, (\sigma^{(j)})^2),\ j = 1, 2, 3.$$

The parameters for the auto-regressive processes are displayed in table 3.1. The MCMC algorithm was conducted for 2000 iterations, with the first half as burn-in samples. In addition, the Stan functions have a smaller burn-in period with 50 burn-in samples.

	c	φ	σ^2
First Eigenvalue	100	0.7	10
Second Eigenvalue	50	0.7	5
Third Eigenvalue	10	0.7	2

Table 3.1: Parameters for the first-order autoregressive processes

3.4.1 Smoothness between Eigenvectors over Time

Under this setup, the eigenvalues are well-separated apart, which means the directions of the eigenvectors can be relatively unambiguously detected. To see how aligned the estimated samples are across time, we consider the metric

$$x_t^{(j)} = |\langle v_t^{(j)}, v_{t+1}^{(j)} \rangle| \tag{3.37}$$

for $j \in \{1, 2, 3\}$ across t. $x_t^{(j)}$ will be close to 1 if $v_t^{(j)}$ and $v_{t+1}^{(j)}$ are aligned, and close to 0 if they are almost orthogonal to each other, which is common in the high-dimensional space.

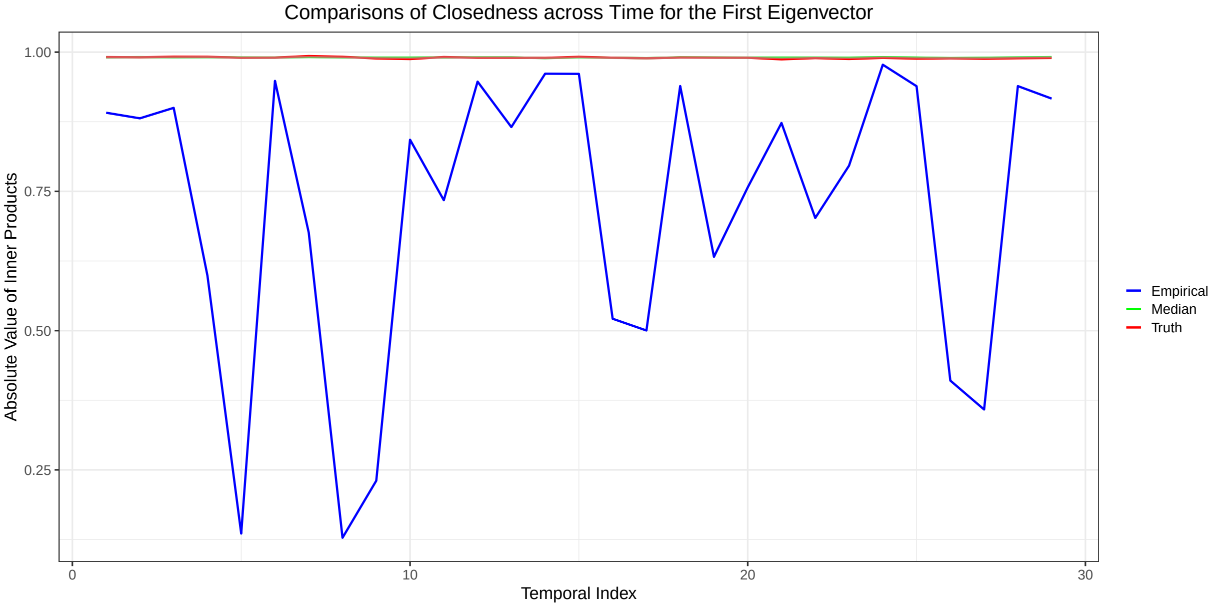
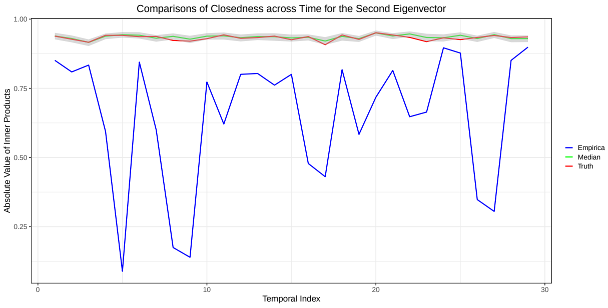


Figure 3.3: First Eigenvector



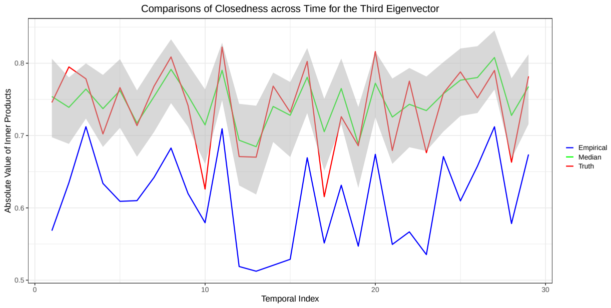


Figure 3.4: Second Eigenvector

Figure 3.5: Third Eigenvector

Computing $x_t^{(j)}$ for all the empirical estimates and posterior samples. The empirical values are denoted by blue lines, while the true values are in red. For the Bayesian samples, we compute the median of all the remaining samples and show them in green, together with a grey ribbon characterizing the uncertainty using the 95% posterior interval. The first two eigenvectors are easily detectable since they correspond to larger eigenvalues. The magnitudes of $A_{11}, A_{22}, B_{11}, B_{22}$ also indicate dynamic processes where the sequential eigenvectors are more closely aligned. For the third eigenvector, its direction is slightly more difficult to decide, and the eigenvectors are not as closely aligned as other eigenvectors. Nevertheless, we can observe that connecting them via a dynamic process can effectively boost the performance, as the median of the estimates are more closer to the truth than the raw principal component estimates.

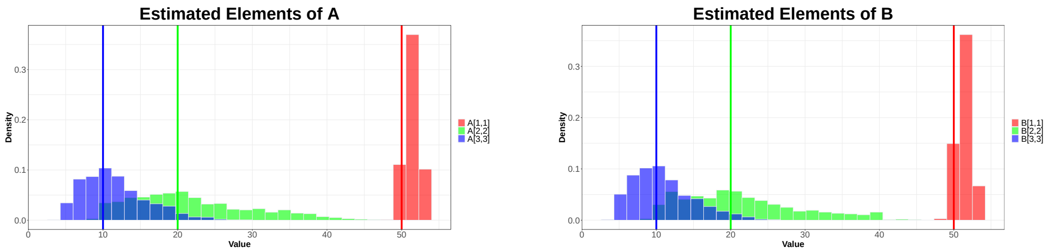
To sum up, in all cases, the median estimates matches the truth much better in contrast to the noisier empirical estimates, and the 95% posterior intervals recover the smoothness between the nearby eigenvectors well. This demonstrates the effectiveness of our shrinkage approach in utilizing the information across time points.

Notice that the method works more effectively when the eigenvalues are spaced out, and less so when the eigenvalues are similar. In that case, the eigenvectors are not clearly identifiable, hence the autoregressive model on the eigenvectors will produce results with high uncertainties.

3.4.2 Estimation of A and B

Next we want to check if the parameters for the generalized Bingham distribution are reasonably recovered. This is not always achieved, especially when the diagonal values of A and B are not spaced apart. However, we generally care more about how the smoothness of the eigenvectors are recovered rather than the parameters themselves.

In figure 3.6, the density plots for the Bayesian samples of the diagonal elements of A and B are shown. The thick vertical lines represent the true values of the respective diagonal elements, which are close to the modes of the samples. Hence, under the situation where eigenvalues are separated out, our method achieves satisfying results for recovering the smoothness parameters.



(a) Elements of A

(b) Elements of B

Figure 3.6: Estimated Samples for A and B

3.4.3 Estimation of Eigenvalues

The eigenvalues are estimated using shrinkage Bayesian first-order auto-regressive models. Intuitively, our estimates would be more smooth than the noisy empirical estimates, since the variances of the noises possess shrinkage priors. The comparisons are demonstrated as below, with the lines and ribbons carrying the same meanings as before. Again we can clearly observe from Figure 3.7, 3.8 and 3.9 that the shrunk results provide smoother estimates which match better with the true eigenvalues, regardless of the magnitudes. Hence our method successfully utilizes information across time points to obtain more accurate results.

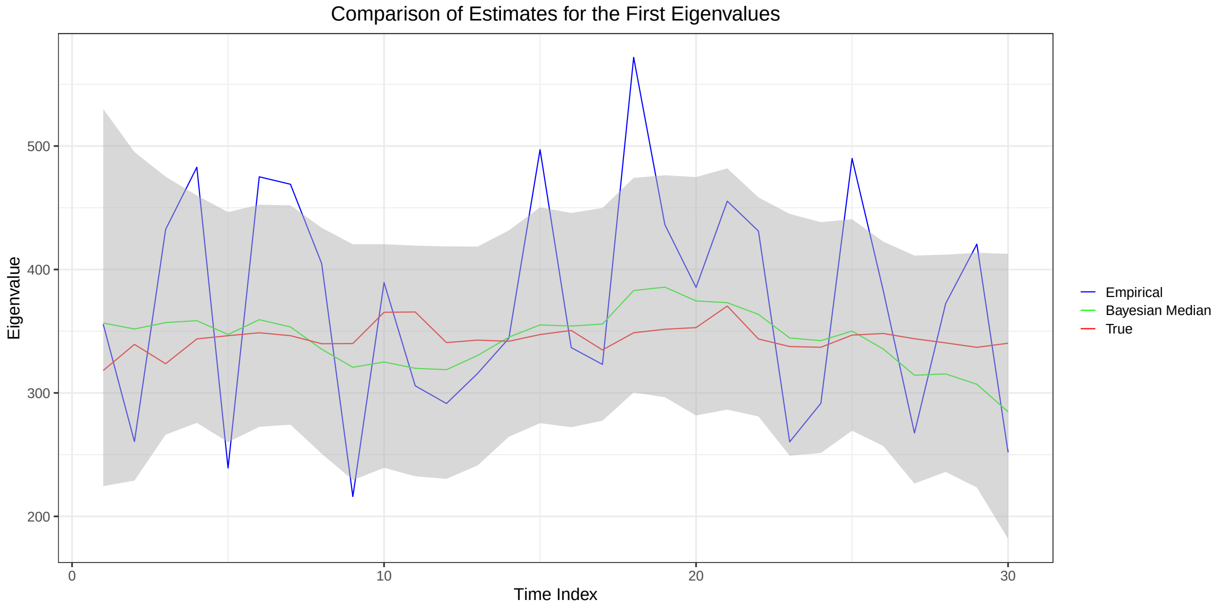


Figure 3.7: First Eigenvalue

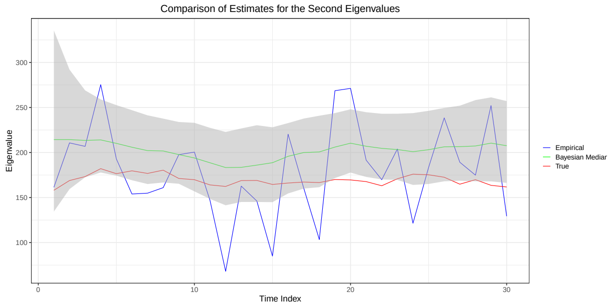


Figure 3.8: Second Eigenvalue

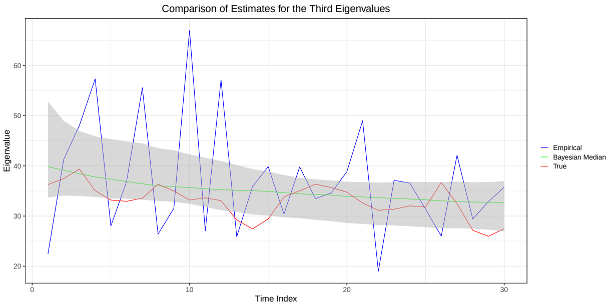


Figure 3.9: Third Eigenvalue

3.5 Example

Our dataset is retrieved from Kaggle, the famous website for data analysis and machine learning competitions, with the link https://www.kaggle.com/sevgisarac/temperature-change. The dataset describes the mean surface temperature change of the domains from Food and Agriculture Organization of the United Nations (FAO). The changes were recorded by country and updated annually from 1961 to 2019. We take the monthly records and divide the data into time points of five years, with the last time point having only four years. This is based on the assumption that the factors that explain the temperature changes are moving slowly. As a result, most of the time points have 60 observations. As for the countries, the original dataset consists of information of 190 countries and 37 other territories. However, some regions have missing records and we decide to remove them. In the end, 137 regions with full historical records are left for further analysis.

The model focuses on estimating the covariance matrices of the residuals, therefore for practical datasets, we need to first remove the mean levels. Let the raw data for time points $1, 2, ..., n$ be represented by $Y_1, Y_2, ..., Y_n$, where each Y_t is a $p \times n_t$ matrix, and each column is an observation of dimension p. For instance, one time point might contain observations for a whole month and each column corresponds to one day. Our covariance model assumes that the means of the variables are 0, so we can demean the raw observations by subtracting the row-wise means. ‘We remove the mean levels at each time point to make the data satisfy our model assumption. The algorithm was implemented with 2000 iterations with the first half as burn-in samples. After conducting exploratory data analysis on the empirical eigenvalues, we decided to fit a five factor model.

Factor Interpretations

We visualize the factor loadings on the world map, trying to discover interpretations of these five latent factors. Figure 3.10 shows that the first factor measures the contrast between most European countries, especially the central European countries, and the Greenland and western African countries. It contributes to the difference between the central European countries and their neighbors, and is probably the ocean current for North Atlantic. The second factor, as shown in Figure 3.11, shows the contrast between northern European countries and countries around the Mediterranean Sea, such as southeastern European countries northern African ones. Since it is well known that the countries around the Mediterranean Sea has little rain in the summer and more in the winter, the factor driving this rainfall phenomenon is more likely to be the second latent factor. Figure 3.12 stands more of northern European countries, as well as Spain, Portugal, Morocco and Western Sahara. Hence the third factor could be the latent factor driving the difference between cold Sweden and hot dessert in Western Sahara. Furthermore, the fourth factor is characterized by the northern European countries, mideastern countries, and Mongolia. This wide-spread range indicates that this factor might not relate to local geographical features, but some cultural or agricultural customs. Lastly, the fifth factor stands for Mongolia mostly. More transparent factor interpretations can be consulted with geographical and Meteorological experts.

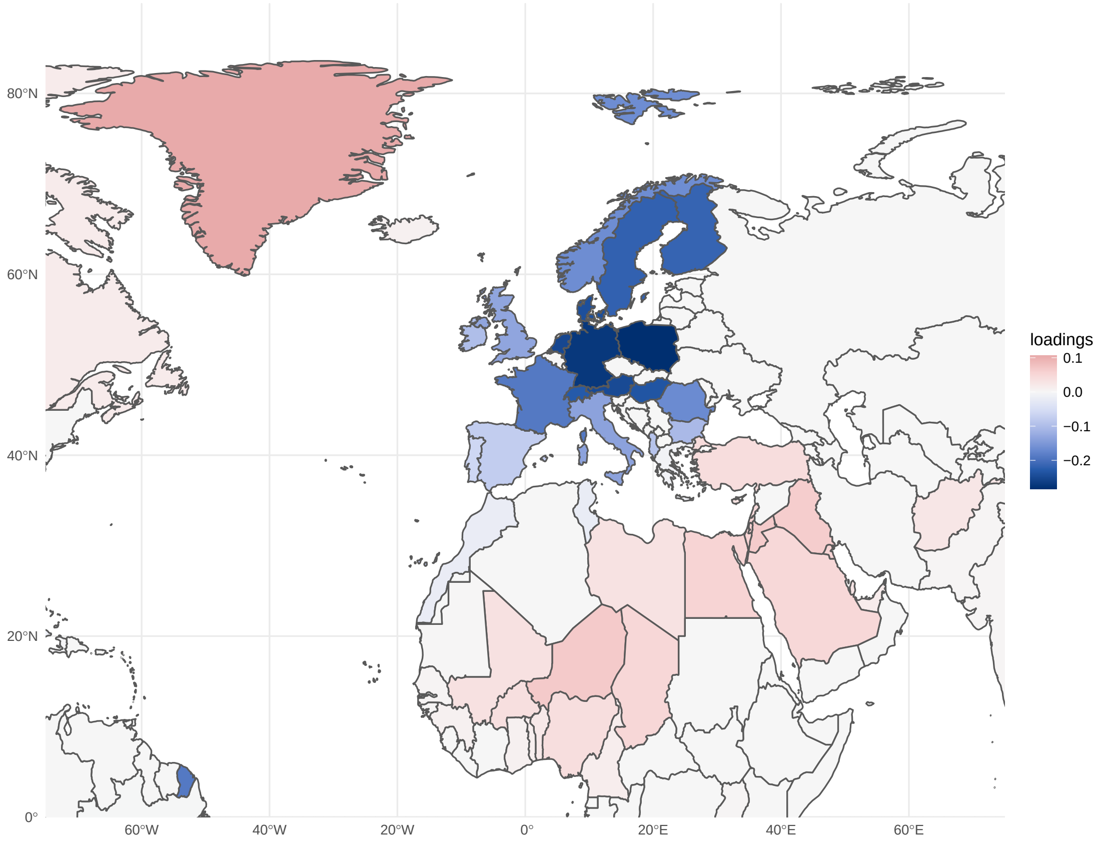


Figure 3.10: Loadings for the First Factor

Dynamic Factor Variances

For the factor variances, we adopt a shrinkage first-order autoregressive model. The results show that there are no significant changes in the factor variances over time for the above-mentioned five factors, indicating that the factors fluctuated on a similar scale over time.

Dynamic Factor Loadings

We also consider the estimated dynamic factor loadings on different factors for the countries. Notice the signs of the eigenvectors are not identifiable, so we can only make sense of the magnitude of the loadings and the relative sign for times series over 1961-2019. We fix the first eigenvector for the first time point, and switch the sign of the latter eigenvectors if the inner product for the nearby eigenvector is negative. Meanwhile, for each latent factor, we selectively visualize the countries with larger average loadings, since they are more exposed to this factor. Figure 3.15 shows that most countries have stable loadings across time. In particular, these countries are located in the European area, middle eastern and north Africa. The loadings of Albania and Bulgaria tend to increase in magnitude, whereas that for Poland is decreasing in magnitude. This indicates that Albania and Bulgaria are more affected by the factor, and Poland is less sensitive to the first latent factor. Moreover, it is surprising that Greenland, which is the farthest north of Europe, is negatively correlated with all other European countries, and positively correlated with some middle eastern countries. Also notice that even though Greenland and the northern European countries are geographically close, but they are the most negatively correlated regions on the first latent factor.

In Figure 3.16, most countries also have stable loadings, with Romania and Austria having decreasing loadings in scale across time, and Saudi Arabia increasing loadings. This means the second latent factor is exerting more influence on the temperature change in the mid-eastern countries, while less influence on some European countries.

In Figure 3.17, it is apparent that the loadings for Norway and Sweden are much larger than the rest, and they are decreasing slightly in scale with time. The loadings for Afghanistan and Western Sahara are dropping over time in magnitude and that for Turkey, Algeria, Poland, Denmark are increasing. Compared with the above three figures, on the fourth and fifth factor, the loadings are changing more severely. Figure 3.18 shows two changing patterns, concave and convex. Concave pattern involves East Asian countries such as Japan, Hong Kong, Macao, Taiwan and mainland China, whereas the convex pattern involves middle eastern countries like Iraq, Kuwait, Saudi Arabia, Qatar, and United Arab Emirates, etc. This shows that the effect of this factor is getting more significant in the East Asian countries, and losing its force in the mideastern countries. Similar conclusion can be found for the last factor, with Mongolia, Republic of Korea, and Democratic People's Republic of Korea as the most highly influenced countries. We can see the dynamics for Iraq, Bahrain and Kuwait all decrease first until around 1992, and bounced back afterwards. Canada and Netherland follow another pattern, where the loadings first went up, then down, and went up again. Meanwhile, Japan, Republic of Korea, and Democratic People's Republic of Korea all moved at similar paces, while it is also true for Mainland China, Hong Kong, Macau and Taiwan. It is worth further investigation for these similar patterns.

In summary, the dynamics for the factor loadings serve as another source to determine the meanings of these latent factors. They also characterize how the influence of these factors change over time on different countries. There are some similar changing pattern worth further discussion and they also provoke more insightful questions about the way the loadings change over time.

3.6 Discussions

The method proposed in this paper successfully achieves information sharing for the eigenvectors and eigenvalues through autoregressive models. It works better when the eigenvalues are spaced out, in which case the eigenvectors would be more identifiable. When the eigenvalues are similar to each other, the directions are difficult to obtain, thus contaminates the autoregressive results. The approach represents a class of models. Apart from the illustrated linear regression and first-order autoregressive models, other models for the eigenvalues can be explored and incorporated to add more models into this class. In addition, since the eigenvalues are strictly positive, we can also propose the eigenvalue model on the logarithmic scale. As for the autoregressive model on the Stiefel manifold, it can be extended by using alternative distributions, such as the Von-Mises Fisher distribution or Watson distribution.

Our current model focuses on the centered observations and assumes the observations have zero means. This is a simplified assumption, and it can be extended to incorporate the mean information. Franks (2020) demonstrates the benefits of obtaining better covariance estimates by incorporating the mean information. Future work can involve modeling jointly the mean levels and the covariance matrices, and taking advantages of the correlation between the two. Similar ideas can be found in Niu and Hoff (2019) and Pourahmadi (1999).

3.7 Appendix

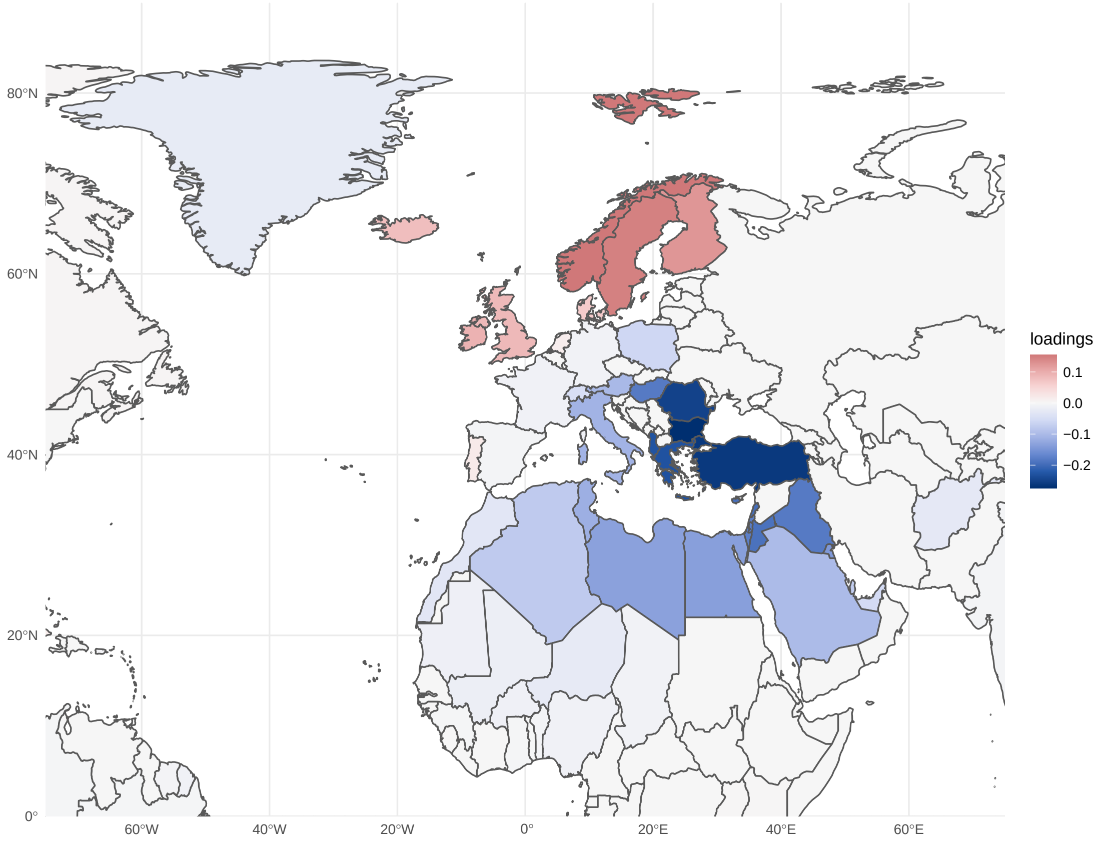


Figure 3.11: Loadings for the Second Factor

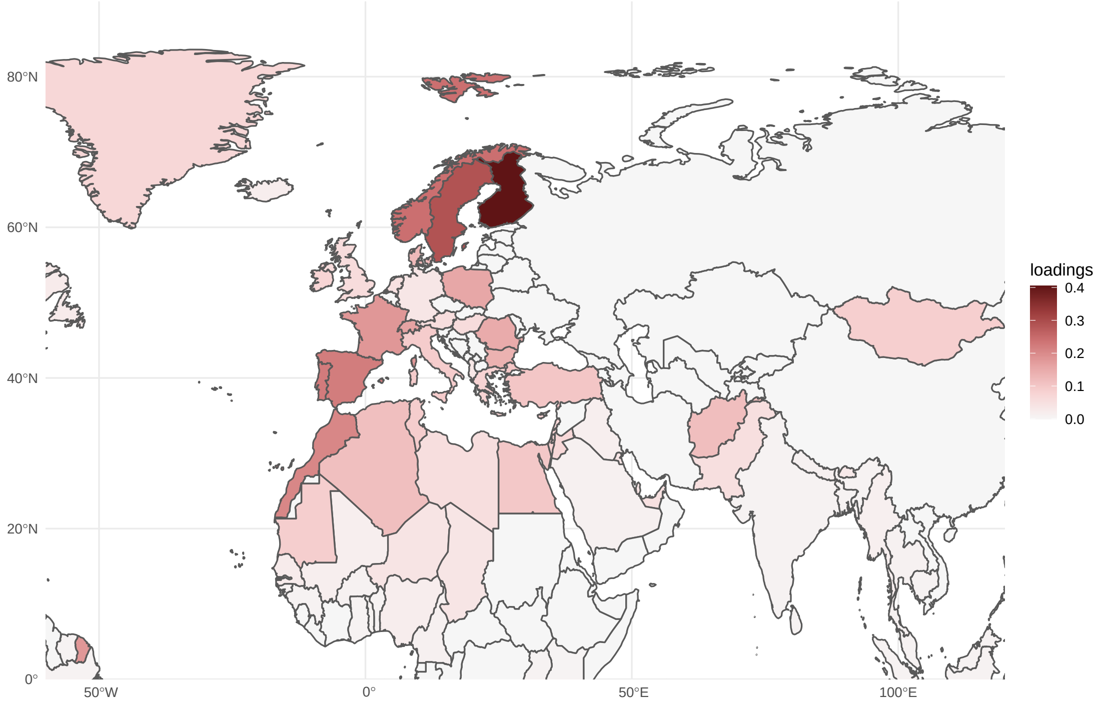


Figure 3.12: Loadings for the Third Factor

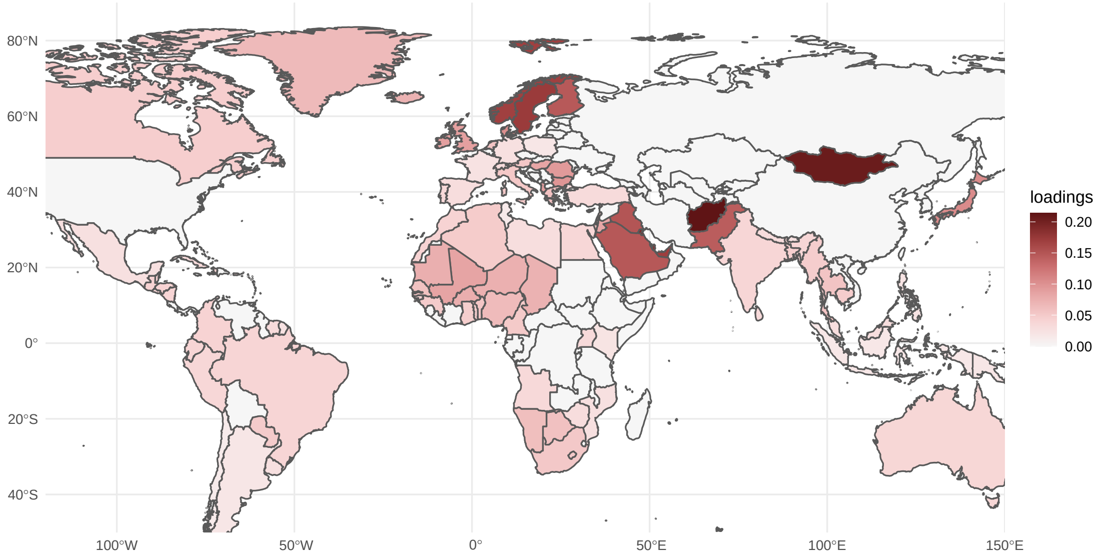


Figure 3.13: Loadings for the Fourth Factor

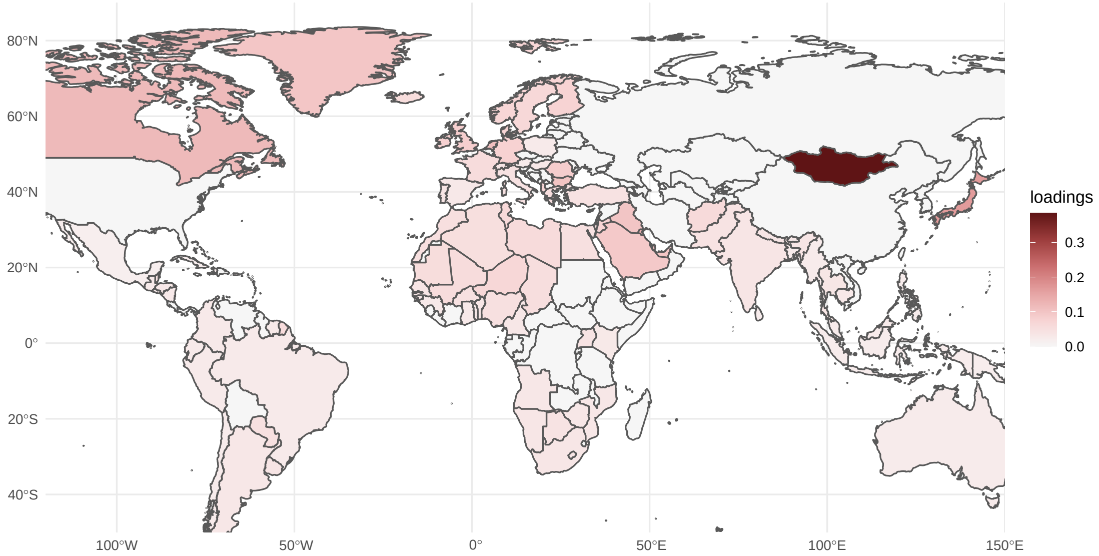


Figure 3.14: Loadings for the Fifth Factor

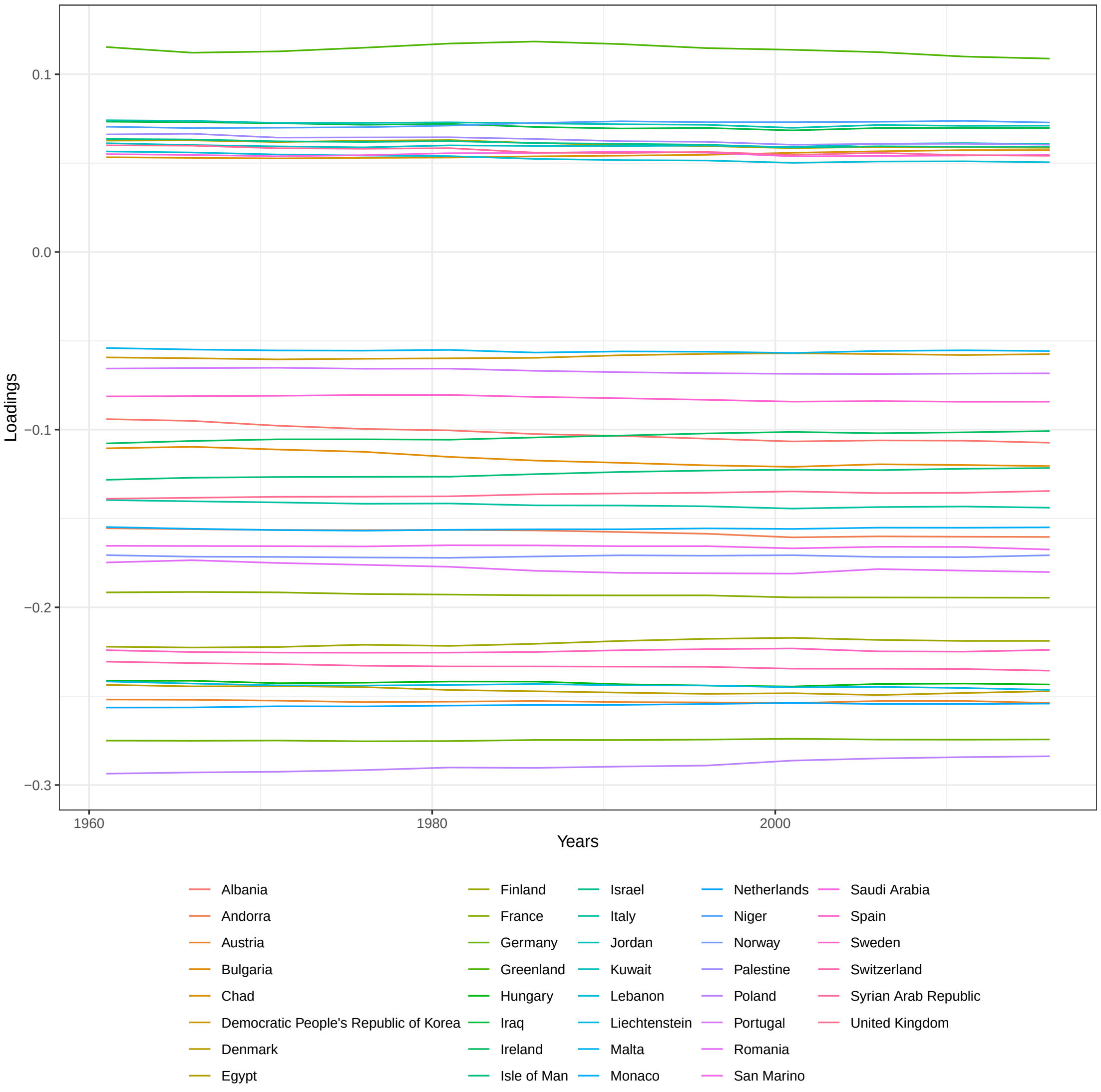


Figure 3.15: Dynamic Loadings on the First Factor

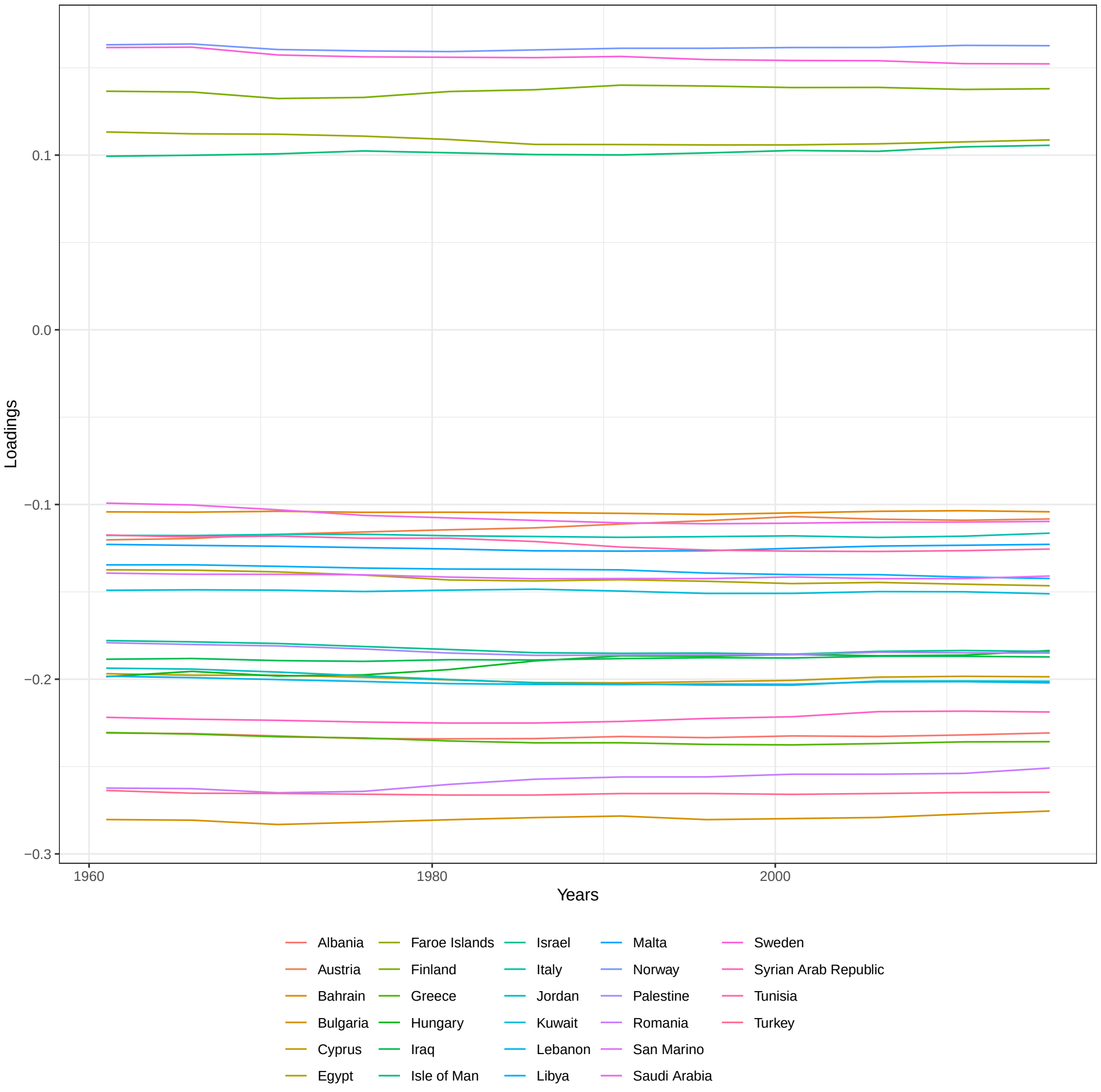


Figure 3.16: Dynamic Loadings on the Second Factor

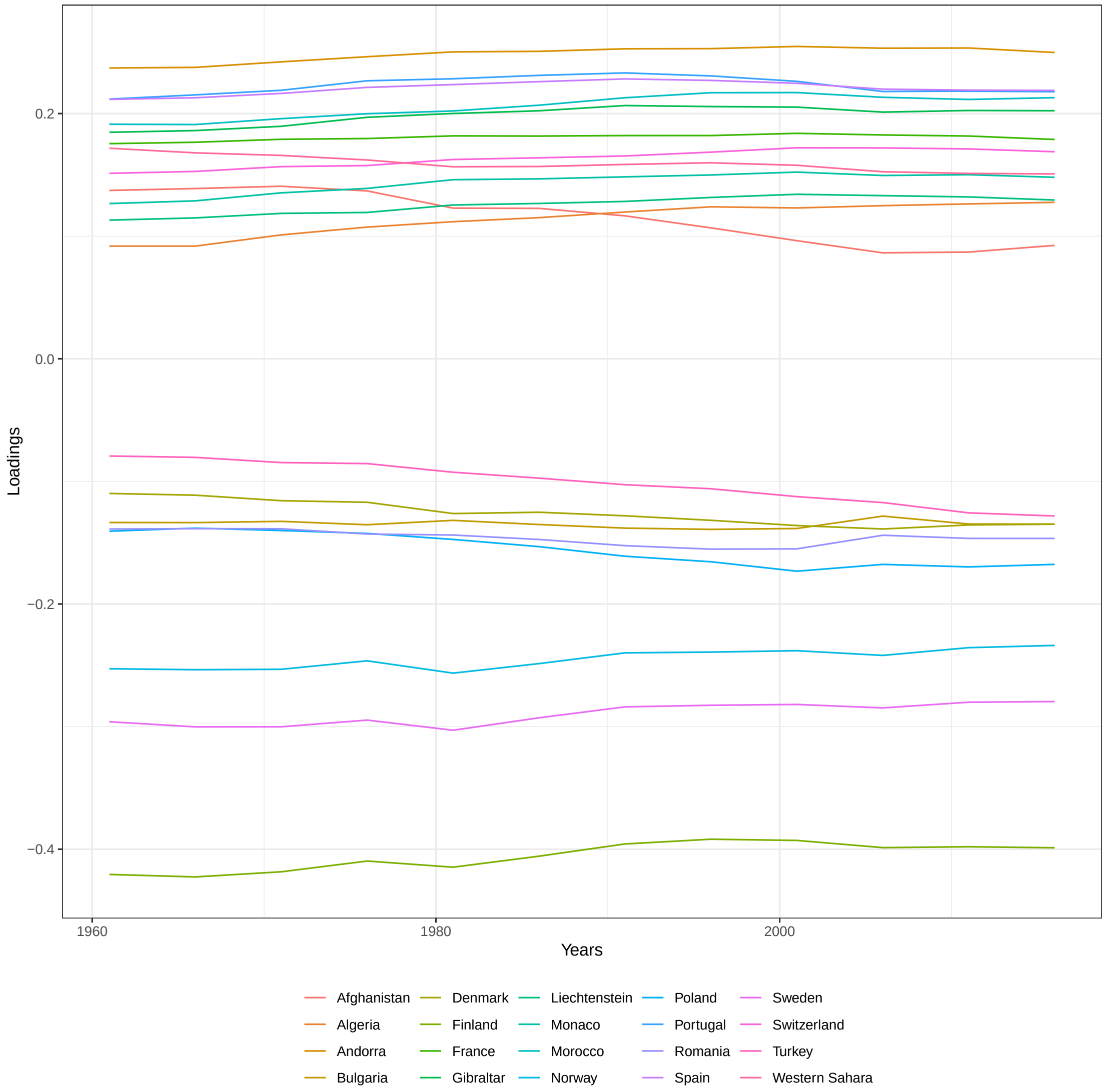


Figure 3.17: Dynamic Loadings on the Third Factor

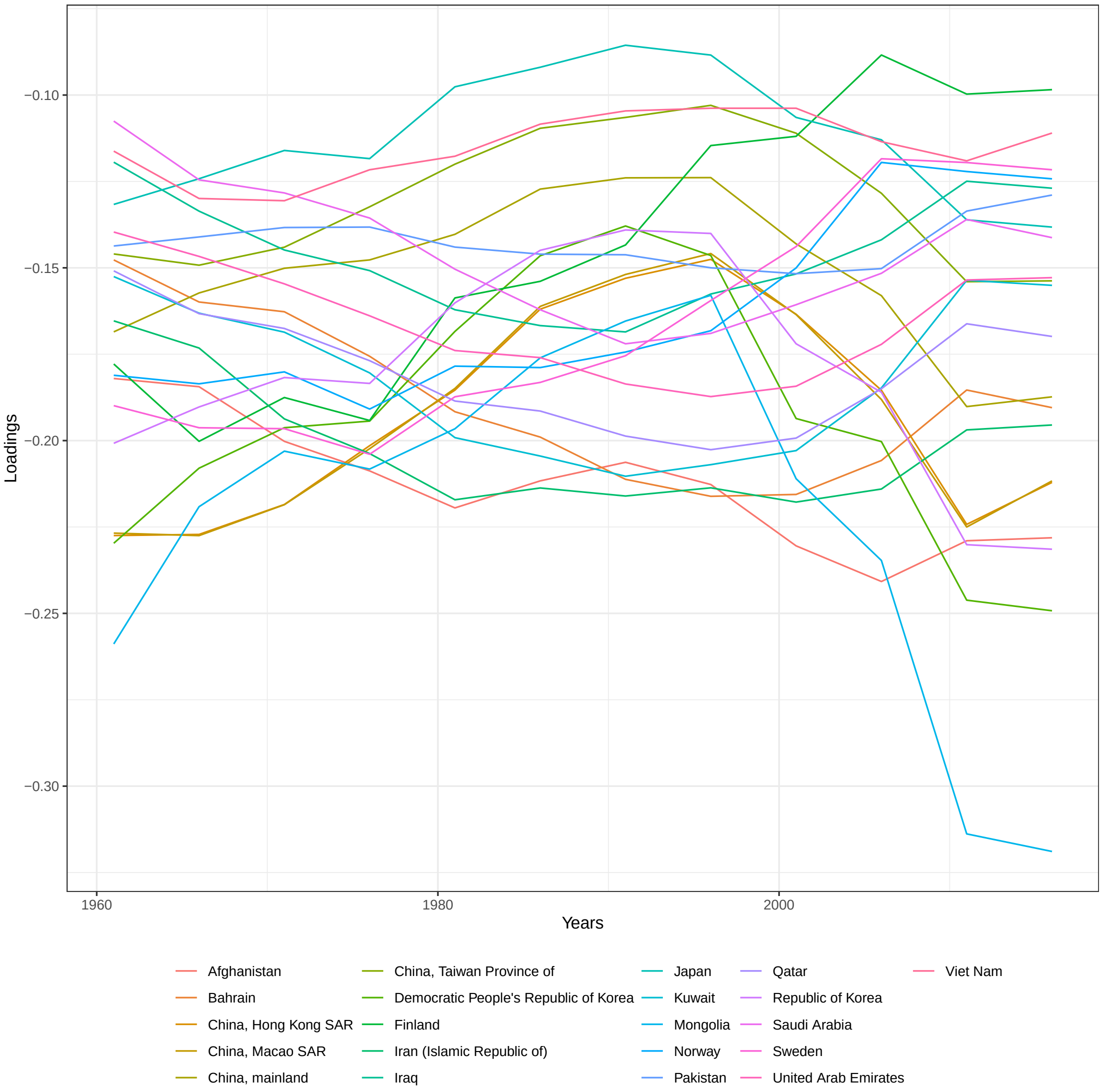


Figure 3.18: Dynamic Loadings on the Fourth Factor

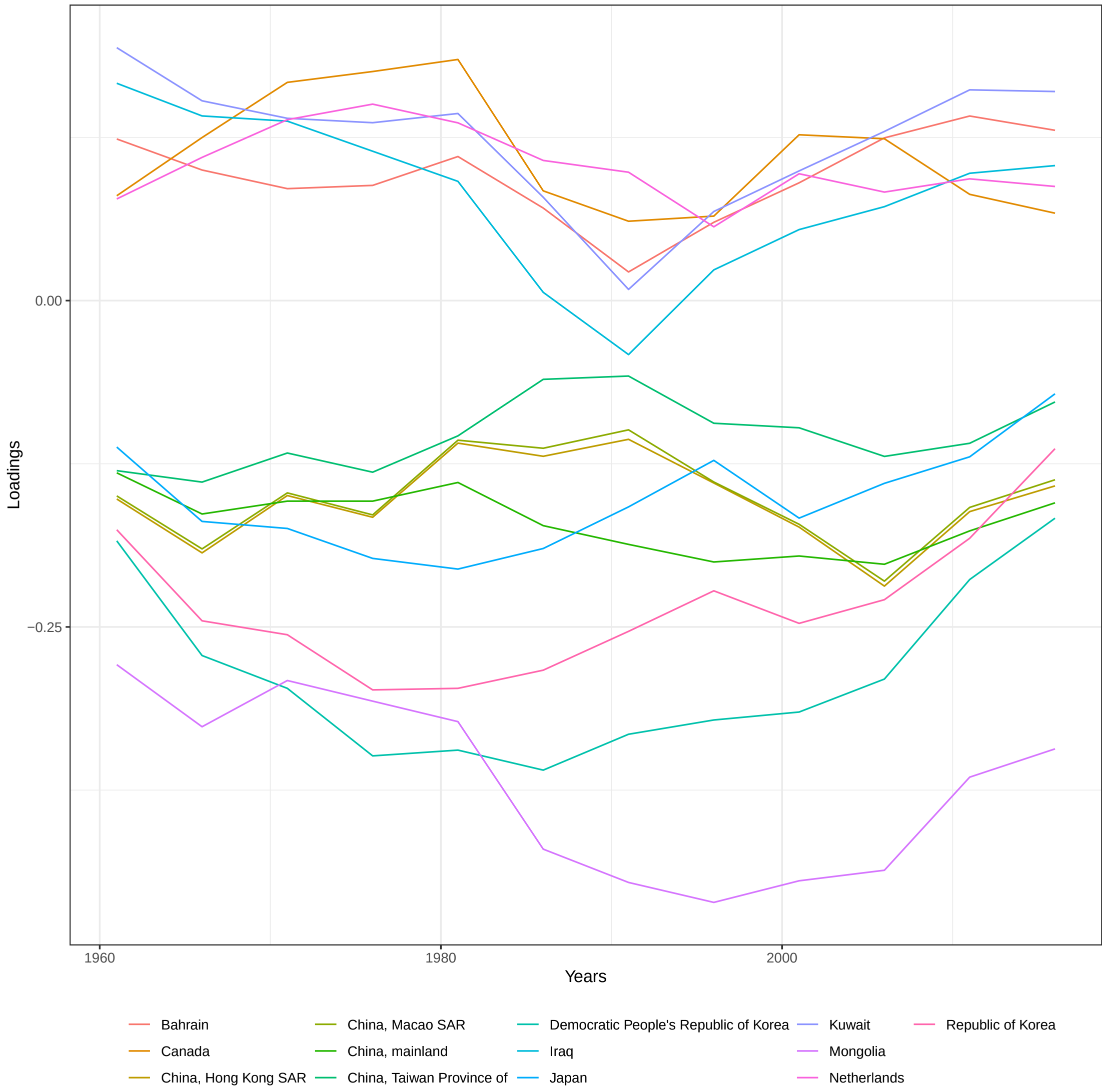


Figure 3.19: Dynamic Loadings on the Fifth Factor

Chapter 4

Bayesian Covariance Modeling for Financial Markets and its Implications

4.1 Introduction

Variance-covariance matrices characterize the linear co-movements of every pair of assets in the financial market. They are of great importance in understanding the financial markets, and serve as an indispensable component in the famous Markowitz mean-variance optimization framework for asset allocation. Since the market environment is constantly evolving over time, the portfolio weights and risk measures need continuously updating. Therefore, accurate time-varying covariances are essential inputs for many dynamic hedging and risk management models, such as Harris et al. (2017) and Engle et al. (2019). When the number of assets under consideration is large relative to the number of historical return observations, the sample covariance matrix is singular and fails to be full rank. In addition, it is a biased and high variance estimator for the true covariance matrix. To overcome this difficulty, various covariance estimation techniques for high-dimensional inference have been proposed. Ledoit and Wolf (2004) provides a shrinkage estimator by a convex linear combination $\delta F + (1-\delta)S$, where $\delta \in (0,1)$, S is the sample covariance matrix, and F a highly structured estimator. Aguilar and West (2000) assumes low rank factor models, and utilizes the vector autoregression for the factor dynamics and stochastic volatility. Engle (2002) introduces a new class of multivariate GARCH estimators, called dynamic conditional correlation, that generalizes the constant conditional correlation model in Bollerslev (1990). Time-varying covariance estimation for high-dimensional data remains an attractive and challenging task for financial practitioners.

Let r_f be the risk-free rate, r_M be the return of the market portfolio, and r_A be the return of the target asset. The static capital asset pricing model (CAPM) states that the excess return of an asset is proportional to that of the market portfolio. Mathematically,

$$E(r_A - r_f) = \beta_A E(r_M - r_f). \tag{4.1}$$

Beta measures the ratio of an asset's excess return with respect to the excess return of the market portfolio. Historically there have been discussions over the constancy of beta. Jensen et al. (1972) established the standard CAPM based on the assumption that beta is constant over a period. However, evidences against this fundamental assumption were discovered to demonstrate the dynamic essence of market beta. In the 1970s, Blume (1971) showed that the betas across time of a portfolio with a large number of securities, rather than the individual assets, were relatively stationary. Blume (1975) further investigated the reasons behind. They use bivariate normal distributions to model β_{it} and $\beta_{i,t+1}$, the true betas at time t and $t+1$ for the ith asset. The same assumption is applied to $\hat{\beta}_{it}$ and $\beta_{i,t+1}$, the estimated beta at time t and the true beta at time $t+1$ for the ith asset. They concluded that it is due mainly to the real non-stationarities in the betas of individual securities across time. On the other hand, Vasicek (1973) proposed a Bayesian approach to find the posterior approximate normal distributions of the betas with prior information.

With the accumulation of empirical evidences, people developed the consensus that static betas do not suffice for explaining the cross-section of average returns on stocks. Jagannathan and Wang (1996) discussed the conditional CAPM model as follows:

$$E[R_{it}|I_{t-1}] = \gamma_{0,t-1} + \gamma_{1,t-1}\beta_{i,t-1}, \tag{4.2}$$

where $\beta_{i,t-1}$ is the conditional beta of asset i defined as

$$\beta_{i,t-1} = Cov(R_{i,t}, R_{m,t}|I_{t-1})/Var(R_{m,t}|I_{t-1}). \tag{4.3}$$

$\gamma_{0,t-1}$ is the risk-free rate at time $t-1$, $\gamma_{1,t-1}$ is the conditional market risk premium and I_{t-1} denotes the information that the investors possess at time $t-1$ to make decisions. Ismaila and Shakranib (2003) investigated evidences that support the dynamics of beta using Islamic unit trusts data in Malaysia. Alternatively, Bali et al. (2017) writes the dynamic conditional beta as

$$E[R_{i,t+1} - r_{f,t+1}|I_t] = E[\beta_{i,t+1}|I_t]E[R_{m,t+1} - r_{f,t+1}|I_t], \tag{4.4}$$

and it further examines its significance in predicting the cross-sectional variation in stock returns. They generated time-varying conditional betas for all stocks trading at NYSE, AMEX, and NASDAQ, and conducted portfolio level analysis and firm-level cross-sectional regressions. The findings confirms that there is a positive and significant link between the dynamic conditional beta and future stock returns. Empirically, Horváth et al. (2020) studied the time-varying beta in factor models in the Chinese market, while Šmídl and Quinn (2007) investigated and compared the performance of static and time-varying beta of Fama-French five factor models in Indonesia and Thailand.

Beta measures the risk that an asset is exposed to based on the market portfolio. The dispersion of beta of p companies is defined as

$$d(\beta) = \sqrt{\frac{1}{p}\sum_{i=1}^{p}\left(\frac{\beta_i}{\mu(\beta)} - 1\right)^2}, \tag{4.5}$$

in Goldberg et al. (2018), where $\mu(\beta) = \frac{1}{p}\sum_{i=1}^{p}\beta_i$. For a vector with equal entries, the dispersion is 0. Therefore, dispersion can be viewed as a measure of how divergent the vector is from the dispersionless vector of the same length. Lahtinen et al. (2018) constructed portfolios of stocks with highly dispersed betas and low dispersion betas, and found that the former outperforms the latter. In practice, the estimation process of beta under small samples will inevitably introduce sampling bias, which boosts the beta dispersion, and this bias will distort the Markowitz minimum variance portfolio, leading to extreme positions in the portfolio composition. To correct that, Goldberg et al. (2018) proposed an optimization-based approach, whereas Goldberg et al. (2020) introduced the GPS adjustment that shrinks empirical beta estimates toward one. Alternative methods for better estimates of beta remains an open question.

Volatility is another crucial measure characterizing the market behavior, it indicates how much the stock market's overall returns fluctuates up and down. Schwert (1989) discussed the reasons why stock market volatility are changing in time. We are interested in the degree to which the volatility vary in time. Furthermore, the relationship between volatility and beta dispersion has aroused interests among researchers. Campbell and Lettau (1999) studied the time series volatility of daily market returns and the dispersion on industry portfolios relative to the market, and they found that the dispersion and volatility move together. Stivers (2003) found a sizeable positive relation between firm return dispersion and future market-level volatility in

U.S. monthly equity returns. Recently, Demirer et al. (2019) controled the state of the economy, and conducted bivariate and multivariate nonlinear causality tests from equity return dispersion to stock market volatility and excess returns. They concluded that both return dispersion and business conditions are valid joint forecasters of stock market volatility and excess returns. The relationship between volatility and dispersion has potential predictive power, and has become a topic worth careful study.

In this paper, we provide a novel model to simultaneously investigate the above-mentioned topics in the financial market. Our Bayesian covariance model is closely related to dynamic factor models, for which there has been a vast amount of literature. Geweke (1977) proposed dynamic factor models for modeling cross-sectional data, and Chow et al. (2011) constructs dynamic factor models with vector autoregressive relations and time-varying cross-regression parameters at the factor level. Recently, Forni et al. (2000) further generalizes the idea to factor models with infinite dynamics and nonorthogonal idiosyncratic components. From the Bayesian perspective, Aguilar and West (2000) develops the Bayesian MCMC algorithm for vector autoregression on the factors and stochastic volatilities. Most of the dynamic factor models are based on vector autoregressive processes due to its convenience. It is also well known that there is identifiability issue since one can always multiply a constant for the factor variance and divide it from the factor loadings. One way to constrain the model is by restricting the norm of the columns of the factor loadings matrix to 1. Now if we adopt the common assumption that the factors are uncorrelated, the factor loadings matrices become orthogonal matrices. In particular, our spirit is the same as that in Franks and Hoff (2019), with a spiked covariance model at each time point, and sharing information across groups. In terms of modeling the time series on the factors, instead of the common practice with vector autoregressive processes, we consider the eigenvector matrices as orthogonal matrices evolving on the Stiefel manifold, as mentioned in 3.10. For the stochastic volatilities, a two-stage transition model is proposed that accounts for both the normal movements and outliers during more volatile periods.

In our approach, the estimated loadings for the first eigenvector provides a scaled version of the market beta, and the estimated beta can be obtained by rescaling the loadings to have a unit mean. The shrinkage effects achieved by borrowing neighboring information will mitigate the high variation induced by the high-dimensional inference and reduce the bias. Moreover, the model offers an intuitive geometric interpretation as a vector rotating over time. The dynamic beta sequence for each individual asset and the beta dispersion can be explored accordingly.

4.1.1 Contributions and Overview

To the best of our knowledge, the model proposed in this paper is the first attempt for Bayesian modeling for time-varying covariance matrices on the Stiefel manifold. Other similar methods including Chikuse (2006) and Yang and Bauwens (2018), which propose state space models with latent variables evolving on the Stiefel manifold. We contribute to the literature of high-dimensional covariance matrix estimation by introducing a full Bayesian autoregressive model on orthogonal matrices, in addition to the common practice of vector autoregressions. Our method offers deeper insights in how factors evolve over time as rotations of orthogonal matrices. The shrinkage parameters are completely data-driven without human intervention, and the effects by utilizing information from neighboring time points provide more reliable results compared with the empirical results. More importantly, our approach separately models the dynamics of eigenvectors and eigenvalues, and allows unrelated priors and separate parameters that prevents the introduction of correlation through prior knowledge and model structure. The model is particular useful for problems where people want to discover the relationship between eigenvectors and eigenvalues, and this separation qualifies it a better candidate for investigating the relationships between market volatility and beta dispersion.

In Section 2 we introduce the model and describe how to construct a time series model on the Stiefel manifold. In Section 3 we describe the details about the Markov Chain Monte Carlo algorithm and sampling procedures for the Gibbs steps. To make sure our model achieves reasonable shrinkage results as expected, in Section 4 we construct a simulation example, for which the model results are more smooth and closer to the underlying known truth. We further apply the new approach on historical returns for S&P500 data in Section 5. We interpret the dynamics of market beta for selected companies, and characterize the relationship between market volatility and beta dispersions.

4.2 Model

In this paper we propose a novel method to address the high-dimensional time-varying covariance estimation problem. Financial returns data for p assets is collected at T time points, with n_t observations at the tth time point. Our goal is to estimate the T corresponding covariance matrices simultaneously. In practice, randomness in the returns can be considered as a result of multiple accumulative effects from different sources. Hence we model the distributions of the returns as multivariate normal distributions. Probabilistically, for time point t, the n_t observations are assumed to be independent and identically distributed following $N(0_p, \Sigma_t)$. We stack the observations column-wise to create the $p \times n_t$ data matrix Y_t at time t, $n_t \ll p$. Then

$$Y_t \sim N(0_{p\times n_t}, \Sigma_t \otimes I), \tag{4.6}$$

and $S_t = Y_t Y_t^T$ follows a possibly degenerate Wishart(Σ_t, n_t) distribution with density

$$p(S_t|\Sigma_t, n_t) \propto l(\Sigma_t : S_t) = |\Sigma_t|^{-n_t/2} \operatorname{etr}(-\Sigma_t^{-1} S_t/2), \tag{4.7}$$

where etr is the exponentiated trace.

Since $p \gg n_t$, the sample covariance matrix won't be a good candidate since it is not a full rank matrix. The main difficulty is that there are too many parameters (in the order of $O(p^2)$) to be estimated with only a few observations. Fortunately, according to Udell and Townsend (2019), high-dimensional data often manifest a low rank structure and can be explained by a few significant factors. Therefore, it is beneficial to postulate a spiked covariance model, which involves a low rank component representing the dominating factors, and a diagonal component that models the uninfluential factors. The diagonal component bridges the gap between the low rank structure and a full rank covariance matrix. Johnstone (2001), Paul (2007) and Baik and Silverstein (2006) study the asymptotic theory of spiked covariance model in high dimensions.

In this paper, we develop a model for quarterly S&P500 data. In particular, we consider the returns data from 1997 to the first quarter of 2021. Let X_t, U_t, Y_t, Σ_t denote the latent factors, the factor loadings, the returns data, and the covariance matrix respectively in the tth quarter for $t = 1, 2, \cdots, 97$. We assume the following spiked covariance model,

$$Y_t = U_t X_t + \epsilon_t, \tag{4.8}$$

$$X_t \sim N(0, \Lambda_t), \tag{4.9}$$

$$\Lambda_t = \operatorname{diag}(\{\lambda_t^{(1)}, \lambda_t^{(2)}, \cdots, \lambda_t^{(r)}\}). \tag{4.10}$$

$$\epsilon_t \sim N(0, \sigma_t^2 I_p), \tag{4.11}$$

$$\Sigma_t = U_t \Lambda_t U_t^T + \sigma_t^2 I_p, \tag{4.12}$$

where p is the number of stocks under consideration, and r the number of dominating factors we want to model. As demonstrated by Fama and French (1992), we can adopt $r = 3$ in the financial context. U_t is a p by r orthogonal matrix denoting the leading r eigenvectors, and $U_t^T U_t = I_r$. Notice that σ_t represents the common variance of the idiosyncratic factors, and

$\{\lambda_t^{(1)} + \sigma_t^2, \lambda_t^{(2)} + \sigma_t^2, \cdots, \lambda_t^{(r)} + \sigma_t^2\}$ represent the variances of the latent factors at time t. The diagonal component $\sigma_t^2 I_p$ is essential to turn the low rank estimator into a full rank matrix, which is a fundamental property for covariance matrices. Meanwhile, Σ_t is fully determined by three groups of parameters: $\{U_t\}, \{\sigma_t^2\}, \{\lambda_t^{(1)}, \lambda_t^{(2)}, \cdots, \lambda_t^{(r)}\}$.

As mentioned above, the scarcity of data adds extra difficulty to the problem. In fact, it is reasonable to assume that the eigenvectors and eigenvalues evolve smoothly, and the high variations in the empirical estimates due mainly to the sampling variability. In order to improve the estimates, we strive to exploit the similarities amongst the $\Sigma_t's$ and propose a shrinkage method to model eigenvectors and eigenvalues separately.

4.2.1 Autoregressive Prior for Eigenvectors

Each factor corresponds to a column vector in $U's$. Since we are considering a temporal problem with r dominating factors, it is natural to model the dynamics of the factors as r vector sequences. Meanwhile, one should not overlook the fact that the factors are assumed to be uncorrelated with each other. A naive way is to model the r sequences separately and de-correlate the estimates. However, this requires extra steps and involves order issues. Hence, instead of modeling them separately and post-process the samples, we propose a new approach that respects the orthogonality amongst the factors in the model.

Importantly, we consider the stocks that have full records for 1997-2021, and there are around 370 stocks after selection. Each time point is a quarter that contains 60-63 trading days. Therefore, estimating Σ_t is essentially a high-dimensional problem where the number of observations is less than the dimension. Assuming the factors evolve smoothly and steadily with time, each variable U_t won't be largely different from its neighbors U_{t-1} and U_{t+1}, provided they exist. In light of this assumption, it would be beneficial to take advantage of the neighboring observations to obtain more accurate estimates for the current time point, and reduce the variabilities associated with the parameters.

A bayesian autoregressive model on the Stiefel manifold would address both aspects simultaneously. Consider the orthogonal matrix U_t at each time point as a random variable on the Stiefel manifold, $U_t^T U_t = I_r$. We propose an autoregressive prior on the sequence of eigenvector matrices $U_t's$, and U_t follows a Bingham distribution parameterized by U_{t-1}. Mathematically, the conditional distribution can be expressed as

$$
\begin{aligned}
U_t|U_{t-1}, A, B \sim c(A, B)\,\mathrm{etr}(BU_t^T U_{t-1} A U_{t-1}^T U_t),& \\
A = \mathrm{diag}(\{a_1, a_2, \cdots, a_r\}), &\quad a_1 \geq a_2 \geq \cdots \geq a_r > 0 \\
B = \mathrm{diag}(\{b_1, b_2, \cdots, b_r\}), &\quad b_1 \geq b_2 \geq \cdots \geq b_r > 0,
\end{aligned}
$$

where $c(A, B)$ is the inverse of the normalizing constant for the generalized Bingham distribution. A and B are diagonal matrices shared across time points, and facilitate the alignment of the columns of $U_t's$.

The idea is analogous to Hoff (2009a), where the population of eigenvectors are modeled as samples from a common distribution parametrized by shared parameters. Our model serves as a similar counterpart for time series modeling. Here, we briefly discuss the interpretations of A and B, and the details can be further found in Hoff (2009a). In general, consider two matrices U and V in the Stiefel manifold $\mathcal{V}_{p,r}$ following the matrix Bingham distribution

$$
U \sim c(A, B)\,\mathrm{etr}(BU^T VAV^T U), \tag{4.13}
$$

where $U = [u_1, u_2, \cdots, u_r]$ and $V = [v_1, v_2, \cdots, v_r]$. Then

$$
\mathrm{tr}(BU^T VAV^T U) = \sum_{i=1}^{r} \sum_{j=1}^{r} a_i b_j (v_i^T u_j)^2 = \sum_{j=1}^{r} b_j u_j^T \left(VAV^T\right) u_j. \tag{4.14}
$$

Based on the principle of maximum likelihood, when a_i and b_j are large, u_j will be close to v_i. For instance, if a_1 and b_1 are larger than the rest, we would expect $a_1 b_1$ to be large, and the sample would be such that $u_1^T v_1$ is large, which means u_1 staying close to v_1. Since U and V both have orthogonal columns, when u_1 is close to v_1, $u_2, u_3, \cdots, u_r$ must be nearly perpendicular to v_1. Alternatively, $a_i = a_{i+1}$ implies $v_i^T u_j$ behaves the same in distribution as $v_{i+1}^T u_j$, and $b_j = b_{j+1}$ implies that u_j follows the same distribution as u_{j+1}.

4.2.2 Autoregressive Prior for Eigenvalues

Since there are only scarce observations at each time point compared with the number of features, the empirical eigenvalues contain lots of noises, making them poor estimates of the true eigenvalues. Moreover, evidences show that the first several empirical eigenvalues can be drastically volatile across time. To introduce smoothness into the model, we leverage volatility models to allow information sharing across all time points. Some common choices involving traditional statistics, such as linear regression models, Gaussian processes models and non-parametric regression models, as well as the latest popular models, such as gradient tree boosting models and deep neural networks. Notice that there is no perfect solution in the choice of models. Nonetheless, simpler models are often preferred for at least two reasons. First of all, the main purpose of the time series model is to bring the estimates closer, so as to reduce the common large sampling variances in high dimensions. Flexible models can easily overfit the data and center around the empirical values. Meanwhile, there are already quite many parameters to be estimated under the current setup. An overly flexible model, especially if parametric, will inevitably add parameters into the context, thus aggravate the scarcity of data and inject more uncertainty in the estimated results.

In the financial market, it can be seen empirically that most of the time the volatility doesn't change much. However, Schwert (1989) analyzed data monthly stock returns data during 1857-1987, and they found extreme high volatility during the Great Depression from 1929 to 1933. Recently, we also had high volatilities during the tech bubble in 2002, the global financial crisis in 2008, and the COVID-19 recession. In order to account for these two paradigms, we propose an autoregressive model with mixture residuals. The mean for the next time point is a linear function of the current value, with an innovation following a two-component Gaussian mixture distribution. Mathematically,

$$\lambda_i^{(j)} = \beta_0^{(j)} + \beta_1^{(j)}\lambda_{i-1}^{(j)} + \epsilon_i^{(j)},\ j = 1, 2, \cdots, r \tag{4.15}$$

$$\epsilon_i^{(j)} \sim \pi^{(j)} N(0, \tau_S^{2^{(j)}}) + (1 - \pi^{(j)}) N(0, \tau_L^{2^{(j)}}). \tag{4.16}$$

Here $\epsilon_i^{(j)}$ is assumed to be drawn from a low variance normal distribution (with a small standard deviation τ_S) with probability $\pi^{(j)}$, and a high variance normal distribution (with a large standard deviation τ_L) with probability $1 - \pi^{(j)}$. All the parameters $\beta_0^{(j)}, \beta_1^{(j)}, \pi^{(j)}, \tau_S^{2^{(j)}}, \tau_L^{2^{(j)}}$ are inferred from data with appropriate priors reflecting the empirical knowledge.

4.2.3 Modeling the Idiosyncratic Variance

To satisfy the full-rank assumption of the covariance matrix, we have the diagonal part $\sigma_t^2 I$ in the model. For time point t, the common trailing eigenvalue is denoted by σ_t^2. We choose not to utilize information from other time points, as opposed to estimating the principal eigenvalues. On one hand, there is sufficient information at time t to provide a good estimate of σ_t^2. Evidences show that the median of the trailing empirical eigenvalues serves as a satisfactory candidate. On the other hand, we reserve this parameter for adjusting for the uniqueness for time point t.

4.2.4 Model Summary

From the time series perspective, the model can be summarized pictorially as follows.

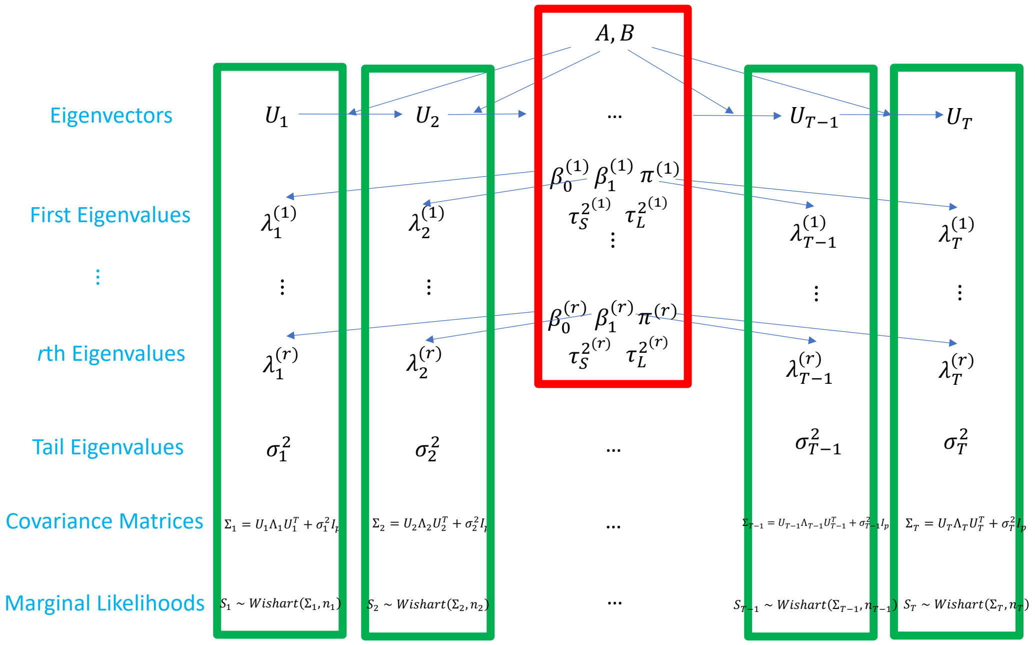

A, B
Eigenvectors
U_1
U_2
...
U_{T-1}
U_T
$\beta_0^{(1)}\ \beta_1^{(1)}\ \pi^{(1)}$
$\tau_S^{2(1)}\ \tau_L^{2(1)}$
⋮
First Eigenvalues
$\lambda_1^{(1)}$
$\lambda_2^{(1)}$
$\lambda_{T-1}^{(1)}$
$\lambda_T^{(1)}$
⋮
$\beta_0^{(r)}\ \beta_1^{(r)}\ \pi^{(r)}$
$\tau_S^{2(r)}\ \tau_L^{2(r)}$
rth Eigenvalues
$\lambda_1^{(r)}$
$\lambda_2^{(r)}$
$\lambda_{T-1}^{(r)}$
$\lambda_T^{(r)}$
Tail Eigenvalues
σ_1^2
σ_2^2
...
σ_{T-1}^2
σ_T^2
Covariance Matrices
$\Sigma_1 = U_1\Lambda_1U_1^T + \sigma_1^2I_p$
$\Sigma_2 = U_2\Lambda_2U_2^T + \sigma_2^2I_p$
...
$\Sigma_{T-1} = U_{T-1}\Lambda_{T-1}U_{T-1}^T + \sigma_{T-1}^2I_p$
$\Sigma_T = U_T\Lambda_TU_T^T + \sigma_T^2I_p$
Marginal Likelihoods
$S_1 \sim Wishart(\Sigma_1, n_1)$
$S_2 \sim Wishart(\Sigma_2, n_2)$
...
$S_{T-1} \sim Wishart(\Sigma_{T-1}, n_{T-1})$
$S_T \sim Wishart(\Sigma_T, n_T)$

Figure 4.1: Model Summary

4.3 Bayesian Parameter Estimation

This section is devoted to the technical details about inferencing the parameters. There are many parameters and they can be grouped up as across-group and group-specific parameters. We first derive the full posterior distribution and move on to the conditional distributions for the Gibbs Sampling. The full Bayesian hierarchical model can be decomposed into four components depending on their functionalities. The observations contribute to the likelihood, which are products of normal distributions for all the observations at different time points. The model parameters can be divided into two main groups, across-group and group-specific. Across-group parameters involve the hyperparameters for the generalized Bingham distribution, as well as the β's and τ's for the first-order autoregressive processes for different eigenvalue sequences. The group-specific parameters at time point t contain the eigenvector U_t, r principal eigenvalues $\lambda_t^{(1)}, \lambda_t^{(2)}, \cdots, \lambda_t^{(r)}$, and the trailing eigenvalue σ_t^2. Standard priors or conjugate priors are adopted to facilitate the inference process. Figure 4.2 shows the organization of the model parameters.

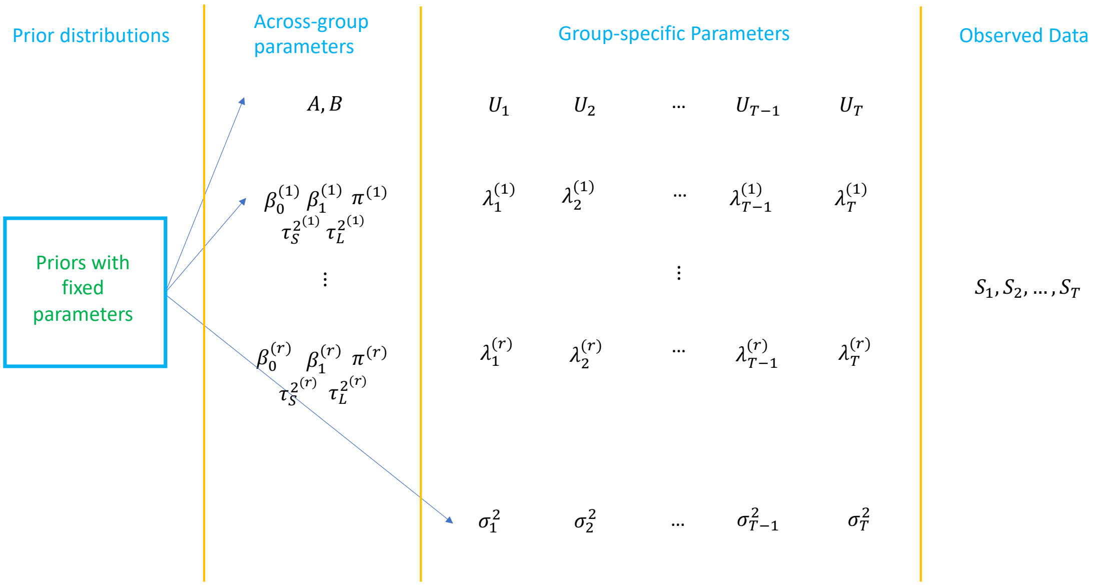

Prior distributions
Across-group parameters
Group-specific Parameters
Observed Data
Priors with fixed parameters
A, B
$\beta_0^{(1)}$ $\beta_1^{(1)}$ $\pi^{(1)}$
$\tau_S^{2(1)}$ $\tau_L^{2(1)}$
$\vdots$
$\beta_0^{(r)}$ $\beta_1^{(r)}$ $\pi^{(r)}$
$\tau_S^{2(r)}$ $\tau_L^{2(r)}$
U_1 U_2 … U_{T-1} U_T
$\lambda_1^{(1)}$ $\lambda_2^{(1)}$ … $\lambda_{T-1}^{(1)}$ $\lambda_T^{(1)}$
$\vdots$
$\lambda_1^{(r)}$ $\lambda_2^{(r)}$ … $\lambda_{T-1}^{(r)}$ $\lambda_T^{(r)}$
σ_1^2 σ_2^2 … σ_{T-1}^2 σ_T^2
$S_1, S_2, \dots, S_T$

Figure 4.2: All the Model Parameters

4.3.1 Derivation of Full Posterior Distribution

The centered likelihoods for the observations are normal, $y_t^{(k)} \sim N(0, \Sigma_t)$ for $k = 1, 2, \cdots, n_t$. The full likelihood for all the observations across all time points is

$$p(S_1, ..., S_T | \Sigma_1, ..., \Sigma_T, n_1, ..., n_T) \propto \prod_{t=1}^{T} \prod_{k=1}^{n_t} f(y_i^{(k)} | \Sigma_i), \tag{4.17}$$

where $f(y_t^{(k)} | \Sigma_t)$ represents the centered multivariate normal likelihood with covariance matrix Σ_t. Furthermore, under the spiked covariance model assumption, the determinant and the inverse of Σ_t can be expressed as:

$$|\Sigma_t| = \det(\Sigma_t) = (\sigma_t^2)^p \prod_{j=1}^{r} \frac{\lambda_t^{(j)} + \sigma_t^2}{\sigma_t^2}, \tag{4.18}$$

and

$$\Sigma_t^{-1} = \frac{1}{\sigma_t^2} \left(I_p - U_t \Omega_t U_t^T \right), \tag{4.19}$$

where Ω_t is a diagonal matrix and $w_t^{(j)} = \frac{\lambda_t^{(j)}}{\lambda_t^{(j)}+\sigma_t^2}$, $j \in \{1, 2, ..., r\}$.
Therefore, the full likelihood can be written explicitly in terms of the parameters as

$$p(S_1, \cdots, S_T|\Sigma_1, \cdots, \Sigma_T, n_1, \cdots, n_T) \propto \prod_{i=1}^{T} (\sigma_t^2)^{-\frac{n_t p}{2}} \operatorname{etr}\left(\frac{1}{2\sigma_t^2}(U_t \Omega_t U_t^T - I)S_t\right) \prod_{j=1}^{r} \left(\frac{\sigma_t^2}{\lambda_t^{(j)} + \sigma_t^2}\right)^{\frac{n_t}{2}} \tag{4.20}$$

The priors, on the other hand, are constructed in an autoregressive fashion. The prior distribution for the eigenvectors is a sequence of generalized Bingham distributions characterizing the evolution:

$$\prod_{t=2}^{T} c(A, B) \operatorname{etr}(BU_t^T U_{t-1} A U_{t-1}^T U_t). \tag{4.21}$$

The priors for the eigenvalues in this application is formed by introducing the two-component Gaussian mixture innovations:

$$\lambda_t^{(j)} - \beta_0^{(j)} - \beta_1^{(j)} \lambda_{t-1}^{(j)} \sim \pi^{(j)} N(0, \tau_S^{2^{(j)}}) + (1 - \pi^{(j)}) N(0, \tau_L^{2^{(j)}}), \quad j = 1, 2, \cdots, r. \tag{4.22}$$

As for the idiosyncratic variances, the family of inverse gamma distributions serves as good conjugate priors. We eventually decide on the uninformative prior on the positive real line to avoid biases when we are not equipped with enough domain knowledge.

Finally, the fixed priors on the across-group parameter are chosen at the discretion of the modeler. Non-informative priors on corresponding domains are selected provided no possession of prior knowledge.

The full posterior distribution is formulated via multiplying the full likelihood 4.20, priors on the eigenvectors 4.21, priors on leading eigenvalues 4.22 and priors on idiosyncratic variances, which is assumed to be non-informative in the current model, as well as all the fixed priors on the across-group parameters. The final result is too complicated to be classified as any standard distribution and it would be extremely slow and troublesome to attempt working in the whole parameter space. Therefore, we are going to apply the Gibbs sampling technique.

4.3.2 Markov Chain Monte Carlo Algorithm

Inference for A and B

Since A and B are diagonal matrices, essentially we are inferencing r parameters each. They can be done separately, subject to the order constraints. The full conditional distribution is

$$p(A, B|U_i's) \sim \prod_{i=2}^{N} c(A, B) \operatorname{etr}(BU_i^T U_{i-1} A U_{i-1}^T U_i). \tag{4.23}$$

In order to estimate A and B, we need to find an adequate numerical approximation of $c(A, B)$. According to corollary 2.1 in Constantine and Muirhead (1976):

If R_1 and S are $k \times k$ and $m \times m$ diagonal matrices respectively, $k \leq m$, with unequal elements ordered in descending order, then

$$\int_{V(k,m)} \exp(\mathrm{tr}(1/2)n\ R_1 H_1^T S H_1)(dH_1)$$
$$\sim 2^k \exp\left((1/2)n \sum_{i=1}^{k} r_i s_i\right) \prod_{i<j}^{k} (\frac{2\pi}{nc_{ij}})^{1/2} \prod_{i=1}^{k} \prod_{j=k+1}^{m} (\frac{2\pi}{nd_{ij}})^{1/2},$$

where $c_{ij} = (r_i - r_j)(s_i - s_j)$ and $d_{ij} = r_i(s_i - s_j)$ for $i = 1, 2, \cdots, k$ and $j = k + 1, \cdots, m$. $V(k, m)$ is the Stiefel manifold consisting of all $m \times k$ matrices H_1 with orthonormal columns.

In the above corollary, we take $n = 2, m = p, k = r, R_1 = A$, the first r diagonal elements of S to be B, and the last $p - r$ elements to 0. We obtain a good approximation of $c(A, B)$ as

$$2^{-r} \pi^{\frac{1}{2}(\frac{r(r+1)}{2} - pr)} \exp\left(-\sum_{i=1}^{r} a_i b_i\right) \prod_{i<j}^{r} (a_i - a_j)^{1/2} (b_i - b_j)^{1/2} \prod_{i=1}^{r} (a_i b_i)^{\frac{p-r}{2}}. \tag{4.24}$$

Notice that as parameters of a generalized Bingham distribution, A and B are non-identifiable under some transformations. As mentioned in Hoff (2009a), the likelihood $p(A, B|U_i's)$ behaves the same as that with $p(kA, \frac{1}{k}B|U_i's)$ for $k > 0$. Meanwhile, $p(A + cI, B + dI|U_i's)$ gets a density proportional to that with A and B, and that suggests only the differences amongst the diagonal elements matter. Taking these properties into consideration, we reparametrize A and B as:

$$\mathrm{diag}(A) = (a_1, \ldots, a_r) = \sqrt{w}\,(\alpha_1, \ldots, \alpha_r) \tag{4.25}$$
$$\mathrm{diag}(B) = (b_1, \ldots, b_r) = \sqrt{w}\,(\beta_1, \ldots, \beta_r)\,, \tag{4.26}$$

where $w > 0, 1 = \alpha_1 > \alpha_2 > \cdots > \alpha_{r-1} > \alpha_r > 0$ and $1 = \beta_1 > \beta_2 > \cdots > \beta_{r-1} > \beta_r > 0$. The final expression using w, $\alpha's$ and $\beta's$ can be coded into a Stan program, which can explore the parameter space well.

Inference for $\beta_0^{(j)}, \beta_1^{(j)}, \pi^{(j)}, \tau_S^{2^{(j)}}, \tau_L^{2^{(j)}}$

Since we have innovations following the Gaussian mixture distribution, it is difficult to explicitly write out the full conditional distribution of the parameters underlying the eigenvalue model.

$$p(\beta_0^{(j)}, \beta_1^{(j)}, \pi^{(j)}, \tau_S^{2^{(j)}}, \tau_L^{2^{(j)}} | \lambda_1^{(j)}, \cdots, \lambda_T^{(j)}) = \prod_{t=2}^{T} f(\lambda_t^{(j)} - \beta_0^{(j)} - \beta_1^{(j)} \lambda_{t-1}^{(j)} | \beta_0^{(j)}, \beta_1^{(j)}, \pi^{(j)}, \tau_S^{2^{(j)}}, \tau_L^{2^{(j)}}), \tag{4.27}$$

where $f(x|\beta_0^{(j)}, \beta_1^{(j)}, \pi^{(j)}, \tau_S^{2^{(j)}}, \tau_L^{2^{(j)}})$ denotes the likelihood of the two component Gaussian mixture distribution. It is difficult to write out the likelihood. However, in Stan we can directly display the model without interfering with the details.

Inference for U_t

The full conditional distribution of U_t varies, based on the location of U_t and the number of neighbors it has. There are three cases, the first time point, the last time point, and any time point in between.

1. U_1.

   It only has one neighbor, U_2. The full conditional distribution is

$$p(U_1|A, B, U_2, U_3, \cdots, U_T) \propto \operatorname{etr}(BU_2^T U_1 A U_1^T U_2) \operatorname{etr}\left(\frac{1}{2\sigma_1^2} U_1 \Omega_1 U_1^T S_1\right) \tag{4.28}$$

2. U_t, $t \in \{2, \cdots, T-1\}$.

   There are two neighbors: U_{t-1} and U_{t+1}.

$$\begin{aligned} &p(U_t|A, B, U_1, \cdots, U_{t-1}, U_{t+1}, \cdots, U_T) \propto \\ &\qquad \operatorname{etr}(BU_t^T U_{t-1} A U_{t-1}^T U_t) \operatorname{etr}(BU_{t+1}^T U_t A U_t^T U_{t+1}) \operatorname{etr}\left(\frac{1}{2\sigma_t^2} U_t \Omega_t U_t^T S_t\right) \end{aligned} \tag{4.29}$$

3. U_T.

   It has only one neighbor, the second last time point U_{T-1}.

$$p(U_T|A, B, U_1, U_2, \cdots, U_{T-1}) \propto \operatorname{etr}(BU_T^T U_{T-1} A U_{T-1}^T U_T) \operatorname{etr}\left(\frac{1}{2\sigma_T^2} U_T \Omega_T U_T^T S_T\right) \tag{4.30}$$

After algebraic manipulations, they can be unified in the general format $\operatorname{etr}(AU^TBU + CU^TDU + EU^TFU)$, where A, C, E are diagonal matrices and B, D, F are $p \times p$ matrices.

The sampling on the Stiefel manifold is challenging and there are attempts from various aspects. Some basic techniques are the rejection sampling and the importance sampling, which are only efficient for a special class of problems. In addition, Laplace approximation and variational Bayes methods are designed in the spirit of replacing the target posterior distribution with a computationally feasible alternative. Another prominent stream of thought, which is well-known as Markov chain Monte Carlo (MCMC), is based on constructing a Markov chain with the target distribution as the stationary distribution. The Metropolis-Hastings algorithm and Gibbs sampling both fall into this category. Recently, a sub-class of MCMC methods gains popularity with their ability to propose long distance moves in the state space and high acceptance rates. Being known as Hamiltonian Monte Carlo (Neal et al. (2011)), the method simulates Hamiltonian dynamics in an augmented parameter space and the projected trajectories are retained as samples.

Hoff (2009b) discusses the Gibbs sampling algorithm for sampling from the matrix Bingham-von Mises-Fisher distribution. Reparameterization was adopted to remove the built-in constraints of the Stiefel manifold. In most cases, we can use Gibbs sampling to sample the column vectors iteratively, and special treatments need to be applied for the full rank case. Pourzanjani et al. (2021) utilizes the idea of Givens representation to develop a nice algorithm in Stan. However, it takes great efforts to theoretically compute the change-of-measure term and it pays to adjust to the topological difference between the transformed parameter space and the original space. Moreover, Jauch et al. (2020b) developed a novel sampling scheme on the basis of the Cayley transformation, and Nirwan and Bertschinger (2019) works on the Householder transformation. In this paper we are going to adopt the latest, and probably the best algorithm devised by Jauch et al. (2020a), which reparametrizes the Stiefel manifold by unconstrained matrices of the same dimension. For a matrix $X \in \mathbb{R}^{p\times k}$, its singular value decomposition is denoted as $X = UDV^T$, let

$$\begin{aligned} Q_X &= X(X^TX)^{-1/2} = UV^T, \\ S_X &= X^TX = VD^TU^TUDV^T = VD^TDV^T, \\ S_X^{1/2} &= VDV^T. \end{aligned}$$

Then $X = Q_X S_X^{1/2}$ and $S_X = S_X^{1/2}S_X^{1/2}$, where Q_X is an orthogonal matrix while $S_X^{1/2}$ is a symmetric positive definite matrix. Analogous to the polar expansion $z = re^{i\phi}$ for complex numbers, $S_X^{1/2}$ is the counterpart for r while Q_X is comparable to $e^{i\phi}$.

The advantage of introducing Q_X and S_x together is that now the mapping from a real, full rank matrix X to the components (Q_X, S_X) of its polar decomposition is one-to-one, and the density f_X can be derived as

$$f_X(X) = f_{S_X|Q_X}(S_X|Q_X)f_{Q_X}(Q_X) \times J(Q_X, S_X; X). \tag{4.31}$$

In contrast with Cayley's transformation and Givens representation, where it is expensive to compute the Jacobian, $J(Q_X, S_X; X)$ is a standard result shown in Chikuse (2012).

$$J\left(Q_X, S_X; X\right) = \frac{\Gamma_k\left(\frac{p}{2}\right)}{\pi^{\frac{pk}{2}}} \left|S_X\right|^{-\frac{p-k-1}{2}}. \tag{4.32}$$

This convenience makes this approach much more attractive than other competitors.

As indicated above, $f_{Q_X}(Q_X)$ would be our target distribution f_Q. Therefore, once the conditional distribution of $f_{S_X|Q_X}$ is determined, we would have a corresponding density on X. It is easily seen that there are various densities $f_X(X)$ that have the margin distribution matching our desired distribution.

As a default choice, Jauch et al. (2020a) recommended $f_{S_X|Q_X}$ to be the density of the Wishart distribution $W_k(p, I_k)$ and it is independent of Q_X. With this choice, the density of the distribution of X simplifies to

$$f_X(X) = \left(\frac{1}{\sqrt{2\pi}}\right)^{pk} \operatorname{etr}(-X^T X/2) f_Q(Q_X). \tag{4.33}$$

In particular, if we consider the problem of sampling uniformly from the Stiefel manifold, $f_Q(Q_X) \propto 1$, then the density of X will be

$$f_X(X) = \left(\frac{1}{\sqrt{2\pi}}\right)^{pk} \operatorname{etr}(-X^T X/2). \tag{4.34}$$

This density shows that all the entries of X are independent standard normal random variables. Notice that this is equivalent to the situation of sampling from the unit sphere. This correspondence motivates the author to select the Wishart distribution as the default choice.

Inference for $\lambda_t^{(j)}$

Inferring $\lambda_t^{(j)}$ considers the conditional distribution formed by the product of the likelihood, together with 4.27. The sampling step is again handled by Stan.

Inference for σ_t^2

Without prior knowledge, we will put a non-informative uniform prior on $[0, \infty)$.

$$p(\sigma_t^2 | U_t, \lambda_t^{(j)}, S_t) \propto (\sigma_t^2)^{-\frac{n_t p}{2}} \operatorname{etr}\left(\frac{1}{2\sigma_t^2}(U_t \Omega_t U_t^T - I) S_t\right) \prod_{j=1}^{r} \left(\frac{\sigma_t^2}{\lambda_t^{(j)} + \sigma_t^2}\right)^{\frac{n_t}{2}} \tag{4.35}$$

This is again a complicated univariate distribution, and we resort to Stan programs.

4.3.3 Summary of Markov Chain Monte Carlo Algorithm

Initialization

The initial values of the parameters are assigned by the empirical values provided by the data. In particular, at time point t, U_t takes the first r empirical eigenvectors and $\{\lambda_t^{(1)}, \lambda_t^{(2)}, ..., \lambda_t^{(r)}\}$

take the leading r empirical eigenvalues, whereas σ_t^2 is initialized as the median of the $p-r$ tail eigenvalues. The initial values of the across-group hyper-parameters are assigned by running the corresponding Gibbs sampling steps once with the initialized group-specific parameters.

Sampling Algorithm

```
Algorithm 5: MCMC Algorithm for Dynamic Covariance Estimation
  Result: Samples of $U_t$'s, $\Lambda_t$'s, $\sigma_t^2$'s, $\Sigma_t$'s.
  Initialization: initialize $U_t$'s, $\Lambda_t$'s, $\sigma_t^2$'s using empirical values;
  for $i$ in 1 : (Burn-in + Iterations) do
      Update the across-group parameters:
      1. Sample $A, B$ with 4.24;
      2. Update $\beta_0^{(j)}, \beta_1^{(j)}, \pi^{(j)}, \tau_S^{2^{(j)}}, \tau_L^{2^{(j)}}$ for $j \in \{1, 2, \cdots, r\}$
      Update the group-specific parameters:
      3. Update $U_t$ with 4.28, 4.29 and 4.30;
      4. Update $\{\lambda_t^{(1)}, \lambda_t^{(2)}, \cdots, \lambda_t^{(r)}\}$ ;
      5. Update $\sigma_t^2$ with 4.35;
      Save the samples for every 5 iterations;
  end
```

4.4 Simulation Results

For the simulation results, we aim to show that our method is able to recover the correct parameters when the data are indeed generated from the specified model, which includes the parameters governing the dynamic process on the Stiefel manifold, the eigenvectors, as well as the eigenvalues. In the following simulation study, we consider a three-factor dynamic model in the 100 dimensional space, namely $p = 100,\ r = 3$. We assume that there are $T = 30$ time points for the temporal evolution process and $n_t = 20$ observations at each time point. The true distribution that governs the first-order Markov dynamics on the Stiefel manifold is a generalized Bingham distribution with $A = B = \text{diag}(\{50, 20, 10\})$. The eigenvalues are generated from stationary distributions of three first-order auto-regressive processes, $j = 1, 2, 3$.

$$\lambda_t^{(j)} = c^{(j)} + \varphi^{(j)} \lambda_{t-1}^{(j)} + \epsilon_t^{(j)},\ \ \epsilon_t^{(j)} \sim N(0, (\sigma^{(j)})^2),$$

	c	φ	σ^2
First Eigenvalue	100	0.7	10
Second Eigenvalue	50	0.7	5
Third Eigenvalue	10	0.7	2

Table 4.1: Parameters for the first-order autoregressive processes

The parameters for the auto-regressive processes are displayed in table 4.1. The MCMC algorithm was conducted for 2000 iterations, with the first half as burn-in samples. In addition, the Stan functions have a smaller burn-in period with 50 burn-in samples.

4.4.1 Smoothness between Eigenvectors over Time

Under this setup, the eigenvalues are well-separated, which means the directions of the eigenvectors are well identified. To see how aligned the estimated samples are across time, we consider the metric

$$x_t^{(j)} = |\langle v_t^{(j)}, v_{t+1}^{(j)} \rangle| \tag{4.36}$$

for $j = 1, 2, 3$ across t. $x_t^{(j)}$ will be close to 1 if $v_t^{(j)}$ and $v_{t+1}^{(j)}$ are aligned, and close to 0 if they are almost orthogonal to each other, which is common in the high-dimensional space.

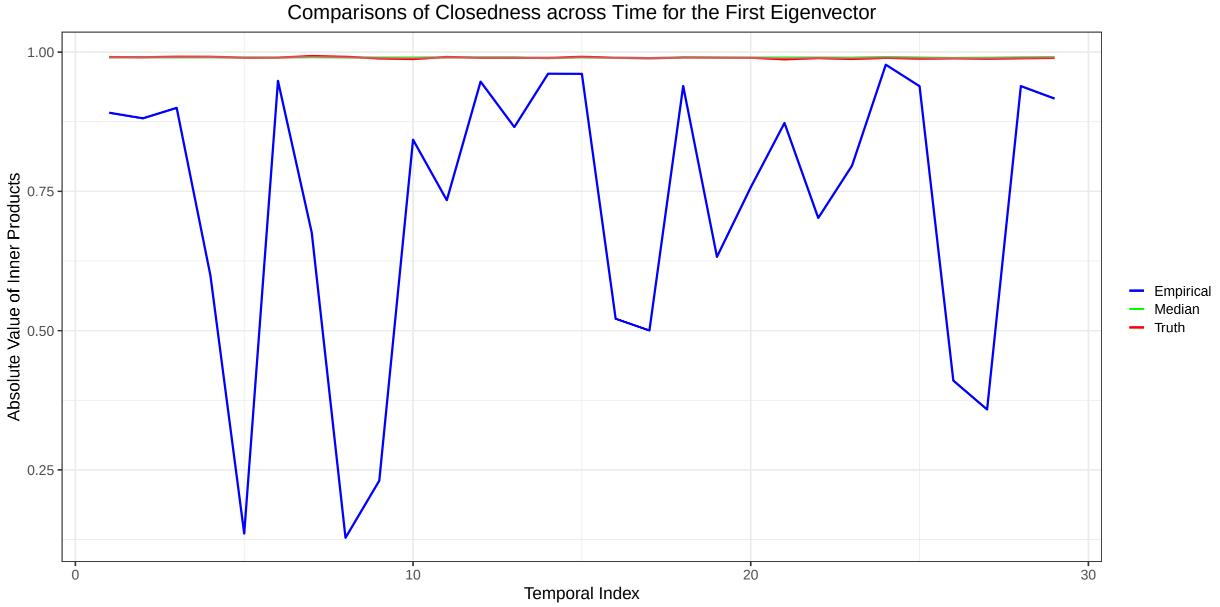
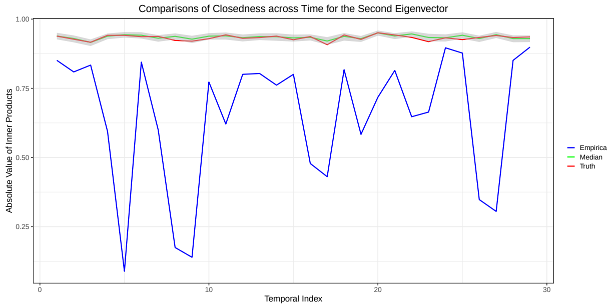


Figure 4.3: First Eigenvector



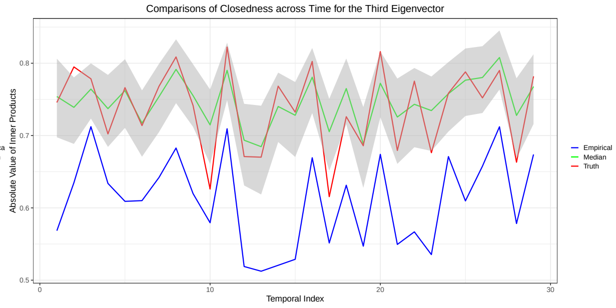


Figure 4.4: Second Eigenvector

Figure 4.5: Third Eigenvector

The results are displayed in Figure 4.3, 4.4 and 4.5. Computing $x_t^{(j)}$ for all the empirical estimates and posterior samples. The empirical values are denoted by blue lines, while the true values are in red. For the Bayesian samples, we compute the median of all the remaining samples and show them in green, together with a grey ribbon characterizing the uncertainty using the 95% posterior interval. The first two eigenvectors are easily detectable since they correspond to larger eigenvalues. The magnitudes of $A_{11}, A_{22}, B_{11}, B_{22}$ also indicate dynamic processes where the sequential eigenvectors are more closely aligned. For the third eigenvector, its direction is slightly more difficult to identify, and the eigenvectors are not as closely aligned. Nevertheless, we can observe that connecting them via a dynamic process can effectively boost the performance, as the median of the estimates are more closer to the truth than the raw principal component estimates.

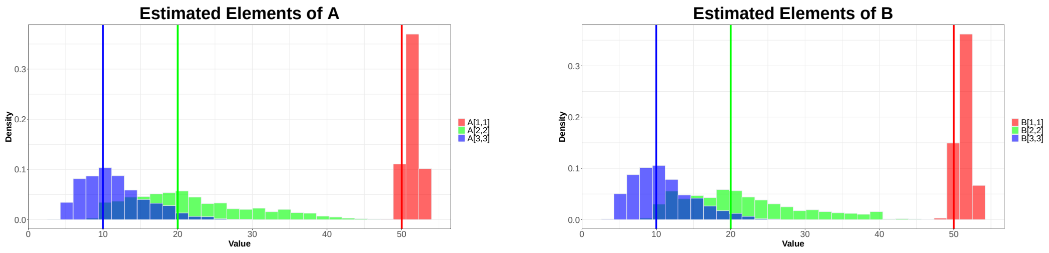
To sum up, in all cases, the median estimates matches the truth much better in contrast to the noisier empirical estimates, and the 95% posterior intervals recover the smoothness between the nearby eigenvectors well. This demonstrates the effectiveness of our shrinkage approach in utilizing the information across time points.

Notice that the method works more effectively when the eigenvalues are spaced out, and less so when the eigenvalues are similar. In that case, the eigenvectors are not clearly identifiable, hence the autoregressive model on the eigenvectors will produce results with high uncertainties.

4.4.2 Estimation of A and B

Next we want to check if the parameters for the generalized Bingham distribution are reasonably recovered. This is not always achieved, especially when the diagonal values of A and B are not spaced apart. However, we generally care more about how the smoothness of the eigenvectors are recovered rather than the parameters themselves.

In figure 4.6, the density plots for the Bayesian samples of the diagonal elements of A and B are shown. The thick vertical lines represent the true values of the respective diagonal elements, which are close to the modes of the samples. Hence, under the situation where eigenvalues are well spaced, our method achieves satisfying results for recovering the smoothness parameters.



(a) Elements of A

(b) Elements of B

Figure 4.6: Estimated Samples for A and B

4.4.3 Estimation of Eigenvalues

The eigenvalues are estimated using shrinkage Bayesian first-order auto-regressive models. Intuitively, our estimates would be more smooth than the noisy empirical estimates, since the variances of the noises possess shrinkage priors. The comparisons are demonstrated as below, with the lines and ribbons carrying the same meanings as before. Again we can clearly observe from Figure 4.7, 4.8 and 4.9 that the shrunk results provide smoother estimates which match better with the true eigenvalues, regardless of the magnitudes. Hence our method successfully utilizes information across time points to obtain more accurate results.

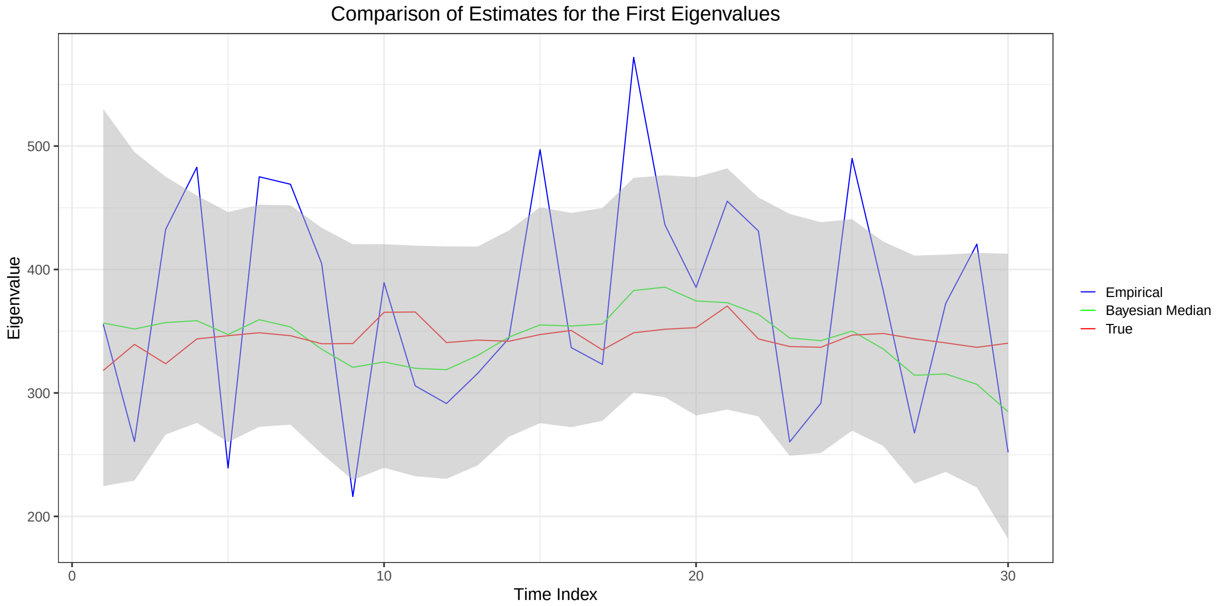
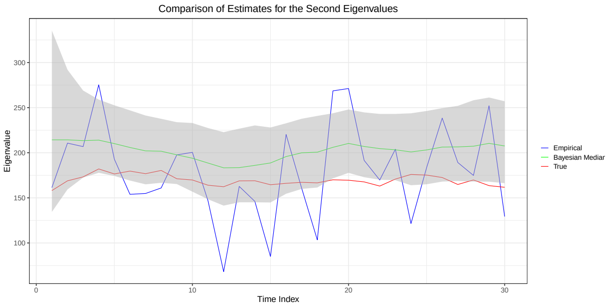


Figure 4.7: First Eigenvalue



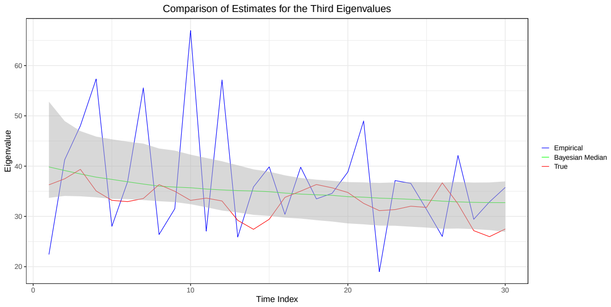


Figure 4.8: Second Eigenvalue

Figure 4.9: Third Eigenvalue

4.4.4 Comparison with Alternative Approaches

Our method automatically infers the smoothness of the eigenvectors. This contrasts with more naive window smoothing techniques by selecting a window that covers t_i, and utilizing all the data in the window to compute the eigenvectors at time t_i. The problem then becomes choosing the optimal window size, which we will characterize using radius. For radius equals to d, the window for t_i would be time points $\{t_i - d, t_i - d + 1, ..., t_i, ..., t_i + d - 1, t_i + d\}$.

In the following, we compare the naive approach with different window sizes, the Bayesian estimates, and the pooled estimate using only observations at t_i. The quantity being examined is the dispersion of the first eigenvector, as defined later in equation 4.5. It is a metric measuring how the eigenvector differs from the direction represented by the vector with all equal elements. The dispersion can be extreme in high dimensions, thus we analyze the logarithm of it instead.

From figure 4.10, it is clear that pooling all the data fails to capture the high dispersions in the starting period. For different radiuses, the estimates stablize as the radius increases since more and more data are utilized for estimation for that time point, with pooling as the extreme case. However, it is uncertain which integer radius, or fractional radius should be adopted in practice when no prior knowledge is provided. The Bayesian approach, on the other hand, provides an automatic implicit determination of optimal radius by sharing the information across all time points. From the plot we can clearly see that our method matches the truth pretty well, showing the possibility of a data-driven approach to automatically detect the smoothness without human intervention.

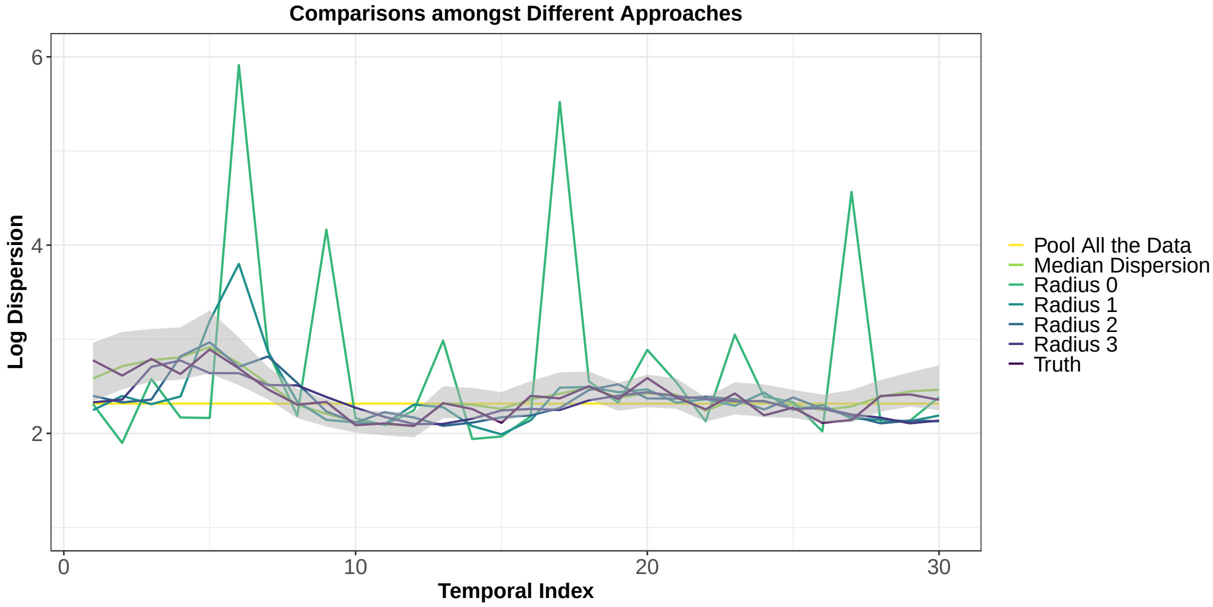


Figure 4.10: Comparisons for Recovery of Dispersion

4.5 Results on S&P500 Returns Data

4.5.1 Estimated Dynamic Betas over Time

In order to obtain the estimated betas from the factor loadings on the first eigenvector, we normalize the loadings to have a unit mean. The following figures show the estimated betas for selected companies. We choose four companies from different industries, Cisco Systems Inc. for technology, Bank of America Corp for finance, Nike Inc. for sports, and Air Products & Chemicals for chemicals. Firstly, it is clear that the Bayesian median estimates are much smoother than the empirical counterparts, and the Bayesian 95% posterior interval, indicated by the gray ribbon, provides quantification for model uncertainties. Moreover, the dynamics of betas admit clear and reasonable interpretations. For Cisco Systems in Figure 4.11, beta was more volatile during the technology bubble between 2000 and 2002, and it stabilizes afterwards until 2018. This happens since it belongs to the technology industry, which was impacted significantly during the tech-bubble. On the other hand, Bank of America achieved high betas in 2009 and 2012. Figure 4.12 shows the peak in 2009, which was definitely impacted by the 2008 global financial crisis, and the peak in 2012 was influenced by the European crisis. In contrast, Nike Inc. belongs to the sports industry, which is less affected by the crises. The same applies to Air Products & Chemicals, which falls in the categories of industrial gases and chemicals. Figure 4.13 and 4.14 confirm the intuition, and the corresponding betas were quite stable around 1 from 1997 to 2021.

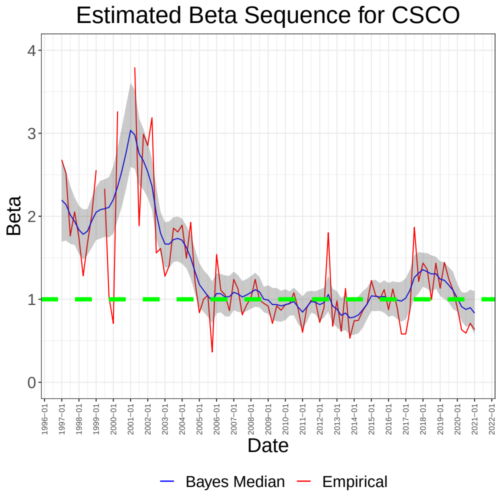


Figure 4.11: Cisco Systems Inc.

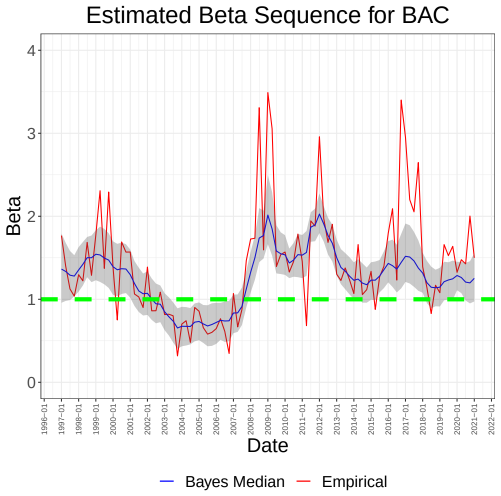


Figure 4.12: Bank of America Corp

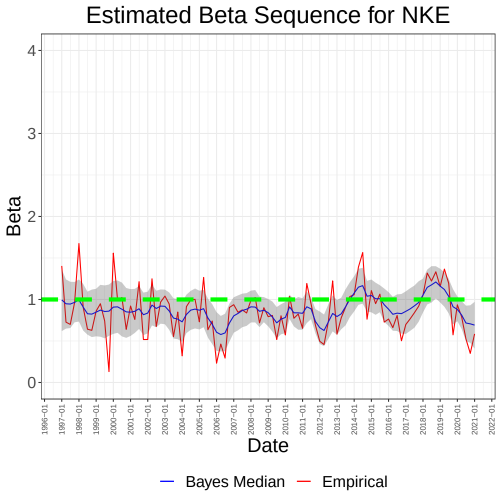


Figure 4.13: Nike Inc.

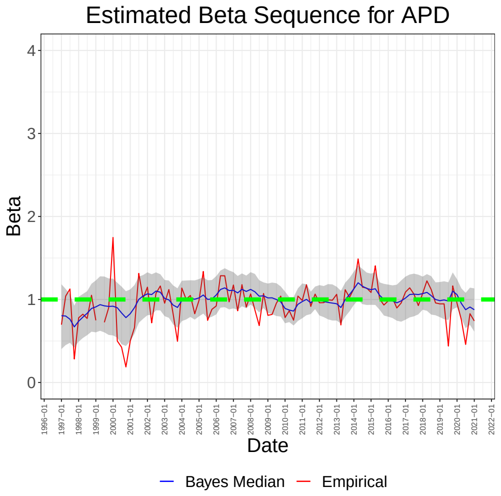


Figure 4.14: Air Products & Chemicals

4.5.2 Volatility Measures

The first dominant latent factor is well-known as the market factor. Since market volatility indicates how fluctuating the market moves as a whole, the variance of the well-diversified portfolio serves as a proxy. Therefore, the estimated eigenvalues serve as a good representation of the market volatility. Notice that for our model, the estimated first eigenvalues are $\lambda_t^{(1)} + \sigma_t^2$. In figure 4.15, we construct a comparison between the Bayesian results and the empirical estimates. The shaded ribbon represents the Bayesian 95% posterior interval, and it covers the empirical estimates well. The Bayesian median, represented by the red line, is comparable to or lower than the empirical volatilities, especially in high volatility periods. The results demonstrate the ability of our two-component Gaussian mixture model for explaining the mechanism of volatility change.

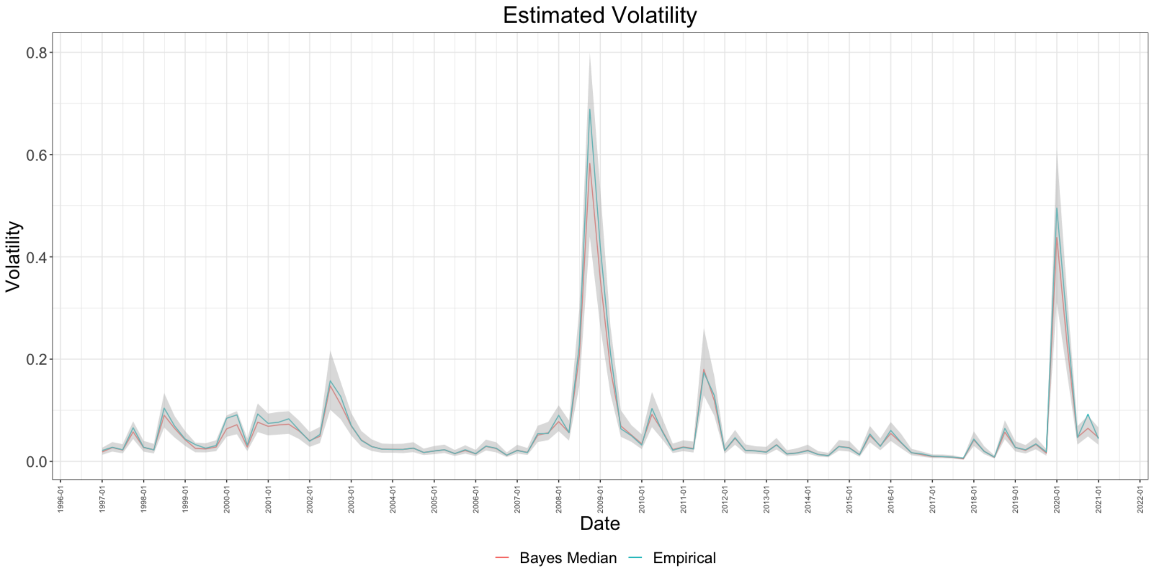


Figure 4.15: Estimated Volatility Time Series

4.5.3 Dispersion

In Figure 4.16, we can see clearly our autoregressive model smooths out the high variabilities in the first eigenvectors and encourages homogeneity. The empirical values contain more noises and are more deviating from the neighbors, whereas our Bayesian results pull the dispersions together. Based on the above simulation results, we believe that our model provides more accurate estimates compared with the empirical values.

We conclude that the dispersion is changing over time, regardless of the sampling variabilities. In particular, it was high during the technology bubble around 2001-2002, and it gradually faded away until 2009, when it slightly increased and then keep decreasing. The dispersion started to climb up last year due to the COVID19 crisis. Beta dispersion will increase when the betas for different companies differ much from each other. A good example was the technology bubble, where the technological companies were heavily affected, while other sectors were less influenced. In 2008, the whole market was shocked, and most companies suffered the crisis. In that case, the companies were affected more evenly and the betas were moving together. Hence we only observe slight increase in the beta dispersion. It is worth noting that the dispersion started to climb up last year, indicating that COVID-19 crisis struck companies differently. It remains an interesting problem to investigate the beta influence of COVID-19 among different companies.

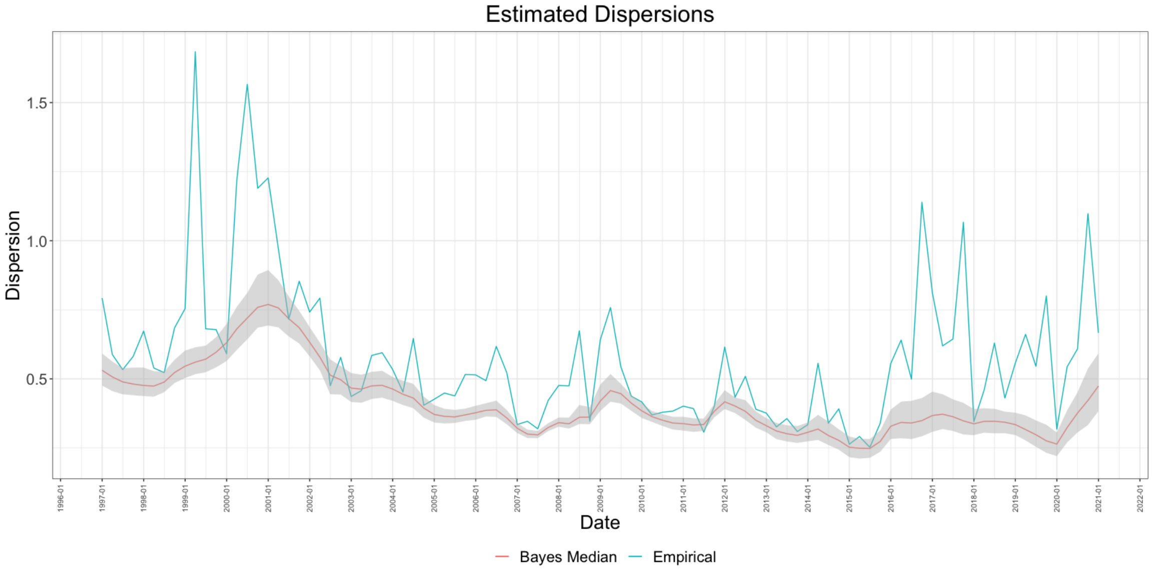


Figure 4.16: Dispersion Time Series

4.5.4 Relationship between Dispersion and Volatility

The volatility measures how volatile the market fluctuates and the dispersion characterizes how different the stocks respond to the market risk. In Figure 4.17, we overlay the estimated time series of dispersion and volatility to discover their relationship. In 2001, the dispersion increased significantly, whereas the volatility stabilized at a low level. However, during 2008-2010, the volatility climbed up to the highest level, when the dispersion went up slightly after the peak in volatility. In comparison, the dispersion dropped while the volatility went up again in 2020 at the same time. From the above three situations, they can move in the same direction, opposite directions, or one is move while the other is stiff. Therefore, the relationship between the dispersion and volatility is complicated, and might depend on some underlying market mechanism.

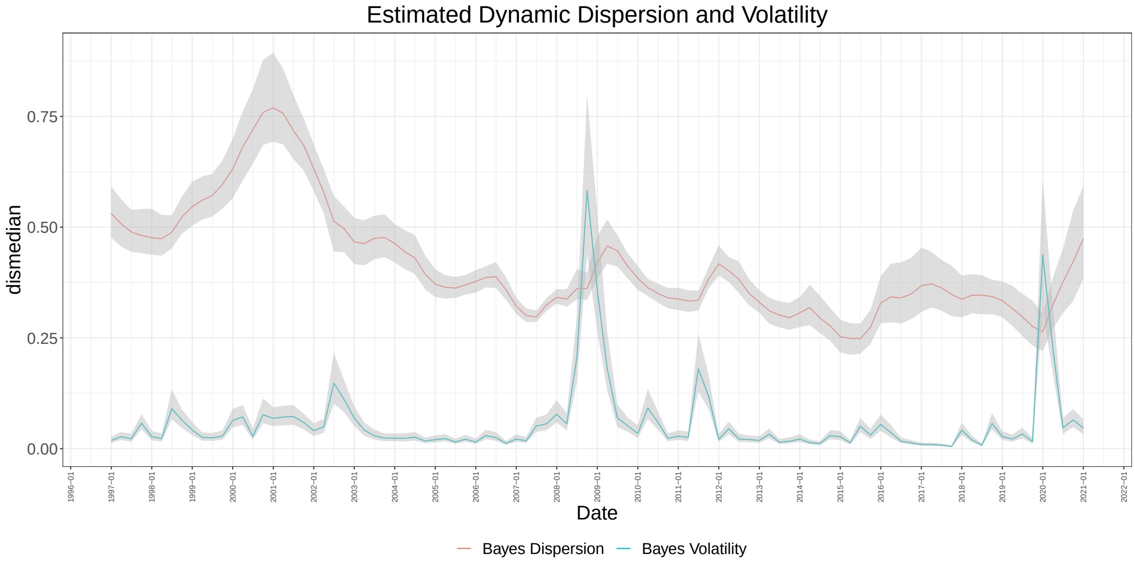


Figure 4.17: Bayes Time Series Plot for Volatility and Dispersion

Now we consider the problem from another perspective. Figure 4.18 was constructed by considering the 90% posterior ellipsis for the logarithmic values of volatility and dispersion at each year. The evolution of the circle with respect to time is of great interest. From 1997-2001, the circle moves towards the direction in which both volatility and dispersion increase. After that, the market was in a regime with low dispersion and volatility from 2004 to 2007, until things changed tremendously in 2008. In 2008 and 2009, the circles became flat ovals, indicating large uncertainties in volatilities. Interestingly, from 2010 to 2019, the market went back to a similar low-volatility low-dispersion regime. The biggest uncertainty was shown in the year 2020, when volatility and dispersion moved in the opposite directions. It is worth mentioning that the year 2021 returns to a similar level as 1997. In total, there is no obvious relationship found between the volatility and dispersion. However, the transition in Figure 4.18 does shed light on some patterns on the market conditions over time.

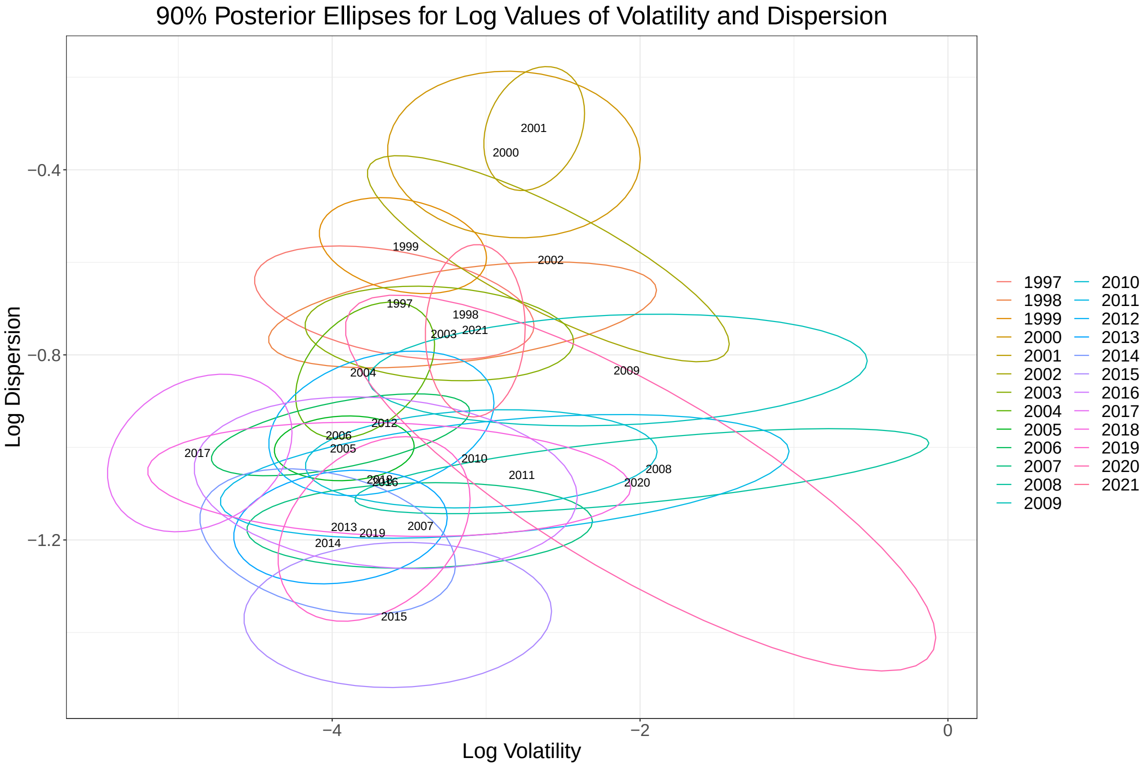


Figure 4.18: Posterior Ellipses for Volatility and Dispersion

4.6 Discussions

In this paper, we were motivated by the relationship between dispersion and volatility in the financial market, and we propose a Bayesian autoregressive model for dynamic covariance estimation that separately models the eigenvalues and eigenvectors. The model considers dynamics of the eigenvectors on the Stiefel manifold, which generalizes the notion of vector rotation to the dynamics of axes, and opens the door to various time series models in the lens of orthogonal matrices. One possible future work is to conduct the one-step forward prediction on the factor loadings. The model can also be applied on incomplete data and provide interpolative results. For illiquid assets such as bonds, whose data is not available for all time points, it is beneficial to get reliable interpolative results for the time points without data. Meanwhile, the model can be easily extended by replacing the distribution on the Stiefel manifold to Matrix Langevin distributions or Watson distributions. We can also switch to other methods for modeling stochastic volatility, such as Gaussian processes or deep neural networks.

Finally, we can also consider exploring the second and third factors and their loadings, even though there might be more uncertainties. It is encouraged to try data with different granularities to further explore the relationships between important financial concepts under different timeframes. Furthermore, the model can be applied to other datasets such as climate data and biological data, and is expected to discover deeper insights in those fields.

Bibliography

Aguilar, O. and West, M. (2000). Bayesian dynamic factor models and portfolio allocation. *Journal of Business & Economic Statistics*, 18(3):338–357.

Arjovsky, M., Shah, A., and Bengio, Y. (2016). Unitary evolution recurrent neural networks. In *International Conference on Machine Learning*, pages 1120–1128. PMLR.

Baik, J. and Silverstein, J. W. (2006). Eigenvalues of large sample covariance matrices of spiked population models. *Journal of multivariate analysis*, 97(6):1382–1408.

Bali, T. G., Engle, R. F., and Tang, Y. (2017). Dynamic conditional beta is alive and well in the cross section of daily stock returns. *Management Science*, 63(11):3760–3779.

Bansal, N., Chen, X., and Wang, Z. (2018). Can we gain more from orthogonality regularizations in training deep cnns? *arXiv preprint arXiv:1810.09102*.

Ben-Israel, A. (1999). The change-of-variables formula using matrix volume. *SIAM Journal on Matrix Analysis and Applications*, 21(1):300–312.

Bingham, C. (1974). An antipodally symmetric distribution on the sphere. *The Annals of Statistics*, pages 1201–1225.

Blume, M. E. (1971). On the assessment of risk. *Journal of Finance*, 26:1–10.

Blume, M. E. (1975). Betas and their regression tendencies. *Journal of Finance*, 30:785–95.

Boik, R. J. (2002). Spectral models for covariance matrices. *Biometrika*, 89(1):159–182.

Bollerslev, T. (1990). Modelling the coherence in short-run nominal exchange rates: a multivariate generalized arch model. *The review of economics and statistics*, pages 498–505.

Butler, R. W., Wood, A. T., et al. (2002). Laplace approximations for hypergeometric functions with matrix argument. *The Annals of Statistics*, 30(4):1155–1177.

Campbell, J. Y. and Lettau, M. (1999). Dispersion and volatility in stock returns: An empirical investigation.

Cayley, A. (1846). About the algebraic structure of the orthogonal group and the other classical groups in a field of characteristic zero or a prime characteristic. *Reine Angewandte Mathematik*, 32(1846):6.

Chen, Y. and Tanaka, K. (2020). Maximum likelihood estimation of the fisher-bingham distribution via efficient calculation of its normalizing constant. *arXiv preprint arXiv:2004.14660*.

Chikuse, Y. (2006). State space models on special manifolds. *Journal of Multivariate Analysis*, 97(6):1284–1294.

Chikuse, Y. (2012). *Statistics on special manifolds*, volume 174. Springer Science & Business Media.

Chow, S.-M., Zu, J., Shifren, K., and Zhang, G. (2011). Dynamic factor analysis models with time-varying parameters. *Multivariate Behavioral Research*, 46(2):303–339.

Chrétien, S. and Guedj, B. (2020). Revisiting clustering as matrix factorisation on the stiefel manifold. In *International Conference on Machine Learning, Optimization, and Data Science*, pages 1–12. Springer.

Cogswell, M., Ahmed, F., Girshick, R., Zitnick, L., and Batra, D. (2015). Reducing overfitting in deep networks by decorrelating representations. *arXiv preprint arXiv:1511.06068*.

Constantine, A. and Muirhead, R. J. (1976). Asymptotic expansions for distributions of latent roots in multivariate analysis. *Journal of Multivariate Analysis*, 6(3):369–391.

Demirer, R., Gupta, R., Lv, Z., and Wong, W.-K. (2019). Equity return dispersion and stock market volatility: Evidence from multivariate linear and nonlinear causality tests. *Sustainability*, 11(2):351.

Engle, R. (2002). Dynamic conditional correlation: A simple class of multivariate generalized autoregressive conditional heteroskedasticity models. *Journal of Business & Economic Statistics*, 20(3):339–350.

Engle, R. F. and Kroner, K. F. (1995). Multivariate simultaneous generalized arch. *Econometric theory*, pages 122–150.

Engle, R. F., Ledoit, O., and Wolf, M. (2019). Large dynamic covariance matrices. *Journal of Business & Economic Statistics*, 37(2):363–375.

Fama, E. F. and French, K. R. (1992). The cross-section of expected stock returns. *the Journal of Finance*, 47(2):427–465.

Fama, E. F. and French, K. R. (2015). A five-factor asset pricing model. *Journal of financial economics*, 116(1):1–22.

Flury, B. K. (1987). Two generalizations of the common principal component model. *Biometrika*, 74(1):59–69.

Forni, M., Hallin, M., Lippi, M., and Reichlin, L. (2000). The generalized dynamic-factor model: Identification and estimation. *Review of Economics and statistics*, 82(4):540–554.

Franks, A. (2020). Reducing subspace models for large-scale covariance regression. *arXiv preprint arXiv:2010.00503*.

Franks, A. M. and Hoff, P. (2019). Shared subspace models for multi-group covariance estimation. *Journal of Machine Learning Research*, 20(171):1–37.

Gatto, R. (2013). The von mises–fisher distribution of the first exit point from the hypersphere of the drifted brownian motion and the density of the first exit time. *Statistics & Probability Letters*, 83(7):1669–1676.

Gelman, A., Lee, D., and Guo, J. (2015). Stan: A probabilistic programming language for bayesian inference and optimization. *Journal of Educational and Behavioral Statistics*, 40(5):530–543.

Geweke, J. (1977). The dynamic factor analysis of economic time series. *Latent variables in socio-economic models*.

Goldberg, L. R., Papanicolaou, A., and Shkolnik, A. (2018). The dispersion bias. *Available at SSRN 3071328*.

Goldberg, L. R., Papanicolaou, A., Shkolnik, A., and Ulucam, S. (2020). Better betas. *The Journal of Portfolio Management*, 47(1):119–136.

Gupta, A. K. and Nagar, D. K. (2018). *Matrix variate distributions*, volume 104. CRC Press.

Harris, R. D., Stoja, E., and Tan, L. (2017). The dynamic black–litterman approach to asset allocation. *European Journal of Operational Research*, 259(3):1085–1096.

Heimberg, G., Bhatnagar, R., El-Samad, H., and Thomson, M. (2016). Low dimensionality in gene expression data enables the accurate extraction of transcriptional programs from shallow sequencing. *Cell systems*, 2(4):239–250.

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

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(2):438–456.

Hoff, P. D. and Niu, X. (2012). A covariance regression model. *Statistica Sinica*, pages 729–753.

Horváth, L., Li, B., Li, H., and Liu, Z. (2020). Time-varying beta in functional factor models: Evidence from china. *The North American Journal of Economics and Finance*, 54:101283.

Hwangbo, N., Zhang, X., Raftery, D., Gu, H., Hu, S.-C., Montine, T. J., Quinn, J. F., Chung, K. A., Hiller, A. L., Wang, D., et al. (2021). An aging clock using metabolomic csf. *bioRxiv*.

Ismaila, A. G. and Shakranib, M. S. (2003). The conditional capm and cross-sectional evidence of return and beta for islamic unit trusts in malaysia. *International Journal of Economics, Management and Accounting*, 11(1).

Jagannathan, R. and Wang, Z. (1996). The conditional capm and the cross-section of expected returns. *The Journal of finance*, 51(1):3–53.

Jauch, M., Hoff, P. D., and Dunson, D. B. (2020a). Monte carlo simulation on the stiefel manifold via polar expansion. *Journal of Computational and Graphical Statistics*, pages 1–23.

Jauch, M., Hoff, P. D., Dunson, D. B., et al. (2020b). Random orthogonal matrices and the cayley transform. *Bernoulli*, 26(2):1560–1586.

Jensen, M. C., Black, F., and Scholes, M. S. (1972). The capital asset pricing model: Some empirical tests.

Johnstone, I. M. (2001). On the distribution of the largest eigenvalue in principal components analysis. *Annals of statistics*, pages 295–327.

Kume, A., Preston, S. P., and Wood, A. T. (2013). Saddlepoint approximations for the normalizing constant of fisher–bingham distributions on products of spheres and stiefel manifolds. *Biometrika*, 100(4):971–984.

Kume, A. and Sei, T. (2018). On the exact maximum likelihood inference of fisher–bingham distributions using an adjusted holonomic gradient method. *Statistics and Computing*, 28(4):835–847.

Lahtinen, K. D., Lawrey, C. M., and Hunsader, K. J. (2018). Beta dispersion and portfolio returns. *Journal of Asset Management*, 19(3):156–161.

Ledoit, O. and Wolf, M. (2004). Honey, i shrunk the sample covariance matrix. *The Journal of Portfolio Management*, 30(4):110–119.

Lin, L., Rao, V., and Dunson, D. (2017). Bayesian nonparametric inference on the stiefel manifold. *Statistica Sinica*, pages 535–553.

Lui, Y. M. (2012). Advances in matrix manifolds for computer vision. *Image and Vision Computing*, 30(6-7):380–388.

Marsaglia, G. et al. (1972). Choosing a point from the surface of a sphere. *The Annals of Mathematical Statistics*, 43(2):645–646.

Mezzadri, F. (2006). How to generate random matrices from the classical compact groups. *arXiv preprint math-ph/0609050*.

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

Muller, M. E. (1959). A note on a method for generating points uniformly on n-dimensional spheres. *Communications of the ACM*, 2(4):19–20.

Neal, R. M. et al. (2011). Mcmc using hamiltonian dynamics. *Handbook of markov chain monte carlo*, 2(11):2.

Nirwan, R. S. and Bertschinger, N. (2019). Rotation invariant householder parameterization for bayesian pca. *arXiv preprint arXiv:1905.04720*.

Niu, X. and Hoff, P. D. (2019). Joint mean and covariance modeling of multiple health outcome measures. *The annals of applied statistics*, 13(1):321.

Pal, S., Sengupta, S., Mitra, R., Banerjee, A., et al. (2020). Conjugate priors and posterior inference for the matrix langevin distribution on the stiefel manifold. *Bayesian Analysis*, 15(3):871–908.

Paul, D. (2007). Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statistica Sinica*, pages 1617–1642.

Pourahmadi, M. (1999). Joint mean-covariance models with applications to longitudinal data: Unconstrained parameterisation. *Biometrika*, 86(3):677–690.

Pourzanjani, A. A., Jiang, R. M., Mitchell, B., Atzberger, P. J., and Petzold, L. R. (2021). Bayesian inference over the stiefel manifold via the givens representation. *Bayesian Analysis*, 1(1):1–28.

Prentice, M. J. (1982). Antipodally symmetric distributions for orientation statistics. *Journal of Statistical Planning and Inference*, 6(3):205–214.

Schwert, G. W. (1989). Why does stock market volatility change over time? *The journal of finance*, 44(5):1115–1153.

Shepard, R., Brozell, S. R., and Gidofalvi, G. (2015). The representation and parametrization of orthogonal matrices. *The Journal of Physical Chemistry A*, 119(28):7924–7939.

Šmídl, V. and Quinn, A. (2007). On bayesian principal component analysis. *Computational statistics & data analysis*, 51(9):4101–4123.

Stivers, C. T. (2003). Firm-level return dispersion and the future volatility of aggregate stock market returns. *Journal of Financial Markets*, 6(3):389–411.

Tan, M., Hu, Z., Yan, Y., Cao, J., Gong, D., and Wu, Q. (2019). Learning sparse pca with stabilized admm method on stiefel manifold. *IEEE Transactions on Knowledge and Data Engineering*.

Turaga, P., Veeraraghavan, A., and Chellappa, R. (2008). Statistical analysis on stiefel and grassmann manifolds with applications in computer vision. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE.

Udell, M. and Townsend, A. (2019). Why are big data matrices approximately low rank? *SIAM Journal on Mathematics of Data Science*, 1(1):144–160.

Vasicek, O. A. (1973). A note on using cross-sectional information in bayesian estimation of security betas. *The Journal of Finance*, 28(5):1233–1239.

Watson, G. S. (1983). Statistics on spheres.

Wu, Y., Hernández-Lobato, J. M., and Zoubin, G. (2013). Dynamic covariance models for multivariate financial time series. In *International Conference on Machine Learning*, pages 558–566.

Yang, Y. and Bauwens, L. (2018). State-space models on the stiefel manifold with a new approach to nonlinear filtering. *Econometrics*, 6(4):48.