8-2022

Bayesian methodologies for constrained spaces.

Siddhesh Kulkarni
University of Louisville

Recommended Citation

Kulkarni, Siddhesh, "Bayesian methodologies for constrained spaces." (2022). *Electronic Theses and Dissertations.* Paper 3956.
https://doi.org/10.18297/etd/3956

BAYESIAN METHODOLOGIES FOR CONSTRAINED SPACES

By

Siddhesh Kulkarni
B.Sc., Savitribai Phule Pune University (Former University of Pune) 2012
M.Sc., Savitribai Phule Pune University (Former University of Pune) 2014
M.S., University of Connecticut 2018

A Dissertation
Submitted to the Faculty of the
School of Public Health and Information Sciences
of the University of Louisville
in Partial Fulfillment of the Requirements
for the Degree of

Doctor of Philosophy
in Biostatistics

Department of Bioinformatics and Biostatistics
University of Louisville
Louisville, Kentucky

August 2022

BAYESIAN METHODOLOGIES FOR CONSTRAINED SPACES

By

Siddhesh Kulkarni
B.S., Savitribai Phule Pune University (Former University of Pune) 2012
M.S., Savitribai Phule Pune University (Former University of Pune) 2014
M.S., University of Connecticut 2018

A Dissertation Approved on

20th July 2022

by the following Dissertation Committee:

Jeremy T. Gaskins, Ph.D., Dissertation Director

Maying Kong, Ph.D.

Ritendranath Mitra, Ph.D.

Michael Sekula, Ph.D.

Brendan Depue, Ph.D.

DEDICATION

I dedicate this dissertation to my parents
Mrs. Mohini Kulkarni and Mr. Shripad Kulkarni
without whose enormous support this journey was not even possible.

ACKNOWLEDGMENTS

I would like to express my deepest gratitude to my conscientious advisor Dr. Jeremy Gaskins for his valuable guidance during my research journey and unconditional support for my career growth. I am very fortunate to be his student. I learned strong fundamentals of research from him which I expect to be a strong foundation for the rest of my professional journey.

Moreover, I would like to thank the rest of my committee members Drs. Maying Kong, Ritendranath Mitra, Brendan Depue, Michael Sekula, and mentor Dr. Subhadip Pal under whose guidance I had a privilege to work and whose suggestions and comments enhanced my research work. I would also like to thank Drs. Shesh Nath Rai and Andrew DeFillippis under whose guidance I started my research assistant journey at the University of Louisville and learned the essentials of collaborative research.

My sincere appreciation also goes to the Department of Bioinformatics and Biostatistics especially Prof. Karunarathna Kulasekera, at the University of Louisville for providing me an opportunity to become a student here and become a member of the research community. I would like to also extend my sincere gratitude to School of Public Health and Information Sciences staff, especially Dr. Tammi Thomas for being a strong backup during this journey.

I will be in all ways indebted to the Delphi center For Teaching and Learning at the University of Louisville, especially IL Barrow, Dr. Marie Brown, Dr. Gale Rhodes, Kristen Brown and Dr. Patty Payette for being my strongest support system. They provided me excellent opportunities for professional growth through graduate assistantship work, celebrated my academic achievements, nurtured my soft skills, and become my family. My memories of the Delphi center will be cherished for lifetime.

My acknowledgments also go to my classmates and friends. I made some of my

meaningful friendships here. Life in Louisville would not have been the same without them.

This work was conducted in part using the resources of the University of Louisville's Research Computing group and the Cardinal Research Cluster.

I am thankful to my mentors at University of Connecticut, Drs. Nalini Ravishankar, Nitish Mukhopadhyay, Dipak Dey and Dimitris Xygalatas for supporting me when setting my feet in this country. A special thank you to Dr. Sanjay Shete of The University of Texas, M.D. Anderson Cancer Center for his encouragement and support during my academic endeavors.

I am forever grateful to Dr. Uttara Naik-Nimbalkar at University of Pune and Dr. Vivek Borkar at Indian Institute of Technology, Bombay because of whom I could reach to pinnacle of my academic life. A warmest thank you to Mrs. Neela Sahasrabuddhe at Sir Parashurambhau College, Pune for being a great supporter during my undergraduate years. I would also like to thank the Jamsetjee Nusserwanjee Tata Endowment Trust for honoring me with one of the most reputed scholarships in India.

Last but not least, my parents who were there as a strong support during my achievements and setbacks, who encouraged and provided me with an environment to reach my goals. Because of their unconditional sacrifices I could reach to this summit.

ABSTRACT

BAYESIAN METHODOLOGIES FOR CONSTRAINED SPACES

Siddhesh Kulkarni

August 10, 2022

Due to advances in technology, there is a presence of directional data in a wide variety of fields. Often distributions to model directional data are defined on manifolds or constrained spaces. Regular statistical methods applied to data defined on special geometries can give misleading results, and this demands new statistical theory. This dissertation addresses two such problems and develops Bayesian methodologies to improve inference in these arenas. It consists of two projects: 1. A Bayesian Methodology for Estimation for Sparse Canonical Correlation, and 2. Bayesian Analysis of Finite Mixture Model for Spherical Data.

In principle, it can be challenging to integrate data measured on the same individuals occurring from different experiments and model it together to gain a larger understanding of the problem. Canonical Correlation Analysis (CCA) provides a useful tool for establishing relationships between such data sets. When dealing with high dimensional data sets, Structured Sparse CCA (ScSCCA) is a rapidly developing methodological area which seeks to represent the interrelations using sparse direction vectors for CCA. There is less development in Bayesian methodology in this area. We propose a novel Bayesian ScSCCA method with the use of a Bayesian infinite factor model. Using a multiplicative half Cauchy prior process, we bring in sparsity at the level of the projection matrix. Additionally, we promote further sparsity in the covariance matrix by using graphical horseshoe prior or diagonal structure. We compare the results for our proposed model with competing frequentist and Bayesian methods and apply the developed method to omics data arising from a breast cancer study.

In the second project, we perform Bayesian Analysis for the von Mises Fisher (vMF) distribution on the sphere which is a common and important distribution used for directional data. In the first part of this project, we propose a new conjugate prior for the mean vector and concentration parameter of the vMF distribution. Further we prove its properties like finiteness, unimodality, and provide interpretations of its hyperparameters. In the second part, we utilize a popular prior structure for a mixture of vMF distributions. In this case, the posterior of the concentration parameter consists of an intractable Bessel function of the first kind. We propose a novel Data Augmentation Strategy (DAS) using a Negative Binomial Distribution that removes this intractable Bessel function. Furthermore, we apply the developed methodology to Diffusion Tensor Imaging (DTI) data for clustering to explore voxel connectivity in human brain.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Constrained Spaces

Advancement in technology have given rise to directional datasets where observations are recorded as directions or angles relative to a system with fixed orientation [Wang and Gelfand, 2013]. Some of such a data lie on circumference of unit circle ($\mathbb{R}^2$) or on unit hypersphere $\mathbb{S}^{p-1} = \{\boldsymbol{y} \in \mathbb{R}^p : \|\boldsymbol{y}\|_2 = 1\}$, where $\|\boldsymbol{y}\|_2 = \sqrt{\boldsymbol{y}^T\boldsymbol{y}}$. For example, data on wind direction are observed as a direction in the plane $\mathbb{R}^2$. It can be represented by an angle θ in the domain of $[0, 2\pi)$ or $[-\pi, \pi)$ measured from a specified origin. Equivalently, this can be also represented as direction unit vector $\mathbf{y} = (\cos\theta, \sin\theta)^T$ with $\|\boldsymbol{y}\|_2 = 1$. This is called as a circular data [Pewsey and García-Portugués, 2021]. Another common instance of directional data is maximum diffusivity directions of water molecules in Diffusion Tensor Imaging. The direction vector describing the flow of water molecules is inherently spherical information. The flow of water molecules is different in the different parts of human brain due to the different properties of the brain tissues in each region which is very much helpful to understand the brain connectivity. Directional data has wide presence in the field of bioinformatics [Mardia et al., 2018], astronomy [Marinucci and Peccati, 2011], medicine [Pardo et al., 2016], genetics [Dortet-Bernadet and Wicker, 2008], image analysis [Esteves et al., 2018], text mining [Banerjee et al., 2005], machine learning [Sra, 2018] and many others [Pewsey and García-Portugués, 2021]. Mardia and Jupp [2000] and Ley and Verdebout [2017] provide a rich literature review of presence of directional statistics in these areas.

The typical support of the distributions used to model directional data are manifolds. Traditional statistical methods such as maximum likelihood estimation and regression analysis rely heavily on vector operations defined on Euclidean space $\mathbb{R}^p$. Inherent to these methods is a notion of geometry that considers a distance between the parameter estimate and true value. However, when working in a (non-Euclidean) manifold, we need to take into account the geometry of the surface to calculate the distances between two points in a way that corresponds to the appropriate geometry. If we measure a mean of sample points on the surface of a manifold using traditional statistical tools, then it may give us a misleading result as the mean might not even lie on the surface of manifold. Further, in least square estimation problems, which is again based on concept of minimizing the error ("distance") in the estimation of parameter of interest in traditional statistics, the usual closed-form solutions are not generally available for manifolds. This intractability raises many challenges in model fitting and inference. Hence, new statistical theory needs to be developed for the directional data [Ley and Verdebout, 2017].

This need to develop new theory has motivated us to address two problems in directional statistics. In our first project we develop Bayesian methodology for sparse Canonical Correlation Analysis (CCA). CCA involves maximizing the correlation among linear summaries of features of datasets measured on same set of subjects. As the linear summaries are constrained to have unit one, their parameter support is a constrained manifold, a hyper-sphere. In our second project we provide and investigate some properties of Bayesian analysis for the von Mises Fisher Distribution, which is one of the prominent distribution in directional statistics. In particular, we provide a new conjugate prior distribution, investigate the finite mixture model using this distribution, and investigate different sampling strategies.

1.2 Probability Modeling of Spherical Data

In this section we will discuss some of the prominent probability distributions used for modeling directional data.

Suppose we have a random p dimensional unit vector $\boldsymbol{y}$ on the hypersphere $\mathbb{S}^{p-1} = \{\boldsymbol{y} \in \mathbb{R}^p : \|\boldsymbol{y}\|_2 = 1\}$. The Fisher-Bingham exponential family of distributions to model this spherical data is given as

$$f_{\boldsymbol{y}}(\boldsymbol{y}; \boldsymbol{\mu}, \kappa, \boldsymbol{A}) \propto \exp\left\{\kappa \boldsymbol{y}^T \boldsymbol{\mu} + \boldsymbol{y}^T \boldsymbol{A} \boldsymbol{y}\right\} \mathbb{I}(\boldsymbol{y} \in \mathbb{S}^{p-1}). \tag{1.1}$$

Here $\boldsymbol{A}$ is symmetric $p \times p$ matrix, and $\boldsymbol{\mu} \in \mathbb{S}^{p-1}$ is mean direction, while $\kappa \geq 0$ is concentration parameter. As $y \in \mathbb{S}^{p-1}$, an assumption of $tr(\boldsymbol{A}) = 0$ holds [Mardia, 1975, Mardia and Jupp, 2000]. A key property of this distribution is that it depends on the random variable $\boldsymbol{y}$ through a quadratic form. This distribution is constructed as a constrained multivariate normal distribution where the constraint is that $\boldsymbol{y} \in \mathbb{S}^{p-1}$.

Several important spherical distributions are derived from this family. When $\kappa = 0$ in (1.1), we obtain the Bingham distribution [Bingham, 1964]. If the constrain $\boldsymbol{A}\boldsymbol{\mu} = \mathbf{0}$ is considered, then we obtain a Kent distribution [Kent, 1982]. This constraint introduces elliptical contours [Pewsey and García-Portugués, 2021]. Overall as new types of directional data are emerging, development of new probability distributions has become an active area of research. Different distributions over this domain are listed in detail by Mardia and Jupp [2009] and Ley and Verdebout [2017]. When $\boldsymbol{A} = \mathbf{0}$ in (1.1) we get one of the simplest and popular distribution for directions data is von Mises Fisher (vMF) distribution. This is a rotationally symmetric distribution. In this distribution the probability density function is dependent on $\boldsymbol{y}$ through $\boldsymbol{y}^T\boldsymbol{\mu}$ which results in the circular contours when $\boldsymbol{y} \in \mathbb{S}^2$ [Pewsey and García-Portugués, 2021]. The PDF of the distribution is

$$f(\boldsymbol{y} \mid \boldsymbol{\mu}, \kappa) = \frac{\kappa^{p/2-1}}{(2\pi)^{p/2} I_{p/2-1}(\kappa)} \exp\left(\kappa \boldsymbol{\mu}^T \mathbf{y}\right) \mathbb{I}(\boldsymbol{y} \in \mathbb{S}^{p-1}), \tag{1.2}$$

where $\kappa \geq 0, \boldsymbol{\mu} \in \mathbb{S}^{p-1}$, $p \geq 2$ and $I_{p/2-1}(\cdot)$ denotes modified Bessel function of the first kind with order $\frac{p}{2} - 1$. The concentration parameter κ quantifies the amount of spread in the distribution around mean $\boldsymbol{\mu}$. For $\kappa = 0$ the distribution is uniform over the sphere. The distribution is unimodal and rotationally symmetric around the direction $\boldsymbol{\mu}$. $\boldsymbol{\mu}^T\mathbf{y}$ is the cosine similarity between $\mathbf{y}$ and $\boldsymbol{\mu}$. Cosine similarity measures angle between two vectors. Its range lies in $[-1, 1]$. If both angle are pointing in the same direction then value will be near unity. This measure of similarity has found immense importance in the field such as text mining, genomics, etc. In our study we will consider data is multivariate over the sphere ($p = 3$). Many types of statistical methods for spatial data rely on the vMF distribution, and in our project, we focus particularly on clustering methods.

Clustering is one of the most popular tools for unsupervised machine learning. In terms of probabilistic modeling, clustering is equivalent to fitting a mixture model to the data. Consider spatial data $\boldsymbol{y}_i$ ($i = 1, 2, \ldots, n$) that we seek to cluster into N distinct components. We let $f_j(\boldsymbol{y} \mid \boldsymbol{\omega}_j)$ denote a probability distribution with parameter $\boldsymbol{\omega}_j$ for cluster j ($j = 1, 2, \cdots, N$). The mixture density is given by

$$f(\boldsymbol{y} \mid \boldsymbol{\pi}, \boldsymbol{\omega}) = \sum_{j=1}^{N} \pi_j f_j(\boldsymbol{y} \mid \boldsymbol{\omega}_j),$$

and the full likelihood is

$$L(\boldsymbol{\pi}, \boldsymbol{\omega} \mid \boldsymbol{Y}) = \prod_{i=1}^{n} \sum_{j=1}^{N} \pi_j f_j(\boldsymbol{y}_i \mid \boldsymbol{\omega}_j). \tag{1.3}$$

This combination of sum and product is intractable. A standard technique to assist in estimation from this mixture model is to introduce auxiliary categorical variables $\mathbf{Z} = (Z_1, \ldots Z_n)^T$ such that

$$P(Z_i = j) = \pi_j \quad (j = 1, 2, \ldots, N;\ i = 1, 2, \ldots, n).$$

The Z_i's are often referred to as the membership variables or cluster indicators. For example, the event $\{Z_i = j\}$ implies that the i^{th} data point is assigned to j^{th} cluster. We can treat this model as a missing data problem, where the cluster membership Z_i is the missing latent data [Celeux et al., 2006].

In fields such as text mining and genomics analysis, involves high dimensional directional data on the unit hypersphere [Banerjee et al., 2005]. As such data is confined to a non-Euclidean space, the most common clustering algorithms which rely on a mixture of multidimensional Gaussian distribution proves inappropriate for modeling. Again, this highlights the need for clustering models based on directional distributions. Dhillon and Modha [2001] is one of the first works which performed clustering over a hypersphere [Sra, 2018]. More examples of such models are mixtures of Kent distributions [Peel et al., 2001], Spherical Topical Model [Reisinger et al., 2010], Dirichlet process vMFMM [Bangert et al., 2010], temporal vMF mixture model [Gopal and Yang, 2014], among others. Clustering models are found in many different areas, a prominent area of application is Neuroimaging [Lashkari et al., 2010, Cabeen and Laidlaw, 2013, Ryali et al., 2013].

Apart from models on spherical data, several mixture models based on different distributions are being proposed to accommodate variety of directional data. Some example of them include mixture of wrapped normal distribution [Agiomyrgiannakis and Stylianou, 2009], Bayesian projected normal mixture models [Wang and Gelfand, 2014, Rodríguez et al., 2020] and general projected mixture normals [Hernandez-Stumpfhauser et al., 2017]. We will not elaborate more on this as it is beyond the scope of our work.

For fitting of vMF mixture models, various approaches based on the E-M algorithm have been considered [Dhillon and Sra, 2003, Banerjee et al., 2003, 2005]. There is comparatively less development in Bayesian methodology. Taghia et al. [2014] provides a variational inference method to fit Bayesian mixture vMF, and Gopal and Yang [2014] consider different variations of Bayesian vMF with graphical modeling approaches based on variational inference and collapsed Gibbs sampling [Pewsey and García-Portugués, 2021]. In this dissertation we further consider the Bayesian vMF mixture model. Further literature review and results could be found in Chapter **??**.

1.3 Canonical Correlation Analysis

In this section we introduce Canonical Correlation Analysis and establish its relation with directional data analysis.

As a powerful tool for data integration, Canonical Correlation Analysis (CCA) has received widespread attention. Originally proposed by Hotelling [1936], it is one of the most prominent techniques to integrate analysis between two or more data views. This technique maximizes the Pearson correlation between a linear combination of each data view to find components which are associated with each other. The key of idea of CCA is to project the complicated high dimensional variables within each view to low-dimensional latent spaces which are correlated across views. This enables analysis of two or more differently dimensional data sets. CCA has been widely used to analyze such multiview data sets in the areas of genomics [Witten and Tibshirani, 2009, Waaijenborg et al., 2008], computer vision [Lin et al., 2006, Zhang et al., 2013], meteorology [Statheropoulos et al., 1998], biomedicine [Li et al., 2009, Zhang et al., 2014], imaging analysis [Lin et al., 2014, Du et al., 2016], among others. Readers are referred to Yang et al. [2019] and Zhuang et al. [2020] for a more extensive review of the CCA literature.

Now, we formalize the mathematics of CCA. We will focus only on the 2-views form of CCA. Let $\mathbf{X}^{(1)} \in \mathbb{R}^{n \times p^{(1)}}$ and $\mathbf{X}^{(2)} \in \mathbb{R}^{n \times p^{(2)}}$ be the full data matrices of the two views. We denote scalar, vector and matrix parameters quantities by lowercase roman, lowercase bold and uppercase bold letters, respectively. The sample size is given by n, and $p^{(m)}$ represents the dimensionality of each view $(m = 1, 2)$. Without loss of generality, we assume that all features are centered in this subsection; that is, $E[\boldsymbol{x}_{i.}^{(m)}] = \mathbf{0}$ for $i = 1, 2, \ldots, n$ and $m = 1, 2$. The matrix $\boldsymbol{\Sigma}$ represents the joint covaraince between the two views $\boldsymbol{x}_{i.}^{(1)}$ and $\boldsymbol{x}_{i.}^{(2)}$ and can be written in a block-wise formulation

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}^{(11)} & \boldsymbol{\Sigma}^{(12)} \\ \boldsymbol{\Sigma}^{(21)} & \boldsymbol{\Sigma}^{(22)} \end{bmatrix}. \tag{1.4}$$

Here, $\boldsymbol{\Sigma}^{(11)}$ represents the covariance matrix for the view 1 data, $\boldsymbol{\Sigma}^{(22)}$ the covariance matrix for the view 2 data, and $\boldsymbol{\Sigma}^{(12)} = \boldsymbol{\Sigma}^{(21)T}$ is the covariance between the two views.

CCA aims to find the optimal vectors $\mathbf{u} \in \mathbb{R}^{p^{(1)}}$ and $\mathbf{v} \in \mathbb{R}^{p^{(2)}}$so that the Pearson correlation between the linear combination of $\mathbf{X}_i^{(1)}\mathbf{u}$ and $\mathbf{X}_i^{(2)}\mathbf{v}$ is maximized. Here $\mathbf{u}$ and $\mathbf{v}$ act as linear summaries which form linear combinations of the features for observation i, respectively. CCA optimization problem can be formulated as

$$\arg\max_{\mathbf{u},\mathbf{v}} \left\{ \frac{\mathbf{u}^t \boldsymbol{\Sigma}^{(11)-1/2} \boldsymbol{\Sigma}^{(12)} \boldsymbol{\Sigma}^{(22)-1/2} \mathbf{v}}{\sqrt{\mathbf{u}^t\mathbf{u}}\sqrt{\mathbf{v}^t\mathbf{v}}} \right\}.$$

This constrained optimization problem can be reformulated as

$$\rho = \max_{\mathbf{u}^*,\mathbf{v}^*} \left\{ \mathbf{u}^{*t} \boldsymbol{\Sigma}^{(11)-1/2} \boldsymbol{\Sigma}^{(12)} \boldsymbol{\Sigma}^{(22)-1/2} \mathbf{v}^* : \mathbf{u}^* \in \mathbb{S}^{p^{(1)}-1}, \mathbf{v}^* \in \mathbb{S}^{p^{(2)}-1} \right\}. \tag{1.5}$$

Note that $\mathbb{S}^{p-1} = \{\boldsymbol{x} \in \mathbb{R}^p : \|\boldsymbol{x}\|_2 = 1\}$, where $\|\boldsymbol{x}\|_2 = \sqrt{\boldsymbol{x}^T\boldsymbol{x}}$, is the compact manifold of the set of p-dimensional vectors with norm 1. These $\mathbf{u}^*$ and $\mathbf{v}^*$ are called the first canonical loadings and represent the directions in which the first canonical correlation is maximized. The above equation transforms CCA model to a Quadratic Constrained Quadratic Program (QCQP), and the canonical correlation is the maximal solution of ρ [Sharma et al., 2012].

In this way, we can observe that the direction vectors reside in a support which is a manifold, and, hence, the CCA problem is intimately related to the directional statistics methodology. In the CCA, primary interest is often in estimating $(\rho, \mathbf{u}^*, \mathbf{v}^*)$

triple, representing the first canonical correlation and direction vectors, but there is sometimes interest in higher order canonical correlation terms that represent the next most impactful areas of dependence between the two views. After finding the $(r-1)$th triple $(\rho_{r-1}, \mathbf{u}^{*(r-1)}, \mathbf{v}^{*(r-1)})$, the rth set of canonical correlation parameters are found from

$$\begin{aligned}\rho_r = \quad & \max_{\mathbf{u}^*,\mathbf{v}^*} \Big\{ \mathbf{u}^{*t}\mathbf{\Sigma}^{(11)-1/2}\mathbf{\Sigma}^{(12)}\mathbf{\Sigma}^{(22)-1/2}\mathbf{v}^* : \mathbf{u}^* \in \mathbb{S}^{p^{(1)}}, \mathbf{v}^* \in \mathbb{S}^{p^{(2)}}, \\ & \qquad \mathbf{u}^{*t}\Sigma^{(11)}\mathbf{u}^{*(j)} = 0, \mathbf{v}^{*t}\Sigma^{(22)}\mathbf{v}^{*(j)} = 0, (j = 1, \dots, r-1) \Big\}. \end{aligned} \tag{1.6}$$

Optimization of (1.6) finds the vectors $\mathbf{u}^* = \mathbf{u}^{*(r)}$ and $\mathbf{v}^* = \mathbf{v}^{*(r)}$ that maximizes the correlation while yielding linear combinations $\mathbf{X}_i^{(1)}\mathbf{u}^{*(r)}$ and $\mathbf{X}_i^{(2)}\mathbf{v}^{*(r)}$ that are uncorrelated with the previous $r-1$ combinations.

Researchers often have high dimensional data where the number of features measured on each subject are frequently much greater than number of subjects itself $(p \gg n)$. This results in inefficiency due to overfitting for traditional CCA based on the empirical estimates of $\mathbf{\Sigma}$ in (1.4). Sparse Canonical Correlation Analysis (SCCA) is a technique which addresses this problem by finding the meaningful features which contribute to the calculation of the canonical correlation (CC) while discarding the uninformative features. In frequentists approach, we have penalized methods which bring in element wise sparsity on directional vectors such as Parkhomenko et al. [2009], Witten and Tibshirani [2009], Waaijenborg et al. [2008], Suo et al. [2017]. Also there are methods which consider underlying structure in the data to bring in sparsity which are termed as Structure Sparse Canonical Correlation Analysis. Some of these methods are Chen et al. [2012], Lin et al. [2013, 2014], among many others. Bayesian CCA has received attention due to advances in Bayesian factor modeling [Wang, 2007, Klami and Kaski, 2007, Klami et al., 2013, Zhao et al., 2016]. However, there is a lack of Bayesian methodology which takes into account the within view sparse interelations. In our first project we propose a model which tries fills the gap in this literature.

CHAPTER 2

A BAYESIAN METHODOLOGY FOR ESTIMATION FOR SPARSE CANONICAL CORRELATION

2.1 Introduction

Advances in technology have resulted in multiple feature sets measured on same subjects which results in multiple data sets. They are also called a multiview datasets. Analyses of these multiview data sets together can improve understanding of the experiment as it facilitates further insight about a common set of subjects by borrowing information from the different views. Understanding the relationships between such multiview data is a challenging task as often this data is high dimensional. These challenges make development of new statistical theory a necessity.

As a powerful tool for data integration, Canonical Correlation Analysis (CCA) has received widespread attention. Originally proposed by Hotelling [1936], it is one of the most prominent techniques to integrate analysis between two or more data views. This technique maximizes the Pearson correlation between a linear combination of each data view to find components which are associated with each other. The key of idea of CCA is to project the complicated high dimensional variables within each view to low-dimensional latent spaces which are correlated across views. This enables analysis of two or more differently dimensional data sets. CCA has been widely used to analyze such multiview data sets in the areas of genomics [Witten and Tibshirani, 2009, Waaijenborg et al., 2008], computer vision [Lin et al., 2006, Zhang et al., 2013], meteorology [Statheropoulos et al., 1998], biomedicine [Li et al., 2009, Zhang et al., 2014], imaging analysis [Lin et al., 2014, Du et al., 2016], among others. Readers are referred to Yang et al. [2019] and Zhuang et al. [2020] for a more extensive review of the CCA literature.

Researchers often have high dimensional data where the number of features measured on each subject are frequently much greater than number of subjects itself ($p \gg n$). This results in inefficiency of traditional CCA due to overfitting. Sparse Canonical Correlation Analysis (SCCA) is a technique which addresses this problem by finding the meaningful features which contribute to the calculation of the canonical correlation (CC) while discarding the uninformative features.

In the frequentist approach to SCCA, there are two main strategies. First is element-wise sparsity which generally employ different types of penalties such as an l_1-norm penalty, fused lasso or their combination on the canonical loadings [Parkhomenko et al., 2009, Witten and Tibshirani, 2009, Waaijenborg et al., 2008, Suo et al., 2017]. A second approach to sparse CCA is Structured Sparse CCA (ScSCCA) where the structure of the interconnections between the views are taken into account to apply penalties on the canonical loadings. For example, known biological relationships between genes could determine such structure. Group lasso based ScSCCA relies on prior knowledge regarding the structure/interconnection in the data to define groups [Chen et al., 2012, Lin et al., 2013, 2014]. However, for a particular biological function or disease, we may not have complete information about underlying relationships among genes, and this incomplete knowledge is difficult to incorporate into analysis. This leads to the development of ScSCCA approaches that involve using graph/network guided fused Lasso penalties [Chen et al., 2012, 2013, Yan et al., 2014, Du et al., 2015, Chen and Liu, 2012]. When prior knowledge about the structure is not available, these methods may use the sample correlation to estimate a graph structure. These methods mainly depend on the sign of the pairwise sample correlation to identify the underlying hidden pattern but may introduce bias in CCA estimation from errors in structure estimation step [Du et al., 2016]. For more elaborate review of frequentist methods, readers are referred to Yang et al. [2019].

Bayesian development of CCA methodology has received attention in recent years. In Klami and Kaski [2007] and Klami et al. [2013], Bayesian Interbattery Factor Analysis (IBFA) models that prove useful for CCA estimation are introduced. In the IBFA model [Tucker, 1958], dependence between views are explained by shared latent factors, so CCA is naturally related to the IBFA framework. The IBFA model decomposes the covariance for a particular view into a factor structure shared across all views and the view-specific noise. The covariance of these noise/residuals terms is typically considered as a diagonal matrix whose elements are sometimes referred to as the specific variances [Johnson and Wichern, 2007]. In more realistic scenarios this assumption of independent noise among the features of same view may not hold and may prove inadequate to capture the interdependence among the features. This motivates us to develop a new Bayesian methodology for modelling of CCA which will explore these interdependences without restricting conditional independence given the factors.

To that end, we propose a novel Bayesian method, which performs joint estimation of canonical correlations as well as within view covariance estimation by using sparsity-inducing priors. A key contribution in this article is the introduction of a novel infinite Bayesian factor shrinkage model through which we gain shrinkage in the projection matrices of the IBFA. This Multiplicative Half Cauchy Process provides flexible and adaptive dimension reduction for the factor loading coefficients. We also estimate the within view covariance for each set of observation using the recently developed graphical horseshoe model by Li et al. [2019]. This is a fully Bayesian model yielding striaghtforward uncertainty quantification through posterior analysis of Markov chain Monte Carlo samples.

The article is organized as follows. In section 2.2 we review the mathematical formulation of CCA. In section 2.3 we introduce our model and justify the choice of priors. Section 2.4 proposes our Markov chain Monte Carlo (MCMC) posterior sampling algorithm and discusses issues related to estimation and inference. Section 2.5 consists of comparisons across competing estimation approaches in different simulation settings. In section 2.6 we analyze a set of breast cancer data to investigate the relationship between copy number and gene expression. Section 2.7 summarizes the project and discusses future directions.

2.2 Factor Model Formation of CCA

2.2.1 Canonical Correlation Analysis

In this section we formalize the mathematics of CCA and discuss a commonly used factor model for CCA. We will focus only on the 2-view form of CCA, but the developed theory can be applied for a multiview set up as we discuss in section 2.7. Let $\mathbf{X}^{(1)} \in \mathbb{R}^{n\times p^{(1)}}$ and $\mathbf{X}^{(2)} \in \mathbb{R}^{n\times p^{(2)}}$ be the full data matrices of the two views. We denote scalar, vector and matrix parameters quantities by lowercase roman, lowercase bold and uppercase bold letters, respectively. The sample size is given by n, and $p^{(m)}$ represents the dimensionality of each view ($m = 1, 2$). Without loss of generality, we assume that all features are centered in this subsection; that is, $E[\boldsymbol{x}_{i.}^{(m)}] = \mathbf{0}$ for $i = 1, 2, \ldots, n$ and $m = 1, 2$. The matrix $\mathbf{\Sigma}$ represents the joint covaraince between the two views $\boldsymbol{x}_{i.}^{(1)}$ and $\boldsymbol{x}_{i.}^{(2)}$ and can be written in a block-wise formulation

$$\mathbf{\Sigma} = \begin{bmatrix} \mathbf{\Sigma}^{(11)} & \mathbf{\Sigma}^{(12)} \\ \mathbf{\Sigma}^{(21)} & \mathbf{\Sigma}^{(22)} \end{bmatrix}. \tag{2.7}$$

Here, $\mathbf{\Sigma}^{(11)}$ represents the covariance matrix for the view 1 data, $\mathbf{\Sigma}^{(22)}$ the covariance matrix for the view 2 data, and $\mathbf{\Sigma}^{(12)} = \mathbf{\Sigma}^{(21)^T}$ is the covariance between the two views.

CCA aims to find the optimal vectors $\mathbf{u} \in \mathbb{R}^{p^{(1)}}$ and $\mathbf{v} \in \mathbb{R}^{p^{(2)}}$so that the Pearson correlation between the linear combination of $\mathbf{X}_i^{(1)}\mathbf{u}$ and $\mathbf{X}_i^{(2)}\mathbf{v}$ is maximized.

Here $\mathbf{u}$ and $\mathbf{v}$ act as linear summaries which form linear combinations of the features for each observation i, respectively. The CCA optimization problem can be formally stated as

$$\arg\max_{\mathbf{u},\mathbf{v}} \left\{ \frac{\mathbf{u}^T \boldsymbol{\Sigma}^{(11)-1/2} \boldsymbol{\Sigma}^{(12)} \boldsymbol{\Sigma}^{(22)-1/2} \mathbf{v}}{\sqrt{\mathbf{u}^t\mathbf{u}}\sqrt{\mathbf{v}^t\mathbf{v}}} \right\}.$$

This constrained optimization problem can be reformulated as

$$\rho = \max_{\mathbf{u}^*,\mathbf{v}^*} \left\{ \mathbf{u}^{*T} \boldsymbol{\Sigma}^{(11)-1/2} \boldsymbol{\Sigma}^{(12)} \boldsymbol{\Sigma}^{(22)-1/2} \mathbf{v}^* : \mathbf{u}^* \in \mathbb{S}^{p^{(1)}}, \mathbf{v}^* \in \mathbb{S}^{p^{(2)}} \right\}. \tag{2.8}$$

Note that $\mathbb{S}^p = \{\boldsymbol{x} \in \mathbb{R}^p : \|\boldsymbol{x}\|_2 = 1\}$, where $\|\boldsymbol{x}\|_2 = \sqrt{\boldsymbol{x}^T\boldsymbol{x}}$, is the compact manifold of the set of p-dimensional vectors with norm 1. These $\mathbf{u}^*$ and $\mathbf{v}^*$ are called the first canonical loadings and represent the directions in which the first canonical correlation is maximized. Hence in this article, we will use the term canonical loadings and direction vectors interchangebly. The above equation transforms CCA model to a Quadratic Constrained Quadratic Program (QCQP), and the canonical correlation is the maximal solution of ρ [Sharma et al., 2012].

Primary interest is often in this $(\rho, \mathbf{u}^*, \mathbf{v}^*)$ triple, representing the first canonical correlation and direction vectors, but there is sometimes interest in higher order canonical correlation terms that represent the next most impactful areas of dependence between the two views. After finding the $(r-1)^{th}$ triple $(\rho_{r-1}, \mathbf{u}^*_{(r-1)}, \mathbf{v}^*_{(r-1)})$, the r^{th} set of canonical correlation parameters are found from

$$\begin{aligned} \rho_r = \quad & \max_{\mathbf{u}^*,\mathbf{v}^*} \Big\{ \mathbf{u}^{*T} \boldsymbol{\Sigma}^{(11)-1/2} \boldsymbol{\Sigma}^{(12)} \boldsymbol{\Sigma}^{(22)-1/2} \mathbf{v}^* : \mathbf{u}^* \in \mathbb{S}^{p^{(1)}}, \mathbf{v}^* \in \mathbb{S}^{p^{(2)}}, \quad (2.9) \\ & \qquad \mathbf{u}^{*T} \Sigma^{(11)} \mathbf{u}^{*(j)} = 0, \mathbf{v}^{*T} \Sigma^{(22)} \mathbf{v}^{*(j)} = 0, (j = 1, \ldots, r-1) \Big\}. \end{aligned}$$

Optimization of (2.9) finds the vectors $\mathbf{u}^* = \mathbf{u}^{*(r)}$ and $\mathbf{v}^* = \mathbf{v}^{*(r)}$ that maximizes the correlation while yielding linear combinations $\mathbf{X}_i^{(1)}\mathbf{u}^{*(r)}$ and $\mathbf{X}_i^{(2)}\mathbf{v}^{*(r)}$ that are uncorrelated with the previous $r-1$ combinations.

With an estimate of the joint covariance matrix, we obtain the estimated canoncial correlations and direction vectors through a singular value decomposition (SVD) of $\boldsymbol{\Sigma}$; derivation and more details of which could be found in Yang et al. [2019]. Briefly, under SVD the $p^{(1)} \times p^{(2)}$ matrix $\mathbf{M} = \boldsymbol{\Sigma}^{(11)-1/2}\boldsymbol{\Sigma}^{(12)}\boldsymbol{\Sigma}^{(22)-1/2}$ from the optimization problem (2.8) is decomposed as $\mathbf{M} = \mathbf{LPQ}^T$. $\mathbf{P} \in \mathbb{R}^{p^{(1)} \times p^{(2)}}$ is a diagonal matrix of singular values, which after ordering turn out to be the canonical correlations $\rho_1, \rho_2, \ldots$. The matrices $\mathbf{L}$ and $\mathbf{Q}$ are the corresponding left and right eigenvectors whose columns live in $\mathbb{S}^{p^{(m)}}$ $(m = 1, 2)$, providing the corresponding canonical loadings.

2.2.2 Latent Factor Model

Bach and Jordan [2005] proposed a latent factor model for two view CCA. Their model is given as

$$\begin{aligned}
\mathbf{z}_{i.} &\sim MVN_d(0, \mathbf{I}) \\
\boldsymbol{x}_{i.}^{(1)} \mid \mathbf{A}^{(1)}, \boldsymbol{\mu}^{(1)}, \mathbf{z}_{i.}, \boldsymbol{\Phi}^{(1)} &\sim MVN_{p^{(1)}}(\boldsymbol{\mu}^{(1)} + \mathbf{A}^{(1)}\mathbf{z}_{i.}, \boldsymbol{\Phi}^{(1)}) \\
\boldsymbol{x}_{i.}^{(2)} \mid \mathbf{A}^{(2)}, \boldsymbol{\mu}^{(2)}, \mathbf{z}_{i.}, \boldsymbol{\Phi}^{(2)} &\sim MVN_{p^{(2)}}(\boldsymbol{\mu}^{(2)} + \mathbf{A}^{(2)}\mathbf{z}_{i.}, \boldsymbol{\Phi}^{(2)}).
\end{aligned} \tag{2.10}$$

Here, d is the dimension of the latent variable $\mathbf{z}_i$ which is less than the data dimensions $(\min(p^{(1)}, p^{(2)}) \gg d)$. In this model $\mathbf{A}^{(1)}$ and $\mathbf{A}^{(2)}$ project the lower dimensional latent space $\mathbf{z}_i$ to the higher dimensional data spaces $\boldsymbol{x}_{i.}^{(1)}$ and $\boldsymbol{x}_{i.}^{(2)}$, respectively. Here $\boldsymbol{\Phi}^{\mathbf{(1)}}$ and $\boldsymbol{\Phi}^{(2)}$ are within view covariance matrices representing variability beyond the factor structure. Generally, these matrices are considered to be diagonal, and these variances are referred to as the specific variance for each feature. For this study we frequently allow these matrices to have non-zero off-diagonal elements, so borrowing from the specific variance terminology, we will refer to the matrices $\boldsymbol{\Phi}^{\mathbf{(1)}}$ and $\boldsymbol{\Phi}^{(2)}$ as a "generalized specificity" for the view. Marginalizing out the latent $\mathbf{z}_{i.}$ yields the joint covariance matrix

$$\mathbf{\Sigma} = \begin{bmatrix} \mathbf{A}^{(1)}\mathbf{A}^{(1)T} + \mathbf{\Phi}^{(1)} & \mathbf{A}^{(1)T}\mathbf{A}^{(2)} \\ \mathbf{A}^{(2)T}\mathbf{A}^{(1)} & \mathbf{A}^{(2)}\mathbf{A}^{(2)T} + \mathbf{\Phi}^{(2)} \end{bmatrix}, \tag{2.11}$$

with the obvious connections to the block structure of (2.7). When $\mathbf{\Phi}^{(m)}$ are diagonal, the above matrix factorization substantially reduces the number of parameters to be estimated in the covariance matrix from approximately $(p^{(1)}+p^{(2)})^2/2$ to $(p^{(1)}+p^{(2)}) \times (d+1)$. Importantly, the covariance between two views are dependent on the product of the two projection matrices, and so $\mathbf{A}^{(1)T}\mathbf{A}^{(2)}$ will be the critical component to CCA. To that end we can write (2.8) as

$$\rho = \max_{\mathbf{u}^* \in \mathbb{S}^{p^{(1)}}, \mathbf{v}^* \mathbb{S}^{p^{(2)}}} \left\{ \mathbf{u}^{*T} (\mathbf{A}^{(1)}\mathbf{A}^{(1)T} + \mathbf{\Phi}^{(1)})^{(-1/2)} \mathbf{A}^{(1)T}\mathbf{A}^{(2)} (\mathbf{A}^{(2)}\mathbf{A}^{(2)T} + \mathbf{\Phi}^{(2)})^{-1/2} \mathbf{v}^* \right\}. \tag{2.12}$$

We can clearly see that the optimization problem and the estimands $(\rho, \mathbf{u}^*, \mathbf{v}^*)$ are a function of projection matrices $\mathbf{A}^{(m)}$ and generalized specificity matrices $\mathbf{\Phi}^{(m)}$.

It is important to note that as with all factor model, this decomposition is not identifiable without any further constraints. One can specify any semi-orthogonal matrix $\mathbf{O}$ such that $\mathbf{O}\mathbf{O}^T = \mathbf{I}_{d \times d}$ and obtain $\tilde{\mathbf{A}}^{(m)} = \mathbf{A}^{(m)}\mathbf{O}$. Substituting $\tilde{\mathbf{A}}^{(m)}$ in (2.11), the overall $\mathbf{\Sigma}$ is unaffected, showing that these model parameters are unidentifiable. Assuming some identifiablity conditions, such as $\mathbf{A}^{(m)}$ as lower triangular matrices, can solve the problem but induces order dependence in the features. Alternatively, specialized structures can be imposed to assign some special role to a few features, but such restrictions method are not generalizable and need domain specific expertise to choose the structure [Carvalho et al., 2008, Zhao et al., 2016].

The unidentifiablity of the projection matrices does not impact CCA estimands because the covariances $\mathbf{\Sigma}^{(11)}, \mathbf{\Sigma}^{(12)}, \mathbf{\Sigma}^{(22)}$ are indentifiable, and the CCA estimands are functions of these covariance matrices [Bhattacharya and Dunson, 2011, Geweke and Zhou, 1996]. Rather than imposing structural zeros in the lower-diagonal portion of the project, sparsity inducing priors can further support model stability and minimize issues related to non-identifiablity.

2.3 Factor Shrinkage Model for Canonical Correlation Estimation

2.3.1 Non-Diagonal Factor Shrinkage Model (NDFSM)

In this section we introduce our Non-Diagonal Factor Shrinking Model (NDFSM) for CCA. An important motivation is that most competing CCA models assume diagonal structure for within view specificity matrices $\mathbf{\Phi}^{(1)}$ and $\mathbf{\Phi}^{(2)}$. This diagonal restriction may not be appropriate in some practical scenarios like genomics data where $\mathbf{\Phi}^{(1)}$ and $\mathbf{\Phi}^{(2)}$ may require non-zero off-diagonal elements due to interactions among the features. By using the IBFA framework, we reparameterize the joint covariance matrix $\mathbf{\Sigma}$ using the projection and generalized specificity matrices that can be naturally modeled through sparse and lower-dimension considerations.

Horseshoe priors [Carvalho et al., 2010] belongs to a wider class of global local shrinkage priors that are characterized by a local shrinkage parameter for recovering large signals and a global shrinkage parameter for adapting to overall sparsity. In addition to their use in standard regression settings, they have also been used in latent factor models [Sekula et al., 2021]. This class of global-local shrinkage priors exhibit a set of common features including heavy tails for robustness and appreciable mass near zero for sparsity, leading to shared optimality properties. This is the key strength and advantage of the horseshoe prior, and hence, we will use it to motivate our prior construction.

In our model we use a horseshoe-like prior on each element of the projection matrices. Our prior for $\mathbf{A}^{(m)} = \{a_{jk}^{(m)}\}$ $(j = 1, \ldots, p^{(m)}$, $k = 1, \ldots, d$ and $m = 1, 2)$

has the following structure:

$$
\begin{aligned}
a_{jk}^{(m)} &\sim N(0, \tau^{2(m)}\eta_k^2\lambda_{jk}^{2(m)}) \\
\lambda_{jk}^{(m)} &\sim C^+(0,1) \\
\tau^{(m)} &\sim C^+(0,1) \\
\eta_k^2 &= \prod_{j=1}^{k}\tilde{\eta}_j^2 \\
\tilde{\eta}_j &\sim C^+(0,\Lambda),\ (j>1);\quad \tilde{\eta}_1 = 1.
\end{aligned}
$$

Each element $a_{jk}^{(m)}$ of the projection matrix $\mathbf{A}^{(m)}$ has variance term $\tau^{2(m)}\eta_k^2\lambda_{jk}^{2(m)}$. Here, η_k^2 is a factor specific shrinkage parameter which controls sparsity of the each column of projection matrix, behaving as a local parameter. $\tau^{2(m)}$ is a view-specific shrinkage parameter which accounts for the overall variability of the view, and hence, acts as a global shrinkage parameter for all $p^{(m)}d$ coefficients in that view. In addition to this global-local structure, we also introduce the hyperlocal shrinkage parameter $\lambda_{jk}^{(m)}$ which account for element-wise variability in the projection matrix. The hyperlocal parameters $\lambda_{jk}^{(m)}$ and global parameters $\tau^{(m)}$ follow a standard half-Cauchy prior. Here $C^+(0,\Lambda)$ represents a random variable with density $p(x) \propto (1 + x^2/\Lambda^2)^{-1}\mathbb{I}(x > 0)$.

As mentioned previously, factor models are a common approach to modeling dependence in high dimensional data and have been frequently used in Bayesian contexts [Archambeau and Bach, 2009, Carvalho et al., 2008, Bhattacharya and Dunson, 2011, Zhao et al., 2016]. The selection of the number of latent factors, d, is an important consideration to any factor model. A common practice is to fit the model with different d and then use a model selection criteria to obtain the top fitting model. It is also possible to put a prior on d and obtain posterior samples through reversible jump MCMC [Lopes and West, 2004, Miller and Harrison, 2018, Yang et al., 2018] However, these algorithms tend to mix poorly when used outside of their original context.

An alternative approach of intentionally over-fitting the factorization model is introduced by Bhattacharya and Dunson [2011] for Gaussian linear factor models. In such an approach they allow the number of factors d to diverge to infinity while using shrinkage priors that force unnecessary components to be adaptively removed by concentrating mass around only meaningful components. As additional factors are added to the model, they play a progressively less important role in explaining the structure of the data, and therefore, the contribution of the parameters associated with those factors should be stochastically decreasing [Legramanti et al., 2020]. In a similar spirit to Bhattacharya and Dunson [2011]'s multiplicative gamma process, we refer to our model structure as a multiplicative half Cauchy process. Since $\Lambda < 1$, the factor-specific shrinkage variances η_k^2s are stochastically decreasing in k, and due to the non-zero mass near zero from the half-Cauchy, η_k tend to be quite small for larger k. Hence, some columns of $\mathbf{A}^{(m)}$ will be shrunk approximately to zero, effectively removing the factor from the model. Consequently, our multiplicative process is able to borrow information across the two views through the projection matrices $\mathbf{A}^{(m)}$ to adaptively determine the number of factors that effectively play a role in the model. While the model is an infinite multiplicative process when $d \to \infty$, in practice one chooses a relatively large d and investigates the behavior of η_d^2 to ensure that it is approximately zero. While our method can accommodate any $\Lambda \in (0, 1)$, we generally recommend the value of $\Lambda = 0.5$

As noted above, there have been previous attempts at latent factor models for CCA analysis through the same IBFA framework. In particular, Wang [2007] uses an automatic relevance determination (ARD) prior find structure and sparsity in the projection matrices along with inverse Wishart priors on $\mathbf{\Phi}^{(1)}$ and $\mathbf{\Phi}^{(2)}$; Klami and Kaski [2007] follow a similar strategy with non-diagnonal structures for $\mathbf{\Phi}^{(m)}$. Similarly, Klami et al. [2013] uses a Inter-Battery Factor Analysis model with an ARD prior to impose structure in the projection matrices. Importantly, this structure encourages a column k to be active in both projections (inducing correlation between views) or in only one view (to induce correlation within the view independent of the other view). As this ARD prior allows both within and across covariance, the authors argue that diagonal $\mathbf{\Phi}^{(m)}$ are sufficient in their models. Zhao et al. [2016] consider an IBF model with a three level regularization in terms of global, factor specific and local shrinkage. The model uses normal scale mixture model and three parameter beta distribution to provide sparsity. The authors also assume diagonal structures for $\mathbf{\Phi}^{(m)}$ with inverse gamma priors on the diagonal elements.

Having specified our sparse prior process for the $\mathbf{A}^{(m)}$ matrices, we turn to the prior structure for the generalized specificity matrices $\mathbf{\Phi}^{(m)}$. It is important to remember from (2.11) that $\mathbf{\Sigma}^{(m)} = \mathbf{A}^{(m)}\mathbf{A}^{(m)T} + \mathbf{\Phi}^{(m)}$, so this matrix represents the covariance between features of the same view that is unexplained by the factor structure. Unlike the previously mentioned methods, we assume that the structure is arbitrary, and do not impose a diagonal structure that assumes the factors explain the entire dependence. However, we do believe that this matrix is likely to be highly sparse as the majority of the structure should be captured by the shared factors, so a prior such as inverse Wishart would be ineffective in this case.

To that end, we apply the graphical horseshoe prior by Li et al. [2019] on the inverse of the generalized specificity matrices $\mathbf{\Phi}^{(m)-1} = \mathbf{\Omega}^{(m)}$ for $m = 1, 2$ as follows:

$$\begin{aligned}
\omega_{ii}^{(m)} &\propto 1 \quad \text{(Flat Prior)} \quad i = 1, \ldots, p^{(m)} \\
\omega_{ij}^{(m)} &\sim N(0, \alpha_{ij}^{2(m)}\beta^2) \quad i < j \\
\alpha_{ij}^{(m)} &\sim C^+(0,1) \quad i < j \\
\beta &\sim C^+(0,1).
\end{aligned} \tag{2.13}$$

The graphical horseshoe model puts horseshoe priors on the off-diagonal elements of the precision matrix and an uninformative prior on the diagonal elements, all under the constraint that $\mathbf{\Omega}^{(m)}$ remain in the space of $p^{(m)} \times p^{(m)}$ positive definite matrices. This also enforces a symmetry constraint $\omega_{ij}^{(m)} = \omega_{ji}^{(m)}$. Despite the flat prior specification on ω_{ii}, the positive definiteness constraint on $\mathbf{\Omega}^{(m)}$ ensures that the full prior is proper [Li et al., 2019]. For the individual $\omega_{ij}^{(m)}$ terms, local shrinkage parameters $\alpha_{ij}^{(m)}$ preserve the magnitude of non-zero elements and shrink the zeros, while β adapts to the sparsity of the entire matrix $\mathbf{\Omega}^{(m)}$ as the global shrinkage parameter.

To complete model specification we require a prior on the mean vectors. We assume the mean-zero Gaussian prior for the view-specific mean vectors $\boldsymbol{\mu}^{(1)}$ and $\boldsymbol{\mu}^{(2)}$ with hyperparameter σ^2 typically taken to be 100. That is, the priors are given as

$$\begin{aligned} \boldsymbol{\mu}^{(1)} &\sim MVN_{p^{(1)}}(0, \sigma^2 \mathbf{I}) \\ \boldsymbol{\mu}^{(2)} &\sim MVN_{p^{(2)}}(0, \sigma^2 \mathbf{I}). \end{aligned} \tag{2.14}$$

2.3.2 Diagonal Factor Shrinkage Model (DFSM)

As many CCA factor models use a diagonal structure for the generalized specificity matrices, we also choose to construct the analogous version of our NDFSM that uses a diagonal structure, called the Diagonal Factor Shrinking Model (DFSM).

In this version instead of using the GHS prior (2.13) on the inverse generalized specificity matrices, we instead restrict them to be diagonal, $\mathbf{\Phi}^{(m)} = \text{diag}(\phi_{11}^{(m)}, \ldots, \phi_{p^{(m)}p^{(m)}}^{(m)})$. The element $\phi_{jj}^{(m)}$, that is the specific variance for feature j in view m, is given a conjugate inverse gamma prior $\phi_{jj}^{(m)} \sim IG(0.1, 0.1)$. The rest of the model structure is the same as NDFSM model, introduced in the previous section.

2.4 Posterior Sampling and Inference

2.4.1 MCMC Algorithm for NDFSM

We use an MCMC Gibbs sampling algorithm to draw posterior samples from our model. All the distributions are obtained through conjugacy. The sampling algorithm iterates between the following steps.

1. Mean Vectors: We will sample the full $(p^{(1)}+p^{(2)})$-dimensional mean vector $\boldsymbol{\mu}^g = (\boldsymbol{\mu}_1^T, \boldsymbol{\mu}_2^T)^T$ by marginalizing the factor scores $\mathbf{z}_i$. Let $\mathbf{X}^g$ be the $n \times (p^{(1)} + p^{(2)})$ matrix of observations obtained by stacking two data view matrices and $\boldsymbol{\Sigma}$ be the grand covariance matrix (2.11) based on the current values of $\mathbf{A}^{(m)}$ and $\boldsymbol{\Phi}^{(m)}$. Let $\bar{\mathbf{X}}^g$ be the $(p^{(1)} + p^{(2)})$-dimensional vector of column means. Then,

$$\begin{aligned} \boldsymbol{\mu}^g \mid \boldsymbol{\Sigma}, \mathbf{X}^g &\sim MVN_{(p^{(1)}+p^{(2)})}(\boldsymbol{\mu}^*, \mathbf{E}^{-1}) \\ \mathbf{E} &= n\boldsymbol{\Sigma}^{-1} + \sigma^{-2}\mathbf{I} \\ \boldsymbol{\mu}^* &= \mathbf{E}^{-1}\boldsymbol{\Sigma}^{-1}\bar{\mathbf{X}}^g. \end{aligned}$$

2. Latent variable $\mathbf{z}_{i.}$: The latent variable $\mathbf{z}_{i.}$ for $i = 1, 2, \ldots, n$ can be updated through conjugacy as follows:

$$\begin{aligned} \mathbf{z}_{i.} \mid \mathbf{A}^{(1)}, \mathbf{A}^{(2)}, \boldsymbol{\Phi}^{(1)}, \boldsymbol{\Phi}^{(2)} &\sim MVN_d(\boldsymbol{\mu}^*, \mathbf{E}^{*-1}) \\ \mathbf{E}^* &= \mathbf{I}_{d\times d} + \mathbf{A}^{(1)T}\boldsymbol{\Phi}^{(1)-1}\mathbf{A}^{(1)} + \mathbf{A}^{(2)T}\boldsymbol{\Phi}^{(2)-1}\mathbf{A}^{(2)} \\ \boldsymbol{\mu}^* &= \mathbf{E}^{*-1}(\mathbf{A}^{(1)T}\boldsymbol{\Phi}^{(1)-1}\boldsymbol{x}_{i.}^{(1)} + \mathbf{A}^{(2)T}\boldsymbol{\Phi}^{(2)-1}\boldsymbol{x}_{i.}^{(2)}). \end{aligned}$$

3. Projection matrices $\mathbf{A}^{(1)}$ and $\mathbf{A}^{(2)}$: To facilitate sampling from the posterior of projection matrices, we use the data augmentation structure of Makalic and Schmidt [2015] for sampling from a half Cauchy distribution.

(a) $\mathbf{A}^{(1)}$ and $\mathbf{A}^{(2)}$: For $m = 1, 2$ and each row $j = 1, \ldots, p^{(m)}$, we sample

$$\begin{aligned}
\mathbf{a}_{j.}^{(m)} \mid \mathbf{\Phi}^{(m)}, \mathbf{Z}, \mathbf{X}^{(m)}, \boldsymbol{\mu}^{(m)} &\sim MVN(\boldsymbol{\mu}, \mathbf{E}^{-1}) \\
\boldsymbol{\mu}^{(m)} &= \tilde{\phi}_j^{-1(m)} \mathbf{E}^{-1} \mathbf{Z}^T \tilde{\mathbf{X}}_{.j}^{(m)} \\
\mathbf{E} &= \tilde{\phi}_j^{-1(m)} \mathbf{Z}^T \mathbf{Z} + \mathbf{\Delta}^{-1} \\
\tilde{x}_{ij}^{(m)} &= x_{ij}^{(m)} - \mu_j^{(m)} - [\mathbf{\Phi}_{j,-j}^{(m)}][\mathbf{\Phi}_{-j-j}^{(m)}]^{-1}(\mathbf{X}_{i,-j}^{(m)} - \boldsymbol{\mu}_{-j}^{(m)} - \mathbf{A}_{-j.}^{(m)} \mathbf{z}_{i.}^T). \\
\tilde{\phi}_j^{(m)} &= \mathbf{\Phi}_{j,j}^{(m)} - \mathbf{\Phi}_{j,-j}^{(m)}[\mathbf{\Phi}_{-j-j}^{(m)}]^{-1}\mathbf{\Phi}_{-j,j}^{(m)}.
\end{aligned}$$

Here, $\mathbf{\Delta}$ is the $d \times d$ diagonal matrix of the shrinkage parameters; for $k = 1, \ldots, d$, $\Delta_{kk} = \tau^{2(m)} \eta_k^2 \lambda_{jk}^{2(m)}$. The element $\tilde{x}_{ij}^{(m)}$ from $\tilde{\mathbf{X}}_{.j}$ is the data residual for observation i after removing the effect of everything except jth response variable in m^{th} view. Here, $\tilde{\phi}_j^{(m)}$ is the variance of jth feature conditionally on the other features. In the above we use the common shorthand where $\mathbf{\Phi}_{ab}$ represents the sub-blocks of the matrix $\mathbf{\Phi}$ given by rows a and columns b; j indicates that only the j^{th} row/column is included and $-j$ denotes that all rows/columns except for the j^{th} are included.

(b) Hyperlocal Shrinkage parameters $\lambda_{jk}^{2(m)}$: For $m = 1, 2$; $j = 1, \ldots, p^{(m)}$; $k = 1, \ldots, d$,

$$\lambda_{jk}^{2(m)} \mid C_{jk}^{(m)}, a_{jk}^{(m)}, \eta_k \sim IG\Big(1, \frac{a_{jk}^{2(m)}}{2\tau^{2(m)}\eta_k^2} + \frac{1}{C_{jk}^{(m)}}\Big).$$

(c) View-Specific Shrinkage Parameter $\tau^{2(m)}$:

$$\tau^{2(m)} \mid F^{(m)}, \mathbf{A}^{(m)}, \eta_k \sim IG\Bigg(\frac{(p^{(m)} \times d) + 1}{2}, \sum_{k=1}^{d} \sum_{j=1}^{p^{(m)}} \frac{a_{jk}^{2(m)}}{2\lambda_{jk}^2 \eta_k^2} + \frac{1}{F^{(m)}}\Bigg).$$

(d) Column-wise Shrinkage Parameter η_k^2: For $j = 2$ to d,

$$\tilde{\eta_j}^2 \quad | \quad \mathbf{A}^{(m)}, \tilde{\eta}_{(-j)}, \tau^{(m)} \sim IG\Big(\frac{(d-(j-1))[p^{(1)}+p^{(2)}]+1}{2}, \\ \sum_{m=1}^{2}\sum_{k=j}^{d}\sum_{i=1}^{p^{(m)}} \frac{a_{ik}^{2(m)}}{2\lambda_{ik}^{2(m)}\tau^{2(m)}\prod_{\substack{k'=1\\k'\neq j}}^{k}\tilde{\eta}_{k'}^2} + \frac{1}{E_j}\Big).$$

After updating $\tilde{\eta_j}$ we compute $\eta_k^2 = \prod_{j=1}^{k}\tilde{\eta_j}^2$.

(e) Data-augmentation parameters $F^{(m)}$, $C_{jk}^{(m)}$, E: For $m = 1, 2$, $j = 1, 2, \ldots p^{(m)}$ and $k = 1, 2, \ldots, d$:

$$\begin{aligned} C_{jk}^{(m)} \mid \lambda_{jk}^{(m)} &\sim IG\Big(1, 1 + \frac{1}{\lambda_{jk}^{2(m)}}\Big) \\ F^{(m)} \mid \tau^{(m)} &\sim IG\Big(1, 1 + \frac{1}{\tau^{2(m)}}\Big) \\ E_j \mid \eta &\sim IG\Big(1, \frac{1}{\Lambda^2} + \frac{1}{\tilde{\eta_j}^2}\Big). \end{aligned}$$

4. View Specific Generalized Specificity Matrices $\mathbf{\Phi}^{(1)}$ and $\mathbf{\Phi}^{(2)}$: Under the NDFSM with the GHS prior, we simply follow the sampling schemes described in Li et al. [2019] based on the sample sum of square matrix given as

$$\mathbf{S}^{(m)} = \sum_{i=1}^{n}(\boldsymbol{x}_{i.}^{(m)} - \boldsymbol{\mu}^{(m)} + \mathbf{A}^{(m)}\mathbf{z}_{i.})^T(\boldsymbol{x}_{i.}^{(m)} - \boldsymbol{\mu}^{(m)} + \mathbf{A}^{(m)}\mathbf{z}_{i.}).$$

2.4.2 MCMC Algorithm for DFSM

For the version of the model that uses the diagonal specificity matrix, we simply replace step 4. of the above with the conjugate sampler. That is $\phi_{jj}^{(m)} \sim IG(0.5n + 0.1, 0.1 + 0.5\sum_{i=1}^{n}\tilde{x}_{ij}^2)$ for each $j = 1, \ldots, p^{(m)}$ and $m = 1, 2$. As in 4. of the previous algorithm, the vector of residuals is determined by $\tilde{\boldsymbol{x}}_{i.} = \boldsymbol{x}_{i.}^{(m)} - \boldsymbol{\mu}^{(m)} + \mathbf{A}^{(m)}\mathbf{z}_{i.}$.

2.4.3 Point Estimation and Inference

We use the corresponding Gibbs sampling algorithm to obtain a large number of posterior samples from our model. The main parameters required for CCA inference are the projection matrices $\mathbf{A}^{(m)}$ and the within view covariances $\mathbf{\Phi}^{(m)}$. These parameters determine the overall covariance structures among and across the views through (2.11) and determine the values of the canonical correlation ρ and the direction vectors $\mathbf{u}^*$ and $\mathbf{v}^*$ through (2.12). For each set of posterior samples of $\mathbf{A}^{(1)}, \mathbf{A}^{(2)}, \mathbf{\Phi}^{(1)}, \mathbf{\Phi}^{(2)}$, we can obtain a sample of $\rho, \mathbf{u}^*, \mathbf{v}^*$, as well as any higher order correlations and directions.

To evaluate mixing and convergence, we inspect traceplots of the CCs, the log-likelihood, and other model parameters to evaluate convergence and select a length for burn-in. Autocorrelation among the MCMC samples increases uncertainty in estimation of parameters, and the effective sample size measures this uncertainty by providing the number of independent samples that would contain an equivalent amount of information as the (correlated) samples from the given MCMC output. Typically, we seek to run the MCMC long enough to obtain an effective sample size of at least 1000 for the key parameters of interests.

To obtain CCA point estimates, we estimate the CCs by taking the sample mean from the estimated ρ_l from the MCMC output (after burn-in and thinning). Similarly, at each iteration we calculate the direction vectors based on orthonormal vectors obtained from the SVD, as described in section 2.2. This type of decomposition is not unique as the vectors could be reflected across the origin. To ensure that the canonical loadings are "pointing" in the same direction across MCMC samples, we first take the mean absolute values across all iterations for each element of the canonical loadings for both views and select the feature (from either view) with the largest absolute loading. We will impose an identifiability constraint on this feature to ensure that it maintains a positive sign in all iterations, so that all direction vectors are pointing in the same direction based on this influential feature. That is, if the loading for the selected feature is negative in a given iteration, we swap the signs of both $\mathbf{u}^*$ and $\mathbf{v}^*$ in that iteration to ensure that the correlation of $\mathbf{X}_i^{(1)}\mathbf{u}^*$ and $\mathbf{X}_i^{(2)}\mathbf{v}^*$ remains the same; if the loading for the selected feature is positive, we make no adjustment. After ensuring comparability across all iterations by imposing this identifiability constraint, an estimated $\hat{\mathbf{u}}^*$ and $\hat{\mathbf{v}}^*$ are obtained by averaging across iterations and dividing by the norm to ensure that they are unit 1.

An important step in CCA is determining which features significantly load onto the direction vectors; that is, which elements of $\mathbf{u}^*$ and $\mathbf{v}^*$ are significantly different from zero. To that end, we utilize a credible interval approach to determine significance. Based on the identifiability-adjusted posterior samples of $\mathbf{u}^*$ and $\mathbf{v}^*$, we obtain a credible interval for each element of each vector and investigate whether or not it contains zero. Recall that this orthonormal direction vectors are complex functions of the parameters $\mathbf{A}^{(m)}$ and $\boldsymbol{\Phi}^{(m)}$ which come from heavy-tailed horseshoe models. Consequently the posteriors for the elements of the direction vectors also tend to have heavy tails. It has been shown in a variety of contexts that a 95% credible interval under a heavy tail prior produce intervals that are overly wide and underpowered for hypothesis testing [van der Pas et al., 2017, Li et al., 2019]. Following the advice of Li et al. [2019] in the context of covariance selection in their GHS model, we use a 50% credible interval to determine if a feature is significantly loaded onto the direction vector. As we will show in the next section, this choice yields good performance in our empirical studies.

A final key inference question is which model should be used, either the general NDFSM that allows correlations between features of the same view through both the latent factor structure and the generalized specificity or the more restrictive DFSM that assumes independence of the features beyond the factors. As will be shown in the next section, we find that in some cases (particularly those with $p \gg n$) the NDFSM may overshrink the projection matrices $\mathbf{A}^{(m)}$ relative to the shrinkage imposed on the specificity matrices $\mathbf{\Phi}^{(m)}$. A consequence of this behavior is that $\mathbf{A}^{(1)}\mathbf{A}^{(2)T}$ will tend to mainly contain zeros, and the canonical correlation ρ will be very low. Fortunately, this is easy to diagnosis by investigating the estimate of ρ and can easily be corrected by instead using the DFSM. DFSM avoids overshrinking $\mathbf{A}^{(m)}$ by imposing maximal shrinkage in $\mathbf{\Phi}^{(m)}$ through zeros in all off-diagonal elements.

To determine which model should be used in a given data set, we recommend the following strategy. First run NDFSM model and check if the overshrinking may be happening by considering $P[\rho_1 < 0.2] > 0.5$, that is, if the event that the first CC is less than 0.2 has probability greater than 0.5, then we suspect overshrinkage may be happening, and instead the base inference of the DFSM output. Note that this threshold of 0.2 is somewhat ad hoc, and other users may prefer a different criteria for switching from the general NDFSM to the more constrained DFSM choice.

2.5 Simulations

2.5.1 Simulation Settings

To validate our proposed methodology across several situations and to compare it with some competing methods, we perform 7 simulation experiments. For synthetic data generation we use the latent model as specified in (2.10). We generated 7 different simulation set up with each containing 100 data sets with different settings of projection matrices and within view covariances.

The projection matrix $\mathbf{A}^{(m)}$ is generated by setting elements $1, 11, 21$ in first column of $\mathbf{A}^{(1)}$ to 1, while elements $1, 11$ in the first column of $\mathbf{A}^{(2)}$ are set to 1 and -1, respectively; all other elements in this column are zero. This first column is responsible for determining the first CC value. The elements in the other columns of the projection matrices are non-zero with probability 0.05 and drawn from standard normal. In some settings, we use an autoregressive structure for a non-diagonal choice of the generalized specificity $\mathbf{\Phi}^{(1)}$ and $\mathbf{\Phi}^{(2)}$ (autocorrelations of 0.4 and 0.2, respectively) and in other settings we use an identity matrix. In all cases, we use dimensions of $p^{(1)} = 100$ and $p^{(2)} = 50$ and consider $n = 300$ for a $n > p$ setting and $n = 50$ for a $p > n$ setting. The mean vector is always zero. We consider the true number of factors to be $d = 1$ so that there is only one non-zero canonical correlation, and let $d = 10$ so that there are 10 non-zero canonical correlations. In the 7th setting, we introduce a "scaling" parameter that we multiply all elements of projection matrix by. The scale is chosen to reduce the contribution of $\mathbf{\Sigma}^{(12)} = \mathbf{A}^{(1)T}\mathbf{A}^{(2)}$ in the decomposition (2.11), reducing the magnitude of the canonical correlation. We summarize the simulation settings, along with the resulting first two canonical correlations ρ_1 and ρ_2 in Table 2.1.

Table 2.1: Simulation Settings

Setting	$p^{(1)} = 100, p^{(2)} = 50$			
1	$n = 300$	AR Dependence	$d = 1$	$\rho_1 = 0.73, \rho_2 = 0.00$
2	$n = 50$	AR Dependence	$d = 1$	$\rho_1 = 0.73, \rho_2 = 0.00$
3	$n = 300$	$\mathbf{\Phi}^{(m)} = \mathbf{I}_{p^{(m)}}$	$d = 1$	$\rho_1 = 0.70, \rho_2 = 0.00$
4	$n = 50$	$\mathbf{\Phi}^{(m)} = \mathbf{I}_{p^{(m)}}$	$d = 1$	$\rho_1 = 0.70, \rho_2 = 0.00$
5	$n = 300$	AR Dependence	$d = 10$	$\rho_1 = 0.73, \rho_2 = 0.60$
6	$n = 50$	AR Dependence	$d = 10$	$\rho_1 = 0.73, \rho_2 = 0.60$
7	$n = 300$	AR Dependence Scaling=0.59	$d = 1$	$\rho_1 = 0.49, \rho_2 = 0.00$

For every dataset, we will obtain estimates $\hat{\rho}_1$ and $\hat{\rho}_2$ of the first two canonical correlations, as well as estimates of the direction vectors $\hat{\mathbf{u}}^*$ and $\hat{\mathbf{v}}^*$ for the first canonical correlation. We measure the accuracy of estimation of canonical correlation value as a root mean squared error (RMSE) between true value and estimated value

$$RMSE(\hat{\rho}_l, \rho_l) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{\rho}_{li} - \rho_l)^2}.$$

Here ρ_l and $\hat{\rho}_{li}$ is the l^{th} canonical correlation and its estimate in i^{th} dataset, and N is the total number of data sets. We also consider the average bias of the CC estimates by considering the average difference between estimate and the true value. As the canonical loadings vectors are the unit space, we calculate the error of canonical loading as a root mean error based on one minus the cosine similarity between of two unit vectors. We refer to this as root mean cosine error (RMCE) and compute it as

$$RMCE(\hat{\mathbf{u}}^*, \mathbf{u}^*) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(1 - (\hat{\mathbf{u}}_i^{*T}\mathbf{u})}).$$

Here $\mathbf{u}^*$ and $\hat{\mathbf{u}}_i^*$ is the direction vector and its estimates in i^{th} dataset. We obtain $RMCE(\hat{\mathbf{v}}^*, \mathbf{v}^*)$ for the direction vector of the second view in an equivalent way.

2.5.2 Competing Methods

For each of the 100 data sets generated according to the seven generation settings, we fit the data acorrding to the following methods. We compare our method with three frequentist and two Bayesian methods, explained as below.

1. NDFSM, DFSM: For both versions of models, MCMC is ran for 15,000 iterations with 5000 burn-in iterations in the low dimensional settings ($n = 50$). Samples are thinned to store 2000 samples. In the higher dimensional settings ($n = 300$), we ran the model for 300,000 iterations with the first 50,000 discarded as burn-in iterations, and a thinned sample of 5000 samples are stored. Implementation is done in R. The thinned samples on average give an effective sample size of 1000–1200 for both the first CC and the log-determinant of joint covariance matrix $\mathbf{\Sigma}$ from (2.11).

2. NDFMS+DFSM: We compare the combined strategy of selecting NDFSM vs DFSM relative to the fixed choice of each model. As noted in section 2.4.3, we base inference on the NDFSM posterior samples unless these samples yields $P[\rho_1 < 0.2] > 0.5$. In this case we suspect the NDFSM results may be impacted

by overshrinkage, and instead use the DFSM output for inference.

3. Bayesian Group Factor Analysis (GFA): The Bayesian Group Factor Model by Klami et al. [2013] provides a Bayesian competitor to our approach that comes from a similar motivation. This method uses a combination of spike-and-slab and ARD priors for the projection matrices and a diagonal structure for the $\mathbf{\Phi}^{(m)}$s. It is encoded in the "GFA" package for R [Leppäaho et al., 2017]. We ran MCMC for 600,000 iterations with 60,000 as burn-in with 2000 samples saved. This gives an effective sample size of approximately 1000 for the first CC. We perform inference using these posterior samples using the same procedures described in section 2.4.3.

4. Graphical Horseshoe (GHS): The GHS [Li et al., 2019] is directly applied to the joint data matrix $\mathbf{X}^g$ obtained by stacking both views, which has dimensions $n \times (p^{(1)} + p^{(2)})$. Hence, this approach directly estimates the overall grand covariance $\mathbf{\Sigma}$ without considering any distinction between features of the different views. Canonical correlations and direction vectors are calculated from the posterior samples of the joint covariance matrix through (2.12) and other inference steps follow as with the other Bayesian methods. We note that this choice of modeling the overall $\mathbf{\Sigma}$ covariance is not a common approach to performing CCA, but as the GHS imposes sparsity in $\mathbf{\Sigma}^{-1}$, it is conceivable that it can produce strong CCA estimates through its own form of regularization. We ran MCMC ran for 60,000 iterations with 5000 burn-in iterations. The sample is thinned to get 2000 samples yielding with effective sample size of approximately 1000.

5. Regularized CCA (RCCA): RCCA [Vinod, 1976] extends the regular CCA method for $p \gg n$ case by adding an ℓ_2 type penalty on the covariance matrix of each view. It is implemented as R package "CCA" by González et al. [2008]. The regularization parameter was chosen using a leave-one out cross

validation mechanism using the default choice from the estim.regul function.

6. Sparse CCA (SCCA): We consider two implementations of SCCA [Witten and Tibshirani, 2009]. Firstly, a Lasso (ℓ_1) penalty is put on the canonical loadings. We refer to this model as SCCA (STD) in our table with “STD” denoting that this is the standard implementation of SCCA. In the second method, a fused Lasso penalty is put on the canonical loadings. This model is referred to as SCCA (O) in the tables with “O” denoting ordered. Both methods are encoded in “PMA” package. We use the CCA.optim function to obtain optimal penalties for both the methods after which the CCA function is used to calculate the direction vectors and CCs.

An additional Bayesian method that we do not consider in our set of competitor methods is the Bayesian group factor Analysis with Structured Sparsity (BASS) model proposed by Zhao et al. [2016]. This model has similar goals and a similar modeling framework to the GFA model [Klami et al., 2013], but the authors provide `C++` codes (but not `R`) for their methodology. Hence, we do not utilize that method for our simulation study, although based on the simulations the model constructions, we would anticipate its performance to be similar to GFA.

2.5.3 Simulation Results: Estimation of Canonical Correlations

We begin by comparing the estimation accuracy for the canonical correlation coefficients across the different methods. Table 2.2 compares the performance for estimation of 1st CC. We first note that our model NDFSM performs strongly across most cases. It has the lowest error in cases 1 and 5, and its error is basically equivalent to the best model in cases 3, 4, and 7. We note that in cases 3 and 4, the best performing DFSM model is the true data generating model since $\mathbf{\Phi}^{(m)}$ are taken to be the identity matrix. In the high dimension cases with non-diagonal specificity (settings

Table 2.2: Comparison Between RMSE of Different Methods for Estimation of 1st and 2nd CC

Method	Simulation Setting						
	1	2	3	4	5	6	7
	RMSE for 1st CC Estimation						
NDFSM	**0.0222**	0.2134	0.0220	0.0770	**0.0222**	0.2649	0.0475
DFSM	0.0288	0.0830	**0.0218**	**0.0693**	0.0297	**0.0439**	0.0972
NDFSM+DFSM	0.0222	0.1088	0.0220	0.0770	0.0222	0.0991	0.0475
GFA	0.0371	0.2922	0.0479	0.2169	0.0362	0.2283	0.1867
GHS	0.0339	**0.0596**	0.0464	0.1603	0.0334	0.0543	0.0435
RCCA	0.1082	0.1647	0.0868	0.1794	0.1057	0.1674	0.1203
SCCA (STD)	0.0495	0.1265	0.0499	0.1431	0.0552	0.1254	0.0897
SCCA (O)	0.1154	0.0659	0.1462	0.1178	0.1148	0.1006	**0.0323**
	RMSE for 2nd CC Estimation						
NDFSM	**0.0194**	0.1115	0.0864	0.2318	**0.0317**	0.2946	**0.0391**
DFSM	0.1851	0.3584	0.0481	0.2067	0.0617	0.0605	0.1855
NDFSM+DFSM	0.0194	0.1563	0.0864	0.2318	0.0317	0.2829	0.0391
GFA	0.2532	**0.1023**	**0.0000**	**0.0002**	0.0956	0.5063	0.2285
GHS	0.3614	0.5939	0.2435	0.4231	0.0348	**0.0430**	0.0435
RCCA	0.4291	0.8657	0.4298	0.8476	0.0842	0.2792	0.5162
SCCA (STD)	0.4446	0.6997	0.4420	0.7124	0.0486	0.1850	0.4229
SCCA(O)	0.1200	0.7570	**0.0000**	0.7055	0.4368	0.1686	0.3971

2 and 6), NDFSM does not perform as well, although DFSM performs well in these cases. The combined strategy substantially reduces the error rate in both of these settings; we will further investigate this effect in section 2.5.4. We also note that the GHS strategy has consistently strong performance in estimating the first CC across all settings, and that the GFA model is consistently outperformed by our models. Among the frequentist approaches, the SCCA methods do fairly well although typically worse than the Bayesian models, and the RCCA tends to be slightly worse than SCCA.

Turning to estimation of the 2nd CC in the lower half of Table 2.2, we see that GFA has strongest performance in settings 2–4 at correctly zeroing out the second CC, although it performs worse in settings 1 and 7 that also have $\rho_2 = 0$. NDFSM has strong estimation of ρ_2 performing among the best in five of the seven settings (not 4 and 6); in these two cases, most methods perform poorly and NDFSM is no worse than the majority. In cases when NDFSM correctly captures ρ_1, it also correctly estimates ρ_2. Conversely, GFA seems to perform better on ρ_2 than ρ_1, and despite the strong performance for the first canonical correlation, GHS has fairly large RMSE for the second. Similarly, the two SCCA have reasonable estimation for ρ_1, but in cases 1–4 with $\rho_2 = 0$, they estimate a much larger CC.

In addition to the RMSE for ρ_l estimation, we also consider the average bias across methods in Table 2.3. The results of bias analysis are generally consistent with the MSE results. We see that our models are generally unbiased for ρ_1, although there is some evidence of bias for ρ_2 (positive when $\rho_2 = 0$ as in case 4 and negative if $\rho_2 \neq 0$ as in case 6). As mentioned above, SCCA is clearly failing to penalize the higher order terms when $\rho_2 = 0$, yielding large estimates of this CC for all STD implementations and the ordered (O) cases when $p > n$.

Table 2.3: Comparison Between Average Bias of Different Methods for Estimation of 1st and 2nd CC

Method	Simulation Settings						
	1	2	3	4	5	6	7
	Average Bias for 1st CC						
NDFSM	0.0052	-0.1009	0.0033	0.0029	0.0054	-0.0281	-0.0060
DFSM	-0.0178	-0.0379	0.0015	0.0026	-0.0189	-0.0029	-0.0593
NDFSM+DFSM	0.0052	-0.0482	0.0033	0.0029	0.0054	-0.0205	-0.0060
GFA	-0.0619	-0.0517	-0.1245	0.0731	-0.0568	-0.3592	-0.2164
GHS	-0.0069	-0.0259	-0.0195	-0.0933	-0.0067	0.0031	-0.0148
RCCA	-0.1059	0.1525	-0.0844	0.1694	-0.1019	0.1556	0.0509
SCCA (STD)	-0.0145	0.0043	-0.0089	0.0470	-0.0239	0.0268	0.0300
SCCA (O)	-0.1030	0.0546	-0.1338	0.0712	-0.0996	0.0496	-0.0040
	Average Bias for 2nd CC						
NDFSM	0.0186	0.1045	0.0851	0.2225	-0.0038	-0.2305	0.0373
DFSM	0.1835	0.3463	0.0446	0.1936	-0.0483	-0.0246	0.1833
NDFSM+DFSM	0.0186	0.1370	0.0851	0.2225	-0.0038	-0.2241	0.0373
GFA	0.1318	0.0318	0.0000	0.0000	-0.1538	-0.2627	0.0201
GHS	0.3612	0.5931	0.2418	0.4141	-0.0047	0.0257	0.3564
RCCA	0.4288	0.8623	0.4296	0.8445	-0.0801	0.2692	0.5035
SCCA (STD)	0.4352	0.6833	0.4405	0.6909	0.0022	0.1038	0.4141
SCCA (O)	0.0525	0.7563	0.0000	0.6611	-0.3523	0.1425	0.3949

Table 2.4: Proportion (%) of Data Sets With Potential Overshrinkage

Setting	1	2	3	4	5	6	7
Proportion with $P(\rho < 0.2) > 0.5$	0	13	0	0	0	2	0

2.5.4 Simulation Results: Diagonal and Non-diagonal Model Selection

In this section we investigate the performance and ther impact of our combined model selection strategy. Recall that our combined strategy switches from the NDFSM model results to the DFSM if there is evidence of overshrinkage based on the posterior probability that $\rho_1 < 0.2$. If this posterior probability exceeds 0.5, then we use the diagonal model results. For the simulations studies, Table 2.4 shows proportion of data sets when this criteria is met and the NDFSM+DFSM strategy bases inference on the more restricted DFSM.

As noted previously, this only appears when we consider cases with more features than observations, representing cases 2 and 6. When it does occur, it is fairly rare, impacting only 13% of cases in setting 2 and 2% of cases in setting 6. When the true generalized specificity matrices are diagonal (case 4), we also do not observe any overshrinking cases. In Table 2.5 we will further investigate estimation error, stratifying by these suspected overshrunk outputs compared to the remaining estimates unaffected by overshrinking.

Table 2.5: RMSE Comparison by Overshrinking Criteria

Method	Setting 2	
	Potentially Overshrunk	Not Overshrunk
N	$N = 13$	$N = 87$
NDFSM	0.5603	0.0758
DFSM	0.1682	0.0626
GFA	0.331	0.2464
GHS	0.0547	0.0596
	Potentially Overshrunk	Not Overshrunk
N	$N = 2$	$N = 98$
NDFSM	0.5735	0.0776
DFSM	0.0093	0.0441
GFA	0.7360	0.5505
GHS	0.0746	0.0538

Viewing the NDFSM rows, there are clear differences in the RMSE between those cases whether the estimates are flagged as overshrunk vs not. To help with comparison we include the DFSM, GFA and GHS results also stratified by the NDFSM overshrinking criteria. Clearly, for the unshrunk cases, NDFSM estimates the first CC with similar accuracy to DFSM and GHS, although it is slightly worse (GFA is consistently poor in these cases). Missestimation in NDFSM is clearly dominated by the poor performance in these 13 and 2 datasets with overshrinking, and our combined strategy will replace these poor estimates with the DFSM estimates that are much more accurate. Consistent with the conclusions from the prior section, this analysis shows that while DFSM may be the best choices for high-dimensional data, when NDFSM combined with an overshrinkage correction produces competitive estimates of the first CC.

2.5.5 Simulation Results: Estimation of First Direction Vector

Table 2.6: Comparison Between RMCE of Different Methods For Estimation of Direction Vectors

Method	Settings						
	1	2	3	4	5	6	7
	RMCE for First Direction Vector of View 1						
NDFSM	**0.0721**	0.3600	0.0966	0.3323	**0.0782**	**0.3399**	**0.1244**
DFSM	0.1953	0.3484	**0.0218**	**0.2703**	0.1976	0.4234	0.5263
NDFSM+DFSM	0.0721	**0.3033**	0.0966	0.2708	0.0782	0.4084	0.1244
GFA	0.4838	0.8467	0.4016	0.7790	0.5233	0.8650	0.8164
GHS	0.2742	0.4646	0.1600	0.4908	0.2869	0.4766	0.4979
RCCA	0.4386	0.8104	0.9983	0.7710	0.4636	0.8296	0.9947
SCCA (STD)	0.3289	0.7819	0.2888	0.8018	0.3552	0.8204	0.4943
SCCA (O)	0.2904	0.8523	0.6602	0.7752	0.5526	0.8283	0.5166
	RMCE for First Direction Vector of View 2						
NDFSM	**0.1004**	0.3323	**0.0940**	**0.2389**	**0.1053**	0.3699	**0.1491**
DFSM	0.1125	**0.2846**	0.1017	0.2529	0.1250	**0.2781**	0.2352
NDFSM+DFSM	0.1004	0.2893	0.0940	0.2389	0.1053	0.4170	0.1491
GFA	0.3596	0.7782	0.3247	0.5625	0.4153	0.7984	0.7261
GHS	0.1683	0.3872	0.1104	0.4274	0.2000	0.4092	0.3117
RCCA	0.3514	0.7586	0.3601	0.7142	0.3846	0.7825	0.6857
SCCA (STD)	0.2478	0.7326	0.3930	0.7423	0.4256	0.7900	0.4221
SCCA (O)	0.2797	0.8013	0.6290	0.7347	0.5501	0.7949	0.4213

When we estimate canonical correlations, we also need to understand in which data direction the correlation is maximized. This is particularly important as it represents the contribution of each variable on the canonical correlation. The root mean cosine error for the direction vectors in view 1 and view 2 associated with the first canonical correlation are summarized in Table 2.6.

This table indicates that all versions of our models—DFSM, NDFSM, and the combined approach—have better performance than the other models. Beyond the relatively minor differences in the CC estimations, these more substantial improvements in the direction estimation make our approach a better alternative when we need to find out the important contributing factors to CC. That is, even in cases when the methods miss-state the magnitude of the relationship between views, NDFSM and DFSM tend to correctly find the combination and weights of features that determine this relationship. GHS and SCCA, which are quite competitive for the estimation of ρ_1, lag behind in this criteria by showing higher values of direction vector RMCEs.

2.5.6 Simulation Results: Significant Variable Loadings

In conjunction with the previous exploration of the accuracy of the canonical loading vectors $\hat{\mathbf{u}}^*$ and $\hat{\mathbf{v}}^*$, we also want to interrogate whether we are able to correctly detect whether a variable is significantly loaded or not. Recall that for Bayesian variable selection we consider a variable to be significantly associated with the CC direction if the 50% credible interval for its factor loading excludes zero. In the penalized methods, SCCA (O) and SCCA (STD), if the component in the estimated (sparse) CC direction is non-zero then we say that the corresponding variable is significantly contributing to the CC calculation.

To characterize the true effect of each variable, we divide the elements of true data-generating direction vectors into 3 groups according to their contribution to the calculation of CC and direction vectors. Features are considered relevant in both the latent and CCA structure, if they have non-zero value in the column of the factor loading matrix $\mathbf{A}^{(m)}$ producing the first canonical correlation. In our data generating approach, there are 5 such features. As the direction vectors are the complex function of $\mathbf{\Sigma}$, it is difficult to understand the direct impact of the AR structure of the generalized specificity on the estimands. We define features that are relevant in the CCA structure if their true loading value is greater than 0.1 in absolute value even as the latent factor $\mathbf{z}$ is not associated with the views; we pick 0.1 because that means their square contribution would be greater than 0.01 (or 1%) of the total direction vector. Recall that variables unrelated to the latent factor structure can be loaded on the canonical correlation if they are highly associated through $\mathbf{\Phi}^{(m)}$ with another variable with non-zero projection value. There are 9 such features in the settings with an AR structure for $\mathbf{\Phi}^{(m)}$, and no such features in settings 3 and 4. If the absolute factor loading is less than 0.1, then it is practically irrelevant, so it goes in the third block of features unrelated to the CC. There are 136 elements in the AR settings and 145 in the independence settings. The better performing methods are those which have high selection rates in block 1 (and to a lesser extent, block 2) and low selection rates in block 3. Due to the low effect sizes of the factor loading in block 2, we do expect lower variable selection rates than in block 1.

Table 2.7 summarizes the results. We can see that in block 1 our methods all consistently recover the true features involved in the latent structure; DFSM case 6 is the only one of our methods with less then 80% accuracy. While GFA also performs well in block 1, GHS selects these most critical variables less than 50% of the time across all methods. This is surprising given the accurate recovery of ρ_1, although it is consistent with the direction vector accuracy results. The penalized approaches perform decently in low dimensional setting but do not perform up to the mark in high dimensional setups.

In block 2, NDFSM continues to have almost perfect recovery in the lower-dimension ($n = 300$) settings, with a substantial drop off when $n = 50$. As would be anticipated, DFSM does worse in this block as these are the features whose role is governed by the non-diagonal specificity matrix. GFA has the best selection in this block when $n > p$ but worse than NDFSM when $n < p$; GHS has low selection rate as in block 1. Here penalized methods are unable to identify features which are result of AR correlations.

In block 3 we expect that selection rates should be low for all models, but GFA shows a very high false positives rate in this case. Our FSM models seem to control the rate of false discoveries, and consistent with their low power throughout, GHS has low selection rates. Penalized methods performed the best in this block as they did not pick any features which are not relevant to the CC structure.

Table 2.7: Percentage Accuracy of Significant Variable Loading

Methods	1	2	3	4	5	6	7
	Features Relevant to Latent and CCA Structure						
NDFSM	100.00	88.80	40.00	98.40	100.00	88.20	100.00
DFSM	100.00	97.80	100.00	98.40	100.00	56.40	96.20
GHS	41.80	39.80	40.20	34.20	41.40	35.60	47.60
GFA	100.00	99.80	100.00	100.00	100.00	73.00	98.00
SCCA (STD)	79.20	45.00	99.40	56.40	98.20	46.80	72.00
SCCA (O)	79.20	45.00	99.40	56.40	98.20	46.80	72.00
	Features Relevant to CCA Structure						
NDFSM	99.80	18.40	-	-	100.00	28.40	100.00
DFSM	38.40	3.60	-	-	33.20	2.00	39.60
GHS	14.63	2.13	-	-	14.75	1.63	21.88
GFA	44.60	61.20	-	-	39.20	66.20	81.40
SCCA (STD)	0.00	0.80	-	-	0.00	0.20	0.20
SCCA (O)	0.00	0.75	-	-	0.00	0.62	0.12
	Non Relevant Features						
NDFSM	6.17	1.55	3.30	10.72	6.37	4.07	6.59
DFSM	3.78	2.96	2.11	6.07	4.58	2.21	4.46
GHS	3.73	0.58	0.03	0.08	4.04	0.88	4.32
GFA	43.69	58.21	46.73	57.41	37.42	63.70	80.84
SCCA (STD)	0.00	0.66	0.00	0.45	0.00	0.60	0.19
SCCA (O)	0.00	0.01	0.00	0.00	0.00	0.01	0.00

In conclusion our models, NDFSM, DFSM and the combined strategy, perform similarly or slightly better than most of the competing models for the estimation of first canonical correlation. However, our models truly show a substantial benefit when considering the estimation of the direction vectors and the selection of the features that are significantly loaded. NDFSM consistently performs well if $n > p$, and when $p > n$, its occasional overshrinkage can be mitigated by using the combined strategy that switches to inference under the diagonal $\boldsymbol{\Phi}^{(m)}$. DFSM surprisingly beats GFA even though models are similar. GHS, even though it is not designed for CCA, performs well in estimation of CCs but under-performs in estimation and variable selection of the direction vectors.

2.6 Data Analysis

Advances in technology have enabled users to collect a vast amount of high quality genetic data, and the integration and joint analysis of multiple types of "-omics" data is an important part of modern biomedical research [Morris and Baladandayuthapani, 2017, Manzoni et al., 2018, Castleberry, 2019]. In particular, studies of multiple data sources on a common set of samples has received widespread attention in genomics, with many authors jointly considering gene expression and copy number variation [Hyman et al., 2002, Pollack et al., 2002]. CCA is a common and effective tool for such analysis.

Breast Cancer (BC) is one of the most widely diagnosed type of cancer. It is the fifth greatest cause of cancer-related deaths with an estimated 2.26 million new cases worldwide [Sung et al., 2021] and is a leading cause of cancer deaths among women worldwide [Ferlay et al., 2020]. The incidence rate of BC varies within race and ethnicity and is affected by several epidemiological risk factors such as demographics, reproductive history, family history, and lifestyle factors [Momenimovahed and Salehiniya, 2019]. In addition to the epidemiological components, further investigation into the genetic factors of BC is an ongoing area of research.

We apply our method to the breast cancer data described in Chin et al. [2006] and available to download from https://tibshirani.su.domains/PMA/. There are $n = 89$ samples/observation on which DNA and RNA data are available. For the view 1 data, we consider the matrix of DNA copy numbers (DNA) for genes located on the 1^{st} chromosome, yielding $p^{(1)} = 136$. The data source also contains genetic expression levels (RNA) for $19,672$ genes, and we select $p^{(2)} = 250$ genes for the view 2 data by selecting the top 50 genes associated with chromosome 1 with the greatest interquartile range. Also we select an additional 200 genes across the other 22 chromosome sites which have highest interquartile range, yielding $p^{(2)} = 250$. We standardize both data views. Among these 250 genes, the 50 genes which are located on the first chromosome are anticipated to be most associated with the view 1 copy number data, some of these are located on the same chromosome. As SCCA (O) utilizes fused lasso that are dependent on the data ordering, we order the features according to their chromosomal location for the copy numbers in view 1 and according to their nucleotide position (within chromosome) for the RNA expressions in view 2.

Using these data views, we apply the same set of methods as in the simulation study. For our method we run MCMC for 100,000 iterations with 25,000 iterations as burn-in iterations, and the sample is thinned to save 5000 samples. When we ran the proposed NDFSM, we encountered the potential overshrinkage scenario discussed previously. Based on the results from the NDFSM MCMC output, we have $P(\rho_1 < 0.2) = 0.7882$. Hence, we follow the discussed strategy and consider inference based on the MCMC output form the DFSM model. We note that these results do not indicate overshirnkage as under this model we have $P(\rho_1 < 0.2) = 0$. The 5000 stored posterior samples provide an effective sample size of 1611 approximately independent posterior samples for ρ. For GFA, we ran the MCMC for 600,000 iterations and obtained a thinned sample of size 10,000 to obtain effective sample size of 4471.

The estimated first and second canonical correlations are shown in Table 2.8.

Clearly, there are substantial differences in the results across the different methods. GFA produces estimates for these two correlations that are almost unity. In fact, it estimates 10 CCs that exceed 0.95, representing a much greater dependence than is represented from the other methods. In contrast, the SCCA approaches have substantially lower estimates between 0.5 and 0.6. GHS and our proposed method yield similar estimates with $\hat{\rho}_1 \approx 0.92$ and $\hat{\rho}_2 \approx 0.90$.

Table 2.8: Canonical Correlation and Variable Selection Analysis for Breast Cancer Data

Method	NDFSM +DFSM	GHS	GFA	SCCA (O)	SCCA (STD)
Estimate of First CC	0.9208	0.9362	0.9773	0.5407	0.6092
Estimate of Second CC	0.9026	0.9090	0.9736	0.5181	0.5902
Sig. Copy Number Loading	60	131	1	59	20
Sig. Gene Loadings	10	0	1	65	50
Sig. Gene Loadings on chromosome 1	4	0	1	10	7
Weight (%) of view 2 direction on chromosome 1	43.86	28.99	23.36	30.52	62.78

We also investigate the behavior of the estimated direction vectors and the number of significant loadings found in this analysis in Table 2.8. As compared to other Bayesian methods, NDFMS+DFSM combination identifies a greater number of significant genes (view 2 components) than GHS and GFA; GHS does not identify any significant genes, and GFA finds only one significant gene. In contrast to GHS failing to find significant loadings in view 2, it selects 131 significant copy numbers from view 1, representing 96% of the view 1 components as being selected as significantly associated with the CC. This is clearly an unreasonable result. GFA selects a single significant copy number with its single significant gene, which is clearly fewer associations than one would expect from this context.

As we expect the 50 genes located on chromosome 1 to be the most active features in determining the correlation to the chromosome 1 copy numbers, we investigate the sum of the square weights (both significant and non-significant) in the view 2 direction vector for the chromosome 1 genes. As the direction vector has norm 1, this sum can be viewed as a percentage weight. In Table 2.8 we can see that among Bayesian methods, GHS and GFA assigns 29% and 23% respectively. While our model selects 4 of the chromosome 1 genes as significant, it assigns 44% of the direction vector to genes on the expected chromosome.

The frequentist method SCCA (O), which utilizes the chromosomal locations and nucleotide positions in the data, selects 65 significant genes for the first CC direction and 10 of these genes are located on chromosome 1. That is, 15% of significant genes are from chromosome 1 compared to 40% (4 of 10) from NDFSM+DFSM. The direction vector under SCCA (O) also assigns a lower weight to the chromosome 1 genes than does NDFSM+DFSM (31% vs 44%). Using the standard implementation of SCCA without the position information, 20 copy numbers are selected (fewer than NDFSM+DFSM and SCCA (O)) and 50 genes are selected (between the two methods) as significantly associated with the first CC. While only 7 of these 50 genes are located on chromosome 1, these genes have very large loadings and account for 63% of the direction vector.

In this data, no method can be understood as the absolute truth. However, under our proposed model the estimated canonical correlations are of a magnitude that might be expected given the biological relationship between the copy numbers and gene expression from a common chromosome. Further, our investigation of the gene expression direction vector indicates the our model places a larger weight on the chromosome 1 genes which are expected to play the largest role, when viewed as a proportion of significant genes and as a proportion of the overall weight, than most completing methods.

2.7 Conclusions

We noted the lack of structured sparse CCA methodology from a Bayesian perspective. The rise of interconnected, high dimensional, sparse data, measured on a small number of samples, demands the development of new statistical theory, and the ability to construct flexible and sparse models make the Bayesian approach a valuable contribution to this work. Our model is one of the few models which provides Bayesian modeling of within view covariance matrix with sparse CCA. To the best of our knowledge ours is the only model which tries to model a sparse structure for the generalized specificity $\mathbf{\Phi}^{(m)}$ without restricting it to be diagonal.

To that end, we apply a graphical horseshoe prior to bring in sparsity in these high dimensional covaraince matrices. As shown in the simulations, this often performs better than GFA and DFSM which both assume diagonal matrices. However, improved performance is not universal, and we do see cases where the NDFSM overshrinks the cross-covariance terms. The NDFSM+DFSM approach, which combines the GHS and diagonal versions of our models with an ad hoc selection rule, proved to be a competitive approach to our models. As this overshrinkage tends to only appear when $p > n$, one might also make the initial decision to only consider the DFSM choice when the sample size is low.

As the CCA problem is generally focused on the case where one is interested in two views of the data, this approach can easily be extended to a multi-view setting along the lines of Zhao et al. [2016]. The IBFA model in (2.10) is easily extended by including additional m with the associated data view $\mathbf{X}^{(m)}$ and the parameters $\boldsymbol{\mu}^{(m)}$, $\mathbf{A}^{(m)}$, and $\mathbf{\Phi}^{(m)}$. Canonical correlations and direction vectors can be estimated for each pair (m, m') of data views based on the common set of MCMC output as discussed throughout the manuscript.

As part of our model specification, we have introduced a new flexible shrinkage prior in the form of a multiplicative half-Cauchy process. Along the lines of the multiplicative shrinkage processes [Bhattacharya and Dunson, 2011, Schiavon et al., 2021], this model flexibly imposes shrinkage on the projection matrix coefficients, while increasingly reducing the roles of each subsequent factor. It would certainly be of interest to further investigate the mathematical and theoretical features of this prior process. However, this can be achieved by considering its role within the context of a single view factor model instead of our IBFA. The additional layer of multiple views and the non-diagonal residual variance complicates the derivations, relative to the exploration in a standard factor model. We have run additional simulations (not shown) comparing our multiplicative half-Cauchy process to the multiplicative gamma process of Bhattacharya and Dunson [2011] and have found our approach to perform comparably and in some cases better.

One of the main challenges of this model, and most Bayesian approaches to CCA, is computational scalability. MCMC Gibbs sampling samples each parameter of the model within each iteration and can be fairly slow in high dimensions. Due to its nature as a completely Bayesian model which scans through all the parameters, the model is computationally very intensive. Approximate Bayes algorithms such as variational Bayes which seeks to find a posterior mode of the parameter distribution has been used in some previous work such as Klami et al. [2013] and Zhao et al. [2016]. While this could be an approach to find a set of parameter estimates more quickly, these estimation algorithms typically fail to appropriately account for uncertainty quantification [Wang and Blei, 2019]. Additionally, the parameters of interest for CCA $(\rho, \mathbf{u}^*, \mathbf{v}^*)$ are complex functionals of the model parameters $(\mathbf{A}^{(m)}, \mathbf{\Phi}^{(m)})$, and so the impact of the approximation to the posterior of $(\mathbf{A}^{(m)}, \mathbf{\Phi}^{(m)})$ relative to the posterior of the CCA parameters may be unclear.

CHAPTER 3

BAYESIAN ANALYSIS OF FINITE MIXTURE MODEL FOR SPHERICAL DATA

3.1 Introduction

Advancements in technology have given rise to directional datasets where observations are recorded as directions or angles relative to a system with fixed orientation [Wang and Gelfand, 2013]. Some examples of such data lie on circumference of unit circle ($\mathbb{R}^2$) or on the unit hypersphere $\mathbb{S}^{p-1} = \{\boldsymbol{y} \in \mathbb{R}^p : \|\boldsymbol{y}\|_2 = 1\}$, where $\|\boldsymbol{y}\|_2 = \sqrt{\boldsymbol{y}^T\boldsymbol{y}}$. For example, in Diffusion Tensor Imaging data consist of the maximum diffusivity directions of water molecules. The direction of the flow of water differs across parts of the human brain due to differences in the properties of the brain tissues. Hence, this data provides an image of the structure of the brain leading to increasing understanding of brain connectivity. In this way directional data has wide presence in the field of bioinformatics [Mardia et al., 2018], astronomy [Marinucci and Peccati, 2011], medicine [Pardo et al., 2016], neurology [Kaufman et al., 2005], genetics [Dortet-Bernadet and Wicker, 2008], image analysis [Esteves et al., 2018], text mining [Banerjee et al., 2005], machine learning [Sra, 2018] and many others [Pewsey and García-Portugués, 2021]. Mardia and Jupp [2000] and Ley and Verdebout [2017] provide a rich literature review on the presence on the directional statistics in these areas.

In this article, we will focus on one of the most popular and pivotal distribution in directional data, the von Mises Fisher (vMF) distribution. This distribution sometimes has different names when applied to hyperspheres of different dimensions. The vMF distribution for circular data is often known as the von Mises distribution, and in the case of spherical data, it may be called the Fisher distribution [Nunez-Antonio and Gutiérrez-Pena, 2005]. The distribution contains two parameters: the mean direction vector $\boldsymbol{\mu}$ and the concentration κ. A detailed description of the distribution is given in the subsequent sections. In our study we will mainly consider data that is multivariate over the sphere ($p = 3$).

Mardia and El-Atoum [1976b] proposed an approach for formal Bayesian inference for vMF distribution. They assumed a known κ and discussed inference for the mean vector under the one sample and two sample problems. Guttorp and Lockhart [1988] developed Bayesian methodology for considering both parameters as unknown. Damien and Walker [1999] introduced an auxiliary variable approach in development of Gibbs sampling. This algorithm provided an approach which tries to eliminate restrictive assumptions on the prior distributions, which had been present in some previous studies. Due to the high levels of autocorrelation among posterior samples, Nunez-Antonio and Gutiérrez-Pena [2005] argued that this auxiliary variable approach is not efficient for relatively large values of concentration parameter. To solve this problem the authors proposed an importance-resampling algorithm. In the first part of this project, we propose a novel Bayesian joint prior on the parameters $\boldsymbol{\mu}$ and κ of vMF. We provide a theoretical justification and investigation of this new distribution.

The problem of clustering data on hyper-sphere has received increased attention recently. As noted in Qin et al. [2016] and Figueiredo [2017], these methods can be roughly divided by whether they utilize Euclidean or spherical geometry principles. The Euclidean geometry based algorithms do not take into account the geometric properties of the sample data and instead utilize Euclidean distance as a measure of similarity [Qin et al., 2016]. One of the popular example of such a model is the K means clustering algorithm [Hartigan and Wong, 1979]. K means is a similarity based algorithm which does not require any assumption relating to an underlying probability model. Probabilistic mixture models fit the mixture of probability distribution to the data. Here conditional probabilities of data points are used to assign the labels to the data [He et al., 2010]. One of the popular models in this area is the Gaussian Mixture Model [Bishop and Nasrabadi, 2006]. This and variants of this model are not effective for data constrained to have norm one due to the methods' tendency to place non-trivial probability mass off the unit sphere domain [Gopal and Yang, 2014].

In contrast, spherical geometry based models use cosine similarity measures to leverage the inherent directional geometry of the data. One of the most common methods in this domain is spherical K means clustering, proposed by Dhillon and Modha [2001], which utilizes cosine distances cluster data on hypersphere. Peel et al. [2001] used mixture of Kent distributions to cluster data on rock mass. Banerjee et al. [2005] developed the vMF Mixture Model (vMFMM). This assumes each cluster in the mixture model follows a vMF distribution. Additional examples of spherical clustering models are the Spherical Topical Model [Reisinger et al., 2010], Dirichlet process vMFMM [Bangert et al., 2010], and temporal VMF mixture model [Gopal and Yang, 2014].

In this article we address a Bayesian vMFMM. Both frequentist and Bayesian clustering models need an efficient algorithm to estimate their parameters. In the case of a frequentist mixture of vMF distributions, the most popular algorithms to estimate parameters are variants of the EM algorithm [Figueiredo, 2017]. In the case of Bayesian methods, different algorithms have been proposed such as variational inference, reversible jump MCMC, collapsed Gibbs sampling, among many others [Reisinger et al., 2010, Gopal and Yang, 2014, Qin et al., 2016]. These algorithms require a numerical approximation to the intractable Bessel function in the vMF distribution.

In the second part of our project we focus on developing an efficient sampling scheme for the Bayesian-vMF mixture model and investigating its properties. We propose a novel Data Augmentation (DA) algorithm which removes the intractability from the sampling distribution of concentration parameter κ.

We organize the article as follows. In Section 3.2 we introduce the von Mises Fisher distribution. After this mathematical background, in Section 3.3 we introduce a novel conjugate prior and investigate properties of this distribution. Section 3.4 introduces Bayesian mixture of vMF distribution and inference under our conjugate prior. In Section 3.5 we talk about a novel data augmentation technique based on an alternative commonly used prior choice. We perform simulation analysis in Section 3.7 and data analysis in Section 3.8.

3.2 Von Mises Fisher Distribution

We consider random variable $\boldsymbol{y}$ with support on the p-dimensional hypersphere $\mathbb{S}^{p-1}$, where $\mathbb{S}^{p-1} = \{\boldsymbol{y} \in \mathbb{R}^p : \|\boldsymbol{y}\|_2 = 1\}$ and $\|\boldsymbol{y}\|_2 = \sqrt{\boldsymbol{y}^T\boldsymbol{y}}$. Then, $\boldsymbol{y} \in \mathbb{S}^{p-1}$ is said to follow p-variate vMF distribution if its probability density function is given by

$$f(\boldsymbol{y} \mid \boldsymbol{\mu}, \kappa) = \frac{\kappa^{p/2-1}}{(2\pi)^{p/2} I_{p/2-1}(\kappa)} \exp\left(\kappa \boldsymbol{\mu}^T \mathbf{y}\right) \mathbb{I}(\boldsymbol{y} \in \mathbb{S}^{p-1}), \tag{3.15}$$

where $\kappa \geq 0, \boldsymbol{\mu} \in \mathbb{S}^{p-1}$, and $p \geq 2$. $I_{p/2-1}(\cdot)$ denotes modified Bessel function of the first kind with order $\frac{p}{2} - 1$. The parameter $\boldsymbol{\mu}$ is the mode of the distribution. We note that the mean $E(\boldsymbol{y})$ does not lie on sphere as it has norm less than one ($\|E(\boldsymbol{y})\| < 1$). However, renormalization of $E(\boldsymbol{y})$ does coincide with $\boldsymbol{\mu} = (\|E(\boldsymbol{y})\|)^{-1}E(\boldsymbol{y})$. Hence, when we refer to $\boldsymbol{\mu}$ as the mean of the vMF, we are referring to this parameter as representing the mean direction of the random variable. The concentration parameter κ quantifies how tightly the function is distributed around its mean direction $\boldsymbol{\mu}$. For $\kappa = 0$ the distribution is uniform over the sphere. For $\kappa > 0$ the distribution is unimodal and rotationally symmetric around the direction $\boldsymbol{\mu}$. $\boldsymbol{\mu}^T\mathbf{y}$ is the cosine similarity between $\mathbf{y}$ and $\boldsymbol{\mu}$, and we note that the density depends on $\boldsymbol{y}$ only through this cosine similarity. The computationally difficult element of the distribution is Bessel function of first order, which is given by

$$I_\nu(\kappa) := \sum_{k=0}^{\infty} \frac{(-1)^\kappa}{k!\ \Gamma(\nu+k+1)} \left(\frac{\kappa}{2}\right)^{(\nu+2k)}.$$

Here $\Gamma(\cdot)$ is the gamma function. The infinite, alternating series is a computationally intractable function of κ, and it complicates the analysis whenever the concentration parameter is treated as unknown.

3.3 A Novel Class of Conjugate Priors for VMF

A conjugate prior provides a remarkable advantage in Bayesian inference as it provides a closed form representation of the posterior distribution. These attractive properties have propelled the use conjugate priors in most applications of Bayesian statistics in a wide variety of fields. One of the benefits to using the conjugate prior framework is that there is a somewhat constructive approach to finding such a prior. For a generic distribution from an exponential family $f(x \mid \theta) = h(x)e^{\theta x - \delta(\theta)}$, a conjugate family of distributions should have the form $\pi(\theta \mid \mu, \lambda) = K(\mu, \lambda)e^{\theta\mu - \lambda\delta(\theta)}$, depending on hyperparameters μ and λ. The resulting posterior under a random sample of size n will be $\pi(\theta \mid \mu + \sum_{i=1}^{n} x_i, \lambda + n)$ [Robert, 2007, Section 3.3.4].

There has been some prior work considering conjugate priors for the vMF distribution [Mardia and El-Atoum, 1976a, Nunez-Antonio and Gutiérrez-Pena, 2005, Bangert et al., 2010]. As noted by Hornik and Grün [2013], these choices have various merits and demerits. As an alternative, Hornik and Grün [2013] construct a conjugate prior based on a Theorem 1 of Diaconis and Ylvisaker [1979] and consider a set of necessary and sufficient condition on the hyperparameters to make the proposed conjugate family proper. However, the construction of their prior is in terms of the natural parameterization $\boldsymbol{\theta}^* = \kappa\boldsymbol{\mu} \in \mathbb{R}^p$, not the usual interpretable parameters

$(\boldsymbol{\mu}, \kappa)$. Consequently, their prior hyperparameters lack a natural interpretation [Pal et al., 2020]. Further, Hornik and Grün [2013] only provides the properties of the the proposed conjugate priors and does not provide an estimation approach or sampling scheme to perform inference.

Considering these gaps in the literature, we propose a novel conjugate prior for $(\boldsymbol{\mu}, \kappa)$ parameterization. Further, we study its various properties and provide an interpretation for its hyperparameters. We define the novel conjugate prior as follows.

Definition 1. *The Conjugate von Mises Fisher (CvMF) distribution is the joint conjugate prior on parameters $\boldsymbol{\mu}$ and κ for vMF distribution and has density is proportional to*

$$g(\boldsymbol{\mu}, \kappa \mid \boldsymbol{\psi}, \lambda) = \left[\frac{\kappa^{\nu} \ \exp\left(\kappa \ \boldsymbol{\mu}^T \boldsymbol{\psi}\right)}{I_{\nu}(\kappa)} \right]^{\lambda} \mathbb{I}(\kappa > 0, \boldsymbol{\mu} \in \mathbb{S}^{p-1}),$$

as long as $g(\boldsymbol{\mu}, \kappa \mid \boldsymbol{\psi}, \lambda)$ is integrable with respect to the measure $d\boldsymbol{\mu}\ d\kappa$ defined on $\mathbb{S}^{p-1} \times \mathbb{R}_+$. Here $\lambda > 0$ and $\boldsymbol{\psi} \in \mathbb{R}^p$.

In Section 3.3.2 we will show that the posterior based on this prior belongs to the same distribution family, proving CvMF to be a conjugate prior. First, we establish the finiteness conditions for the kernel by establishing conditions on $\|\boldsymbol{\psi}\|$ and λ.

Theorem 1. *Consider the function*

$$g(\boldsymbol{\mu}, \kappa \mid \boldsymbol{\psi}, \lambda) = \left[\frac{\kappa^{\nu} \ \exp\left(\kappa \ \boldsymbol{\mu}^T \boldsymbol{\psi}\right)}{I_{\nu}(\kappa)} \right]^{\lambda} \mathbb{I}(\kappa > 0, \boldsymbol{\mu} \in \mathbb{S}^{p-1}),$$

for $\boldsymbol{\psi} \in \mathbb{R}^d$, $\lambda > 0$. The following statements hold regarding the integrability of $g(\mu, \kappa \mid \boldsymbol{\psi}, \lambda)$ with respect to the measure $d\boldsymbol{\mu}\ d\kappa$.

1. *If $\|\boldsymbol{\psi}\| < 1$, then*

$$\int_{\mathbb{R}_+} \int_{\mathbb{S}^{p-1}} g(\boldsymbol{\mu}, \kappa \mid \boldsymbol{\psi}, \lambda)\ d\boldsymbol{\mu}\ d\kappa < \infty.$$

2. *If* $\|\boldsymbol{\psi}\| = 1$*, then*

$$\int_{\mathbb{R}_+} \int_{\mathbb{S}^{p-1}} g(\boldsymbol{\mu}, \kappa \mid \boldsymbol{\psi}, \lambda)\ d\boldsymbol{\mu}\ d\kappa = \infty.$$

3. *If* $\|\boldsymbol{\psi}\| > 1$*, then*

$$\int_{\mathbb{R}_+} \int_{\mathbb{S}^{p-1}} g(\boldsymbol{\mu}, \kappa \mid \boldsymbol{\psi}, \lambda)\ d\boldsymbol{\mu}\ d\kappa < \infty,$$

if and only if $\nu > 0.5$ *and* $\lambda < \frac{2\nu-1}{2\nu+1}$.

The proof of this Theorem can be found in Section C.

As we focus our attention on the case with $p = 3$ ($\nu = 0.5$) for the hypersphere, this theorem implies that we must consider $\boldsymbol{\psi} \in \mathbb{R}^d$ such that $\|\boldsymbol{\psi}\| < 1$ to ensure that this distribution is proper. After establishing the finitness property of the kernel, we further proceed to establish interpretations of the hyperparameters. We start with establishing a property related to the modality of the conjugate prior. The following theorem establishes conditions on unimodality of the distribution.

Theorem 2. *Let* $\kappa > 0$*,* $\boldsymbol{\mu} \in \mathbb{S}^{p-1}$*. Consider the distribution CvMF with probability density function proportional to*

$$g(\boldsymbol{\mu}, \kappa \mid \boldsymbol{\psi}, \lambda) = \left[\frac{\kappa^{\nu}\ \exp\left(\kappa\ \boldsymbol{\mu}^T \boldsymbol{\psi}\right)}{I_{\nu}(\kappa)} \right]^{\lambda} \mathbb{I}(\kappa > 0, \boldsymbol{\mu} \in \mathbb{S}^{p-1}),$$

for $\nu = \frac{p}{2} - 1$, $\boldsymbol{\psi} \in \mathbb{R}^d$. *We assume that* $0 < \|\boldsymbol{\psi}\| < 1$ *and* $\lambda > 0$ *such that the density is finite, then there exists an unique mode* $(\hat{\boldsymbol{\mu}}, \hat{\kappa})$ *located at* $\hat{\boldsymbol{\mu}} = \frac{\boldsymbol{\psi}}{\|\boldsymbol{\psi}\|}$ *and* $\hat{\kappa} = R_\nu^{-1}(\|\boldsymbol{\psi}\|)$.

In the above, $R_\nu(x) = \frac{I_{\nu+1}(x)}{I_\nu(x)}$ is the ratio of Bessel functions, and $R_\nu^{-1}(\cdot) : (0, 1) \to \mathbb{R}_+$ is its inverse function. Lemma 1 verifies the existance of this inverse function, and the proof of the Theorem can be found in Section C. In particular, we note that mode of κ is a function of $\boldsymbol{\psi}$ through $\|\boldsymbol{\psi}\|$. This implies that only the parameter $\boldsymbol{\psi}$ determines the mode of this distribution, not the λ hyperparameter.

Having established the unimodality of the distribution, we now characterize the concentration behaviour of the distribution through the concept of the level sets. Let $g(\boldsymbol{\mu}, \kappa \mid \boldsymbol{\psi}, \lambda)$ be the unnormalized probability density kernel as given in Definition 1, and this density achieves its maximum at the unique $(\hat{\boldsymbol{\mu}}, \hat{\kappa})$ as per Theorem 2. Let the level set of level $l \in (0, 1)$ be given as

$$S_l = \left\{ (\boldsymbol{\mu}, \kappa) \in \mathbb{S}^{p-1} \times \mathbb{R}_+ : \frac{g(\boldsymbol{\mu}, \kappa \mid \boldsymbol{\psi}, 1)}{g(\hat{\boldsymbol{\mu}}, \hat{\kappa} \mid \boldsymbol{\psi}, 1)} > l \right\}. \tag{3.16}$$

This is a level set containing the mode $(\hat{\boldsymbol{\mu}}, \hat{\kappa})$ for all l. Note that in (3.16) we fix λ=1.

Let $P_{\boldsymbol{\psi},\lambda}(\cdot)$ denotes the probability distribution corresponding to the kernel $g(\boldsymbol{\mu}, \kappa \mid \boldsymbol{\psi}, \lambda)$. Assuming the conditions of Theorem 2 are met, the distribution $P_{\boldsymbol{\psi},\lambda}(\cdot)$ will have its mode at $(\hat{\boldsymbol{\mu}}, \hat{\kappa})$. To assess the properties of the distribution of mass in $P_{\boldsymbol{\psi},\lambda}(\cdot)$ around the mode, we provide Theorem 3 which characterizes the changes in the probability distribution as λ changes.

Theorem 3. *Let* $\boldsymbol{\psi} \in \mathbb{R}^p$ *such that* $0 < \|\boldsymbol{\psi}\| < 1$ *and* $\lambda > 0$. *Let* $P_{\boldsymbol{\psi},\lambda}(\cdot)$ *denote the probability measure corresponding to the kernel*

$$g(\boldsymbol{\mu}, \kappa \mid \boldsymbol{\psi}, \lambda) = \left[\frac{\kappa^\nu \ \exp\left(\kappa \, \boldsymbol{\mu}^T \boldsymbol{\psi}\right)}{I_\nu(\kappa)} \right]^\lambda \mathbb{I}(\kappa > 0, \boldsymbol{\mu} \in \mathbb{S}^{p-1}),$$

for $\nu = \frac{p}{2} - 1$.

1. *If S_l denotes the l^{th} level set for some $l \in (0, 1)$, then $P_{\boldsymbol{\psi},\lambda}(S_l)$ is an increasing function in $\lambda > 0$.*

2. *Let A be any set containing the mode $(\hat{\boldsymbol{\mu}}, \hat{\kappa})$, then*

$$\lim_{\lambda \to \infty} P_{\boldsymbol{\psi},\lambda}(A) = 1.$$

The proof of this Theorem can be found in Section C.

Importantly, Theorem 3 implies that the parameter λ controls the concentration of mass around the mode. As λ increases, more mass is concentrated around mode.

From the above discussion we can see here that the conjugate distribution is parameterized by two parameters. Just like the vMF distribution has two parameters which for direction and spread, in our conjugate prior the parameter $\boldsymbol{\psi}$ controls the mode of $(\boldsymbol{\mu}, \kappa)$, while λ characterizes the probability concentration around the mode. This justifies the following parameter names.

Definition 2. *In the context of the probability distribution $CvMF(\cdot; \boldsymbol{\psi}, \lambda)$, the parameters λ and $\boldsymbol{\psi}$ are labeled as the concentration parameter and modal parameter, respectively.*

Figure 3.1 displays a contour plot of κ for λ parameters from 5 to 55. with $\nu = 0.5$, $\mu = (1, 0, 0), \boldsymbol{\psi} = (0.25, 0.25, 0.25)$. It can be observed that the modal point of the distribution remains same while the spread changes according to the value of λ. Further, it can be observed that as value of λ increases, the spread increases affirming the results proved in Theorem 3.

3.3.1 Properties of Marginal Density of κ

In this section we will investigate the properties of marginal density of κ under the CvMF distribution after marginalizing over the mean vector $\boldsymbol{\mu}$. This is useful to

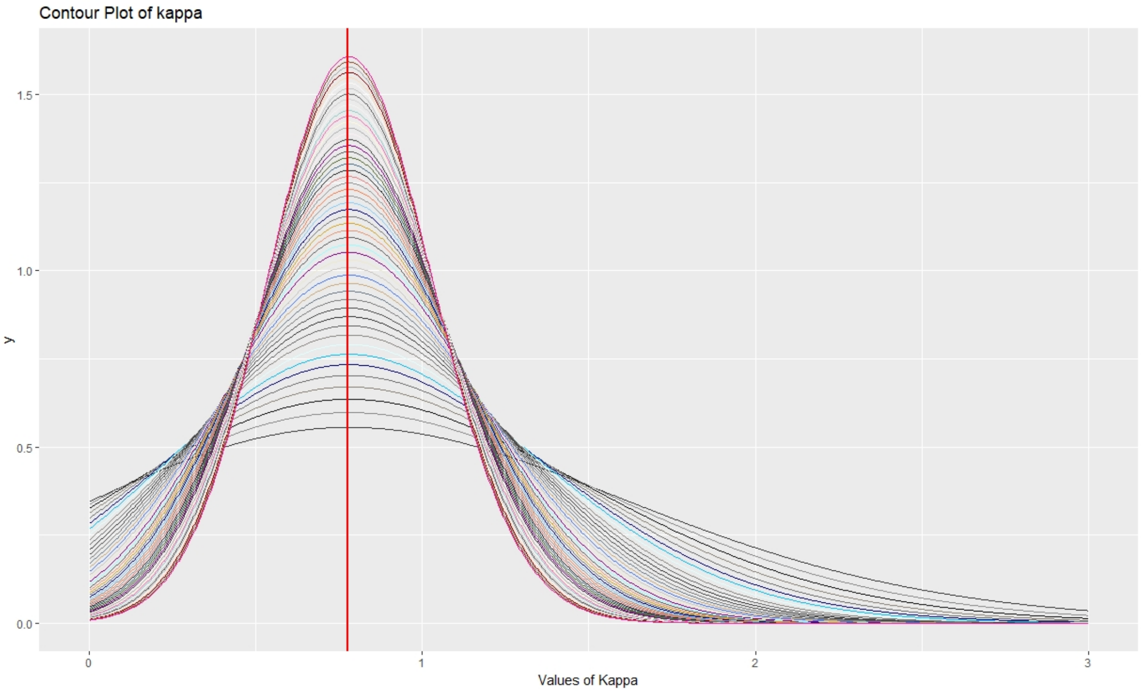


Figure 3.1: Contour Plot of $\pi(\mu, \kappa; \boldsymbol{\psi}, \lambda)$ as a function of κ for different values of λ. Other parameters are set at $\nu = 0.5$, $\mu = (1, 0, 0), \boldsymbol{\psi} = (0.25, 0.25, 0.25)$. The mode of the distributions are located at the point $\hat{\kappa} = 0.77$.

derive efficient sampling scheme for the parameter κ which has intractable marginal density. The marginal density of κ can be calculated (up to a normalizing constant) from $g(\boldsymbol{\mu}, \kappa \,|, \lambda)$ through

$$\begin{aligned}
\pi(\kappa \mid \boldsymbol{\psi}, \lambda) \;&\propto\; \int_{\mathbb{S}^{p-1}} g(\boldsymbol{\mu}, \kappa \mid \boldsymbol{\psi}, \lambda) d\boldsymbol{\mu} \;=\; \int_{\mathbb{S}^{p-1}} \left[\frac{\kappa^{\nu}\ \exp\left(\kappa\ \boldsymbol{\mu}^T \boldsymbol{\psi}\right)}{I_\nu(\kappa)} \right]^{\lambda} d\boldsymbol{\mu} \\
&= \frac{\kappa^{\nu\lambda}}{[I_\nu(\kappa)]^\lambda} \left[\int_{\mathbb{S}^{p-1}} \exp\left(\kappa \boldsymbol{\mu}^T \boldsymbol{\psi} \lambda\right) d\boldsymbol{\mu} \right] \\
&= \frac{\kappa^{\nu\lambda}}{[I_\nu(\kappa)]^\lambda} \left[\int_{\mathbb{S}^{p-1}} \exp\left(\|\kappa \boldsymbol{\psi} \lambda\|\ \boldsymbol{\mu}^T \hat{\boldsymbol{\mu}}\right)\ d\boldsymbol{\mu} \right] \\
&= \frac{\kappa^{\nu\lambda}}{[I_\nu(\kappa)]^\lambda} \left[\int_{\mathbb{S}^{p-1}} \exp\left(\|\kappa \boldsymbol{\psi} \lambda\|\ \boldsymbol{\mu}^T \hat{\boldsymbol{\mu}}\right)\ d\boldsymbol{\mu} \right] \\
&= \frac{(2\pi)^{p/2}}{\|\boldsymbol{\psi}\lambda\|^{\nu}} \frac{\kappa^{\nu\lambda-\nu}}{[I_\nu(\kappa)]^\lambda} I_\nu(\kappa\lambda \, \|\boldsymbol{\psi}\|) \\
&\propto \frac{\kappa^{\nu\lambda-\nu}}{[I_\nu(\kappa)]^\lambda} I_\nu(\kappa\lambda \, \|\boldsymbol{\psi}\|), \qquad (3.17)
\end{aligned}$$

where $\hat{\boldsymbol{\mu}} = \frac{\boldsymbol{\psi}}{\|\boldsymbol{\psi}\|}$ assuming $0 < \|\boldsymbol{\psi}\| < 1$.

We note that as part of the derivation in (3.17), we find the conditional distribution of $\boldsymbol{\mu}$, given κ. The conditional distribution $\pi(\boldsymbol{\mu} \mid \kappa, \lambda, \boldsymbol{\psi})$ is proportional to

$$\boldsymbol{\mu} \mid \kappa, \lambda, \boldsymbol{\psi} \propto \exp\left\{\boldsymbol{\mu}^T(\kappa\lambda\boldsymbol{\psi})\right\},$$

implying

$$\boldsymbol{\mu} \mid \kappa, \lambda, \boldsymbol{\psi} \sim \text{vMF}\left(\frac{\boldsymbol{\psi}}{\|\boldsymbol{\psi}\|}, \|\kappa\lambda\boldsymbol{\psi}\|\right) \tag{3.18}$$

This clearly implies that our conjugate prior can be deconstructed as a product of a well-known distribution $\pi(\boldsymbol{\mu} \mid \kappa, \boldsymbol{\psi}, \lambda)$ for the density of the mean vector conditionally on the concentration and the marginal prior $\pi(\kappa \mid \boldsymbol{\psi}, \lambda)$ in (3.17) with unknown properties.

To that end, the following Theorem 4 establishes various properties of this marginal density of κ. In particular, we establish that $\pi(\kappa \mid \boldsymbol{\psi}, \lambda)$ is increasing under certain parameter values. We further show unimodality of the marginal density under the alternative parameter values. We also provide a theoretical guarantee that this inflection point is certain to occur between 0 and the mode of density.

Theorem 4. *Let $\pi(\kappa \mid \boldsymbol{\psi}, \lambda)$ denotes the marginal density for κ as given by (3.17). Then,*

1. *$\pi(\kappa \mid \boldsymbol{\psi}, \lambda)$ is decreasing function of κ if $\lambda||\boldsymbol{\psi}||^2 \leq 1$.*

2. *$\pi(\kappa \mid \boldsymbol{\psi}, \lambda)$ has unique modal point if $\lambda||\boldsymbol{\psi}||^2 > 1$.*

3. *The function $\pi(\kappa \mid \psi, \lambda)$ has inflection point κ_{in} between 0 and the mode $\hat{\kappa}$. There will be no inflection point if $\lambda\|\boldsymbol{\psi}\|^2 \leq 1$.*

Readers are referred to C for proof of the theorem. This theorem provides insight into the nature of the marginal distribution of the CvMF distribution.

3.3.2 Posterior Properties of vMF Distribution Using New Conjugate Prior

Assuming a random sample of data from a vMF distribution, the posterior distribution can be easily calculated using this new conjugate prior. We let $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n \in \mathbb{S}^{p-1}$ be an i.i.d. sample of data generated from the vMF distribution with mode $\boldsymbol{\mu} \in \mathbb{S}^{p-1}$ and concentration $\kappa > 0$. For a Bayesian analysis of the data, we consider following model

$$\begin{aligned} \boldsymbol{y}_i \mid \boldsymbol{\mu}, \kappa &\overset{i.i.d.}{\sim} \text{vMF}\left(\boldsymbol{\mu}, \kappa\right) \text{ for } i = 1, \ldots, n \\ \boldsymbol{\mu}, \kappa \mid \boldsymbol{\psi}, \lambda &\sim \text{CvMF}(\boldsymbol{\psi}, \lambda), \end{aligned}$$

where $\boldsymbol{\psi} \in \mathbb{R}^d$, $\lambda > 0$, such that $\|\boldsymbol{\psi}\| < 1$. We use the notation $\boldsymbol{Y} = \{\boldsymbol{y}_i\}_{i=1}^n$ to denote all the observed data with $\overline{\mathbf{Y}} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{y}_i$, and we continue to let $\nu = \frac{p}{2} - 1$. The data likelihood can be written as

$$L(\boldsymbol{\mu}, \kappa \mid \boldsymbol{Y}) = \prod_{i=1}^n \frac{\kappa^\nu \ \exp\left(\kappa \ \boldsymbol{\mu}^T \boldsymbol{y}_i\right)}{(2\pi)^{p/2} \ I_\nu(\kappa)}.$$

Therefore, the joint posterior of $(\boldsymbol{\mu}, \kappa)$ given the data is specified as

$$\pi(\boldsymbol{\mu}, \kappa \mid \boldsymbol{Y}, \boldsymbol{\mu}, \lambda, \boldsymbol{\psi}) \propto \frac{\kappa^{\nu(n+\lambda)} \ \exp\left\{\boldsymbol{\mu}^T(\kappa \sum_{i=1}^n \boldsymbol{y}_i + \kappa\lambda\boldsymbol{\psi})\right\}}{(2\pi)^{np/2} \ [I_\nu(\kappa)]^{(n+\lambda)}}. \tag{3.19}$$

Here we can observe that the posterior has same form as the prior distribution given in (1). This posterior distribution can be written as CvMF$(\boldsymbol{\psi}^*, \lambda^*)$ where $\boldsymbol{\psi}^* = \left(\frac{n}{n+\lambda}\overline{\mathbf{Y}} + \frac{\lambda}{n+\lambda}\boldsymbol{\psi}\right)$ and $\lambda^* = n + \lambda$. In particular, we can appeal to our Theorems 1 and 2 to note that the posterior will be proper as $\|\boldsymbol{\psi}^*\| < 1$ always, following the triangle inequality. Hence, the posterior modes exist, and Theorem 2 immediately implies that $\hat{\boldsymbol{\mu}} = \frac{\boldsymbol{\psi}^*}{\|\boldsymbol{\psi}^*\|}$ and $\hat{\kappa} = R_\nu^{-1}(\|\boldsymbol{\psi}^*\|)$.

Note that $\boldsymbol{\psi}^*$ is as a convex combination of prior modal parameter and the sample mean when we use CvMF prior, and λ^* is a combination of the sample size and prior sample size. We note that the posterior modal estimates $\hat{\boldsymbol{\mu}} = \frac{\boldsymbol{\psi}^*}{\|\boldsymbol{\psi}^*\|}$ and $\hat{\kappa} = R_\nu^{-1}(\|\boldsymbol{\psi}^*\|)$ are not weighted averages of the data and hyperparameters.

3.4 Bayesian von Mises Fisher Mixture Model Using Conjugate Prior

3.4.1 Finite Mixture of vMF distributions

As discussed in the introduction, we wish to implement our methodology in the context of a finite mixture model for directional data. To that end, we assume that the data are made up of N (finite) clusters, where within each cluster data follow a unique vMF distribution. Following the notations in Qin et al. [2016], we let $f_j(\boldsymbol{y} \mid \boldsymbol{\omega}_j)$ denote a vMF distribution with parameters $\boldsymbol{\omega}_j = (\boldsymbol{\mu}_j, \kappa_j)$ for $j = 1, 2, \ldots, N$. Banerjee et al. [2005] proposed the standard von Mises Fisher Mixture Model (vMFMM) with density is given by

$$f(\boldsymbol{y} \mid \boldsymbol{\pi}, \boldsymbol{\omega}_j) = \sum_{j=1}^{N} \pi_j f_j(\boldsymbol{y} \mid \boldsymbol{\omega}_j),$$

where $f(\boldsymbol{y} \mid \boldsymbol{\mu}_j, \kappa_j) = \frac{\kappa_j^{p/2-1}}{(2\pi)^{p/2} I_{p/2-1}(\kappa_j)} \exp\left(\kappa_j \boldsymbol{\mu}_j^T \boldsymbol{y}\right) \mathbb{I}(\boldsymbol{y} \in \mathbb{S}^{p-1})$. Here π_j is the cluster allocation probability for jth cluster. The full data likelihood is then given as

$$L(\{\boldsymbol{\mu}_j\}_{j=1}^N, \{\kappa_j\}_{j=1}^N, \boldsymbol{\pi}, \mathbf{Z} \mid \boldsymbol{y}) = \prod_{i=1}^{n} \sum_{j=1}^{N} \left\{ \pi_j \frac{\kappa_j^{p/2-1}}{(2\pi)^{p/2} I_{p/2-1}(\kappa_j)} \exp\left(\kappa_j \boldsymbol{\mu}_j^T \mathbf{y}_i\right) \right\}. \quad (3.20)$$

This combination of sum and product is intractable, representing a common challenge to fitting mixture models.

A standard technique to reformulate this mixture model is to introduce auxiliary categorical variables $\mathbf{Z} = (Z_1, \ldots Z_n)^T$ such that

$$P(Z_i = j) = \pi_j, \quad j = 1, 2, ..., N; i = 1, 2, ..., n.$$

The Z_i's are often referred to as the membership variables; for example, the event $\{Z_i = j\}$ implies that the i^{th} data point is assigned to j^{th} cluster and follows the vMF distribution determined by the parameters $(\boldsymbol{\mu}_j, \kappa_j)$. The likelihood (3.20) of the parameters $\{\kappa_j\}_{j=1}^N$, $\{\boldsymbol{\mu}_j\}_{j=1}^N$, $\boldsymbol{\pi}$ after introduction of this auxiliary variables can be rewritten as

$$L(\{\boldsymbol{\mu}_j\}_{j=1}^N, \{\kappa_j\}_{j=1}^N, \boldsymbol{\pi}, \mathbf{Z} \mid \boldsymbol{y}) = \prod_{j=1}^N \prod_{i=1}^n \left\{ \pi_j \frac{\kappa_j^{p/2-1}}{(2\pi)^{p/2} I_{p/2-1}(\kappa_j)} \exp\left(\kappa_j \boldsymbol{\mu}_j^T \boldsymbol{y}_i\right) \right\}^{\mathbb{I}(Z_i=j)} . \tag{3.21}$$

3.4.2 Bayesian Model Based on Conjugate Prior

In order to formulate the Bayesian model, we consider our proposed conjugate prior for $(\boldsymbol{\mu}_j, \kappa_j)$, while we assume a standard Dirichlet distribution for $\boldsymbol{\pi}$. The full model can be represented as the following hierarchy, which we refer to as the Bayesian Conjugate von Mises Fisher Mixture Model.

$$\begin{aligned}
\boldsymbol{y}_i \mid Z_i = z_i, \boldsymbol{\mu}_i, \kappa_i &\sim \text{vMF}(\boldsymbol{\mu}_{z_i}, \kappa_{z_i}) \text{ for } i = 1, \dots n, \\
Z_i \mid \boldsymbol{\pi} &\sim \text{Categorical}(\boldsymbol{\pi}) \text{ for } i = 1, \dots n, \\
\boldsymbol{\mu}_j, \kappa_j \mid \boldsymbol{\psi}, \lambda &\sim \text{CvMF}(\boldsymbol{\psi}, \lambda) \\
\boldsymbol{\pi} &\sim \text{Dirichlet}(\boldsymbol{\eta}),
\end{aligned} \tag{3.22}$$

where the hyperparameters $\boldsymbol{\psi} \in \mathbb{R}^d$, $\lambda > 0$, and $\boldsymbol{\eta} = (\eta_1, \dots \eta_N)^T$ such that $\eta_j > 0$ for $j = 1, \dots N$. We obtain the full posterior as

$$\begin{aligned}
&\pi(\{\boldsymbol{\mu}_j\}_{j=1}^N, \{\kappa_j\}_{j=1}^N, \boldsymbol{\pi}, \mathbf{Z} \mid \boldsymbol{y}) \\
\propto &\prod_{j=1}^N \prod_{i=1}^n \left\{ \pi_j \frac{\kappa_j^{p/2-1}}{(2\pi)^{p/2} I_{p/2-1}(\kappa_j)} \exp\left(\kappa_j \boldsymbol{\mu}_j^T \boldsymbol{y}_i\right) \right\}^{\mathbb{I}(Z_i=j)} \pi(\boldsymbol{\mu}_j, \kappa_j)\ \pi(\boldsymbol{\pi}).
\end{aligned} \tag{3.23}$$

3.4.3 MCMC Algorithm for Conj-MH

Based on the above mentioned model, we propose an MCMC algorithm to draw samples from this posterior. The sampling algorithm iterates between the following steps.

1. Mixture Probability $\boldsymbol{\pi}$: The vector of mixture probabilities $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_N)$ is sampled from
$$\boldsymbol{\pi} \mid \boldsymbol{Y} \sim \text{Dirichlet}(n_1 + \eta, \cdots, n_N + \eta),$$
where $n_j = \sum_{i=1}^{n} I(Z_i = j)$.

2. Cluster Membership $\mathbf{Z}$: Cluster membership Z_i can be sampled for $i = 1, 2, \ldots, n$ as
$$P(Z_i = j \mid \boldsymbol{\pi}, \boldsymbol{y}_i, \{\boldsymbol{\mu}_j\}_{j=1}^{N}, \{\kappa_j\}_{j=1}^{N}) \propto \frac{f(\boldsymbol{y}_i \mid \boldsymbol{\mu}_j, \kappa_j)\pi_j}{\sum_{j=1}^{N} f(\boldsymbol{y}_i \mid \boldsymbol{\mu}_j, \kappa_j)\pi_j}.$$

3. Concentration parameter κ: For $j = 1, 2, \ldots, N$, the sampling distribution is
$$\pi(\kappa_j \mid \mathbf{Z}, \boldsymbol{Y}, \boldsymbol{\psi}) \propto \frac{\kappa_j^{\nu(n_j+\lambda)-\nu}}{[I_\nu(\kappa_j)]^{\lambda+n_j}} I_\nu\left(\kappa_j \left\Vert \lambda\boldsymbol{\psi} + \sum_{i=1}^{n} I(Z_i = j)\boldsymbol{y}_i \right\Vert\right).$$
We use random walk Metropolis Hastings (MH) algorithm to sample from this distribution.

4. Direction Vector $\boldsymbol{\mu}_j$: We sample mean vector for $j = 1, 2, \ldots, N$ as
$$\boldsymbol{\mu}_j \mid \mathbf{Z}, \boldsymbol{y}, \kappa_j \sim \text{vMF}\left(\frac{\boldsymbol{\Delta}_j^*}{||\boldsymbol{\Delta}_j||^*}, ||\boldsymbol{\Delta}_j||^*\right),$$
where $\boldsymbol{\Delta}_j^* = \kappa_j\lambda\ \boldsymbol{\psi} + \sum_{i=1}^{n} I(Z_i = j)\kappa_j\boldsymbol{y}_i$.

Note that in the κ_j step we utilize an MH step for this non-standard distribution. For this reason we refer to the implementation of this model as Conj-MH. This Metropolis Hastings algorithm is a flexible and popular strategy for obtaining posterior samples under non-standard distributions. One of the key steps is the selection of an appropriate proposal distribution for the candidate parameter value. However, this can be challenging and typically requires user tuning. Further, MH sampling is often less efficient than competing methods due in part to repeated samples across iterations. We note that it is potentially possible to develop a more sophisticated sampling approach for this distribution. In particular, one could use the results from Theorem 4 to develop an Adaptive Rejection Sampling step for this step [Gilks and Wild, 1992].

3.5 An Alternative Prior Distribution with a Data Augmentation Algorithm

3.5.1 Bayesian Data Augmented von Mises Fisher Mixture Model

In this section, we consider a Bayesian mixture model which assumes vMF and Gamma priors for all the cluster specific parameters $\{\boldsymbol{\mu}_j\}_{j=1}^N$ and $\{\kappa_j\}_{j=1}^N$. This model is similar to Gopal and Yang [2014] except that we use Gamma priors for κ instead of the log-normal distribution they employ. The main goal is to derive a data augmentation sampling scheme which will provide closed-form, known sampling distributions without resorting to an MH algorithm. We call this model as Gam-DA. The full model can be represented as the following hierarchy.

$$\begin{aligned}
\boldsymbol{y}_i \mid Z_i = z_i, \boldsymbol{\mu}_i, \kappa_i &\sim \text{vMF}(\boldsymbol{\mu}_{z_i}, \kappa_{z_i}) \text{ for } i = 1, \dots n, \\
Z_i \mid \boldsymbol{\pi} &\sim \text{Categorical}(\boldsymbol{\pi}) \text{ for } i = 1, \dots n, \\
\boldsymbol{\mu}_j &\sim \text{vMF}(\boldsymbol{\theta}, \zeta) \quad j = 1, 2, ..., N \\
\kappa_j &\sim \text{Gamma}(\alpha, \beta) \quad j = 1, 2, ..., N \\
\boldsymbol{\pi} &\sim \text{Dirichlet}(\boldsymbol{\eta}),
\end{aligned} \tag{3.24}$$

where the hyper parameters $\boldsymbol{\theta} \in \mathbb{S}^2$, $\zeta, \alpha, \beta > 0$ and $\boldsymbol{\eta} = (\eta_1, \dots \eta_N)^T$ such that $\eta_j > 0$ for $j = 1, \dots, N$.

From the likelihood in (3.20) and the prior structure (3.24), we get that the full posterior is proportional to

$$\begin{aligned} & \pi(\{\boldsymbol{\mu}_j\}_{j=1}^N, \{\kappa_j\}_{j=1}^N, \boldsymbol{\pi}, \mathbf{Z} \mid \boldsymbol{y}_i) \\ \propto \quad & \pi(\boldsymbol{\pi}) \prod_{j=1}^{N} \prod_{i=1}^{n} \left\{ \pi_j \frac{\kappa_j^{p/2-1}}{(2\pi)^{p/2} I_{p/2-1}(\kappa_j)} \exp\left(\kappa_j \boldsymbol{\mu}_j^T \mathbf{y}\right) \right\}^{\mathbb{I}(Z_i = j)} \pi(\boldsymbol{\mu}_j)\ \pi(\kappa_j) \\ = \quad & \pi(\boldsymbol{\pi}) \prod_{j=1}^{N} \left\{ \pi_j^{n_j} \frac{\kappa_j^{n_j(p/2-1)}}{(I_{p/2-1}(\kappa_j))^{n_j}} \exp\left(\kappa_j \boldsymbol{\mu}_j^T \sum_{i=1}^{n} \mathbb{I}\left(Z_i = j\right) \boldsymbol{y}_i \right) \right\} \pi(\boldsymbol{\mu}_j)\ \pi(\kappa_j). \end{aligned} \tag{3.25}$$

Sampling of $\mathbf{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}$ are uncomplicated and follow roughly the same sampling steps as in the MCMC algorithm under the CvMF choice. However, the posterior distribution of κ_j, which is proportional to

$$\frac{\kappa_j^{n_j(p/2-1)+\alpha-1}}{(I_{p/2-1}(\kappa_j))^{n_j}} \exp\left(\kappa_j \boldsymbol{\mu}_j^T \sum_{i=1}^{n} \mathbb{I}\left(Z_i = j\right) \boldsymbol{y}_i - \beta\kappa_j \right), \tag{3.26}$$

is not a standard distribution. To that end, we propose a data augmentation algorithm to facilitate sampling of the concentration.

The data augmentation approach was first proposed by Tanner and Wong [1987] to make simulation feasible and simple. The central idea of the DA algorithm is to sample from an intractable density, say $\pi(\boldsymbol{\omega})$, by constructing a joint density $\pi(\boldsymbol{\omega}, V)$ such that

$$\int \pi(\boldsymbol{\omega}, V)\ dV = \pi(\boldsymbol{\omega}),$$

and that the conditionals $\pi(\boldsymbol{\omega} \mid V)$ and $\pi(V \mid \boldsymbol{\omega})$ can be (easily) sampled from. In the context of Bayesian analysis, $\pi(\boldsymbol{\omega})$ typically refers to posterior density or conditional sampling density for a parameter of interest $\boldsymbol{\omega}$. Though the underlying idea behind

DA is uncomplicated, it is often nontrivial and a matter of some art to construct an appropriate choice for the distribution $\pi(\boldsymbol{\omega}, V)$. A commonly used strategy is to build an appropriate conditional distribution $\pi(V \mid \boldsymbol{\omega})$ so that $\pi(\boldsymbol{\omega}, V) = \pi(V \mid \boldsymbol{\omega})\ \pi(\boldsymbol{\omega})$. We call V the augmented random variable. In certain cases, novel distributions must be created to provide a distribution $\pi(V \mid \boldsymbol{\omega})$ that befits the need of the specific DA algorithm. In general the DA algorithm is an effective technique to address the intractability of a distribution that may cause sampling difficulties. Readers are referred to Hobert [2011] for detailed methodologies for finding appropriate augmentation variables.

3.5.2 Negative Binomial Data Augmentation Algorithm

We recall from (3.2) that the Bessel function of first kind can be written as

$$I_{p/2-1}(\kappa) := \sum_{k=0}^{\infty} \frac{1}{k!\ \Gamma(|p/2-1|+k+1)} \left(\frac{\kappa}{2}\right)^{(p/2-1+2k)}.$$

In the particular case of interest when data are spherical $p = 3$, the function can be simplified to

$$I_{1/2}(\kappa) = \sqrt{\frac{2}{\pi\kappa}} \sinh(\kappa),$$

where $\sinh(\kappa) = \frac{1}{2}e^{\kappa}(1 - e^{-\kappa})$ is the usual hyperbolic sine function. From (3.26), we can see that the intractable Bessel function appears in the denominator of the sampling distribution for κ. In terms of sinh, we can get the likelihood of the parameters $\{\kappa_j\}_{j=1}^N$, $\{\boldsymbol{\mu}_j\}_{j=1}^N$, $\boldsymbol{\pi}$ as

$$L(\{\boldsymbol{\mu}_j\}_{j=1}^N, \{\kappa_j\}_{j=1}^N, \boldsymbol{\pi}, \mathbf{Z} \mid \boldsymbol{Y}) = \prod_{j=1}^{N}\prod_{i=1}^{n} \left\{\pi_j \frac{\kappa_j}{4\pi \sinh(\kappa_j)} \exp\left(\kappa_j \boldsymbol{\mu}_j^T \boldsymbol{y}_i\right)\right\}^{\mathbb{I}(Z_i=j)}. \quad (3.27)$$

Now, we introduce an augmented variable such that the Bessel function will be part of numerator of its distribution, and we will get rid of this computational complexity. We introduce an augmentation variable for each cluster. Here we augment the likelihood with latent variables $V_1, \ldots, V_N$ for each cluster where

$$V_j \mid \kappa_j, n_j \sim \text{Negative-Binomial}(n_j, 1 - e^{-2\kappa_j}) \text{ for } j = 1, \ldots N.$$

Note that the probability mass function of the V_j distribution can be represented as

$$\begin{aligned} P(V_j = v_j \mid \kappa_j, n_j) &= \binom{v_j + n_j - 1}{v_j} e^{-2\kappa_j v_j} (1 - e^{-2\kappa_j})^{n_j} \\ &= \binom{v_j + n_j - 1}{v_j} e^{-\kappa_j(2v_j + n_j)} (e^{\kappa_j} - e^{-\kappa_j})^{n_j} \\ &= 2^{n_j} \binom{v_j + n_j - 1}{v_j} e^{-\kappa_j(2v_j + n_j)} \left(\sinh(\kappa_j)\right)^{n_j}. \end{aligned}$$

We can write the likelihood function for the augmentation variables as

$$L(\mathbf{V} \mid N, \kappa) = \prod_{j=1}^{N} 2^{n_j} \binom{v_j + n_j - 1}{v_j} e^{-\kappa_j(2v_j + n_j)} \left(\sinh(\kappa_j)\right)^{n_j}.$$

Note that after augmentation of the latent variables the resulting posterior becomes free from the $\sinh(\cdot)$ function as the p.m.f. of V contains the $(\sinh(\kappa_j))^{n_j}$ in the numerator. The complete likelihood of the parameters $\{\kappa_j\}_{j=1}^N$, $\{\boldsymbol{\mu}_j\}_{j=1}^N$, $\boldsymbol{\pi}$ after the introduction of this auxiliary variable is given by

$$\begin{aligned} & L(\{\boldsymbol{\mu}_j\}_{j=1}^N, \{\kappa_j\}_{j=1}^N, \{\pi_{\mathbf{j}}\}_{j=1}^N, \{v_j\}_{j=1}^N, \mathbf{Z} \mid \boldsymbol{Y}) \\ \propto & \prod_{j=1}^{N} \left[2^{n_j} \binom{v_j + n_j - 1}{v_j} e^{-\kappa_j(2v_j + n_j)} \left(\sinh(\kappa_j)\right)^{n_j} \prod_{i=1}^{n} \left\{ \pi_j \frac{\kappa_j}{4\pi \sinh(\kappa_j)} \exp\left(\kappa_j \boldsymbol{\mu}_j{}^T \boldsymbol{y}_i\right) \right\}^{I(Z_i = j)} \right] \\ \propto & \prod_{j=1}^{N} \left[2^{n_j} \binom{v_j + n_j - 1}{v_j} e^{-\kappa_j(2v_j + n_j)} \pi_j^{n_j} \kappa_j^{n_j} \exp\left\{ \kappa_j \boldsymbol{\mu}_j^T \sum_{i=1}^{n} \mathbb{I}\left(Z_i = j\right) \boldsymbol{y}_i \right) \right\} \right]. \end{aligned}$$

3.5.3 MCMC Algorithm for Gam-DA

After introducing the data augmentation variables $V_1, \ldots, V_n$, we are able to provide an MCMC sampling scheme that relies only sampling from known distributions. The sampling algorithm iterates between the following steps.

1. Mean vector: We sample mean vector for $j = 1, 2, \ldots, N$ from

$$\boldsymbol{\mu}_j \mid \mathbf{Z}, \boldsymbol{y}, \kappa_j, \sim \text{vMF}\left(\frac{\boldsymbol{\Delta}_j}{\|\boldsymbol{\Delta}_j\|}, \|\boldsymbol{\Delta}_j\|\right),$$

   where $\boldsymbol{\Delta}_j = \zeta\boldsymbol{\theta} + \sum_{i=1}^{n} I(Z_i = j)\kappa_j \boldsymbol{y}_i$.

2. Augmented Variable: We sample Data Augmentation variable V_j for $j = 1, \ldots, N$ from

$$V_j \mid \kappa_j, n_j \sim \text{Negative-Binomial}(n_j, 1 - e^{-2\kappa_j}) \text{ for } j = 1, \ldots N.$$

3. Concentration parameter: We sample the concentration parameter κ_j for $j = 1, 2, \ldots, N$. Using the data augmented posterior, the conditional sampling distribution will be

$$\pi(\kappa_j \mid \mathbf{Z}, \boldsymbol{y}, \boldsymbol{\mu}_j, v_j) \quad \propto \quad \kappa_j^{n_j+\alpha-1} \exp\left\{-\kappa_j\left(-\boldsymbol{\mu}_j^T \sum_{i=1}^{n} \mathbb{I}\left(Z_i = j\right) \boldsymbol{y}_i + (2v_j + n_j) + \beta\right)\right\}.$$

   This is the typical kernel of a Gamma distribution, showing that the augmented variable V_j has removed the intractability from κ_j posterior distribution. Hence, this sampling distribution is

$$\kappa_j \mid \mathbf{Z}, \boldsymbol{Y}, \boldsymbol{\mu}_j, v_j \sim \text{Gamma}(n_j + \alpha, -\boldsymbol{\mu}_j^T \sum_{i=1}^{n} \mathbb{I}\left(Z_i = j\right) \boldsymbol{y}_i + (2v_j + n_j) + \beta).$$

4. Mixture Probability $\boldsymbol{\pi}$: The vector of mixture probabilities $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_N)$ is

sampled from

$$\boldsymbol{\pi} \mid \boldsymbol{Y} \sim \text{Dirichlet}(n_1 + \eta, \cdots, n_n + \eta),$$

where $n_j = \sum_{i=1}^{n} I(Z_i = j)$.

5. Cluster Membership $\mathbf{Z}$: We sample the cluster membership Z_i for $i = 1, 2, \ldots, n$ through

$$P(Z_i = j \mid \boldsymbol{\pi}, \boldsymbol{Y}, \{\boldsymbol{\mu}_j\}_{j=1}^{N}, \{\kappa_j\}_{j=1}^{N}, \mathbf{V}) \propto \frac{f(\boldsymbol{y}_i \mid \boldsymbol{\mu}_j, \kappa_j)\pi_j}{\sum_{j=1}^{N} f(\boldsymbol{y}_i \mid \boldsymbol{\mu}_j, \kappa_j)\pi_j}.$$

3.5.4 Bayesian Metropolis Hastings von Mises Fisher Mixture Model

To investigate the impact of the data augmentation choice, relative to a naive sampling step for κ_j, we also implement a version of this model where update κ_j from (3.26) using a random walk Metropolis Hastings step. We call this version of sampling model as Gam-MH.

3.6 Point Estimation and Inference

We use the developed data augmentation sampling algorithm to obtain a large number of posterior samples from each of our models/sampling schemes. The main parameters required for inference are the $\{\boldsymbol{\mu}_j\}_{j=1}^{N}, \{\kappa_j\}_{j=1}^{N}, \{\pi_j\}_{j=1}^{N}, \mathbf{Z}$. These parameters determine the mean of the cluster, spread of the cluster, probability of allocation for the cluster, and the cluster membership labels for each observation.

To evaluate mixing and convergence, we inspect traceplots of the log-likelihood to evaluate global convergence of the algorithm and to select a length for burn-in. Autocorrelation among the MCMC samples increases uncertainty in estimation of parameters. The effective sample size provides the number of independent samples that would contain an equivalent amount of information as the (correlated) samples from the given MCMC output. Typically, we seek to run the MCMC long enough to obtain an effective sample size of at least 1000 for the key parameters of interests.

Label switching is a typical characteristic of Bayesian mixture model fitting. The term label switching was coined by Redner and Walker [1984] to describe the phenomenon of the invariance of likelihood under relabeling of mixture components. In the Bayesian context this leads to symmetric and multi-modal posterior distributions which results in nonsensical parameter estimates when consider posterior mean without further adjustment [Stephens, 2000]. There are several methods that have been proposed to address this issue [Frühwirth-Schnatter, 2001, Stephens, 2000, Marin et al., 2007, Sperrin et al., 2010, Papastamoulis, 2014]. Readers are referred to Jasra et al. [2005] and Papastamoulis [2016] for a summary of some of these methods. In our project we use the Kullback-Leibler (KL) based method proposed by Stephens [2000]. In this algorithm, an initial estimate for the cluster-specific parameters is chosen, and the cluster labels at each iteration are permuted to minimize the KL distance between the samples of the given iteration and the overall estimates. This algorithm iterates between updating the global estimates and recomputing the permutations at each level. Readers are refereed to Stephens [2000] for further details.

After we obtain the relabeled MCMC posterior samples using the Stephens algorithm, we obtain $\{\hat{\kappa}_j\}_{j=1}^N$ point estimates by taking the sample mean for each κ_j from the MCMC output (after burn-in and thinning). We obtain $\{\hat{\boldsymbol{\mu}}_j\}_{j=1}^N$ point estimates by taking sample average of estimated $\boldsymbol{\mu}_j$ and then dividing it by the norm to make $\|\hat{\boldsymbol{\mu}}_j\| = 1$. As part of the Stephens relabelling algorithm, we obtain an estimated cluster membership for each observation which we take as our point estimate for $\mathbf{Z}$.

Table 3.1: Simulation Settings

Setting #	1	2	3	4	5	6
ϕ	30	45	60	30	45	60
κ	50	50	50	25	25	25

3.7 Simulations

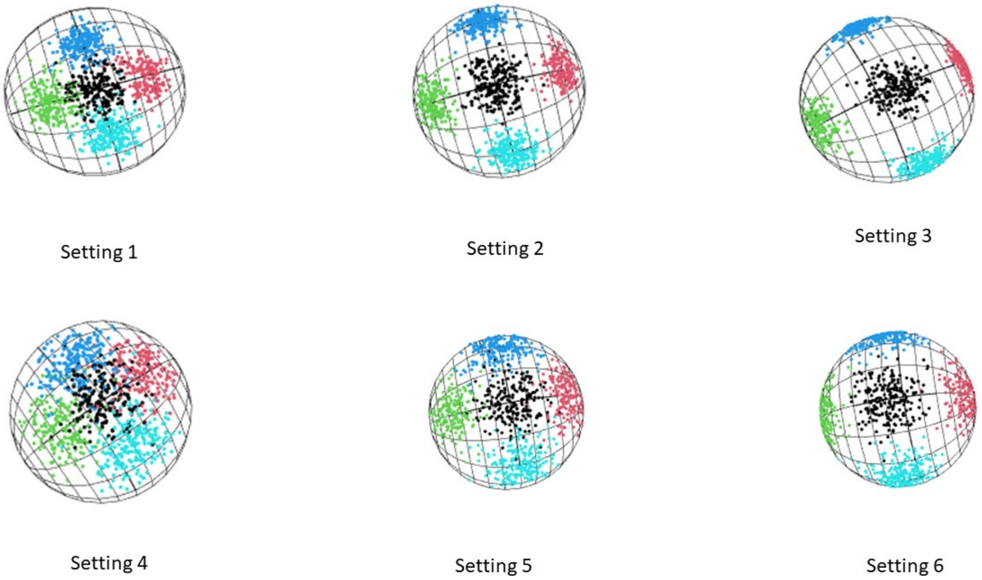
3.7.1 Simulation Setting

In this section we compare the performance of our proposed methodology across several situations and compare it with competing methods. In particular, we asses how well the newly proposed algorithm recovers the parameters of the underlying cluster distributions, as well as how well it recovers the true cluster memberships of the generated data.

To perform our experiments we consider 6 different experiment settings, as shown in Table 3.1. These settings differ in the concentration parameters, as well as in the amount of separations among clusters. Each setting consists of 100 data sets. Here ϕ denotes angle for separation of the clusters. Throughout, we assume a sample size of $n = 1000$, and $N = 5$ clusters with an allocation probability of $\pi_j = 0.2$. The mean of the clusters are determined from

$$\begin{aligned}
\boldsymbol{\mu}_1 &= (1, 0, 0)\\
\boldsymbol{\mu}_2 &= (\cos\phi, \sin\phi, 0)\\
\boldsymbol{\mu}_3 &= (\cos\phi, -\sin\phi, 0)\\
\boldsymbol{\mu}_4 &= (\cos\phi, 0, \sin\phi)\\
\boldsymbol{\mu}_5 &= (\cos\phi, 0, -\sin\phi).
\end{aligned}$$

The value of $\phi = 30^\circ, 45^\circ, 60^\circ$ determines how far apart the clusters are separated. Further varying $\kappa = 25, 50$ determines how closely the data within each cluster are



Setting 1

Setting 2

Setting 3

Setting 4

Setting 5

Setting 6

Figure 3.2: Plot of a representative sample data from each of the 6 simulation settings

.

concentrated. As κ becomes smaller, the spread of the data becomes larger, resulting in overlap of clusters. This proves challenging to determine the cluster memberships of the sample points. We generate the data using rvmf function in package Directional. Figure 3.2 denotes a plot of one representative data sets from each simulation setting. We can clearly see that first row corresponds to higher κ_j hence the data points are more clustered near their cluster direction vectors than in the second row with the lower κ_js. As ϕ increases clusters are well separated from each other.

For each method we fit models with different values of N ranging from 2 to 8 and choose an appropriate model using a standard model selection criteria. For the Bayesian mixture model we use the Deviance Information Criteria (DIC) [Gelman et al., 2014] given by

$$\begin{aligned} D(\boldsymbol{\omega}) &= -2\log(p(\boldsymbol{y} \mid \boldsymbol{\omega})) \\ DIC &= \overline{D(\boldsymbol{\omega})} + 0.5 \text{ var}(D(\boldsymbol{\omega})). \end{aligned}$$

Here $\overline{D(\boldsymbol{\omega})}$ denotes the posterior expectation of the deviance $D(\boldsymbol{\omega})$ with respect to the parameters $\boldsymbol{\omega}$. For frequentist methods, we use the Bayesian Information Criteria (BIC) [Schwarz, 1978] for model selection according to

$$BIC \quad = \quad -2\log(\hat{L}) + (4N-1)\log n.$$

As is standard, we choose the model with N producing the minimum BIC/DIC within each estimation approach.

We calculate the accuracy of cluster membership by using the Rand Index [RI; Rand, 1971] between the estimated cluster and the true data generating cluster. RI is a measure of similarity between two clusterings and is computed as follows. We let the agreement A be the number of pairs of observations that are both in the same cluster in the data generation clustering and in the same cluster in the estimated clustering. Similarly, we define disagreement D as the number of pairs of observations belonging to different clusters in the data generation model and as well as in the estimated clusters. Then, RI is defined by

$$RI = \frac{A+D}{n(n-1)/2},$$

which is clearly a ratio bounded between 0 and 1. The maximum value 1 indicates the estimated clustering exactly matches the true clustering from data generation. As the accuracy of the pairwise cluster memberships can be considered whether or not methods recover the true number of clusters N, we calculate the RI for both the best model obtained using DIC/BIC criteria and for the model fitted with true number of clusters.

In order to evaluate the accuracy of the parameter recovery of all cluster distributions, we consider the results when the model is fit using the correct number of clusters ($N = 5$). We calculate the Mean Total Cosine Error (MTCE) for estimation of the mean vectors as

$$MTCE(\{\hat{\boldsymbol{\mu}}_j\}_{j=1}^N, \{\boldsymbol{\mu}_j\}_{j=1}^N) = \frac{1}{n^*}\sum_{i=1}^{n*}\sum_{j=1}^{N}(1 - (\hat{\boldsymbol{\mu}}_{ji}^T\boldsymbol{\mu}_j)^2).$$

Here n^* is total number of datasets considered in each simulation setting, and $\hat{\boldsymbol{\mu}}_{ji}$ represents the point estimate for the jth cluster using the ith replicated dataset. We measure the accuracy of estimation of the concentration parameters as a Mean Total Squared Error (MTSE) between true values and estimated values through

$$MTSE(\{\hat{\kappa}\}_{j=1}^N, \{\kappa\}_{j=1}^N) = \frac{1}{n^*}\sum_{i=1}^{n*}\sum_{j=1}^{N}(\hat{\kappa}_{ji} - \kappa_j)^2.$$

$\hat{\kappa}_{ji}$ and κ_j are the estimated and true the concentration parameters for jth cluster in ith datset. To best match the estimates of parameters with their true parameter counterpart, we permute the order of point estimates and calculate these errors MTCE and MTSE at each possible parameter. We assume the permutation of $\hat{\boldsymbol{\mu}}$ which minimizes MTCE to be the set of estimates with the correct labeling.

In addition to parameter and cluster recovery, we additionally assess the predictive accuracy of the model by considering the likelihood evaluated on a new dataset from the same data generating model. For each of the $n^* = 100$ simulated datasets, we use the posterior estimates and consider the estimated data distribution to be

$$f(\boldsymbol{Y} \mid \hat{\boldsymbol{\pi}}, \{\hat{\boldsymbol{\mu}}_j\}_{j=1}^n, \{\hat{\kappa_j}\}_{j=1}^N) = \sum_{j=1}^{N}\hat{\pi}_j f_j(\boldsymbol{Y} \mid \hat{\boldsymbol{\mu}}_j, \hat{\kappa}_j),$$

by plugging in the parameter estimates of each parameter. We then evaluate the likelihood for a new dataset of the same size ($n = 1000$) using this estimated distribution. We consider the loglikelihood version of this quantity, and report the average of this predictive distribution criteria across the 100 simulated datasets. As this criteria will behave similarly to a likelihood function, the best performing method can be taken to be the one with the highest value. This method of comparison is based on one utilized in Gaskins [2019].

3.7.2 Competing Methods

For each of the 100 data sets generated according to the six generation settings, we fit the data according to the following methods. We compare our methods with two frequentist methods. All the implementation is done in R.

1. Bayesian Data Augmented von Mises Fisher Mixture Model (Gam-DA) and Bayesian Metropolis Hastings von Mises Fisher Mixture Model (Gam-MH): We ran MCMC for 5000 iterations with 1000 as burn-in iterations. Samples are thinned to store 2000 samples. The thinned samples on average give an effective sample size of 1000–1200 based on the log-likelihood of the model. Here we choose hyperparameters $\alpha = 1$, $\beta = 1$, $\boldsymbol{\theta} = (1, 0, 0)$, $\zeta = 1$.

2. Bayesian Conjugate von Mises Fisher Mixture Model (Conj-MH): We ran MCMC for 5000 iterations with 1000 burn-in iteration. Samples are thinned to store 2000 samples. The thinned samples on average give an effective sample size of 1000–1200 based on log-likelihood of the model. We choose hyperparameters $\boldsymbol{\psi} = (0.9797, 0, 0)$ and $\lambda = 2.5$ which yeilds the prior mode $(1, 0, 0)$ and 25 and prior sample size 2.5 according to Theorem 2.

3. Spherical K Means (SK Means): Dhillon and Modha [2001] proposed the spherical K means algorithm in the context of text mining. This algorithm is based on cosine similarity measure. The algorithm partitions the high dimensional unit sphere using a collection of great hypercircles. We use skmeans function available in R package SK Means to apply the algorithm on our data [Hornik et al., 2012].

4. Mixture of von Mises Fisher Distribution (MovMF): Banerjee et al. [2005] developed a frequentist mixture model of von Mises Fisher distributions. It uses EM Algorithm to fit the model. This model is coded in package movMF [Hornik and Grün, 2022] which we use for the analysis.

3.7.3 Simulation Results: Model Selection Accuracy

Table 3.2: Average Number of Estimated Clusters ($N = 5$ is true value)

Method	Setting					
	1	2	3	4	5	6
Gam-DA	5.3	5.3	5.4	4.3	6.0	5.5
Conj-MH	6.8	6.2	6.2	7.2	6.3	5.8
Gam-MH	5.1	5.1	5.1	5.3	5.1	5.1
SK Means	5.0	5.0	5.0	4.9	5.0	5.0
MovMF	5.0	5.0	5.1	4.7	5.0	5.0

Table 3.2 shows average number of clusters estimated by each method in each simulation settings. Recall that the true value is $N = 5$ in all cases. We observe that the frequentist methods selects the true model fairly accurately. The Bayesian estimates are somewhat less accurate but typically in the neighbourhood of the true number of clusters. The exception is that Conj-MH consistently overestimates the number of clusters in low concentration parameter setup ($\kappa_j = 25$). This is due to an "elbow effect" in its DIC plot where the DIC rapidly decreases over N until the true model ($N = 5$) but continues to show minor decreases for larger N. Table 3.3 summarizes the percentage of simulated datasets in which the true model is selected. The frequentest methods show fairly high percentages by selecting the true model almost 100% of time in the higher κ settings. Among the Bayesian models, MH sampling of the alternative prior model performs better than the DA sampling of the same model. The conjugate prior model typically performs worse than these choices and has consistently poor performance in the $\kappa = 25$ settings.

Table 3.3: Comparison Based on Percentage Cluster Selection

Method	Setting					
	1	2	3	4	5	6
Gam-DA	73	71	67	21	82	91
Conj-MH	14	40	42	6	35	48
Gam-MH	93	92	93	74	90	91
SK Means	100	100	100	87	100	100
MovMF	100	100	96	79	100	97

3.7.4 Simulation Results: Cluster Estimation Accuracy

Table 3.4 shows cluster estimation accuracy using the average Rand Index across datasets. We show the RIs computed both for the estimated number of clusters and for the model fit to the true number of clusters. As expected, cluster estimation accuracy is lower in the $\kappa_j = 25$ settings than in the more concentrated ($\kappa_j = 50$) cases; similarly, cluster estimation is improved as ϕ increases and the cluster means are located further apart. Unlike the model selection performance, the Bayesian methods tend to show equivalent or slightly better RI values than the SK Means and MovMF approaches. This suggests that in the cases when the Bayesian results overestimate the number of clusters, these additional clusters typically only consist of one or two observations. There is very little difference between the RI values among the Bayesian methods.

3.7.5 Simulation Results: Mean Accuracy

We study the accuracy of the estimation of the cluster in Table 3.5. As noted previously, this is a total cosine error summed over all cluster mean estimates using those results computed under the true number of clusters $N = 5$. Here we can see that all the methods are doing well in total mean estimation, although in the most challenging case (setting 4) where the data are more disperse and the clusters are more closely located, Conj-MH performs the best of all competing methods.

Table 3.4: Comparison Based on Rand Index

Method	Selection	Setting		
		1	2	3
Gam-DA	Model Selected	0.9473	0.9880	0.9998
	$N = 5$	0.9548	0.9959	0.9998
Conj-MH	Model Selected	0.9506	0.9931	0.9978
	$N = 5$	0.9548	0.9953	0.9997
Gam-MH	Model Selected	0.9553	0.9956	0.9989
	$N = 5$	0.9556	0.9960	0.9997
SK Means	Model Selected	0.9565	0.9960	0.9997
	$N = 5$	0.8851	0.9960	0.9201
MovMF	Model Selected	0.9559	0.9960	0.9923
	$N = 5$	0.9559	0.9960	0.9134
Method	Selection	Setting		
		4	5	6
Gam-DA	Model Selected	0.8310	0.9625	0.9923
	$N = 5$	0.8351	0.9625	0.9923
Conj-MH	Model Selected	0.8679	0.9576	0.9891
	$N = 5$	0.8718	0.9610	0.9914
Gam-MH	Model Selected	0.8410	0.9614	0.9917
	$N = 5$	0.8378	0.9621	0.9923
SK Means	Model Selected	0.8747	0.9632	0.9925
	$N = 5$	0.8443	0.8865	0.9016
MovMF	Model Selected	0.8462	0.9622	0.9923
	$N = 5$	0.7889	0.8368	0.8810

Table 3.5: Comparison of Different Methods Based on Total Mean Estimation Error

Method	Setting					
	1	2	3	4	5	6
Gam-DA	0.0016	0.0011	0.0010	0.0826	0.0030	0.0022
Conj-MH	0.0014	0.0010	0.0010	0.0056	0.0028	0.0022
Gam-MH	0.0016	0.0011	0.0010	0.0158	0.0031	0.0022
SK Means	0.0015	0.0011	0.0010	0.0090	0.0029	0.0023
MovMF	0.0017	0.0011	0.0010	0.0144	0.0035	0.0023

3.7.6 Simulation Results: Kappa Accuracy

Table 3.6: Comparison Based on Total κ Estimation Error

Method	Setting					
	1	2	3	4	5	6
Gam-DA	1122.56	595.52	551.44	751.15	99.90	64.25
Conj-MH	611.75	366.83	611.03	1133.13	143.11	82.01
Gam-MH	400.42	152.46	139.78	1754.79	84.32	41.18
MovMF	166.83	79.20	73.06	358.07	34.88	21.43

Table 3.6 shows total squared error for κ estimation. We note that SK Means does not produce an estimate of the concentration parameter, so we do not include it in these comparisons. Here, the MovMF method tends to outperform the various Bayesian algorithms. In particular, when the clusters are closed together ($\phi = 30°$ in settings 1 and 4) estimation of the concentration parameters is very poor. Comparing across Bayesian approaches, we see the data augmentation scheme typically performs worse than corresponding MH sampler.

Table 3.7: Comparison Based on Total Kappa Bias Estimation

Method	Setting					
	1	2	3	4	5	6
Gam-DA	-13.66	-10.17	-9.75	-9.75	-3.41	-2.66
Conj-MH	-6.28	-28.02	-44.24	15.97	-3.47	-11.13
Gam-MH	1.12	0.78	0.87	3.47	0.46	0.52
MovMF	4.87	3.21	2.89	10.86	2.23	1.79

To further understand this behavior, we consider the mean bias in Table 3.7; these values are computed as $\frac{1}{N n^*} \sum_{i=1}^{n*} \sum_{j=1}^{N} (\hat{\kappa}_{ji} - \kappa_j)$. While MovMF indicates a mostly small bias (except in setting 4), some of the Bayesian methods display a consistent bias. The data augmentation algorithm yields estimates that show a substantial negative bias (toward too clusters that are too disperse), whereas the MH implementation of this same prior structure are close to unbiased. We suspect that the (true) value of κ_j used in these settings may be ill suited to our DA algorithm.

Recall that our augmentation variables are $V_j \sim$ Negative-Binomial$(n_j, 1-e^{-2\kappa_j})$, but as our true κ_j is 25 and 50, the corresponding success probabilities are approximately one. Hence, V_js are zero and show almost no variability across zeros. This may limit our ability to distinguish among large values of κ_j. Conversely, the conjugate prior has tendency to over-estimate κ_j, especially in the settings when the true $\kappa_j = 25$.

3.7.7 Simulation Results: Predictive Analysis

Table 3.8: Comparison Based on Predictive Likelihood Estimation

Method	Selection	Settings		
		1	2	3
Gam-DA	Model Selected	-468	-582	-619
	$N = 5$	-428	-557	-564
Conj-MH	Model Selected	-1188	-2180	-3622
	$N = 5$	-412	-570	-607
Gam-MH	Model Selected	-461	-703	-639
	$N = 5$	-411	-566	-582
SK Means	Model Selected	-757	-1438	-1923
	$N = 5$	-980	-1438	-2143
MovMF	Model Selected	-394	-531	-544
	$N = 5$	-390	-526	-539
Method	Selection	Settings		
		4	5	6
Gam-DA	Model Selected	-818	-1108	-1208
	$N = 5$	-812	-1130	-1208
Conj-MH	Model Selected	-1196	-2151	-1802
	$N = 5$	-812	-1110	-1237
Gam-MH	Model Selected	-1050	-1188	-1326
	$N = 5$	-883	-1108	-1233
SK Means	Model Selected	-885	-1496	-1955
	$N = 5$	-1104	-1717	-2176
MovMF	Model Selected	-785	-1108	-1211
	$N = 5$	-777	-1097	-1199

To measure differences in the predictive performance, we compute an estimate of the mixture density by plugging in the estimated cluster probabilities and vMF parameters from the output of the selected model and the $N = 5$ model. As shown in Table 3.8. the Bayesian methods tend to perform better than the frequntist method

SK Means but fall slightly behind of MovMF, as determined by a higher/less negative predictive likelihood.

To conclude, our Bayesian approaches based on the new conjugate prior and the MH and DA sampling schemes for the standard prior all perform well overall. They produce an accurate clustering of the observations with equivalent or improved Rand Index compared to the frequentist methods. The poorer recovery of N is mainly associated with the inclusion of small clusters with few observations. All methods performs similar in mean estimation with the conjugate prior approach performing best in the most challenging settings. MovMF proved superior in κ estimation relative to the Bayes estimators, which leads to somewhat better predictive densities than the Bayesian competitors.

3.8 Diffusion Tensor Imaging Data Analysis

Diffusion Tensor Imaging is a non-invasive way of studying brain structural connectivity. It utilizes the diffusion of water molecules to generate images where the external magentic field interacts with protons in water molecules. Statistically, the movement of the water molecule is believed to follow a Brownian motion process. Water diffusion in brain tissue depends on the multiple factors such as the environment, extracellular structures, physical orientation of tissue, nerve fiber direction, etc. The flow of water molecules in the brain structure is anisotropic, that is, it has directionality associated with it. The diffusion of water is greater in the axis parallel to the orientation of the nerve fiber, meaning that the studying the directions of diffusion can reveal the anatomy of the brain. Tracking the orientation of water molecules enable us to study the structure of the brain. This is called as “Tractography” [Silva, 2016].

To model the directional data of the water molecule, a tensor model is fitted on each voxel of the brain image. At each voxel, the tensor matrix is of dimension 3×3, and we extract the principle eigenvector describing the direction along which maximum diffusion is taking place. [Johansen-Berg and Behrens, 2013]. These eigenvectors are in the sphere $\mathbb{S}^2$. Voxels in the brain may be considered as connected based upon similarities in maximum diffusion direction. Consequently, it is of interest to cluster voxels with similar diffusivity directions to potentially reveal voxels that are interconnected [McGraw et al., 2006, Silva, 2016].

In this project we apply our developed methods to cluster the voxels in the brain. The data is collected by Neuro Imaging Laboratory of Cognitive Affective and Motoric Processes (NILCAMP) at the University of Louisville. Data was collected as part of a larger project examining cognitive processes. One individual's data was selected randomly from the sample of 20 individuals. Scans were acquired using a Siemens Tim Trio 3T scanner with a 12-channel head coil. Imaging parameters for structural MRI images using a magnetization-prepared rapid gradient echo (MPRAGE) sequence were as follows: voxel size $1.0 \times 1.0 \times 1.0\ mm^3$, repetition time (TR) 2,500 ms, echo time (TE) 3.5 ms, flip angle (FA) 8o, 192 slices. Imaging parameters for DTI were as follows: 64 diffusion directions, voxel size $2.5 \times 2.5 \times 3.2\ mm^3$, TR 10,000 ms, TE 91 ms, b-value 1,000 s/mm^2, 64 slices. We have a final image grid of size $96 \times 96 \times 49$.

Diffusion data analysis was carried out using tract-based spatial statistics [Smith et al., 2006], a tool provided by FMRIB Software Library [Jenkinson et al., 2012]. First, diffusion weighted images were corrected for eddy-current-induced distortions. Fractional anisotropy (FA) images were then created by fitting a tensor model to the raw diffusion data using FSL's Diffusion Toolbox (FDT), and subsequently brain-extracted using BET [Smith, 2002]. All subjects' FA data were then aligned into a common space using FMRIB's nonlinear image registration tool [Andersson et al., 2007], which uses a b-spline representation of the registration warp field [Rueckert et al., 1999]. Next, the mean FA image was created and thinned to produce a mean FA skeleton, which represents the centers of all tracts common to the group. Each subject's aligned FA data was then projected onto this skeleton and the resulting data fed into voxelwise cross-subject statistics.

After this primary data cleaning, we extracted the principal eigenvectors and eigenvalues for the data of the selected individual. Initially, we have 451,584 principal eigenvectors. Then we calculate the Generalized Fractional Anisotropy (GFA) values based on eigenvalues. This is a scalar measure between 0 and 1, which describes the degree of anisotropy in diffusion [Glenn et al., 2015]. We select our sample by keeping only those voxels whose first eigenvector has a GFA value greater than 0.4. This ensures that the eigenvector is a good summary of the behavior within the retained voxels. While there is no consensus in the literature about the optimal GFA threshold to use, Kunimatsu et al. [2004] suggest cutoffs in the range of 0.2 and 0.4. Our threshold of 0.4 gives us a sample size of 736 eigenvectors depicting the directions for maximum diffusivity from 736 voxels.

Now we apply all the methods under study on the given data. We fit the models with N ranging from 1 to 11. For the Bayesian methods we ran MCMC to 500000 iterations with 100000 burn iterations. This ensures the ESS is more than 1000 for the log-likelihood under each value of N. We apply the DIC criteria and choose the number of clusters; for SK Means and MovMF, we apply BIC to detect the number of clusters. The label switching algorithm is applied to the MCMC output, and estimates of the cluster-specific parameters are computed for all methods as discussed in the simulation study.

The model selection statistics are shown in Table 3.9. Note that both Gam-DA and Conj-MH selected the model with $N = 3$ clusters. In contrast, the MovMF model selects the model with 6 clusters, matching the Conj-MH choice.

In Table 3.10 we show the Rand Index computed between each pair of methods at the selected value of N. Clearly, the two methods Gam-DA and Conj-MH that estimate 3 clusters yield a similar clustering with an RI of 0.7402. Similarly Gam-MH

Table 3.9: BIC/DIC Values For Competing methods

N	SK Means	MovMF	Gam-DA	Conj-MH	Gam-MH
1	13291.71	3574.70	3561.03	3561.06	3560.74
2	**3381.77**	3364.57	3332.22	3337.65	3332.45
3	3432.19	3345.33	**3331.51**	**3308.45**	3296.88
4	3436.97	3323.81	3372.90	3308.60	3254.05
5	3438.32	3339.22	3409.83	3326.00	3210.47
6	3432.63	**3315.11**	3390.86	3322.28	**3209.73**
7	3420.99	3331.29	3389.59	3320.75	3217.08
8	3408.61	3334.30	3372.34	3326.88	3224.95
9	3430.13	3353.09	3376.73	3325.31	3222.48
10	3444.11	3360.07	3377.71	3328.56	3212.68
11	3454.10	3375.77	3376.14	3328.97	3213.00

Table 3.10: Rand Index Comparison Among Different Methods

Method	SK Means	MovMF	Gam-DA	Conj-MH	Gam-MH
SK Means	-	0.5496	0.6155	0.6068	0.5381
MovMF	0.5496	-	0.7148	0.5921	0.7282
Gam-DA	0.6155	0.7148	-	0.7402	0.8414
Conj-MH	0.6068	0.5921	0.7402	-	0.7374
Gam-MH	0.5381	0.7282	0.8414	0.7374	-

and MovMF estimates same number of cluster their RI index 0.7282. Gam-DA and Gam-MH have highest RI of 0.8414, indicating the most similar pairwise behaviour, even as they have different values of N. In general, Bayesian methods tend to have higher RIs among themselves than the RIs between Bayesian and frequentist methods.

Table 3.11: Parameter for Each Estimated Cluster by Methods

Method	Direction Vector			Concentration	Mixture Probabilities
SK Means Cluster 1	0.9447	0.2546	0.2068	-	0.5082
SK Means Cluster 2	-0.8880	0.3900	0.2438	-	0.4918
MovMF Cluster 1	0.0421	0.9977	-0.0530	1.7603	0.3497
MovMF Cluster 2	-0.9409	-0.3374	0.0308	37.2946	0.0558
MovMF Cluster 3	-0.2044	0.1973	0.9588	2.2710	0.2501
MovMF Cluster 4	-0.9477	0.3173	0.0344	24.6689	0.1004
MovMF Cluster 5	0.9308	-0.3392	0.1362	12.0399	0.1087
MovMF Cluster 6	0.9454	0.3220	0.0508	25.2309	0.1353
Gam-DA Cluster 1	0.9867	0.1367	0.0877	7.9653	0.2487
Gam-DA Cluster 2	-0.1115	0.7914	0.6010	5.2717	0.3936
Gam-DA Cluster 3	-0.9820	0.1712	0.0794	5.9637	0.3576
Conj-MH Cluster 1	-0.5945	0.6671	0.4490	1.3483	0.7609
Conj-MH Cluster 2	0.9498	-0.2794	0.1406	16.0821	0.0846
Conj-MH Cluster 3	0.9427	0.3302	0.0483	18.0994	0.1545
Gam-MH Cluster 1	-0.9448	0.3267	0.0263	25.2858	0.0726
Gam-MH Cluster 2	0.9501	0.2938	0.1051	61.6790	0.2792
Gam-MH Cluster 3	-0.9457	-0.3234	0.0324	37.0420	0.0616
Gam-MH Cluster 4	0.9571	0.2632	0.1213	89.2448	0.2268
Gam-MH Cluster 5	0.0905	0.8353	0.5423	172.9198	0.1565
Gam-MH Cluster 6	0.9782	-0.1545	0.1384	63.1159	0.2033

Table 3.11 depicts the parameters of each clusters by methods. We can observe even though MovMF and Gam-MH estimates same numbers of clusters, Gam-MH estimates concentration parameters at higher range resulting in data concentrated around mean of each cluster in dense manner . Mixing probabilities for both methods are fairly same. Further we can see the difference between parameters of Gam-DA and Conj-MH. Both methods estimates same number of clusters but Conj-MH assign higher weight to cluster 1 through mixing probability of 0.7609. Also Concentration parameter for that cluster is very low depicting that the data is widely spread. On the

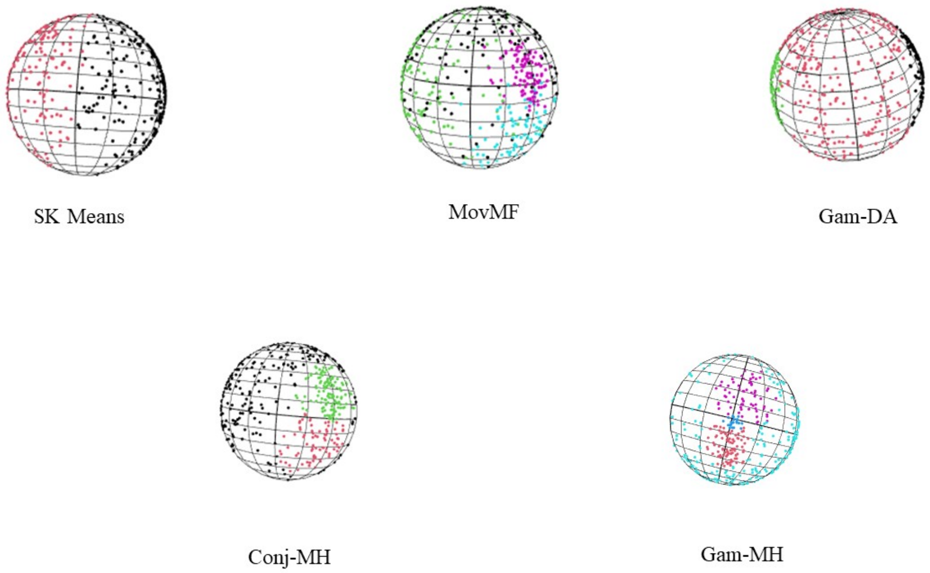


Figure 3.3: Brain data observations color-coded by estimated cluster.

other hand Gam-DA assigns fairly equal probabilities for two clusters. Concentration parameters for them To get a visual perception, we also plot the data directions in Figure in 3.3 on the sphere with color coding from the estimated clusters. Here SK means, Gam-DA and Conj-MH shows clear separation of clusters. Conj-DA shows some overlap of clusters. Also as the diagram is two dimensional we could only see 4 clusters here.

Figure 3.4 shows how the data at a certain region of the hypersphere are clustered according to different methods. This is done by fixing the view. In SK Means, the data is clearly divided in two clusters. MovMF shows some overlapping clusters, while Gam-DA, Conj-MH, Gam-MH include most of the data into one cluster while on edge we can observe two different clusters.

In terms of biological interpretations, our initial analysis shows that clusters are situated in the same region named cerebellar peduncle. The cerebellar peduncle are the brain structures connecting the cerebellum to the brain stem and cerebrum. A further analysis is being carried out to interpret the biological significance of the

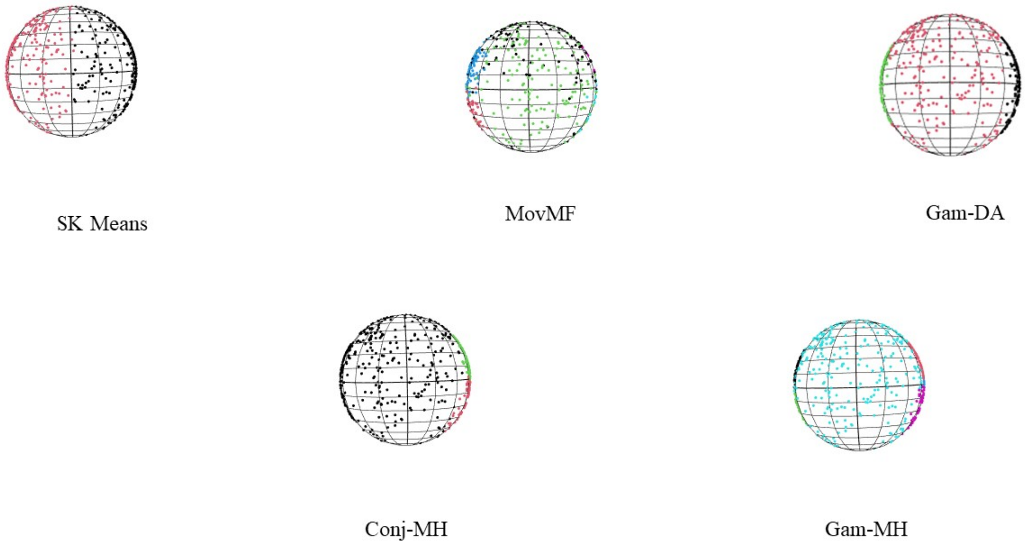


Figure 3.4: Brain data observations color-coded by estimated cluster for a data at a certain region.

clusters.

3.9 Conclusions

In the first half of this project, we have proposed a new conjugate prior for the mean and concentration parameter of the von Mises Fisher distribution. We provided a theoretical investigation of a number of key properties for this new prior distribution including posterior propriety and the unimodality of the distribution. We also examine the spread of the distribution as it relates to the λ parameter. While the conditional distribution of the mean vector follows a standard vMF, the marginal density of the concentration parameter is not a standard distribution. We have investigated certain properties, in particular we prove the existence of an inflection point where the density changes from log-convex to log-concave. This is important as this implies that we may be able to build an adaptive rejection sampling scheme to update the concentration in the MCMC sampling, rather than the less efficient Metropolis step that we currently implement. We leave this investigate as future work.

In the second half of the project, we developed a novel data augmentation model for Bayesian mixture vMF model. In this model, we augment the likelihood with a variable for each cluster which removes the intractable Bessel function from the posterior distribution. However, our empirical results showed that these estimates were substantially less accurate and negatively biased, unlike the results from the Metropolis sampling. This surprising result requires further investigation. We suspect this may be related to the relatively large values of κ_j considered in the simulations and the corresponding insensitivity of the V_j distribution to κ_j of this magnitude, but this must be verified using a wider range of simulation settings.

There are a variety of model extensions that are potentially of interest. One possibility would be to an infinite Bayesian Mixture model of VMF distributions. Additionally, the use of the data augmentation algorithm or our novel conjugate prior can further be developed and applied in the context of more complex models such as directional regression, directional time series, etc. In fact, there are some variations of the Bayesian mixture of vMFw in the literature that could benefit from combination with our new algorithms and priors. Gopal and Yang [2014] introduced the Bayesian Hierarchical vMFMM (H-vMFmix) which is designed for big data. The Temporal vMFMM (T-vMFmix) is another directional mixture designed to model clusters in data that evolve across time [Gopal and Yang, 2014]. Extending our proposed methodologies for these methods could be an interesting methodology path.

In the era of fast computations, there is always important to know if the given algorithm is fast enough to give us the trustworthy results. In other words the rate of convergence is of utmost important and a vital property of any sampling algorithm. This rate is typically desired to at least Geometrically fast. The standard method of establishing the existence of a Central Limit Theorem is to prove that the underlying Markov chain converges at a geometric rate [Roberts and Rosenthal, 1997]. Hence, it will be further interesting to study the convergence properties of our algorithm to provide theoretical assurance and accuracy of the result estimation.

CHAPTER 4

DISCUSSION

The lack of development for Bayesian methodologies in constrained spaces area motivated us to consider research problems in this area. In this dissertation, the first chapter provides a definition of constrained spaces. It also introduce the vMF distribution and the problem of CCA and provides a brief literature review.

The second chapter develops a novel infinite Bayesian factor model which is then utilized for the development of Bayesian Sparse Canonical Correlation Analysis. We demonstrated that our model performs superior to some competing frequentist and Bayesian models in the estimation of the key CCA parameters of interest, the first two canonical correlations and the first direction vectors. We applied this model on a genomics dataset arising from a breast cancer study.

The third chapter investigates some theoretical results relating to Bayesian analysis of the von Mises Fisher distribution and its mixture model. In this chapter we developed a novel prior for the distribution, provided a strategy for interpretation of its hyperparameters, and proved different properties of the distribution. In the second half of this project we provide a data augmentation algorithm which removed the intractability from the posterior distribution of concentration parameter. We implemented the developed methodologies for finite mixture models. We compared our developed models with some competing methods as well as assessed the accuracy of estimation of parameters through simulation studies. We later applied this model on diffusion tensor imaging data to understand the brain structure through the clustering of the voxels.

For the CCA project, future directions for research include further refining the algorithms used to improve scalability for analysis of the very high dimensional data. An additional avenue for extension is to generalize this methodology for three or more views. This project also introduced a novel multiplicative prior process based on half-Cauchy distribution. Establishing theoretical properties of this process is a challenging task, and one considered for future work.

For the vMF project, a possible future direction could be development of novel conjugate prior as well as data augmentation algorithm for more complex extensions such as directional regression, time series, etc. Furthermore, it would be interesting to develop these models in the context of infinite mixture model. A theoretical study of the convergence of MCMC can be conducted, and we have already taken first steps in proving geometric erogodicity of the discussed data augmentation algorithm. Penalized clustering on hyper-sphere could be an interesting direction in which this work can be extended.

To conclude, this dissertation work opens up new avenues for theoretical study in Bayesian methodologies for constrained spaces.

REFERENCES

Milton Abramowitz and Irene A Stegun. *Handbook of Mathematical Functions: With Formulas, Graphs, and Mathematical Tables*, volume 55. Courier Corporation, 1965.

Yannis Agiomyrgiannakis and Yannis Stylianou. Wrapped gaussian mixture models for modeling and high-rate quantization of phase data of speech. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(4):775–786, 2009.

Jesper LR Andersson, Mark Jenkinson, Stephen Smith, et al. Non-linear registration, aka spatial normalisation fmrib technical report tr07ja2. *FMRIB Analysis Group of the University of Oxford*, 2(1):e21, 2007.

Cédric Archambeau and Francis R Bach. Sparse probabilistic projections. In *Advances in Neural Information Processing Systems*, pages 73–80, 2009.

Francis R Bach and Michael I Jordan. A probabilistic interpretation of canonical correlation analysis. 2005.

Arindam Banerjee, Inderjit Dhillon, Joydeep Ghosh, and Suvrit Sra. Generative model-based clustering of directional data. In *Proceedings of the Ninth ACM SIGKDD International Conference On Knowledge Discovery and Data Mining*, pages 19–28, 2003.

Arindam Banerjee, Inderjit S Dhillon, Joydeep Ghosh, Suvrit Sra, and Greg Ridgeway. Clustering on the unit hypersphere using von mises-fisher distributions. *Journal of Machine Learning Research*, 6(9), 2005.

Mark Bangert, Philipp Hennig, and Uwe Oelfke. Using an infinite von mises-fisher mixture model to cluster treatment beam directions in external radiation therapy.

In *2010 Ninth International Conference on Machine Learning and Applications*, pages 746–751. IEEE, 2010.

Anirban Bhattacharya and David B Dunson. Sparse bayesian infinite factor models. *Biometrika*, pages 291–306, 2011.

Christopher Bingham. *Distributions On the Sphere and on the Projective Plane*. Yale University, 1964.

Christopher M Bishop and Nasser M Nasrabadi. *Pattern Recognition and Machine Learning*, volume 4. Springer, 2006.

Ryan P Cabeen and David H Laidlaw. White matter supervoxel segmentation by axial dp-means clustering. In *International MICCAI Workshop on Medical Computer Vision*, pages 95–104. Springer, 2013.

Carlos M Carvalho, Jeffrey Chang, Joseph E Lucas, Joseph R Nevins, Quanli Wang, and Mike West. High-dimensional sparse factor modeling: Applications in gene expression genomics. *Journal of the American Statistical Association*, 103(484): 1438–1456, 2008.

Carlos M Carvalho, Nicholas G Polson, and James G Scott. The horseshoe estimator for sparse signals. *Biometrika*, 97(2):465–480, 2010.

Alissa Castleberry. *Integrated Analysis of Multi-Omics Data Using Sparse Canonical Correlation Analysis*. PhD thesis, The Ohio State University, 2019.

Gilles Celeux, Florence Forbes, Christian P Robert, and D Mike Titterington. Deviance information criteria for missing data models. *Bayesian Analysis*, 1(4):651–673, 2006.

Jun Chen, Frederic D Bushman, James D Lewis, Gary D Wu, and Hongzhe Li.

Structure-constrained sparse canonical correlation analysis with an application to microbiome data analysis. *Biostatistics*, 14(2):244–258, 2013.

Xi Chen and Han Liu. An efficient optimization algorithm for structured sparse CCA, with applications to eqtl mapping. *Statistics in Biosciences*, 4(1):3–26, 2012.

Xi Chen, Liu Han, and Jaime Carbonell. Structured sparse canonical correlation analysis. In *Artificial Intelligence and Statistics*, pages 199–207, 2012.

Koei Chin, Sandy DeVries, Jane Fridlyand, Paul T Spellman, Ritu Roydasgupta, Wen-Lin Kuo, Anna Lapuk, Richard M Neve, Zuwei Qian, Tom Ryder, et al. Genomic and transcriptional aberrations linked to breast cancer pathophysiologies. *Cancer Cell*, 10(6):529–541, 2006.

Paul Damien and Stephen Walker. A full Bayesian analysis of circular data using the von Mises distribution. *The Canadian Journal of Statistics*, pages 291–298, 1999.

Inderjit S Dhillon and Dharmendra S Modha. Concept decompositions for large sparse text data using clustering. *Machine Learning*, 42(1-2):143–175, 2001.

Inderjit S Dhillon and Suvrit Sra. Modeling data using directional distributions. Technical report, Citeseer, 2003.

Persi Diaconis and Donald Ylvisaker. Conjugate priors for exponential families. *The Annals of Statistics*, pages 269–281, 1979.

Jean-Luc Dortet-Bernadet and Nicolas Wicker. Model-based clustering on the unit sphere with an illustration using gene expression profiles. *Biostatistics*, 9(1):66–80, 2008.

Lei Du, Jingwen Yan, Sungeun Kim, Shannon L Risacher, Heng Huang, Mark Inlow, Jason H Moore, Andrew J Saykin, Li Shen, et al. GN-SCCA: Graphnet based sparse canonical correlation analysis for brain imaging genetics. In *International Conference on Brain Informatics and Health*, pages 275–284. Springer, 2015.

Lei Du, Heng Huang, Jingwen Yan, Sungeun Kim, Shannon L Risacher, Mark Inlow, Jason H Moore, Andrew J Saykin, Li Shen, and Alzheimer's Disease Neuroimaging Initiative. Structured sparse canonical correlation analysis for brain imaging genetics: An improved graphnet method. *Bioinformatics*, 32(10):1544–1551, 2016.

Carlos Esteves, Christine Allen-Blanchette, Ameesh Makadia, and Kostas Daniilidis. Learning so (3) equivariant representations with spherical cnns. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 52–68, 2018.

J Ferlay, M Ervik, F Lam, M Colombet, L Mery, M Piñeros, A Znaor, I Soerjomataram, and F Bray. Global cancer observatory: Cancer today. lyon: International agency for research on cancer; 2018, 2020.

Adelaide Figueiredo. Clustering directions based on the estimation of a mixture of von Mises-Fisher distributions. *The Open Statistics & Probability Journal*, 8(1), 2017.

Sylvia Frühwirth-Schnatter. Markov Chain Monte Carlo estimation of classical and dynamic switching and mixture models. *Journal of the American Statistical Association*, 96(453):194–209, 2001.

Jeremy Gaskins. Hyper markov laws for correlation matrices. *Statistica Sinica*, 29 (1):165–184, 2019.

Andrew Gelman, Jessica Hwang, and Aki Vehtari. Understanding predictive information criteria for bayesian models. *Statistics and computing*, 24(6):997–1016, 2014.

John Geweke and Guofu Zhou. Measuring the pricing error of the arbitrage pricing theory. *The Review of Financial Studies*, 9(2):557–587, 1996.

Walter R Gilks and Pascal Wild. Adaptive rejection sampling for Gibbs sampling. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 41(2):337–348, 1992.

G Russell Glenn, Joseph A Helpern, Ali Tabesh, and Jens H Jensen. Quantitative assessment of diffusional kurtosis anisotropy. *NMR in Biomedicine*, 28(4):448–459, 2015.

M Goldstein and RM Thaler. Recurrence techniques for the calculation of Bessel functions. *Mathematics of Computation*, 13(66):102–108, 1959.

Ignacio González, Sébastien Déjean, Pascal Martin, and Alain Baccini. CCA: An R package to extend canonical correlation analysis. *Journal of Statistical Software*, 23(12):1–14, 2008.

Siddharth Gopal and Yiming Yang. vonMises Fisher clustering models. In *International Conference on Machine Learning*, pages 154–162, 2014.

Peter Guttorp and Richard A Lockhart. Finding the location of a signal: A Bayesian analysis. *Journal of the American Statistical Association*, 83(402):322–330, 1988.

John A Hartigan and Manchek A Wong. A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series c (Applied Statistics)*, 28(1):100–108, 1979.

Xiaofei He, Deng Cai, Yuanlong Shao, Hujun Bao, and Jiawei Han. Laplacian regularized Gaussian mixture model for data clustering. *IEEE Transactions on Knowledge and Data Engineering*, 23(9):1406–1418, 2010.

Daniel Hernandez-Stumpfhauser, F Jay Breidt, and Mark J van der Woerd. The general projected normal distribution of arbitrary dimension: Modeling and Bayesian inference. *Bayesian Analysis*, 12(1):113–133, 2017.

James P Hobert. The data augmentation algorithm: Theory and methodology. *Handbook of Markov Chain Monte Carlo*, pages 253–293, 2011.

K Hornik and B Grün. On conjugate families and Jeffreys priors for von Mises-Fisher distributions. *J Stat Plan Inference*, 143(5):992–999, May 2013.

Kurt Hornik and Bettina Grün. *movMF: Mixtures of von Mises-Fisher Distributions*, 2022. URL https://CRAN.R-project.org/package=movMF. R package version 0.2-7.

Kurt Hornik, Ingo Feinerer, Martin Kober, and Christian Buchta. Spherical k-means clustering. *Journal of Statistical Software*, 50(10):1–22, 2012. doi: 10.18637/jss.v050.i10.

Harold Hotelling. Relations between two sets of variates. *Biometrika*, 28(3/4):321–377, 1936. ISSN 00063444. URL http://www.jstor.org/stable/2333955.

Elizabeth Hyman, Päivikki Kauraniemi, Sampsa Hautaniemi, Maija Wolf, Spyro Mousses, Ester Rozenblum, Markus Ringnér, Guido Sauter, Outi Monni, Abdel Elkahloun, et al. Impact of DNA amplification on gene expression patterns in breast cancer. *Cancer Research*, 62(21):6240–6245, 2002.

Ajay Jasra, Chris C Holmes, and David A Stephens. Markov Chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling. *Statistical Science*, 20(1):50–67, 2005.

Mark Jenkinson, Christian F Beckmann, Timothy EJ Behrens, Mark W Woolrich, and Stephen M Smith. Fsl. *Neuroimage*, 62(2):782–790, 2012.

Heidi Johansen-Berg and Timothy EJ Behrens. *Diffusion MRI: from quantitative measurement to in vivo neuroanatomy*. Academic Press, 2013.

R.A. Johnson and D.W. Wichern. *Applied Multivariate Statistical Analysis*. Applied Multivariate Statistical Analysis. Pearson Prentice Hall, 2007.

AL Jones. An extension of an inequality involving modified bessel functions. *Journal of Mathematics and Physics*, 47(1-4):220–221, 1968.

Cari G Kaufman, Valérie Ventura, and Robert E Kass. Spline-based non-parametric regression for periodic functions and its application to directional tuning of neurons. *Statistics in Medicine*, 24(14):2255–2265, 2005.

John T Kent. The Fisher-Bingham distribution on the sphere. *Journal of the Royal Statistical Society: Series B (Methodological)*, 44(1):71–80, 1982.

Arto Klami and Samuel Kaski. Local dependent components. In *Proceedings of the 24th International Conference on Machine Learning*, pages 425–432, 2007.

Arto Klami, Seppo Virtanen, and Samuel Kaski. Bayesian canonical correlation analysis. *Journal of Machine Learning Research*, 14(Apr):965–1003, 2013.

Akira Kunimatsu, Shigeki Aoki, Yoshitaka Masutani, Osamu Abe, Naoto Hayashi, Harushi Mori, Tomohiko Masumoto, and Kuni Ohtomo. The optimal trackability threshold of fractional anisotropy for diffusion tensor tractography of the corticospinal tract. *Magnetic Resonance in Medical Sciences*, 3(1):11–17, 2004.

Danial Lashkari, Ed Vul, Nancy Kanwisher, and Polina Golland. Discovering structure in the space of FMRI selectivity profiles. *Neuroimage*, 50(3):1085–1098, 2010.

Sirio Legramanti, Daniele Durante, and David B Dunson. Bayesian cumulative shrinkage for infinite factorizations. *Biometrika*, 107(3):745–752, 2020.

Eemeli Leppäaho, Muhammad Ammad-ud din, and Samuel Kaski. GFA: exploratory analysis of multiple data sources with group factor analysis. *The Journal of Machine Learning Research*, 18(1):1294–1298, 2017.

Christophe Ley and Thomas Verdebout. *Modern Directional Statistics*. Chapman and Hall/CRC, 2017.

Yi-Ou Li, Tülay Adali, Wei Wang, and Vince D Calhoun. Joint blind source separation by multiset canonical correlation analysis. *IEEE Transactions on Signal Processing*, 57(10):3918–3929, 2009.

Yunfan Li, Bruce A Craig, and Anindya Bhadra. The graphical horseshoe estimator for inverse covariance matrices. *Journal of Computational and Graphical Statistics*, 28(3):747–757, 2019.

Dongdong Lin, Jigang Zhang, Jingyao Li, Vince D Calhoun, Hong-Wen Deng, and Yu-Ping Wang. Group sparse canonical correlation analysis for genomic data integration. *BMC Bioinformatics*, 14(1):1–16, 2013.

Dongdong Lin, Vince D Calhoun, and Yu-Ping Wang. Correspondence between fmri and snp data by group sparse canonical correlation analysis. *Medical Image Analysis*, 18(6):891–902, 2014.

Zhonglin Lin, Changshui Zhang, Wei Wu, and Xiaorong Gao. Frequency recognition based on canonical correlation analysis for SSVEP-based BCIs. *IEEE Transactions on Biomedical Engineering*, 53(12):2610–2614, 2006.

Hedibert Freitas Lopes and Mike West. Bayesian model assessment in factor analysis. *Statistica Sinica*, pages 41–67, 2004.

Yudell L Luke. Inequalities for generalized hypergeometric functions. *Journal of Approximation Theory*, 5(1):41–65, 1972.

Enes Makalic and Daniel F Schmidt. A simple sampler for the horseshoe estimator. *IEEE Signal Processing Letters*, 23(1):179–182, 2015.

Claudia Manzoni, Demis A Kia, Jana Vandrovcova, John Hardy, Nicholas W Wood, Patrick A Lewis, and Raffaele Ferrari. Genome, transcriptome and proteome: the rise of omics data and their integration in biomedical sciences. *Briefings in Bioinformatics*, 19(2):286–302, 2018.

Kanti V Mardia. Statistics of directional data. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 349–393, 1975.

Kanti V Mardia and SAM El-Atoum. Bayesian inference for the von Mises-Fisher distribution. *Biometrika*, 63(1):203–206, 1976a.

Kanti V Mardia and SAM El-Atoum. Bayesian inference for the von Mises-Fisher distribution. *Biometrika*, 63(1):203–206, 1976b.

Kanti V Mardia and Peter E Jupp. *Directional Statistics*, volume 494. John Wiley & Sons, 2009.

Kanti V Mardia, Jesper Illemann Foldager, and Jes Frellsen. Directional statistics in protein bioinformatics. In *Applied Directional Statistics*, pages 17–40. Chapman and Hall/CRC, 2018.

KV Mardia and PE Jupp. Directional statistics. john willey and sons. *Inc., Chichester*, 2000.

Jean-Michel Marin, Christian P Robert, et al. *Bayesian core: A Practical approach to Computational Bayesian Statistics*, volume 268. Springer, 2007.

Domenico Marinucci and Giovanni Peccati. *Random Fields on the Sphere: Representation, Limit Theorems and Cosmological Applications*, volume 389. Cambridge University Press, 2011.

Tim McGraw, Baba C Vemuri, Bob Yezierski, and Thomas Mareci. Von mises-fisher mixture model of the diffusion odf. In *3rd IEEE International Symposium on Biomedical Imaging: Nano to Macro, 2006.*, pages 65–68. IEEE, 2006.

Jeffrey W Miller and Matthew T Harrison. Mixture models with a prior on the number of components. *Journal of the American Statistical Association*, 113(521): 340–356, 2018.

Zohre Momenimovahed and Hamid Salehiniya. Epidemiological characteristics of and risk factors for breast cancer in the world. *Breast Cancer: Targets and Therapy*, 11:151, 2019.

Jeffrey S Morris and Veerabhadran Baladandayuthapani. Statistical contributions to bioinformatics: Design, modelling, structure learning and integration. *Statistical Modelling*, 17(4-5):245–289, 2017.

Thanit Nanthanasub, Boriboon Novaprateep, and Narongpol Wichailukkana. The logarithmic concavity of modified Bessel functions of the first kind and its related functions. *Advances in Difference Equations*, 2019(1):1–14, 2019.

Gabriel Nunez-Antonio and Eduardo Gutiérrez-Pena. A bayesian analysis of directional data using the von mises–fisher distribution. *Communications in Statistics—Simulation and Computation®*, 34(4):989–999, 2005.

Subhadip Pal, Subhajit Sengupta, Riten Mitra, and Arunava Banerjee. Conjugate priors and posterior inference for the matrix langevin distribution on the stiefel manifold. *Bayesian Analysis*, 15(3):871–908, 2020.

Panagiotis Papastamoulis. Handling the label switching problem in latent class models via the ECR algorithm. *Communications in Statistics-Simulation and Computation*, 43(4):913–927, 2014.

Panagiotis Papastamoulis. label.switching: An r package for dealing with the label switching problem in mcmc outputs. *Journal of Statistical Software, Code Snippets*, 69(1):1–24, 2016. doi: 10.18637/jss.v069.c01. URL https://www.jstatsoft.org/index.php/jss/article/view/v069c01.

Arturo Pardo, Eusebio Real, Venkat Krishnaswamy, José Miguel López-Higuera, Brian W Pogue, and Olga M Conde. Directional kernel density estimation for classification of breast tissue spectra. *IEEE Transactions on Medical Imaging*, 36 (1):64–73, 2016.

Elena Parkhomenko, David Tritchler, and Joseph Beyene. Sparse canonical correlation analysis with application to genomic data integration. *Statistical Applications in Genetics and Molecular Biology*, 8(1), 2009.

David Peel, William J Whiten, and Geoffrey J McLachlan. Fitting mixtures of kent distributions to aid in joint set identification. *Journal of the American Statistical Association*, 96(453):56–63, 2001.

Arthur Pewsey and Eduardo García-Portugués. Recent advances in directional statistics. *Test*, 30(1):1–58, 2021.

Jonathan R Pollack, Therese Sørlie, Charles M Perou, Christian A Rees, Stefanie S Jeffrey, Per E Lonning, Robert Tibshirani, David Botstein, Anne-Lise Børresen-Dale, and Patrick O Brown. Microarray analysis reveals a major direct role of dna copy number alteration in the transcriptional program of human breast tumors. *Proceedings of the National Academy of Sciences*, 99(20):12963–12968, 2002.

Xiangju Qin, Pádraig Cunningham, and Michael Salter-Townshend. Online trans-dimensional von mises-fisher mixture models for user profiles. *The Journal of Machine Learning Research*, 17(1):7021–7071, 2016.

William M Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336):846–850, 1971.

Richard A Redner and Homer F Walker. Mixture densities, maximum likelihood and the EM algorithm. *SIAM Review*, 26(2):195–239, 1984.

Joseph Reisinger, Austin Waters, Bryan Silverthorn, and Raymond J Mooney. Spherical topic models. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 903–910, 2010.

Christian P Robert. *The Bayesian choice: From Decision-Theoretic Foundations to Computational Implementation*, volume 2. Springer, 2007.

Gareth Roberts and Jeffrey Rosenthal. Geometric ergodicity and hybrid markov chains. *Electronic Communications in Probability*, 2:13–25, 1997.

Carlos E Rodríguez, Gabriel Núñez-Antonio, and Gabriel Escarela. A bayesian mixture model for clustering circular data. *Computational Statistics & Data Analysis*, 143:106842, 2020.

Daniel Rueckert, Luke I Sonoda, Carmel Hayes, Derek LG Hill, Martin O Leach, and David J Hawkes. Nonrigid registration using free-form deformations: application to breast mr images. *IEEE transactions on medical imaging*, 18(8):712–721, 1999.

Srikanth Ryali, Tianwen Chen, Kaustubh Supekar, and Vinod Menon. A parcellation scheme based on von mises-fisher distributions and markov random fields for segmenting brain regions using resting-state fmri. *Neuroimage*, 65:83–96, 2013.

Lorenzo Schiavon, Antonio Canale, and David B Dunson. Generalized infinite factorization models. *arXiv preprint arXiv:2103.10333*, 2021.

Gideon Schwarz. Estimating the Dimension of a Model. *The Annals of Statistics*, 6 (2):461 – 464, 1978. doi: 10.1214/aos/1176344136.

Michael Sekula, Jeremy Gaskins, and Susmita Datta. Single-cell differential network analysis with sparse Bayesian factor models. *Frontiers in Genetics*, 12:810816–810816, 2021.

Abhishek Sharma, Abhishek Kumar, Hal Daume, and David W Jacobs. Generalized multiview analysis: A discriminative latent space. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2160–2167. IEEE, 2012.

Adelino R Silva. Probabilistic tractography using particle filtering and clustered directional data. In *Advances in Neurotechnology, Electronics and Informatics*, pages 47–62. Springer, 2016.

Stephen M Smith. Fast robust automated brain extraction. *Human Brain Mapping*, 17(3):143–155, 2002.

Stephen M. Smith, Mark Jenkinson, Heidi Johansen-Berg, Daniel Rueckert, Thomas E. Nichols, Clare E. Mackay, Kate E. Watkins, Olga Ciccarelli, M. Zaheer Cader, Paul M. Matthews, and Timothy E.J. Behrens. Tract-based spatial statistics: Voxelwise analysis of multi-subject diffusion data. *NeuroImage*, 31(4): 1487–1505, 2006.

Matthew Sperrin, Thomas Jaki, and Ernst Wit. Probabilistic relabelling strategies for the label switching problem in Bayesian mixture models. *Statistics and Computing*, 20(3):357–366, 2010.

Suvrit Sra. Directional statistics in machine learning: a brief review. *Applied Directional Statistics: Modern Methods and Case Studies*, 225:6, 2018.

M Statheropoulos, N Vassiliadis, and A Pappa. Principal component and canonical correlation analysis for examining air pollution and meteorological data. *Atmospheric Environment*, 32(6):1087–1095, 1998.

Matthew Stephens. Dealing with label switching in mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(4):795–809, 2000.

Hyuna Sung, Jacques Ferlay, Rebecca L Siegel, Mathieu Laversanne, Isabelle Soerjomataram, Ahmedin Jemal, and Freddie Bray. Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal For Clinicians*, 71(3):209–249, 2021.

Xiaotong Suo, Victor Minden, Bradley Nelson, Robert Tibshirani, and Michael Saunders. Sparse canonical correlation analysis. *arXiv preprint arXiv:1705.10865*, 2017.

Jalil Taghia, Zhanyu Ma, and Arne Leijon. Bayesian estimation of the von-mises fisher mixture model with variational inference. *IEEE Transactions On Pattern Analysis and Machine Intelligence*, 36(9):1701–1715, 2014.

Martin A Tanner and Wing Hung Wong. The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, 82(398): 528–540, 1987.

Ledyard R Tucker. An inter-battery method of factor analysis. *Psychometrika*, 23 (2):111–136, 1958.

Stéphanie van der Pas, Botond Szabó, and Aad van der Vaart. Uncertainty quantification for the horseshoe (with discussion). *Bayesian Analysis*, 12(4):1221–1274, 2017.

Hrishikesh D Vinod. Canonical ridge and econometrics of joint production. *Journal of Econometrics*, 4(2):147–166, 1976.

Sandra Waaijenborg, Philip C Verselewel de Witt Hamer, and Aeilko H Zwinderman. Quantifying the association between gene expressions and DNA-markers by penalized canonical correlation analysis. *Statistical Applications in Genetics and Molecular Biology*, 7(1), 2008.

Chong Wang. Variational bayesian approach to canonical correlation analysis. *IEEE Transactions on Neural Networks*, 18(3):905–910, 2007.

Fangpo Wang and Alan E Gelfand. Directional data analysis under the general projected normal distribution. *Statistical Methodology*, 10(1):113–127, 2013.

Fangpo Wang and Alan E Gelfand. Modeling space and space-time directional data using projected Gaussian processes. *Journal of the American Statistical Association*, 109(508):1565–1580, 2014.

Yixin Wang and David M Blei. Frequentist consistency of variational Bayes. *Journal of the American Statistical Association*, 114(527):1147–1161, 2019.

Daniela M Witten and Robert J Tibshirani. Extensions of sparse canonical correlation analysis with applications to genomic data. *Statistical Applications in Genetics and Molecular Biology*, 8(1), 2009.

Jingwen Yan, Lei Du, Sungeun Kim, Shannon L Risacher, Heng Huang, Jason H Moore, Andrew J Saykin, Li Shen, and Alzheimer's Disease Neuroimaging Initiative. Transcriptome-guided amyloid imaging genetic analysis via a novel structured sparse learning algorithm. *Bioinformatics*, 30(17):i564–i571, 2014.

Linxiao Yang, Jun Fang, Huiping Duan, Hongbin Li, and Bing Zeng. Fast low-rank Bayesian matrix completion with hierarchical Gaussian prior models. *IEEE Transactions on Signal Processing*, 66(11):2804–2817, 2018.

Xinghao Yang, Liu Weifeng, Wei Liu, and Dacheng Tao. A survey on canonical correlation analysis. *IEEE Transactions on Knowledge and Data Engineering*, 2019.

Zhen-Hang Yang, Jing-Feng Tian, and Ya-Ru Zhu. New sharp bounds for the modified Bessel function of the first kind and Toader-Qi mean. *Mathematics*, 8(6):901, 2020.

Yu Zhang, Guoxu Zhou, Jing Jin, Minjue Wang, Xingyu Wang, and Andrzej Cichocki. L1-regularized multiway canonical correlation analysis for ssvep-based bci. *IEEE Transactions on Neural Systems and Rehabilitation engineering*, 21(6):887–896, 2013.

YU Zhang, Guoxu Zhou, Jing Jin, Xingyu Wang, and Andrzej Cichocki. Frequency recognition in ssvep-based bci using multiset canonical correlation analysis. *International Journal of Neural Systems*, 24(04):1450013, 2014.

Shiwen Zhao, Chuan Gao, Sayan Mukherjee, and Barbara E Engelhardt. Bayesian group factor analysis with structured sparsity. *The Journal of Machine Learning Research*, 17(1):6868–6914, 2016.

Xiaowei Zhuang, Zhengshi Yang, and Dietmar Cordes. A technical review of canonical correlation analysis for neuroscience applications. *Human Brain Mapping*, 41(13): 3807–3833, 2020.

APPENDIX

This section includes the theoretical results and proofs for Chapter 3.

A Some Results for Bessel Function of First Kind $I_\nu(\cdot)$

In this section we state and recall some crucial properties of the Bessel function of the first kind $I_\nu(\cdot)$ and the ratio of the Bessel function of the first kind. We will utilize the further results to develop required theory. We recall following results from Yang et al. [2020].

For $\nu = 0$, bounds for $I_\nu(\cdot)$ are given such that

$$\sqrt{\frac{\sinh(x)}{x}\left[\cosh(qx)\right]^{\frac{1}{q}}} \leq I_0(x) \leq \sqrt{\frac{\sinh(x)}{x}\left[\cosh(px)\right]^{\frac{1}{p}}}, \tag{28}$$

for $p \geq \frac{2}{3}$ and $q \leq \frac{\log(2)}{\log(\pi)}$. Further, we state the important recurrence relation for the differentiation of $I_\nu(\cdot)$ [Goldstein and Thaler, 1959]:

$$\begin{aligned}
\frac{\partial}{\partial x} I_\nu(x) &= \frac{1}{2}\left[I_{\nu-1}(x) + I_{\nu+1}(x)\right] \\
\frac{\partial}{\partial x} I_\nu(x) &= \frac{\nu}{x} I_\nu(x) + I_{\nu+1}(x) \\
\frac{\partial}{\partial x} I_\nu(x) &= -\frac{\nu}{x} I_\nu(x) + I_{\nu-1}(x). \qquad (29)
\end{aligned}$$

We also state an useful approximation for large arguments [Abramowitz and Stegun, 1965]

$$I_\nu(x) \approx \frac{e^x}{\sqrt{2\pi x}}. \tag{30}$$

Similarly for small arguments,

$$I_\nu(x) \approx \frac{x^\nu}{2^\nu \Gamma(\nu+1)} \tag{31}$$

B Lemmas and Their Proofs

The following Lemmas are stated and proved to assess the properties of ratio of the Bessel functions. Later some theorems are build upon these Lemmas.

Lemma 1. *$x \longmapsto \frac{I_{\nu+1}(x)}{I_\nu(x)}$ is increasing function.*

Proof. 1 This result is proven in Jones [1968]. □

Lemma 2. *$\frac{I_{\nu+1}(\kappa)}{I_\nu(\kappa)} > \frac{I_{\nu+2}(\kappa)}{I_{\nu+1}(\kappa)}$ for all $\kappa > 0$, $\nu > 0$. The result is equivalent to $[I_{\nu+1}(\kappa)]^2 > I_{\nu+2}(\kappa)\ I_\nu(\kappa)$.*

Proof. 2 This result is called log concavity of ratio of Bessel function. Theorem 3 of Nanthanasub et al. [2019] proves this result. □

Lemma 3. *Let b, κ be positive real numbers and n be a positive integer. If $I_\nu(\cdot)$ denotes the modified Bessel function of the first kind of order $\nu \geq 0$, then the function $h(b) := \frac{1}{b}\frac{I_{\nu+1}(b\kappa)}{I_\nu(b\kappa)}$ is a decreasing function.*

Proof. 3 From the definition of the function $h(b)$, it follows that

$$\begin{aligned}
\frac{\partial}{\partial b}\log(h(b)) &= -\frac{1}{b} + \kappa\frac{I'_{\nu+1}(b\kappa)}{I_{\nu+1}(b\kappa)} - \kappa\frac{I'_\nu(b\kappa)}{I_\nu(b\kappa)} \\
&\overset{(\star)}{=} -\frac{1}{b} + \kappa\frac{\{I'_{\nu+1}(b\kappa)\nu+2\}}{I_{\nu+1}(b\kappa)} - \kappa\frac{\{I'_\nu(b\kappa)\nu+1\}}{I_\nu(b\kappa)} \\
&= \kappa\left\{\frac{I_{\nu+2}(b\kappa)}{I_{\nu+1}(b\kappa)} - \frac{I_{\nu+1}(b\kappa)}{I_\nu(b\kappa)}\right\} \\
&< 0,
\end{aligned} \tag{32}$$

where the equality in $(\star)$ is due to the recurrence relation of the derivative of the Modified Bessel function of the first kind [Abramowitz and Stegun, 1965]. The inequality in (32) follows from the log concavity of ratio of modified Bessel function of the first kind [Nanthanasub et al., 2019, Theorem 3] along with the fact that $\kappa > 0$. The statement of the lemma is immediate from (32). □

Lemma 4. *Let a, κ be positive real numbers and n be a positive integer and $\nu \geq 0$. Consider the function $g(a) := a\frac{I_{\nu+1}(an\kappa)}{I_{\nu}(an\kappa)}$ where $I_{\nu}(\cdot)$ denotes the modified Bessel functions of the first kind. Then,*

1. *the function $a \mapsto g(a)$ is an increasing function for all $a > 0$.*

2. *If $na^2 \leq 1$, then, $g(a) \leq \frac{I_{\nu+1}(\kappa)}{I_{\nu}(\kappa)}$ for all $\kappa > 0$.*

Proof. 1. According to the definition of the function $g(a)$, it follows that,

$$\begin{aligned}\frac{\partial}{\partial a}\log(g(a)) &= \frac{1}{a} + n\kappa\left\{\frac{I'_{\nu+1}(an\kappa)}{I_{\nu+1}(an\kappa)} - \frac{I'_{\nu}(an\kappa)}{I_{\nu}(an\kappa)}\right\} \\ &= \frac{1}{a} + n\kappa\left\{\frac{I''_{\nu+1}(an\kappa)\nu}{I_{\nu+1}(an\kappa)} - \frac{I''_{\nu}(an\kappa)\nu - 1}{I_{\nu}(an\kappa)}\right\} \\ &= n\kappa\left\{\frac{I_{\nu}(an\kappa)}{I_{\nu+1}(an\kappa)} - \frac{I_{\nu-1}(an\kappa)}{I_{\nu}(an\kappa)}\right\} \\ &> 0, \end{aligned} \tag{33}$$

where the last inequility follows as $I^2_{\nu}(x) > I_{\nu-1}(x)I_{\nu+1}(x)$ for all $\nu \geq 0$ and $x > 0$ [Nanthanasub et al., 2019, Theorem 3]. Therefore, the function $a \mapsto g(a)$ is an increasing function for $a > 0$.

2. Based on the assumption $na^2 \leq 1$ in the statement of the lemma, we have $a \leq \frac{1}{\sqrt{n}}$. As the function $a \mapsto g(a)$ is increasing (from part(a) of Lemma 4), it appears that

$$g(a) \leq g\left(\frac{1}{\sqrt{n}}\right) = \frac{1}{\sqrt{n}}\frac{I_{\nu+1}(\sqrt{n}\kappa)}{I_{\nu}(\sqrt{n}\kappa)}. \tag{34}$$

Additionally, from Lemma 3 the function $h(b) = \frac{1}{b}\frac{I_{\nu+1}(b\kappa)}{I_{\nu}(b\kappa)}$ is a decreasing function for all $\nu \geq 0$ and $\kappa > 0$. Therefore, for all $n > 1$,

$$\frac{1}{\sqrt{n}} \frac{I_{\nu+1}(\sqrt{n}\kappa)}{I_\nu(\sqrt{n}\kappa)} = h(\sqrt{n}) < h(1) = \frac{I_{\nu+1}(\kappa)}{I_\nu(\kappa)}. \tag{35}$$

The claim $g(a) \leq \frac{I_{\nu+1}(\kappa)}{I_\nu(\kappa)}$ then follows from (34) and (35).

□

Lemma 5. *Let κ, a be positive real numbers. Let $I_\nu(\cdot)$ denote the modified Bessel function of the first kind, then the function $h_1(a) := a\frac{I_{\nu+2}(a\kappa)}{I_{\nu+1}(a\kappa)} - a\frac{I_{\nu+1}(a\kappa)}{I_\nu(a\kappa)}$ is a decreasing function for all $\kappa > 0$.*

Proof. Using the notation $R_\nu(x) = \frac{I_{\nu+1}(x)}{I_\nu(x)}$, we obtain from the part (b) of Lemma 4, that

$$\frac{\partial h_1(a)}{\partial a} = \frac{1}{x}\frac{\partial}{\partial a}\left[ax\left\{R_{\nu+1}(ax) - R_\nu(ax)\right\}\right] < 0. \tag{36}$$

Therefore, the function $a \mapsto h_1(a)$ is a decreasing function. □

Lemma 6. *Let a be a positive real number. Consider the function*

$$g_a(\kappa) := \frac{I_{\nu+1}(\kappa) I_\nu(a\kappa)}{I_\nu(\kappa) I_{\nu+1}(a\kappa)},$$

for $\kappa > 0$. The function $g_a(\cdot)$ has the following properties:

1. *The function $\kappa \mapsto g_a(\kappa)$ is an increasing function when $a > 1$.*

2. *If $a > 1$, then $\frac{1}{a} \leq g_a(\kappa) \leq 1$.*

3. *Let n be a positive integer. If $n < a^2 \leq n^2$, then there exists a unique positive number $\kappa^\star$ such that $g_a(\kappa^\star) = \frac{a}{n}$.*

Proof. 1. If we denote

$$h_1(a) := a\frac{I_{\nu+2}(a\kappa)}{I_{\nu+1}(a\kappa)} - a\frac{I_{\nu+1}(a\kappa)}{I_\nu(a\kappa)},$$

then from the definition of $g_a(\kappa)$, we obtain that

$$\begin{aligned}
\frac{\partial}{\partial\kappa}\log\left(g_a(\kappa)\right\} &= \frac{I'_{\nu+1}(\kappa)}{I_{\nu+1}(\kappa)} + a\frac{I'_{\nu}(a\kappa)}{I_{\nu}(a\kappa)} - \frac{I'_{\nu}(\kappa)}{I_{\nu}(\kappa)} - a\frac{I'_{\nu+1}(a\kappa)}{I_{\nu+1}(a\kappa)} \\
&= \frac{I'_{\nu+1}(\kappa)\nu+2}{I_{\nu+1}(\kappa)} - \frac{I'_{\nu}(\kappa)\nu+1}{I_{\nu}(\kappa)} \\
&\qquad + a\left\{\frac{I'_{\nu}(a\kappa)\nu+1}{I_{\nu}(a\kappa)} - \frac{I'_{\nu+1}(a\kappa)\nu+2}{I_{\nu+1}(a\kappa)}\right\} \\
&= \left\{\frac{I_{\nu+2}(\kappa)}{I_{\nu+1}(\kappa)} - \frac{I_{\nu+1}(\kappa)}{I_{\nu}(\kappa)}\right\} - a\left\{\frac{I_{\nu+2}(a\kappa)}{I_{\nu+1}(a\kappa)} - \frac{I_{\nu+1}(a\kappa)}{I_{\nu}(a\kappa)}\right\} \\
&= h_1(1) - h_1(a) \\
&> 0,
\end{aligned}$$

where the last inequality follows from the assumption that $a > 1$ and the fact that the function $a \mapsto h_1(a)$ is decreasing (see Lemma 5).

2. If we assume $a > 1$, then from part (a) we get that the function $\kappa \mapsto g_a(\kappa)$ is an increasing function. Therefore,

$$g_a(\kappa) \leq \lim_{\kappa\to\infty} g_a(\kappa) = \lim_{\kappa\to\infty} \frac{I_{\nu+1}(\kappa)I_{\nu}(a\kappa)}{I_{\nu}(\kappa)I_{\nu+1}(a\kappa)}.$$

Applying the asymptotic approximation $\lim_{x\to\infty}\sqrt{2\pi x}\,e^{-x}I_{\alpha}(x) = 1$ for $\alpha \geq 0$ [Abramowitz and Stegun, 1965], it follows from (37) that

$$g_a(\kappa) \leq \lim_{\kappa\to\infty} \frac{\sqrt{2\pi\kappa}e^{-\kappa}I_{\nu+1}(\kappa)}{\sqrt{2\pi\kappa}e^{-\kappa}I_{\nu}(\kappa)}\frac{\sqrt{2\pi a\kappa}e^{-a\kappa}I_{\nu+1}(a\kappa)}{\sqrt{2\pi a\kappa}e^{-a\kappa}I_{\nu}(a\kappa)} = 1. \tag{37}$$

In Lemma 3, we have established that the function $h(b) = \frac{1}{b}\frac{I_{\nu+1}(b\kappa)}{I_{\nu}(b\kappa)}$ is decreasing for $\kappa > 0$ and $b > 0$. Therefore, $\frac{1}{a}\frac{I_{\nu+1}(a\kappa)}{I_{\nu}(a\kappa)} = h(a) < h(1) = \frac{I_{\nu+1}(\kappa)}{I_{\nu}(\kappa)}$ when $a > 1$. Consequently, when $a > 1$,

$$\frac{1}{a} < \frac{I_{\nu+1}(\kappa)I_{\nu}(a\kappa)}{I_{\nu}(\kappa)I_{\nu+1}(a\kappa)}. \tag{38}$$

It follows from (37) and (38) that $\frac{1}{a} < g_a(\kappa) \leq 1$.

3. From part (a) and part (b) it appears that the function $g_a : \mathbb{R}_+ \mapsto (\frac{1}{a}, 1)$ is monotone, one-to-one function when $a > 1$. Therefore, the corresponding inverse function $g_a^{-1} : (\frac{1}{a}, 1) \mapsto \mathbb{R}_+$ exists and is one-to-one. According to the assumption $n < a^2 < n^2$, it follows that $\frac{a}{n} \in (\frac{1}{a}, 1)$. As a result, there is a unique point $\kappa^\star \in \mathbb{R}_+$ such that $g_a(\kappa^\star) = \frac{a}{n}$.

□

Lemma 7. *$h_*(\kappa) = \frac{I_\nu(\kappa)}{\kappa^\nu}$ is a log convex function, i.e.,*

$$\frac{I_\nu(\alpha\kappa_1 + (1-\alpha)\kappa_2)}{(\alpha\kappa_1 + (1-\alpha)\kappa_2)^\nu} \leq \left\{\frac{I_\nu(\kappa_1)}{\kappa_1^\nu}\right\}^\alpha \left\{\frac{I_\nu(\kappa_2)}{\kappa_2^\nu}\right\}^{(1-\alpha)}$$

for any $\kappa_1 > 0$, $\kappa_2 > 0$ and $0 < \alpha < 1$.

Proof. From the definition of $h_*(\kappa)$, we obtain that $\log(h_*(\kappa)) = \log I_\nu(\kappa) - \nu\log(\kappa)$. Therefore,

$$\frac{\partial}{\partial\kappa}(\log(h_*(\kappa)) = \left[\frac{I'_{\nu+1}(\kappa)}{I_\nu(\kappa)} - \frac{\nu}{\kappa}\right] = \left[\frac{I_{\nu+1}(\kappa)}{I_\nu(\kappa)} + \frac{\nu}{\kappa} - \frac{\nu}{\kappa}\right].$$

We also have

$$\frac{\partial^2}{\partial^2\kappa}(\log(h_*(\kappa)) = \frac{\partial}{\partial\kappa}\left[\frac{I_{\nu+1}(\kappa)}{I_\nu(\kappa)}\right] = R'_\nu(\kappa) > 0.$$

□

We will utilize this result to prove a lemma about log convexity of g as below.

Lemma 8. *$g(\kappa \mid \lambda, \zeta) = \left[\frac{\kappa^\nu \ \exp(\kappa\,\zeta)}{I_\nu(\kappa)}\right]^\lambda$ is a log convex function. Furthermore,*

$$\frac{g(\kappa \mid \lambda, \zeta)}{g(\alpha\hat{\kappa} + (1-\alpha)\kappa \mid \lambda, \zeta)} \leq \frac{g(\kappa \mid \lambda, \zeta)}{[g(\hat{\kappa} \mid \lambda, \zeta)]^\alpha [g(\hat{\kappa} \mid \lambda, \zeta)]^{1-\alpha}} \leq \left[\frac{g(\kappa \mid \lambda, \zeta)}{g(\hat{\kappa} \mid \lambda, \zeta)}\right]^\alpha$$

for all $\kappa \neq \hat{\kappa}$, α any constant in $(0, 1)$, and $\hat{\kappa}$ is the unique mode of $g(\kappa \mid \lambda, \zeta)$.

Proof.

$$\begin{aligned}
g(\alpha\hat{\kappa} + (1-\alpha)\kappa) &= \left[\frac{(\alpha\hat{\kappa} + (1-\alpha)\kappa)^{\nu}\ \exp\{(\alpha\hat{\kappa} + (1-\alpha)\kappa)\zeta\}}{I_{\nu}(\kappa)}\right]^{\lambda} \\
&= \left[\frac{\exp(\alpha\hat{\kappa}\zeta)\exp((1-\alpha)\kappa\zeta)}{h_*(\alpha\hat{\kappa} + (1-\alpha)\kappa)}\right]^{\lambda} \\
&\overset{(*)}{\geq} \left[\frac{\exp(\alpha\hat{\kappa}\zeta)\exp((1-\alpha)\kappa\zeta)}{[h_*(\hat{\kappa})]^{\alpha}[h_*(\hat{\kappa})]^{1-\alpha}}\right]^{\lambda} \\
&= \left[\frac{\exp(\hat{\kappa}\zeta)}{h_*(\hat{\kappa})}\right]^{\lambda\alpha}\left[\frac{\exp(\kappa\zeta)}{h_*(\hat{\kappa})}\right]^{\lambda(1-\alpha)} \\
&= [g(\hat{\kappa} \mid \lambda,\zeta)]^{\alpha}[g(\kappa \mid \lambda,\zeta)]^{1-\alpha},
\end{aligned}$$

where the inequality at (*) follows from Lemma 7. As $g(\kappa \mid \lambda,\zeta) > 0$, we get

$$\frac{g(\kappa \mid \lambda,\zeta)}{g(\alpha\hat{\kappa} + (1-\alpha)\kappa \mid \lambda,\zeta)} \leq \frac{g(\kappa \mid \lambda,\zeta)}{[g(\hat{\kappa} \mid \lambda,\zeta)]^{\alpha}[g(\hat{\kappa} \mid \lambda,\zeta)]^{1-\alpha}} \leq \left[\frac{g(\kappa \mid \lambda,\zeta)}{g(\hat{\kappa} \mid \lambda,\zeta)}\right]^{\alpha},$$

for all $\kappa \neq \hat{\kappa}$, as required. □

Lemma 9. *Let*

$$g(\kappa \mid \zeta,\lambda) = \left[\frac{\kappa^{\nu}\ \exp(\kappa\ \zeta)}{I_{\nu}(\kappa)}\right]^{\lambda}\mathbb{I}(\kappa > 0),$$

be a probability density kernel with $\kappa > 0$, $\lambda > 0$, and $\zeta \in (0,1)$. Further, we let $f(\kappa \mid \lambda,\zeta) = g(\kappa \mid \lambda,\zeta)/K_{\lambda,\zeta}$, where $K_{\lambda,\zeta} = \int_0^{\infty} g(\kappa \mid \lambda,\zeta)d\kappa$. Then,

1. *For any $\lambda > 0$ and $\zeta \in (0,1)$, there is an unique maximum $\hat{\kappa}$ such that $g(\hat{\kappa} \mid \lambda,\zeta) > g(\kappa \mid \lambda,\zeta)$ for $\kappa \neq \hat{\kappa}, \kappa > 0$. Further, $\hat{\kappa}$ is a function of ζ and does not depend on λ as long as $\lambda > 0$.*

2. *For all $\zeta \in (0,1)$,*

$$\lim_{\lambda\to\infty}\int_{\hat{\kappa}+\epsilon}^{\infty} f(\kappa \mid \lambda,\zeta)\ d\kappa = 0.$$

3. *For all $\zeta \in (0,1)$,*

$$\lim_{\lambda\to\infty}\int_{0}^{\hat{\kappa}-\epsilon} f(\kappa \mid \lambda,\zeta)\ d\kappa = 0.$$

Proof.

1. We have

$$g(\kappa \mid \zeta, \lambda) = \left[\frac{\kappa^{\nu}\ \exp(\kappa\,\zeta)}{I_{\nu}(\kappa)}\right]^{\lambda},$$

where $\zeta \in (0,1)$ and $\kappa > 0$. Note that $\log(g(\kappa \mid \zeta, \lambda)) = \lambda[\nu\log(\kappa) + \kappa\zeta - \log(I_{\nu}(\kappa))]$. Therefore,

$$\begin{aligned}\frac{\partial}{\partial\kappa}\log(g(\kappa \mid \zeta,\lambda)) &= \lambda\left[\frac{\nu}{\kappa} + \zeta - \frac{I'_{\nu+1}(\kappa)}{I_{\nu}(\kappa)}\right]\\ &= \lambda\left[\frac{\nu}{\kappa} + \zeta - \frac{\nu}{\kappa} - \frac{I_{\nu+1}(\kappa)}{I_{\nu}(\kappa)}\right]\\ &= \lambda(\zeta - R_{\nu}(\kappa)), \qquad (39)\end{aligned}$$

and $\frac{\partial^2}{\partial\kappa^2}\log(g(\kappa \mid \zeta,\lambda)) = -\lambda R'_{\nu}(\kappa)$ where $R_{\nu}(\kappa) = \frac{I_{\nu+1}(\kappa)}{I_{\nu}(\kappa)}$. From Lemma 1 we know that $\kappa \mapsto R_{\nu}(\kappa)$ is a strictly increasing function from $\mathbb{R}_+$ to $(0,1)$ because

$$\lim_{\kappa\to 0} R_{\nu}(\kappa) = \lim_{\kappa\to 0}\frac{I_{\nu+1}(\kappa)}{I_{\nu}(\kappa)} = 0 \quad \text{and} \quad \lim_{\kappa\to\infty} R_{\nu}(\kappa) = \lim_{\kappa\to\infty}\frac{I_{\nu+1}(\kappa)}{I_{\nu}(\kappa)} = 1.$$

As a result, $R_{\nu}^{-1}(\cdot) : (0,1) \to \mathbb{R}_+$ is well defined and one-to-one. Therefore, $\hat{\kappa} = R_{\nu}^{-1}(\zeta)$ is the unique solution for $\frac{\partial}{\partial\kappa}\log(g(\kappa \mid \zeta,\lambda)) = 0$. It is implied from Lemma 1 that $R'_{\nu}(\kappa) > 0$ for all κ. Therefore, $-\frac{\partial^2}{\partial\kappa^2}\log(g(\kappa \mid \zeta,\lambda))|_{\hat{\kappa}=R_{\nu}^{-1}(\|\boldsymbol{\psi}\|)} < 0$. All together we conclude that the distribution has unique mode at $\hat{\kappa} = R_{\nu}^{-1}(\zeta)$.

We can also conclude that $\hat{\kappa}$ is a function of ζ as it does not depend on λ as long as $\lambda > 0$. Therefore, $g(\hat{\kappa} \mid \lambda, \zeta) > g(\kappa \mid \lambda, \zeta)$ for $\kappa \neq \hat{\kappa}$, $\kappa > 0$.

2.

$$\begin{aligned}
\int_{\hat\kappa+\epsilon}^{\infty} f(\kappa \mid \lambda,\zeta)\;d\kappa &= \int_{\hat\kappa+\epsilon}^{\infty}\left[\frac{f(\kappa \mid 1,\zeta)}{f(\alpha\hat\kappa+(1-\alpha)\kappa \mid 1,\zeta)}\right]^{\lambda} f(\alpha\hat\kappa+(1-\alpha)\kappa \mid 1,\zeta)\;d\kappa \\
&\leq \int_{\hat\kappa+\epsilon}^{\infty}\left[\frac{g(\kappa \mid 1,\zeta)}{g(\hat\kappa \mid 1,\zeta)}\right]^{\alpha\lambda} f(\alpha\hat\kappa+(1-\alpha)\kappa \mid 1,\zeta)\;d\kappa \\
&\leq \int_{\hat\kappa+\epsilon}^{\infty}\left[\frac{g(\hat\kappa_* \mid 1,\zeta)}{g(\hat\kappa \mid 1,\zeta)}\right]^{\alpha\lambda} f(\alpha\hat\kappa+(1-\alpha)\kappa \mid 1,\zeta)\;d\kappa.
\end{aligned}$$

The inequalities follow by Lemma 7 for arbitrary α in $(0,1)$ and $\hat\kappa_* = \underset{\hat\kappa>\kappa+\epsilon}{\operatorname{argmax}}\, g(\kappa \mid 1,\zeta)$. Note that,

$$\begin{aligned}
\lim_{\kappa\to\infty} g(\kappa \mid 1,\zeta) = 0 \quad \exists \quad \hat{\kappa_*} \geq \hat\kappa+\epsilon \\
g(\hat\kappa_* \mid 1,\zeta) \geq g(\kappa \mid 1,\zeta) \qquad \text{for all} \quad \kappa \geq \hat\kappa+\epsilon
\end{aligned}$$

$$\begin{aligned}
\int_{\hat\kappa+\epsilon}^{\infty} f(\kappa \mid \lambda,\zeta)\;d\kappa &< \left[\frac{g(\hat\kappa_* \mid 1,\zeta)}{g(\hat\kappa \mid 1,\zeta)}\right]^{\alpha\lambda}\int_{\hat\kappa+\epsilon}^{\infty} f(\alpha\hat\kappa+(1-\alpha)\kappa \mid 1,\zeta)\;d\kappa \\
&= \left[\frac{g(\hat\kappa_* \mid 1,\zeta)}{g(\hat\kappa \mid 1,\zeta)}\right]^{\alpha\lambda}\int_{\hat\kappa+(1-\alpha)\epsilon}^{\infty} f(y \mid 1,\zeta)\frac{dy}{(1-\alpha)}. \\
&< \left[\frac{g(\hat\kappa_* \mid 1,\zeta)}{g(\hat\kappa \mid 1,\zeta)}\right]^{\alpha\lambda}.
\end{aligned}$$

where we have $y = \alpha\hat\kappa+(1-\alpha)\kappa$, Finally,

$$\lim_{\lambda\to\infty}\int_{\kappa>\hat\kappa+\epsilon} f(\kappa \mid \lambda,\zeta)\;d\kappa < \lim_{\lambda\to\infty}\left[\frac{g(\hat\kappa_* \mid 1,\zeta)}{g(\hat\kappa \mid 1,\zeta)}\right]^{\alpha\lambda} = 0 \qquad \text{as} \quad \hat\kappa_* \neq \hat\kappa.$$

3. We prove part (c), similarly as above. Following Lemma 7, for arbitrary α in

$(0,1)$, we can write

$$\begin{aligned}
\int_0^{\hat{\kappa}-\epsilon} f(\kappa \mid \lambda, \zeta)\; d\kappa &\leq \int_0^{\hat{\kappa}-\epsilon} \left[\frac{f(\kappa \mid 1, \zeta)}{f(\alpha\hat{\kappa} + (1-\alpha)\kappa \mid 1, \zeta)}\right]^{\lambda} f(\alpha\hat{\kappa} + (1-\alpha)\kappa \mid \lambda, \zeta)\; d\kappa \\
&= \int_0^{\hat{\kappa}-\epsilon} \left[\frac{g(\kappa \mid 1, \zeta)}{g(\hat{\kappa} \mid 1, \zeta)}\right]^{\alpha\lambda} f(\alpha\hat{\kappa} + (1-\alpha)\kappa \mid \lambda, \zeta)\; d\kappa \\
&\leq \int_0^{\hat{\kappa}-\epsilon} \left[\frac{g(\hat{\kappa}_* \mid 1, \zeta)}{g(\hat{\kappa} \mid 1, \zeta)}\right]^{\alpha\lambda} f(\alpha\hat{\kappa} + (1-\alpha)\kappa \mid \lambda, \zeta)\; d\kappa,
\end{aligned}$$

where $\hat{\kappa}_* = \underset{0\leq\kappa\leq\hat{\kappa}-\epsilon}{\operatorname{argmax}} g(\kappa \mid 1, \zeta)$. Note that $\lim_{\kappa\to 0} g(\kappa \mid 1, \zeta) = 0$, which implies that $0 < \hat{\kappa_*} \leq \hat{\kappa} - \epsilon$. As a result,

$$\begin{aligned}
\int_0^{\hat{\kappa}-\epsilon} f(\kappa \mid \lambda, \zeta)\; d\kappa &\leq \left[\frac{g(\hat{\kappa}_* \mid 1, \zeta)}{g(\hat{\kappa} \mid 1, \zeta)}\right]^{\alpha\lambda} \int_0^{\hat{\kappa}-(1-\alpha)\epsilon} f(\alpha\hat{\kappa} + (1-\alpha)\kappa \mid \lambda, \zeta)\; d\kappa \\
&\leq \left[\frac{g(\hat{\kappa}_* \mid 1, \zeta)}{g(\hat{\kappa} \mid 1, \zeta)}\right]^{\alpha\lambda} \int_0^{\infty} f(y \mid 1, \zeta) \frac{dy}{(1-\alpha)} \\
&\leq \left[\frac{g(\hat{\kappa}_* \mid 1, \zeta)}{g(\hat{\kappa} \mid \lambda, \zeta)}\right]^{\alpha\lambda} F(\hat{\kappa} - (1-\alpha)\epsilon),
\end{aligned}$$

where we use the substitution $y = \alpha\hat{\kappa} + (1-\alpha)\kappa$. Here, $F(\cdot)$ is the cumulative distribution function of $f(\cdot \mid 1, \zeta$. Finally,

$$\lim_{\lambda\to\infty} \int_0^{\hat{\kappa}-\epsilon} f(\kappa \mid \lambda, \zeta)\; d\kappa \leq \lim_{\lambda\to\infty} \left[\frac{g(\hat{\kappa}_* \mid 1, \zeta)}{g(\hat{\kappa} \mid 1, \zeta)}\right]^{\alpha\lambda} = 0,$$

as the $\hat{\kappa}$ is the unique mode and $\hat{\kappa}_* \neq \hat{\kappa}$.

□

Proof of Theorem 1

Proof. 1. From Definition 1 it follows that

$$\begin{aligned}
\int_{\mathbb{R}_+}\int_{\mathbb{S}^{p-1}} g(\boldsymbol{\mu},\kappa \mid \boldsymbol{\psi},\lambda)d\boldsymbol{\mu}\; d\kappa &= \int_{\mathbb{R}_+}\int_{\mathbb{S}^{p-1}} \left[\frac{\kappa^{\nu}\ \exp\left(\kappa\ \boldsymbol{\mu}^T\boldsymbol{\psi}\right)}{I_\nu(\kappa)}\right]^{\lambda} d\boldsymbol{\mu}\; d\kappa \\
&= \int_{\mathbb{R}_+} \frac{\kappa^{\nu\lambda}}{[I_\nu(\kappa)]^{\lambda}}\left[\int_{\mathbb{S}^{p-1}} \exp\left(\kappa\boldsymbol{\mu}^T\boldsymbol{\psi}\lambda\right)d\boldsymbol{\mu}\right]\ d\kappa \\
&= \int_{\mathbb{R}_+} \frac{\kappa^{\nu\lambda}}{[I_\nu(\kappa)]^{\lambda}}\left[\int_{\mathbb{S}^{p-1}} \exp\left(\|\kappa\boldsymbol{\psi}\lambda\|\ \boldsymbol{\mu}^T\hat{\boldsymbol{\mu}}\right)d\boldsymbol{\mu}\right]\ d\kappa \\
&= \int_{\mathbb{R}_+} \frac{\kappa^{\nu\lambda}(2\pi)^{p/2}}{[I_\nu(\kappa)]^{\lambda}}\frac{I_\nu(\|\kappa\boldsymbol{\psi}\lambda\|)}{\|\kappa\boldsymbol{\psi}\lambda\|^{\nu}}\ d\kappa \\
&= \frac{(2\pi)^{p/2}}{\|\boldsymbol{\psi}\lambda\|^{\nu}}\int_{\mathbb{R}_+}\frac{\kappa^{\nu\lambda-\nu}}{[I_\nu(\kappa)]^{\lambda}}I_\nu(\kappa\lambda\,\|\boldsymbol{\psi}\|)\ d\kappa, \qquad (40)
\end{aligned}$$

where $\hat{\boldsymbol{\mu}} = \frac{\boldsymbol{\psi}}{\|\boldsymbol{\psi}\|}$. Let $M > 0$ be any positive number. Then it follows from (40) that

$$\begin{aligned}
\int_{\mathbb{R}_+}\int_{\mathbb{S}^{p-1}} g(\boldsymbol{\mu},\kappa \mid \boldsymbol{\psi},\lambda)\ d\boldsymbol{\mu}\ d\kappa = \frac{(2\pi)^{p/2}}{\|\boldsymbol{\psi}\lambda\|^{\nu}}\Bigg[&\int_M^{\infty}\frac{\kappa^{\nu\lambda-\nu}}{[I_\nu(\kappa)]^{\lambda}}I_\nu(\|\kappa\boldsymbol{\psi}\lambda\|) \\
&+\int_0^M \frac{\kappa^{\nu\lambda-\nu}}{[I_\nu(\kappa)]^{\lambda}}I_\nu(\|\kappa\boldsymbol{\psi}\lambda\|)\Bigg]d\kappa.
\end{aligned} \qquad (41)$$

In order to bound the above integrals, consider Luke [1972, Eq.6.25]

$$\frac{\kappa^{\nu}}{2^{\nu}\Gamma(\nu+1)} < I_\nu(\kappa) < \frac{(1+\exp\left(-2\kappa\right))}{2}\frac{\kappa^{\nu}e^{\kappa}}{2^{\nu}\Gamma(\nu+1)} < \frac{\kappa^{\nu}e^{\kappa}}{2^{\nu}\Gamma(\nu+1)}. \qquad (42)$$

On the other hand we have from Pal et al. [2020, Lemma 9] that for $M > 0$ and $\nu > \frac{1}{2}$,

$$I_\nu(\kappa) \geq \frac{\exp(\kappa)}{\sqrt{(\kappa)}}G(M), \qquad (43)$$

where $G(M) = \sqrt{M}\exp(-M)I_\nu(M)$. Using (43), we obtain that

$$\begin{aligned}
&\frac{(2\pi)^{p/2}}{\|\boldsymbol{\psi}\lambda\|^\nu}\int_M^\infty \frac{\kappa^{\nu\lambda-\nu}}{[I_\nu(\kappa)]^\lambda} I_\nu(\|\kappa\boldsymbol{\psi}\lambda\|)\ d\kappa \\
\leq\ & \frac{(2\pi)^{p/2}}{\|\boldsymbol{\psi}\lambda\|^\nu}\int_M^\infty \left[\frac{\sqrt{\kappa}}{\exp(\kappa)G(M)}\right]^\lambda I_\nu(\|\kappa\boldsymbol{\psi}\lambda\|)\kappa^{\nu\lambda-\nu}\ d\kappa \\
\overset{(\dagger)}{<}\ & \frac{(2\pi)^{p/2}(\lambda\,\|\boldsymbol{\psi}\|)^\nu}{\|\boldsymbol{\psi}\lambda\|^\nu\,2^\nu G(M)^\nu}\int_M^\infty \kappa^{\lambda/2+\nu(\lambda-1)+\nu}\exp(-\lambda\kappa+\lambda\kappa\,\|\boldsymbol{\psi}\|)\ d\kappa \\
=\ & \frac{(2\pi)^{p/2}(\lambda\,\|\boldsymbol{\psi}\|)^\nu}{\|\boldsymbol{\psi}\lambda\|^\nu\,2^\nu G(M)^\nu}\int_M^\infty \kappa^{\lambda/2+\nu(\lambda-1)+\nu}\exp(-\lambda\kappa(1-\|\boldsymbol{\psi}\|))\ d\kappa, \qquad (44)
\end{aligned}$$

where inequality in (†) follows from (42). The above integral is finite if $\|\boldsymbol{\psi}\| < 1$, since $\lambda/2 + \nu(\lambda-1)+\nu > 0$. On other hand, consider that

$$\begin{aligned}
&\frac{(2\pi)^{p/2}}{\|\boldsymbol{\psi}\lambda\|^\nu}\int_0^M \frac{\kappa^{\nu\lambda-\nu}}{[I_\nu(\kappa)]^\lambda} I_\nu(\|\kappa\boldsymbol{\psi}\lambda\|)\ d\kappa \\
\overset{(\dagger\dagger)}{\leq}\ & \frac{(2\pi)^{p/2}}{\|\boldsymbol{\psi}\lambda\|^\nu}\int_0^M \frac{\kappa^{\nu\lambda+\nu\lambda-\nu+\nu}}{(2^\nu\Gamma(\nu+1))^\lambda}\frac{\exp(\kappa\,\|\boldsymbol{\psi}\lambda\|)}{2^\nu\Gamma(\nu+1)}\ d\kappa \\
=\ & \frac{(2\pi)^{p/2}}{\|\boldsymbol{\psi}\lambda\|^\nu\,2^{\nu\lambda+1}\Gamma(\nu+1)^{\nu\lambda+1}}\int_0^M \kappa^{\nu\lambda+\nu\lambda-\nu+\nu}\exp(\kappa\,\|\boldsymbol{\psi}\lambda\|)\ d\kappa, \qquad (45)
\end{aligned}$$

where (††) follows from (42). Altogether from (41), (44) and (45) we conclude that $\int_{\mathbb{R}_+}\int_{\mathbb{S}^{p-1}} g(\boldsymbol{\mu},\kappa \mid \boldsymbol{\psi},\lambda)\ d\boldsymbol{\mu}\ d\kappa < \infty$ when $\|\boldsymbol{\psi}\| < 1$.

2. From (41),

$$\begin{aligned}
&\int_{\mathbb{R}_+}\int_{\mathbb{S}^{p-1}} g(\boldsymbol{\mu},\kappa \mid \boldsymbol{\psi},\lambda)\ d\boldsymbol{\mu}\ d\kappa \\
>\ & \frac{(2\pi)^{p/2}}{\|\boldsymbol{\psi}\lambda\|^\nu}\int_M^\infty \frac{\kappa^{\nu\lambda-\nu}}{[I_\nu(\kappa)]^\lambda} I_\nu(\|\kappa\boldsymbol{\psi}\lambda\|)\ d\kappa \\
>\ & \frac{(2\pi)^{p/2}2^{\nu\lambda}}{\|\boldsymbol{\psi}\lambda\|^\nu}\int_M^\infty \kappa^{-\nu\lambda}\kappa^{\nu\lambda-\lambda} exp(-\kappa\lambda)I_\nu(\|\kappa\boldsymbol{\psi}\lambda\|)\ d\kappa,
\end{aligned}$$

where the last inequality follows from (42). Additionally, using (43) we obtain that

$$\begin{aligned}
& \int_{\mathbb{R}_+}\int_{\mathbb{S}^{p-1}} g(\boldsymbol{\mu},\kappa \mid \boldsymbol{\psi},\lambda)\ d\boldsymbol{\mu}\ d\kappa \\
\overset{\dagger}{\geq}\ & \frac{(2\pi)^{p/2}2^{\nu\lambda}}{\|\boldsymbol{\psi}\lambda\|^{\nu}} \int_M^\infty \kappa^{-\lambda} exp(-\kappa\lambda)\frac{exp(\|\kappa\boldsymbol{\psi}\lambda\|)G(M)}{\sqrt{\kappa\,\|\boldsymbol{\psi}\|\,\lambda}}\ d\kappa \\
=\ & \frac{(2\pi)^{p/2}2^{\nu\lambda}G(M)}{\|\boldsymbol{\psi}\lambda\|^{\nu}\sqrt{\|\boldsymbol{\psi}\|\,\lambda}} \int_M^\infty \kappa^{-\lambda-\frac{1}{2}}\exp(\kappa\lambda(\|\boldsymbol{\psi}\|-1))\ d\kappa, \\
=\ & \infty,
\end{aligned}$$

because $\|\boldsymbol{\psi}\| = 1$.

3. From Definition (41) we have

$$\int_{\mathbb{R}_+}\int_{\mathbb{S}^{p-1}} g(\boldsymbol{\mu},\kappa \mid \boldsymbol{\psi},\lambda)\ d\boldsymbol{\mu}\ d\kappa = \frac{(2\pi)^{p/2}}{\lambda^\nu}\int_0^\infty \frac{\kappa^{\nu\lambda-\nu}}{[I_\nu(\kappa)]^\lambda} I_\nu(\kappa\lambda)\ d\kappa. \tag{46}$$

We obtain from (30) that

$$\frac{I_\nu(\kappa)}{\frac{e^\kappa}{\sqrt{2\pi\kappa}}} \to 1,$$

for all $\nu \geq 0$ as $\kappa \to \infty$. Therefore for any given small $\epsilon \in (0,1)$, there is M_ϵ such that,

$$(1-\epsilon)\ \frac{e^\kappa}{\sqrt{2\pi\kappa}} < I_\nu(\kappa) < (1+\epsilon)\ \frac{e^\kappa}{\sqrt{2\pi\kappa}} \quad \text{for all} \quad \kappa > M_\epsilon.$$

Similarly,

$$(1-\epsilon)\ \frac{e^{\lambda\kappa}}{\sqrt{2\pi\lambda\kappa}} < I_\nu(\lambda\kappa) < (1+\epsilon)\ \frac{e^{\lambda\kappa}}{\sqrt{2\pi\lambda\kappa}} \quad \text{for all} \quad \kappa > M_\epsilon/\lambda.$$

As a result, if $\kappa > M_* = \max\{M_\epsilon/\lambda, M_\epsilon\}$ then,

$$\begin{aligned}
(1-\epsilon)\ \frac{e^\kappa}{\sqrt{2\pi\kappa}} &< I_\nu(\kappa) < & (1+\epsilon)\ \frac{e^\kappa}{\sqrt{2\pi\kappa}}, \\
(1-\epsilon)\ \frac{e^{\lambda\kappa}}{\sqrt{2\pi\lambda\kappa}} &< I_\nu(\lambda\kappa) < & (1+\epsilon)\ \frac{e^{\lambda\kappa}}{\sqrt{2\pi\lambda\kappa}}
\end{aligned} \tag{47}$$

On other hand know, from (31) we have that $\frac{I_\nu(\kappa)}{\frac{\kappa^\nu}{2^\nu\Gamma(\nu+1)}} \to 1$ as $\kappa \to 0$. Therefore, for any $0 < \epsilon_1 < 1$, there exists M_{ϵ_1} such that.

$$\begin{aligned} (1-\epsilon_1)\,\frac{\kappa^\nu}{2^\nu\Gamma(\nu+1)} &< I_\nu(\kappa) < \quad (1+\epsilon_1)\,\frac{\kappa^\nu}{2^\nu\Gamma(\nu+1)} \quad \text{for all} \quad \kappa < M \\ (1-\epsilon_1)\,\frac{\kappa^{\lambda\nu}}{(2^\nu\Gamma(\nu+1))^\lambda} &< I_\nu(\lambda\kappa) < \quad (1+\epsilon_1)\,\frac{\kappa^{\lambda\nu}}{(2^\nu\Gamma(\nu+1))^\lambda} \quad \text{for all} \quad \kappa < M/\lambda. \end{aligned}$$

As a result for all $0 < \kappa < m_* = \min\{M_{\epsilon_1}/\lambda, M_{\epsilon_1}\}$,

$$\begin{aligned} (1-\epsilon_1)\,\frac{\kappa^\nu}{2^\nu\Gamma(\nu+1)} &< I_\nu(\kappa) < \quad (1+\epsilon_1)\,\frac{\kappa^\nu}{2^\nu\Gamma(\nu+1)} \\ (1-\epsilon_1)\,\frac{(\lambda\kappa)^\nu}{(2^\nu\Gamma(\nu+1))^\lambda} &< I_\nu(\lambda\kappa) < \quad (1+\epsilon_1)\,\frac{(\lambda\kappa)^\nu}{(2^\nu\Gamma(\nu+1))^\lambda}. \end{aligned}$$

Using these bounds we further calculate,

$$\begin{aligned} &\int_{\mathbb{R}_+}\int_{\mathbb{S}^{p-1}} g(\boldsymbol{\mu},\kappa \mid \boldsymbol{\psi},\lambda)\; d\boldsymbol{\mu}\; d\kappa \\ =\;& \frac{(2\pi)^{p/2}}{\lambda^\nu}\left[\int_0^{m_*}\frac{\kappa^{\nu(\lambda-1)}}{[I_\nu(\kappa)]^\lambda} I_\nu(\lambda\kappa)\; d\kappa + \int_{m_*}^{M^*}\frac{\kappa^{\nu(\lambda-1)}}{[I_\nu(\kappa)]^\lambda} I_\nu(\lambda\kappa)\; d\kappa + \int_{M^*}^{\infty}\frac{\kappa^{\nu(\lambda-1)}}{[I_\nu(\kappa)]^\lambda} I_\nu(\lambda\kappa)\; d\kappa\right] \\ =\;& \frac{(2\pi)^{p/2}}{\lambda^\nu}\left[\int_0^{m_*}\frac{\kappa^{\nu(\lambda-1)}}{[I_\nu(\kappa)]^\lambda} I_\nu(\lambda\kappa)\; d\kappa + \int_{M^*}^{\infty}\frac{\kappa^{\nu(\lambda-1)}}{[I_\nu(\kappa)]^\lambda} I_\nu(\lambda\kappa)\; d\kappa + G(m_*, M_*)\right]. \qquad (48) \end{aligned}$$

Here, $G(m_*, M_*) = \int_{m_*}^{M^*} \frac{\kappa^{\nu(\lambda-1)}}{[I_\nu(\kappa)]^\lambda} I_\nu(\lambda\kappa)\; d\kappa$. Now, substituting inequalities (47) and (48) in the above equation, we get

$$\begin{aligned}
&\int_{\mathbb{R}_+}\int_{\mathbb{S}^{p-1}} g(\boldsymbol{\mu},\kappa \mid \boldsymbol{\psi},\lambda)\; d\boldsymbol{\mu}\; d\kappa \\
<\;& \frac{(2\pi)^{p/2}}{\lambda^{\nu}}\Bigg[\int_0^{m_*} \kappa^{\nu(\lambda-1)}\frac{(\lambda\kappa)^{\lambda}(1+\epsilon_1)(2^{\nu}\Gamma(\nu+1))^{\lambda}}{2^{\nu}\Gamma(\nu+1)\kappa^{\lambda\nu}(1+\epsilon_1)^{\lambda}}\; d\kappa \\
&+\int_{M^*}^{\infty}\kappa^{\nu(\lambda-1)}\frac{(1+\epsilon)e^{\lambda\kappa}(\sqrt{2\pi\lambda\kappa})^{\lambda}}{\sqrt{2\pi\lambda\kappa}e^{\lambda\kappa}(1+\epsilon)^{\lambda}}\; d\kappa + G(m_*,M_*)\Bigg]. \\
<\;& \frac{(2\pi)^{p/2}}{\lambda^{\nu}}\Bigg[\frac{2^{(\lambda-1)\nu}(\Gamma\nu+1)^{\lambda-1}(1+\epsilon_1)}{(1+\epsilon_1)^{\lambda}}m_* \\
&+\int_{M^*}^{\infty}\frac{(1+\epsilon)}{(1-\epsilon)^{\lambda}}(2\pi)^{(\lambda-1)/2}\kappa^{((\lambda-1)(\nu+0.5))}\; d\kappa + G(m_*,M_*)\Bigg].
\end{aligned}$$

If $\lambda < \frac{2\nu-1}{2\nu+1}$ and $\nu > 1/2$, then $(\lambda-1)(\nu+0.5)+1<0$, the above integral is finite. Conversely, if $\lambda \geq \frac{2\nu-1}{2\nu+1}$ then from (47), (48) and (48) it follows that

$$\begin{aligned}
&\int_{\mathbb{R}_+}\int_{\mathbb{S}^{p-1}} g(\boldsymbol{\mu},\kappa \mid \boldsymbol{\psi},\lambda)\; d\boldsymbol{\mu}\; d\kappa \\
>\;& \frac{(2\pi)^{p/2}}{\lambda^{\nu}}\Bigg[\int_0^{m_*} \kappa^{\nu(\lambda-1)}\frac{(\lambda\kappa)^{\lambda}\;(1-\epsilon_1)\;(2^{\nu}\Gamma(\nu+1))^{\lambda}}{2^{\nu}\Gamma(\nu+1)\;\kappa^{\lambda\nu}(1+\epsilon_1)^{\lambda}}\; d\kappa \\
&+\int_{M^*}^{\infty}\kappa^{\nu(\lambda-1)}\frac{(1-\epsilon)\;e^{\lambda\kappa}(\sqrt{2\pi\kappa})^{\lambda}}{\sqrt{2\pi\;\lambda\;\kappa}\;e^{\lambda\kappa}(1+\epsilon)^{\lambda}}\; d\kappa + G(m_*,M_*)\Bigg]. \\
&= \frac{(2\pi)^{p/2}}{\lambda^{\nu}}\Bigg[\frac{2^{(\lambda-1)\nu}(\Gamma(\nu+1))^{\lambda-1}(1-\epsilon_1)}{(1+\epsilon_1)^{\lambda}}m_* \\
&+\int_{M^*}^{\infty}\frac{(1-\epsilon)}{(1+\epsilon)^{\lambda}}(2\pi)^{(\lambda-1)/2}\kappa^{((\lambda-1)(\nu+0.5))+1}d\kappa \\
&+G(m_*,M_*)\Bigg] \\
&= \infty.
\end{aligned}$$

□

Proof of Theorem 2

Proof. We start by considering the fact that

$$\boldsymbol{\mu}^T\boldsymbol{\psi} \leq \sqrt{\|\boldsymbol{\mu}\|^2 \|\boldsymbol{\psi}\|^2} = \|\boldsymbol{\psi}\|, \tag{49}$$

since $\boldsymbol{\mu} \in \mathbb{S}^{p-1}$ and $\|\boldsymbol{\mu}\| = 1$. The equality is achieved in (49) at

$$\hat{\boldsymbol{\mu}} = \frac{\boldsymbol{\psi}}{\|\boldsymbol{\psi}\|}.$$

Therefore irrespective of value of $\kappa > 0$,

$$g(\boldsymbol{\mu}, \kappa \mid \boldsymbol{\psi}, \lambda) < \left[\frac{\kappa^\nu \ \exp\left(\kappa \ \|\boldsymbol{\psi}\|\right)}{I_\nu(\kappa)}\right]^\lambda = g(\hat{\boldsymbol{\mu}}, \kappa \mid \boldsymbol{\psi}, \lambda). \tag{50}$$

For $\boldsymbol{\psi} \in \mathbb{R}^d$, $\lambda > 0$, such that $\|\boldsymbol{\psi}\| < 1$. Note that

$$\log(g(\hat{\boldsymbol{\mu}}, \kappa \mid \boldsymbol{\psi}, \lambda)) = \lambda[\nu \log(\kappa) + \kappa \|\boldsymbol{\psi}\| - \log(I_\nu(\kappa))].$$

Therefore,

$$\begin{aligned}
\frac{\partial}{\partial \kappa}\log(g(\hat{\boldsymbol{\mu}}, \kappa \mid \boldsymbol{\psi}, \lambda)) &= \lambda\left[\frac{\nu}{\kappa} + \|\boldsymbol{\psi}\| - \frac{I'_\nu(\kappa)}{I_\nu(\kappa)}\right] \\
&= \lambda\left[\frac{\nu}{\kappa} + \|\boldsymbol{\psi}\| - \frac{\nu}{\kappa} - \frac{I_{\nu+1}(\kappa)}{I_\nu(\kappa)}\right] \\
&= \lambda(\|\boldsymbol{\psi}\| - R_\nu(\kappa)), \qquad (51)
\end{aligned}$$

and $\frac{\partial^2}{\partial \kappa^2}\log(g(\hat{\boldsymbol{\mu}}, \kappa \mid \boldsymbol{\psi}, \lambda)) = -\lambda R'_\nu(\kappa)$ where $R_\nu(\kappa) = \frac{I_{\nu+1}(\kappa)}{I_\nu(\kappa)}$. From Lemma 1 we know that $\kappa \mapsto R_\nu(\kappa)$ is strictly increasing function from $\mathbb{R}_+$ to $(0, 1)$ because

$$\lim_{\kappa\to 0} R_\nu(\kappa) = \lim_{\kappa\to 0}\frac{I_{\nu+1}(\kappa)}{I_\nu(\kappa)} = 0 \quad \text{and} \quad \lim_{\kappa\to\infty} R_\nu(\kappa) = \lim_{\kappa\to\infty}\frac{I_{\nu+1}(\kappa)}{I_\nu(\kappa)} = 1.$$

As a result, $R_\nu^{-1}(\cdot) : (0,1) \to \mathbb{R}_+$ is well defined and one-to-one. Therefore $\hat{\kappa} = R_\nu^{-1}(\|\boldsymbol{\psi}\|)$ is the unique solution for $\frac{\partial}{\partial\kappa}\log(g(\hat{\boldsymbol{\mu}}, \kappa \mid \boldsymbol{\psi}, \lambda)) = 0$. It is implied from Lemma 1 that $R'_\nu(\kappa) > 0$ for all κ. Therefore,

$$-\frac{\partial^2}{\partial\kappa^2}\log(g(\hat{\boldsymbol{\mu}}, \kappa \mid \boldsymbol{\psi}, \lambda))|_{\hat{\kappa}=R_\nu^{-1}(\|\boldsymbol{\psi}\|)} < 0.$$

In total, we conclude that the distribution has unique mode at $\hat{\boldsymbol{\mu}} = \frac{\boldsymbol{\psi}}{\|\boldsymbol{\psi}\|}$ and $\hat{\kappa} = R_\nu^{-1}(\|\boldsymbol{\psi}\|)$. □

Proof of Theorem 3

Proof. 1. From the definition of the function, we observe that

$$g(\boldsymbol{\mu}, \kappa \mid \boldsymbol{\psi}, \lambda) = [g(\boldsymbol{\mu}, \kappa \mid \boldsymbol{\psi}, 1)]^\lambda.$$

From Theorem 2, we know that $g(\boldsymbol{\mu}, \kappa \mid \boldsymbol{\psi}, \lambda)$ has a unique mode at $\hat{\boldsymbol{\mu}} = \frac{\boldsymbol{\psi}}{\|\boldsymbol{\psi}\|}$ and $\hat{\kappa} = R_\nu^{-1}(\|\boldsymbol{\psi}\|)$. Irrespective of the value of $\lambda > 0$, let S_l be the l^{th} level set of $g(\boldsymbol{\mu}, \kappa \mid \boldsymbol{\psi}, 1)$ as defined in (3.16) for some $l \in (0,1)$. Note that $g(\boldsymbol{\mu}, \kappa \mid \boldsymbol{\psi}, 1) > g(\boldsymbol{\mu}_*, \kappa_* \mid \boldsymbol{\psi}, 1)$ for $(\boldsymbol{\mu}, \kappa) \in S_l$ and $(\boldsymbol{\mu}_*, \kappa_*) \in S_l^c$, where S_l^c denotes complimentary set of S_l. Consequently, the function $\lambda \mapsto r_\lambda(\boldsymbol{\mu}_*, \kappa_*)$ is an increasing function in $\lambda > 0$, for any $(\boldsymbol{\mu}_*, \kappa_*) \in S_l^c$, where

$$r_\lambda(\boldsymbol{\mu}_*, \kappa_*) = \iint\limits_{S_l} \left[\frac{g(\boldsymbol{\mu}, \kappa \mid \boldsymbol{\psi}, 1)}{g(\boldsymbol{\mu}_*, \kappa_* \mid \boldsymbol{\psi}, 1)}\right]^\lambda d\boldsymbol{\mu}\, d\kappa. \tag{52}$$

Now consider the fact that

$$\begin{aligned}
\frac{P_{\boldsymbol{\psi},\lambda}(S_l^c)}{P_{\boldsymbol{\psi},\lambda}(S_l)} &= \frac{\iint\limits_{S_l^c} \left[g(\boldsymbol{\mu}_*, \kappa_* \mid \psi, 1)\right]^{\lambda} d\boldsymbol{\mu}_* \, d\kappa_*}{\iint\limits_{S_l} \left[g(\boldsymbol{\mu}, \kappa \mid \boldsymbol{\psi}, 1)\right]^{\lambda} d\boldsymbol{\mu} d\kappa} \\
&= \iint\limits_{S_l^c} \frac{1}{\iint\limits_{S_l} \frac{g(\boldsymbol{\mu},\kappa|\psi,1)}{g(\boldsymbol{\mu}_*,\kappa_*|\psi,\lambda)} d\boldsymbol{\mu} \, d\kappa} d\boldsymbol{\mu}_* \, d\kappa_* \\
&= \iint\limits_{S_l^c} \frac{1}{r_\lambda(\boldsymbol{\mu}_*, \kappa_*)} d\boldsymbol{\mu}_* \, d\kappa_*.
\end{aligned}$$

Therefore, it follows that $\frac{P_{\boldsymbol{\psi},\lambda}(S_l^c)}{P_{\boldsymbol{\psi},\lambda}(S_l)}$ is a decreasing function of $\lambda > 0$ as $\frac{1}{r_\lambda(\boldsymbol{\mu}_*,\kappa_*)}$ is decreasing for all $(\boldsymbol{\mu}_*, \kappa_*) \in S_l^c$.

2. Let A be an open set such that $(\hat{\boldsymbol{\mu}}, \hat{\kappa}) \in A$. Therefore there exists an open ball, $B_\epsilon(\hat{\boldsymbol{\mu}}, \hat{\kappa}) = \{d((\boldsymbol{\mu}, \kappa), (\hat{\boldsymbol{\mu}}, \hat{\kappa})) < \epsilon\}$ such that $B_\epsilon(\hat{\boldsymbol{\mu}}, \hat{\kappa}) \subset A$. We are considering the following norm for space $\mathbb{S}^{p-1} \times \mathbb{R}_+$,

$$d((\boldsymbol{\mu}, \kappa), (\hat{\boldsymbol{\mu}}, \hat{\kappa})) = \sqrt{(\kappa - \hat{\kappa})^2 + (\boldsymbol{\mu} - \hat{\boldsymbol{\mu}})^T (\boldsymbol{\mu} - \hat{\boldsymbol{\mu}})}.$$

Now consider that

$$B_\epsilon(\hat{\boldsymbol{\mu}}, \hat{\kappa})^c \nsubseteq \left\{(\boldsymbol{\mu}, \kappa) : |\kappa - \hat{\kappa}| > \frac{\epsilon}{2}\right\} \cup \left\{(\boldsymbol{\mu}, \kappa) : \|\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}\| > \frac{\epsilon}{2}\right\}.$$

As a result

$$P((\boldsymbol{\mu}, \kappa) \in B_\epsilon(\hat{\boldsymbol{\mu}}, \hat{\kappa})^c) \leq P\left(|\kappa - \hat{\kappa}| > \frac{\epsilon}{2}\right) + P\left(\|\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}\| > \frac{\epsilon}{2}\right). \tag{53}$$

Note that

$$\begin{aligned}
P_{\boldsymbol{\psi},\lambda}\left(|\kappa-\hat{\kappa}|>\frac{\epsilon}{2}\right) &= \int_{|\kappa-\hat{\kappa}|>\frac{\epsilon}{2}}\int_{\mathbb{S}^{p-1}} \frac{g(\boldsymbol{\mu},\kappa \mid \psi,\lambda)}{K_{\boldsymbol{\psi},\lambda}} d\boldsymbol{\mu}\ d\kappa \\
&= \int_{|\kappa-\hat{\kappa}|>\frac{\epsilon}{2}}\int_{\mathbb{S}^{p-1}} \left[\frac{\kappa^{\nu}\ \exp\left(\kappa\ \boldsymbol{\mu}^T\boldsymbol{\psi}\right)}{I_\nu(\kappa)}\right]^{\lambda} \frac{1}{K_{\boldsymbol{\psi},\lambda}} d\boldsymbol{\mu}\ d\kappa \\
&\leq \int_{|\kappa-\hat{\kappa}|>\frac{\epsilon}{2}}\int_{\mathbb{S}^{p-1}} \left[\frac{\kappa^{\nu} \exp\left(\kappa\ \|\boldsymbol{\psi}\|\right)}{I_\nu(\kappa)}\right]^{\lambda} \frac{1}{K_{\boldsymbol{\psi},\lambda}} d\boldsymbol{\mu}\ d\kappa.
\end{aligned}$$

The inequality follows from (49). As the $d\boldsymbol{\mu}$ is the normalized Haar measure on $\mathbb{S}^{p-1}$, Lemma 8 gives

$$\lim_{\lambda\to\infty} P_{\boldsymbol{\psi},\lambda}(|\kappa-\hat{\kappa}|>\frac{\epsilon}{2}) \leq \lim_{\lambda\to\infty}\int_{|\kappa-\hat{\kappa}|>\frac{\epsilon}{2}} \left[\frac{\kappa^{\nu}\ \exp\left(\kappa\ \|\boldsymbol{\psi}\|\right)}{I_\nu(\kappa)}\right]^{\lambda} d\kappa = 0. \tag{54}$$

On the other hand, if $\|\boldsymbol{\mu}-\hat{\boldsymbol{\mu}}\|>\frac{\epsilon}{2}$ then $\boldsymbol{\mu}^T\hat{\boldsymbol{\mu}}<(1-\frac{\epsilon^2}{8})$. From $\hat{\boldsymbol{\mu}}=\frac{\boldsymbol{\psi}}{\|\boldsymbol{\psi}\|}$,

$$\begin{aligned}
P_{\boldsymbol{\psi},\lambda}(\|\boldsymbol{\mu}-\hat{\boldsymbol{\mu}}\|>\frac{\epsilon}{2}) &= \int_{\mathbb{R}_+}\int_{\|\boldsymbol{\mu}-\hat{\boldsymbol{\mu}}\|>\frac{\epsilon}{2}} \left[\frac{\kappa^{\nu}\ \exp\left(\kappa\ \boldsymbol{\mu}^T\boldsymbol{\psi}\right)}{I_\nu(\kappa)}\right]^{\lambda} \frac{1}{K_{\|\boldsymbol{\psi}\|,\lambda}} d\boldsymbol{\mu}\ d\kappa \\
&= \int_{\mathbb{R}_+}\int_{\|\boldsymbol{\mu}-\hat{\boldsymbol{\mu}}\|>\frac{\epsilon}{2}} \left[\frac{\kappa^{\nu}\ \exp\left(\kappa\,\|\boldsymbol{\psi}\|\ \boldsymbol{\mu}^T\hat{\boldsymbol{\mu}}\right)}{I_\nu(\kappa)}\right]^{\lambda} \frac{1}{K_{\|\boldsymbol{\psi}\|,\lambda}} d\boldsymbol{\mu}\ d\kappa \\
&< \int_{\mathbb{R}_+}\int_{\|\boldsymbol{\mu}-\hat{\boldsymbol{\mu}}\|>\frac{\epsilon}{2}} \left[\frac{\kappa^{\nu}\ \exp\left(\kappa\,\|\boldsymbol{\psi}\|\,(1-\frac{\epsilon^2}{8})\right)}{I_\nu(\kappa)}\right]^{\lambda} \frac{1}{K_{\|\boldsymbol{\psi}\|,\lambda}} d\boldsymbol{\mu}\ d\kappa \\
&= \int_{\mathbb{R}_+} \left[\frac{\kappa^{\nu}\ \exp\left(\kappa\,\|\boldsymbol{\psi}\|\,(1-\frac{\epsilon^2}{8})\right)}{I_\nu(\kappa)}\right]^{\lambda} \frac{1}{K_{\|\boldsymbol{\psi}\|,\lambda}} d\kappa.
\end{aligned}$$

Hence from (55) it follows that

$$P_{\boldsymbol{\psi},\lambda}(\|\boldsymbol{\mu}-\hat{\boldsymbol{\mu}}\|>\frac{\epsilon}{2}) < \int_{\mathbb{R}_+}\left\{\frac{1}{K_{\|\boldsymbol{\psi}\|,\lambda}}\left[\frac{\kappa^{\nu}\ \exp\left(\kappa\,\|\boldsymbol{\psi}\|\right)}{I_\nu(\kappa)}\right]^{\lambda}\right\}\exp(-\lambda\frac{\epsilon^2}{8}\,\|\boldsymbol{\psi}\|\,\kappa)\ d\kappa.$$

As

$$\int \frac{1}{K_{\|\boldsymbol{\psi}\|,\lambda}} \left[\frac{\kappa^{\nu}\ \exp\left(\kappa \|\boldsymbol{\psi}\|\right)}{I_{\nu}(\kappa)}\right]^{\lambda} = 1,$$

and $\epsilon > 0$, $\|\boldsymbol{\psi}\| > 0$, and $\kappa > 0$, using dominated convergence theorem we obtain that

$$\lim_{\lambda\to\infty} P_{\boldsymbol{\psi},\lambda}(\|\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}\| > \frac{\epsilon}{2}) = 0. \tag{55}$$

Therefore, it follows from (53), (54) and (55) that

$$\begin{aligned} \lim_{\lambda\to\infty} P_{\boldsymbol{\psi},\lambda}(A) &\geq \lim_{\lambda\to\infty} P_{\boldsymbol{\psi},\lambda}(B_{\epsilon}(\hat{\boldsymbol{\mu}}, \hat{\kappa})) \\ &= 1 - \lim_{\lambda\to\infty} P_{\boldsymbol{\psi},\lambda}(B_{\epsilon}(\hat{\boldsymbol{\mu}}, \hat{\kappa})^{c}) \\ &= 1. \end{aligned}$$

□

Proof of Theorem 4

Proof. 1. We have from (3.17)

$$\log(\pi(\kappa \mid \boldsymbol{\psi}, \lambda)) = 1 + \log(I_{\nu}(\lambda\kappa \|\boldsymbol{\psi}\|)) - \lambda \log(I_{\nu}(\kappa)) + (\lambda - 1)\nu \log(\kappa).$$

Therefore,

$$\begin{aligned} \frac{\partial \log(\pi(\kappa \mid \boldsymbol{\psi}, \lambda))}{\partial\kappa} &= \lambda \|\boldsymbol{\psi}\| \frac{I'_{\nu}(\lambda\kappa \|\boldsymbol{\psi}\|)}{I_{\nu}(\lambda\kappa \|\boldsymbol{\psi}\|)} - \lambda \frac{I'_{\nu}(\kappa)}{I_{\nu}(\kappa)} + \frac{(\lambda - 1)\nu}{\kappa} \\ &\overset{*}{=} \frac{\nu}{\kappa} + \lambda \|\boldsymbol{\psi}\| \frac{I_{\nu+1}(\lambda \|\boldsymbol{\psi}\| \kappa)}{I_{\nu}(\lambda \|\boldsymbol{\psi}\| \kappa)} - \lambda\frac{\nu}{\kappa} - \lambda\frac{I_{\nu+1}(\kappa)}{I_{\nu}(\kappa)} + \frac{(\lambda - 1)\nu}{\kappa} \\ &= \lambda \|\boldsymbol{\psi}\| \frac{I_{\nu+1}(\lambda \|\boldsymbol{\psi}\| \kappa)}{I_{\nu}(\lambda \|\boldsymbol{\psi}\| \kappa)} - \lambda\frac{I_{\nu+1}(\kappa)}{I_{\nu}(\kappa)}, \end{aligned}$$

where $(*)$ follows from the recursion relation in (29). By Lemma 4,

$$\frac{\partial}{\partial \kappa}\log \pi(\kappa \mid \boldsymbol{\psi}, \lambda) = \lambda \left\|\boldsymbol{\psi}\right\| \frac{I_{\nu+1}(\lambda\kappa \left\|\boldsymbol{\psi}\right\|)}{I_{\nu}(\lambda\kappa \left\|\boldsymbol{\psi}\right\|)} - \lambda\frac{I_{\nu+1}(\kappa)}{I_{\nu}(\kappa)} < \lambda\frac{I_{\nu+1}(\kappa)}{I_{\nu}(\kappa)} - \lambda\frac{I_{\nu+1}(\kappa)}{I_{\nu}(\kappa)}.$$

This term will be negative only if $\lambda||\boldsymbol{\psi}||^2 < 1$. Hence $\frac{\delta}{\delta\kappa}\log \pi(\kappa \mid \boldsymbol{\psi}, \lambda) < 0$ if $\lambda||\boldsymbol{\psi}||^2 < 1$.

2. We will prove this theorem based on Lemma 6. We have from (3.17)

$$\log(\pi(\kappa \mid \boldsymbol{\psi}, \lambda)) = 1 + \log(I_\nu(\lambda\kappa \left\|\boldsymbol{\psi}\right\|)) - \lambda\log(I_\nu(\kappa)) + (\lambda - 1)\nu\log(\kappa).$$

$$\begin{aligned}
\frac{\partial\log(\pi(\kappa \mid \boldsymbol{\psi}, \lambda))}{\partial\kappa} &= \lambda \left\|\boldsymbol{\psi}\right\| \frac{I'_{\nu}(\lambda\kappa \left\|\boldsymbol{\psi}\right\|)}{I_{\nu}(\lambda\kappa \left\|\boldsymbol{\psi}\right\|)} - \lambda\frac{I'_{\nu}(\kappa)}{I_{\nu}(\kappa)} + \frac{(\lambda-1)\nu}{\kappa} \\
&\overset{*}{=} \frac{\nu}{\kappa} + \lambda \left\|\boldsymbol{\psi}\right\| \frac{I_{\nu+1}(\lambda \left\|\boldsymbol{\psi}\right\| \kappa)}{I_{\nu}(\lambda \left\|\boldsymbol{\psi}\right\| \kappa)} - \lambda\frac{\nu}{\kappa} - \lambda\frac{I_{\nu+1}(\kappa)}{I_{\nu}(\kappa)} + \frac{(\lambda-1)\nu}{\kappa} \\
&= \lambda \left\|\boldsymbol{\psi}\right\| \frac{I_{\nu+1}(\lambda \left\|\boldsymbol{\psi}\right\| \kappa)}{I_{\nu}(\lambda \left\|\boldsymbol{\psi}\right\| \kappa)} - \lambda\frac{I_{\nu+1}(\kappa)}{I_{\nu}(\kappa)},
\end{aligned}$$

where $(*)$ follows from the recursion relation in (29). Equating above equation to 0 to calculate mode we get

$$\begin{aligned}
\left\|\boldsymbol{\psi}\right\| \frac{I_{\nu+1}(\left\|\boldsymbol{\psi}\right\| \kappa)}{I_{\nu}(\left\|\boldsymbol{\psi}\right\| \kappa)} &= \frac{I_{\nu+1}(\kappa)}{I_{\nu}(\kappa)} \\
\implies g_{\lambda\|\boldsymbol{\psi}\|}(\kappa) &= \frac{I_{\nu+1}(\kappa)I_{\nu}(\left\|\boldsymbol{\psi}\right\| \kappa)}{I_{\nu}(\kappa)I_{\nu+1}(\lambda \left\|\boldsymbol{\psi}\right\| \kappa)} = \left\|\boldsymbol{\psi}\right\|. \qquad (56)
\end{aligned}$$

By Lemma 6 we can see that (56) is increasing function of κ when $\lambda\left\|\boldsymbol{\psi}\right\|^2 \geq 1$. Further by Lemma 6, $\lim_{\kappa\to\infty} g_{\lambda\|\boldsymbol{\psi}\|}(\kappa) = 1$ and $\frac{1}{\lambda\|\boldsymbol{\psi}\|} \leq g_{\lambda\|\boldsymbol{\psi}\|}(\kappa) \leq 1$. By the same Lemma we can prove that $g_{\lambda\|\boldsymbol{\psi}\|}(\kappa) = \frac{\lambda\|\boldsymbol{\psi}\|}{\lambda}$ has the unique solution if $\frac{1}{\lambda\|\boldsymbol{\psi}\|} \leq \frac{\lambda\|\boldsymbol{\psi}\|}{\lambda} \leq 1$. Also we have proved in Theorem 4 (a) that $\pi(\kappa \mid \boldsymbol{\psi}, \lambda)$ is decreasing function of κ if $\lambda||\boldsymbol{\psi}||^2 < 1$. This proves that $\pi(\kappa \mid \boldsymbol{\psi}, \lambda)$ has unique modal point.

3.

$$\log(\pi(\kappa \mid \boldsymbol{\psi}, \lambda)) = 1 + \log(I_\nu(\lambda\kappa \|\boldsymbol{\psi}\|)) - \lambda \log(I_\nu(\kappa)) + (n-1)\nu \log(\kappa).$$

$$\begin{aligned} \frac{\partial \log(\pi(\kappa \mid \boldsymbol{\psi}, \lambda))}{\partial \kappa} &= \lambda \|\boldsymbol{\psi}\| \frac{I_{\nu+1}(\lambda \|\boldsymbol{\psi}\| \kappa)}{I_\nu(\lambda \|\boldsymbol{\psi}\| \kappa)} - \lambda \frac{I_{\nu+1}(\kappa)}{I_\nu(\kappa)} \\ &= \lambda \|\boldsymbol{\psi}\| R_\nu(\lambda \|\boldsymbol{\psi}\| \kappa) - \lambda R_\nu(\kappa) \end{aligned}$$

where $R_\nu(\kappa) = \frac{I_{\nu+1}(\kappa)}{I_\nu(\kappa)}$. Now, we derive the inflection point as the solution of equation:

$$\frac{\partial^2 \log(\pi(\kappa \mid \boldsymbol{\psi}, \lambda))}{\partial \kappa^2} = \lambda^2 \|\boldsymbol{\psi}\|^2 \; R'_\nu(\lambda \|\boldsymbol{\psi}\| \kappa) - \lambda \; R'_\nu(\kappa).$$

If κ_{in} be the inflection point then it will satisfy $\frac{\partial^2 \log(\pi(\kappa|\lambda,\boldsymbol{\psi}))}{\partial \kappa^2}|_{\kappa_{\text{in}}} = 0$. This gives us

$$\frac{R'_\nu(\lambda \;\; \|\boldsymbol{\psi}\| \kappa_{\text{in}})}{R'_\nu(\kappa_{\text{in}})} = \frac{1}{\lambda \|\boldsymbol{\psi}\|^2} = \frac{\lambda}{\lambda^2 \|\boldsymbol{\psi}\|^2}. \tag{57}$$

Let $a = \lambda \|\boldsymbol{\psi}\|$, then by Lemma 6 part (a), we know that, $g_a(\kappa_{\text{in}}) = \frac{I_{\nu+1}(\kappa_{\text{in}}) I_\nu(\lambda \;\; \|\boldsymbol{\psi}\| \kappa)}{I_\nu(\kappa) I_{\nu+1}(\lambda \|\boldsymbol{\psi}\| \kappa)} = \frac{R_\nu(\kappa)}{R_\nu(\lambda \|\boldsymbol{\psi}\| \kappa)}$ is an increasing function of κ for $\lambda \|\boldsymbol{\psi}\| > 1$. In other words we can say that

$$\begin{aligned} \frac{\partial \log(g_a(\kappa))}{\partial \kappa} &> 0, \\ \frac{\partial log[R_\nu(\kappa)]}{\partial \kappa} - \frac{\partial log[R_\nu(\lambda \|\boldsymbol{\psi}\| \kappa)]}{\partial \kappa} &> 0, \\ \frac{R'_\nu(\kappa)}{R_\nu(\kappa)} &> n \|\boldsymbol{\psi}\| \frac{R'_\nu(n \|\boldsymbol{\psi}\| \kappa)}{R_\nu(n \|\boldsymbol{\psi}\| \kappa)}, \\ \frac{R_\nu(\lambda \|\boldsymbol{\psi}\| \kappa)}{R_\nu(\kappa)} &> \lambda \|\boldsymbol{\psi}\| \frac{R'_\nu(\lambda \|\boldsymbol{\psi}\| \kappa)}{R'_\nu(\kappa)}. \end{aligned}$$

For κ_{in} we get

$$\begin{aligned}\frac{1}{g_a(\kappa_{in})} &> \lambda \left\| \boldsymbol{\psi} \right\| \frac{R'_\nu(\lambda \left\| \boldsymbol{\psi} \right\| \kappa_{in})}{R'_\nu(\kappa_{in})}, \\ \frac{1}{g_a(\kappa_{in})} &\overset{*}{>} \frac{1}{\left\| \boldsymbol{\psi} \right\|}. \end{aligned} \tag{58}$$

The last inequality follows from (57). When $\hat{\kappa}$ is a mode, then by Lemma 6 part (c) we have $g_a(\hat{\kappa}) = \frac{\lambda \|\boldsymbol{\psi}\|}{\lambda}$. If $\kappa_{\text{in}} > \hat{\kappa}$, due to increasing nature of the function in κ, we have $g_a(\kappa_{\text{in}}) > g_a(\hat{\kappa})$, i.e., $g_a(\kappa_{\text{in}}) > \lambda \left\| \boldsymbol{\psi} \right\| / \lambda$ or $\frac{1}{g_a(\kappa_{\text{in}})} < \frac{1}{\|\boldsymbol{\psi}\|}$. But this contradicts the (58). Hence if only we have $0 < \kappa_{\text{in}} < \hat{\kappa}$ then (58) will be satisfied.

□

CURRICULUM VITA
Siddhesh Kulkarni

Education

University of Louisville Louisville, KY

PhD in Biostatistics *2018-2022*

University of Connecticut Storrs, CT

Master of Science in Statistics (Emphasis: Business Analytics) *January 2018*

Savitribai Phule Pune University (Formerly University of Pune) Pune, MH, India

Master of Science in Statistics (Emphasis : Biostatistics) *May 2014*

Bachelor of Science. Major: Statistics. Minor: Physics and Mathematics *April 2012*

Relevant Research and Manuscripts

1. **Kulkarni S.**, Pal S., Gaskins J. (2022, ongoing): "A Bayesian Methodology for Estimation for Sparse Canonical Correlation."

2. **Kulkarni S.**, Pal S., Depue B., Gaskins J. (2022, ongoing): "Efficient Sampling Schemes for Bayesian Mixture of von Mises Fisher Distribution."

3. Beckerson W, Anderson J. O, **Kulkarni S.**, Yoder Himes D.: "It is About Time: Exploring the dose-dependent effects of active learning on student social personality in an upper-level biology course." (Accepted at Journal of College Science Teaching)

4. Singam, Narayana Sarma V., Bahjat AlAdili, Alok R. Amraotkar, Amanda R. Coulter, Ayesha Singh, **Siddhesh Kulkarni**, Riten Mitra, Omar Noori Daham, Allison E. Smith, and Andrew P. DeFilippis. "In-vivo platelet activation and aggregation during and after acute atherothrombotic myocardial infarction in patients with and without Type-2 diabetes mellitus treated with ticagrelor." Vascular Pharmacology (2022): 107000.

Selected Honors

- Best Presenter Award at KY ASA Spring 2022 Meeting Spring 2022
- Inducted in the national statistics honor society 'Mu Sigma Rho' for academic achievements. Summer 2021
- American Statistical Association Mary G. and Joseph Natrella Scholarship. Summer 2021
- American Statistical Association Biopharmaceutical Chapter Scholarship. Summer 2021
- National Science Foundation (Travel) Awards. Summer 2017, 2021, 2022
- J.N. Tata Endowment Fellowship for Higher Education in Fall 2016. Gift Scholarship awarded to very selective pool of J.N Tata Fellowship awardees for exceptional performance. Spring 2019
- Full scholarship to attend Summer Institute in Biostatistics at University of Washington, Seattle. Summer 2018
- Multiple travel and other funding support for different conferences.
- Won the Second Prize on all India bases in 'On the Spot Essay Writing Competition' organized by Ministry of Statistics and Program Implementation, Govt. of India. Summer 2013
- Multiple university level awards in project competitions in India.

Selected Presentations

- **A Bayesian Methodology For Estimation For Sparse Canonical Correlation**
  - Quality and Productivity Research Conference (Virtual/hybrid). Summer 2022
  - Kentucky ASA Chapter Spring 2022 Meeting. Spring 2022
  - ENAR 2022 Spring Meeting. Spring 2022

- Colloquium at Dept. of Bioinformatics and Biostatistics, University of Louisville (Invited). Fall 2021
- 34th New England Statistics Symposium at University of Rhode Island, Providence, RI. Fall 2021
- International Chinese Statistical Association's Conference' (virtual). Fall 2021
- Joint Statistical Meeting 2021 (virtual). Fall 2021
- Quality and Productivity Research Conference 2021 at Florida State University Tallahassee (Invited). Summer 2021
- Kentucky ASA Chapter Spring 2021 Meeting (virtual). Spring 2021

- **Data Augmentation Algorithm for Mixture of von Mises Fisher Distribution**

  - Bernoili-IMS One World Symposium (Virtual). Summer 2020
  - Regional Graduate Research Conference, (University of Louisville), Louisville, KY. (Finalist at poster competition). Spring 2020
  - Research Louisville! (University of Louisville), Louisville, KY. Fall 2019
  - Joint Statistical Meeting 2019, Denver, CO. Summer 2019